МИНИСТЕРСТВО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

МИНИСТЕРСТВО ПРОМЫШЛЕННОСТИ, НАУКИ И ТЕХНОЛОГИЙ
РОССИЙСКОЙ ФЕДЕРАЦИИ

РОССИЙСКАЯ АССОЦИАЦИЯ НЕЙРОИНФОРМАТИКИ

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ИНЖЕНЕРНО-ФИЗИЧЕСКИЙ ИНСТИТУТ
(ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ)

НАУЧНАЯ СЕССИЯ МИФИ–2002

НЕЙРОИНФОРМАТИКА–2002

**IV ВСЕРОССИЙСКАЯ
НАУЧНО-ТЕХНИЧЕСКАЯ
КОНФЕРЕНЦИЯ**

ЛЕКЦИИ
ПО НЕЙРОИНФОРМАТИКЕ

Часть 2

По материалам Школы-семинара
«Современные проблемы нейроинформатики»

Москва 2002

УДК 004.032.26 (06)
ББК 32.818я5
М82

НАУЧНАЯ СЕССИЯ МИФИ–2002. IV ВСЕРОССИЙСКАЯ НАУЧНО-ТЕХНИЧЕСКАЯ КОНФЕРЕНЦИЯ «НЕЙРОИНФОРМАТИКА–2002»: ЛЕКЦИИ ПО НЕЙРОИНФОРМАТИКЕ. Часть 2. – М.: МИФИ, 2002. – 172 с.

В книге публикуются тексты лекций, прочитанных на Школе-семинаре «Современные проблемы нейроинформатики», проходившей 23–25 января 2002 года в МИФИ в рамках IV Всероссийской конференции «Нейроинформатика–2002».

Материалы лекций связаны с рядом проблем, актуальных для современного этапа развития нейроинформатики, включая ее взаимодействие с другими научно-техническими областями.

Ответственный редактор
Ю. В. Тюменцев, кандидат технических наук

ISBN 5–7262–0400–X

Содержание

УДК 004.032.26 (06) Нейронные сети

УДК 004.032.26 (06) Нейронные сети

ПРЕДИСЛОВИЕ

В этой книге (она выходит в двух частях) содержатся тексты лекций, прочитанных на Школе-семинаре «Современные проблемы нейроинформатики», проходившей 23–25 января 2002 года в МИФИ в рамках IV Всероссийской научно-технической конференции «Нейроинформатика–2002».

Как и для первой Школы [1], основной целью было дать представление слушателям о современном состоянии и перспективах развития важнейших направлений, связанных с теорией и практикой нейроинформатики, ее применениями, а также с некоторыми смежными вопросами. При подготовке программы Школы особенно приветствовались лекции, лежащие по охватываемой тематике «на стыке наук», рассказывающие о проблемах не только собственно нейроинформатики (т. е. проблемах, связанных с нейронными сетями, как естественными, так и искусственными), но и о взаимодействиях нейроинформатики с другими областями мягких вычислений (нечеткие системы, генетические и другие эволюционные алгоритмы и т. п.), с системами, основанными на знаниях, с традиционными разделами математики, инженерной теории и практики. При этом изложение материала должно было строиться с таким расчетом, чтобы содержание лекции не только было бы интересным для членов нейросетевого сообщества, но и доступно более широкой аудитории, особенно студентам-старшекурсникам и аспирантам (в определенной степени моделью такого рода изложения могут служить брошюры знаменитой серии «Математика, кибернетика», выпускавшейся в течение 30 лет издательством «Знание»).

Предлагаемая подборка текстов лекций — это не учебник, охватывающий всю нейроинформатику или хотя бы значительную ее часть. Целью лекторов, приглашенных из числа ведущих специалистов в области нейроинформатики и ее приложений, было дать живую картину работы «на переднем крае» нейроинформатики, рассказать о ее взаимодействии с другими научно-техническими областями, причем сделать это, по-возможности, на примерах проблем, наиболее актуальных и активно изучаемых на данный момент.

Как и положено работам «с переднего края», каждая из них содержит, хотя и в разной степени, элементы дискуссионности. Не со всеми положениями, выдвигаемыми авторами, можно безоговорочно согласиться,

но это только повышает ценность предлагаемых материалов — они стимулируют возникновение дискуссии, поиск альтернативных ответов на поставленные вопросы, альтернативных решений сформулированных задач.

В программу Школы-семинара «Современные проблемы нейроинформатики» на конференции «Нейроинформатика–2002» вошли лекции В. Г. Редько, игумена Феофана (Крюкова), Ю. И. Нечаева, С. А. Шумского, С. А. Терехова и Н. Г. Макаренко [1].

Открывался данный цикл лекцией **В. Г. Редько** «Эволюционная кибернетика». И это было не случайно.

Наука, техника, многие другие области человеческой деятельности немыслимы без создания и исследования моделей, в том числе и такого важнейшего их класса, как модели символьные, базирующиеся на одной из знаковых систем — это и всевозможные математические и другие формальные модели, и различного рода компьютерные программы, и тексты на естественных языках, и разнообразные комбинации этих элементов.

Уже сама возможность применения символьных (в частности, математических) моделей в естественных науках, в технике, представляет собой факт достаточно нетривиальный. Вопрос можно поставить и шире, как это делается в лекции В. Г. Редько: «Почему *человеческая* логика применима к познанию *природы*?»

Эти проблемы — взаимоотношений математики и естествознания, причин применимости человеческой логики к познанию природы, и вообще — «непостижимой эффективности математики в естественных науках» (по известному выражению Юджина Вигнера) обсуждали и продолжают обсуждать многие видные ученые. Наряду с работами Ю. Вигнера, М. Клайна и А. Пуанкаре, упоминаемыми в лекции В. Г. Редько, по этим вопросам можно также рекомендовать обратиться к книгам [2–9].

Создание теоретических моделей для достаточно сложных объектов и процессов — в высшей степени непростая задача. Традиционный путь решения такой задачи состоит в получении требуемой модели сразу на заданном уровне сложности. То обстоятельство, что вначале, чаще всего, решается серия так называемых «модельных задач», сути дела не меняет, поскольку эти модельные задачи представляют собой просто усечен-

[1] Первые три из перечисленных лекций публикуются в части 1, а оставшиеся три — в части 2 сборника «Лекции по нейроинформатике.»

ные различным образом варианты основной задачи, но концептуально ее «дух» всегда остается неизменным. Базой для подобного рода процесса решения служит изучение строения требуемого объекта (процесса) и его составных частей, взаимодействия этих частей между собой, а также объекта в целом с окружающей средой (см., например, [10–12]).

Можно не углубляться в изучение внутреннего строения объекта, его «природы», а рассматривать его как «черный ящик», про который известно лишь, как он реагирует на некие представляющие интерес воздействия, возмущающие и/или управляющие. И мы получаем таким способом еще одну разновидность упомянутого выше подхода, поскольку суть дела опять же не изменилась — по-прежнему мы пытаемся получить модель объекта сразу на требуемом уровне сложности.

В значительной степени наука, а вместе с ней и инженерная теория, в течение всей своей истории развивались именно так в попытках познания мира и создания искусственных объектов.

Но есть и в последнее время довольно активно начинает развиваться другой подход, в своих концептуальных установках диаметрально противоположный первому. Он состоит в том, чтобы в качестве исходных взять некоторые очень простые модели и добавить к ним механизмы развития, позаимствованные у Природы. Тогда задача получения модели сложной системы (а в ряде случаев и самой этой системы!) сводится к «выведению», «выращиванию» такой модели эволюционным путем из модели более простой системы (или совокупности моделей простых систем).

Основное содержание лекции В. Г. Редько как раз и посвящено изложению ряда основных концепций этого (второго) направления и основной вопрос, которым задается здесь автор — «. . . нельзя ли промоделировать эволюцию познавательных способностей животных и подойти к моделированию эволюционного возникновения интеллекта?»

Эволюционное направление как в создании моделей систем, так и самих систем представляется весьма перспективным и многообещающим. Оно открывает возможность заменить процесс создания модели сразу как целого процессом подготовки некоторой «затравки», на которую «напускаются» механизмы эволюционного развития. Такой путь может оказаться перспективным с точки зрения преодоления пресловутого «порога сложности», возникающего при создании систем.

Нельзя сказать, что данному направлению раньше совсем не уделялось внимания. Напротив, предыстория его довольно богата.

Известно, что первые вычислительные машины появились в связи с потребностями выполнения больших объемов вычислений, например, в баллистике, авиационной и ракетной технике, атомной технике и др.

Но уже с самого начала, примерно с середины 50-х годов, ЭВМ пытались использовать не только для проведения расчетов, но и для моделирования интеллектуальных систем. Уже тогда сформировались основные направления работ в этой области, существующие и в настоящее время.

Сразу же сформировалось два конкурирующих направления исследований, получивших наименование нисходящего и восходящего подходов.

Сторонники *нисходящего подхода* пытались воспроизводить (моделировать) достаточно сложные интеллектуальные операции и виды деятельности (игры — шашки, шахматы; доказательство теорем; поиск решений и т. п.). Работы в этом направлении привели, в частности, к появлению экспертных систем и, шире, систем, основанных на знаниях (см., например, [14–16])).

Исследователи, работавшие в рамках *восходящего подхода*, пытались идти от простых аналогов нервной системы примитивных существ с очень малым числом нейронов к сложнейшей нервной системе человека. Это направление привело, в частности, к появлению обширного класса моделей, именуемых искусственными нейронными сетями (см., например, [15, 17–20]; см. также «тему номера» в журнале «Компьютерра» [21]).

Но тогда же, практически одновременно с упомянутыми двумя, возник еще и третий подход к созданию интеллектуальных систем, называемый *эволюционным программированием*. Целью его было, как отмечал А. Г. Ивахненко в предисловии к русскому переводу книги [22] (оригинал ее был издан в 1966 году), «заменить процесс моделирования человека моделированием процесса его эволюции».

Ранняя история данного направления связана с работами Л. Фогеля и его сотрудников [22] по сообществам эволюционирующих конечных автоматов (в определенной степени развитием работ данного направления стали книги [23–25]), работами 60-х годов М. Л. Цетлина по моделям автоматов, адаптивно приспосабливающихся к окружающей среде, а также работы 60–70-х годов М. М. Бонгарда по адаптивному поведе-

нию искусственных организмов на плоскости, разбитой на клетки [2]. Наряду с этими работами следует также упомянуть активное обсуждение проблемы «Автоматы и жизнь», проходившее в 60-е годы с участием таких видных отечественных и зарубежных ученых, как Н. М. Амосов, И. И. Артоболевский, Н. Винер, В. М. Глушков, А. А. Дородницын, А. Г. Ивахненко, А. Е. Кобринский, А. Н. Колмогоров, У. Р. Эшби и др. Спектр мнений по данной проблеме был самый широкий — от безудержного оптимизма («Только автомат? Нет, мыслящее существо!») до полнейшего пессимизма («Машина не может жить, плесень не способна мыслить!») [3]. Некоторые материалы дискуссии «Автоматы и жизнь» (статьи и доклады разных лет) содержатся в сборнике [30].

В тот же период времени начались исследования по такой сложнейшей проблеме, как *самовоспроизводящиеся искусственные системы*; одними из первых здесь были работы Дж. фон Неймана по самовоспроизводящимся автоматам [31].

Идейно близки к перечисленным работам и быстро развивающиеся сейчас направления — генетические алгоритмы, генетическое программирование, эволюционные вычисления [26–29].

Идеи и методы эволюционного моделирования активно использовались в возникшем в конце 80-х годов интереснейшем направлении, именуемом «Искусственная жизнь» (Artificial Life, или просто ALife), основные элементы которого также рассматриваются в лекции В. Г. Редько.

Обсуждение ряда элементов ALife есть в тематическом разделе («теме номера») журнала «Компьютерра» [32]. В одной из статей этого номера рассказывается об эволюционном процессе, реализованном аппаратно — на уровне электронных микросхем. Здесь же содержится целый ряд ссылок по теме ALife на ресурсы Интернет.

В лекции В. Г. Редько приводится целый ряд примеров модельной реализации идей ALife на программном или аппаратном уровне. Список этот, разумеется, не может претендовать на исчерпывающую полноту.

Хотелось бы обратить внимание читателей на один достаточно показательный пример, не вошедший в этот список.

[2] Ссылки на работы М. Л. Цетлина и М. М. Бонгарда можно найти в лекции В. Г. Редько и списке литературы к ней.

[3] Заголовки разделов в сборнике [30].

Речь идет о работах Марка Тилдена (Mark W. Tilden) из Лос-Аламосской национальной лаборатории США (Los Alamos National Laboratory) по направлению, которое он называет «Живые машины». Русский перевод (в сокращении) одной из статей М. Тилдена (совместно с Б. Хасслахером) был опубликован в журнале «Природа» [33].

М. Тилден с сотрудниками построили около сотни действующих образцов «биоморфных машин» («биоморфов», или «жизнеподобных»), главная задача которых — преодолевать незнакомые сложные ландшафты в поисках «пищи». Управляющее ядро этих машин представляет собой аналоговую нейросеть осцилляторного типа с очень небольшим числом нейронов в ней (как правило, менее десятка). Эти машины продемонстрировали очень высокую приспосабливаемость к меняющемуся рельефу местности.

Кроме статьи [33], информацию о работах М. Тилдена можно найти по адресам Интернет, перечисленным под номером [34] в списке литературы в конце предисловия. Среди этих ресурсов можно найти патент М. Тилдена на нейросеть, используемую им в биоморфных машинах.

Пересказывать содержание этой многоплановой и интересной лекции здесь нет никакой необходимости, укажем лишь ряд дополнительных источников, с помощью которых можно более глубоко проработать затронутые в лекции вопросы.

Различные аспекты зарождения и развития жизни на Земле, общие законы функционирования живого освещаются в книгах [35–43]. Принципы биологической эволюции, ее механизмы и модели рассматриваются в книгах [44–69]. Об эволюционном возникновении интеллекта можно прочитать в книгах [70, 71], об организации психики человека, происхождении, формировании и развитии высших потребностей познания — в книгах [72, 73]. Попытка мысленно представить эволюционное возникновение иерархии биологических систем управления сделана в прекрасной книге В. Ф. Турчина [13].

Общая схема адаптивного поведения, рассматриваемая В. Г. Редько, основывается на *функциональной системе*, разработанной советским нейрофизиологом П. К. Анохиным [74]. Функциональная система характеризует такие свойства схемы управления поведением, как целенаправленность, мотивацию для формирования цели, доминанту по А. А. Ухтомскому для мобилизации ресурсов животного на достижение приоритетной

цели (в том числе и мобилизацию интеллектуальных ресурсов — концентрацию внимания), а также ряд других.

Как показано в лекции **игумена Феофана (Крюкова)** «Модель внимания и памяти, основанная на принципе доминанты», важнейшая роль в этом перечне свойств принадлежит доминанте.

В лекции описаны *шесть основных проблем внимания*: проблема селективности стимулов (почему из нескольких одновременно предъявленных стимулов одни привлекают внимание и получают таким образом доступ к высшей сенсорной обработке, а другие не получают?); проблема долговременной памяти (каков механизм взаимодействия внимания и долговременной памяти?); проблема интеграции (как и где происходит реконструкция интегрального образа для стимулов, обрабатывавшихся параллельно?); проблема инерции (какова основа сохранения длительного внимания в случаях, когда стимулы предъявляются кратковременно?); проблема торможения и подавления помех (что происходит со стимулами, которым не оказывается внимания?); проблема Центрального Управителя (существует ли отдельная структура для координации процессов внимания и памяти или же здесь работают процессы самоорганизации новой коры?).

В лекции показано, что на основе принципа *доминанты А. А. Ухтомского* удается найти ответы на все шесть перечисленных выше вопросов. Показано, что в основе учения о доминанте лежит физическое явление фазовых переходов, а также трактовка нейронной сети как системы связанных нелинейных осцилляторов. Приводятся доказательства того, что неравновесные фазовые переходы действительно происходят в мозге.

Нейрофизиологический материал, необходимый для понимания материала лекции игумена Феофана (Крюкова), можно почерпнуть, например, в общем курсе биологии [39], а также в книгах [75,76]. Об исследованиях мозга говорится в книгах [77,78]. О связях высшей нервной деятельности с психологией рассказывается в книге [79], здесь рассматривается и роль доминанты А. А. Ухтомского для понимания процессов высшей нервной деятельности.

На важность и перспективность использования в обработке информации *колебательных моделей*, включая и колебательные (осцилляторные) нейронные сети, автор данной лекции обращал внимание нейросетевого сообщества в ходе «Дискуссии о нейрокомпьютерах», состоявшейся

в рамках конференции «Нейроинформатика–99» (см. [80], с. 29–33, выступление В. И. Крюкова). Им утверждалось, в частности, что «. . . материальным носителем биологической памяти, если таковой существует, является не синаптическая система, а скорее целостная нервная ткань, как это предсказывается, исходя из принципа доминанты».

Того же мнения о значимости колебательных нейронных сетей придерживается и Р. М. Борисюк, который на той же самой дискуссии в ответе на вопросы о наиболее значительных достижениях в теории нейронных сетей и в понимании работы мозга, полученных в течение 90-х годов (см. [80], с. 13–16) отметил: «Одним из основных достижений можно считать создание теории осцилляторных нейронных сетей и демонстрацию того, что принцип синхронизации нейронной активности является важным принципом обработки информации в структурах мозга. Детальная разработка этой теории, имеющей глубокие корни в работах выдающегося физиолога А. А. Ухтомского, была начата в нашей стране В. И. Крюковым, а на Западе в работах К. фон-дер Мальсбурга (Christoph von der Malsburg). Дальнейшее развитие теории показало, что на основе принципа синхронизации можно решать задачи распознавания образов, запоминания информации, интеграции признаков объекта в цельный образ, формирования и управления фокусом внимания и др.».

Вопросам, связанным с осцилляторными нейронными сетями, постоянно уделялось внимание и на конференциях «Нейроинформатика» (см. [81–87]).

Здесь уместно будет отметить, что работы М. Тилдена по «живым машинам», упоминавшиеся выше, также основываются на использовании осцилляторных нейронных сетей.

В лекции игумена Феофана (Крюкова) в противовес традиционной коннекционистской архитектуре нейросетевых систем предлагается доминантная архитектура обработки информации в мозге. Кроме того, в ней ставится вопрос о неудовлетворительности существующей концептуальной базы (парадигмы [4]) нейроинформатики и делается вывод о необходимости смены этой парадигмы: «Почти все теоретики мозга ищут

[4]Концепция *парадигмы* в науке была сформулирована Томасом Куном в начале 60-х годов: «. . . Под парадигмами я подразумеваю признанные всеми научные достижения, которые в течение определенного времени дают научному сообществу модель постановки проблем и их решений (см. [88], с. 11)». Смена одной парадигмы на другую трактуется Т. Куном как *научная революция*.

не истину, а подтверждения хеббовской программы, приняв гипотезу за незыблемый факт. А истина лежит совсем в другом месте — в учении А. А. Ухтомского о доминанте».

В лекции В. Г. Редько отмечается, что удивительная эффективность функционирования живых организмов, гармоничность и согласованность работы органов («компонент») живых существ обеспечивается биологическими управляющими системами. Относительно этих систем возникает целый ряд вопросов, в том числе и такой важнейший, как пути возникновения интеллекта.

Другой аспект этой же проблемы рассматривался в лекции игумена Феофана (Крюкова), где показано, как на основе принципа доминанты А. А. Ухтомского можно адекватно моделировать такие, не менее важные, свойства живых существ, как память и внимание.

Но ведь управляющие системы встречаются не только в живых системах, но и в системах, создаваемых человеком, они являются важнейшим элементом, определяющим в значительной мере уровень возможностей той или иной системы.

Лекция **Ю. И. Нечаева** «Нейросетевые технологии в бортовых интеллектуальных системах реального времени» посвящена вопросам создания управляющих систем именно такого рода, а также систем анализа и интерпретации измерительной информации о поведении динамического объекта.

Эта лекция представляет собой один из примеров того междисциплинарного подхода, что упоминался выше как весьма желательный для Школы-семинара.

Предметом рассмотрения в лекции Ю. И. Нечаева являются *бортовые интеллектуальные системы*, обеспечивающие управление динамическим объектом, идентификацию экстремальных ситуаций, оценку параметров динамического объекта и внешней среды.

Эти задачи решаются с привлечением целого ряда новых подходов, в число которых входят: геометрическая интерпретация динамических моделей на основе теории хаотических систем и принципов самоорганизации; нейросетевые технологии; методы построения систем, основанных на знаниях; методы нечеткой (размытой) логики и нечетких систем; методы теории возможностей; эволюционное моделирование (генетические алгоритмы и т. п.); различные комбинированные технологии (нейро-

нечеткие, нейро-генетические и т. д.).

Целесообразность применения этой совокупности методов и средств, взаимодействие их между собой, последовательно демонстрируется на конкретных примерах задач для динамических объектов, таких как управление движением подводного аппарата, идентификация экстремальных ситуаций для плавучих динамических объектов, оценка динамических характеристик объекта и внешней среды, создание интеллектуальных нейросетевых датчиков.

В лекции Ю. И. Нечаева показано, что сложности, присущие традиционным подходам к созданию бортовых измерительных и управляющих систем, могут быть в значительной мере преодолены, если воспользоваться технологиями мягких вычислений (включая нейросети, нечеткие системы, генетические алгоритмы и т.п.). Рациональное использование этих технологий позволяет обеспечить измерительным и управляющим системам гибкость и способность адаптироваться к изменяющимся условиям внешней и внутренней среды динамического объекта.

Дополнительные сведения по затронутым в лекции Ю. И. Нечаева вопросам можно получить в следующих книгах: по нелинейной динамике, хаотическим системам, самоорганизации — в [90–103] (см. также журнал «Компьютерра» [89] с темой номера «Хаос»); по системам, основанным на знаниях — в [14–16]; по нечеткой логика, нечетким системам — в [104–113] (см. также журнал «Компьютерра» [114] с темой номера «Нечеткая логика»); теория возможностей — в [115–117]; по нейросетевым технологиям — в [15, 17–21]; по генетическим алгоритмам, эволюционному моделированию — в [22–28]; по смешанным технологиям мягких вычислений — в [28, 29]; по информационной обработке и управлению на основе технологий мягких вычислений — в [118–126]. Значительное число программ и публикаций по таким темам, как искусственные нейронные сети, нечеткие системы, генетические алгоритмы, а также их применениям можно найти через портал научных вычислений, адрес которого содержится в позиции [127] списка литературы к предисловию.

В начале данного предисловия было сказано о двух диаметрально противоположных подходах к построению моделей систем — традиционном и эволюционном. Эти два подхода вовсе не исключают, а скорее дополняют друг друга.

Примерами, основанными на традиционном подходе являются и лекция игумена Феофана (Крюкова), и лекция Ю. И. Нечаева. А именно, в лекции игумена Феофана (Крюкова) используется подход, типичный для науки: изучение объекта, его особенностей и т. п. В лекции Ю. И. Нечаева наряду с данным вариантом широко применяется и подход типа «черный ящик», реализующийся в искусственных нейросетях (но здесь широко используются и обычные математические модели движения динамических объектов, записанные в виде систем дифференциальных уравнений).

Еще дальше идет **С. А. Шумский** в своей лекции «Байесова регуляризация обучения». В ней речь идет о системе типа «черный ящик», для которой есть только некие описывающие ее эмпирические данные.

Рассматривается задача *машинного обучения*, цель решения которой — выявление закономерностей в эмпирических данных.

Как отмечает С. А. Шумский: «В противоположность математическому моделированию, изучающему следствия из известных законов, машинное обучение стремится воссоздать причины, наблюдая порожденные ими следствия — эмпирические данные».

Отсюда следует, что рассматриваемая задача относится к классу обратных задач, которые в общем случае являются плохо определенными или некорректными. Вследствие повышенной чувствительности некоторых из решений таких задач к данным, для нахождения устойчивых решений приходится применять процедуру так называемой *регуляризации*, которая приводит к ограничению класса допустимых решений.

При этом надо, с одной стороны, не потерять чувствительность к данным, чтобы оставалась возможность объяснения всех имеющихся фактов, а с другой — не переусложнить модель так, что она станет реагировать не только на требуемую закономерность, но и на случайные события в обучающей выборке. Или, как замечает С. А. Шумский, «пройти между Сциллой переупрощения и Харибдой переусложнения».

В лекции С. А. Шумского подробно рассматривается один из наиболее эффективных способов решения этой проблемы — *байесова регуляризация*, основанная не на оценке ожидаемой ошибки, как это принято в традиционных методах математической статистики, а на выборе наиболее правдоподобной (с учетом имеющихся данных) модели.

Иллюстрируется данный подход на задачах оценки параметров, интерполяции функций и кластеризации; одна из практически интересных задач здесь — определение рационального числа элементов в скрытом слое искусственной нейросети.

По теме лекции С. А. Шумского можно рекомендовать следующую дополнительную литературу: некорректные задачи и регуляризация — [128, 129]; традиционная математическая статистика — [130–133]; байесовский подход [134] (здесь управление трактуется как процесс обучения, подробно рассматривается теорема Байеса и ее применение).

Есть задачи, они особенно часто встречаются в ряде областей численного анализа и оптимизации, для решения которых есть, казалось бы, все необходимое — теоретическая база, алгоритмы, даже компьютерные программы. Но тем не менее, решение почти каждой такой задачи представляет собой «штучную работу», в значительной степени опирающуюся на ранее полученный опыт решения аналогичных задач.

Пример решения именно такого рода задачи демонстрируется в лекции **С. А. Терехова** «Нейросетевые аппроксимации плотности распределения вероятности в задачах информационного моделирования». Здесь, как и в лекции С. А. Шумского, изучается проблема построения эмпирических моделей на основе числовых данных. При этом рассматривается обучение без учителя на примерах, в условиях неопределенности в характере модели.

Эта задача аппроксимации плотности распределения вероятности, описывающего множество многомерных экспериментальных данных.

К такой постановке сводятся многие важные прикладные задачи: задача распознавания образов, проблема заполнения пропусков в таблицах данных, вероятностный прогноз и т. п.

В лекции С. А. Терехова дается сопоставление нескольких подходов к аппроксимации плотности распределения, в числе которых параметрические методы аппроксимации и методы непараметрической статистики. Рассматриваются также *байесовы сети*, представляющие собой одно из наиболее важных достижений последнего десятилетия в области искусственного интеллекта.

В качестве еще одного подхода предлагается заменить задачу аппроксимации эквивалентной ей задачей классификации. Здесь опять возникает проблема регуляризации, о которой, хотя и в несколько ином плане

говорилось в лекции С. А. Шумского.

Дополнительную информацию по затронутым в лекции С. А. Терехова вопросам можно получить из книг [130–133] (математическая статистика), а также [15, 17–20] (искусственные нейросети и их применений). Популярное изложение материала о байесовых сетях, а также пакет расширения (Bayes Net Toolbox) для Matlab содержится по адресам, указанным в позиции [135] списка литературы к предисловию.

Наряду с лекцией Ю. И. Нечаева, лекция **Н. Г. Макаренко** «Фракталы, аттракторы, нейронные сети и все такое» представляет собой яркий образец междисциплинарного подхода. Ценность его — в демонстрации глубоких взаимосвязей между различными областями науки, в том числе и такими, что возникли и развивались вначале совершенно независимо друг от друга.

Изложение в лекции Н. Г. Макаренко начинается с изложения концепции дробной размерности и фрактала. Затем вводятся системы итеративных функций в пространстве компактов.

Изучение предельной динамики систем итеративных функций ведет к теории дискретных динамических систем. Далее показано, что процесс аппроксимации аттрактора системы итеративных функций эквивалентен работе бинарной нейронной сети.

Как замечает Н. Г. Макаренко: «Таким образом, термины “фрактал” в геометрии и “странный аттрактор” в динамике оказываются синонимами, а систему итеративных функций (СИФ) можно рассматривать как рекуррентную асимметричную нейросеть. С другой стороны, Фернандо Ниньо в 2000 году установил, что случайная итеративная нейронная сеть (гипернейрон) топологически эквивалентна динамической системе с заданным аттрактором. Круг замкнулся, образовав Единый Контекст, объединяющий *фракталы*, *СИФ*, *аттракторы* и *нейронные сети*. Цель лекции — показать взаимную связь этих предметов, потому что *единое лучше, чем всё вместе, но по-отдельности*».

Дополнительные сведения по фракталам можно найти в книгах [136, 137], по динамическим системам — в книгах [138–141].

* * *

Как это уже было в [1], помимо традиционного списка литературы каждая из лекций сопровождается списком интернетовских адресов, где можно найти информацию по затронутому в лекции кругу вопросов, включая и

дополнительные ссылки, позволяющие расширить, при необходимости, зону поиска.

Вызвано это тем, что ссылки в списке литературы на традиционные «письменные» источники обычно трудно «разрешимы», материалы, на которые они указывают, в современной ситуации мало доступны, особенно вне столиц. В то же время, в Интернете можно найти сейчас информацию практически по любой тематике, часто — те же статьи, которые включены в список литературы — надо только знать, где их искать. Включение в лекции ссылок на интернетовские ресурсы дает подобного рода сведения тем, кто заинтересуется соответствующей тематикой и захочет более подробно разобраться в ней. Учитывая все расширяющиеся возможности доступа к Интернету, это обеспечивает доступ к разнообразным данным практически всем желающим.

Перечень проблем нейроинформатики и смежных с ней областей, требующих привлечения внимания специалистов из нейросетевого и родственных с ним сообществ, далеко не исчерпывается, конечно, вопросами, рассмотренными в предлагаемом сборнике.

В дальнейшем предполагается расширение данного списка за счет рассмотрения насущных проблем собственно нейроинформатики, проблем «пограничного» характера, особенно относящихся к взаимодействию нейросетевой парадигмы с другими парадигмами, развиваемыми в рамках концепции мягких вычислений, проблем использования методов и средств нейроинформатики для решения различных классов прикладных задач. Не будут забыты и взаимодействия нейроинформатики с такими важнейшими ее «соседями», как нейробиология, нелинейная динамика (синергетика — в первую очередь), численный анализ (вейвлет-анализ и др.) и т.п.

Замечания, пожелания и предложения по содержанию и форме лекций, перечню рассматриваемых тем и т.п. просьба направлять электронной почтой по адресу tium@mai.ru Тюменцеву Юрию Владимировичу.

Литература

1. *Лекции по нейроинформатике*: По материалам Школы-семинара «Современные проблемы нейроинформатики» // III Всероссийская научно-техническая конференция «Нейроинформатика-2001», 23–26 января 2001 г. / Отв. ред. *Ю. В. Тюменцев*. – М.: Изд-во МИФИ, 2001. – 212 с.

2. *Адамар Ж.* Исследование психологии процесса изобретения в области математики: Пер. с франц. – М.: Сов. радио, 1970. – 152 с.

3. *Блехман И. И., Мышкис А. Д., Пановко Я. Г.* Механика и прикладная математика: Логика и особенности приложений математики. 2-е изд., испр. и доп. – М.: Наука, 1990. – 360 с.

4. *Вейль Г.* Математическое мышление: Сб. статей: Пер. с англ. и нем. – М.: Наука, 1989. – 400 с.

5. *Кац М., Улам С.* Математика и логика: Ретроспектива и перспективы: Пер. с англ. – М.: Мир, 1971. – 251 с. (Серия «Современная математика: Популярная серия»)

6. *Клайн М.* Математика: Утрата определенности: Пер. с англ. – М.: Мир, 1984. – 434 с.

7. *Курант Р., Роббинс Г.* Что такое математика? Элементарный очерк идей и методов: Пер. с англ., 3-е изд. – Ижевск: НИЦ «Регулярная и хаотическая динамика», 2001. – 592 с.

8. *Пойа Д.* Математическое открытие: Решение задач — основные понятия, изучение и преподавание: Пер. с англ. – М.: Наука, 1970. – 452 с.

9. *Пойа Д.* Математика и правдоподобные рассуждения: Пер. с англ. 2-е изд., испр. – М.: Наука, 1975. – 464 с.

10. *Калашников В. В.* Сложные системы и методы их анализа. – М.: Знание, 1980. – 64 с. (Новое в жизни, науке, технике. Серия «Математика, кибернетика», вып. 9, 1980)

11. *Калашников В. В., Немчинов Б. В., Симонов В. М.* Нить Ариадны в лабиринте моделирования. – М.: Наука, 1993. – 192 с. (Серия «Кибернетика: неограниченные возможности и возможные ограничения»)

12. *Шрейдер Ю. А., Шаров А. А.* Системы и модели. – М.: Радио и связь, 1982. – 152 с. (Серия «Кибернетика»)

13. *Турчин В. Ф.* Феномен науки: Кибернетический подход к эволюции. 2-е изд. – М.: ЭТС, 2000. – 368 с.

14. *Нильсон Н.* Принципы искусственного интеллекта: Пер. с англ. – М.: Радио и связь, 1985. – 376 с.

15. Компьютер обретает разум: Пер. с англ. Под ред. *В. Л. Стефанюка*. – М.: Мир, 1990. – 240 с.

16. Будущее искусственного интеллекта / Ред.-сост. *К. Е. Левитин* и *Д. А. Поспелов*. – М.: Наука, 1991. – 302 с.

17. *Горбань А. Н., Россиев Д. А.* Нейронные сети на персональном компьютере. – Новосибирск: Наука, 1996. – 276 с.

18. *Нейрокомпьютер как основа мыслящих ЭВМ*: Сб. науч. статей / Отв. ред. *А. А. Фролов* и *Г. И. Шульгина*. – М.: Наука, 1993. – 239 с.

19. *Уоссерман Ф.* Нейрокомпьютерная техника: Теория и практика: Пер. с англ. – М.: Мир, 1992. – 240 с.

20. *Ежов А. А., Шумский С. А.* Нейрокомпьютинг и его приложения в экономике и бизнесе. – М.: МИФИ, 1998. – 222 с.

21. *Нейросети* (тема номера, 4 статьи) // Компьютерра. – №4 (333), 8 февраля 2000 г. – с. 19–31.
URL: http://www.computerra.ru/offline/2000/333/

22. *Фогель Л., Оуэнс А., Уолш М.* Искусственный интеллект и эволюционное моделирование: Пер. с англ. – М.: Наука, 1969. – 231 с.

23. *Букатова И. Л.* Эволюционное моделирование и его приложения. – М.: Наука, 1979. – 231 с.

24. *Букатова И. Л.* Эволюционное моделирование: Идеи, основы теории, приложения. – М.: Знание, 1981. – 64 с. (Новое в жизни, науке, технике. Серия «Математика, кибернетика», вып. 10, 1981)

25. *Букатова И. Л., Михасев Ю. И., Шаров А. М.* Эвоинформатика: Теория и практика эволюционного моделирования. – М.: Наука, 1991. – 206 с.

26. Special Issue *"Evolutionary Computations"* / Ed.: *David B. Fogel* and *Lawrence J. Fogel* // IEEE Transactions on Neural Networks. – January 1994. – v. 5, No. 1. – pp. 1–147.

27. Special Issue *"Genetic Algorithms"* / Eds.: *Anup Kumar* and *Yash P. Gupta* // Computers and Operations Research. – January 1995. – v. 22, No. 1. – pp. 3–157.

28. Special Issue *"Artificial Intelligence, Evolutionary Programming and Operations Research"* / Eds.: *James P. Ignizio* and *Laura I. Burke* // Computers and Operations Research. – June 1996. – v. 23, No. 6. – pp. 515–622.

29. Special Issue *"Neuro-Fuzzy Techniques and Applications"* Eds.: *George Page* and *Barry Gomm* // Fuzzy Sets and Systems: Intern. J. of Soft Computing and Intelligence. – Apr. 8, 1996. – v. 79, No. 1. – pp. 1–140.

30. Кибернетика: Итоги развития / Ред.-сост.: *В. Д. Пекелис*. – М.: Наука, 1979. – 200 с. (Серия «Кибернетика: неограниченные возможности и возможные ограничения»)

31. *фон Нейман Дж.* Теория самовоспроизводящихся автоматов: Пер. с англ. – М.: Мир, 1971. – 382 с.

32. *Искусственная жизнь* (тема номера, 5 статей) // Компьютерра. – № 11 (289), 16 марта 1999 г. – с. 17–31.
URL: http://www.computerra.ru/offline/1999/289/

33. *Хасслахер Б., Тилден М.* Живые машины // Природа. – 1995. – № 4. – с. 32–46. Это сокращенный русский вариант статьи: *B.Ḧasslacher* and *M. W. Tilden*. Living machines // Robotics and Autonomous Systems. – 1995. – v. 15. – pp. 143–169.

34. Ресурсы Интернет, касающиеся работ М. Тилдена:
• информация о действующих образцах биоморфных машин:
URL: http://biosat.lanl.gov/
URL: http://cism.jpl.nasa.gov/biocomputing/workshop
URL: http://citeseer.nj.nec.com/6446.html
URL: http://www.beam-online.com/Robots/Galleria_other/tilden.html
URL: http://www.geocities.com/SouthBeach/6897/beam2.html
• патент на нейросеть, применяемую в биоморфных машинах:
URL: http://microcore.solarbotics.net/patent.html
• популярное объяснение ее устройства:
URL: http://bftgu.solarbotics.net/starting_nvnet.html
• нейроконтроллера на ее основе:
URL: http://biosat.lanl.gov/pubs/SPIE/ABSTRACT_SPIE_19981.html
• а также пример применения в шагающем роботе-жуке:
URL: http://tnewton.solarbotics.net/robot2.html
URL: http://www.iguana-robotics.com/RobotUniverse/BiomorphicRobots.htm
• Здесь — большое интервью с М. Тилденом:
URL: http://fargo.itp.tsoa.nyu.edu/∼kevin/tilden/

35. *Ичас М.* О природе живого: Механизмы и смысл: Пер. с англ. – М.: Мир, 1994. – 496 с.

36. *Медников Б. М.* Аксиомы биологии: Biologia axiomatica. – М.: Знание, 1982. – 136 с. (Серия «Наука и прогресс»)

37. *Рьюз М.* Философия биологии: Пер. с англ. – М.: Прогресс, 1977. – 319 с.

38. *Чернов Г. Н.* Законы теоретической биологии. – М.: Знание, 1990. – 64 с. (Новое в жизни, науке, технике. Серия «Биология», вып. 1, 1990)

39. *Вилли К., Детье В.* Биология: Биологические процессы и законы: Пер. с англ. – М.: Мир, 1975. – 822 с.

40. *Кемп П., Армс К.* Введение в биологию: Пер. с англ. – М.: Мир, 1988. – 671 с.

41. *Сингер М., Берг П.* Гены и геномы. В двух томах. Том 1: Пер. с англ. – М.: Мир, 1998. – 373 с.

42. *Сингер М., Берг П.* Гены и геномы. В двух томах. Том 2: Пер. с англ. – М.: Мир, 1998. – 391 с.

43. *Франк-Каменецкий М. Д.* Самая главная молекула. – М.: Наука, 1983. – 160 с. (Библиотечка «Квант». Вып. 25)

44. *Антонов А. С.* Генетические основы эволюционного процесса. – М.: Знание, 1983. – 64 с. (Новое в жизни, науке, технике. Серия «Биология», вып. 4, 1983)

45. *Кайданов Л. З.* Генетика популяций. – М.: Высшая школа, 1996. – 320 с.

46. *Кейлоу П.* Принципы эволюции: Пер. с англ. – М.: Мир, 1986. – 128 с.

47. *Арена биологической эволюции*: Сборник. – М.: Знание, 1986. – 64 с. (Новое в жизни, науке, технике. Серия «Биология», вып. 6, 1986)

48. *Бердников В. А.* Эволюция и прогресс. – М.: Наука, 1991. – 192 с. (Серия «Человек и окружающая среда»)

49. *Борзенков В. Г.* Философские основания теории эволюции. – М.: Знание, 1987. – 64 с. (Новое в жизни, науке, технике. Серия «Биология», вып. 1, 1987)

50. *Георгиевский А. Б., Попов Е. Б.* «Белые пятна» эволюции. – М.: Просвещение, 1987. – 96 с. (Серия «Мир знаний»)

51. *Голубев В. С.* Эволюция: От геохимических систем до ноосферы. – М.: Наука, 1992. – 110 с. (Серия «Человек и окружающая среда»)

52. *Горбань А. Н., Хлебопрос Р. Г.* Демон Дарвина: Идея оптимальности и естественный отбор. – М.: Наука, 1988. – 208 с. (Серия «Проблемы науки и технического прогресса»)

53. *Грант В.* Эволюция организмов: Пер. с англ. – М.: Мир, 1980. – 407 с.

54. *Грант В.* Эволюционный процесс: Критический обзор эволюционной теории: Пер. с англ. – М.: Мир, 1991. – 488 с.

55. *Докинз Р.* Эгоистичный ген: Пер. с англ. – М.: Мир, 1993. – 318 с.

56. *Камшилов М. М.* Эволюция биосферы. 2-е изд., доп. – М.: Наука, 1979. – 256 с. (Серия «Человек и окружающая среда»)

57. *Лима-де-Фариа А.* Эволюция без отбора: Автоэволюция формы и функции: Пер. с англ. – М.: Мир, 1991. – 455 с.

58. *Моран П.* Статистические процессы эволюционной теории: Пер. с англ. – М.: Наука, 1973. – 288 с.

59. *Назаров В. И.* Финализм в современном эволюционном учении. – М.: Наука, 1984. – 284 с.

60. *Нейфах А. А., Лозовская Е. Р.* Гены и развитие организма. – М.: Наука, 1984. – 188 с. (Серия «От молекул до организма»)

61. *Пианка Э.* Эволюционная экология: Пер. с англ. – М.: Мир, 1981. – 400 с.

62. Проблемы теории молекулярной эволюции / *В. А. Ратнер, А. А. Жарких, Н. А. Колчанов, С. Н. Родин, В. В. Соловьев, В. В. Шамин*. Отв. ред. *Р. И. Салганик*. – Новосибирск: Наука, 1985. – 263 с.

63. *Северцов А. С.* Основы теории эволюции. – М.: Изд-во МГУ, 1987. – 320 с.

64. *Скворцов А. К.* Микроэволюция и пути видообразования. – М.: Знание, 1982. – 64 с. (Новое в жизни, науке, технике. Серия «Биология», вып. 9, 1982)

65. *Солбриг О., Солбриг Д.* Популяционная биология и эволюция: Пер. с англ. – М.: Мир, 1982. – 488 с.

66. *Татаринов Л. П.* Палеонтология и эволюционное учение. – М.: Знание, 1985. – 64 с. (Новое в жизни, науке, технике. Серия «Биология», вып. 9, 1985)

67. *Татаринов Л. П.* Эволюция и креационизм. – М.: Знание, 1988. – 64 с. (Новое в жизни, науке, технике. Серия «Биология», вып. 8, 1988)

68. Эволюция: Сборник: Пер. с англ. под ред. *М. В. Мины*. – М.: Мир, 1981. – 265 с.

69. *Яблоков А. В., Юсуфов А. Г.* Эволюционное учение: Дарвинизм. 4-е изд., стер. – М.: Высшая школа, 1998. – 336 с.

70. *Кликс Ф.* Пробуждающееся мышление: У истоков человеческого интеллекта. Пер. с нем. – М.: Прогресс, 1983. – 302 с.

71. *Сергеев Б. Ф.* Ступени эволюции интеллекта. – Л.: Наука, 1986. – 192 с. (Серия «От молекулы до организма»)

72. *Веккер Л. М.* Психика и реальность: Единая теория психических процессов. – М.: Смысл, 2000. – 685 с.

73. *Симонов П.В., Ершов П.М., Вяземский Ю.П.* Происхождение духовности – М.: Наука, 1989. – 352 с. (Серия «Общество и личность»)

74. *Анохин П. К.* Системные механизмы высшей нервной деятельности. – М.: Наука, 1979. – 453 с.

75. *Алейникова Т. В., Думбай В. Н., Кураев Г. А., Фельдман Г. Л.* Физиология центральной нервной системы. 2-е изд., доп. и испр. – Ростов н/Д.: Феникс, 2000. – 384 с.

76. *Данилова Н. Н., Крылова А. Л.* Физиология высшей нервной деятельности. – Ростов н/Д.: Феникс, 1999. – 400 с.

77. *Блум Ф., Лейзерсон А., Хофстедтер Л.* Мозг, разум и поведение: Пер. с англ. – М.: Мир, 1988. – 248 с.

78. Мозг: Сборник: Пер. с англ. под ред. и с предисл. *П. В. Симонова.* – М.: Мир, 1982. – 280 с.

79. *Симонов П.В.* Мотивированный мозг: Высшая нервная деятельность и естественнонаучные основы общей психологии. – М.: Наука, 1987. – 269 с.

80. *Дискуссия о нейрокомпьютерах* // Всероссийская научно-техническая конференция «Нейроинформатика-99», 19–21 января 1999 г. / Отв. ред. *А. А. Фролов* и *А. А. Ежов*. – М.: Изд-во МИФИ, 2000. – 224 с.

81. *Борисюк Р. М., Виноградова О. С., Денэм М., Казанович Я. Б., Хоппенштедт Ф.* Модель детекции новизны на основе частотного кодирования информации // 2-я Всероссийская научно-техн. конференция «Нейроинформатика-2000», 19–21 января 2000 г. – М.: Изд-во МИФИ, 2000. – с. 145–156.

82. *Борисюк Р. М., Виноградова О. С., Денэм М., Казанович Я. Б., Хоппенштедт Ф.* Модель детекции новизны на основе осцилляторной нейронной сети с разреженной памятью // III Всероссийская научно-техн. конференция «Нейроинформатика-2001», 24–26 января 2001 г. – М.: Изд-во МИФИ, 2001. – с. 183–190.

83. *Кузьмина М. Г., Манькин Э. А., Сурина И. И.* Оценка памяти в замкнутых однородных цепочках осцилляторов // 2-я Всероссийская научно-техн. конференция «Нейроинформатика-2000», 19–21 января 2000 г. – М.: Изд-во МИФИ, 2000. – с. 94–99.

84. *Кузьмина М. Г., Манькин Э. А., Сурина И. И.* Модель осцилляторной сети, имитирующая основанное на синхронизации функционирование зрительной коры // III Всероссийская научно-техн. конференция «Нейроинформатика-2001», 24–26 января 2001 г. – М.: Изд-во МИФИ, 2001. – с. 191–200.

85. *Лагутина Н. С.* Модель импульсного нейрона. Колебания в простейшей сети из трех нейронов. Самоорганизация полносвязной сети импульсных нейронов // III Всероссийская научно-техн. конференция «Нейроинформатика-2001», 24–26 января 2001 г. – М.: Изд-во МИФИ, 2001. – с. 200–205.

86. *Мирошников С. А.* Интеграция импульсных и осцилляторных сетей в нейропсихологической системе // III Всероссийская научно-техн. конференция «Нейроинформатика-2001», 24–26 января 2001 г. – М.: Изд-во МИФИ, 2001. – с. 205–213.

87. *Сухов А. Г., Бездудная Т. Г., Медведев Д. С.* Ритмическая активность как фактор самоорганизации и пластичности нейронной сети // III Всероссийская научно-техн. конференция «Нейроинформатика-2001», 24–26 января 2001 г. – М.: Изд-во МИФИ, 2001. – с. 213–220.

88. *Кун Т.* Структура научных революций. 2-е изд.: Пер. с англ. – М.: Прогресс, 1977. – 300 с. (Серия «Логика и методология науки»)

89. *Хаос* (тема номера, 3 статьи) // Компьютерра. – № 47 (275), 1 декабря 1998 г. – с. 20–35.
URL: http://www.computerra.ru/offline/1998/275/

90. *Баблоянц А.* Молекулы, динамика и жизнь: Введение в самоорганизацию материи: Пер. с англ. – М.: Мир, 1990. – 375 с.

91. *Заславский Г. М., Сагдеев Р. З.* Введение в нелинейную физику: От маятника до турбулентности и хаоса. – М.: Наука, 1988. – 368 с.

92. *Лоскутов А. Ю., Михайлов А. С.* Введение в синергетику. – М.: Наука, 1990. – 272 с.

93. *Малинецкий Г. Г.* Хаос. Структуры. Вычислительный эксперимент: Введение в нелинейную динамику. – М.: Эдиториал УРСС, 2000. – 256 с.

94. *Малинецкий Г. Г., Потапов А. Б.* Современные проблемы нелинейной динамики. – М.: Эдиториал УРСС, 2000. – 336 с.

95. *Николис Дж., Пригожин И.* Познание сложного. Введение: Пер. с англ. – М.: Мир, 1990. – 344 с.

96. *Табор М.* Хаос и интегрируемость в нелинейной динамике: Пер. с англ. – М.: Эдиториал УРСС, 2001. – 320 с.

97. *Хакен Г.* Синергетика: Пер. с англ. – М.: Мир, 1980. – 404 с.

98. *Хакен Г.* Синергетика. Иерархия неустойчивостей в самоорганизующихся системах и устройствах: Пер. с англ. – М.: Мир, 1985. – 423 с.

99. *Хакен Г.* Информация и самоорганизация. Макроскопический подход к сложным системам: Пер. с англ. – М.: Мир, 1991. – 240 с.

100. *Шустер Г.* Детерминированный хаос. Введение: Пер. с англ. – М.: Мир, 1988. – 240 с.

101. *Эбелинг В., Энгель А., Файстель Р.* Физика процессов эволюции. Синергетический подход: Пер. с нем. – М.: Эдиториал УРСС, 2001. – 328 с.

102. *Эткинс П.* Порядок и беспорядок в природе: Пер. с англ. – М.: Мир, 1987. – 224 с.

103. *Эфрос А. Л.* Физика и геометрия беспорядка. – М.: Наука, 1982. – 176 с. (Библиотечка «Квант», вып. 19)

104. *Борисов А. Н., Алексеев А. В., Меркурьева Г. В., Слядзь Н. Н., Глушков В. И.* Обработка нечеткой информации в системах принятия решений. – М.: Радио и связь, 1989. – 304 с.

105. *Заде Л.* Понятие лингвистической переменной и его применение к принятию приближенных решений: Пер. с англ. – М.: Мир, 1976. – 165 с. (Серия «Новое в зарубежной науке: Математика», вып. 3 / Ред. серии *А. Н. Колмогоров* и *С. П. Новиков*)

106. Классификация и кластер / Под ред. *Дж. Вэн Райзина*: Пер. с англ. – М.: Мир, 1980. – 389 с.

107. *Кофман А.* Введение в теорию нечетких множеств: Пер. с франц. – М.: Радио и связь, 1982. – 432 с.

108. *Кузьмин В. Б.* Построение групповых решений в пространствах четких и нечетких бинарных отношений. – М.: Наука, 1982. – 168 с. (Серия «Теория и методы системного анализа»)

109. *Малышев Н. Г., Берштейн Л. С., Боженюк А. В.* Нечеткие модели для экспертных систем в САПР. – М.: Энергоатомиздат, 1991. – 136 с.

110. *Мелихов А. Н., Берштейн Л. С., Коровин С. Я.* Ситуационные советующие системы с нечеткой логикой. – М.: Наука, 1990. – 272 с.

111. *Орлов А. И.* Задачи оптимизации и нечеткие переменные. – М.: Знание, 1980. – 64 с. (Новое в жизни, науке, технике. Серия «Математика, кибернетика». Вып.8, 1980)

112. *Орловский С. А.* Проблемы принятия решений при нечеткой исходной информации. – М.: Наука, 1981. – 208 с. (Серия «Оптимизация и исследование операций»)

113. Прикладные нечеткие системы / Под. ред. *Т. Тэрано, К. Асаи* и *М. Сугэно*: Пер. с япон. – М.: Мир, 1993. – 368 с.

114. *Нечеткая логика* (тема номера, 4 статьи) // Компьютерра. – № 38 (415), 9 октября 2001 г. – с. 18–31.
URL: http://www.computerra.ru/offline/2001/415/

115. *Дюбуа Д., Прад А.* Теория возможностей. Приложения к представлению знаний в информатике: Пер. с франц. – М.: Радио и связь, 1990. – 288 с.

116. Нечеткие множества и теория возможностей: Последние достижения / Под ред. *Р. Р. Ягера*: Пер. с англ. – М.: Радио и связь, 1986. – 408 с.

117. *Пытьев Ю.П.* Возможность: Элементы теории и применения. – М.: Эдиториал УРСС, 2000. – 192 с.

118. Special Issue *"Fuzzy Information Processing"* / Ed.: *Dan Ralescu* // Fuzzy Sets and Systems: Intern. J. of Soft Computing and Intelligence. – Feb. 10, 1995. – v. 69, No. 3. – pp. 239–354.

119. Special Issue *"Fuzzy Signal Processing"* / Eds.: *Anca L. Ralescu* and *James G. Shanahan* // Fuzzy Sets and Systems: Intern. J. of Soft Computing and Intelligence. – Jan. 15, 1996. – v. 77, No. 1. – pp. 1–116.

120. Special Issue *"Fuzzy Multiple Criteria Decision Making"* / Eds.: *C. Carlsson* and *R. Fullér* // Fuzzy Sets and Systems: Intern. J. of Soft Computing and Intelligence. – March 11, 1996. – v. 78, No. 2. – pp. 139–241.

121. Special Issue *"Fuzzy Modelling"* / Ed.: *J. M. Barone* // Fuzzy Sets and Systems: Intern. J. of Soft Computing and Intelligence. – May 27, 1996. – v. 80, No. 1. – pp. 1–120.

122. Special Issue *"Fuzzy Optimization"* / Ed.: *J.-L. Verdegay* // Fuzzy Sets and Systems: Intern. J. of Soft Computing and Intelligence. – July 8, 1996. – v. 81, No. 1. – pp. 1–183.

123. Special Issue *"Fuzzy Methodology in System Failure Engineering"* / Ed.: *Kai-Yuan Cai* // Fuzzy Sets and Systems: Intern. J. of Soft Computing and Intelligence. – Oct. 8, 1996. – v. 83, No. 2. – pp. 111–290.

124. Special Issue *"Analytical and Structural Considerations in Fuzzy Modelling"* / Ed.: *A. Grauel* // Fuzzy Sets and Systems: Intern. J. of Soft Computing and Intelligence. – Jan. 16, 1999. – v. 101, No. 2. – pp. 205–313.

125. Special Issue *"Soft Computing for Pattern Recognition"* / Ed.: *Nikhil R.Pal* // Fuzzy Sets and Systems: Intern. J. of Soft Computing and Intelligence. – Apr. 16, 1999. – v. 103, No. 2. – pp. 197–367.

126. Special Issue *"Fuzzy Modeling and Dynamics"* / Eds.: *Horia-Nicolai Teodorescu, Abraham Kandel, Moti Schneider* // Fuzzy Sets and Systems: Intern. J. of Soft Computing and Intelligence. – Aug. 16, 1999. – v. 106, No. 1. – pp. 1–97.

127. Портал научных вычислений (Matlab, Fortran, C++ и т.п.)
URL: http://www.mathtools.net/

128. *Тихонов А.Н., Арсенин В.Я.* Методы решения некорректных задач. 3-е изд., испр. – М.: Наука, 1986. – 288 с.

129. *Тихонов А.Н., Гончарский А.В., Степанов В.В., Ягода А.Г.* Численные методы решения некорректных задач. – М.: Наука, 1990. – 232 с.

130. *Айвазян С. А., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика: Основы моделирования и первичная обработка данных. Справочное издание. – М.: Финансы и статистика, 1983. – 471 с.

131. *Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика: Классификация и снижение размерности. Справочное издание. – М.: Финансы и статистика, 1989. – 607 с.

132. *Бендат Дж., Пирсол А.* Прикладной анализ случайных данных: Пер. с англ. – М.: Мир, 1989. – 540 с.

133. *Боровков А. А.* Математическая статистика: Оценка параметров, проверка гипотез. – М.: Наука, 1984. – 472 с.

134. *Моррис У. Т.* Наука об управлении: Байесовский подход. Пер. с англ. – М.: Мир, 1971. – 304 с.

135. Bayes net toolbox for Matlab:
URL: http://www.cs.berkeley.edu/∼murphyk/Bayes/bnt.html
A Brief Introduction to Graphical Models and Bayesian Networks:
URL: http://www.cs.berkeley.edu/∼murphyk/Bayes/bayes.html

136. *Пайтген Х.-О., Рихтер П. Х.* Красота фракталов. Образы комплексных динамических систем: Пер. с англ. – М.: Мир, 1993. – 176 с.

137. *Шредер М.* Фракталы, хаос, степенные законы. Миниатюры из бесконечного рая: Пер. с англ. – М.: Мир, 1993. – 176 с.

138. *Боуэн Р. Методы символической динамики*. Сб. статей: Пер. с англ. под ред. *В.М.Алексеева*. – М.: Мир, 1979. – 245 с. (Серия «Новое в зарубежной науке: Математика», вып. 13 / Ред. серии *А.Н.Колмогоров* и *С.П.Новиков*)

139. *Каток А. Б., Хассельблат Б.* Введение в современную теорию динамических систем: Пер. с англ. – М.: Факториал, 1999. – 768 с.

140. *Палис Ж., Ди Мелу В.* Геометрическая теория динамических систем. Введение: Пер. с англ. – М.: Мир, 1986. – 301 с. (Серия «Современная математика: Вводные курсы»)

141. Теория систем: Математические методы и моделирование. Сб. статей: Пер. с англ. – М.: Мир, 1989. – 384 с. (Серия «Новое в зарубежной науке: Математика», вып. 44 / Ред. серии *А. Н. Колмогоров* и *С. П. Новиков*)

Редактор материалов выпуска,
кандидат технических наук *Ю. В. Тюменцев*

E-mail: tium@mai.ru

С. А. ШУМСКИЙ
Физический институт им. Лебедева РАН, ООО «НейрОК», Москва
E-mail: shumsky@neurok.ru

БАЙЕСОВА РЕГУЛЯРИЗАЦИЯ ОБУЧЕНИЯ

Аннотация

Байесовский подход к обучению, основанный на первых принципах теории вероятности, представляет собой наиболее последовательную парадигму в теории статистического обучения. С практической точки зрения, байесовское обучение органично включает в себя процедуру регуляризации, предлагая реальную альтернативу традиционным методам контроля сложности моделей, основанным на кросс-валидации.

S. A. SHUMSKY
Lebedev Physics Institute RAS, NeurOK LLC, Moscow
E-mail: shumsky@neurok.ru

BAYESIAN REGULARIZATION OF LEARNING

Abstract

Bayesian approach based on the first principles of the probability theory is the most consistent paradigm of statistical learning. From practical perspective Bayesian learning offers intrinsic regularization procedure providing a viable alternative to traditional cross-validation technique.

Введение

Машинное обучение (*machine learning*) ставит своей задачей выявление закономерностей в эмпирических данных. В противоположность математическому моделированию, изучающему следствия из известных законов, машинное обучение стремится воссоздать причины, наблюдая порожденные ими следствия — эмпирические данные. Обучение, таким образом, относится к классу обратных задач и в общем случае является

плохо определенной или *некорректной* задачей. Такие задачи отличаются особой чувствительностью некоторых решений к данным и нахождение устойчивых решений подразумевает процедуру *регуляризации* — ограничения класса допустимых решений.

Обучающиеся модели по определению должны быть чувствительны к данным, адаптируя в процессе обучения свои настроечные параметры для наилучшего объяснения всех известных фактов. Однако, хорошее качество объяснения имеющихся данных еще не гарантирует соответствующее качество предсказаний [1]. Излишне сложные модели способны адаптироваться не только к типичным закономерностям, но и к случайным событиям в данной обучающей выборке. Как следствие, такие модели обладают плохой предсказательной способностью: большая чувствительность к данным приводит к большому разбросу в предсказаниях. Модель в этом случае оказывается неспособной *обобщить* данные, отделив общие закономерности от случайных флуктуаций. Поэтому ограничение сложности моделей является необходимым элементом теории обучения. Качество обучения напрямую зависит от нашей способности пройти между Сциллой переупрощения и Харибдой переусложнения.

На практике наибольшее распространение получили методики регуляризации, основанные на тех или иных способах оценки ожидаемой ошибки обучения на новых данных — *ошибки обобщения*. Этот подход интуитивно кажется наиболее естественным, поскольку минимизация последней и является истинной целью обучения, тогда как практически мы имеем возможность измерять лишь эмпирическую *ошибку обучения*.

Такое интуитивно обоснованное обучение подразумевает два этапа: настроечные параметры модели определяются минимизацией ошибки обучения, тогда как выбор между моделями различной сложности определяется, исходя из оценки ошибки обобщения. Имеющиеся данные при этом также делятся на две категории. Часть данных используют для настроек модели, а на остальных проверяют достигнутое качество обучения. Этот этап называют *валидацией* модели. Чтобы избежать зависимости от конкретного разбиения данных на обучающую и валидационную выборки, используют метод *кросс-валидации*, оценивая оптимальную сложность модели в большом числе экспериментов с разными спо-

[1] Например, биржевые обозреватели, уверенно объясняющие наблюдаемое движение цен, становятся гораздо менее категоричными в части прогнозов на будущее.

собами разбиения данных. Трудоемкость метода кросс-валидации ограничивает его применимость, например в системах реального времени или для действительно сложных моделей, требующих длительного обучения.

Байесова регуляризация, предмет данного обзора, является альтернативной методикой оптимизации сложности модели. Она основана не на оценке ожидаемой ошибки, а на выборе наиболее *правдоподобной* модели, в пользу которой свидетельствуют имеющиеся данные. Такой подход имеет ряд преимуществ. Во-первых, он исходит из первых принципов теории вероятностей и теории статистического обучения, гарантирующих уменьшение ошибки обобщения. Во-вторых, он подразумевает оценку вариаций параметров модели и соответственно — оценку точности своих предсказаний. В-третьих, поставленная таким образом задача в некоторых практически важных случаях может быть решена с минимальным числом дополнительных упрощающих предположений. И, наконец, как следствие, *last but not least*: байесова регуляризация может быть встроена непосредственно в алгоритмы обучения. Причем, такие регуляризованные алгоритмы уже не подразумевают этапа валидации, единообразно используя все имеющиеся данные как для выбора оптимальной сложности модели, так и для настройки ее параметров.

В следующем разделе («Обучение по Байесу», с. 33–51) мы подробно остановимся на идеологической стороне байесовской регуляризации и основанных на ней алгоритмах обучения. Затем, в разделе «Оценка параметров по Байесу» (с. 51–57) мы применим общий подход к простейшей задаче оценки зашумленной величины. Байесов подход в этом случае дает, например, четкий критерий достаточности экспериментальных данных для проверки теоретической гипотезы. Раздел «Байесова интерполяция функций» (с. 57–69) посвящен байесовской регуляризации аппроксимации функций, проблеме, к которой сводится большинство прикладных задач машинного обучения. Соответствующие алгоритмы обучения применимы, в частности, для многослойных персептронов. В разделе «Байесова классификация» (с. 70–82) мы рассмотрим другую практически важную задачу — кластеризацию данных. В частности, покажем как «по Байесу» определять оптимальное число кластеров. В конце каждого раздела дана краткая историко-библиографическая справка по развитию затронутых в нем идей. Чтобы облегчить изложение, все детали вынесены в раздел Подробности.

Обучение по Байесу

В этом разделе мы обсудим *процедуру байесовской регуляризации*, ее обоснование и связь с другими концепциями обучения, а также опишем в общем виде алгоритм обучения со встроенной байесовской регуляризацией.

Начнем с формализации основных понятий: обучения, регуляризации и байесовской статистики.

Обучение. Основные понятия

Интуитивно, задачей *обучения* является *обобщение* эмпирических данных, предполагающее возможность предсказывать новые события, основываясь на известном опыте прошлого. Такие предсказания в наиболее общем случае имеют вероятностный характер [2]: обобщением имеющегося набора данных $D = \{\mathbf{d}_1, \ldots, \mathbf{d}_N\}$ служит некая гипотеза h вероятностного происхождения данных $P_h(\mathbf{d}) \equiv P(\mathbf{d}\,|h)$.

Такая гипотеза обладает предсказательной силой, поскольку позволяет не только оценить меру *правдоподобия* (*likelihood*) имеющихся данных $P(D\,|h)$, но и предсказать вероятность любого нового набора данных $P(D'\,|h)$. Расчет подобного рода вероятностей различных исходов экспериментов при заданном способе порождения данных $P(\mathbf{d}\,|h)$ является предметом теории вероятности. Например, вычислить вероятность выпадения определенного числа «решек» при многократном бросании монеты с известной степенью «кривизны» (монеты, а не вычисления!). Здесь $\mathbf{d}_n$ — исход n-го бросания монеты, D — результат N опытов, а $P(\mathbf{d}\,|h)$ — вероятность выпадения «решек» при данной степени кривизны монеты h.

Обучение предполагает решение *обратной задачи*: по имеющимся данным следует выяснить вероятность различных гипотез о способе порождения этих данных $P(h\,|D)$. В случае с монетой, например, требуется оценить вероятность различной степени ее «кривизны» по известной (конечной) выборке исходов экспериментов.

[2] Детерминистские функции являются частным случаем, когда вероятностные распределения вырождаются в дельта-функции.

Обычно эту *апостериорную* (*posterior*) вероятность используют для выбора наиболее вероятной гипотезы в качестве кандидата для предсказания будущих событий такого рода:

$$h_{MP} = \arg\max_h P(h|D) \ .$$

В традиционной статистике, рассматривающей, по сути, идентичный круг задач выбора наилучшей аппроксимации эмпирических данных, базовым является другой критерий оптимальности — *принцип максимума правдоподобия*:

$$h_{ML} = \arg\max_h P(D|h) \ ,$$

не предполагающий решения обратной задачи. Как мы увидим далее, такое приближение действительно оправдано в рамках обычных предположений традиционной статистики, а именно, когда количество данных намного превышает эффективное число параметров модели. Между тем, при относительно небольшом количестве данных принцип максимального правдоподобия может приводить к парадоксам. Например, при бросании монеты наиболее правдоподобной оценкой ее кривизны является эмпирическая частота выпадения «решек». И если в серии из 5 исходов случайно не выпадет ни одной «решки», то мы вынуждены будем считать ее «бесконечно кривой», тогда как на самом деле вероятность такого события даже для идеальной монеты не слишком мала.

Байесов подход к обучению, основанный на решении обратной задачи, более последователен и, соответственно, применим к более широкому классу моделей с большими возможностями моделирования сложных явлений. Тем более, что в общем виде эта задача решается «в одну строку» и ее решение, следующее из общих принципов теории вероятностей, было известно уже в XVIII веке. Действительно, если трактовать как выбор гипотезы, так и наблюдение данных в вероятностном смысле и записать согласно определению условных вероятностей $P(D,h) = P(h|D)\,P(D) = P(D|h)\,P(h)$, получим теорему преподобного Байеса:

$$P(h|D) = \frac{P(D|h)\,P(h)}{P(D)} = \frac{P(D|h)\,P(h)}{\sum_h P(D|h)\,P(h)} \ . \qquad (1)$$

(В разделе Подробности, в качестве примера, дано Байесово решение задачи о монете.)

Для фиксации терминологии запишем эту основополагающую формулу в словесном виде:

$$Posterior = \frac{Likelihood \cdot Prior}{Evidence}$$

Регуляризация обучения

Как видим, решение обратной задачи требует формализации наших *априорных* (*prior*) предположений $P(h)$ о степени вероятности той или иной гипотезы. Подобного рода ограничение на множество гипотез, в котором ищется решение, в теории обратных задач называют *регуляризацией*. Необходимость ее связана с конечным объемом эмпирических данных. Если мы не будем ограничены в средствах, то всегда сможем подобрать гипотезу, идеально объясняющую имеющиеся данные, но с плохими способностями к обобщению: $P(D'|h) \ll P(D|h)$. Иными словами, такие гипотезы (называемые по латыни *ad hoc*)[3] чрезвычайно чувствительны к конкретному набору обучающих данных. Чувствительность к данным есть индикатор того, что задача обучения по своей природе *некорректна*, и как всякая некорректная обратная задача требует регуляризации. В ограниченном классе гипотез чрезмерную чувствительность решения к обучающей выборке можно преодолеть.

В качестве иллюстрации приведем результаты определения частоты зашумленного синуса методом наименьших квадратов без регуляризации (рис. 1) и с регуляризацией (рис. 2). В первом случае ответ чрезвычайно чувствителен к шумовой компоненте данных. В зависимости от реализации шума, наименьшую ошибку может показать любая из бесконечного набора частот. Ограничение сложности модели, в данном случае — добавление к ошибке штрафного члена, пропорционального квадрату частоты, выявляет решение, наименее чувствительное к шуму.

Выбор метода регуляризации, то есть класса гипотез, в свою очередь, является *мета-гипотезой* H более высокого порядка, которые в теории машинного обучения принято называть моделями: $P(h) = P_H(h) \equiv P(h|H)$. Так, в задаче интерполяции функций модель фиксирует выбранный метод параметризации функций, например, персептрон с задан-

[3] Ad hoc гипотеза — гипотеза, специально созданная для объяснения именно данного конкретного явления. — Прим. ред.

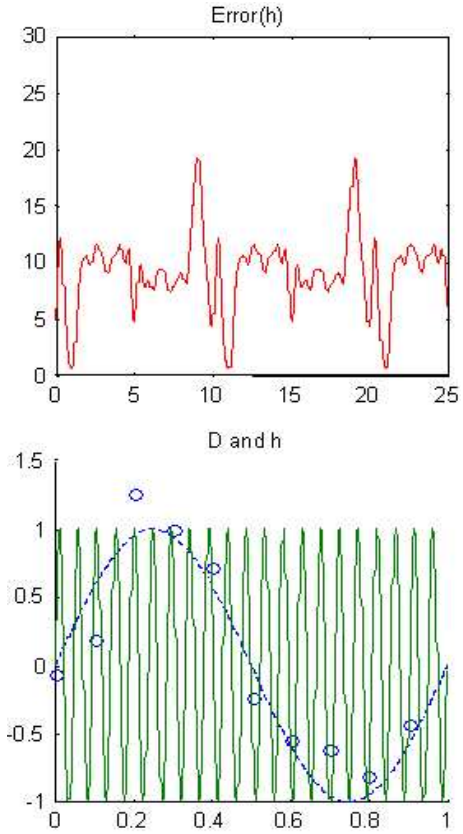

Рис. 1. Определение частоты зашумленного синуса $y = \sin(h\,x) + 0.2\,\eta$. Здесь модель задает характер шума η и вид функции $\sin(h\,x)$, где в роли гипотезы h выступает частота. Функция ошибки (вверху) имеет множество локальных минимумов. Без регуляризации наиболее правдоподобным может оказаться любой из них, в данном примере $h = 21$. На нижнем рисунке показано соответствующее решение (сплошная кривая) и истинная функция $h = 1$ (пунктир).

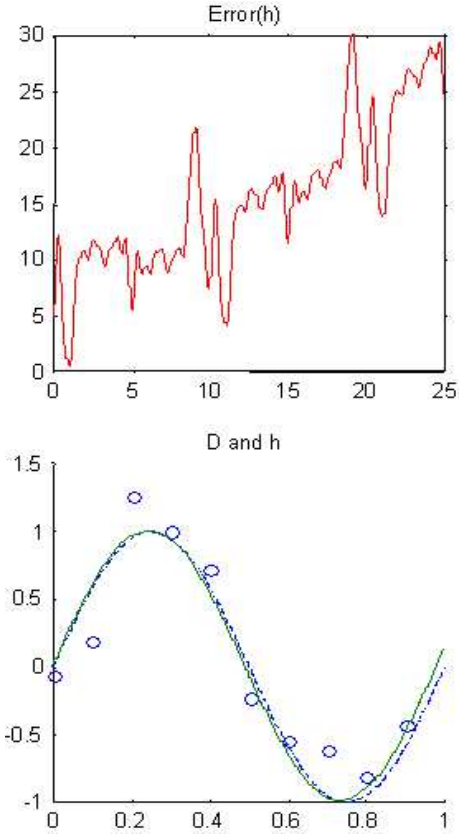

Рис. 2. Регуляризация модели — добавление к ошибке члена, штрафующего большие частоты, помогает выявить наиболее устойчивое к шуму решение, близкое к реальному прототипу.

ной топологией связей или сплайны определенного порядка. Конкретные значения подгоночных параметров соответствуют гипотезам. Гипотезы всегда выбираются в рамках той или иной модели и, с этой точки зрения, все вероятности в формуле Байеса зависят от H:

$$P\left(h\,|D\,,H\right)=\frac{P\left(D\,|h,H\right)P\left(h\,|H\right)}{P\left(D\,|H\right)}\;.$$

В дальнейшем, однако, как и в выражении (1), мы иногда для краткости не будем обозначать эту зависимость от модели.

Фундаментальный характер теоремы Байеса позволяет в едином ключе сравнивать между собой не только гипотезы, но и различные модели регуляризации. Тем самым, байесовский подход позволяет расширить рамки традиционной теории регуляризации, не предполагающей сравнение между собой *регуляризирующих функционалов* $P\left(h\,|H\right)$.

Насколько правдоподобно выглядит объяснение данных моделью определяет знаменатель формулы Байеса

$$P\left(D\,|H\right)=\sum_{h}P\left(D\,|h\,,H\right)P\left(h\,|H\right)=\sum_{h}P\left(D,h\,|H\right)\;. \qquad (2)$$

Поэтому его и называют *Evidence*, что можно перевести как *свидетельство* или *доказательство* в пользу данной модели H. Формула Байеса, но уже на уровне моделей:

$$P\left(H\,|D\right)=\frac{P\left(D\,|H\right)P\left(H\right)}{P\left(D\right)}$$

дает возможность сравнивать между собой различные «априорные» ограничения $P\left(h\,|H\right)$, присущие различным типам моделей. А именно:

$$H_{MP}=\arg\max_{H}P\left(H\,|D\right)\;.$$

Решение обратной задачи для модели требует, естественно, выбора Prior уже на множестве моделей, т. е. задания некой мета-модели более высокого порядка. И так далее. На практике, разумеется, ограничиваются конечным числом ступеней в иерархии моделей, заменяя на каком-то уровне наиболее вероятную модель наиболее правдоподобной.

Например, в простейшей двухуровневой схеме Байесовского обучения полагают, что в отсутствие каких-то предпочтений между несколькими различными способами моделирования данных $P(H) = \text{const}$ и мы имеем возможность обоснованно выбрать тот из них, в пользу которого свидетельствуют эмпирические данные, т.е. модель с максимальным значением Evidence:

$$H_{ML} = \arg\max_{H} P(D|H) \ .$$

Этот *принцип максимизации значения Evidence* и определяет в данной работе байесовскую регуляризацию обучения.

Предварительное обсуждение

Необходимость явного задания априорной функции распределения нередко трактуется сторонниками традиционной статистики как препятствие к практическому использованию байесовского подхода. На самом деле, как мы видим, ситуация, скорее, обратная. Ведь выбор той или иной модели интерполяции данных в любом случае задает какой-то Prior. Байесов формализм просто не дает замести эти неявные предположения под ковер. Напротив, возможность обоснованно выбирать оптимальные модели порождения данных следует считать существенным преимуществом последовательного байесовского подхода к обучению.

Подчеркнем, что оптимальная модель, по Байесу, состоит из *ансамбля* гипотез. Считается, что в предсказаниях участвуют все гипотезы, каждая со своей апостериорной вероятностью. Как будет показано ниже, ансамбль в целом обладает лучшей обобщающей способностью, чем любой из его представителей [4]. На качественном уровне этот факт иллюстрируется рис. 3. Далее мы обсудим вопрос о связи байесовской достоверности с обобщающей способностью модели более подробно.

Заметим в скобках, что регуляризация методом кросс-валидации также оценивает ошибку обобщения ансамблей, а не отдельных гипотез. Байесовская регуляризация лишь выражает эту точку зрения более систематически.

[4] Читатель, знакомый с теорией игр, заметит прозрачную аналогию предсказаний ансамблем со смешанными стратегиями, позволяющими добиваться лучших результатов, чем чистые стратегии.

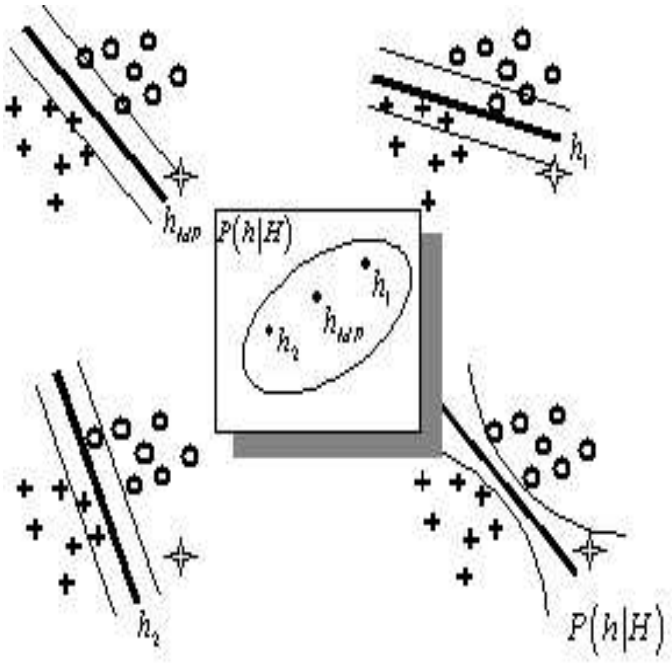

Рис. 3. Иллюстрация байесовского подхода к предсказаниям. Данные представляют собой набор точек из двух классов. Гипотеза классифицирует данные в соответствии с их расположением относительно линии разделения классов, в данном случае — прямой. Звездой отмечена новая точка, отсутствующая в обучающей выборке. Наиболее вероятная гипотеза h_{MP} классифицирует эту точку как «круг». Однако, среди других возможных гипотез нет единства: некоторые, такие как h_1, голосуют за «крест», другие, как h_2 — за «круг». Тем самым, предсказание ансамблем гипотез дает возможность понять, что новая точка лежит далеко от обучающей выборки и оценить надежность ее классификации.

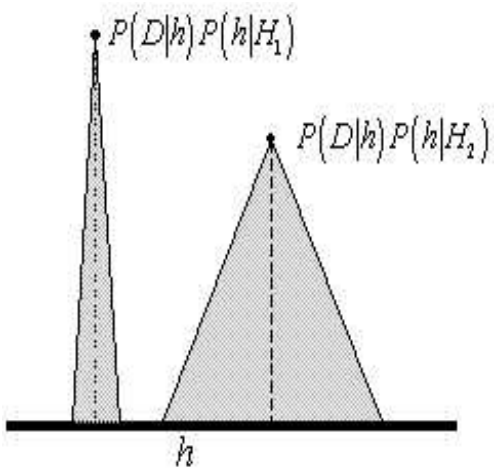

Рис. 4. Из двух моделей, H_1 и H_2, более предпочтительной, по Байесу, является вторая — с большей Evidence (площадь под кривой), несмотря на то, что наилучшая гипотеза в H_1 лучше объясняет данные. Зато H_1 гораздо более чувствительна к вариациям своих параметров, чем H_2.

При таком подходе вполне естественно, что наилучшей моделью считается не та, в которой существует наиболее правдоподобная гипотеза, а та, в которой доля правдоподобных гипотез достаточно велика. Максимизация Evidence выражает именно эту точку зрения (см. рис. 4). Поскольку интеграл Evidence определяется не только высотой, но и шириной апостериорного пика в пространстве гипотез, то наиболее вероятная гипотеза в оптимальной, по Байесу, модели должна не просто соответствовать данным, но и быть одновременно наиболее робастной, т. е. наименее чувствительной к вариациям своих параметров.

Наиболее близки байесовской трактовке обучения стохастические алгоритмы с фиктивной «температурой», где гипотезы играют роль состояний с энергией, равной их эмпирической ошибке [5]. Вообще говоря,

[5] Например, схема Метрополиса и метод имитации отжига.

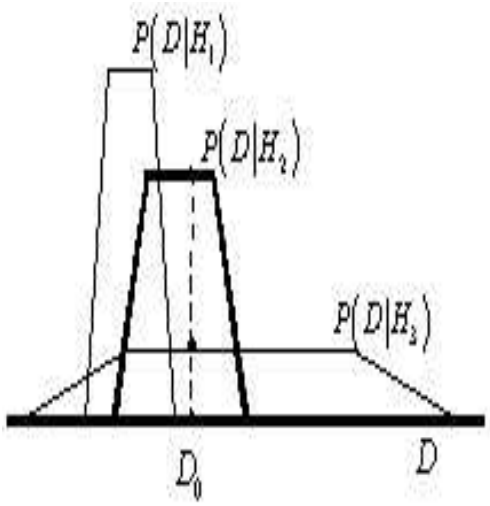

Рис. 5. Максимизация Evidence предполагает выбор наиболее простой модели объяснения данных. Модель H_1 не соответствует данным. Модель H_3 может объяснить не только имеющиеся данные, но и широкий круг других исходов эксперимента. Условие нормировки автоматически понижает ее Evidence. По Байесу, эмпирические данные свидетельствуют в пользу модели H_2.

существует глубокая аналогия между байесовским подходом в теории обучения и статистической физикой. Обе теории основаны на усреднении по ансамблю состояний с заданным каноническим распределением вероятностей. При этом максимизация Evidence аналогична минимизации функции свободной энергии, в чем у нас еще будет случай убедиться.

Можно сказать, что максимизация Evidence реализует известный принцип *бритвы Оккама*: предпочтение отдается наиболее простой модели, способной объяснить эмпирические данные. Этот факт иллюстрирует рис. 5. Как видно из этого рисунка, байесов подход отсеивает не только модели, не соответствующие наблюдаемым данным, но и излишне слож-

ные модели, могущие объяснить большее разнообразие данных [6]. Этот вопрос заслуживает более подробного рассмотрения, поскольку, помимо прочего, проливает свет на соотношение байесовской регуляризации с минимизацией ошибки обобщения.

Связь с ошибкой обобщения и минимальной длиной описания

На уровне гипотез ошибка обобщения тем меньше, чем ближе наша гипотеза порождения данных $P(D|h)$ к истинной функции распределения $P(D|h_0)$. Отличие между ними, по мере Кулбака-Леблера, равно:

$$|P(D|h) - P(D|h_0)| = \sum_D P(D|h_0) \log \frac{P(D|h_0)}{P(D|h)} =$$

$$= \sum_D P(D|h_0) [L(D|h) - L(D|h_0)] \geqslant 0 ,$$

где *эмпирический риск* $L(D|h) = -\log P(D|h)$ аддитивен по числу примеров и пропорционален эмпирической ошибке. Усредненный по бесконечному набору выборок *ожидаемый риск* $\sum_D P(D|h_0) L(D|h)$ соответствует ошибке обобщения.

Именно эту последнюю (ненаблюдаемую!) величину мы хотели бы минимизировать в процессе обучения.

Асимптотически, с ростом объема выборки эмпирический риск стремится к ожидаемому, но при конечном числе примеров это разные сущности, и минимизация одного не обязательно приводит к минимизации другого.

Цель регуляризации — ввести новую измеримую величину, *регуляризированный риск*, которая вела бы себя аналогично ненаблюдаемому ожидаемому риску. В байесовском подходе регуляризирующий функционал принимает вид априорной вероятности гипотез $P(h)$. При этом максимизируется совместная вероятность данных и гипотезы, т. е. минимизируется регуляризированный риск вида:

$$L(D,h) = -\log P(D,h) = -\log P(D|h) - \log P(h) \ .$$

[6] Модели, которые в принципе могут объяснить любые данные, К. Поппер предлагал считать «ненаучными», так как никакой эксперимент не в состоянии их опровергнуть. По Байесу, в силу условия нормировки, их Evidence действительно стремится к нулю.

Параметры регуляризации можно подбирать, исходя из оценок ожидаемого риска, т. е. методом *валидации* со всеми его недостатками, упомянутыми во Введении. Существует, однако, теоретически обоснованный «внутренний» критерий выбора функционала $P(h)$, основанный на принципе *минимальной длины описания* (*Minimum Description Length*), тесно связанный с байесовским подходом. Дело в том, что эмпирический риск $L(D|h) = -\log P(D|h)$ можно трактовать как длину оптимального кодирования данных с помощью гипотезы h, а $L(h) = -\log P(h)$ — как длину кодирования самой этой гипотезы. Таким образом, регуляризированный риск представляет собой суммарную длину описания данных и гипотезы $L(D,h) = L(D|h) + L(h)$, а его минимизация соответствует поиску наиболее компактного представления данных. Оказывается, именно суммарная длина описания и определяет качество предсказаний, ограничивая сверху ожидаемый риск. Этот фундаментальный факт был открыт Риссаненом [Rissanen 1978]: чем короче суммарная длина описания, тем лучше обобщающая способность гипотезы (см. Подробности) [7].

Оптимальная по Байесу гипотеза дает по определению как раз наиболее компактное представление данных в рамках выбранной модели:

$$h_{MDL} = \arg\min_h L(D,h) = \arg\max_h \{\log P(h|D) + \log P(D)\} = h_{MP} .$$

То же самое справедливо и на уровне моделей. Модель, наиболее компактно представляющая данные, обладает и наилучшей обобщающей способностью. Таким образом, максимизация Evidence соответствует принципу минимальной длины описания, но только применительно к ансамблю гипотез. Причем, длина описания данных с помощью всего ансамбля меньше, чем длина описания с помощью наилучшей гипотезы:

$$\begin{aligned} L(D|H) &= -\log P(D|H) = \\ &= -\log \sum_h \exp\left[-L(D,h|H)\right] < L(D,h_{MP}|H) \ , \end{aligned}$$

поскольку любой член суммы положительных слагаемых под логарифмом меньше, чем вся сумма. Следовательно, согласно Риссанену, и обоб-

[7] Качественно, это следует из теории сложности Колмогорова, согласно которой случайные данные несжимаемы. Сжатие возможно лишь при наличии скрытых закономерностей, и чем большего сжатия удается достичь с помощью некоторой гипотезы, тем менее вероятно, что это простая случайность.

щающая способность предсказаний с помощью ансамбля выше, чем обобщающая способность любой, даже наилучшей из его гипотез!

Таким образом можно, не обращаясь к эмпирическим методикам оценки ошибки обобщения, выбирать модель с минимальной ошибкой обобщения, а именно ту, которая описывает данные наиболее компактным образом. В этом и состоит суть байесовской регуляризации, в которой модель представлена ансамблем гипотез.

Здесь опять уместно обратиться к термодинамической аналогии. В терминах длины описания формулу Байеса можно записать в виде канонического распределения, известного из статистической физики:

$$\begin{aligned} P\left(h\,|D,H\right) &= \frac{1}{P\left(D\,|H\right)}\exp\left[-L\left(D,h\,|H\right)\right], \\ P\left(D\,|H\right) &= \sum_{h}\exp\left[-L\left(D,h\,|H\right)\right]. \end{aligned} \tag{3}$$

Здесь в качестве безразмерной «энергии» гипотезы h выступает суммарная длина описания $L\left(D,h\,|H\right)$, а «статистической сумме» соответствует Evidence $P\left(D\,|H\right)$. Длина описания данных моделью $L\left(D\,|H\right) = -\log P\left(D\,|H\right)$ является аналогом «свободной энергии».

Таким образом, максимизация Evidence эквивалентна минимизации длины описания данных моделью и соответствует минимизации свободной энергии в статистической физике [8]. Эту термодинамическую аналогию мы используем при выводе итерационного алгоритма байесовского обучения в следующем параграфе.

EM-алгоритм

Байесовское обучение можно проводить итерационно: при данных параметрах регуляризации H^t оценить вероятности гипотез $P^t\left(h\,|D,H^t\right)$, максимизируя соответствующую Evidence, подправить параметры регуляризации H^{t+1}, и так далее. Рассмотрим этот весьма распространенный способ обучения несколько подробнее.

[8] А также максимизации энтропии при заданных ограничениях (например, при заданном значении средней энергии).

Оптимизация модели, т. е. минимизация длины описания

$$L\left(D\,|H\right) = -\log\sum_h P\left(D,h\,|H\right),$$

подразумевает вычисление «статсуммы» Evidence. Избежать этой непростой операции можно, воспользовавшись аналогией со статистической физикой, согласно которой «свободная энергия» $L\left(D\,|H\right)$ должна быть равна разнице усредненной по каноническому распределению «энергии» $L\left(D,h\,|H\right)$ и энтропии этого распределения. В соответствии с этим, определим функционал свободной энергии следующим образом:

$$\begin{aligned} F\left(\mathcal{P},H\right) &= \left\langle L\left(D,h\,|H\right)\right\rangle_{\mathcal{P}} - S\left(\mathcal{P}\right) = \\ &= \sum_h \mathcal{P}\left(h\right)\left[-\log P\left(D,h\,|H\right) + \log\mathcal{P}\left(h\right)\right], \end{aligned} \tag{4}$$

где усреднение проводится по неизвестной пока функции распределения $\mathcal{P}\left(h\right)$ в пространстве гипотез. Минимум этого функционала должен, по идее, достигаться для апостериорного распределения Байеса и совпадать при этом с длиной описания $L\left(D\,|H\right)$. Действительно, в этом легко убедиться, переписав функционал (4) в эквивалентном виде:

$$F\left(\mathcal{P},H\right) = L\left(D\,|H\right) + \sum_h \mathcal{P}\left(h\right)\log\frac{\mathcal{P}\left(h\right)}{P\left(h\,|D,H\right)}\ .$$

Второй член здесь соответствует расстоянию Кулбака между $\mathcal{P}\left(h\right)$ и апостериорным байесовским распределением, откуда и следует, что

$$\arg\min_{\mathcal{P}} F\left(\mathcal{P},H\right) = P\left(h\,|D,H\right), \quad \min_{\mathcal{P}} F\left(\mathcal{P},H\right) = L\left(D\,|H\right)\ .$$

Таким образом, ценой введения дополнительной переменной $\mathcal{P}\left(h\right)$ мы избавились от суммирования под знаком логарифма. Суммирование логарифмов при усреднении — потенциально гораздо более простая задача. К тому же, решение для $\mathcal{P}\left(h\right)$ дается в явном виде формулой Байеса.

На этом факте строится следующая схема последовательной минимизации свободной энергии, содержащая на каждой итерации два этапа — на уровне гипотез и на уровне моделей. По названию своих этапов этот алгоритм обучения известен как *Expectation Maximization*, или EM-алгоритм. А именно:

- этап **Expectation**:
  $\mathcal{P}^t(h) = \arg\min_{\mathcal{P}} F(\mathcal{P}, H^t)$;
- этап **Maximization**:
  $H^{t+1} = \arg\min_{H} F(\mathcal{P}^t, H)$.

Таким образом, на каждом этапе мы фиксируем одну группу параметров и оптимизируем другую. Эти этапы повторяются, пока алгоритм не сойдется. Сходимость EM-алгоритма гарантируется тем, что свободная энергия (длина описания) ограничена снизу и на каждом шаге не возрастает.

Названия этапов определяется их содержанием.

На этапе **Expectation** производится оценка апостериорной функции распределения гипотез при текущих параметрах регуляризации. Ответ дается формулой Байеса:

$$\mathcal{P}^t(h) = \frac{P(D, h \,|H^t)}{P(D \,|H^t)} \ .$$

На этапе **Maximization** производится уточнение параметров регуляризации путем минимизации усредненной по найденному распределению «энергии» (поскольку энтропия не зависит от H):

$$H^{t+1} = \arg\min_{H} \langle L(D, h \,|H)\rangle_{\mathcal{P}^t} = \arg\max_{H} \langle \log P(D, h \,|H)\rangle_{\mathcal{P}^t} \ .$$

Заметим, что как и любой другой градиентный способ обучения, EM-алгоритм сходится к локальному минимуму, не обязательно совпадающему с глобальным.

Резюме

В этом разделе мы рассмотрели основы байесовской теории регуляризации обучения, не использующей процедуру кросс-валидации. Этот подход последовательно извлекает имеющуюся в данных информацию, исходя из первых принципов теории вероятности.

Модель порождения данных в байесовской трактовке представлена ансамблем гипотез. Обучение увеличивает наше знание относительно

такой модели. Ему предшествует некий априорный ансамбль гипотез, а результатом является более компактный апостериорный ансамбль гипотез. Предсказания модели подразумевают усреднение по этому ансамблю. При этом, качество предсказаний ансамбля выше, чем качество предсказания его наилучшей гипотезы. Оптимальному апостериорному ансамблю соответствует максимальная Evidence.

Мы обсудили также эквивалентность формулы Байеса принципу минимальной длины описания данных, а тем самым и связь байесовского подхода с минимизацией ошибки обобщения, поскольку имеются строгие результаты, согласно которым уменьшение длины описания данных сопровождается уменьшением ошибки обобщения.

Наконец, мы описали конкретный алгоритм обучения, реализующий идеи байесовской регуляризации. На очереди — примеры применения общей теории к различным классам задач.

История и библиография

Статистическое сравнение и проверка гипотез появились в научном арсенале в XVIII веке. Пионером здесь является, по-видимому, английский математик, врач и писатель Джон Арбутнот, который отверг естественную гипотезу о равновероятности рождения мальчиков и девочек на основании демографических данных, согласно которым за все 82 года наблюдения мальчиков рождалось больше, чем девочек. Арбутнот аргументировал свои выводы тем, что если бы вероятность рождения мальчиков была $\frac{1}{2}$, то данная выборка имела бы исчезающе малую вероятность 2^{-82}.

В 1734 году Французская академия присудила премию за исследования по орбитам планет Даниилу Бернулли. Подобно Арбутноту, Бернулли отверг гипотезу о случайности орбит планет, изучая распределение пересечения их осей с единичной сферой. Позже, в 1812 году, Лаплас показал, что орбиты комет, напротив, равномерно распределены на этой сфере, чем обосновал гипотезу о том, что кометы являются пришельцами из внешнего космоса, а не элементами Солнечной системы.

Преподобный Томас Байес, ученик де Муавра, доказал свою знаменитую теорему где-то около 1750 года при рассмотрении задачи, «обратной

проблеме Бернулли». Имелся в виду Якоб Бернулли [9], автор основополагающего трактата по теории вероятностей *Искусство предположений* (*Acta conjectandi*, 1713). Опубликована работа Байеса была лишь после его смерти [1] (Bayes, 1763). Современный вид, как и свое имя, теорема приобрела в трудах Лапласа [19] (Laplace, 1819).

Несмотря на свою простоту и очевидность, она стала настоящим яблоком раздора в математической статистике. Споры вокруг ее практической применимости не затихают до сих пор. Противники байесовской статистики считают ее бесполезной в силу произвольности выбора априорных вероятностей. В итоге, долгое время эта теорема была практически исключена из статистических исследований [10].

Определяющим принципом в статистике, начиная с 20-х годов XX века, стал принцип наибольшего правдоподобия Рональда Фишера [7] (Fisher, 1912). Наиболее правдоподобные оценки действительно обладают рядом привлекательных асимптотических свойств. В частности, при данной функции распределения $P(D|h_0)$, наиболее правдоподобная оценка гипотезы h_{ML} асимптотически ведет себя как нормальная величина со средним значением h_0 и минимально возможной дисперсией $\propto 1/N$. Иными словами, статистика Фишера была асимптотической теорией и байесовский подход был ей идейно чужд.

Ограниченность данных, как мы знаем, начинает сказываться, когда длина их описания становится сравнимой с длиной описания гипотез. В этом случае и требуется обращение к байесовской регуляризации. Фишеровская же статистика предполагала сравнение гипотез, определенных с точностью до конечного, обычно небольшого, числа параметров, на основании стремящегося к бесконечности числа примеров.

В последней трети XX века развитие статистики шло по пути постепенного отказа от этих ограничений. Асимптотический подход сменился анализом обучения на конечных выборках, а жесткая параметризация гипотез — общими ограничениями на класс функций, в которых отыскивается решение. Смена фишеровской парадигмы завершилась к 80-м годам.

[9] Дядя упомянутого выше Даниила Бернулли.

[10] Хотя, было известно, что роль выбора априорных ограничений можно свести к нулю, если использовать апостериорные вероятности от предыдущих экспериментов в качестве априорных вероятностей для последующих [24] (Mises, 1939).

Облик новой теории статистического обучения определили четыре открытия, сделанные в 60-е годы [35] (Vapnik, 1995):

- непараметрическая статистика ознаменовала отказ от жесткого регламентирования функционального вида решения [26] (Parzen, 1962), [31] (Rosenblatt, 1956), [48] (Ченцов, 1962);
- метод регуляризации эффективно сужает класс решений без их жесткой параметризации [47] (Тихонов, 1963), [43] (Иванов, 1962), [28] (Phillips, 1962);
- неасимптотическая теория распознавания образов связывает разнообразие множества гипотез, на котором происходит поиск решения, с ошибкой обобщения [41], [42] (Вапник, 1968 и 1974);
- теория алгоритмической сложности связывает разнообразие и сложность множеств с длиной описания порождающих их программ [17] (Kolmogoroff, 1965), [33] (Solomonoff, 1960), [4] (Chaitan, 1966).

Байесовский подход, уже доказавший к тому времени свою практическую пользу, прекрасно вписался в новый стиль мышления. Он, как мы убедились, теснейшим образом связан практически со всеми составляющими новой теории обучения. Настолько тесно, что имя преподобного Байеса стало сегодня одним из наиболее часто употребляемых в теории обучения.

На практике байесовское сравнение моделей применял еще кембриджский геофизик сэр Джеффрис, не акцентируя внимание на том, какой Prior «истинный» [13] (Jeffreys, 1939). В компьютерную эру, по мере накопления баз данных, байесовское сравнение моделей завоевывает популярность в эконометрике [40] (Zellner, 1984), геофизике [27] (Patrick, 1982), обработке сигналов, теории распознавания образов [8] (Gull, 1988), [32] (Skilling, 1991), [10] (Hanson, 1991) и других областях. В качестве обзорных можно порекомендовать работы [16] (Kashyap, 1977), [12] (Janes, 1986), [21] (Loredo, 1989) [11].

EM-алгоритм впервые подробно обсуждался с позиции неполного описания данных в статье [5] (Dempster, 1977). Градиентный характер

[11] Дополнительные материалы можно найти на сайте байесовского общества URL: http://www.bayesian.org

EM-алгоритма был выявлен в работе [25] (Neal, 1999), где, по-видимому впервые, была предложена его формулировка через минимизацию функции свободной энергии.

Оценка параметров по Байесу. Семь раз отмерь...

Как говорится, «семь раз отмерь — один отрежь». Спрашивается: «В каком месте резать»? В данном разделе мы рассмотрим этот вопрос с позиций байесовского обучения.

Рассмотрим следующую классическую задачу оценки параметров. Пусть у нас имеется набор d-мерных векторов: $D = \left\{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(N)}\right\}$ — результаты измерений с погрешностью для некоторой величины. Мы полагаем, что погрешности вносятся неким случайным шумом, т. е. модель происхождения данных имеет вид:

$$P\left(\mathbf{y}\,|\mathbf{h}, H\right) = \mathbf{h} + \eta\left(H\right) \ ,$$

где η — модель шумовых погрешностей, а роль гипотезы h играет наша оценка $\mathbf{h}$ истинного значения измеряемой величины.

Посмотрим сначала, как влияет выбор модели искажения данных H на оценку параметра $\mathbf{h}$, т. е. того, «где резать». Степень соответствия этой модели имеющимся данным покажет Evidence. Она же будет критерием сравнения разных моделей шума.

Оценка параметра в разных моделях

Допустим, у нас есть две модели — гауссов и лапласов шум амплитуды β^{-1}, соответственно:

$$P\left(\mathbf{y}\,|\mathbf{h}, \beta, H_G\right) = \left(\frac{\beta}{2\pi}\right)^{d/2} \exp\left(-\frac{\beta}{2}\sum_{n,i}\left(y_i^{(n)} - h_i\right)^2\right),$$

$$P\left(\mathbf{y}\,|\mathbf{h}, \beta, H_L\right) = \left(\beta/2\right)^d \exp\left(-\beta\sum_i |y_i - h_i|\right) \ .$$

В силу независимости погрешностей отдельных измерений, правдоподобие объяснения всего массива данных в обеих моделях есть произведение вероятностей:

$$P\left(D\,|\mathbf{h},\beta,H_G\right)=\left(\frac{\beta}{2\pi}\right)^{Nd/2}\exp\left(-\frac{\beta}{2}\sum_{n,i}\left(y_i^{(n)}-h_i\right)^2\right),$$

$$P\left(D\,|\mathbf{h},\beta,H_L\right)=(\beta/2)^{Nd}\exp\left(-\beta\sum_{n,i}\left|y_i^{(n)}-h_i\right|\right)\ .$$

По Байесу, вероятностное распределение оценки дается выражением:

$$P\left(\mathbf{h}\,|D,\beta,H\right)=\frac{P\left(D\,|\mathbf{h},\beta,H\right)P\left(\mathbf{h}\right)}{\int d\mathbf{h}\;P\left(D\,|\mathbf{h},\beta,H\right)P\left(\mathbf{h}\right)}\ .$$

Пусть для начала у нас нет никаких априорных знаний об истинном значении оцениваемого параметра, т. е. $P\left(\mathbf{h}\right)=const$. Тогда можно считать, что вероятность гипотез в обеих моделях нам известна, а наиболее вероятную оценку получаем приравнивая нулю ее логарифмическую производную по $\mathbf{h}$. Легко показать, что для гауссовой модели наилучшая оценка — центр тяжести имеющихся измерений

$$0=\frac{\partial}{\partial h_i}\sum_{n,i}\left(y_i^{(n)}-h_i\right)^2\quad\Longrightarrow\quad\mathbf{h}^{ML}=\langle\mathbf{y}\rangle=\frac{1}{N}\sum_n\mathbf{y}^{(n)}\ ,$$

тогда как для лапласовской — медиана (когда для каждой компоненты число измерений, превышающих оценочное, равно числу измерений меньших оценочного):

$$0=\frac{\partial}{\partial h_i}\sum_{n,i}\left|y_i^{(n)}-h_i\right|\quad\Longrightarrow\quad h_i^{ML}=\frac{1}{N}med\left\{y_i^{(n)}\right\}\ .$$

Такую оценку называют еще *робастной*, поскольку она слабо чувствительна к большим выбросам (лапласовский шум допускает гораздо большие выбросы, чем гауссов).

Заметим, что обе оценки не зависят от амплитуды шума. Однако, чтобы выбрать какая из них больше соответствует реальности, нам необходимо вычислить Evidence, которая зависит от этого параметра модели.

Покажем как это делается на примере гуассовой модели.

Оценка шума

Логарифм Evidence для гауссовой модели равен:

$$\ln P\left(D\,|\beta, H_G\right) = \ln \int d\mathbf{h}\; P\left(D\,|\mathbf{h}, \beta, H_G\right) = \\ = \frac{(N-1)\,d}{2} \ln\left(\frac{\beta}{2\pi}\right) - \frac{d}{2}\ln N - \frac{\beta N d}{2}\sigma_y^2\,,$$

где $\sigma_y^2 \equiv \left\langle \left(y_i^{(n)}\right)^2 \right\rangle - \left\langle y_i^{(n)} \right\rangle^2$ — разброс значений каждой компоненты измеряемой величины. Отсюда оптимальная оценка уровня шума для гауссовой модели:

$$0 = \frac{\partial}{\partial\beta} \ln P\left(D\,|\beta, H_G\right) \quad \Longrightarrow \quad \beta_G^{-1} = \frac{N}{(N-1)}\sigma_y^2\,. \tag{5}$$

Как видим, наиболее правдоподобное значение дисперсии случайной величины несколько больше ее эмпирической оценки. Известно, что такая оценка дисперсии — *несмещенная*, т. е. ее среднее по различным выборкам равно истинному, и на многих калькуляторах даже введена специальная функция σ_{N-1} для этой оценки дисперсии.

Логарифм Evidence оптимальной гауссовой модели равен, таким образом:

$$\ln P\left(D\,|\beta_G, H_G\right) = \frac{(N-1)\,d}{2} \ln\left(\frac{(N-1)}{2\pi e N \sigma_y^2}\right) - \frac{d}{2}\ln N\,.$$

Первое слагаемое, пропорциональное числу данных, есть длина описания данных в рамках оптимальной модели, а второе — длина описания оптимальной модели. Заметим, что такое выражение для оптимальной сложности модели является весьма общим [12]. Как показал Риссанен [Rissanen 1978], в любой параметрической модели среднее число бит, приходящееся на описание одного параметра оптимальной модели, равно $\log\sqrt{N}$. Действительно, нет нужды тратить лишние биты, зная, что ошибка оценки параметров модели убывает со скоростью $\propto 1\Big/\sqrt{N}$.

[12] Этот член с неизбежностью возникает при взятии d-мерного интеграла для Evidence (см. Подробности, параграф об информационных критериях).

Проверка априорных гипотез

Рассмотрим теперь другую задачу. Допустим, что какое-то значение оцениваемого вектора априори выделено, например, является неким теоретическим предсказанием, подлежащим экспериментальной проверке. Перенеся начало координат в это выделенное значение, нашу задачу можно сформулировать следующим образом. Мы хотим проверить гипотезу $h_0 : \mathbf{h} = 0$, против альтернативной гипотезы $h_1 : \mathbf{h} \neq 0$. Очевидно, что принцип Maximal Likelihood здесь не подходит, так как согласно ему гипотеза h_0 выиграет лишь в случае $\langle \mathbf{y} \rangle = 0$, имеющем нулевую вероятность. Ясно, что мы должны как-то учесть имеющуюся у нас априорную информацию, не навязывая ее, тем не менее, в качестве результата эксперимента.

Выберем поэтому некоторую функцию распределения в пространстве гипотез, например гауссову с неизвестным пока параметром регуляризации α:

$$P(\mathbf{h}\,|\alpha) = \frac{1}{Z_\alpha} \exp\left(-\frac{\alpha}{2}\mathbf{h}^2\right), \quad Z_\alpha = \left(\frac{\alpha}{2\pi}\right)^{-d/2} .$$

Если для шума также выбрана гауссова модель

$$P(D\,|\mathbf{h}, \beta) = \frac{1}{Z_\beta} \exp\left(-\frac{\beta}{2} \sum_n \left(\mathbf{y}^{(n)} - \mathbf{h}\right)^2\right),$$

$$Z_\beta = \left(\frac{\beta}{2\pi}\right)^{-Nd/2},$$

то апостериорная вероятность оценки будет иметь вид:

$$P(\mathbf{h}\,|D, \beta, \alpha) = \frac{P(D\,|\mathbf{h}, \beta)\, P(\mathbf{h}\,|\alpha)}{\int d\mathbf{h}\; P(D\,|\mathbf{h}, \beta)\, P(\mathbf{h}\,|\alpha)} =$$

$$= \frac{1}{Z_{\alpha,\beta}} \exp\left(-\frac{\beta}{2} \sum_{n,i} \left(y_i^{(n)} - h_i\right)^2 - \frac{\alpha}{2} \sum_i h_i^2\right) .$$

(Значение нормировочного интеграла $Z_{\alpha,\beta}$, как и детали последующих выкладок, можно найти в разделе Подробности.)

Наиболее вероятная оценка, максимизирующая апостериорное распределение, есть:

$$\mathbf{h}_{MP} = \frac{\beta N}{\beta N + \alpha} \langle \mathbf{y} \rangle \ .$$

А значение Evidence

$$P\left(D\,|\beta,\alpha\right) = \int d\mathbf{h}\ P\left(D\,|\mathbf{h},\beta\right) P\left(\mathbf{h}\,|\alpha\right) = \frac{Z_{\alpha,\beta}}{Z_\alpha Z_\beta}$$

достигает максимума при следующих значениях α и β [13]:

$$\beta_{ML}^{-1} = \frac{N}{N-1}\sigma_y^2 \ ,$$

$$\alpha_{ML} = \begin{cases} \left(\langle \mathbf{y} \rangle^2 - \sigma_{\mathbf{y}}^2/(N-1)\right)^{-1}, & \langle \mathbf{y} \rangle^2 > \sigma_{\mathbf{y}}^2/(N-1) \\ \infty \ , & \langle \mathbf{y} \rangle^2 \leq \sigma_{\mathbf{y}}^2/(N-1) \end{cases}$$

где $\sigma_{\mathbf{y}}^2 = \sigma_y^2 d$ — полная дисперсия всех компонент данных. Таким образом, первоначальную гипотезу можно считать подтвержденной, если квадрат среднего отклонения от теоретического значения существенно, как минимум в N раз, меньше полной дисперсии данных. В этом случае эмпирические данные не дают достаточных оснований для пересмотра теоретического значения оцениваемой величины:

$$\mathbf{h}_{MP} = 0, \quad \langle \mathbf{y} \rangle^2 \leq \sigma_{\mathbf{y}}^2/(N-1) \ .$$

Существенные отклонения от теоретического значения, согласно байесовскому подходу, дают такие экспериментальные данные, при которых среднее значение превышает пороговый уровень, обратно пропорциональный корню числа данных. Наиболее вероятная оценка сдвинута к априорному значению и отличается от среднего тем больше, чем ближе к пороговому уровень шума в данных:

$$\mathbf{h}_{MP} = \left(1 - \frac{\sigma_{\mathbf{y}}^2}{(N-1)\langle \mathbf{y} \rangle^2}\right) \langle \mathbf{y} \rangle \ , \quad \langle \mathbf{y} \rangle^2 > \sigma_{\mathbf{y}}^2/(N-1) \ . \qquad (6)$$

[13] Выбранные для оптимальных параметров регуляризации обозначения напоминают, что максимизация Evidence соответствует принципу Maximal Likelihood в пространстве моделей.

Заметим, что в математической статистике уверенность в гипотезе $\mathbf{h}_{MP} = 0$ также зависит, согласно t-критерию Стьюдента, от того, насколько мала величина $t_N^2 = \langle \mathbf{y} \rangle^2 (N-1) \Big/ \sigma_{\mathbf{y}}^2$. Байесовское рассмотрение показывает, что наличие минимальных априорных знаний относительно выделенного значения величины вносит пороговый эффект в процесс проверки гипотез.

Резюме

Как видим, вопрос «где резать?» не столь уж и тривиален. Однако байесовский подход позволяет дать на него обоснованный ответ при различных моделях зашумления данных, одновременно оценивая их достоверность. Мы выяснили, что при проверке априорных гипотез существует пороговый эффект, отличающий значимые экспериментальные данные от незначимых. Аналогичный эффект обнуления значений незначимых параметров модели мы встретим в следующем разделе при обсуждении более сложной проблемы интерполяции функций.

История и библиография

Оценка математического ожидания и дисперсии неизвестного распределения по конечной выборке является одной из классических задач математической статистики. Отметим лишь некоторые относящиеся к нашему рассмотрению результаты.

Известно, например, что эмпирическое арифметическое среднее является несмещенной состоятельной оценкой для любого распределения с конечным математическим ожиданием. Это означает, что для различных выборок эта оценка колеблется около истинного значения, а при бесконечной выборке стремится к нему. Однако, таких оценок существует множество и в статистике принято выбирать такие, которые при этом обладают наименьшей дисперсией.

Оказывается, что эмпирическое среднее обладает наименьшей дисперсией лишь для гауссова распределения (в полном соответствии с нашим рассмотрением) [15] (Kagan, 1965). Более того, большим сюрпризом для статистиков оказалось, что если отказаться от свойства несмещенности, то даже для многомерного гауссова распределения при размерности

векторов больше двух можно найти оценку с меньшей дисперсией [34] (Stein, 1956). Пример такой оценки, смещенной к началу координат аналогично (6), приведен в книге [46] (Секей, 1990). Байесовский подход позволяет найти обоснованную смещенную оценку в случае, когда для такого смещения имеется причина.

Что касается дисперсии, то хорошо известно, что оценка (5) является несмещенной, а соответствующий множитель $N/(N-1)$ называется *множителем Бесселя*, (см., например, [44], [45] (Кокс 1978, 1984)).

Байесова интерполяция функций. Без кросс-валидации

Задачу интерполяции функций можно рассматривать как обобщение задачи оценки параметра. Вместо оценки одного зашумленного значения, она подразумевает восстановление зашумленной функции. Соответственно, данные в этом случае являются примерами функциональной зависимости, т. е. парами значений $D = \left\{\mathbf{y}^{(n)}, \mathbf{x}^{(n)}\right\}_{n=1}^{N}$, и мы пытаемся смоделировать условное распределение вероятности $P(\mathbf{y}\,|\mathbf{x}, h, H)$ — зависимость зашумленных *выходов* $\mathbf{y}$ от *входов* $\mathbf{x}$. В качестве гипотезы h мы будем рассматривать функцию $\mathbf{h}(\mathbf{x}, \mathbf{w})$, с настроечными параметрами $\mathbf{w}$, а модель H определяет ограничения на вид функций и параметры шумовой компоненты. Такую задачу восстановления зашумленной функции называют также *регрессионным анализом*.

Постановка задачи

Для простоты рассмотрим случай d-мерных входов и скалярного выхода. В качестве пространства гипотез выберем W-параметрическое семейство функций $h : y = h(\mathbf{x}, \mathbf{w})$ с заданными ограничениями на значения ее параметров $\mathbf{w}$. Например, в случае нейросетевой аппроксимации $\mathbf{w}$ есть набор всех настроечных *синаптических весов* (*synaptic weights*). На эту функцию накладывается шум $y = h(\mathbf{x}, \mathbf{w}) + \eta(\beta)$ с характерной «температурой» β^{-1}, и функцией распределения $P_\eta(x) \propto \exp(-\beta E(x))$.

Правдоподобие объяснения имеющихся данных $P(D|h,H)$ примет вид:

$$P(D|\mathbf{w},\beta) = \prod_{n=1}^{N} P\left(y^{(n)}\left|\mathbf{x}^{(n)},\mathbf{w},\beta\right.\right) \propto$$
$$\propto \exp\left[-\beta\sum_{n} E\left(y^{(n)} - h\left(\mathbf{x}^{(n)},\mathbf{w}\right)\right)\right] .$$

Запишем это выражение в более компактной форме:

$$P(D|\mathbf{w},\beta) = \frac{1}{Z_\beta}\exp\left(-\beta E_D(\mathbf{w})\right), \quad Z_\beta = \int dD\,\exp\left(-\beta E_D(\mathbf{w})\right) .$$

Аналогичным образом можно сформулировать и априорные ограничения $P(h|H)$, характеризуемые *параметром регуляризации* α:

$$P(\mathbf{w}|\alpha) = \frac{1}{Z_\alpha}\exp\left(-\alpha E_W(\mathbf{w})\right), \quad Z_\alpha = \int d\mathbf{w}\,\exp\left(-\alpha E_W(\mathbf{w})\right) .$$

Решение в общем виде

В соответствии с байесовским подходом, решение задачи в общем виде дается апостериорным распределением вероятностей:

$$P(\mathbf{w}|D,\beta,\alpha) = \frac{P(D|\mathbf{w},\beta)\,P(\mathbf{w}|\alpha)}{P(D|\beta,\alpha)} ,$$
$$P(D|\beta,\alpha) = \int d\mathbf{w}\,P(D|\mathbf{w},\beta)\,P(\mathbf{w}|\alpha) ,$$

причем оптимальные значения параметров α,β максимизируют значение Evidence:

$$(\beta_{ML},\alpha_{ML}) = \arg\max_{\beta,\alpha} P(D|\beta,\alpha) = \arg\max_{\beta,\alpha}\frac{Z_{\alpha,\beta}}{Z_\alpha Z_\beta}$$

выраженное через статсуммы:

$$Z_{\alpha,\beta} = \int d\mathbf{w}\,\exp\left(-\beta E_D(\mathbf{w}) - \alpha E_W(\mathbf{w})\right) ,$$

$$Z_\alpha = \int d\mathbf{w}\ \exp\left(-\alpha E_W\left(\mathbf{w}\right)\right), \quad Z_\beta = \int dD\ \exp\left(-\beta E_D\left(\mathbf{w}\right)\right).$$

Наилучшая гипотеза в наилучшей модели соответствует функции $h\left(\mathbf{x}, \mathbf{w}_{MP}\right)$:

$$\begin{aligned}\mathbf{w}_{MP} &= \arg\max_{\mathbf{w}} P\left(\mathbf{w}\,|D, \beta_{ML}, \alpha_{ML}\right) \\ &= \arg\min_{\mathbf{w}} \left(\beta_{ML} E_D\left(\mathbf{w}\right) + \alpha_{ML} E_W\left(\mathbf{w}\right)\right).\end{aligned}$$

Вычисление методом перевала

Все, что нам нужно для решения, это суметь вычислить определенные выше статсуммы для данного типа моделей. При этом статсуммы Z_α и Z_β не зависят от данных и их можно вычислить точно, выбрав для моделирования подходящие функции $E_D\left(\mathbf{w}\right)$ и $E_W\left(\mathbf{w}\right)$. Например, для гауссова шума:

$$\begin{aligned}Z_\beta &= \int dD\ \exp\left(-\beta E_D\left(\mathbf{w}\right)\right) = \\ &= \prod_n \int dy^{(n)}\ \exp\left[-\frac{\beta}{2}\left(y^{(n)} - h\left(\mathbf{x}^{(n)}, \mathbf{w}\right)\right)^2\right] = \left(\frac{\beta}{2\pi}\right)^{-N/2}\end{aligned}$$

Аналогично, для гауссовой величины Prior:

$$Z_\alpha = \int d\mathbf{w}\ \exp\left(-\alpha E_W\left(\mathbf{w}\right)\right) = \int d\mathbf{w}\ \exp\left(-\frac{\alpha}{2}\mathbf{w}^2\right) = \left(\frac{\alpha}{2\pi}\right)^{-W/2}.$$

Сложнее обстоит дело с интегралом $Z_{\alpha,\beta}$, поскольку функция $E_D\left(\mathbf{w}\right)$ зависит от настроечных весов $\mathbf{w}$ сложным образом — через функцию $h\left(\mathbf{x}, \mathbf{w}\right)$. В этом существенное отличие аппроксимации функций от оценки параметров, где интеграл $Z_{\alpha,\beta}$ также был гауссовым. Можно, однако, попытаться вычислить этот интеграл приближенно, *методом перевала*, воспользовавшись тем, что он содержат в экспоненте большой множитель $N \gg 1$ и, следовательно, имеет острый пик вблизи своего максимума. Раскладывая выражение под экспонентой в ряд в окрестности $\mathbf{w}_{MP}$ и ограничиваясь квадратичными членами, получим следующее приближенное выражение для логарифма Evidence:

$$\ln P\left(D\,|\beta,\alpha\right) = \ln Z_{\alpha,\beta} - \ln Z_\alpha - \ln Z_\beta =$$

$$= -\alpha E_W^{MP} - \beta E_D^{MP} - \frac{1}{2}\ln|\mathbf{A}| + \frac{W}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{N}{2}\ln\left(2\pi\right)\ , \quad (7)$$

где $|\mathbf{A}|$ — детерминант матрицы вторых производных функции $\beta E_D\left(\mathbf{w}\right) + \alpha E_W\left(\mathbf{w}\right)$ в точке ее минимума

$$\beta\,\nabla E_D\left(\mathbf{w}_{MP}\right) = -\alpha\,\nabla E_W\left(\mathbf{w}_{MP}\right) = -\alpha\mathbf{w}_{MP}\ , \quad (8)$$

$$\mathbf{A} = \nabla\nabla\left(\beta E_D\left(\mathbf{w}\right) + \alpha E_W\left(\mathbf{w}\right)\right)_{\mathbf{w}_{MP}}\ . \quad (9)$$

Детерминант матрицы равен произведению ее собственных значений. В случае квадратичного $E_W\left(\mathbf{w}\right)$ их легко выразить через собственные значения λ_i матрицы $\beta\nabla\nabla E_D\left(\mathbf{w}_{MP}\right)$, которую при обучении нейросетей можно вычислять методом *обратного распространения ошибок* (*error back propagation*):

$$|\mathbf{A}| = \prod_{i=1}^{W}\left(\,\lambda_i + \alpha\right).$$

Приравнивая нулю производные логарифма Evidence по α и β, находим их оптимальные значения:

$$2\alpha_{ML}E_W^{MP} = \sum_{i=1}^{W}\frac{\lambda_i}{\lambda_i + \alpha_{ML}} \equiv \mathcal{W}\ , \quad (10)$$

$$2\beta_{ML}E_D^{MP} = N - \mathcal{W}\ . \quad (11)$$

Здесь $\mathcal{W}$ играет роль эффективного числа параметров, участвующих в обучении — таких, для которых собственные значения $\lambda_i > \alpha$. Действительно, если $\lambda_i \ll \alpha$, значит точность определения веса w_i из имеющихся эмпирических данных существенно ниже, чем характерный масштаб синаптических весов.

Выражения (10),(11) аналогичны известному физическому факту: средняя энергия на одну степень свободы равна $T/2$. В нашем случае удвоенная суммарная безразмерная «энергия» оптимальной модели равна общему числу примеров:

$$2\beta_{ML}E_D^{MP} + 2\alpha_{ML}E_W^{MP} = N\ .$$

Однако благодаря тому, что часть информации тратится на определение параметров гипотезы, ожидаемое значение дисперсии ошибки превышает ее эмпирическую оценку тем больше, чем сложнее модель данных:

$$\frac{1}{\beta_{ML}} = \frac{1}{N-\mathcal{W}} \sum_n \left(y^{(n)} - h^{(n)} \right)^2 = \frac{N}{N-\mathcal{W}} \sigma_y^2 \, .$$

Эта оценка дисперсии обобщает аналогичное выражение (5) при измерении зашумленного параметра, с той разницей, что более сложная модель требует большего числа данных для фиксирования своих параметров. Как видим, оценка каждого существенного параметра модели уменьшает число «неиспользованных» данных на единицу.

Длина описания данных наилучшей моделью равна:

$$\begin{aligned} & L\left(D\,|\beta_{ML}, \alpha_{ML}\right) = -\ln P\left(D\,|\beta_{ML}, \alpha_{ML}\right) \simeq \\ \simeq & N\left(2 + \frac{1}{2}\ln\left(2\pi\right) + \frac{1}{2}\ln\sigma_y^2\right) + \frac{1}{2}\mathcal{W} + \frac{1}{2}\sum_{i=1}^{W} \ln\left(1 + \frac{\lambda_i}{\alpha}\right) \end{aligned} \qquad (12)$$

Первое слагаемое, пропорциональное N, отвечает описанию отклонений значений данных от предсказаний модели. Остальные два — длине описания модели, пропорциональной эффективному числу ее параметров $\mathcal{W}$. Действительно, поскольку члены последней суммы, для которых $\lambda_i \ll \alpha$, пренебрежимо малы, можно считать, что число членов в ней равно $\mathcal{W}$. Поскольку, кроме того, все λ_i пропорциональны числу примеров N, мы опять получаем, что длина описания оптимальной модели $\sim \mathcal{W} \ln\sqrt{N}$ (более развернутую интерпретацию см. в Подробностях).

Предварительное обсуждение

В предыдущем разделе мы рассмотрели пример аппроксимации зашумленного синуса, чтобы проиллюстрировать необходимость регуляризации обучения. Теперь мы можем убедиться в этом, исходя из полученных выше выражений. Так, если вовсе отказаться от регуляризации, устремив $\alpha \to 0$, то Evidence (7) также устремится к нулю, а длина описания (12) — к бесконечности. Это возможно даже для гипотез, описываемых лишь одним параметром!

Напомним, что все проделанные (следуя [22] (MacKay, 1992)) выкладки относятся лишь к одному локальному максимуму апостериорной плотности в пространстве гипотез, тогда как таких максимумов может быть много и их вклады в Evidence суммируются. Наличие многократно вырожденных состояний может быть следствием симметрии модели. Следовательно, чем симметричнее модель, т. е. чем больше кратность повторения пиков, тем больше ее Evidence. С этой точки зрения, максимизация Evidence содержит в себе и «эстетическую» компоненту.

Заметим также для справки, что полученное нами приближенное выражение (12) для длины описания данных оптимальной моделью является более подробной версией часто встречающегося в литературе асимптотического *байесова информационного критерия* сравнения гипотез (см. Подробности: *Bayesian Information Criterion*).

Итерационное обучение

Систему уравнений (8)–(11) для определения оптимальных параметров модели можно решать итерациями в духе EM-алгоритма. На первом этапе каждой итерации фиксируются параметры наилучшей модели и находится наилучшая гипотеза $\mathbf{w}_{MP}$, минимизирующая функцию:

$$\beta E_D(\mathbf{w}) + \alpha E_W(\mathbf{w}) = \frac{1}{2}\left((N-\mathcal{W})\frac{E_D(\mathbf{w})}{E_D^{MP}} + \mathcal{W}\frac{E_W(\mathbf{w})}{E_W^{MP}}\right).$$

В этой точке вычисляются новые значения E_W^{MP}, E_D^{MP}, а также вычисляется матрица вторых производных, определяющая новое значение $\mathcal{W}$. Эти итерации продолжаются до достижения стационарной точки.

С практической точки зрения, на начальных стадиях для ускорения обучения можно вообще не вычислять матрицу вторых производных, заменяя приближенно $\mathcal{W} \simeq W$. В дальнейшем можно ограничиться следующим приближением. Для гауссовой эмпирической ошибки $E_D(\mathbf{w}) = \frac{1}{2}\sum_n \left(\varepsilon^{(n)}\right)^2$ матрица вторых производных равна:

$$\frac{\partial^2 E_D}{\partial w_i \partial w_j} = \sum_n \left(\frac{\partial \varepsilon^{(n)}}{\partial w_i}\frac{\partial \varepsilon^{(n)}}{\partial w_j} + \varepsilon^{(n)}\frac{\partial^2 \varepsilon^{(n)}}{\partial w_i \partial w_j}\right).$$

Поскольку среднее значение *невязки* $\varepsilon^{(n)} \equiv y^{(n)} - h^{(n)}$ для оптимальной модели стремится к нулю, вторым слагаемым можно пренебречь, как это

делается в методе Левенберга–Марквардта. Тем самым, матрицу вторых производных можно приближенно выразить через первые производные, вычисляемые при поиске $\mathbf{w}_{MP}$.

Лапласовский Prior и прореживание модели

Регуляризация обучения, как мы убедились выше, приводит к эффективному уменьшению числа параметров модели до значения, соответствующего эмпирическим данным. Некоторые линейные комбинации весов являются «лишними» и в процессе обучения автоматически уменьшаются. Этот эффект для случая двух синаптических весов иллюстрирует рис. 6.

Это относится, однако, не к индивидуальным весам, а к их комбинациям. Сами веса могут при этом быть не малы. Между тем, для некоторых приложений желательно сделать модель как можно более «прозрачной», уменьшив число ее параметров до необходимого минимума. Это позволяет не просто построить модель данных, но и в явном виде выявить присущие им закономерности.

Так, *прореживание* (*pruning*) нейросети — избавление от лишних весов — позволяет выявлять значимые для моделирования входы и выделять наиболее существенные факторы, определяющие поведение модели.

Для построения таких моделей можно использовать лапласовский Prior:

$$E_W(\mathbf{w}) = \sum_{i=1}^{W} |w_i| \,,$$

$$Z_\alpha = \int d\mathbf{w}\, \exp\left(-\alpha E_W(\mathbf{w})\right) = (\alpha/2)^{-W} \,.$$

В отличие от гауссовой, лапласовская модель характеризуется одинаковой чувствительностью эмпирической ошибки ко всем синаптическим весам. Действительно, в стационарной точке

$$\nabla\left(\beta E_D(\mathbf{w}) + \alpha E_W(\mathbf{w})\right)\big|_{\mathbf{w}_{MP}} = 0 \,,$$

откуда:

$$\left|\frac{\partial E_D(\mathbf{w}_{MP})}{\partial w_i}\right| = \frac{\alpha}{\beta} \,.$$

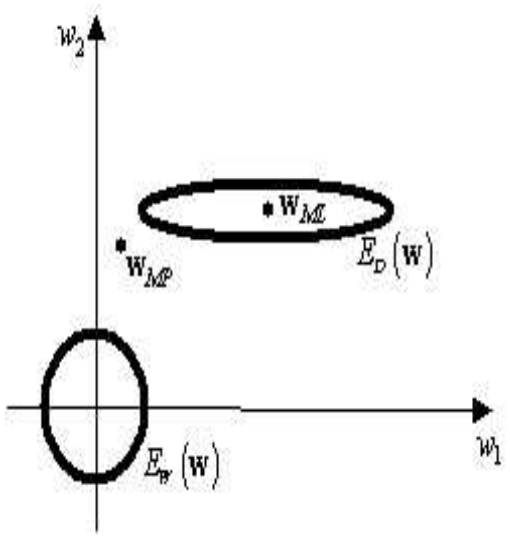

Рис. 6. Пространство параметров после перехода в систему главных осей матрицы вторых производных функции ошибки. Обозначены контуры гауссова Prior E_W и эмпирического Likelihood E_D. Горизонтальному направлению соответствует малое собственное значение $\lambda_1 \ll \alpha$. Соответственно, комбинация весов w_1 плохо определена имеющимися данными, и в наиболее правдоподобной гипотезе эта компонента практически исчезает. Напротив, комбинация весов w_2 надежно определяется из данных, и ее оценка слабо искажается регуляризацией

Веса, которые не могут обеспечить такую чувствительность обращаются в нуль согласно градиентному алгоритму обучения

$$\frac{\partial \mathbf{w}}{\partial t} = -\nabla \left(\beta E_D\left(\mathbf{w}\right) + \alpha E_W\left(\mathbf{w}\right)\right) = -\beta \nabla E_D\left(\mathbf{w}\right) - \alpha \ ,$$

который приводит к линейному по времени затуханию любого веса, чувствительность к которому у ошибки меньше α/β.

Через конечное время такой вес с неизбежностью обращается в нуль. В этом существенное отличие лапласовской регуляризации от гауссовой.

Найдем оптимальные параметры модели в этом случае. Теперь логарифм Evidence выглядит следующим образом:

$$\ln P(D|\beta,\alpha) = \ln Z_{\alpha,\beta} - \ln Z_\alpha - \ln Z_\beta =$$
$$= -\alpha E_W^{MP} - \beta E_D^{MP} - \frac{W}{2}\ln\beta + W\ln\alpha + \frac{N}{2}\ln\beta -$$
$$- \left[\frac{N}{2}\ln(2\pi) + W\ln 2 + \frac{1}{2}\ln|\nabla\nabla E_D(\mathbf{w})|\right],$$

где выражение в квадратных скобках уже не зависит от α и β. Таким образом, в этом случае детерминант матрицы вторых производных уже не влияет на процедуру оптимизации! Выражения для оптимальных параметров:

$$\alpha_{ML} E_W^{MP} = W,$$
$$2\beta_{ML} E_W^{MP} = N - W,$$

аналогичны (10),(11), но зависят уже не от эффективного, а от общего количества ненулевых весов, определяемого в этом случае в процессе поиска стационарной точки $\mathbf{w}_{MP}$.

Таким образом, лапласовская регуляризация подразумевает, во-первых, более простой алгоритм обучения, не требующий вычисления матрицы вторых производных, и, во-вторых, приводит к оптимальному прореживанию модели, оставляя в ней лишь наиболее значимые для объяснения данных параметры.

Оценка ошибок предсказаний

Итак, у нас есть рецепт нахождения наиболее вероятной гипотезы с оптимальными параметрами регуляризации, способной предсказывать значения выходов для любых входов. Однако, заблуждение, по словам Спинозы, это «истина, взятая вне пределов своей применимости». Т. е. предсказание без оценки ошибок — это еще не предсказание.

Байесовский подход позволяет получить не только предсказания, но и их ожидаемый разброс. Во-первых, найденное значение β_{ML}^{-1} дает оценку шумовой составляющей в данных, т. е. нижнюю границу разброса предсказаний (ведь шум по определению не предсказуем). Однако, шум — не

единственный источник неопределенности. Вспомним, что в байесовской модели в предсказаниях участвуют все гипотезы с их апостериорными вероятностями, а не только наиболее вероятная из них. Соответственно, чем шире пик функции $P(\mathbf{w}\,|D, \beta_{ML}, \alpha_{ML})$ вокруг своего максимума в точке $\mathbf{w}_{MP}$, тем больше разброс предсказаний ансамбля. Такая ситуация характерна для областей данных, далеких от имеющихся примеров. Рассмотрим этот вопрос на количественном уровне.

Байесовские предсказания дают не просто значение функции, а плотность вероятности ее распределения:

$$P(y\,|\mathbf{x}, D) = \int d\mathbf{w}\, P(y\,|\mathbf{x}, \mathbf{w})\, P(\mathbf{w}\,|D)\,.$$

При гауссовом шуме в квадратичном приближении для логарифма апостериорной вероятности получаем (опуская все константы):

$$P(y\,|\mathbf{x}, D) \propto \int d\mathbf{w}\, \exp\left[-\frac{\beta}{2}\left(y - h(\mathbf{x}, \mathbf{w})\right)^2 - \frac{1}{2}\Delta\mathbf{w}^T\mathbf{A}\Delta\mathbf{w}\right].$$

Если ширина пика в пространстве гипотез, определяемого матрицей вторых производных A, достаточно мала, можно ограничиться первым членом разложения функции $h(\mathbf{x}, \mathbf{w})$ в его окрестности:

$$h(\mathbf{x}, \mathbf{w}) \simeq h(\mathbf{x}, \mathbf{w}_{MP}) + \mathbf{g}\Delta\mathbf{w}, \quad \mathbf{g} \equiv \nabla_{\mathbf{w}} h|_{\mathbf{w}_{MP}}\,.$$

В этом приближении предсказания ансамбля гипотез будут нормально распределены вокруг предсказания наилучшей гипотезы:

$$P(y\,|\mathbf{x}, D) = \frac{1}{\sqrt{2\pi\,\sigma_y^2}} \exp\left(-\frac{(y - y_{MP})^2}{2\sigma_y^2}\right).$$

И разброс предсказаний характеризуется соответствующей дисперсией [Bishop 1995]:

$$\sigma_y^2 = \beta_{ML}^{-1} + \mathbf{g}^T\mathbf{A}^{-1}\mathbf{g}\,.$$

Если наиболее вероятная гипотеза определяется достаточно уверенно, т. е. разброс в ансамбле гипотез невелик, то предсказания модели имеют минимальный разброс, определяемый уровнем шума. В противном случае разброс предсказаний возрастает пропорционально разбросу в пространстве гипотез. Рис. 7 иллюстрирует этот подход к определению ошибок предсказаний.

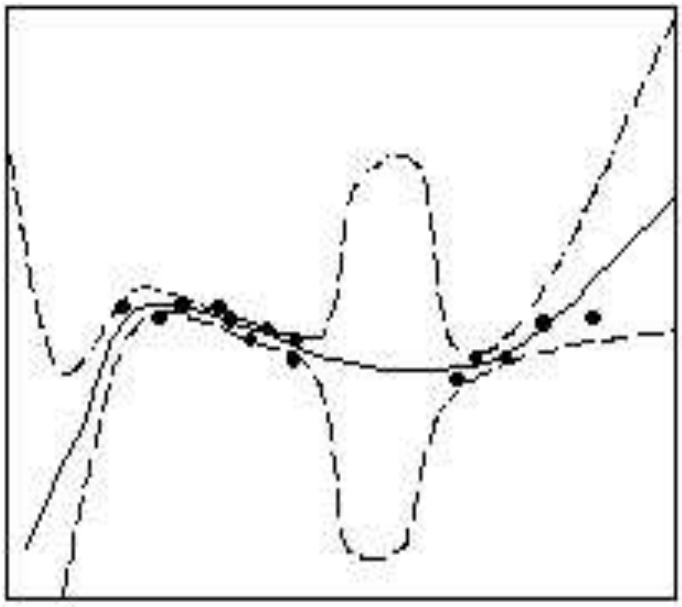

Рис. 7. Предсказания оптимальной модели и оценка разброса предсказаний. Последний возрастает вдали от области эмпирических данных.

Резюме

Таким образом, мы убедились в конструктивности байесовский подхода применительно к проблеме аппроксимации функций. Он дополняет обычные градиентные алгоритмы обучения итеративной подстройкой параметров регуляризации в духе EM-алгоритма. Этапу Expectation соответствует минимизация регуляризированной ошибки градиентными методами, а этапу Maximization — оценка оптимальных параметров регуляризации, причем для описанных выше моделей эта оценка выписывается в явном виде. В частности, лапласовская регуляризация порождает очень простой алгоритм обучения, приводящий, помимо прочего, к упрощению структуры оптимальной гипотезы — уменьшению числа синаптических

весов до необходимого минимума.

При этом, напомним, байесовский подход не предполагает кросс-валидации! Все примеры используются для обучения одновременно и синаптических весов, и параметров регуляризации. Сходимость такого EM-алгоритма гарантирована. В случае же кросс- валидации оптимизация модели вся построена на эвристиках, не гарантирующих, к тому же, нахождение оптимальной модели. Как выбирать параметры регуляризации на каждом новом цикле валидационных экспериментов? Каков размер валидационных выборок? Сколько циклов валидации достаточно для обоснованного определения качества модели с текущими параметрами регуляризации? На все эти вопросы нет теоретически обоснованных ответов. В байесовском подходе, напротив, количество итераций соответствует сложности данной задачи. Кроме того, гарантируется, что каждая следующая итерация улучшает модель.

История и библиография

Понятие регрессии в научный обиход ввел Френсис Гальтон, систематически применявший статистические методы при анализе биологических данных, многие из которых предоставлял ему его двоюродный брат Чарльз Дарвин. Гальтона также считают основоположником генетики человека. Его исследования зависимости между ростом детей и их родителей привлекли к себе всеобщее внимание. Отсюда — временна́я окраска термина, показывающего насколько те или иные характеристики возобновляются, т. е. *регрессируют* в следующих поколениях. В дальнейшем регрессией стали называть любую функциональную зависимость между случайными величинами.

Регрессионный анализ в XX веке сначала широко распространился в биологии, став основным инструментом *биометрики*. В 30-х годах Рональд Фишер по аналогии ввел термин *эконометрика*.

Авторегрессия — зависимость значения временного ряда от его же значений в предшествующие моменты времени — является основным методом прогнозирования поведения сложных динамических систем [38] (Weigend, 1994). Причем, иногда довольно сложные временные ряды могут быть описаны с помощью линейной авторегрессии, как это было продемонстрировано Юлом в 1927 году на примере предсказания ежегодного

числа солнечных пятен. В течение десятилетий линейная регрессия доминировала в решении прикладных задач. При этом линейность модели уже настолько сильно ограничивает класс решений, что дополнительной регуляризации обучения, как правило, не требовалось.

Однако, для многих практически важных задач линейной регрессии оказывается недостаточно. После открытия в 1986 году эффективного метода обучения многослойных персептронов [36] (Rumelhart, 1986), последние приобрели широкую популярность в качестве инструмента нелинейной регрессии и аппроксимации функций. Для таких моделей с потенциально очень большим числом параметров вопросы регуляризации обучения выходят на первый план.

Наиболее простым, а потому — распространенным на практике, методом регуляризации является ограничение числа скрытых нейронов с последующим сравнением моделей методом кросс-валидации. Но со временем все большую популярность приобретают идеи встраивания регуляризации непосредственно в алгоритм обучения. Так, *затухание весов* (*weight decay*) эквивалентное гауссовой регуляризации, появилось практически одновременно с методом обучения персептронов [11] (Hinton, 1987). Далее последовали различные модификации регуляризирующих функционалов [18] (Lang, 1990) и алгоритмов прореживания весов [20] (Le Cun, 1990), [9] (Hassibi, 1993). Однако, до проникновения байесовской идеологии в нейросетевое сообщество, параметры регуляризации подбирались методом кросс-валидации. В программной статье [22] (MacKay, 1992) была впервые изложена процедура обучения персептронов с внутренней оптимизацией гауссовой регуляризации. Лапласовская регуляризация, как инструмент прореживания нейросетей, была предложена в [39] (Williams, 1995).

Подробное обсуждение байесовской интерполяции функций можно найти в прекрасной книге [3] (Bishop, 1995), откуда, кстати, заимствованы иллюстрации к этому разделу.

Байесова кластеризация. Сколько кластеров «на самом деле»?

В предыдущем разделе мы рассмотрели случай, когда компоненты данных в задаче естественным образом разбиваются на входные и зависящие от них выходные. Если такое разбиение отсутствует, то все компоненты данных равнозначны, и моделирование сводится к задаче *аппроксимации плотности данных*. Существуют три основных подхода к этой проблеме — *параметрический*, *непараметрический* и промежуточный, иногда называемый *полупараметрическим*.

Параметрическая аппроксимация предполагает конкретный функциональный вид функции плотности с конечным числом подгоночных параметров: $M = const$. Например, предположению о многомерном гауссовом распределении соответствует *анализ главных компонент*. Такие модели зачастую страдают от недостатка гибкости.

Непараметрическая аппроксимация, напротив, использует для предсказания непосредственно сами данные. Типичный пример такого подхода — *ядерное сглаживание*, в котором плотность представлена совокупностью сферических источников вокруг каждой точки данных. Соответственно, сложность таких моделей растет пропорционально числу данных: $M = O(N)$, что приводит к трудностям при работе с большими базами данных.

Полупараметрические модели, как легко догадаться, призваны быть «золотой серединой». Они достаточно гибки, так как их сложность может по мере необходимости возрастать, и в то же время практичны, поскольку число свободных параметров всегда остается гораздо меньше числа данных $M = O(N^{\gamma})$, $\gamma < 1$. Однако эти достоинства имеют свою цену — сложный по сравнению с двумя другими подходами процесс обучения модели.

Примером такого рода моделей являются *гауссовы смеси*, которые и станут предметом нашего рассмотрения в этом разделе. Эмпирическая плотность в данном случае также аппроксимируется совокупностью сферических источников, соответствующих, однако, не каждой точке данных, а крупномасштабным флуктуациям плотности — кластерам. Вопрос, которым мы зададимся, касается оптимальной сложности модели: «Сколькими кластерами лучше всего описываются данные?»

Постановка задачи

Пусть набор эмпирических данных состоит из N d-мерных векторов: $D = \left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\right\}$. Требуется на основании этой выборки найти наилучшую *кластерную модель* порождения этих данных, $P(\mathbf{x}\,|h)$. Мы будем искать решение в виде смеси M независимых источников данных, каждый из которых относительно прост. А именно, вероятность порождения данных источником зависит лишь от расстояния до его центра:

$$P(\mathbf{x}\,|h) = \sum_{m=1}^{M} P(\mathbf{x}\,|m)\, P(m)\ ,$$
$$P(\mathbf{x}\,|m) \propto \exp\left(-\beta E\left(|\mathbf{x} - \mathbf{w}_m|\right)\right)\ .$$

Мы также будем полагать для определенности, что предполагаемые источники данных — гауссовы. Такая модель называется *гауссовой смесью*:

$$E\left(|\mathbf{w}_m - \mathbf{x}|\right) = \frac{1}{2}\left(\mathbf{w}_m - \mathbf{x}\right)^2\ .$$

Гипотезе h о происхождении данных соответствует набор координат этих источников и их относительная интенсивность:

$$h = \{\mathbf{w}_1, \ldots, \mathbf{w}_M,\ P(1), \ldots, P(M)\}\ .$$

Модель H определяет число M и дисперсию β^{-1} источников, а в общем случае — и конкретный вид функции ошибки E.

Как обычно, максимизация $P(h\,|D, H)$ дает наилучшую гипотезу, а максимизация $P(D\,|H)$ — наилучшую модель данных.

Оптимальная гипотеза

Допустим, у нас нет априорных предпочтений относительно различных гипотез $P(h\,|H) = \mathrm{const}$. В этом случае оптимальная гипотеза максимизирует правдоподобие данных:

$$h_{ML} = \arg\max_h \ln P(D\,|h\ , H) = \sum_n \ln \sum_m P\left(\mathbf{x}^{(n)}\,|m\right) P(m)\ .$$

Здесь мы встречаемся со знакомой ситуацией, когда под логарифмом производится суммирование по неким альтернативам (см. выше параграф про EM-алгоритм). Мы уже знаем, что такая задача сводится к минимизации «свободной энергии»:

$$F(\mathcal{P},h)=\sum_{m,n}\mathcal{P}(m\,|n\,)\ln\mathcal{P}(m\,|n\,)-\sum_{m,n}\mathcal{P}(m\,|n\,)\ln P\left(\mathbf{x}^{(n)},m\right)=$$
$$=-\sum_{n}\ln\sum_{m}P\left(\mathbf{x}^{(n)},m\right)+\sum_{m,n}\mathcal{P}(m\,|n\,)\ln\frac{\mathcal{P}(m\,|n\,)}{P\left(m\left|\mathbf{x}^{(n)}\right.\right)}\ ,$$

причем решение для $\mathcal{P}(m\,|n\,)$ дается формулой Байеса:

$$\mathcal{P}(m\,|n\,)=P\left(m\left|\,\mathbf{x}^{(n)}\right.\right)=\frac{P\left(\mathbf{x}^{(n)}\right|m\right)P(m)}{\sum_{m}P\left(\mathbf{x}^{(n)}\right|m\right)P(m)}\ ,$$

а наилучшая гипотеза — максимизацией усредненного по этому распределению логарифма совместной вероятности:

$$\mathbf{w}_m=\arg\max_{\mathbf{w}_m}\sum_{m,n}\mathcal{P}(m\,|n\,)\ln P\left(\mathbf{x}^{(n)},m\right)=$$
$$=\arg\max_{\mathbf{w}_m}\sum_{m,n}\mathcal{P}(m\,|n\,)\ln P\left(\mathbf{x}^{(n)}\,|m\right)\ ,$$
$$P(m)=\arg\max_{P(m)}\sum_{m,n}\mathcal{P}(m\,|n\,)\ln P\left(\mathbf{x}^{(n)},m\right)=$$
$$=\arg\max_{P(m)}\sum_{m,n}\mathcal{P}(m\,|n\,)\ln P(m)\ .$$

Отсюда легко получить:

$$\sum_{n}\mathcal{P}(m\,|n\,)\,\frac{\partial E\left(\left|\mathbf{w}_m-\mathbf{x}^{(n)}\right|\right)}{\partial\mathbf{w}_m}=0\ \ \Longrightarrow\ \ \mathbf{w}_m=\frac{\sum\limits_{n}P\left(m\left|\mathbf{x}^{(n)}\right.\right)\mathbf{x}^{(n)}}{\sum\limits_{n}P\left(m\left|\mathbf{x}^{(n)}\right.\right)}\ ,$$

$$P(m)=\frac{1}{N}\sum_{n}\mathcal{P}(m\,|n\,)\ .$$

Иными словами, оптимальная гипотеза находится путем последовательных итераций следующего EM-алгоритма:

- **E шаг**: Фиксируем источники $\{\mathbf{w}_m, P(m)\}$ и находим вероятности принадлежности к ним точек данных:

$$P\left(m \,\middle|\, \mathbf{x}^{(n)}\right) = \frac{P(m)\exp\left(-\beta E\left(\left|\mathbf{x}^{(n)} - \mathbf{w}_m\right|\right)\right)}{\sum\limits_m P(m)\exp\left(-\beta E\left(\left|\mathbf{x}^{(n)} - \mathbf{w}_m\right|\right)\right)} \,. \tag{13}$$

- **M шаг**: Фиксируем распределение данных по источникам $P\left(m \,\middle|\, \mathbf{x}^{(n)}\right)$ и находим новые характеристики источников:

$$\mathbf{w}_m = \frac{\sum\limits_n P\left(m \,\middle|\, \mathbf{x}^{(n)}\right)\mathbf{x}^{(n)}}{\sum\limits_n P\left(m \,\middle|\, \mathbf{x}^{(n)}\right)} \,, \tag{14}$$

$$P(m) = \frac{1}{N}\sum_n P\left(m \,\middle|\, \mathbf{x}^{(n)}\right) \,. \tag{15}$$

Повторяем эти итерации до гарантированной сходимости.

Заметим, что в пределе $\beta \to \infty$ приведенный выше алгоритм совпадает с хорошо известной кластеризацией методом K-means. А именно, на каждом шаге, во-первых, определяется жесткая привязка точек к своим кластерам:

$$P\left(m \,\middle|\, \mathbf{x}^{(n)}\right) = \delta_{m,m^{(n)}}, \quad m^{(n)} = \arg\min_m \left|\mathbf{x}^{(n)} - \mathbf{w}_m\right|$$

и, во-вторых, новые центры кластеров помещаются в центры тяжести принадлежащих им точек:

$$\mathbf{w}_m = \sum_n \delta_{m,m^{(n)}}\mathbf{x}^{(n)} \,.$$

Гауссовы смеси с конечным β осуществляют мягкую или нечеткую кластеризацию.

Сколько кластеров в данных?

Хотя выше мы считали число источников в модели известным (и равным M), на самом деле этот параметр явным образом не определен. Действительно, описанный выше EM-алгоритм допускает слияние источников.

А именно, если в какой-то момент положения двух источников совпадут: $\mathbf{w}_m^t = \mathbf{w}_k^t$, то согласно (13)–(15), эти источники сольются, т. е. и на всех последующих итерациях мы получим $\mathbf{w}_m^{t+T} = \mathbf{w}_k^{t+T}$. Таким образом, реальное число кластеров в модели может существенно отличаться от начального.

В зависимости от значения β и конкретной конфигурации данных, некоторые источники будут притягиваться друг к другу и сливаться. Численные эксперименты (см. рис. 8) показывают, что флуктуация плотности данных приводит к слиянию источников, оказавшихся в ее окрестности, если их радиус $\beta^{-1/2}$ превышает масштаб этой флуктуации. Иными словами, такой масштаб неоднородностей в данных модель уже не различает.

Таким образом, чем больше β, т. е. меньше радиус взаимодействия источников, тем больше возможное число кластеров в модели. Напротив, достаточно малые β, такие, что характерный радиус источников превышает масштаб разброса всех данных, приводят к слиянию источников в один большой кластер.

Можно взглянуть на эту ситуацию следующим образом. Допустим, что мы смотрим на имеющееся распределение данных с расстояния, пропорционального $\beta^{-1/2}$. На большом расстоянии все данные сливаются в одно пятно. По мере нашего приближения, становятся различимы все новые и новые детали неоднородностей в кластерной структуре данных, и число различимых кластеров возрастает. В пределе $\beta \to \infty$ становятся различимы отдельные точки данных, т. е. «равновесное» число кластеров стремится к числу примеров, хотя их реальное количество будет, естественно, ограничено начальным значением [14].

Оптимальная модель данных соответствует в этой аналогии оптимальному масштабу, с которого структура данных видна наилучшим образом, т. е. когда мелкомасштабные флуктуации не мешают разглядеть общую картину.

[14] Эта картина полностью аналогична череде термодинамических фазовых переходов по мере «замерзания» системы, описываемой соответствующей свободной энергией. Высокой температуре соответствует один глобальный минимум — один источник в центре тяжести всех данных. С понижением «температуры» рельеф функции свободной энергии становится все более изрезанным, и количество ее локальных минимумов, соответствующих решениям EM-алгоритма, быстро возрастает.

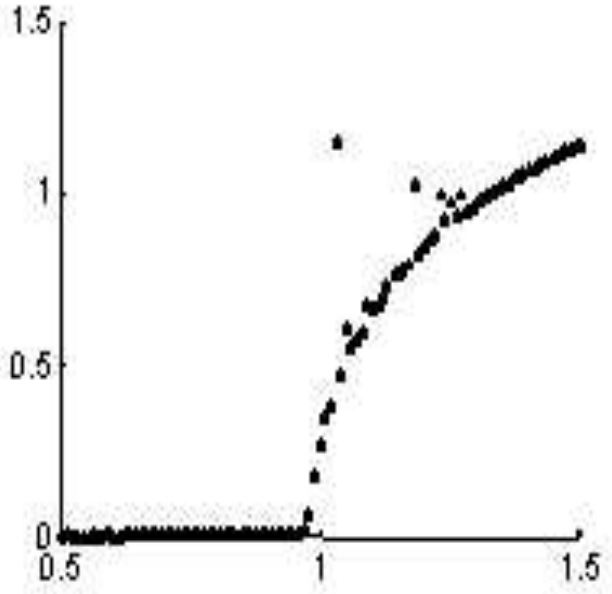

Рис. 8. Зависимость конечного расстояния между двумя центроидами от параметра β. Данные представляли собой 1000 случайных точек с двумерным гауссовым распределением единичной дисперсии. В качестве начального положения центроидов выбирались координаты двух случайно выбранных точек данных. При $\beta < 1$ источники сливаются, т. е. данные воспринимаются как единый кластер. При $\beta > 1$ становятся различимы флуктуации плотности более мелкого масштаба.

В двумерном или трехмерном случае человек легко находит наиболее информативный масштаб огрубления данных. Однако для многомерных данных определение «оптимального разрешения» модели уже не столь тривиально. Сколько кластеров в данных «на самом деле»? Байесовский подход позволяет дать ответ на этот непростой вопрос.

Оптимальная модель

Оптимальный параметр β и соответствующее ему число кластеров определяется максимизацией Evidence, которую можно вычислить приближенно, используя метод перевала:

$$P(D|\beta) = \int dh\, P(D|h,\beta) = \int dh\, \exp(\ln P(D|h,\beta)) \simeq$$
$$\simeq (2\pi)^{|h|/2} |\mathbf{A}|^{-1/2} P(D|h_{ML},\beta) \ .$$

Здесь $|h| = Md + M$ — размерность пространства гипотез, а $|\mathbf{A}|$ — детерминант матрицы вторых производных в точке максимума Likelihood: $\mathbf{A} \equiv -\nabla\nabla \ln P(D|h,\beta)|_{h_{ML}}$. Заглянув в раздел Подробности, можно убедиться, что эта матрица в нашем случае диагональна, и ее детерминант равен:

$$|\mathbf{A}| = (\beta N)^{Md} N^M \left(\prod_m P(m)\right)^{d-1}$$

С учетом этого факта, приравнивая нулю производную

$$0 = \frac{\partial}{\partial\beta} \ln P(D|\beta) \simeq \frac{\partial}{\partial\beta} \ln P(D|h_{ML},\beta) - \frac{1}{2}\frac{\partial}{\partial\beta} \ln|\mathbf{A}| \ ,$$

получим следующее выражение для оптимальной β:

$$\beta_{ML}^{-1} = -\frac{1}{(N-M)\,d} \sum_{m,n} P\left(m \,\middle|\, \mathbf{x}^{(n)}\right) \left(\mathbf{w}_m - \mathbf{x}^{(n)}\right)^2 .$$

Заметим, что если бы мы определяли β из условия максимума Likelihood, то получили бы аналогичный результат, только без уменьшения числа данных на число источников M. Что касается самого значения Evidence для оптимальной модели, то его главные члены, растущие с числом данных, даются следующим выражением:

$$\ln P(D|\beta_{ML}) \simeq$$
$$\simeq \frac{Nd}{2} \ln\beta_{ML} - \sum_{m,n} P\left(m \,\middle|\, \mathbf{x}^{(n)}\right) \ln\frac{P\left(m \,\middle|\, \mathbf{x}^{(n)}\right)}{P(m)} - \frac{M(d+1)}{2} \ln N \quad (16)$$

Более точное выражение с учетом членов следующего порядка малости приводится в разделе Подробности.

Как видим, качество модели возрастает с ростом β_{ML}, чему соответствует уменьшение масштаба ошибки. Таким образом, первый член способствует увеличению числа кластеров. Однако, второй и третий члены, напротив, «штрафуют» излишне сложные модели с большим числом кластеров. Оптимальная модель представляет собой баланс сложности модели и точности воспроизведения ею структуры данных. При наличии нескольких вариантов кластеризации предпочтение следует отдавать той модели, которой соответствует наибольшая Evidence.

Заметим, что при выводе (16) мы считали слившиеся источники единым кластером, т. е. как число M во всех формулах, так и все распределения вероятностей по кластерам соответствуют различающимся между собой источникам.

Численные эксперименты

Задача кластеризации позволяет сравнить наше интуитивное представление о качестве модели с формальным определением последнего с помощью Evidence. С этой целью мы приведем результаты трех серий численных экспериментов, в которых описанный выше EM-алгоритм применялся к трем различным двумерным распределениям данных: равномерному, гауссову и гауссовой смеси.

Начальное число источников в моделях варьировалось от 1 до 30, что составляет примерно корень от числа данных (1000 точек). По равновесным значениям числа кластеров и параметра β вычислялась Evidence, максимум которой определял наилучшую модель для всех трех выборок данных.

Рис. 9–11 показывают поведение Evidence в численных экспериментах, а также демонстрируют конфигурации источников в наилучших моделях.

Согласно численным экспериментам, для однородного распределения наилучшие модели представляют собой квазикристаллические решетки, однородно покрывающие пространство данных. Для гауссова распределения наилучшая модель, как и следовало ожидать, представляет собой один гауссовый источник. Дальнейшее усложнение модели, по Бай-

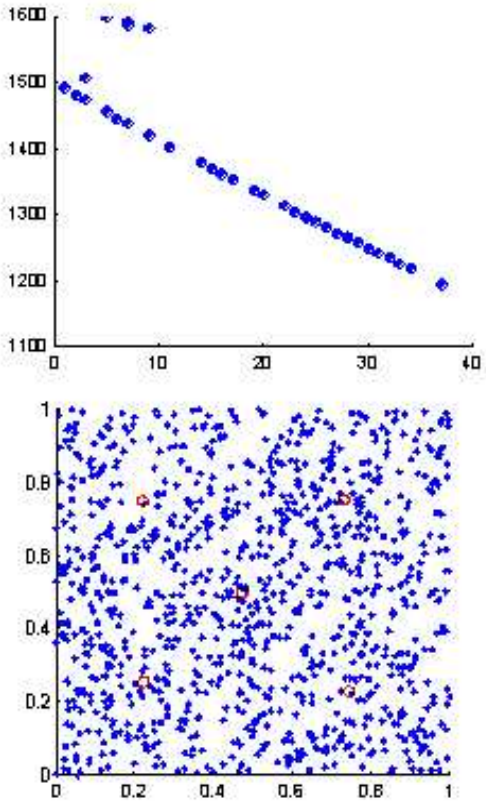

Рис. 9. Верхний график: значения Evidence для точек с однородным распределением в единичном квадрате как функция от конечного числа источников. Максимумы соответствуют моделям с однородным распределением кластеров. Внизу — пример кластеризации с наибольшей Evidence.

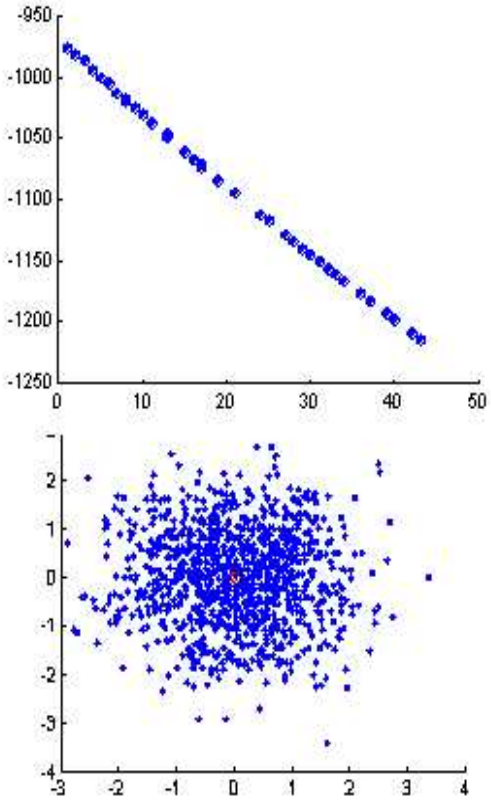

Рис. 10. Верхний график: значения Evidence для точек с гауссовым распределением как функция от конечного числа источников. Максимум соответствует модели с одним кластером. Внизу — пример кластеризации с наибольшей Evidence.

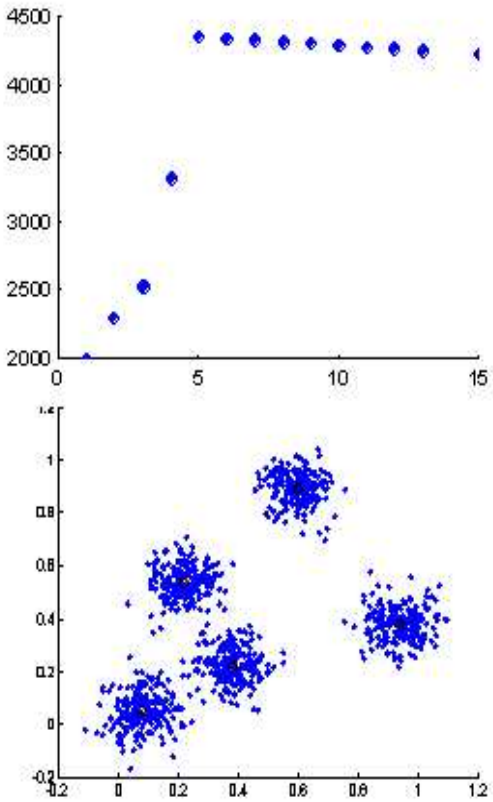

Рис. 11. Верхний график: значения Evidence для точек, порожденных пятью гауссовыми источниками. Максимум соответствует модели с пятью кластерами. Внизу — пример кластеризации с наибольшей Evidence.

есу, нецелесообразно. Для данных, порожденных гауссовой смесью с несколькими источниками, оптимальная модель правильно определяет реальное число источников. Таким образом, наши численные эксперименты показывают, что в данном случае Байесов критерий выбора наилучшей модели соответствует нашей интуиции.

Резюме

В этом разделе мы применили байесовский подход к задаче моделирования плотности данных смесью однотипных гауссовых источников. Параметром регуляризации, косвенным образом определяющим число кластеров в модели, является дисперсия источников β^{-1}. Байесовская регуляризация позволяет нам определить оптимальный масштаб огрубленного представления данных.

История и библиография

Кластеризация является одной из базовых методик обработки данных. Этому вопросу посвящена обширная литература (см., например [14] (Jain, 1988), в том числе по нечеткой кластеризации [2] (Bezdek, 1981) и смесям [23] (McLachlan, 1988) [15].

Интересно, что принцип минимальной длины описания был впервые применен именно к задаче кластеризации [37] (Wallace, 1968) под названием Minimum Message Length. Задача кластеризации при этом определялась как минимизация длины сообщения, состоящего из нескольких компонент:

- число кластеров;
- число точек принадлежащих каждому кластеру;
- центроиды кластеров;
- принадлежность каждой точки тому или иному кластеру.

[15] Библиографию по кластеризации можно найти по адресу:
URL: ftp://ftp.sas.com/pub/neural/clus-bib.txt

С позиций минимизации свободной энергии, наиболее близкой нашему изложению, этот вопрос был рассмотрен в [30] (Rose, 1990) (без определения оптимальной конфигурации). Там же приводится итерационный алгоритм кластеризации, по сути, идентичный EM-алгоритму.

Заключение

В заключение резюмируем в чем суть байесовского подхода, чем он отличается от традиционного и что дает практикующим специалистам в области машинного обучения и data mining.

Байесовская статистика исходит из решения задачи обучения в наиболее общем виде. Теоретически, традиционная статистика является частным случаем байесовской, когда регуляризация обучения ограничивается выбором функционального вида модели. Все гипотезы в рамках данной параметризации считаются априори равновероятными. Байесова статистика допускает более широкий класс гипотез с произвольными априорными ограничениями.

Методологически, традиционная статистика ставит своей целью нахождение одной наилучшей гипотезы, тогда как в байесовской статистике обучение приводит лишь к сужению множества допустимых гипотез от априорного к апостериорному. Разброс предсказаний байесовской модели диктуется существующим разбросом в пространстве гипотез. В традиционной же статистике, разброс предсказаний единственной гипотезы определяется по набору искусственно сгенерированных валидационных выборок.

Наконец, с практической точки зрения байесова регуляризация конструктивна, благодаря тому, что для многих классов задач существуют априорные плотности, допускающие аналитическое интегрирование (суммирование) по гипотезам. В идеальной ситуации для Likelihood находят подходящий *сопряженный* Prior, такой, что Posterior имеет тот же функциональный вид, что и Prior, только с другими параметрами (как в примере с бросанием монеты):

$$P\left(h\,|D\right) = \frac{P\left(D\,|h\right)P_0\left(h\,|D_0\right)}{P\left(D\,|D_0\right)} = P_0\left(h\,|D_0 + D\right)\ .$$

В этом случае Evidence также удается выразить в замкнутом виде. В других случаях ситуацию удается свести к идеальной, используя асимптотические разложения для большого числа примеров.

Сравнение Evidence для разных моделей (способов регуляризации обучения) позволяет обоснованно выбирать наилучшую из них. Выбор параметров регуляризации можно совместить с обучением модели в едином итерационном алгоритме, извлекающем информацию из данных более последовательно и экономно, чем процедура кросс-валидации.

Подробности

Бросание монеты (к разделу «Обучение по Байесу»)

Рассмотрим в качестве примера задачу о бросании несимметричной монеты, к которой Байес впервые и применил свой метод. Пусть вероятность выпадения решки в отдельном испытании равна h. Распределение Бернулли дает решение прямой задачи — вероятность выпадения N_h решек в N испытаниях:

$$P\left(D\,|h\right) = \left(\begin{array}{c} N \\ N_h \end{array} \right) h^{N_h} \left(1-h\right)^{N-N_h} \equiv P\left(N_h\,|N,h\right) \ .$$

Допустим, вслед за Байесом, что априори все значения кривизны в допустимом интервале $h \in [0,1]$ равновероятны. Тогда мы получим следующее апостериорное распределение вероятностей для «кривизны» монеты h:

$$\begin{aligned} P\left(h\,|D\right) &= \frac{P\left(N_h\,|N,h\right)}{\int_0^1 dh\, P\left(N_h\,|N,h\right)} = \\ &= \frac{(N+1)!}{N_h!\,(N-N_h)!} h^{N_h} \left(1-h\right)^{N-N_h} \equiv P\left(h\,|N_h,N\right) \end{aligned}$$

с ожидаемым значением кривизны:

$$\langle h \rangle = \int dh\, h\, P\left(h\,|D\right) = \frac{1+N_h}{2+N} \ ,$$

не равным нулю, даже если $N_h = 0$. Точность определения кривизны растет с числом испытаний как $O\left(1\Big/\sqrt{N}\right)$:

$$\langle h \rangle \underset{N \gg 1}{\to} \frac{N_h}{N} ,$$

$$\Delta h^2 = \left\langle h^2 \right\rangle - \langle h \rangle^2 \underset{N \gg 1}{\to} \frac{1}{N} \frac{N_h}{N} \left(1 - \frac{N_h}{N} \right) .$$

Если мы теперь учтем приобретенный опыт, т. е. примем за априорную функцию $P(h\,|D)$ и проведем дополнительно N' испытаний, в которых выпадет N'_h решек, то получим новое, уточненное апостериорное распределение в точности такого же вида:

$$P(h\,|D', D) = \frac{P(D'\,|h)\,P(h\,|D)}{P(D'\,|D)} =$$

$$= \frac{P(N'_h\,|N', h)\,P(h\,|N_h, N)}{\int_0^1 dh\,P(N'_h\,|N', h)\,P(h\,|N_h, N)} = P(h\,|N_h + N'_h, N + N') .$$

Таким образом, следующие друг за другом серии испытаний можно считать единым экспериментом, в котором постепенно происходит накопление наших знаний о свойствах монеты — концентрация плотности распределения гипотез вокруг истинного значения кривизны.

Принцип минимальной длины описания (к разделу «Обучение по Байесу»)

Согласно теории кодирования Шеннона, при известном распределении $P(X)$ случайной величины X длина оптимального кода для передачи конкретного значения x по каналу связи стремится к $L(x) = -\log P(x)$. *Энтропия источника* $S(P) = -\sum_x P(x) \log P(x)$ является минимальной ожидаемой длиной закодированного сообщения. Любой другой код, основанный на неправильном представлении об источнике сообщений приведет к большей ожидаемой длине сообщения. Иными словами, чем лучше наша модель источника, тем компактнее могут быть закодированы данные.

В задаче обучения источником данных является некая неизвестная нам истинная функция распределения $P(D|h_0)$. Отличие между ней и модельным распределением $P(D|h)$ по мере Кулбака–Леблера:

$$|P(D|h) - P(D|h_0)| = \sum_D P(D|h_0) \log \frac{P(D|h_0)}{P(D|h)} =$$
$$= \sum_D P(D|h_0) [L(D|h) - L(D|h_0)] \geqslant 0$$

представляет собой разницу ожидаемой длины кодирования данных с помощью гипотезы и минимально возможной. Эта разница всегда неотрицательна и равна нулю лишь при полном совпадении двух распределений. Иными словами, гипотеза тем лучше, чем короче средняя длина кодирования данных.

Теория Шеннона предполагает, что код h известен как отправителю, так и получателю сообщений. В теории обучения известным предполагается лишь некоторое априорное распределение вероятностей $P(h)$ [16]. Соответственно, закодированное сообщение в этом случае должно иметь две составляющие: описание способа раскодирования h длиной $L(h) = -\log P(h)$ и закодированные этим способом данные длиной $L(D|h) = -\log P(D|h)$. В соответствии с принципом *минимальной длины описания* (*Minimum Description Length*), оптимальная гипотеза минимизирует именно эту суммарную *длину описания данных* $L(D,h) = L(D|h) + L(h) = -\log P(D,h)$:

$$h_{MDL} = \arg\min_h L(D,h) \ .$$

Именно суммарная длина описания дает правильную оценку ожидаемого риска

$$\langle L(D|h) \rangle_D \equiv \sum_D P(D|h_0) L(D|h) \ ,$$

соответствующего *ошибке обобщения* на новых данных, а вовсе не *эмпирический риск* $L(D|h)$, соответствующий *ошибке обучения*. Последний может быть сведен к нулю в достаточно сложной модели, т. е. не дает

[16] В этом смысле, теорию обучения можно считать обобщением теории оптимального кодирования.

представления о реальном риске предсказаний. Тогда как для суммарной длины описания можно доказать (см. [Vapnik 1995]), что с вероятностью не меньшей $1-\eta$:

$$\langle L\left(D\,|h\right)\rangle_D \leqslant 2\left(L\left(D,h\right)\ln 2 - \ln\eta\right)\ .$$

Поскольку все риски пропорциональны числу данных, второй член в правой части пренебрежимо мал по сравнению с остальными при большом N. Таким образом, для удельных ошибок в расчете на один пример $l\left(\cdot\right) \equiv N^{-1}L\left(\cdot\right)$ получаем:

$$\langle l\left(D\,|h\right)\rangle_D \leqslant 2\ln 2 \cdot l\left(D,h\right)\ .$$

Иными словами, именно совместная длина описания данных и гипотезы ограничивает ошибку обобщения. Заметим, что этот результат не зависит ни от числа примеров, ни от конкретного вида функций, среди которых ищется решение, ни от способа регуляризации, ни, наконец, от того, какова ошибка обучения.

Следовательно, минимизация совместной длины описания является очень общим, теоретически обоснованным принципом, который можно положить в основу процесса обучения.

Проверка априорных гипотез (к разделу «Оценка параметров по Байесу»)

Статсуммы, через которые выражается Evidence, в случае гауссовых шума и Prior, имеют вид:

$$Z_\alpha = \left(\frac{\alpha}{2\pi}\right)^{-d/2},$$

$$Z_\beta = \left(\frac{\beta}{2\pi}\right)^{-Nd/2},$$

$$Z_{\alpha,\beta} = \exp\left(-\frac{\beta N d}{2}\sigma_y^2 - \frac{\alpha\,\beta N}{2\left(\beta N+\alpha\right)}\sum_i \langle y_i\rangle^2\right)\left(\frac{\left(\beta N+\alpha\right)}{2\pi}\right)^{-d/2}.$$

Логарифм Evidence в этом случае равен:

$$\ln P\left(D\,|\beta,\alpha\right) =$$
$$= \ln \int d\mathbf{h}\; P\left(D\,|\mathbf{h},\beta\right) P\left(\mathbf{h}\,|\alpha\right) = \ln Z_{\alpha,\beta} - \ln Z_{\alpha} - \ln Z_{\beta} =$$
$$= \left(-\frac{\beta N d}{2}\sigma_y^2 - \frac{\alpha\,\beta N}{2\left(\beta N+\alpha\right)}\langle \mathbf{y}\rangle^2\right) - \frac{d}{2}\ln\left(\frac{\left(\beta N+\alpha\right)}{2\pi}\right) +$$
$$+\frac{Nd}{2}\ln\left(\frac{\beta}{2\pi}\right) + \frac{d}{2}\ln\left(\frac{\alpha}{2\pi}\right)$$

и достигает максимума при следующем значении α:

$$0 = \frac{\partial}{\partial\alpha}\left[-\frac{\beta N}{2}\langle \mathbf{y}\rangle^2\frac{\alpha}{\left(\beta N+\alpha\right)} + \frac{d}{2}\ln\left(\frac{\alpha}{\left(\beta N+\alpha\right)}\right)\right] =$$
$$= \left[-\frac{\beta N}{2}\langle \mathbf{y}\rangle^2 + \frac{d}{2}\left(\frac{\left(\beta N+\alpha\right)}{\alpha}\right)\right]\left[\frac{\partial}{\partial\alpha}\left(\frac{\alpha}{\left(\beta N+\alpha\right)}\right)\right] .$$

Если нуль определяется первым сомножителем, то оптимальное значение α равно:

$$\alpha_{ML} = \frac{\beta N}{\frac{\beta N}{d}\langle \mathbf{y}\rangle^2 - 1} , \quad \frac{\beta N}{d}\langle \mathbf{y}\rangle^2 > 1 .$$

В противном случае первый сомножитель всегда положителен, и нуль производной достигается за счет второго сомножителя, равного нулю при бесконечном α:

$$\alpha_{ML} = \infty, \quad \frac{\beta N}{d}\langle \mathbf{y}\rangle^2 \leqslant 1 .$$

Аналогичным образом находим оптимальную оценку β:

$$\beta_{ML}^{-1} = \frac{N}{N-1}\sigma_y^2 .$$

Bayesian Information Criterion (к разделу «Байесова интерполяция функций»)

В пределе большого числа примеров априорная вероятность гипотез гораздо более гладкая функция от h, чем Likelihood. Вычисляя длину описания модели

$$L\left(D\,|H\right) = -\ln\int dh P\left(h\right)\,\exp\left(-L\left(D\,|h,H\right)\right) ,$$

возьмем интеграл в пространстве гипотез методом перевала. В итоге получим приближенное выражение:

$$L\left(D\,|H\right) \simeq L\left(D\,|h_{ML}, H\right) + \frac{1}{2}\ln|\mathbf{H}|\,, \quad \mathbf{H} \equiv \nabla\nabla L\left(D\,|h, H\right)|_{ML}\ .$$

Каждый член *гессиана* $\mathbf{H}$ пропорционален N, поскольку ошибка аддитивна, а ранг этой матрицы равен эффективному числу свободных параметров гипотез M. Следовательно, главный член в логарифме детерминанта гессиана равен $\ln|\mathbf{H}| \simeq M\ln N$.

Отсюда получаем так называемый *байесовский информационный критерий* (Bayesian Information Criterion — BIC):

$$L\left(D\,|H\right) \simeq L\left(D\,|h_{ML}, H\right) + \frac{M}{2}\ln N\ .$$

Его можно трактовать следующим образом. Интеграл Evidence равен произведению своего максимума на соответствующий объем в пространстве гипотез:

$$P\left(D\,|H\right) = \int dh\, P\left(D\,|h, H\right) P\left(h\,|H\right) \simeq P\left(D\,|h_{ML}, H\right) \frac{\Delta h_{posterior}}{\Delta h_{prior}}\ .$$

Коэффициент сжатия фазового объема за счет содержащейся в данных информации называют иногда *фактором Оккама*. Именное его логарифм фигурирует в BIC. Поскольку характерная точность определения параметров модели $\Delta h_m \propto 1\Big/\sqrt{N}$, а пространство гипотез M-мерно, то логарифм фактора Оккама масштабируется согласно BIC:

$$\ln\frac{\Delta h_{prior}}{\Delta h_{posterior}} \sim \frac{M}{2}\ln N\ .$$

Согласно Риссанену, асимптотически это минимальное количество информации, необходимое для выбора наилучшей гипотезы с наилучшей точностью. В байесовской интерпретации — это количество информации, сужающей ансамбль гипотез в модели оптимальной сложности. Фактически же, речь идет об одном и том же.

Заметим, что BIC не противоречит тому, что длина описания данных ансамблем меньше, чем совместная длина описания данных отдельной гипотезой, которая и определяет обобщающую способность:

$$L\left(D\,|h_{ML}, H\right) < L\left(D\,|H\right) < L\left(D, h_{ML}\,|H\right)\ .$$

Действительно, последнее неравенство можно переписать в виде:

$$L\left(D\left|h_{ML},H\right.\right)<L\left(D\left|h_{ML},H\right.\right)+\ln\frac{\Delta h_{prior}}{\Delta h_{posterior}}<$$
$$<L\left(D\left|h_{ML},H\right.\right)+\ln\Delta h_{prior}\ .$$

Соответственно, разница ожидаемой ошибки предсказаний наиболее правдоподобной гипотезы и оптимального ансамбля равна $\ln\Delta h_{posterior}$.

Оптимизация кластерной модели (к разделу «Байесова кластеризация»)

При вычислении детерминанта $|\mathbf{A}|$ следует принять во внимание, что в точке экстремума все ненулевые вторые производные находятся на главной диагонали матрицы $\mathbf{A}$:

$$\frac{\partial^2}{\partial w_{i,m}\partial w_{i',m'}}\ln P\left(D\left|h,\beta\right.\right)\Big|_{h_{ML}}=$$
$$=\frac{\partial^2}{\partial w_{i,m}\partial w_{i',m'}}\sum_{m,n}\mathcal{P}\left(m\left|n\right.\right)\ln P\left(\mathbf{x}^{(n)}\left|m\right.\right)=$$
$$=-\frac{\beta}{2}\sum_{n}\mathcal{P}\left(m\left|n\right.\right)\frac{\partial^2\left(\mathbf{w}_m-\mathbf{x}^{(n)}\right)^2}{\partial w_{i,m}\partial w_{i',m'}}=$$
$$=-\beta\sum_{n}\mathcal{P}\left(m\left|n\right.\right)\delta_{m,m'}\delta_{i,i'}=-\beta N\,P\left(m\right)\delta_{m,m'}\delta_{i,i'}$$
$$\frac{\partial^2}{\partial P_m\partial P_{m'}}\ln P\left(D\left|h,\beta\right.\right)\Big|_{h_{ML}}=-\sum_{n}\frac{\mathcal{P}\left(m\left|n\right.\right)}{P^2\left(m\right)}\delta_{m,m'}=-\frac{N\delta_{m,m'}}{P\left(m\right)}\ .$$

Таким образом, детерминант матрицы $\mathbf{A}$ равен произведению всех ее диагональных членов:

$$|\mathbf{A}|=\left(\beta N\right)^{Md}N^M\left(\prod_m P\left(m\right)\right)^{d-1}\ .$$

Что касается значения Evidence в оптимальной модели, то оно определяется найденным выше детерминантом и значением Likelihood, для которого имеем, с учетом найденного значения β_{ML}:

$$\ln P\left(D\,|h_{ML},\beta_{ML}\right) =$$
$$= -\frac{(N-M)\,d}{2} + \frac{Nd}{2}\ln\frac{\beta}{2\pi} - \sum_{m,n}\mathcal{P}\left(m\,|n\right)\ln\frac{\mathcal{P}\left(m\,|n\right)}{P\left(m\right)}\ ,$$

$$\ln P\left(D\,|\beta_{ML}\right) \simeq \ln P\left(D\,|h_{ML},\beta_{ML}\right) - \frac{1}{2}\ln|\mathbf{A}| + \frac{1}{2}\left(Md+M\right)\ln\left(2\pi\right)\ .$$

Литература

1. *Bayes T.* An essay towards solving a problem in the doctrine of chances // Philos. Trans., London. – 1764. – v.53, pp. 376–398. Ibid., 1958. – v. 54. – pp. 298–310. Reprint: Biometrica, v. 45. – pp.293–315,
2. *Bezdek J.* Pattern recognition with fuzzy objective function algorithms. – Plenum, 1981.
3. *Bishop C. M.* Neural networks for pattern recognition. – Oxford: Clarendon Press, 1995.
4. *Chaitan G.* On the length of programs for computing finite binary sequences // J. Assoc. Comput. Mach. – 1966. – v. 13. – pp. 547–569.
5. *Dempster A., Laird N., and Rubin D.* Maximum likelihood from incomplete data via the EM algorithm // Journal of the Royal Statistical Society. B. – 1977. – v. 39. – pp. 1–38.
6. *Finetti B.* Bayessianism // Intern. Statist. Rev. – 1974. – v. 42. – pp. 117–130.
7. *Fisher R.* On an absolute criterion for fitting frequency curves // Messenger of Mathematics. – 1912. – v. 41. – pp. 155–160.
8. *Gull S.* Bayesian inductive inference and maximum entropy // In: Maximum entropy and Bayesian methods in science and engineering, vol. 1: Foundations / Eds.: *Erickson G.* and *Smith C.*. – Kluwer, 1988.
9. *Hassibi B., Stork D.* Second order derivates for network pruning: optimal brain surgeon // In: *Hanson S., Cowan J.*, and *Giles C.* (Eds.) Advances in Neural Information Processing Systems, Volume 5. – Morgan Kaufmann. – 1993. – pp. 164–171.

10. *Hanson R., Stutz J., Cheeseman P.* Bayesian classification theory. – NASA Ames TR FIA-90-12-7-01. – 1991.

11. *Hinton G.* Learning translation invariant recognition in massively parallel networks // In: *de Bakker J., Nijman A.*, and *Treleaven P.* (Eds.) Proceedings PARLE Conference on Parallel Architectures and Languages Europe. – Springer-Verlag, 1987. – pp. 1–13.

12. *Janes E.* Bayesian methods: general background // In: Maximum Entropy and Bayesian Methods in Applied Statistics / Ed.: *J. Justice*. – Cambridge University Press, 1986.

13. *Jeffreys H.* Theory of probability. – Oxford Univ. Press, 1939.

14. *Jain A., Dubes R.* Algorithms for Clustering Data. – Prentice Hall, 1988.

15. *Kagan A.M., Linnik Yu. V., Rao C.R.* On a characterization on the normal low based on a property of the sample average // Sankhya, Ser. A. – 1965. – v. 27, No. 3–4. – pp. 405–406.

16. *Kashyap R.* A Bayesian comparison of different classes of dynamic models using empirical data // IEEE Trans. Automatic Control. – 1977. – AC-22, No.5. – pp. 715–727.

17. *Kolmogoroff A.* Three approaches to the quantitative definitions of information // Problems of Inform. Transmission. – 1965. – v. 1, No. 1. – pp.1–7.

18. *Lang K., Hinton G.* Dimensionality reduction and prior knowledge in *E*-set recognition // In: *Touretzky D.* (Ed.) Advances in Neural Information Processing Systems, Volume 2. – Morgan Kaufmann. – 1990. – pp. 178–185.

19. *Laplace P.* A philosophical essay on probabilities. – Dover, 1819.

20. *Le Cun Y., Denker J., Solla S.* Optimal brain damage // In: *Touretzky D.* (Ed.) Advances in Neural Information Processing Systems, Volume 2. – Morgan Kaufmann. – 1990. – pp. 598–605.

21. *Loredo T.* From Laplace to supernova SN 1987A: Bayesian inference in astrophysics // In: Maximum Entropy and Bayesian Methods, Ed.: *P. Fougere*. – Kluwer, 1989.

22. *MacKay D.* Bayesian interpolation // Neural Computation. – 1992. – v. 4. – pp. 415–447. A practical Bayesian framework for backprop networks, Ibid., pp. 448–472.

23. *McLachlan G., Basford K.* Mixture models: Inference and applications to clustering. – Marcel Dekker, 1988.

24. *Mises R. von.* Probability, statistics and truth. – MacMillan, 1939.

25. *Neal R., Hinton G.* A view of the EM algorithm that justifies incremental, sparse, and other variants // In *M. I. Jordan* (Ed). Learning in Graphical Models. – Cambridge, MA: MIT Press, 1999. – pp. 355-368.

26. *Parzen E.* On estimation of probability function and mode // Annals of Math. Statistics. – 1962. – v. 33, No. 3.

27. *Patrick J., Wallace C.* Stone circle geometries: An information theory approach // In: Archeoastronomy in the Old World, Ed.: *D. Heggie*. – Cambridge University Press, 1982.

28. *Phillips D.* A technique for numerical solution of certain integral equation of the first kind // J. Assoc. Comput. Math. – 1962. – v. 9. – pp. 84–96.

29. *Rissanen J.* Modeling by shortest data description // Automatica. – 1978. – v. 14. – pp. 465–471.

30. *Rose K., Gurevitz E., Fox C.* Statistical Mechanics and Phase Transitions in Clustering // Phys. Rev. Lett. – 1990. – v. 65, No. 8. – pp. 945–948.

31. *Rosenblatt M.* Remarks on some nonparametric estimation of density function // Annals of Math. Statistics. – 1956. – v. 27. – pp. 642–669.

32. *Skilling J.* On parameter estimation and quantified MaxEnt // In: Maximum Entropy and Bayesian Methods / Eds. *Grandy W.* and *Schick L.* – Kluwer, 1991.

33. *Solomonoff R.* A preliminary report on general theory of inductive inference. – Tech. Report ZTB-138, Zator Company, Cambridge, MA. – 1960.

34. *Stein C.* Inadmissibility of the usual estimator for the mean of multivariable normal distribution // Proc. 3rd Berkeley Symp. On Math. Statist. and Probab. – Univ. of California Press, Berkeley. – v. 1. – 1956. – pp. 197–206.

35. *Vapnik V.* The Nature of statistical learning theory. – Springer, 1995.

36. *Rumelhart D, Hinton G., Williams R.* Learning internal representation by error propagation // In: Rumelhart D, McClelland, and PDP Research Group (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations. – MIT Press, 1986. – pp. 318–362.

37. *Wallace C., Boulton D.* An information measure for classification // Computing Journal. – 1968. – v. 11. – pp. 185–195.

38. *Weigend A., Neil A.*, Eds. Time series prediction: Forecasting the future and understanding the past. – Addison-Wesley, 1994.

39. *Williams P.* Bayesian regularization and pruning using a Laplace prior // Neural Computation. – 1995. – v. 7, No. 1. – pp. 117–143.

40. *Zellner A.* Basic issues in econometrics. – Chicago, 1984.

41. *Вапник В.Н., Червонинкис А.Я.* О равномерной сходимости частот к их вероятностям // ДАН. – 1968. – т. 181, №4.

42. *Вапник В.Н., Червонинкис А.Я.* Теория распознавания образов. – М: Наука, 1974.

43. *Иванов В.К.* О линейных некорректных задачах // ДАН. – 1962. – v. 145, №2.

44. *Кокс Д., Хинкли Д.* Теоретическая статистика. – М.: Мир, 1978.

45. *Кокс Д., Снелл Э.* Прикладная статистика. Принципы и примеры. – М.: Мир, 1984.

46. *Секей Г.* Парадоксы в теории вероятностей и математической статистике. – М.: Мир, 1990.

47. *Тихонов А.Н.* О регуляризации некорректно поставленных задач //ДАН. – 1963. – т. 153, №1. с. 49–53.

48. *Ченцов Н.Н.* Оценка неизвестной плотности вероятности из наблюдений // ДАН. – 1962. – т. 147. – с. 45–48.

Сергей Александрович Шумский, кандидат физико-математических наук, старший научный сотрудник ФИАН им. П. Н. Лебедева РАН, вице-президент ООО «НейрОК». Научные интересы — физика плазмы и термоядерного синтеза, статистическая механика распределенных вычислений, теория и приложения нейрокомпьютинга.

С. А. ТЕРЕХОВ
Снежинский физико-технический институт, г. Снежинск;
ООО НейрОК, г. Москва,
E-mail: alife@narod.ru

НЕЙРОСЕТЕВЫЕ АППРОКСИМАЦИИ ПЛОТНОСТИ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТИ В ЗАДАЧАХ ИНФОРМАЦИОННОГО МОДЕЛИРОВАНИЯ

Аннотация

С прикладных позиций рассматривается задача аппроксимации плотности распределения, описывающего множество многомерных экспериментальных данных. Предложены эффективные нейросетевые методики аппроксимации формы плотности. Приведены примеры постановок задач анализа данных на основе аппроксимации плотности. Обсуждаются приложения подхода.

S. A. TEREKHOFF
Snezhinsk Institute of Physics and Technology (SFTI), Snezhinsk;
NeurOK LLC, Moscow,
E-mail: alife@narod.ru

NEURAL APPROXIMATIONS OF PROBABILITY DENSITY IN INFORMATIONAL MODELLING

Abstract

The problem of probability density approximation based on set of multivariate experimental data is considered from the practical informatics point of view. Effective neural approximation techniques for the density are proposed. Statements for several data analysis problems are presented using density approximation approach. Applications of the method are discussed.

Плотность распределения вероятности и ее роль в информационном моделировании

Рассмотрим проблему построения эмпирических моделей на основе числовых данных. Речь пойдет об обучении без учителя на примерах в условиях неопределенности о характере модели. Пусть данные составляют совокупность имеющихся результатов экспериментов или наблюдений над некоторой сложной [1] системой или устройством. В центре рассмотрения лежит матрица наблюдений D_{jk}, в которой $j = 1, \ldots, N$ — номер наблюдения (одна строка в таблице или запись в базе данных), а $k = 1, \ldots, M$ — номер наблюдаемой переменной x_k (признака, фактора, свойства и т. д.). Матричным элементом является действительное число — результат наблюдения. Здесь мы ограничимся случаем непрерывных переменных и не будем пока касаться дискретных (ординальных и категориальных) наблюдений.

На этом этапе относительно механизма порождения данных будем предполагать следующее:

- наблюдаемое значение является реализацией некоторой случайной величины;
- наблюдаемые данные порождены стационарным процессом (системой), т. е. рассматриваемые случайные величины не зависят от времени;
- различные наблюдения не зависят друг от друга;
- факт наблюдения не влияет на свойства исследуемой системы (процесса).

Таким образом, считаем, что наблюдаемые данные порождены некоторой вероятностной [1] средой. При этом можно выделить несколько

[1] Подчеркнем разницу между экспериментом и наблюдением. При проведении *эксперимента* условия внешнего воздействия и параметры исследуемой системы управляются экспериментатором. При простом *наблюдении* система изучается при тех параметрах и условиях, в которых она (случайно) оказалась в момент наблюдения. Различие весьма существенно при изучении, например, социальных, рыночных систем или живых организмов. В подобных случаях эксперимент, как правило, невозможен.

основных причин, по которым предлагается согласиться с вероятностной [2] трактовкой данных:

- процесс измерения сопряжен с экспериментальными погрешностями;
- изучаемая система является сложной [2], т. е. несводимой к сумме свойств отдельных компонент, а наблюдаемое многообразие данных может быть равновероятно объяснено великим множеством структурных описаний, при этом нельзя достоверно предпочесть ни одно из них;
- объем измерений конечен и не может считаться исчерпывающим описанием системы.

В этих условиях мы будем говорить о статистической модели, описывающей совокупность данных, как об *информационной* модели изучаемой сложной системы.

Наиболее полным статистическим описанием наблюдаемых данных является *совместная плотность распределения вероятности* точек в векторном пространстве признаков $P(x_1, x_2, \ldots, x_M)$. Рассмотрим особенности задачи ее аппроксимации. В отличие от традиционных постановок задач сглаживания данных [3], когда в распоряжении у исследователя имеются пары значений «аргумент–функция», при аппроксимации плотности даны только координаты точек в многомерном пространстве. Поэтому будем считать, что аппроксиматором плотности P множества точек D в параллелепипеде V является всякая функция A, такая, что:

- A равна нулю вне V;
- A нормирована на V : $\int A(X)dV = 1$;
- Отношение интегралов от A по двум объемам V_1 и V_2 из V, содержащим точки из D, «стремится» к отношению числа точек из D в этих объемах.

[2] Содержательными являются также и другие (не вероятностные) трактовки. В частности, система может рассматриваться, как описываемая некоторым числом детерминированных скрытых факторов, а стохастичность наблюдаемых переменных (которых обычно больше, чем факторов) является внешней по отношению к самой системе. Другой подход — нелинейные динамические системы, порождающие наблюдаемый динамический хаос.

Намеренно ограничившись столь нестрогим определением, подчеркнем важные с практической точки зрения свойства задачи:

- всякая попытка восстановления плотности вдали за границами объема V, содержащего точки наблюдений, потребует дополнительных предположений и ограничений;
- внутри исследуемого объема задача восстановления плотности также является некорректно поставленной [4], хотя бы уже потому, что решение не единственно;
- у задачи, в некотором смысле, нет «наилучшего» решения, имея в виду использование оцененной плотности для генерации и объяснения новых данных.

Заметим, что некоторые простые, кажущиеся естественными, попытки построить функционалы, оптимизация которых ведет к устранению некорректности задачи, часто приводят к решениям с весьма скромной практической пользой. Так например, если в качестве такого функционала выбрать популярный принцип максимального правдоподобия, т. е.

$$\max \sum_{j \in D} \log(A(D_j))$$

и не предпринять никаких дополнительных мер по регуляризации, то легко получим, что функция

$$A(x) = \frac{1}{\|D\|} \sum_{j \in D} \delta(x - D_j)$$

является неограниченно «правдоподобной», удовлетворяя наилучшим образом всем условиям, наложенным ранее на аппроксиматор. К сожалению, универсальных автоматических рецептов для регуляризации не существует.

В следующих разделах мы вернемся к практическим методам регуляризации задачи аппроксимации, а сейчас зададимся вопросом, *что дает исследователю знание совместной плотности распределения признаков изучаемого объекта*?

В некотором смысле, описание на языке плотности распределения соответствует описанию на языке волновой функции в квантовой механике.

Полнота статистического описания на основе совместной плотности вероятности подразумевает, что в плотности P содержится практически вся[3] информация об исследуемой системе, которую можно почерпнуть из имеющихся данных. Это, в свою очередь, означает, что все наблюдаемые следствия о значениях и распределениях различных величин при различных условиях могут быть получены путем вычислений функционалов от P. Искомые функционалы в общем случае могут быть оценены методом Монте–Карло с различными вариациями [5].

Рассмотрим семейство важных *функциональных запросов*, представляющих прикладной интерес.

1. **Однофакторные условные распределения.** Как распределен какой-то их признаков, если значения остальных признаков известны достоверно? В случае, если выделенный признак является выходным (зависимым от остальных) и кодирующим некоторый отклик, эта задача соответствует *задаче распознавания образов*. В качестве искомого кода образа, задаваемого остальными признаками, принимается наиболее вероятное или среднее значение в распределении признака-кода. Такая плотность вероятности называется условной:

$$P(x_k|x_1^*,\ldots,x_{k-1}^*,x_{k+1}^*,\ldots,x_M^*) =$$

$$= \frac{P(x_1^*,\ldots,x_{k-1}^*,x_k,x_{k+1}^*,\ldots,x_M^*)}{\int dx_k P(x_1^*,\ldots,x_{k-1}^*,x_k,x_{k+1}^*,\ldots,x_M^*)}$$

Таким образом, имеет место пропорциональность:

$$p(x_k) \sim P(x_1^* \ldots x_k \ldots x_M^*)$$

В общем случае, в отличие от традиционной задачи аппроксимации поверхности отклика или распознавания образов, при использовании для оценки условной плотности (как функции одной переменной x_k) некоторой аппроксимации совместной плотности P получается *истинное распределение возможных значений результата*, а не только его наиболее вероятное значение.

[3] В действительности, всегда происходит некоторая потеря информации вследствие ошибок аппроксимации. Заметим, также, что ограничение информации при аппроксимации плотности объемом информации, содержащимся в исходных данных — является, с точки зрения автора, ответом на дискуссию с А. А. Ежовым [6].

Аналогично рассматривается и задача распределения пары признаков или любого другого их числа:

$$P(x_1, \ldots, x_q | x^*_{q+1}, \ldots, x^*_M) \sim P(x_1, \ldots, x_q, x^*_{q+1}, \ldots, x^*_M).$$

2. **Пропуски в данных.** Какие значения может принимать некоторый признак, если значения части других признаков известны достоверно, а оставшихся - неизвестны вовсе? К этой задаче сводится известная *проблема заполнения пропусков в таблицах данных*, в контексте нейронных сетей подробно рассмотренная в работе [7]. Соответствующие распределения даются маргинальными интегралами от совместной плотности:

$$P(x_k | x^*_q \ldots x^*_M) \sim \int dx_1, \ldots, dx_h P(x_1..x_h, \ldots, x_k, \ldots, x^*_q \ldots x^*_M).$$

3. **Вероятностный прогноз.** Какова форма плотности распределения условных вероятностей выделенных признаков, если известны однофакторные плотности распределения остальных переменных. Результат:

$$P(x_1 \ldots x_q | p(x_{q+1}) \ldots p(x_M)) \sim$$
$$\sim \int dx_{q+1} p(x_{q+1}) \ldots dx_M p(x_M) P(x_1, \ldots, x_M).$$

4. **Обратные задачи.** Каковы должны быть величины переменных $x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_M$, чтобы наиболее вероятным значением для переменной x_k было число $x^\star_k$? Эта задача сводится к (численному) поиску решений нелинейного уравнения на максимум величины условной вероятности

$$\max_{x_k} P(x_k | x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_M) = x^\star_k$$

как функции $M-1$ переменной. Эффективность решения такой задачи во многом определяется функциональной простотой выбранной аппроксимации плотности.

Большой интерес представляют также различные комбинации перечисленных задач (например, случай, когда часть переменных известны точно, для других известны лишь оценки распределений, а прочие — неизвестны).

Важно отметить, что все выражения, в которых используется совместная плотность, требуют ее определения с точностью до некоторого постоянного множителя — везде встречаются только отношения плотностей. Эта деталь существенно облегчает нашу задачу!

Подходы к аппроксимации плотности распределения

Переходим теперь собственно к вопросу построения аппроксимаций. Основной объем литературы по вопросам аппроксимации плотности вероятности можно условно разбить на такие крупные разделы, как:

- **Параметрические методы аппроксимации**. Подходы этого раздела основаны на предположении о конкретном функциональном виде распределения, который зависит от некоторых параметров. Эти параметры далее выбираются на основе статистических критериев максимального правдоподобия или максимума апостериорной вероятности описания данных моделью. Теория и методики этого раздела подробно обсуждаются в текстах по математической статистике и обработке результатов экспериментов.

- **Методы непараметрической статистики**. Сюда относятся выборочные частотные гистограммы (мало применимые в многомерном случае, см. ниже) и широкий класс методов, основанных на аппроксимации плотности смесью базисных функций [8–10]. Частным случаем такой аппроксимации являются гауссовы смеси [11], радиальные базисные функции, а также вейвлет-методы [12].

Плотность распределения является функцией многих переменных — по числу исследуемых признаков. В условиях, когда имеется дополнительная информация о степени зависимости или независимости признаков, эта плотность может быть факторизована на функции меньшего числа переменных. При дополнительных сведениях об обусловливании одних признаков другими удобнее говорить об условных [4] плотностях распределений, которые также зависят от меньшего числа переменных (т. е.,

[4] Существует точка зрения (см. например, [13]), что вероятность, как мера наших ожиданий (belief), всегда является условной, т. е. обусловленной текущим объемом информации, имеющимся у исследователя.

только от обусловливающих признаков). Такое факторизованное описание дается *байесовыми сетями событий* [5].

Байесовы сети являются, по-видимому, самой революционной технологией последнего десятилетия [5] в области искусственного интеллекта. Основное их достоинство — универсальное и интуитивно понятное представление моделей, основанных на данных. В противоположность нейронным сетям, графические байесовы модели допускают прямую интерпретацию. Байесова сеть состоит из узлов, соответствующих переменным задачи, и ребер, отражающих зависимости между переменными. Отсутствие ребра между двумя узлами означает независимость между отвечающими им переменными.

Баейсовы сети — столь обширная и крайне интересная научная область, что, не имея возможности более подробно обсуждать здесь этот вопрос, автор предлагает ознакомиться с ним по работе [14] [6].

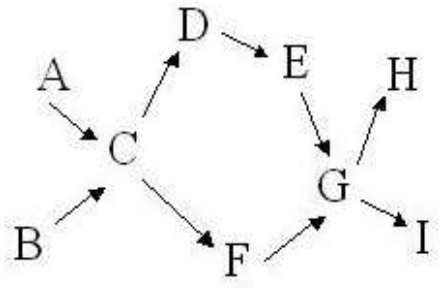

Рис. 1. Байесова сеть для 9 переменных

[5] Билл Гейтс заявил в интервью газете «Лос-Анжелес Таймс» (28 октября 1996 года), что байесовы сети более перспективны и значимы, чем DOS — если сравнивать их по степени влияния на общество (Гейтс в своих выступлениях и публикациях не прочь напомнить миру, что самым что ни на есть революционным событием XX века было изобретение операционной системы, которую смогли использовать даже домохозяйки). — См. URL: http://www.cs.berkeley.edu/~murphyk/Bayes/la.times.html

[6] См. URL: http://bayes.wustl.edu/etj/prob/book.pdf.tar.gz

Важно, что в результате построения байесовой сети происходит редукция плотности. Так, например [15], для байесовой сети, отражающей зависимости между 9 переменными (рис. 1), плотность факторизуется следующим образом:

$$P(A,B,C,D,E,F,G,H,I) = P(A)P(B)P(C|A,B)P(D|C)\times \\ \times P(E|D)P(F|C)P(G|E,F)P(H|G)P(I|G)$$

Каждая из полученных функций зависит от одной или двух переменных. Далее, если ограничиться вместо описания распределений вероятности только их математическими ожиданиями, тогда формально связь между обусловливающими (входными) и обусловливаемыми (выходными) признаками может быть эффективно представлена многослойной нейронной сетью [16], известной своими хорошими аппроксимационными свойствами (см. например, [17]). И, наконец, если в распоряжении исследователя имеются формальные логические соотношения между входными и выходными переменными, то их можно отразить в форме экспертных систем или функциональных уравнений.

В данной работе предлагается подход на основе замены задачи аппроксимации эквивалентной задачей классификации, при этом рассмотрение ведется на исходном уровне совместной многомерной плотности распределения вероятности. *Суть метода* состоит в построении наилучшего решающего правила, позволяющего отличить наблюдаемую совокупность данных от некоторой искусственной выборки данных с известной плотностью распределения. Этот метод является, по-видимому, относительно новым [18] и мало распространен [7]. Рассмотрим его в самой простой форме — на примере.

Пример 1. Аппроксимация плотности на отрезке

Пусть в нашем распоряжении имеется выборка из суммы двух различных гауссовых распределений, априорные вероятности которых равны:

$$P_{exact} = \frac{0.5}{\sqrt{2\pi\cdot 1^2}}\exp\Big(-\frac{|x-1|^2}{2\cdot 1}\Big) + \frac{0.5}{\sqrt{2\pi\cdot 0.5^2}}\exp\Big(-\frac{|x-3|^2}{2\cdot 0.5}\Big)$$

[7] Автору посчастливилось предложить этот метод независимо от других публикаций.

График точного значения плотности представлен сплошной линией на рис. 3. Объем выборки — $2N$ точек. Для аппроксимации такой плотности распределения на отрезке $[-3, 5]$ дополнительно (искусственно) разыграем еще $2N$ равномерно распределенных случайных точек. Их плотность распределения равна константе:

$$P_0 = \frac{1}{x_{right} - x_{left}} = \frac{1}{8}$$

Объединим оба множества в одно — их $4N$ наблюдений. Первым $2N$ точкам сопоставим значение «класс A», равное 1. Равномерно распределенным $2N$ точкам припишем «класс B» со значением 0. Класс B выполняет эффективную функцию шума. Наша задача состоит в построении классификатора, отличающего точки сигнала (класс A) от этого шума. Выход классификатора будет принимать значения в интервале $[0, 1]$. В качестве классификатора предлагается использовать многослойную нейронную сеть с сигмоидальными нейронами [16], обучаемую методом *RProp* (Приложение А).

Итак, на вход нейросетевого классификатора подается координата одной их точек совокупной выборки, и при обучении требуем, чтобы выход нейросети был равен 0 либо 1 в зависимости от класса, к которому принадлежит данная точка. Класс каждой точки достоверно известен (по построению).

Без сомнений, у такого классификатора нет шансов безошибочно обучиться, так как предъявляемые ему точки обоих классов эффективно перемешаны на отрезке. Для каждой точки в отдельности нет способа достоверно сказать, порождена она сигналом или шумом. К чему же будет сходиться обученный классификатор?

Для ответа на этот вопрос выделим окрестность dx некоторой точки x на исследуемом отрезке. В этой окрестности число точек класса A близко к $P(x)dx$, а число голосующих за класс B стремится к $P_0(x)dx$. Ясно, что минимизирующее квадрат уклонения (т. е. оптимальное) значение на выходе классификатора стремится к величине

$$y(x) = \frac{P(x)}{P(x) + P_0(x)}$$

равной отношению числу голосов, поданных за “1” к общему числу го-

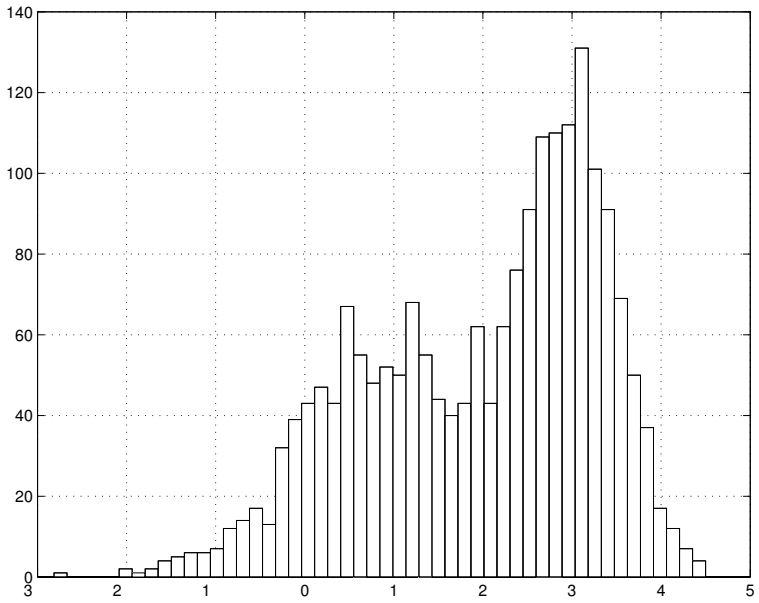

Рис. 2. Гистограмма обучающих данных класса A (сигнал)

лосов. Отсюда для формы плотности легко находим

$$P(x) = P_0 \frac{y(x)}{1 - y(x)}$$

Простой анализ чувствительности дает

$$\delta P \sim P \frac{1}{y(1-y)} \delta y$$

т. е. ошибка аппроксимации сосредоточена в основном в областях насыщения сигмоиды выходного нейрона классификатора.

Результаты аппроксимации этим методом приведены на рис. 3, где также приведен график точной формулы двух гауссианов. Достаточно легко видеть, что такой аппроксимации весьма трудно достичь, сглаживая гистограмму.

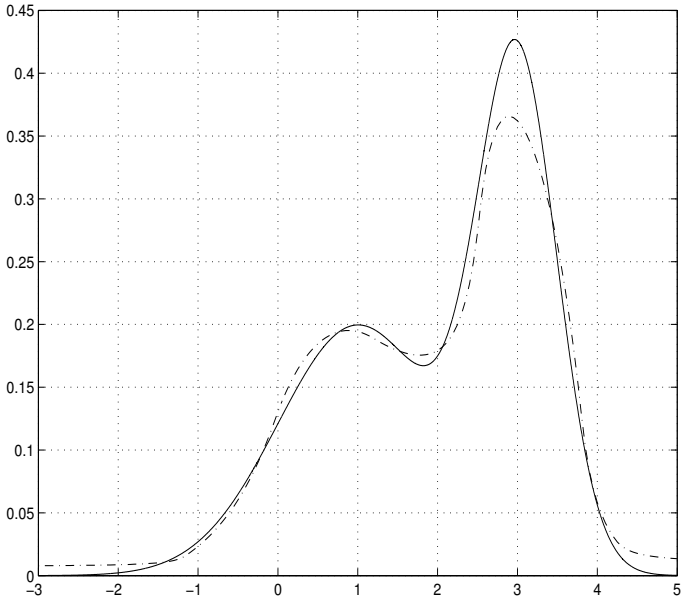

Рис. 3. Результат нейросетевой аппроксимации плотности распределения (пунктир) в сравнении с точной формой плотности (сплошная линия)

В приведенном примере, в действительности, нигде не использовался факт одномерности распределения. В точности такой же метод используется для аппроксимации многомерных данных. Остановимся теперь на особенностях и видимых проблемах предложенного метода аппроксимации плотности.

- **«Проклятие размерности».** Простые оценки показывают, что при росте числа переменных надежды на надежное статистическое описание данных тают на глазах. Так, для построения достоверной гистограммы в пространстве 10 измерений даже с двумя интервалами по каждому из них требуется порядка $30 \times 2^{10} = 30,000$ точек. Такие объемы данных имеются только в приложениях, относящихся к реальному времени (поступление финансовых тиков на

биржах, или онлайн-диагностика). У экспертов, изучающих некоторое устойчивое явление или систему, обычно имеются базы данных лишь с 100–5000 записями. На такой взрывной характер зависимости объема требуемых данных от размерности обратили внимание еще пионеры «раскопки данных» *(data mining)*, такие как *John Tukey* [19]. В нашем подходе эта фундаментальная проблема, разумеется, не устраняется, но частично ослабляется тем фактом, что очень сложные, существенно многомерные, формы поверхности плотности встречаются *редко*, поэтому на практике *почти всегда* нейросетевой классификатор эффективно находит главные особенности и направления вариации функции.

- **Качество обобщения и регуляризация.** Суть проблемы состоит в том, что при улучшении качества аппроксимации обучающих данных возникает переход к их прямому запоминанию, при этом теряются обобщающие свойства модели. К настоящему моменту разработан широкий круг статистических алгоритмов оценки качества обобщения (см. [16], с. 376). Обычно оценка ошибки представляется в общей форме

$$\textit{Ошибка обобщения} = \textit{Ошибка обучения} + \\ + \textit{Штраф за сложность модели}$$

Соотношение между двумя составляющими ошибки наиболее последовательно оценивается при байесовом обучении. Строго говоря, в нашем подходе используются два типа регуляризации — выбор аппроксиматора контролируемой сложности и регуляризация шумом в данных. При использовании нескольких выборок из шумового распределения каждая точка сигнала, образно говоря, окружается облаком шума, что препятствует прямому запоминанию.

- **Эффективные обучающие выборки.** Здесь проблема заключается в том, что число примеров, генерируемых аппроксимируемой плотностью конечно и задано заранее, в то время как имеется полная свобода в генерации данных «шумового» распределения. Какой выбор данных предпочесть? Ответ[8] на этот вопрос рассмотрен в следующем разделе.

[8]По-прежнему, уровень строгости и обоснованности даваемых рекомендаций да-

Бутстреп-выборки

Для построения семейства равномощных обучающих выборок их экспериментальных точек и базового однородного распределения необходимо иметь возможность порождать новые наборы экспериментальных данных. Генерация новых выборок из базового равномерного распределения p_0 не представляет проблем, а что делать с исследуемым экспериментальным распределением? Один из возможных ответов на этот вопрос — метод бутстреп [9] — предложен, по-видимому, Брэдом Эфроном [20] в начале 70-х годов.

Метод, в целом, основан на следующем наблюдении. Для множества точек в многомерном пространстве плотность в форме суммы дельта-функций обладает максимальным правдоподобием (см. начало лекции). Тогда, если для порождения новых выборок пользоваться таким распределением, это будет эквивалентно выборкам из данного множества некоторого нового множества точек *с возвратом*. В выборке будут повторения, но, в некотором смысле, такая выборка статистически распределена так же, как и исходное множество точек. Вводные лекции по бутстреп-методам имеются в Интернет [21], поэтому мы не будем излишне подробно останавливаться на теоретических вопросах. Отметим лишь, что число различных бутстреп-выборок длины N из совокупности N наблюдений равно C_{2N-1}^{N-1}. Это значение легко получить, если рассмотреть задачу размещения $N-1$ перегородок в цепочке из N шариков. Важно, что число вариантов для практических интересных объемов данных заведомо велико (см. табл. 1.).

Имея в своем распоряжении неограниченное число выборок из шумового распределения и соответствующие бутстреп-выборки для полез-

лек от уровня формулировок теорем. Проблема здесь — методологическая. За редким исключением невозможно высказать абсолютно верное (и при этом практически полезное!) утверждение относительно некоторого конкретного набора данных. 10-летний опыт автора показывает, что каждый раз в новой задаче данные устроены совершенно по-разному, и часто исследование приходится начинать практически с нуля. Собственно, это и имел в виду John Tukey, еще в 60-х годах отделяя область анализа данных от матстатистики [19].

[9] Бутстреп — транслитерация английского bootstrap (дословно «тянуть за застежки ботинок»), означающая «использовать существующий вариант системы или процесса для создания нового варианта» (словарь *Lingvo 5.0*). В компьютерной литературе этим термином называются самозагружающиеся программы.

ТАБЛИЦА 1. Зависимость числа различных бутстреп-выборок от размера исходной совокупности данных

N	5	10	20	30
Число выборок	125	93539	6.93×10^{10}	$5.9\times.10^{16}$

ного сигнала, можно применять постраничное обучение. Это обучение снижает риск систематического смещения результата, вызванного фиксированностью данных, так как на разных эпохах обучения нейросети могут использоваться различные обучающие выборки.

Поскольку многомерные задачи аппроксимации могут приводить к необходимости обучения нейросетей с большим числом неизвестных весовых коэффициентов, в этой работе мы отказались от быстросходящихся методов на основе обращения гессиана (или его приближений) и использовали адаптивный метод *RProp* (см. Приложение А).

Численные эксперименты

В этом разделе мы попытаемся проиллюстрировать особенности предложенного метода аппроксимации на реалистичных задачах. Соответствующие базы данных доступны в Интернет [22], поэтому возможна свободная независимая апробация.

Задача Banana

В этой игрушечной двумерной задаче предлагается построить аппроксимацию плотности по сложно распределенным на плоскости выборкам точек из нескольких пятен сложной формы (см. рис. 4). Данные для экспериментирования доступны в Интернет [10].

В исходной постановке точки считаются принадлежащими двум классам (точки и плюсы на рис. 4), требуется решить задачу классификации. В расширенной постановке переменная признака класса, принимающая

[10] URL: http://ida.first.gmd.de/ raetsch/data/banana/banana.data.tar.gz

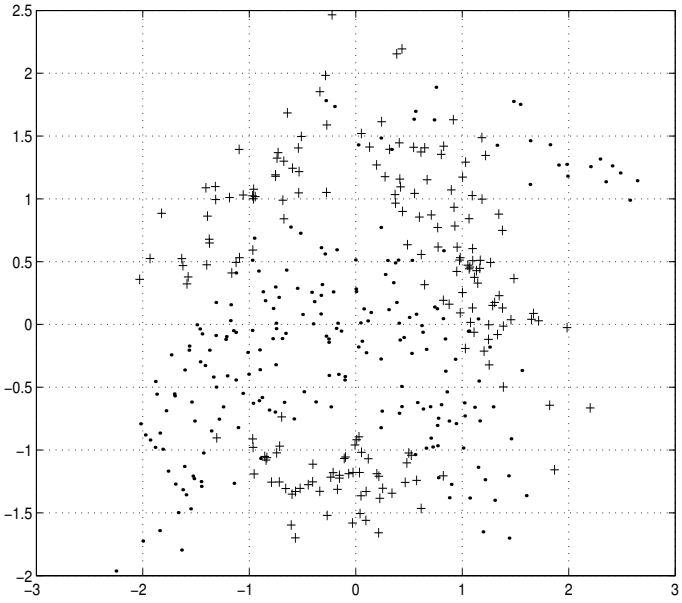

Рис. 4. Исходные данные в задаче Banana

значения 0 или 1, добавляется к исходным двум координатам точек, после чего предлагается построить аппроксимацию плотности в совокупном пространстве трех измерений. После успешного обучения нейросетевого классификатора, с полученным распределением плотности можно решить несколько различных задач. При этом мы будем придерживаться последовательного байесового подхода к вычислениям, а именно:

- Построенная аппроксимация плотности является «объективной» в том смысле, что в ней отражена лишь та информация, которая имелась в данных.

- При решении конкретной задачи исследователь *добавляет* новую априорную информацию в ее условии и интересуется тем, как знание этой дополнительной информации в сочетании с «объективной» плотностью отразится на апостериорных условных распределениях.

Выясним, например, как распределены примеры класса 0 (точки на рис. 4), имеющие координату y, равную 0.5? В постановке задачи содержатся дополнительные сведения о распределении двух из трех переменных (а именно, две переменные известны достоверно). Для ответа на вопрос задачи формально требуется вычислить свертку трехмерной плотности распределения с двумя дельта-функциями — или, что то же самое, нормированную зависимость плотности от одной переменной при фиксированных значениях двух оставшихся. Типичные результаты приведены на рис. 5.

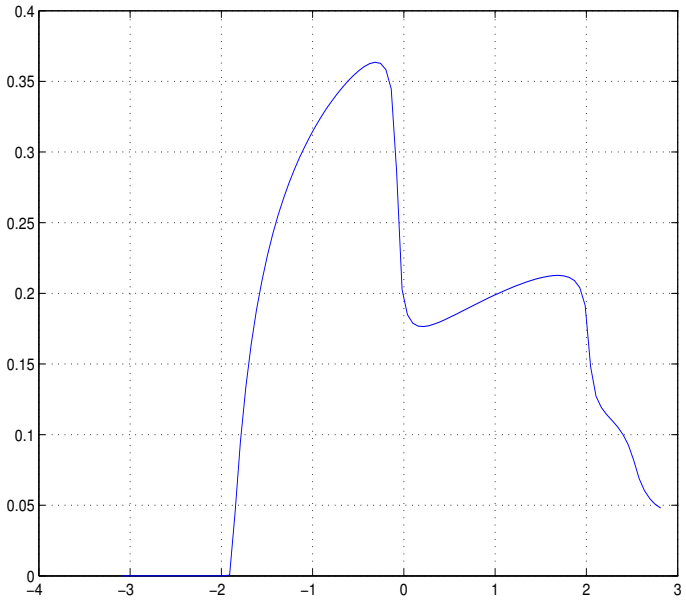

Рис. 5. Апостериорное распределение точек выделенного класса с известным значением одной из координат

Другой вопрос: каковы относительные вероятности встретить примеры разных классов в окрестности координаты $(1.5, 0.5)$? Ответ — появление плюсика в этой точке в 1.5 раза вероятнее, чем точки.

Задача прогноза загрузки процессора ЭВМ (CompAct)

Рассмотрим более реалистичную задачу прогноза доступности процессора ЭВМ для пользователя в реальных условиях эксплуатации (этой ЭВМ). База данных для выбранной задачи (DELVE repository) содержит измерения параметров системной и аппаратной активности в компьютере *Sun SparcStation 20/712* со $128Mb$ памяти, работающем в многопользовательском режиме в условиях университета. Пользователи обычно выполняют большое количество разнообразных задач от доступа в Интернет и редактирования файлов до проведения ресурсоемких расчетов. Данные о состоянии системы записывались 1 раз в 5 секунд. В результате получена 8192 запись с 13 параметрами (см. табл. 2).

Суть традиционной постановки задачи сводится к прогнозу последней переменной — пользовательского КПД системы — по значениям остальных параметров. Исчерпывающее решение этой задачи дано в проекте *Delve*[11]. В обсуждаемой в контексте этой лекции методике для решения требуется[12] построить 13-мерную совместную плотность распределения параметров. Для аппроксимации плотности была использована нейронная сеть с 26 нейронами в скрытом слое. Однако, имея в распоряжении такую аппроксимацию, можно рассмотреть и более интересные аналитико-информационные задачи, нежели просто прогноз переменной *usr*. Например, пусть все параметры кроме размера очереди процессов принимают свои средние значения. Как в этих условиях доля пользовательского процессорного времени зависит от размера очереди процессов? Результирующая зависимость приведена на рис. 6.

Видно, что эта зависимость носит нелинейный характер — вследствие эффектов, вовсе не очевидных из общих соображений оценки нагрузки и т. п.

Такого типа результат не может быть получен путем простой нейросетевой регрессии.

[11] URL: http://www.cs.toronto.edu/ delve/data/datasets.html

[12] Следует отметить, что если пользователя интересует лишь прогноз 13-й переменной по значениям остальных 12-ти, то такая частная задача может быть решена путем применения обычной регрессии (например, с использованием нейронной сети). Нашей же целью является получение решения общей задачи прогноза характера распределения части признаков по известной информации об остальных переменных.

Таблица 2. Параметры задачи о загрузке ЭВМ.

Обозначение	Описание параметра	Мин.	Макс.
lread	Число операций чтения-передач в сек. между системной и пользовательской памятью	0	1845
lwrite	Число операций записи-передач в сек. между системной и пользовательской памятью	0	575
scall	Число системных вызовов всех типов в сек.	109	12493
sread	Число системных вызовов на чтение в сек.	6	5318
swrite	Число системных вызовов на запись в сек.	7	5456
fork	Число системных вызовов fork в сек.	0	20
exec	Число системных вызовов exec в сек.	0	60
rchar	Число символов в сек., передаваемых посредством системных вызовов на чтение	278	2526649
wchar	Число символов в сек., передаваемых посредством системных вызовов на запись	1498	1801623
runqsz	Размер очереди процессов	1	2823
freemen	Число системных страниц, доступных для пользовательских процессов	55	12027
freeswap	Число блоков диска, доступных для своппинга страниц	2	2243187
usr	Доля времени (в процентах) в течении которого процессоры работают в режиме непосредственного обслуживания пользователя (user mode)	0	99

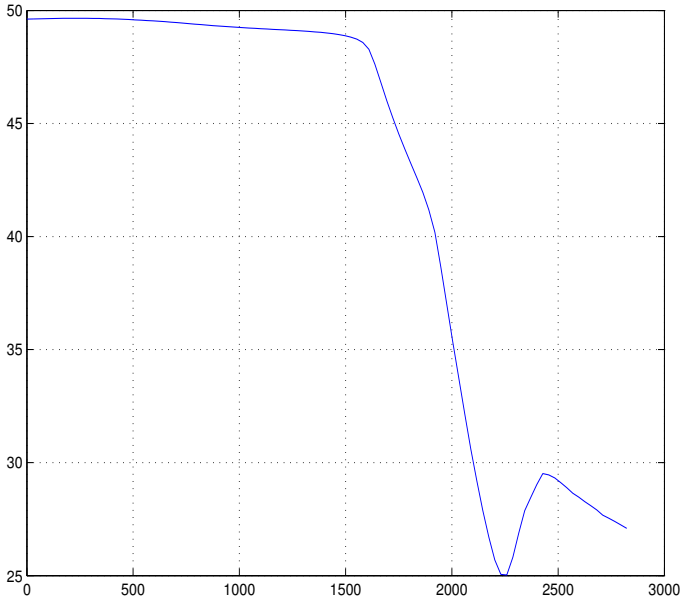

Рис. 6. Зависимость доли пользовательского времени процессора от размера очереди процессов при типичных значениях остальных параметров

Обсуждение

Статистическое описание на основе совместной плотности распределения вероятности приобретает «второе дыхание» вследствие бурного развития вычислительной техники. Так, если совсем забыть об оптимизации кода, то 1 000 000 вычислений значения нейросетевой аппроксимации функции плотности от 10 переменных занимает порядка 15 секунд в системе *MatLab* на «типичном» современном персональном компьютере. Это позволяет применять для оценок условных вероятностей прямые методы Монте-Карло, традиционно считающиеся самыми вычислительно трудоемкими.

Одна из основных проблем, препятствующих созданию полностью автоматизированных методов на основе аппроксимации плотности, как всегда, уходит корнями в «проклятие размерности» — в многомерном пространстве при типичных объемах данных и типичной сложности задачи плотность распределения сконцентрирована в крошечных областях, объем которых ничтожно мал в сравнении с априорным исследуемым объемом. Тем самым крайне затрудняется получение «полезной» статистики.

При решении практических задач обсуждаемыми методами роль эксперта — специалиста в предметной области — состоит в формулировке априорного знания о параметрах задачи в терминах распределений вероятности и интерпретации полученных апостериорных распределений, а роль специалиста в вычислительной математике и информатике сводится к построению эффективных аппроксимаций и получению достоверных оценок при вычислении интегралов методами Монте-Карло. Роль компьютера — выполнять операции над числами.

Благодарности

При подготовке этой лекции автор общался со многими специалистами в фирме «НейрОК» и за ее пределами, всем им большое спасибо. Ю. В. Тюменцев взял на себя нелегкий труд приведения всех лекций в единую систему. Особо хочется поблагодарить А. Н. Горбаня и Н. Г. Макаренко за полезные обсуждения и советы, а также Livermore Software Technology Corporation за финансовую поддержку. Сотрудники Н. Г. Макаренко оказали неоценимую помощь в использовании системы LaTeX.

Литература

1. *Вентцель Е.С.* Теория вероятностей. 6-е изд., стер. – М.: Высшая школа, 1999. – 576 с.
2. *Терехов С.А.* Нейросетевые информационные модели сложных инженерных систем // В сб. «Нейроинформатика» / А. Н. Горбань, В. Л. Дунин–Барковский, А. Н. Кирдин, Е. М. Миркес, А. Ю. Новоходько, Д. А. Россиев, С. А. Терехов, М. Ю. Сенашова, В. Г. Царегородцев. – Новосибирск: Наука, 1998. – с. 101–136.

3. *Бердышев В. И., Петрак Л. В.* Аппроксимация функций. Сжатие численной информации. Приложения. – Екатеринбург, 1999.

4. *Тихонов А. Н., Арсенин В. Я.* Методы решения некорректных задач. 3-е изд., испр. – М.: Наука, 1986. – 288 с.

5. *Neal R. M.* Probabilistic inference using Markov chain Monte Carlo methods. – Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto, 25 Sep 1993.

6. *Ежов А. А.* // Дискуссия о нейрокомпьютерах — 10 лет спустя. Материалы круглого стола «Нейроинформатика–99». – М.: Изд-во МИФИ, 2000. – с. 58.

7. *Ghahramani Z., Jordan M. I.* Learning from incomplete data. – MIT AI Memo-1509, 1994.

8. *Jordan M. I., Jacobs R. A.* Hierarchical mixtures of experts and the EM algorithm. – MIT AI Memo 1440, 1993.
ftp://publications.ai.mit.edu/ai-publications/1000-1499/AIM-1440.ps.Z

9. *Zeevi A. J., Meir R.* Density estimation through convex combinations of densities: Approximation and estimation bounds // Neural Networks. – 1996. – v. 10. – pp. 99–109.

10. *Li J. Q., Barron A. R.* Mixture density estimation // NIPS'99.

11. *Moerland P.* Mixtures of latent variable models for density estimation and classification. – IDIAP-RR 00-25, 2000.
URL: ftp://ftp.idiap.ch/pub/reports /2000 /rr00-25.ps.gz

12. *Терехов С. А.* Вейвлеты и нейронные сети // В сб.: «Лекции по нейроинформатике». – М. Изд-во МИФИ, 2001. – с. 142–182.

13. *D'Agostini G.* Bayesian reasoning in high energy physics — principles and applications. – CERN Yellow Report 99–03, July 1999.

14. *Heckerman D.* A Tutorial on learning with Bayesian networks. – Microsoft Technical Report MSR–TR–95–6, 1995.
URL: http://bayes.wustl.edu/etj/prob/book.pdf.tar.gz

15. *Minka Th.* Independence diagrams. – Tech. Rep. MIT, 1998.
URL: http://www-white.media.mit.edu/ tpminka/papers/diagrams.html

16. *Bishop C. M.* Neural networks for pattern recognition. – Oxford University Press, 1995.

17. *Горбань А. Н.* Обобщенная аппроксимационная теорема и вычислительные возможности нейронных сетей // Сибирский журнал вычислительной математики. – 1998. – Т.1, № 1. – с. 11–24.

18. *Likas A.* Probability density estimation using artificial neural networks // Computer Physics Communications. – 2001. – v. 135, No. 2, pp. 167–175.

19. *Donoho D.L.* Aide-Memoire. High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality // AMS “Math Challenges of the 21st Century”, Stanford, August 8, 2000.

20. *Efron B., Tibshirani R.* Cross-validation and the bootstrap: Estimating the error rate of a prediction rule. – Stanford University Technical Report, May 1995.

21. *Holmes S.* Course 208 Lectures. Introduction to the Bootstrap. – Stanford University, 1999.
URL: http://www-stat.stanford.edu/ susan/courses/s208/web1.html

22. *Blake C. L., Merz C. J.* UCI Repository of Machine Learning Databases. – 1998.
URL: http://www.ics.uci.edu/ mlearn/MLRepository.html

23. *Гилл Ф., Мюррей У., Райт М.* Практическая оптимизация: Пер. с англ. – М.: Мир, 1985. – 509 с.

24. *Riedmiller M., Braun H.* A direct adaptive method for faster backpropagation learning: The RPROP algorithm // Proc. of the IEEE Intern. Conf. on Neural Networks (ICNN) / Ed.: *H. Ruspini*. – San Francisco, 1993. – pp. 586–591.

25. *Neal R. M.* Learning stochastic neural networks. – Technical Report CRG-TR-90-7, 1990, Dept. of Computer Science, University of Toronto.

26. *Terekhoff S. A.* Direct, inverse and combined problems in complex engineered system modeling by artificial neural networks // Proc. SPIE AeroSense Conference, Orlando, Florida, April 21–24, 1997. – Vol. 3077, Paper 71.

27. *Терехов С. А.* Лекции по теории и приложениям искусственных нейронных сетей. – Снежинск, 1994–1998.
URL: http://alife.narod.ru/lectures/neural/Neu_index.htm

Приложение А. Эффективное обучение больших нейронных сетей

Задачи аппроксимации плотности при больших размерностях пространства признаков и, соответственно, больших объемах данных могут приводить к необходимости обучения нейросетей с большим числом неизвестных параметров (синаптических весов межнейронных связей). Это обстоятельство может ограничивать применимость мощных методов обучения типа Левенберга–Маркара или BFGS, так как в них требуется решение

плохо обусловленных систем линейных уравнений высокой размерности [23]. С другой стороны, традиционные схемы градиентного спуска с дифференцированием нейросети методом обратного распространения крайне медленно сходятся.

Напомним, что обучение нейросети обычно связывается с движением против градиента ошибки с целью ее минимизации:

$$\triangle W_{ij}(n) = -\varepsilon \frac{\partial E(n)}{\partial W_{ij}}$$

Более удачное направление поиска для сложных поверхностей ошибки может быть получено путем использования «момента», т. е. памяти о направлении движения на предыдущем шаге:

$$\triangle W_{ij}(n) = -\varepsilon \frac{\partial E(n)}{\partial W_{ij}} + \mu \triangle W_{ij}(n-1)$$

В методе *RProp* (**R**esilient **Prop**agation — «эластичное» распространение), предложенном в [24], коррекция каждого синаптического веса зависит только от знака производной и от поправки на предыдущем шаге. При этом поправка для каждого веса индивидуальна и адаптивна.

В случае, если компонента градиента не изменила знак по сравнению с предыдущей итерацией, оптимизация проходит «гладко», и можно увеличить шаг *для этой компоненты*. В противном случае шаг уменьшается (см. рис. 7):

$$\sigma = \operatorname{sign}\Big(\frac{\partial E(k-1)}{\partial w_{ij}} \; \frac{\partial E(k)}{\partial w_{ij}}\Big)$$

$$\Delta_{ij}^{(k)} = \begin{cases} \min\{\eta^{+}\Delta_{ij}^{(k-1)},\, \Delta_{max}\}, & \sigma = 1 \\ \max\{\eta^{-}\Delta_{ij}^{(k-1)},\, \Delta_{min}\}, & \sigma = -1 \\ \Delta_{ij}^{(k-1)}, & \sigma = 0 \end{cases}$$

Поправка к значению веса синапса вычисляется по формуле:

$$\triangle W_{ij}^{(k)} = \begin{cases} -\operatorname{sign}\Big(\frac{\partial E(k)}{\partial w_{ij}}\Big)\Delta_{ij}^{(k)}, & \sigma \geqslant 0 \\ -\triangle W_{ij}(k), & \sigma < 0 \end{cases}$$

На практике используются следующие типичные значения для параметров, вариация которых мало меняет картину:

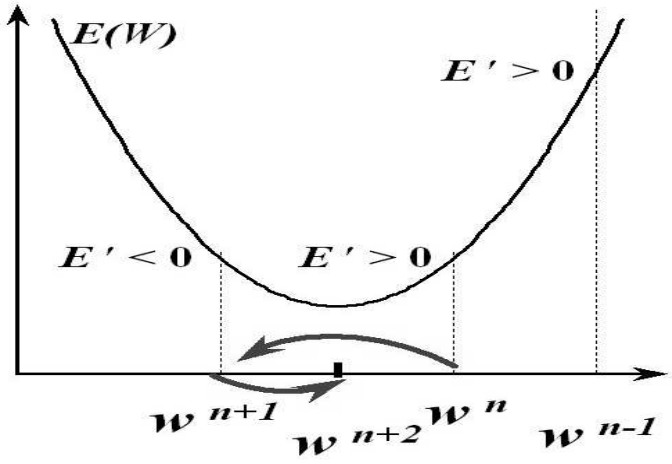

Рис. 7. Идея алгоритма оптимизации методом *RProp*

$$h^+ = 1.2,\ h^- = 0.5,\ D_{min} = 0.0001,\ D_{max} = 50.$$

Как видно из изложения, метод *RProp* является не чем иным, как алгоритмом оптимизации функции многих переменных. Никакая специфика оптимизации именно ошибки нейросети не использована. Приведем поэтому в завершение лекции этот алгоритм оптимизации в виде программы для системы *Matlab*.

```
% RPROP        Поиск безусловного минимума функции
%              многих переменных методом RProp
%
% (c) 2002, Сергей А. Терехов
%
function [x, fx, epoch, dnorm] = ...
    rprop( fgrad, x, max_epochs, min_dnorm )

    % -- Константы метода RProp
    delta_0   = 0.01;
    delta_max = 50;
```

```
delta_min = 1e-6;
delta_inc = 1.2;
delta_dec = 0.5;

% -- Инициализация
delta_x = delta_0*ones(size(x));
grad_x = zeros(size(x));
grad_sign = zeros(size(grad_x));
alldone = 0;
epoch = 0;

% -- Основной цикл поиска минимума
while ~alldone
    % -- Вычислить градиент оптимизируемой
    %    функции в точке с координатами x
    [grad_x, fx] = feval( fgrad, x );

    % -- Применить RProp алгоритм для
    %    вычисления индивидуального шага по
    %    каждой координате
    wrk = grad_x.*grad_sign;
    delta_x = ((wrk>0)*delta_inc + ...
        (wrk<0)*delta_dec + ...
        (wrk==0)).*delta_x;
    grad_sign = zeros( size(grad_sign) );
    grad_sign = grad_sign + ...
        ( (wrk>=0)&(grad_x>0) ) - ...
        ( (wrk>=0)&(grad_x<0) );
    delta_x = min(delta_x, delta_max);
    delta_x = max(delta_x, delta_min);
    dx = - delta_x.*sign(grad_x);

    % -- Сделать шаг по координатам
    x = x + dx;

    % -- Контроль останова итераций
```

```
        dnorm = norm( dx ) / ...
            ( sqrt(prod(size(dx))) + eps );
        epoch = epoch + 1;
        alldone = (epoch >= max_epochs) | ...
            (dnorm <= min_dnorm);
    end
return;

% TEST       Тестовая функция для алгоритма RProp
%
function [grad, fun] = test( x )
    ndim = length( x );
    fun = 0;
    grad = zeros( size(x) );
    for j = 1:ndim
        fun = fun + 0.5*( x(j) - j )^2;
        grad(j) = x(j) - j;
    end
return
```

Вызов программы из командной строки *Matlab* весьма прост:

```
[x, fx, epoch, dnorm] = ...
    rprop( 'test', [11 3 19 124], 100, 1e-6 );
```

Если программа не работает, попробуйте удалить комментарии на русском языке (или замените их английскими). Для обучения нейросети или решения других задач оптимизации функций большого числа переменных от пользователя требуется запрограммировать вид функции, названной здесь `test` (и, по возможности, ознакомиться с книгой [23]).

Сергей Александрович ТЕРЕХОВ, кандидат физико-математических наук, заведующий лабораторией искусственных нейронных сетей СФТИ, заместитель генерального директора ООО «НейрОК». Область научных интересов — анализ данных при помощи искусственных нейронных сетей, генетические алгоритмы, марковские модели, байесовы сети, методы оптимизации, моделирование сложных систем. Автор 1 монографии и более 50 научных публикаций.

Н. Г. МАКАРЕНКО
Институт математики, Алма-Ата, Казахстан
E-mail: makarenko@math.kz

ФРАКТАЛЫ, АТТРАКТОРЫ, НЕЙРОННЫЕ СЕТИ И ВСЕ ТАКОЕ

Аннотация

Известно, что *единое лучше, чем всё вместе, но врозь*. Лекции представляет собой попытку продемонстрировать этот тезис на примере интригующих связей между теорией фракталов, системами гиперболических итеративных функций, дискретными динамическими системами и нейронными сетями. Изложение рассчитано на широкий круг слушателей, которые не являются математиками.

N. MAKARENKO
Institute of Mathematics, Kazakhstan, Alma-Ata
E-mail: makarenko@math.kz

FRACTALS, ATTRACTORS, NEURAL NETWORKS AND ALL THAT

Abstract

It is known that *a single whole is better than all individual things together*. The lecture is to be an attempt to demonstrate this thesis on the example of intriguing connections between fractal theory, hyperbolic iterated function systems, discrete dynamical systems and neural networks. The exposition is counted on wide circle of listeners, who are not mathematicians.

УДК 004.032.26 (06) Нейронные сети

Введение

Предлагаемый фрагмент из еще не завершенного романа публикуется досрочно в качестве раздумья об исподней сущности переживаемого момента, чем и как может он обернуться нам под воздействием нашей неосторожности

Леонид Леонов
«Мироздание по Дымкову»

«Все фигуры, которые я исследовал и назвал фракталами, в моем представлении обладали свойством быть нерегулярными, но самоподобными», — писал Бенуа Мандельброт, который в 1975 году ввел термин *фрактал* (от латинского *fractus* — дробный). Позднее оказалось, что фракталами являются и давно известные в анализе нерегулярные функции, вызывавшие отвращение аналитиков прошлого века [1] [2]. Процесс построения классических фракталов, таких как множество Кантора или ковер Серпинского, который можно принять за их определение, предельно прост. Сперва следует выбрать основную процедуру-генератор, а затем итеративно, до бесконечности применять ее к произвольному компакту. То, что получится в пределе — это и есть фрактал. Идея о том, что объектами геометрии могут быть предельные образы итеративной динамики, не вызывала особого восторга. Можно представить себе (и даже изобразить) на вещественной оси предел числовой последовательности. Однако, мускулы нашего тренированного воображения отказывают, когда пределом является компактный глобально несвязный объект, зачастую обладающий инфернальными свойствами: во многих случаях у него нет даже тени! Да и сами процедуры генерации фрактала напоминали скорее хирургию с элементами садизма, нежели приемы аналитики. Фракталы были чужды уютному евклидову миру с его регулярными структурами.

[1] Если ссылка на примечание представляет собой число, заключенное в круглые скобки, например, (3), то она обозначает номер примечания, помещенного в конце данного текущего раздела. Ссылка в виде числа, не заключенного в такие скобки — обычное подстраничное примечание. — Прим. ред.

[2] «С омерзением и ужасом я отворачиваюсь от этой зловредной язвы — непрерывных функций, нигде не имеющих производных», — писал Эрмит Стилтьесу в 1893 году.

Ситуация изменилась в 1981 году, когда Хатчинсон поместил пришельцев в их родную среду — пространство компактов. Именно здесь *Система Итеративных Функций* (СИФ) — сжимающих отображений, порождает фракталы простым и естественным путем, при котором ампутация отдельных фрагментов связных множеств трансформируется в корректную форму непрерывного сжатия в подходящей метрике. Реминисценции, навеянные предельной динамикой СИФ, ведут к теории *дискретных динамических систем*. Аналогами аффинных СИФ являются здесь нелинейные преобразования, порожденные, например, сечениями Пуанкаре многомерных фазовых потоков. Нелинейность приводит к существенной зависимости от начальных данных так, что малые погрешности экспоненциально растут в фазовом потоке и, начиная с некоторого момента времени, будущее состояние системы становится лишь ограниченно предсказуемым. Этот процесс чаще всего происходит в диссипативных системах, траектории которых заполняют низкоразмерное инвариантное притягивающее подмножество — *аттрактор* в фазовом пространстве. На аттракторе траектории разбегаются в неустойчивых направлениях и сжимаются в устойчивых. Вследствие диссипации сжатие преобладает и в устойчивых направлениях аттрактор копирует сам себя: сечение фазового потока приобретает самоподобную структуру канторова множества с дробной размерностью. Такой аттрактор называют *странным* или фрактальным.

Самоподобная структура фрактала позволяет восстановить СИФ по фрагменту, по крайней мере, в приближении конечного числа итераций. В 1991 году Бресслофф и Штарк показали, что процесс аппроксимации аттрактора СИФ эквивалентен работе бинарной нейронной сети. Таким образом, термины «фрактал» в геометрии и «странный аттрактор» в динамике оказываются синонимами, а СИФ можно рассматривать как рекуррентную асимметричную нейросеть. С другой стороны, Фернандо Ниньо в 2000 году установил, что случайная итеративная нейронная сеть (гипернейрон) топологически эквивалентна динамической системе с заданным аттрактором. Круг замкнулся, образовав Единый Контекст, объединяющий **фракталы**, **СИФ**, **аттракторы** и **нейронные сети**. Цель лекции — показать взаимную связь этих предметов, потому что *единое лучше, чем всё вместе, но по-отдельности*. Более серьезные мотивы указаны в эпиграфе ко введению.

Основой предлагаемой Лекции послужили *Заметки о фракталах* [3] и доклады автора на Фрактальном семинаре Института математики. В тексте опущены все доказательства, и это дает читателю восхитительную возможность *видеть, но не верить!* Поскольку предварительных знаний не всегда хватает как раз для понимания предварительных сведений — они помещены в конце текста, в Глоссарии. Он не обязателен для понимания основного текста: если вас начинает смущать терминология, следуйте простому правилу: *никогда не жуйте пилюлю, которую вас заставляют проглотить*. В конце каждого раздела помещен путеводитель по Литературе, список которой совсем не претендует на полноту.

Автор искренне благодарен своим молодым коллегам: Светлане Ким, Кате Данилкиной и Ерболу Куандыкову, которые взяли на себя неблагодарный труд по набору и редактированию Лекции.

Размерности, площади и объемы

> Необходимо число, различитель *инаковости*, без которого невозможно отличить одно от другого
>
> *Николай Кузанский*
> *«Игра в шар»*

Мы начнем с идей итальянского математика *Никколо Фонтана Тарталья* (1500–1557 гг.). Известно, что на три неколлинеарные точки плоскости можно натянуть треугольник, а четырем некомпланарным точкам в $\mathbb{R}^3$ соответствует тетраэдр. Это выпуклые геометрические тела, для которых существует мера — площади и объемы. Согласно Тарталья, *пространство является n-мерным, если для его $n+1$ точек, не принадлежащих одной гиперплоскости, существует полиэдр ненулевого объема*. Наиболее существенным здесь является то, что в основу определения размерности было положено понятие меры: именно этот момент служит основой современных конструкций. В этом определении неявно используется предположение, что необходимые $n+1$ точки *могут быть всегда выбраны так,* чтобы они не принадлежали одной гиперплоскости. Однако принцип общего положения, который столь широко используется в

[3] URL: http://www.keldysh.ru/dpt_17/works/mak/index.htm

современной математике, требует по меньшей мере понятий непрерывного отображения и многообразия. Поэтому последующие определения размерности рассматривались в именно в таком контексте.

До появления стандартной терминологии произвольное гладкое многообразие называли просто *непрерывностью*. Понятие размерности вводилось аналитически как минимальное число параметров, необходимых для идентификации точки в *непрерывности*. На привычном физическом языке — это n координат точки в $\mathbb{R}^n$. Однако почти сразу же оказалось, что эта **параметрическая размерность** неудовлетворительна по нескольким причинам. Прежде всего выяснилось, что такое определение не различает прямую и плоскость, поскольку между ними можно установить однозначное соответствие. Впервые это показал *Георг Кантор*, используя следующие соображения. Пусть точка $x = (x_1, x_2) \in I = [0,1] \times [0,1]$. Ее координаты $0 \leqslant x_1, x_2 \leqslant 1$ можно представить в виде бинарных дробей: $x_1 = 0, \alpha_1\alpha_2\ldots; x_2 = 0, \beta_1\beta_2\ldots$, где α_i, β_i — нули либо единицы. Для обеспечения единственности условимся записывать обрывающиеся разложения так, чтобы все двоичные цифры, начиная с некоторого места были тогда нулями. Сопоставим паре разложений x_1 и x_2 последовательность $z = 0, \alpha_1\beta_1\alpha_2\beta_2\ldots$, наследующую *лексикографический* порядок исходных оригиналов. Мы получили отображение $z : I \to [0,1]$. Следовательно, существует способ параметризации точек плоскости только одной координатой — извилистой линией! Разумеется, это возможно лишь в том случае, если мы не связываем себя условием непрерывности так, чтобы двум бесконечно близким точкам прямой, соответствовали две бесконечно близкие точки плоскости. *Анри Пуанкаре* заметил, что отказ от непрерывности трудно примирить с интуитивными логическими принципами [(1)]. То, что было предложено им взамен, основано на весьма остроумной концепции **физической непрерывности**. Здесь аналогом точки является элемент, неотличимый от соседних, но совокупность таких элементов образует непрерывную цепь, оба конца которой легко различимы. Размерность непрерывности вводится с помощью понятия *купюры*. Последняя образуется совокупностью элементов, изъятых из непрерывности. Если в результате такой операции непрерывность разделяется на две части, а сама купюра состоит лишь из конечного числа элементов, не образующих связного множества, то размерность непрерывности равна единице. Например, линия одномерна, потому что

делится на две части нуль-мерной купюрой — точкой. Для того, чтобы разделить плоскость, необходима одномерная купюра — линия и т. д. Таким образом, *математическая непрерывность имеет n измерений, если ее можно разбить на части, произведя в ней одно или несколько сечений, которые сами являются непрерывностями $n-1$ измерения*. Это рекуррентное определение размерности, которое предполагает, что объемы — части пространства, поверхности — границы объемов, линии — границы поверхностей, а точки — границы линий.

Формализация этих идей привели *Брауэра, Урысона* и *Менгера* к индуктивному определению **топологической размерности**. Размерность любого конечного или счетного множества точек есть $d_t = 0$. Размерность любого связного множества[4] точек есть $d_t + 1$, если его можно разрезать на два несвязных куска, исключив как минимум d_t-мерное множество точек, т. е. сделав d_t-мерный разрез$^{(2)}$. Ясно, что топологическая размерность всегда есть целое число.

Следующее понятие размерности опирается на идею покрытия и практические способы вычисления длин и площадей, предложенные еще *Борхардом* и *Минковским*. Рассмотрим плоскую кривую $\mathcal{C}$, имеющую в каждой точке непрерывную касательную. Начнем перемещать отрезок длины 2ε так, чтобы его середина двигалась вдоль кривой, а сам он оставался нормалью: касательная определена, следовательно определена и нормаль. Отрезок заметает площадь $S(\varepsilon)$ (мы считаем, конечно, что площадь существует). То, что получается при этом, называется *шарфом Минковского*. При довольно широких допущениях можно показать, что существует $\lim S(\varepsilon)/2\varepsilon$, при $\varepsilon \to 0$, который мы и назовем длиной $\mathcal{C}$. Аналогичным образом можно определить площадь как предел: $\lim V(\varepsilon)/2\varepsilon$, $\varepsilon \to 0$, где $V(\varepsilon)$ — объем тела (*сосиски Минковского*), заметаемого отрезками 2ε, нормальными к поверхности.

Таким образом, меру геометрического объекта некоторой размерности можно определить как скорость изменения меры другого объекта, большей размерности, покрывающего исходный, при *линейном* убывании некоторого характерного масштаба. Заметим, что это определение не зависит от выбора координат.

[4] Топологическое пространство называют *связным*, если его нельзя представить в виде объединения двух непустых множеств.

Для определения размерности используется следующее обобщение. Регулярный многомерный объект можно представить в форме прямого произведения объектов низшей размерности. Тогда его мера (объем) выражается в виде степени характерного размера элемента покрытия — $\varepsilon^{\dim}$. Для сохранения аддитивности меры удобнее перейти от покрытия к разбиению. Степенная зависимость меры от ε называется **скейлингом**. Тогда размерность $(\dim)$ можно определить как скорость изменения меры при измельчении разбиения — т. е. уменьшения ε по *степенному* закону. Поскольку при этом нас интересует только функциональная связь, вместо суммарного объема элементов разбиения (покрытия) часто используют *число* таких элементов. Покрытия (разбиения) можно реализовать многими способами. Поэтому для получения корректных численных оценок берется *минимальное* число элементов покрытия или разбиения. В последнем случае подсчитываются только непустые кубы или шары.

Все эти соображения лежат в основе размерности по *Колмогорову*, которую чаще называют **емкостью**. Пусть (M, ρ) полное метрическое пространство с метрикой ρ и $\dim M = d$. Рассмотрим непустое компактное подмножество $\mathfrak{F} \subset M$. Для $\varepsilon > 0$ пусть $\mathcal{B}(x, \varepsilon)$ — замкнутый d-мерный шар с центром в точке $x \in M$.

Подсчитаем наименьшее число $N(\varepsilon)$ таких шаров, необходимых для покрытия $\mathfrak{F}$: $N(\varepsilon) =$ *наименьшему целому* m *такому, что* $\mathfrak{F} \subset \cup^m \mathcal{B}(x_n, \varepsilon)$ для $\{x_n \,|\, n = 1, 2, \ldots, m\} \subset M$. Такое число всегда существует, потому что $\mathfrak{F}$ — компакт, т. е. ограниченное и замкнутое множество. Следовательно, из каждого его бесконечного покрытия можно выбрать конечное подпокрытие. Тогда искомое число есть минимальная длина всех таких конечных последовательностей.

Если существует *скейлинг* $N(\varepsilon) \sim \varepsilon^{-d_c}$ при $\varepsilon \to 0$, то d_c — **колмогоровская емкость** множества. Заметим, что $d_c \leqslant d$ и не обязательно целое число! Это на первый взгляд кажется парадоксальным, потому что мы использовали d-мерные шары, где d — *целое* число. Однако, d_c определяется из скейлинга, который выполняется *асимптотически*, в пределе исчезающих ε. Для корректного определения d_c следует учесть еще нормировку, равную объему d-мерной единичной сферы $V(1)$. Тогда емкость можно определить как такое число d_c, при котором существует отличный от нуля предел[(3)]: $\lim\{V(1)N(\varepsilon)\varepsilon^d\}$ при $\varepsilon \to 0$. При использовании кубов вместо шаров необходимость в нормировке исчезает.

Размерность Хаусдорфа d_H отличается от емкости тем, что следует брать разные шары, радиус которых $r_i \leqslant \varepsilon$. Тогда *крупнозернистая* α-мерная мера Хаусдорфа для любого $\alpha > 0$ определяется как

$$m_\alpha(\mathfrak{F}, \varepsilon) = \inf \sum_i (r_i)^\alpha,$$

где $\inf$ берется по всем покрытиям. Когда ε убывает, сумма возрастает или, во всяком случае, не убывает. Поэтому предел при $\varepsilon \to 0$, конечный или бесконечный, существует и называется *α-мерной мерой Хаусдорфа* для $\mathfrak{F}$: $m_\alpha(\mathfrak{F}) = \lim_{\varepsilon \to 0} m_\alpha(\mathfrak{F}, \varepsilon)$. *Хаусдорфова размерность* $\mathfrak{F}$ характеризует поведение $m_\alpha(\mathfrak{F})$ как функцию от α:

$$\dim_H \mathfrak{F} = \sup\{\alpha | m_\alpha(\mathfrak{F}) = \infty\} = \inf\{\alpha | m_\alpha(\mathfrak{F}) = 0\}.$$

Таким образом, $\dim_H$ отделяет значения α, дающие бесконечную меру, от значений α, приводящих к нулевой мере. В качестве примера рассмотрим фрагмент кривой длины L. Ясно, что для его покрытия (разбиения) требуется $N = L/\varepsilon$ шаров. Тогда мера $m = (L/\varepsilon)\varepsilon^\alpha = L\varepsilon^{\alpha-1}$. Если $\alpha < 1$, то $m \to \infty$ при $\varepsilon \to 0$. Если же $\alpha > 1$, то $m \to 0$ при $\varepsilon \to 0$. Единственным «правильным» значением будет $\alpha = 1$. Заметим, что d_H для некоторых относительно простых множеств может существенно отличаться от d_c. Тем не менее, мы будем использовать оба термина как синонимы, полагая $d_c = d_H$.

Приведенным выше определениям можно придать информационный смысл. Представим себе кассу из двух ящиков. В одном из них находится монета. Очевидно необходим всего один вопрос с возможными ответами «да» или «нет», чтобы идентифицировать ее нахождение. Четыре ящика требуют двух вопросов, а если касса состоит из $2^5 = 32$ ящиков, таких вопросов надо задать уже 5. На первый взгляд кажется, что каждый следующий вопрос следует выбирать только после получения ответа на предыдущий. Однако, венгерский математик *Реньи* показал, что можно задать всего один «сложный» вопрос, допускающий несколько «одновременных» ответов «да», «нет», позволяющих однозначно идентифицировать необходимый номер. Будем считать, что информация, которую содержит один ответ, равна одному биту. Тогда для идентификации системы с N возможными и равновероятными состояниями нам необходима информация в $\mathcal{I} = \log_2 N \equiv \lg N$ бит. Мы пришли к формуле, которую впервые получил *Хартли* в 1928 году.

Вопросы и ответы можно реализовать по-разному. Для нас удобно представить себе вопрос как элемент покрытия (нечто вроде фишки при игре в лото), идентифицирующий «точку» множества с точностью ε. Если $N(\varepsilon)$ — минимальное число таких элементов, $\mathfrak{I} = \log_2 N(\varepsilon)$ — количество информации, необходимое для идентификации всего множества с той же точностью. В этом контексте размерность:

$$d_c = -\lim_{\varepsilon \to 0} \mathfrak{I} / \log \varepsilon$$

является *скоростью изменения информации при бесконечном увеличении разрешения*.

Примечания

1. *А. Пуанкаре* писал по поводу концепций, не согласованных с интуицией: «Они заменяют определяемый предмет и интуитивное понятие этого предмета конструкцией, сделанной из более простых материалов. Мы видим, что из этих материалов действительно можно собрать такую конструкцию, но никогда непонятно, почему собрали эти материалы именно так, а не иначе. Я не хочу сказать, что арифметизация математики — плохая вещь, я утверждаю лишь, что она не составляет всего» [1]. Несмотря на указанные недостатки, параметрическая размерность осталась в физике, где она согласована с числом степеней свободы.

2. Такая размерность называется большой индуктивной размерностью — Ind. Кроме того, существует малая индуктивная размерность (ind) : считают, что $\mathrm{ind}X = n$, если $\forall x \in X$ существует сколь угодно малая окрестность $U : x \in U$, граница которой ∂U имеет $\mathrm{ind}(\partial U) = n-1$. Начало этой индуктивной цепочки — пустое множество $\varnothing$, для которого $\mathrm{ind}\varnothing = -1$.

3. Точнее, при $\varepsilon \to 0$ существует соотношение: $N(\varepsilon) \propto V(1)\varepsilon^d$, в котором знак $\propto$ понимается в следующем смысле. Для двух функций f и g, $f(\varepsilon) \propto g(\varepsilon)$, если $\lim\{\ln f(\varepsilon)/\ln g(\varepsilon)\} = 1$ при $\varepsilon \to 0$. Поэтому **емкость** определяется как предел, *если он вообще существует*, отношения

$$\{\ln N(\varepsilon) - \ln V(1)\}/\{\ln(1/\varepsilon)\}.$$

Очевидно, что $\ln V(1)/\ln(1/\varepsilon) \to 0$ при $\varepsilon \to 0$, поэтому этот член обычно опускают. Существуют две теоремы [2], которые позволяют перейти к дискретному варианту для ε. Первая из них утверждает, что емкость компактного подмножества $\mathfrak{F}$ совпадает с пределом:

$$\lim_{n\to\infty} \{\ln N(\varepsilon_n)\}/\{\ln(1/\varepsilon_n)\},$$

где $\varepsilon_n = cr^n$, $0 < r < 1$, $c > 0$, $n = 1, 2, \ldots$, так что радиусы шаров можно менять дискретно. Вторая теорема позволяет использовать боксы размером $1/2^n$. Если $N_n(\mathfrak{F})$ число таких боксов, пересекающих $\mathfrak{F}$, то емкость определяется как предел при $n \to \infty$:

$$\lim\{\ln N_n(\mathfrak{F})\}/\{\ln 2^n\}.$$

Путеводитель по литературе. Подход Пуанкаре к понятию размерности изложен им в книге [1]. О работах Минковского и Борхарда можно узнать из монографии Лебега [3]. Элементарное изложение современных понятий о математической непрерывности содержат учебные курсы [4, 5]. В монографии [6] можно найти много интересного по истории вопроса. Наконец, подробное описание всевозможных размерностей можно найти в обзорах [7–9]. С информацией Шеннона и ее применением можно познакомиться по книге [10] или замечательным беседам Реньи [11].

Дробные размерности

> Оставаясь на почве любой физической теории, мы не можем интерпретировать формулы, содержащие дробные показатели основных единиц.
>
> *В. Вильямс*

История началась с попыток английского физика *Ричардсона* измерить длину побережья Британии. Располагая подробной картой, он аппроксимировал линию побережья ломаной L_b, составленной из отдельных хорд длиной b, все вершины которой лежали на берегу. *Ричардсон* полагал,

что в этом случае существует предел $L_b \to L$ при $b \to 0$, как бывает для гладких кривых. Однако, оказалось, что с уменьшением b суммарная длина ломаной растет до бесконечности по закону:

$$L_b = \lambda b^{1-D},\ D < 2.$$

Если построить на каждом звене b квадрат, то суммарная площадь квадратов: $b^2 N = b^2 L_b / b = \lambda b^{2-D} \to 0$ при $b \to 0$ и $D < 2$. Таким образом, береговая линия имеет бесконечную длину и порождает нулевую площадь! Первое, что приходит в голову, это сомнения в корректности такой аппроксимации. В математике уже давно известны случаи, когда метод хорд не работал.

Классический пример привел еще *Лебег*: возьмем равносторонний треугольник ABC и соединим середины трех сторон (см. рис. 1).

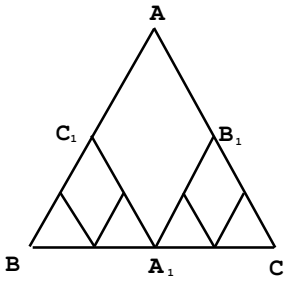

Рис. 1. Парадокс Лебега

Очевидно, $AB + AC = BC_1 + C_1A_1 + A_1B_1 + B_1C$. Продолжая процесс разбиения, мы приближаем ломаную к стороне BC. Если считать последнюю пределом ломаной, мы получим $AB + AC = BC$. Для того, чтобы разобраться в деталях парадокса, необходима нетривиальная математика: понятие тонкой сходимости, трансфинитные повторные пределы и т. п. Мы заменим все это здравым смыслом и простыми геометрическими соображениями. В действительности, сторона BC не является пределом зигзагообразной ломаной — эта ломаная действительно приближается к BC по положению, *но не по направлению*. Все

дело здесь в степени гладкости. Вспомним, что длина кривой выражается интегралом, под знаком которого стоят производные от функции, задающей кривую в параметрическом виде. Таким образом, для вычисления длины необходима гладкость кривой по меньшей мере $\mathcal{C}^1$, т. е. ее уравнение имеет непрерывную первую производную. Длина спрямляемой кривой действительно ограничена верхней гранью суммарных длин вписанных ломаных. На такой кривой $x = x(s), s \in [0,1]$ конечной длины L в качестве параметра удобно выбрать длину дуги. В этом случае справедливо условие **Липшица**: $|x(s_1) - x(s_2)| \leqslant |s_1 - s_2|$ или при $s \to t, s = Lt : |x(t_2) - x(t_1)| \leqslant L|t_2 - t_1|$ и касательная вдоль кривой меняется непрерывно. Углы, которые образует произвольная хорда с касательными, не превосходят наибольшего из углов, образованного разными касательными в концах дуги. Поэтому, когда число вписанных в дугу звеньев ломаной возрастает, последняя приближается к дуге не только по положению, *но и по направлению*$^{(1)}$. В случае, если гладкость меньше чем $\mathcal{C}^1$, хорды при измельчении стремятся встать перпендикулярно к стороне треугольника, что и приводит к расходящемуся пределу.

Существует аналог парадокса *Лебега*. В конце прошлого века *Герман Шварц* показал, что триангуляция цилиндра единичного радиуса и единичной высоты дает для площади боковой грани произвольную величину. Причины обоих парадоксов — в функциях, гладкость которых ниже той, с которой мы привыкли иметь дело в обычном анализе.

В общем случае формула линейных приращений: $\triangle y = \mu \triangle x + O(\triangle x^2)$ заменяется на $\triangle y = \mu_H (\triangle x)^H$, где H — **показатель Гельдера**, а μ и μ_H — обычная и гельдеровская производные. Показатель $H = 1$ для гладких функций. В случае $H < 1$ вместо касательной имеется криволинейный конус $\triangle y \approx \triangle x^H$. Объекты, для которых $H < 1$, а показатель D в формуле *Ричардсона* строго больше единицы, называются *фракталами*. Рассмотрим несколько известных примеров.

Пример 1. **Множество Кантора.** Пусть $\mathfrak{F}_0 = [0,1]$. Выбросим из $\mathfrak{F}_0$ интервал $(1/3, 2/3)$, а то что останется, обозначим $\mathfrak{F}_1$. Затем выбросим из $\mathfrak{F}_1$ интервалы $(1/9, 2/9)$ и $(7/9, 8/9)$ и получим $\mathfrak{F}_2$ (рис. 2). Продолжая этот процесс, мы придем к убывающей последовательности замкнутых интервалов $\{\mathfrak{F}_n\}$. Множество $\mathfrak{F} = \cap \mathfrak{F}_n$ называют ***канторовым множеством***. Пусть $\mathfrak{S}$ — множество «выброшенных» кусков отрезка $[0,1]$ при построении $\mathfrak{F}$, т. е. $\mathfrak{S} = (1/3, 2/3) \cup (1/9, 2/9) \cup (7/9, 8/9) \cup \ldots$ Тогда

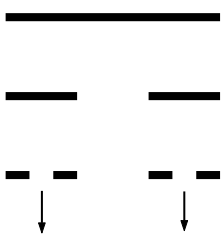

Рис. 2. Множество Кантора

лебегова мера $\mathfrak{S}$ равна сумме: $1/3 + 2 \times 1/9 + 4 \times 1/27 + \ldots = 1$. Таким образом, при построении $\mathfrak{F}$ мы выбросили всю длину отрезка! Но осталось бесконечное множество точек, *канторово множество*: $[0,1]\backslash\mathfrak{S}$, имеющее мощность континуума и *лебегову* меру нуль[(2)]. Например, ему принадлежат точки $\{0, 1, 1/3, 2/3, 1/9, \ldots\}$ — то есть концы выбрасываемых интервалов, но не только они! Все точки $\mathfrak{F}_\circ \in \mathfrak{F}$ можно описать следующим образом. Запишем каждое из чисел $x \in [0,1]$ в троичной системе: $x = a_1/3 + a_2/3^2 + \ldots + a_n/3^n + \ldots$, где $a_n = 0, 1$ или 2. Некоторые числа в этом представлении допускают двойную запись, например: $1/3 = 1/3 + 0/3^2 + \ldots$ и $1/3 = 0/3 + 2/3^2 + 2/3^3 + \ldots$. Легко убедиться, что $\mathfrak{F}$ принадлежат только те x, которые могут быть записаны хотя бы одним способом так, что в последовательности числителей $a_1, a_2, \ldots, a_n, \ldots$ ни разу не встречается единица. Совокупность таких последовательностей имеет мощность континуума[(3)].

Вычислим размерность самоподобия канторова множества. При первой итерации имеем $\varepsilon = 1/3, N = 2$, при второй — $\varepsilon = 1/9, N = 2^2; \ldots$, при k-ой — $\varepsilon = 1/3^k, N = 2^k$. Поэтому $d_H = \ln 2/\ln 3 \approx 0,63$ — меньше, чем размерность исходного прообраза.

Существует функциональный аналог множества Кантора — *функция Кантора*. Распределим равномерно на $\mathfrak{F}$ единичную массу (меру) с плотностью μ. Тогда функция $F(x) = \int_0^x d\mu(x)$ описывает распределение меры на канторовом носителе. Она является непрерывной возрастающей функцией, которая тем не менее почти всюду имеет нулевую производную (т. е. горизонтальна!). Ее называют «чертовой лестницей» (рис. 3).

Пример 2. **Кривая Коха.** Возьмем равносторонний треугольник и определим следующую элементарную операцию: каждая сторона делит-

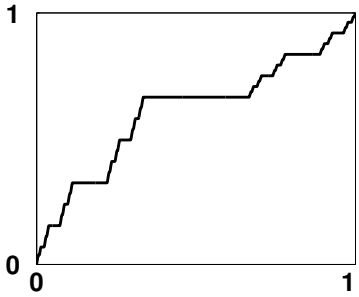

Рис. 3. Чертова лестница

ся на $r = 3$ части, после чего средний сегмент заменяется на равносторонний треугольник. Операция повторяется n раз. То, что получится при $n \to \infty$, и есть триада (снежинка) *Коха* (рис. 4)

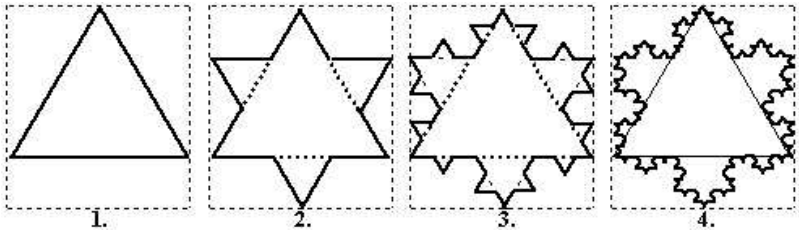

Рис. 4. Триада Коха. Четыре итерации

Ясно, что это множество самоподобно, точно так же, как и сам процесс его построения. На любом шаге его линейный элемент длины l заменяется на $N = 4$ элемента длиной $lr^{-1} = 1/3$ каждый. Поэтому размерность подобия или емкость $d_c = \ln N / \ln r = \ln 4 / \ln 3 \approx 1,2619$. Триада Коха является математической моделью кривой побережья, с которой работал *Ричардсон*. Любопытно, что в $\mathbb{R}^2$ можно построить множества с дробной размерностью, которые обладают инфернальным свойством *призраков* или *вурдалаков*: они *не отбрасывают тени*. Точнее, для таких объектов лебегова мера проекций на некоторые (или даже на все) направления равна нулю.

Примечания

1. Наиболее важным обстоятельством здесь является переход к пределу даже для гладкой кривой. Действительно, сегмент дуги рассматриваемый непосредственно, ничем не отличается от его половины удвоенной микроскопом. Целое однородно с частью и как заметил Пуанкаре [1]: «Здесь заключается противоречие, или скорее это было бы противоречием, если бы число членов предполагалось конечным; в самом деле, ясно, что часть, которая содержит менее членов сравнительно с целым, не может быть подобной целому».

2. Напомним, что множество $\mathfrak{X}$ на $\mathfrak{R}$ имеет меру нуль, если его можно покрыть интервалами, сумма длин которых произвольно мала. Например, $\mathfrak{X}$ состоит из рациональных чисел из $(0, 1)$. Они могут быть записаны в виде последовательности: $a_1, a_2, a_3, \ldots$, например так:

   $$1/2,\ 1/3,\ 2/3,\ 3/4,\ 1/5,\ 2/5, \ldots$$

   Для любого $\varepsilon > 0$, a_1 можно заключить в интервал длины $\varepsilon/2$; a_2 — в интервал длины $\varepsilon/4; \ldots$ a_n — в интервал длины $\varepsilon/2^n$. Эти интервалы покрывают $\mathfrak{X}$ и сумма их длин равна ε. Следовательно, множество рациональных чисел имеет меру нуль.

3. Действительно, любому набору $\{a_1, \ldots, a_n, \ldots \mid a_i = 0 \text{ или } 2\}$ можно поставить в соответствие последовательность $b_1, \ldots, b_n, \ldots$, где $b_n = 0$, если $a_n = 0$ и $b_n = 1$, если $a_n = 2$. Но такая последовательность — просто двоичный код числа, принадлежащего $[0, 1]$.

Путеводитель по литературе. Лучшими, по моему мнению, являются обзоры [7,9] написанные для физиков. Работы Ричардсона изложены в монографии [12], а в [13] им посвящена целая глава. Статьи Мандельброта в [14,15] весьма содержательны, но не подходят для первого чтения. Небольшая книга Лебега [3] содержит обсуждение упомянутых парадоксов. Описание функции Кантора можно найти в [16], а примеры фракталов-вурдалаков приведены в дополнении к книге [17]. Наконец, некоторые вопросы, связанные с тонкой сходимостью, строго рассмотрены в [18].

Фракталы, неполная автомодельность и контекстно-свободные грамматики

Plus ça change, plus c'est la même chose.
Чем больше оно изменяется, тем более остается тем же.

Французская пословица

Вернемся к построению *триады Коха*. Здесь все легко считается. Обозначим через L ребро треугольника на первом шаге. На n-ом шаге размер одной стороны $b = L/3^n$, а периметр $L_b = 3L(4/3)^n$ (рис. 4).

Когда $n \to \infty, b \to 0$, $L_b \to \infty$. Поскольку $n = \lg(L/b)/\lg 3$, имеем:

$$L_b = 3L10^{\alpha \lg(L/b)} = 3L(L/b)^{\alpha},$$

где $\alpha = (\lg 4 - \lg 3)/\lg 3 \approx 0,2618$.

Мы легко получим формулу *Ричардсона*, если положим:

$$\lambda = L^{1-\alpha} = L^D;\ L_b = 3\lambda b^{1-D};\ D = 1 + \alpha = 1,2619.$$

Аналогия с береговой линией становится полной, поскольку число звеньев триады $N = L_b/b = \lambda b^{-D}$ и суммарная площадь квадратов, построенных на периметре, $Nb^2 = \lambda b^{2-D} \to 0$ при $b \to 0\,(D < 2)$. Мы можем получить конечную величину, если удастся найти такое D_*, что $Nb^{D_*} = \lambda$, где $1 < D_* < 2$; при этом $Nb = \infty$ и $Nb^2 = 0$, когда $b \to 0$.

Легко усмотреть основные особенности построения триады:

1. **Однородность**: каждая сторона треугольника n-го шага порождает одинаковое число звеньев $n + 1$.

2. **Самоподобие**: число звеньев $n+1$ шага зависит только от отношения звеньев n-го и $n + 1$ шагов.

Оказывается, что этих двух условий, заданных локально, достаточно для получения *формулы Ричардсона*[(1)].

Рассмотрим произвольную кривую, которая аппроксимируется системой ломаных линий с уменьшающейся длиной звена. Пусть малая

окрестность кривой содержит две соседних вершины ломаной с длиной звена a. Пусть N_{ab} — число вершин ломаной со звеном $b < a$, расположенных между теми же вершинами. Свойство *локального* самоподобия означает, что N_{ab} асимптотически при $a/b \to \infty$ можно представить разложением, главный член которого не зависит от a и b, а является функцией их отношения: $N_{ab} = f(a/b)$ при фиксированном $a/b \gg 1$. Возьмем новую ломаную с длиной звена $c \ll b$. В силу локальной однородности и самоподобия число ее звеньев, приходящихся на длину a, равно $f(a/c)$. С другой стороны, это же число равно произведению $(N_{ab}) \times (N_{bc})$. Здесь первый сомножитель — число b–звеньев внутри одного a–звена, а второй — число c–звеньев внутри одного b–звена. Таким образом, получаем функциональное уравнение:

$$f(a/c) = f(a/b)f(b/c).$$

Замена переменных $a/b = x$; $a/c = y$ дает:

$$f(y)/f(x) = f(y/x).$$

Дифференцируя обе части по y и полагая $y = x$, получим: $f'(x)/f(x) = (1/x)f'(1) = D/x$ или $f(x) = x^D$.

Следовательно, при упомянутых предположениях для длины ломаной справедлива асимптотика:

$$L_b = a^D b^{1-D} + \ldots$$

Напомним, что процедура называется **автомодельной**, если ее характеристики на двух смежных уровнях связаны друг с другом преобразованием подобия. Асимптотика показывает, что фракталы обладают свойством неполной автомодельности. Поясним этот момент. В общем случае длина аппроксимирующей ломаной для непрерывной кривой между двумя точками, разделенными расстоянием a, зависит от a и длины звена b. Из соображений размерности $L_a = af(a/b)$. Для гладкой кривой при $a/b \to \infty$ $(b \to 0)$ функция f стремится к конечному пределу $f(\infty)$. Именно тогда величина $f(\infty)a$ является длиной отрезка гладкой кривой. Например, для окружности с диаметром a, $f(\infty) = \pi/2$. Таким образом, длина окружности автомодельна по параметру a/b при $a/b \to \infty$. Для фрактальных кривых конечного предела $f(a/b)$ не существует. Однако

имеет место **неполная автомодельность**:

$$f(a/b) \propto (a/b)^{D-1}.$$

Разумеется, D зависит от геометрии кривой и не определяется из соображений размерности.

Существует любопытный подход к неполной автомодельности, связанный с *символической динамикой*. Рассмотрим **бинарное слово** (двоичную последовательность):

$$z = z_0 z_1 \dots z_{r-1},$$

где z_i — нули, либо единицы. Длина слова $\log_2 z = r$, а $s \leqslant r$ число мест, занятых единицей. Определим гомоморфизм $\mathbf{T} : [0,1] \to [0,1]$ уравнениями: $\mathbf{T}(0) = 0^r$, $\mathbf{T}(1) = z$. Пусть, например, $z = 101$. Возьмем точку $x_0 = 1$ и рассмотрим последовательность итераций: $x_1 = \mathbf{T}(x_0), \dots, x_k = \mathbf{T}(x_{k-1})$. В результате получим последовательность:

$$\begin{gathered} x_0 = 1 \\ x_1 = \mathbf{T}(1) = 101 \\ x_2 = \mathbf{T}(\mathbf{T}(1)) = \mathbf{T}(101) = \mathbf{T}(1)\mathbf{T}(0)\mathbf{T}(1) = 101000101 \end{gathered}$$

Можно убедиться, что неподвижной точкой x_f гомоморфизма $\mathbf{T}$ будет предел $\lim_{k\to\infty} x_k$, удовлетворяющий уравнению $\mathbf{T}(x) = x$:

$$x_f = 10100010100000000010100010100\dots$$

Это символическая запись *множества Кантора*. Действительно, 101 представляет собой *код* первого шага построения этого фрактала, когда мы удаляем середину (0) единичного интервала. На втором шаге та же процедура применяется к получившимся подинтервалам: $[0, 1/3] \to 101$; $[2/3, 1] \to 101$ и т. д.

Примечания

1. Или на другом языке — получения **ренормгруппы**, относительно которой самоподобные системы являются инвариантными множествами. *Канторову пыль* $\mathfrak{K}$ можно рассматривать как подмножества интервала $[0,1]$. Пусть преобразование S множества $\mathfrak{F}$ на вещественной оси дает образ $S(\mathfrak{F})$ и обладает следующим свойством:

образ объединения двух множеств равен объединению их образов. Рассмотрим два отображения:

$$\mathbf{S}_1(\mathfrak{F}) = [x/3 : x \in \mathfrak{F}]$$
$$\mathbf{S}_2(\mathfrak{F}) = [(x+2)/3 : x \in \mathfrak{F}]$$

$\mathbf{S}_1$ отображает *канторово множество* в его левую половину, а $\mathbf{S}_2$ — в правую. Очевидно, $\mathbf{T}(\mathfrak{F}) = S_1 \cup S_2$ — оставляет пыль неизменной. Преобразования, оставляющие множество неизменным, $\mathbf{T}(\mathfrak{K}) = \mathfrak{K}$ называются **симметриями** $\mathfrak{K}$. Если приложить $\mathbf{T}$ к интервалу $[0, 1]$ несколько раз, то $\lim_{n\to\infty} \mathbf{T}^n([0, 1]) = \mathfrak{K}$. Фактически $\mathfrak{K}$ можно генерировать из любого ограниченного множества вещественной оси, даже из любой единственной точки x_0 : $\mathbf{T}(x_0)$ тогда состоит из двух точек, $\mathbf{T}^2(x_0)$ — из четырех и т. д. На языке теории ренорм-групп $\mathfrak{K}$ — единственная притягивающая «точка» или инвариантное множество $\mathbf{T}$. В этом контексте $d_c = \ln 2/\ln 3$ результат того, что $\mathbf{T}$ объединение двух сжатий. Можно показать, что $\mathbf{T}, \mathbf{T}^2, \ldots$ — единственные симметрии канторова множества.

Путеводитель по литературе. Монография [13] — единственный известный мне источник, где фракталы изложены с позиций неполной автомодельности. Связь свойств самоподобия и однородности с теорией групп обсуждается в обзоре [19]. Связям фракталов с теорией чисел и символической динамикой посвящены работы [2, 20, 23].

Фракталы и системы итеративных функций

> Но как же оно образуется, если не содержится в том же своем прообразе?
>
> *Николай Кузанский*
> *«Игра в шар»*

Отображение $f : X \to X$, действующее в метрическом пространстве (X, d), называется *сжимающим*, если существует $s \in [0, 1)$ такое, что

$$d(f(x), f(y)) \leqslant sd(x, y)$$

для всех x и y из X; s называется коэффициентом сжатия отображения.

Системой итеративных функций (СИФ) $(X, \{f_i\}), i = 1, 2, \ldots, k$ называют набор сжимающих отображений[5] $\{f_i\}$, в компактном метрическом пространстве (X, d). Коэффициентом сжатия СИФ называют $s = \max\{s_i : i = 1, \ldots, k\}$.

Для пространства (X, d) можно определить другое метрическое пространство $(H(X), h)$, называемое «пространством фракталов».

Пусть $H(X)$ — множество непустых компактных подмножеств X. Определим на нем *метрику Хаусдорфа* h следующим образом. Пусть $B(x, r)$ — замкнутый шар, радиуса r с центром в точке x. Для произвольного множества $A \in X$ *дилатацией*[6] A_r радиуса r множества A называется $\bigcup_{x \in A} B(x, r)$. Таким образом, дилатация множества A — это добавление к A всех точек, лежащих на расстоянии $\leqslant r$ от его границы. Пусть A, B — непустые компактные подмножества из X. Тогда расстояние(метрика) Хаусдорфа определяется как (рис. 5)

$$h(A, B) = \min\{r > 0 | A \subset B_r; B \subset A_r\}.$$

Таким образом, это минимальное из двух чисел: первое из них получается расширением множества A, до тех пор, пока его образ не поглотит B, второе — дилатацией B, пока она не поглотит A[7].

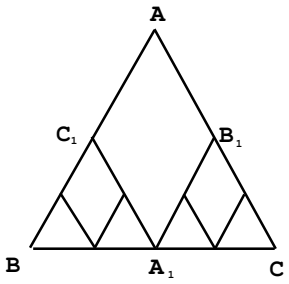

Рис. 5. Метрика Хаусдорфа

[5] Такую СИФ иногда называют гиперболической.
[6] Т.е. расширением.
[7] Другое определение дано в глоссарии.

Можно показать, что если (X, d) — полное метрическое пространство, то $(H(X), h)$ также является полным метрическим пространством. Определим преобразование $T : H(X) \to H(X)$ как

$$T(B) = \bigcup_{i=1}^{k} f_i(B), \forall B \in H(X),$$

где $T(B) = \{T(x)|x \in B\}$ — *оператор Хатчинсона*.

Пусть $B \in H$ и $T^{\circ n}$ — композиция [8] порядка n оператора T:

$$T^{\circ 0}(B) = B,\ T^{\circ(j+1)}(B) = T^{\circ j}(T(B)).$$

Последовательность множеств, полученная в результате итерирования $T(B)$, т. е.

$$\{B, T(B), T^{\circ 2}(b), T^{\circ 3}(B), ..., T^{\circ n}(B), ...\},$$

называется *орбитой* B для $(H(X), T)$. Пару $(H(X), T)$ можно рассматривать как *детерминированную дискретную динамическую систему* с пространством состояний $H(X)$ и преобразованием T.

Согласно *теореме Банаха о неподвижной точке*, действие сжимающего преобразования T на произвольную начальную точку $B_0 \in H$ в пространстве $(H(X), h)$ приводит к последовательности точек $B_0, B_1 = T(B_0), B_2 = T(B_1) \equiv T^{\circ 2}(B_0), \ldots$, которая сходится к некоторой точке $A \in H$. Она является единственным решением уравнения $T(A) = A$. Очевидно, что

$$\lim_{n\to\infty} T^{\circ n}(B) = A.$$

Неподвижная точка A называется *аттрактором* СИФ или *фракталом*.

При практическом использовании СИФ для построения фракталов оператор Хатчинсона применяют только к граничным точкам компакта.

Пример 1. Пусть $X = [0, 1]$, $f_1 = 1/3x$, $f_2 = 1/3x + 2/3$. Пусть $T = f_1 \cup f_2$. Последовательность итераций $T(X)$ при $n \to \infty$

$$T([0, 1]) = \{[0, 1/3], [2/3, 1]\},$$

$$T^{\circ 2} = \{[0, 1/9], [2/9, 1/3], [2/3, 7/9], [8/9, 1]\}, \ldots$$

[8] Не путать со степенью!

очевидно приводит к канторову множеству: $\mathfrak{F} = \bigcap T^{\circ n}[0,1]$.

ПРИМЕР 2. Пусть $X \in \mathbb{R}^2$. Определим СИФ как

$$f_1\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

$$f_2\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1/2 \\ 0 \end{bmatrix},$$

$$f_3\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}.$$

Аттрактор этой СИФ (ковер Серпинского) приведен на рис. 6. Возьмем

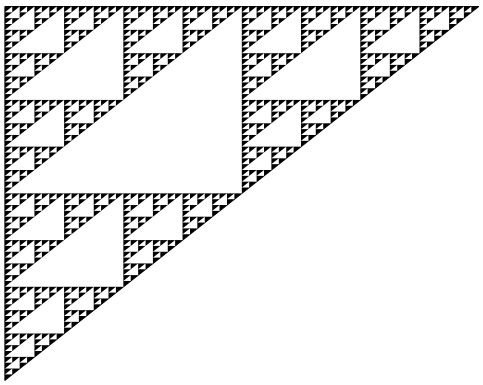

РИС. 6. Фрактал Серпинского

в качестве исходного множества при построении ковра прямоугольный треугольник B с координатами вершин $(0,0), (0,1), (1,1)$. Тогда все f_i — аффинные преобразования сжатия B с коэффициентом 0.5. Преобразование f_3 сдвигает образ B вправо на $1/2$, f_2 — вверх на эту же величину и f_1 оставляет его на месте. Таким образом, для каждого n $T^{\circ n}(B)$ репродуцирует сжатие и сдвиги. Заметим, что любое аффинное преобразование в $\mathbb{R}^2$ кодируется всего шестью коэффициентами, поэтому СИФ реализует фрактальное сжатие изображения ковра. Множество

$T(B) : f_1(B) \cup f_2(B) \cup f_3(B)$ представляет собой композицию (коллаж) трех уменьшенных копий B. Нахождение коэффициентов аффинных преобразований по фрагменту коллажа называют обратной задачей в теории фракталов. В ее решении важную роль играет следующая

Теорема о коллаже. Пусть $\{(X, d); f_1, f_2, \ldots, f_k\}$ — гиперболические СИФ с коэффициентом сжатия s, B — произвольное непустое компактное множество, принадлежащее $H(X)$ и A — аттрактор СИФ. Пусть

$$h(B, \bigcup_{i=1}^{k} f_i(B)) \equiv h(B, T(B)) < \varepsilon.$$

Тогда

$$h(A, B) \leqslant h(B, T(B)) \frac{\varepsilon}{1 - s}.$$

Таким образом, для решения обратной задачи следует выбирать B, «похожее» на фрагмент коллажа.

Системой *случайных итеративных функций* (ССИФ)

$$\{X; f_1, f_2, \ldots, f_k; p_1, p_2, \ldots, p_k.\}$$

называют СИФ, снабженную набором вероятностей $\{p_i | i = 1, 2, \ldots, k\}$ для каждого f_i, где $p_i > 0$ и $p_1 + p_2 + \ldots + p_k = 1$. Пусть $\sigma = (\sigma_1, \ldots, \sigma_N)$; $\sigma_n \in \{1, \ldots, k\}$. Тогда орбита ССИФ определяется как $x_{n+1} = f_{\sigma_n}(x_n)$, где f_{σ_n} выбирается с вероятностью[(1)] p_{σ_n}. Предположим, что преобразования f_i, $i = 1, 2, 3$ при построении ковра Серпинского выбираются с фиксированными вероятностями $p_1 > p_2 > p_3$. Тогда, образы $f_1(B)$ при итерациях T будут появляться «в среднем» чаще и, следовательно, соответствующие фрагменты в ковре окажутся «более черными» на каждом масштабе. При этом условие нормировки вероятностей при переходе к второй итерации индуцирует новые вероятности:

$$(p_1) \to (p_1 p_1, p_1 p_2, p_1 p_3)$$
$$(p_2) \to (p_2 p_1, p_2 p_2, p_2 p_3)$$
$$(p_3) \to (p_3 p_1, p_3 p_2, p_3 p_3)$$

Эта цепочка продолжается с ростом номера итерации и мы получим ковер Серпинского в черно-серо-белых тонах, причем интенсивность окраски (мера) приобретает скейлинговые свойства. На другом языке такую картину называют *мультифрактальной мерой*.

Заметим, что кроме ССИФ существуют их разновидности — рандомизированные СИФ. Это обычные итеративные функции, коэффициенты которых выбираются случайно на каждом шаге итерации из некоторого множества. Пример аттрактора такой СИФ приведен на рис. 7.

Примечания

1. Рассмотрим некоторое измеримое множество B, в которое может заходить орбита ССИФ. *Эргодическая теорема Элтона* утверждает, что для каждой последовательности символов $\{\sigma_n\}$ частота, с которой орбита $\{x_n\}$ посещает B, равна мере этого множества $\mu_F(B)$.

$$\lim_{N\to\infty} \frac{\sharp\{x_n \in B:\ 1 \leqslant n \leqslant N\}}{N} = \mu_F(B), \quad B \subset \mathbf{B}(X),$$

где $\sharp$ заменяет слово «количество». Другими словами, последовательность точек орбиты ССИФ $x_n = f_{\sigma_n} \circ f_{\sigma_{n-1}} \circ \ldots \circ f_{\sigma_1}(x_0)$ сходится к глобальному аттрактору в смысле расстояния Хаусдорфа для почти каждой точки $x_0 \in X$ и для любого набора вероятностей.

Путеводитель по литературе. Оригинальная работа Хатчинсона [21] трудна для первого чтения, однако существует ее упрощенный вариант [22]. Наиболее полное изложение СИФ и ССИФ дано в книге Барнсли [2] и в недавно переведенной книге Кроновера [23]. Хорошие введения в теорию детерминированных и случайных СИФ можно найти на *Web*-страничках, см., например, [24]. Основные идеи мультифрактального формализма можно найти в [20, 22].

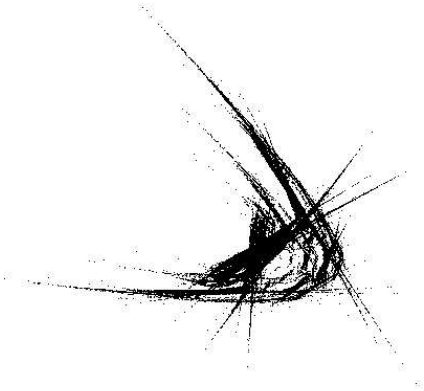

Рис. 7. Случайный Фрактал

Динамические системы и странные аттракторы

Поэтому, если тебе сначала все
покажется пустым бредом, знай,
что причиной твоя слабость!

Николай Кузанский
«Берилл»

Дискретная *динамическая система* задается отображением[1]

$$x_{n+1} = f(x_n), x \in R^n.$$

Пространство, в котором работает отображение, обычно называют *фазовым*. Траектория такой системы получается последовательным применением оператора f:

$$x_1 = f(x_0);\ x_2 = f(x_1) = f(f(x_0)) = f^{\circ 2}(x_0) \ldots$$

Для некоторых систем траектории образуют множества, весьма экзотического вида. Одно из таких множеств приведено на рис. 8. Оно образовано итерациями системы:

$$\begin{cases} x_{n+1} = y_n - sign(x_n)|x_n|^{1/2} \\ y_{n+1} = 0.4 - x_n \end{cases}$$

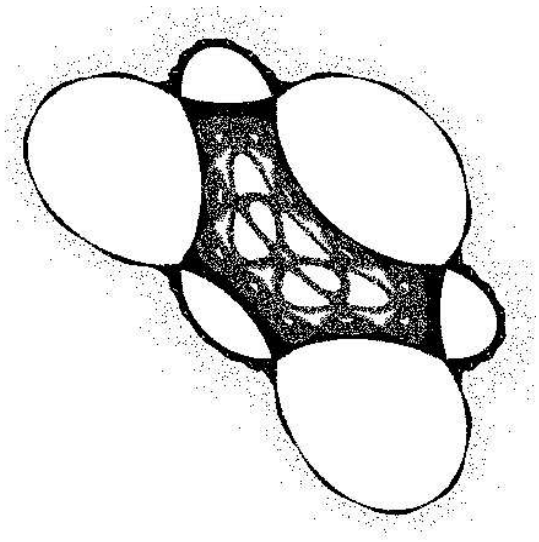

Рис. 8. Фазовый портрет дискретной системы

Как и в случае СИФ, *неподвижные точки* x_f отображения f определяются решением уравнения: $f(x_f) = x_f$. Неподвижные точки могут быть *устойчивыми*, если к ним сходиться последовательность точек $f^{\circ n}(x_0)$, $n \to \infty$, и неустойчивыми, в противном случае. Легко понять, в каких случаях возникает устойчивость. Рассмотрим «возмущенную» неподвижную точку $x = x_f + \delta x$, которую можно получить с помощью малого возмущения отображения f:

$$x_{n+1} = x_f + \delta x_{n+1} = f(x_f + \delta x_n) \simeq f(x_f) + \delta x_n f'(x_f).$$

Поскольку $\delta x_{n+1} = f'(x_f)\delta x_n$, понятно, что $|\delta x_n|$ будет увеличиваться, если $|f'(x_f)| > 1$. Таким образом, производная $|f'|$ играет роль локального коэффициента сжатия или расширения для нелинейного (в общем случае) отображения f. Очевидно, что устойчивой неподвижной точке соответствует сжатие: $|f'(x_f)| < 1$. Для многомерного случая надо вычислять производные по всем координатам, т. е. якобиан $J = det|\frac{\partial x^i_{n+1}}{\partial x^j_n}|$ отображения f в точке $x_f = (x^1_f, x^2_f, \ldots, x^n_f)$. В этом случае условие

устойчивости $J < 1$ соответствует сжатию в J раз объема исходного «параллелепипеда», натянутого на координатные оси, которое происходит под действием отображения f.

Разновидностью неподвижных точек являются *периодические точки* отображения f: говорят, что точка x — *периодическая* с периодом p, если существует такое минимальное число p, что $f^{\circ p}(x) = x$. Все эти объекты *инвариантны* относительно действия отображения или его итераций. Это весьма напоминает фракталы — предельные образы СИФ, которые остаются инвариантными под действием оператора Хатчинсона. Для того, чтобы сделать эту аналогии более точной, введем некоторые дополнительные определения.

Точка y называется *ω-предельной точкой* для точки x, если

$$f^{\circ n}(x) \to y,\ n \to \infty.$$

Иными словами, предельная точка — это «конец» траектории, которая начинается в x. Все ω-предельные точки образуют *ω-предельное* множество. Следующим важным понятием является обобщение идеи инвариантности неподвижных и периодических точек относительно действия f на множества. А именно, множество $B \in R^n$ называют *инвариантным*, если для всех $n \geqslant 0,\ f^{\circ n}(B) = B$. *Окрестностью* множества B называют открытое множество U, которое содержит B и все его предельные и граничные точки. Можно представить себе это множество как шар без границ, внутри которого находится все B и все точки, которые служат пределами сходящихся последовательностей точек из B. Наконец, расстоянием между точкой x и множеством B называют наименьшее из всех расстояний, когда точка y пробегает по всему B, т. е. $l(x, B) = \inf_{y \in B} \|x - y\|$.

Замкнутое инвариантное множество $A \subseteq X$ называется *притягивающим множеством*, если для него существует окрестность U такая, что $\forall x \in U\ f^{\circ n}(x) \to A$ при $n \to \infty$. Наибольшее U, которое обладает таким свойством, называется *бассейном притяжения* для A. Иногда свойство притяжения формулируют на языке *асимптотической устойчивости.* Множество A называют *устойчивым по Ляпунову* если начальное расстояние $l(x, A) < \delta$ между любой точкой и множеством со временем становится меньше любого заданного числа ε, т. е. $l(f^{\circ n}(x), A) < \varepsilon$. Если, кроме того, $\varepsilon \to 0$ при $n \to \infty$ множество называют *асимптотически устойчивым.* Именно таким и является притягивающее множество.

Наиболее важное понятие — *аттрактор* — основано на выделении минимальной[9] структуры притягивающего множества. *Аттрактором* A называют притягивающее множество, содержащее точку x, для которой $\omega(x) = A$. Иными словами, аттрактор содержит целиком *всю траекторию* $f^{\circ n}(x)$. Иногда говорят, что A содержит всюду плотную траекторию x, понимая под этим, что в окрестности любой точки траектории на аттракторе, можно найти другую точку этой же траектории. Кроме того, если наугад выбранная точка аттрактора оказалась «пустой», не надо отчаиваться: в ее локальной окрестности обязательно найдется точка траектории. Часто на аттракторе можно ввести меру, т. е. некоторое число, аналогичное объему в евклидовой геометрии[10]. Это число в ряде интересных случаев можно оценить как относительную долю времени, которое проводит траектория в некотором компактном подмножестве аттрактора[11]. Тогда динамическая неразложимость аттрактора означает существование на нем инвариантной меры$^{(2)}$, которая не может быть представлена как взвешенное среднее нескольких ненулевых инвариантных мер.

Аттракторы возникают, как правило, для диссипативных динамических систем[12]. В том случае, если фазовым пространством является R^2, такие аттракторы имеют тривиальную структуру: это неподвижные точки и предельные циклы. Однако, в $R^n,\ n \geqslant 3$, диссипативные системы могут иметь экзотические аттракторы с фрактальной структурой.

Рассмотрим в качестве примера так называемый *аттрактор Хенона*. Исходная динамическая диссипативная система описывается нелинейными дифференциальными уравнениями в R^4. Для упрощения исследования эволюцию траекторий многомерных систем часто отображают на так называемое *сечение Пуанкаре* в фазовом пространстве. Этот прием заключается в следующем. Рассмотрим плоскость, ортогональную фазовым траекториям. Каждый момент прохождения орбиты через плоскость будем отмечать парой координат (x_n, y_n) той точки, в которой траектория «протыкает» плоскость. Тогда динамика исходной системы редуцируется

[9] Т.е. неразложимой на более мелкие элементы с теми же свойствами.

[10] Аналогичное лишь по свойствам: евклидов объем фрактального аттрактора равен нулю!

[11] Для *эргодической* меры, эта доля пропорциональна мере самого множества.

[12] Динамическая система называется *консервативной*, если отображение сохраняет ее фазовый объем, и *диссипативной* — в противном случае.

к дискретным преобразованиям $(x_n, y_n) \to (x_{n+1}, y_{n+1})$ точек плоскости. В частности, преобразование Хенона в R^2 задается оператором

$$T:\ x_{i+1} = y_i + 1 - ax_i^2,\ y_{i+1} = bx_i.$$

Этот оператор представляет собой композицию трех преобразований:

1. $T':\ x' = x,\ y' = y + 1 - ax^2$
2. $T'':\ x'' = bx',\ y'' = y'$
3. $T''':\ x''' = y'',\ y''' = x''$

Первое преобразование складывает фигуру и сохраняет площадь, второе — сжимает фигуру относительно оси x и уменьшает площадь, умножая ее на постоянный множитель $b < 1$, третье — поворачивает и сохраняет площадь, но меняет ее знак. Якобиан отображения равен

$$\frac{\partial(x_{i+1}, y_{i+1})}{\partial(x_i, y_i)} = -b.$$

Значение $|b| < 1$ соответствует сжатию. На рис. 9 приведен аттрактор преобразования Хенона для значений параметров $a = 1.4;\ \ b = 0.3$. Известно, что бокс-размерность этого аттрактора равна 1.26, так что это — фрактал.

Фрактальные аттракторы связаны с так называемыми *сценариями динамического хаоса*(2). Хаотические системы описываются полностью детерминированными уравнениями, однако прогноз их решений не может быть продолжен дальше некоторого ограниченного интервала времени. Это эффект так называемой *существенной зависимости* от начальных условий, которые всегда задаются с конечной точностью. Рассмотрим простой пример динамической системы[13], заданной на интервале $I = [0, 1]$ отображением: $x_{n+1} = 2x_n,\ \mod 1$. Динамика системы сводится просто к удвоению координаты точки и отбрасыванию единицы (это и обозначается как $\mod 1$), если полученное значение координаты становиться больше, чем размеры «фазового» пространства. Выберем начальную точку из интервала I и запишем ее координату (адрес) в виде двоичного кода. Это всегда можно сделать единственным образом, если

[13]Ее называют иногда *сдвигом Бернулли*.

заранее оговорить выбор способа записи для чисел, допускающих два варианта адреса. Итак, пусть например, $x_0 = 0,101101$, где адрес записан с точностью до шестого разряда. Последнее означает, что мы не знаем, какой из двух символов 0 или 1 стоит в следующем разряде. Заметим, что каждый разряд уточняет адрес:

$$\begin{aligned} 0,1 &\Rightarrow 1/2 < x_0 < 1 \\ 0,10 &\Rightarrow 1/2 < x_0 < 3/4 \\ 0,101 &\Rightarrow 5/8 < x_0 < 3/4 \ \cdots \end{aligned}$$

Вспомним теперь, что умножение в двоичной арифметике — это просто сдвиг запятой вправо и отбрасывание единицы, с учетом модуля. Таким образом, эволюция нашей системы сводится просто к уточнению адреса начальной точки! Однако, через 6 итераций последний известный символ уйдет влево, исчерпав всю информацию, которая содержалась в начальных данных! Следовательно, на седьмой итерации мы не сможем сказать, в какую половину интервала I попадет очередная точка траектории. Понятно, что увеличение начальной точности не изменит ситуацию — увеличится лишь число «детерминированных» шагов. Легко понять причины потери точности: общее решение имеет вид: $x_n = 2^n x^0$, так что расстояние между сколь угодно близкими начальными точка-

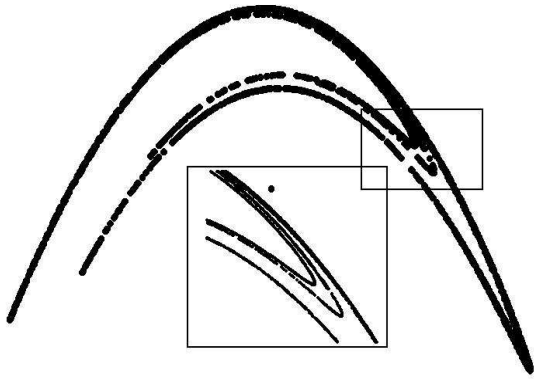

Рис. 9. Аттрактор Хенона

ми будет удваиваться при каждой итерации, пока не достигнет размеров всего интервала. Этот эффект называют *разбеганием близких траекторий*, существенной зависимостью от начальных данных или баттерфляй-эффектом [14]. Впервые этот эффект был обнаружен *Эдвардом Лоренцем* в 1961 году при численном решении простой системы из трех нелинейных уравнений, моделирующих конвекционные потоки в атмосфере. Уравнения были полностью детерминированными, т. е. они не содержали каких-либо случайных величин. Траектории системы заполняли фрактальное подмножество — аттрактор с хаусдорфовой размерностью 2.06. При этом, две численные траектории, стартующие из двух близких начальных точек очень быстро расходились, демонстрируя совершенно различное поведение. Поскольку начальные данные для метеосистем всегда получаются с конечной точностью, Лоренцу стало ясно что прогноз погоды будет всегда ограничен тем диапазоном времени, на котором расхождение близких траекторий еще не превышает некоторой заданной величины. Такой временной масштаб — его называют *горизонтом прогноза*, по-видимому, не превышает недели$^{(4)}$. Существенно, что ограниченная предсказуемость является следствием нелинейности уравнений. В более общем случае оказалось, что динамический сценарий, которому следуют траектории хаотических систем определяется *типом* нелинейности, и почти не зависит от вида самого уравнения! Однако, обсуждение этих вопросов выходит за рамки данной лекции. В заключение еще раз заметим, что странные аттракторы хаотических динамических систем по своей сути являются синонимами предельных образов (аттракторов или фракталов) СИФ. Фракталы в геометрии — результат работы системы линейных сжимающих отображений (СИФ); в динамике, сжатие определяется нелинейными диссипативными уравнениями.

Примечания

1. В физике под динамической системой понимают обычно систему (автономных) обыкновенных дифференциальных уравнений $x' = f(x),\ x \in R^n$, где $x' \equiv dx/dt$, и $f : U \to R^n$ определено для некоторого открытого подмножества $U \in R^n$. Пространство R^n называют *фазовым пространством*, а $f(x)$ — *векторным полем*, так как ре-

[14] Из названия статьи *Э. Лоренца* «Предсказуемость: может ли взмах крылышек бабочки в Бразилии привести к образованию торнадо в Техасе?»

шение уравнения есть кривая с касательным вектором x'. В более общем случае под динамической системой в фазовом пространстве M понимают *однопараметрическую группу* преобразований $g^t : M \to M$, где параметром обычно служит время: непрерывное $(t \in R)$ или дискретное $(t \in Z)$. Точка $x_0 \in M$ под действием g^t переходит в точку $x_t = g^t(x_0)$, при этом групповое свойство означает, что $g^t g^s = g^{t+s}$. Последовательность точек $\{x_t\}$ образует *орбиту* группы. Если M — дифференцируемое многообразие, и g^t — взаимно однозначное дифференцируемое отображение, такое, что обратное к нему g^{-t} обладает аналогичными свойствами (такое отображение называют *диффеоморфизмом*), то пару $(M, \{g^t\})$ называют *фазовым потоком* или же *каскадом* — для случая, когда время дискретно. Легко показать, что «физическое» определение сводится к абстрактному. Рассмотрим динамическую систему заданную уравнением $x' = ax,\ x \in R$. Его решение можно записать в виде $x_t = g^t x_0 = e^{at} x_0$, так что поток индуцируется действием непрерывной группы $g^t = e^{at}$ однопараметрических преобразований в R.

2. Для того, чтобы ввести меру на траекториях динамической системы, заданной парой X, f, где X — топологическое пространство, выделяют некоторые интересные подмножества $B \subset X$, которые образованы обычно неподвижными точками или периодическими орбитами. Такие «элементарные» подмножества (их называют *борелевыми*) образуют «строительные блоки» для введения меры: она определяется на них, а затем обобщается на более сложные подмножества, которые можно построить из объединения борелевых «кирпичей». Например, для интервалов на прямой мера — это просто длина интервала. На плоскости, произведение интервалов дает прямоугольники, мера на которых — площадь. Объединение и пересечение прямоугольников порождает произвольные многоугольники. Их мера вычисляется по площадям «борелевых» прямоугольников, с учетом условия аддитивности. Говорят, что преобразование $f : X \to X$ *сохраняет меру* μ, если $\mu(f^{-1}(B)) = \mu(B)$. Такая запись позволяет рассматривать не только взаимно однозначные преобразования. Наиболее интересным утверждением о преобразованиях, сохраняющих меру на компактах, является *теорема Пуанкаре*

о возвращении. Пусть $B \in X$, μ — измеримо, X — компакт и обратимое преобразование f сохраняет меру на $X : \mu\{f(B)\} = \mu\{B\}$. Тогда существует некоторое n (возможно, зависящее от x_0), такое, что почти все точки $x_0 \in B$ возвращаются в B, т. е. $f^{\circ n}(x_0) \in B$. Иными словами, орбита любой точки $p \in B : \{p, f(p), f^{\circ 2}(p) \ldots\}$ содержит некоторые образы $f^{\circ k}(p)$, которые вновь попадают в B. Предположим, что это неверно и часть точек, образующих подмножество $A \subset B$, *никогда не возвращаются*. Тогда множества $\{A, f(A), f^{\circ 2}(A), \ldots\}$ *никогда не пересекаются!* Действительно, пусть точка $P \in f(A) \cap f^{\circ 3}(A)$. Очевидно, что $f^{-1}(P) \in A$, но она же принадлежит и $f^{\circ 2}(A)$, так как $P \in f^{\circ 3}(A)$. Но это невозможно, поскольку A не содержит возвращающихся точек. Следовательно, орбита $f^{circn}(A)$ разбивает X на бесконечную последовательность непересекающихся фрагментов, которые могут иметь только нулевую меру: X-компакт и его мера конечна.

3. Динамическая система (X, g) называется *хаотической*, если выполняется следующие условия:

   - пусть $x \in X$ и U — открытое подмножество, содержащее X. Если для некоторого $\delta > 0$ существует такое n и такая точка $y \in U$, что $d(g^n(x), g^n(y)) > \delta$, то g обладает *существенной зависимостью* от начальных условий;
   - g — *транзитивно*, т. е. для любой пары открытых множеств U, V существует такое n, что $g^n(U) \cap g^n(V) \neq \varnothing$;
   - периодические точки g плотны в X, т. е. в любой окрестности любой точки в X существует по крайней мере одна периодическая точка (и, следовательно, бесконечно много периодических точек).

4. Любопытно, что за 175 лет до Лоренца, ученик Эйлера, академик С. Я. Румовский писал в статье «Рассуждения о предсказании погод»: «Что же думать о предсказаниях погоды на целый год? Они не что иное суть, как тщетное и пустое умствование, которым легкомысленным людям во многих случаях вред нанести может». Статья была опубликована в № 1 Докладов Академии наук за 1786 год. Они назывались тогда «Новые ежемесячные сочинения к пользе и увеселению служащие».

Путеводитель по литературе. Связь фракталов и хаотических систем великолепно описана в книге Кроновера [23]. На сайтах [24] можно найти хорошие примеры дискретных преобразований на плоскости. Доступное изложение основ теории нелинейных динамических систем содержится в монографиях [25, 26] и журнальных обзорах [27, 28]. Строгому обсуждение концепции аттрактора посвящена статья Милнора [29]. Хорошие вводные курсы по топологической динамики можно найти на сайтах [30, 31], а современное изложение эргодической теории динамических систем на *Web*-страничке [31].

Нейронные сети, СИФ и гипернейрон

— Что это? Никак игрушка!
— Подберите фалды! . . .
— Смотрите издали! . . .

Козьма Прутков

Как известно, основной элемент модели нейронной сети — формальный нейрон — вычисляет скалярное произведение входного вектора

$$\mathbf{x} = (x_0, x_1, \ldots, x_n)^T$$

и вектора синаптических весов

$$\mathbf{w} = (\omega_0, \omega_1, \ldots, \omega_n)^T,$$

которое преобразуется затем функцией активации g, вычисляющей выход нейрона y:

$$y(\mathbf{x}, \mathbf{w}) = g(\mathbf{x} \cdot \mathbf{w})$$

В качестве g обычно используется либо функция Хевисайда (ступенчатая), либо сигмоида.

Пусть гиперболическая СИФ задана системой $\{(X, d); f_1, f_2, \ldots, f_q\}$. Пусть A_F — аттрактор СИФ. Существует два способа приближенного построения этого фрактала. Первый из них начинается с выбора одной или нескольких начальных точек x_0 и вычисления их дискретных орбит длиной N. Пусть, как и раньше, $\sigma = (\sigma_1, \ldots, \sigma_N);\ \sigma_n \in \{1, \ldots, q\}$. Тогда

при достаточно большом N множество:

$$A(x_0, N) = \bigcup_{\sigma} f_{\sigma_N} \dots f_{\sigma_1}(x_0)$$

является хорошим приближением A_F. Этот способ можно оптимизировать, если выбрать в качестве x_0 неподвижную точку одного из f_k так, что $x_0 \in A_F$. Заметим, что вычисление N орбит на один шаг вперед не эквивалентно вычислению одной орбиты на N шагов. Другой способ заключается в определении траектории начального компактного множества $A_0 \in H(X, d)$:

$$A_{n+1} = F(A_n) \equiv \bigcup_{i}^{N} f_i(A_n);\ n = 0, \dots, \infty.$$

Последовательность A_n сходится к A_F в метрике Хаусдорфа, так что при достаточно больших N, $F^{\circ N}(A_0) \approx A_F$. Рис. 10 иллюстрирует этот подход. Коллаж состоит из трех начальных множеств, которые эволюционируют под действием F и приводят к аппроксимации фрактала, показанной на рис. 11.

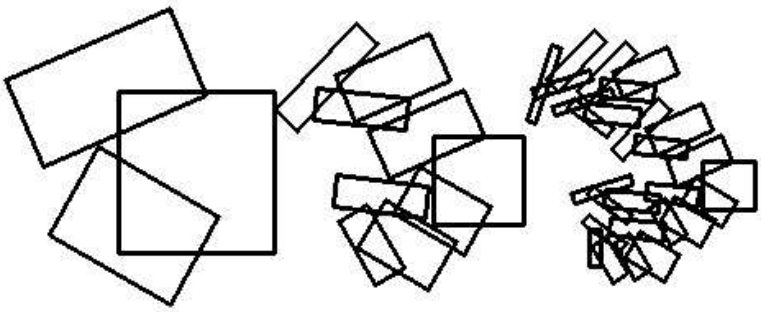

Рис. 10. Первые три итерации коллажа

Предположим, что мы собираемся построить на экране компьютера аттрактор СИФ. Будем считать, что экран состоит из квадратных пикселов и горизонтальное и вертикальное разрешения равны.

Пусть $X \equiv S^u \subset \mathbb{R}^2$, S^u — единичный квадрат с разбиением, соответствующим выбранному разрешению. Чтобы упростить обозначения,

Рис. 11. Аппроксимация фрактала, порожденного эволюцией коллажа

мы будем использовать значок s вместо координат пиксела x_{ij}. Снабдим каждое A_n индикаторной функцией (нейроном) $y : S^u \to \{0,1\}$ такой, что для каждого пиксела $s \in S^u$:

$$y_s(n) = \begin{cases} 1\,, & s \in A_n \\ 0\,, & s \notin A_n \end{cases}$$

Определим «веса»

$$\omega_{ss'} = \begin{cases} 1\,, & \text{если } \ f_i(s') = s \ \text{ для некоторого } i \\ 0\,, & \text{в противном случае} \end{cases}$$

и зададим динамику $y_s(n)$ как

$$y_s(n+1) = g\Big(\sum_{s' \in S} \omega_{ss'} y_{s'}(n)\Big),$$

где g — ступенчатая функция.

Очевидно, что по виду и смыслу полученное выражение определяет бинарную нейронную сеть с $|S^u|$ нейронами y_s и синаптическими весами $\omega_{ss'}$, которая реализует СИФ. Заметим, что каждый нейрон такой сети имеет только бинарные веса и максимальное количество нейронов, с которыми он связан, равно q. Несмотря на то, что число весов может быть очень большим, подавляющее большинство из них будут нулями. Так что результирующая сеть будет пространственно разреженной.

Детерминированная СИФ и эквивалентная ей нейронная сеть позволяют получать лишь черно-белые изображения аттракторов. Для генерации полутонов следует использовать ССИФ. Поскольку принципы построения соответствующей нейросети остаются прежними, мы опускаем здесь этот вариант.

Дальнейшее обобщение этих результатов приводит к рассмотрению случайной итеративной нейронной сети, которая в определенном смысле эквивалентна дискретной динамической системе. Представим себе формальный нейрон, который устроен следующим образом. На каждый его вход подается вектор $\mathbf{u}_i \in R^m$. Весами служат множество $\mathbf{W}$ векторов $\mathbf{w}_j \in R^k$, функцией активации $T_{\sigma,\mathbf{W}}$ является ССИФ, а выходом вектор $\mathbf{y} \in R^m$, так что $\mathbf{y} = T_{\sigma,\mathbf{W}}(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k)$, где множество σ нумерует выбор отображений, входящих в T с соответствующими вероятностями, так же как и в разделе «Фракталы и системы итеративных функций» (см. с. 139–144). Такое устройство называют вероятностным *гипернейроном*. *Случайная итеративная нейронная сеть* (СИНС) состоит из трех гипернейронов и имеет архитектуру обычной рекуррентной нейронной сети, которая реализует аттрактор ССИФ. Динамику СИНС можно описать парой (X, A_f), где X — пространство состояний СИНС и A_f ее глобальный аттрактор [15]. Пусть (Y, A_g) — дискретная динамическая система с аттрактором A_g. В начале 2000 года Фернандо Ниньо (The University of Memphis) показал в своей диссертации, что для любой заданной пары (Y, A_g) существует (X, A_f) такая, что хаусдорфово расстояние $h(A_f, A_g)$ между аттракторами динамической системы и СИНС можно сделать сколь угодно малым. Иными словами, СИНС является аппроксиматором дискретной динамической системы с заданным аттрактором! Необходимый для аппроксимации набор ССИФ можно най-

[15] Такой аттрактор всегда существует в силу упомянутой в разделе «Фракталы и системы итеративных функций» теоремы Элтона.

ти генетическим алгоритмом. К сожалению, подробное обсуждение этой интересной проблемы выходит за рамки лекции.

Путеводитель по литературе. Автору известны всего три работы [37–39], специально посвященные связи СИФ и нейронных сетей. Описание гипернейрона можно найти в диссертации Ф. Ниньо [40].

Глоссарий

> «Запятирующее тире» — отражает кумулятивное диффундирование смыслокачества в новизну подтекстовых проводок в ситуации резкого спада количества определяющих указателей.
>
> *Павел Таранов*
> *«Маневры общения»*

Этот раздел содержит предварительные сведения, т. е. определения и понятия, чуть более содержательные, чем просто определения. Они собраны из разных областей математики с целью сделать чтение ***Лекций*** по возможности независимым от сопутствующей литературы. Мы будем использовать ниже следующую символику:

- Запись $x \in \mathfrak{X}$ означает, что элемент x принадлежит множеству $\mathfrak{X}$.
- Запись $\forall\, x$ читается: для любых x (для всех x), запись $\exists\, y$ — существует y, а запись $\forall x \,\exists\, y \,|$ означает: для любого x существует y, такой что...
- Запись $\mathfrak{X} = \{x|A\}$ обозначает множество таких элементов x, для которых справедливо условие A.
- Запись $A \subset B$ ($A \subseteq B$) означает, что A содержится в B (содержится в B или совпадает с ним, соответственно).
- Наконец, $A \Rightarrow B$ читается: A влечет B.

Множество возникает благодаря объединению отдельных вещей в одно целое. Для того, чтобы избежать парадоксов, полагают, что элементы $a, b, c, \ldots$ особым *не подлежащим определению образом* порождают множество $\mathfrak{X}$. Иными словами $\mathfrak{X}$ задано, если известно, входит туда элемент или нет[(1)]. Природа самих элементов безразлична: ими могут быть, например, Луна и прошлогодний снег.

Объединением множеств $\mathfrak{A}$ и $\mathfrak{B}$ называют множество $\mathfrak{C} = \mathfrak{A} \cup \mathfrak{B}$, состоящее из элементов $c \in \mathfrak{C}$, каждый из которых принадлежит $\mathfrak{A}$ или (и) $\mathfrak{B}$.

Пересечением $\mathfrak{A}$ и $\mathfrak{B}$ называют множество $\mathfrak{C} = \mathfrak{A} \cap \mathfrak{B}$, каждый элемент которого c входит и в $\mathfrak{A}$, и в $\mathfrak{B}$. Эти определения легко обобщаются. Для произвольного множества индексов $\mathfrak{I} = \{i\}$ каждому i поставим в соответствие множество $\mathfrak{A}_i$. Тогда объединение

$$\mathfrak{C} = \cup\mathfrak{A}_i = \{c \in \mathfrak{C} | \exists i \in \mathfrak{I} \Rightarrow c \in \mathfrak{A}_i\}.$$

Аналогично определяется пересечение:

$$\mathfrak{C} = \cap\mathfrak{A}_i = \{c \in \mathfrak{C} | \forall i \in \mathfrak{I} \Rightarrow c \in \mathfrak{A}_i\}.$$

Пустое множество обозначается $\varnothing$.

Семейство множеств $\mathfrak{A} \equiv \{A_i\}$ называют ***покрытием*** множества $\mathfrak{B}$, если $\mathfrak{B} \subseteq \cup\{A_i | A_i \in \mathfrak{A}\}$, т. е. каждая точка $\mathfrak{B}$ принадлежит некоторому элементу $\mathfrak{A}$. Частный случай покрытия, когда $A_i \cap A_j = \varnothing$, $\forall i, j$ называют ***разбиением*** $\mathfrak{B}$.

Два множества $\mathfrak{A}$ и $\mathfrak{B}$ можно поставить во взаимно однозначное соответствие, если $\forall a \in \mathfrak{A}$ $\exists b \in \mathfrak{B}$ и наоборот. Так определяется ***отношение эквивалентности*** $\mathfrak{A} \equiv \mathfrak{B}$; говорят еще, что $\mathfrak{A}$ и $\mathfrak{B}$ имеют одинаковую мощность. Если $\mathfrak{A} \equiv \{1, 2, \ldots\}$ то $\mathfrak{A}$ называют ***счетно-бесконечным***.

Для описания мощности множества служат ***кардинальные числа***. Если множество конечно, это просто число элементов в нем. В случае бесконечного множества $\mathfrak{A}$, кардинальное число определяют присвоением имен кардинальным числам конкретных множеств, с которыми можно сравнить $\mathfrak{A}$. Кардинальное число счетно-бесконечного множества обозначают символом *алеф-нуль* — $\aleph_0$. Примером ***несчетного*** множества являются вещественные числа интервала $[0, 1]$. Его кардинальное число $|c|$ называют ***мощностью континуума***. Ясно, что мощность всей вещественной оси тоже $|c|$; в этом легко убедиться с помощью преобразования:

$x \to 1/2(1 + \operatorname{th} x)$. Кантор показал, что точки квадрата и точки отрезка эквивалентны, т. е. мощность $\mathfrak{R}^2$ равна $|c|$.

Соответствие между различными множествами удобно описывать функциями или отображениями: $f(*)$ или $f : \mathfrak{A} \to \mathfrak{Y}$. Если $A \in \mathfrak{X}$, то и $f[A] = \{f(x)|x \in A\}$. Аналогично, если $B \in \mathfrak{Y}$, то $f^{-1}[B] = \{x|f(x) \in B\}$ — подмножество в $\mathfrak{X}$. Подмножество $\{f(x)\} \in \mathfrak{Y}$ — это ***область значений*** функции, а $\{x\} \in \mathfrak{X}$ — ее ***область определения***. Функциональные отношения между множествами имеют определенную классификацию:

- Если образ всего множества $\mathfrak{A}$ совпадает с $\mathfrak{B}$ или, что тоже самое, каждый элемент из $\mathfrak{B}$ является образом по крайней мере одного элемента из $\mathfrak{A}$, то говорят, что f — ***сюръекция*** или отображение **на**.

- Функцию $f : \mathfrak{A} \to \mathfrak{B}$ называют ***инъективной*** или взаимно-однозначной если для $\forall y \in \mathfrak{B}$ существует не более одного элемента $x \in \mathfrak{A}$ такого, что $y = f(x)$. Очевидно, что в этом случае $f(x) = f(y) \Rightarrow x = y$.

- ***Биективное*** отображение или ***биекция*** — это функция, которая является одновременно сюръективной и инъективной.

Если $\mathfrak{A}$ и $\mathfrak{B}$ совпадают, то $f : \mathfrak{A} \to \mathfrak{A}$ и элемент x, удовлетворяющий условию $x = f(x)$, называется ***неподвижной точкой*** отображения f. Например, каждое непрерывное отображение $f : [a, b] \to [a, b]$ отрезка в себя имеет хотя бы одну неподвижную точку[(2)].

Абстрактное множество $\mathfrak{G}$ называют ***группой***, если:

1. Для любой пары его элементов g_1, g_2 определено произведение или бинарная операция $g_1 * g_2 \in \mathfrak{G}$, ассоциативная, но не обязательно коммутативная.

2. Существует единица группы $e : g * e = e * g = g$, $\forall g \in \mathfrak{G}$.

3. Для любого $g \in \mathfrak{G}$ существует $g^{-1} : g * g^{-1} = e$.

Рассмотрим группу преобразований плоскости $(x, y) \to (x^*, y^*)$, заданную в матричном виде: $\mathbf{x}^* = \mathbf{Ax+B}$. Если $\mathbf{A}$ — ортогональная матрица (т.е. $\det \mathbf{A} = 1$), то обычную геометрию можно определить как совокупность свойств, инвариантных относительно действия нашей группы. Координаты в $\mathfrak{R}^n$, например, определены лишь с точностью до ортогональ-

ных преобразований. С другой стороны, в ***арифметическом*** пространстве X^n точка x определяется *абсолютно* набором n чисел $(x^1, \ldots, x^n)$ с точностью до *тождественных* преобразований.

Аффинная геометрия удовлетворяет более скромному требованию $\det \mathbf{A} \neq 0$. Можно задать группу в общем виде: $x^* = f(x, y)$; $y^* = g(x, y)$, где f и g — нелинейные C^N–функции, т. е. функции, обладающие непрерывными производными до N-го порядка. Если такими же свойствами обладают обратные преобразования (а их существование гарантирует группа), то мы получим геометрию ***гладких многообразий***. Такое многообразие определяет объект с точностью до произвольных гладких деформаций. Вырожденный случай C^0 — это уже топология.

Метрическое пространство $(\mathbf{M}, d)$ — это множество $\mathbf{M}$ вместе с вещественнозначной функцией d, определенной на декартовом произведении $\mathbf{M} \times \mathbf{M}$, называемой метрикой на $\mathbf{M}$ и удовлетворяющей следующим требованиям:

- $d(x, y) \geqslant 0;\ d(x, y) = 0 \Rightarrow x = y$
- $d(x, y) = d(y, x)$
- $d(x, z) \leqslant d(x, y) + d(y, z)$

Для $\mathbf{M} \equiv \Re^n$ обычная евклидова метрика:

$$d(x, y) = \Big[\sum_i (x_i - y_i)^2\Big]^{1/2}$$

легко обобщается до $\boldsymbol{L}_\mathbf{p}$ метрики Минковского:

$$d_p(x, y) = \Big[\sum_i (x_i - y_i)^p\Big]^{1/p}$$

Располагая метрикой, нетрудно ввести понятие *сходимости* по аналогии с тем, что было в X^n. Говорят, что последовательность точек $\{x_n\}$, $n = 1, 2, \ldots$, метрического пространства $(\mathbf{M}, d)$ сходится к $x \in \mathbf{M}$, если $d(x, x_n) \to 0$, при $n \to \infty$. Различные метрики порождают разные сходимости.

Метрика d_1 эквивалентна d_2 в $\mathbf{M}$, если существуют две постоянные $0 < c_1 < c_2 < \infty$ такие, что:

$$c_1 d_1(x,y) \leqslant d_2(x,y) \leqslant c_2 d_1(x,y),\ \forall (x,y) \in \mathbf{M} \times \mathbf{M}$$

Последовательность $\{x_n\} \in (\mathbf{M}, d)$ называют ***последовательностью Коши***, если $\forall \varepsilon > 0\ \exists N \,|\, n,m > N \Rightarrow d(x_n, x_m) < \varepsilon$. Заметим, что эта последовательность не обязательно сходится к определенному пределу. Пусть $(\mathbf{M}, d)$ — метрическое пространство и $\{x_n\}$ — последовательность Коши, сходящаяся к точке $x \in \mathbf{M}$. Пусть также $f : \mathbf{M} \to \mathbf{M}$ — непрерывная функция. Тогда:

$$\lim_{n\to\infty} f(x_n) = f(x).$$

Пространство $(\mathbf{M}, d)$, в котором любая последовательность Коши сходится, называется ***полным***.

Пусть $(\mathbf{M}, d)$ — метрическое пространство и $r > 0$ — действительное число. Тогда ***открытым*** шаром в точке $a \in \mathbf{M}$ называется подмножество $B(a,r) \subset \mathbf{M}:\ B(a,r) = \{x \in \mathbf{M} \,|\, d(a,r) < r\}$. Шар называется ***замкнутым***, если $d(a,r) \leqslant r$.

Открытым множеством в $(\mathbf{M}, d)$ называется подмножество $A \subset \mathbf{M}$, для любой точки $x \in A$ которого существует такое $r > 0,\ B(x,r) \subset A$. Пустое множество $\varnothing$ и само пространство $\mathbf{M}$ — открыты.

Замкнутым множеством в $\mathbf{M}$ называют дополнение открытого множества; $\varnothing$ и $\mathbf{M}$ объявляются замкнутыми. Метрическое пространство $\mathbf{M}$ называют ***компактным***, если оно удовлетворяет *аксиоме Лебега—Бореля*: из любого покрытия $\mathbf{M}$ открытыми множествами можно выделить *конечное* подпокрытие.

$(\mathbf{M}, d)$ называют ***вполне ограниченным***, если $\forall \varepsilon > 0$ существует конечное покрытие пространства $\mathbf{M}$ шарами диаметром $< \varepsilon$.

Понятия компактности и ограниченности заменяют понятие *конечности* в чистой теории множеств. Компактным (вполне ограниченным) множеством в $(\mathbf{M}, d)$ называют такое подмножество $A \subset \mathbf{M}$, для которого подпространство A компактно (вполне ограничено). Наконец, для $(\mathbf{M}, d)$ следующие три условия эквивалентны: $\mathbf{M}$ — компактно, или любая бесконечная последовательность в $\mathbf{M}$ имеет по крайней мере одну предельную точку, или $\mathbf{M}$ — полное и вполне ограниченное.

Пусть $H(\mathbf{M})$ — пространство, «точками» которого являются компактные множества из $(\mathbf{M}, d)$. Пусть $A, B \in H$. Определим метрику h в H как

$$h(A,B) = \max\{\max_{x\in A}\min_{y\in B} d(x,y),\ \max_{y\in A}\min_{x\in B} d(x,y)\}$$

Величина $h(A,B)$ называется ***расстоянием Хаусдорфа*** и удовлетворяет всем аксиомам метрики. Можно показать, что $(H(\mathbf{M}), h)$ является полным метрическим пространством. В тексте Лекции дано эквивалентное определение этой метрики.

Пусть $(\mathbf{M}, d)$ — метрическое пространство и $s > 0$. Отображение $f : \mathbf{M} \to \mathbf{M}$ называют ***подобием*** с масштабным коэффициентом s, если $d(f(x), f(y)) = sd(x,y)$, $\forall x, y \in \mathbf{M}$ и ***сжимающим*** отображением с коэффициентом сжатия s, если существует такое число $0 \leqslant s < 1$, что $d(f(x), f(y)) < sd(x,y)$, $\forall x, y \in \mathbf{M}$. Иными словами, f — сжатие, если расстояние между образами двух произвольных точек в s раз меньше исходного. Определим для f множество итераций вперед $f^{\circ n} : \mathbf{M} \to \mathbf{M}$, $n = 1, 2, \ldots$ соотношениями:

$$f^{\circ 0}(x) = x;\ f^{\circ 1}(x) = f(x);\ f^{\circ n+1}(x) = f * f^{\circ n}(x) = f(f^{\circ n}(x)).$$

Если f — обратима, аналогично определяются итерации назад:

$$f^{-\circ n}(x) = (f^{\circ n})^{-1}(x).$$

Пусть $f : \mathbf{M} \to \mathbf{M}$ — сжатие в полном метрическом пространстве $(\mathbf{M}, d)$. Тогда справедлива ***Теорема о сжимающем отображении***: f имеет единственную[(3)] неподвижную точку $f(x) = x_f \in \mathbf{M}$ и $\forall y \in \mathbf{M}$ последовательность $f^{\circ n}(y)$ сходится к

$$x_f :\ \lim_{n\to\infty} f^{\circ n}(y) = x_f.$$

Множество $B \subset \mathbf{M}$ называют ***плотным*** в $\mathbf{M}$, если любое $x \in \mathbf{M}$ является пределом последовательности элементов из B.

Пусть X — множество точек x. Систему $\mathfrak{B}$ подмножеств X называют **σ-алгеброй**, если

1. $\varnothing,\ X \in \mathfrak{B}$
2. $A \in \mathfrak{B} \Rightarrow X - A \in \mathfrak{B}$

3. $A_n \in \mathfrak{B},\ n = 1, 2, \ldots \Rightarrow \bigcup A_n \in \mathfrak{B}, \bigcap A_n \in \mathfrak{B}$

Наименьшая из σ-алгебр в метрическом пространстве $\mathbb{R}^n$ называется ***борелевой σ-алгеброй.***

Пара $(X, \mathfrak{B})$ называется измеримым пространством. Пусть $(X, \mathfrak{B})$ — измеримое пространство. Вещественная функция $\mu = \mu(A)$, $A \in \mathfrak{B}$, принимающая значения из интервала $[0, \infty]$, называется **мерой**, если

- $\mu(\varnothing) = 0$
- $\mu(A) \geqslant 0,\ \forall A \in \mathfrak{B}$
- если $\{A_n\}_{n=1}^{\infty}$ — непересекающиеся множества из $\mathfrak{B}$, то $\mu(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$

Тройка $(X, \mathfrak{B}, \mu)$ называется **пространством с мерой**. Например, *лебегова мера* на вещественной оси — это просто длина интервала вида (a, b), $[a, b)$, $(a, b]$, $[a, b]$. Мера Лебега не является универсальной: существуют множества, не являющиеся μ-измеримыми. Тем не менее, μ-мера является эталоном не только во многих разделах математики, но и в эргодической теории гладких динамических систем.

Множество $\mathfrak{S}$ ***имеет меру нуль***, если оно может быть покрыто системой интервалов, сумма длин которых произвольно мала. Иными словами, такое $\mathfrak{S}$ можно вложить в борелево множество $\mathfrak{B}$ мерой меньшей чем $\varepsilon, \forall \varepsilon > 0$. Если $\mathfrak{S}$ можно получить из борелева множества $\mathfrak{B}$ добавлением или выбрасыванием множества точек нулевой меры, то $\mu(\mathfrak{S}) = \mu(\mathfrak{B})$. Такие множества называют μ-измеримыми. Выражение «почти для всех x» понимают в смысле *для всех x, за исключением множества меры нуль*.

Если $\mu(X) = 1,\ \mu$ называют вероятностной мерой, а $(X, \mathfrak{B}, \mu)$ — **вероятностным пространством**.

Пусть $(X, \mathfrak{B}, \mu)$ — вероятностное пространство. Преобразование $T : X \to X$ называется **измеримым**, если для любого $A \in \mathfrak{B}$, $T^{-1}A = \{x : Tx \in A\} \in \mathfrak{B}$.

Говорят, что измеримое преобразование T **сохраняет меру** μ, если для каждого $A \in \mathfrak{B}$, $\mu(T^{-1}A) = \mu(A)$. Меру μ называют в этом случае **T-инвариантной**.

Четверка $(X, \mathfrak{B}, \mu, T)$, где T — преобразование, сохраняющее меру, называется **динамической системой**.

Примечания

1. При этом абсолютно неясен способ, посредством которого мы могли бы выяснить, входит элемент a в A или нет. Например, если A задано запахом, а B — вкусом, можем ли мы решить эквивалентны они или нет? Приведем пример одного парадокса Рассела. Пусть A множество всех натуральных чисел. Предположим, что каждое из этих чисел можно определить фразой, содержащей менее 20 русских слов. Содержит ли A весь натуральный ряд? Считая, что русский язык содержит не более чем n слов и, следовательно, не более чем n^{20} нужных нам фраз, можно полагать, что A конечно. Но тогда определим *наименьшее натуральное число, не входящее в множество* A. Это число по определению не входит в A, но должно входить в него, поскольку определяется фразой, не превышающей 20 слов. Очевидно, A содержит весь натуральный ряд, а парадокс возникает потому, что мы дополнили русский язык фразой, выделенной курсивом.

2. Определим новую функцию: $g(x) = f(x) - x$ для каждой $x \in [a, b]$. Она непрерывна и $g(a) > 0$, $g(b) < 0$, если концы интервала не являются неподвижными точками. Но если непрерывная функция меняет знак, найдется хотя бы одна точка x в которой $g(x) = 0 \Rightarrow f(x) = x$.

3. Действительно, пусть $f(x)$ имеет две неподвижных точки: $x^\star, y^\star$. Тогда, $d(f(x^\star), f(y^\star)) \leqslant sd(x^\star, y^\star)$. Но тогда $d(f(x^\star), f(y^\star)) = d(x^\star, y^\star) \leqslant sd(x^\star, y^\star)$. Следовательно, $d(x^\star, y^\star) = 0$.

Путеводитель по литературе. Основные понятия теории множеств можно найти в [33–35]. Небольшая книжка [41] содержит простое введение в теорию меры, история развития которой изложено в работе Лебега [3]. Отличное введение в теорию гладких многообразий содержит учебник [5]. Метрика Хаусдорфа в пространстве компактов и ее связь с фракталами описана в [2]. Сводку многих приведенных определений можно найти в справочнике [36].

Summa sine laude [16]

Разумеется, все что написано выше можно было бы изложить гораздо лучше, но ведь можно было и хуже. . .

Литература

1. *Пуанкаре А.* О науке. – М.: Наука, 1983. – 560 с.
2. *Barnsly M.* Fractals everywhere. – Academic Press, 1988. – 394 pp.
3. *Лебег А.* Об измерении величин. – М.: ГУПИМП, 1960. – 204 с.
4. *Стинрод Н., Чинн У.* Первые понятия топологии. – М.: Мир, 1967. – 224 с.
5. *Chillingworth D.* Differential topology with a view to applications. – Pitman Press, 1976. – 291 pp.
6. *Горелик Г. Е.* Размерность пространства. – М.: Изд-во МГУ, 1983. – 216 с.
7. *Зельдович Я. Б., Соколов Д. Д.* Фракталы, подобие, промежуточная асимптотика // УФН. – 1985. – т. 146. – с. 493–506.
8. *Perdang J.* Astrophysical fractals: An overview and prospects // Vistas in Astronomy. – 1990. – v. 33. – pp. 249–294.
9. *Соколов И. М.* Размерности и другие геометрические критические показатели в теории протекания // УФН. – 1986. – v. 150. – pp. 221–253.
10. *Голдман С.* Теория информации. – М.: Мир, 1967.
11. *Реньи А.* Трилогия о математике. – М.: Мир, 1980. – 376 с.
12. *Федер Е.* Фракталы. – М.: Мир, 1991. – 260 с.
13. *Баренблатт Г. И.* Подобие, автомодельность, промежуточная асимптотика. – Л.: Гидрометеоиздат, 1982. – 255 с.
14. *Мандельброт Б.* Общие свойства фракталов // В сб.: Фракталы в физике. – М.: Мир, 1988. – с. 9–47.
15. *Mandelbrot B.* Self-affine fractals and fractal dimension // Physica Scripta. – 1985. – v. 32. – pp. 257–260.
16. *Рихтмайер Р.* Принципы современной математической физики. т. 2. – М.: Мир, 1982. – 486 с.
17. *Иванов Л. Д.* Вариации множеств и функций. – М.: Наука, 1975.

[16] Итог без похвал (лат.)

18. *Янг Л.* Лекции по вариационному исчислению и теории оптимального управления. – М.: Мир, 1974. – 488 с.

19. *Pfeifer P., Obert M.* Fractals: Basic concepts and terminology // Chapter 1.2 in: Fractal Approach to Heterogeneous Chemistry: Surfaces, Colloids, Polymers / Ed.: David Avnir. – John Wiley & Sons Ltd., Chichester. – 1989. – p. 11.–40.

20. *Godreche C., Luck J. M.* Multifractal analysis in reciprocal space and the nature of the Fourier transform of self-similar structures // J. Physics. A.: Math. Gen. – 1990. – v. 23. – pp. 3769–3797.

21. *Hatchinson J.* Fractals and self-similarity // Indiana Univ. J. Math. – 1981. – v. 30. – pp. 713–747.

22. *Hatchinson J.* Fractals: A mathematical framework. – Department of Mathematics, School of Mathematical Sciences, Australian National University, Dec. 1996.
URL: http://www.csu.edu.au/ci/vol02/jeh2frac/jeh2frac.html

23. *Кроновер Р. М.* Фракталы и хаос в динамических системах. – М.: Постмаркет, 2000. – 350 с.

24. *Bourke P.* An introduction to fractals. – May 1991.
URL: http://astronomy.swin.edu.au/pbourke/fractals/fracintro
Paul Bourke's Home Page:
URL: http://astronomy.swin.edu.au/pbourke/

25. *Малинецкий Г. Г., Потапов А. Б.* Современные проблемы нелинейной динамики. – М.: Эдиториал, 2000. – 335 с.

26. *Лоскутов А. Ю., Михайлов А. С.* Введение в синергетику. – М.: Наука, 1990. – 272 с.

27. *Eckmann J. P., Ruelle D.* Ergodic theory of chaos and strange attractors // Rev. Mod. Phys. – 1985. – v. 57. – pp. 617–656.

28. *Schaw R.* Strange attractors, chaotic behavior, and information flow // Z. Naturforsch. – 1981. – v. 36a. – pp. 80–112.

29. *Milnor J.* On the concept of attractor // Commun. Math. Phys. – 1985. – v. 99. – pp. 177–195.

30. *Sprott J.* Strange attractors: Creating patterns in chaos. – New York: M & T Books, 1993. – 591 pp.
URL: http://sprott.physics.wisc.edu/sa.htm
http://sprott.physics.wisc.edu/fractals/booktext/sabook.pdf
Sprott's Home Page:
URL: http://sprott.physics.wisc.edu/sprott.htm

31. *Dynamical Systems* // In: Topics in Mathematics. Mathematical Archives, 1996–2001.
URL: http://archives.math.utk.edu/topics/dynamicalSystems.html
Surveys in dynamical systems available on-line:
URL: http://www.math.sunysb.edu/dynamics/surveys.html

32. *Young Lai-Sang*. Ergodic theory of chaotic dynamical systems. – Univ. of California, Department of Mathematics, Los Angeles, CA. – Oct. 1997. – 14 pp.
URL: http://www.math.ucla.edu/∼lsy/expository.html
Prof. Lai-Sang Young's Home Page:
URL: http://www.math.ucla.edu/∼lsy/

33. *Александров П. С.* Введение в теорию множеств и общую топологию. – М.: Наука, 1977. – 368 с.

34. *Келли Дж. Л.* Общая топология. – М.: Наука, 1968. – 432 с.

35. *Колмогоров А. Н., Фомин С. В* Элементы теории функций и функционального анализа. – М.: Наука, 1989. – 496 с.

36. *Фор Р., Кофман А., Дени-Папен М.* Современная математика. – М.: Мир, 1966. – 271 с.

37. *Severyanov V. M.* Automata network dynamical systems for construction of fractal objects // Electronic Proceedings of IMACS ACA'98 — the 4th International IMACS Conference on Applications of Computer Algebra. Czech Technical University, Prague, Czech Republic, August 9–11, 1998. – 8 pp. URL: http://www.math.unm.edu/ACA/1998/sessions/dynamical/sever

38. *Stark J.* Iterated function systems as neural networks // Neural Networks. – 1991. – v. 4. – pp. 679–690.

39. *Bressloff P. C., Stark J.* Neural networks, learning automata and iterated function systems // In: Fractals and Chaos / Eds.: A. J. Crilly et al. – 1991. – Springer-Verlag. – pp. 145–164.

40. *Niño F.* Random iterated neural networks: Properties, evolutionary design and applications. – Doctoral Dissertation, University of Memphis, May 2000. – 103 pp.
URL: http://www.msci.memphis.edu/∼ninol/diss_fn.ps
Fernando Niño's Home Page:
URL: http://www.msci.memphis.edu/∼ninol/index.html
URL: http://www.msci.memphis.edu/∼ninol/research.html

41. *Брудно А. Л.* Теория функций действительного переменного. – М.: Наука, 1971. – 119 с.

Николай Григорьевич Макаренко, ведущий научный сотрудник, кандидат физико-математических наук, руководитель группы в Лаборатории компьютерного моделирования (Институт математики, Алма-Ата, Казахстан). Область научных интересов: фрактальная геометрия, вычислительная топология, алгоритмическое моделирование, детерминированный хаос, нейронные сети, физика Солнца. Имеет более 50 научных публикаций.

УДК 004.032.26 (06) Нейронные сети

НАУЧНАЯ СЕССИЯ МИФИ–2002

НЕЙРОИНФОРМАТИКА–2002

IV ВСЕРОССИЙСКАЯ
НАУЧНО-ТЕХНИЧЕСКАЯ
КОНФЕРЕНЦИЯ

ЛЕКЦИИ
ПО НЕЙРОИНФОРМАТИКЕ
Часть 2

Оригинал-макет подготовлен Ю. В. Тюменцевым

ЛР №020676 от 09.12.97 г.
Подписано в печать 10.12.2001 г. Формат 60 × 84 1/16
Печ. л. 10,75. Тираж 250 экз. Заказ №

Московский государственный инженерно-физический институт
(технический университет)
Типография МИФИ
115409, Москва, Каширское шоссе, 31