ПРЕДИСЛОВИЕ

И хочешь знать,
Что ждет впереди.

В. Лебедев-Кумач

Дорогой читатель, вы держите в руках первую книгу на русском языке, посвященную современной математической теории машинного обучения и предсказания. И это довольно удивительно, так как эта тематика стала необычайно популярной в последние годы, по ней читаются курсы лекций в различных университетах страны и, главное, соответствующие методы и результаты активно используются на практике. Но, как говорится, лучше поздно, чем никогда.

Сама идея предсказания будущего (по прошлому, или по настоящему, или просто так) всегда притягивала человечество. Достаточно вспомнить великого французского математика Лапласа, писавшего два века назад: «Мы можем рассматривать настоящее состояние Вселенной как следствие его прошлого и причину его будущего. Разум, которому в каждый определенный момент времени были бы известны все силы, приводящие природу в движение, и положение всех тел, из которых она состоит, будь он также достаточно обширен, чтобы подвергнуть эти данные анализу, смог бы объять единым законом движение величайших тел Вселенной и мельчайшего атома; для такого разума ничего не было бы неясного, и будущее существовало бы в его глазах точно так же, как прошлое». Конечно, принцип неопределенности лишил этот тезис Лапласа былой убедительности, но зато появление компьютеров дало надежду на построение «обширного разума».

Оставалось совсем мало — «обучить машину».

Этим с энтузиазмом занялась кибернетика — наука более чем популярная полвека назад и почти забытая теперь. Тогда (50 лет назад) люди хотели научить компьютер хорошо играть в шахматы и впоследствии в этом преуспели, а вот, например, задача распознавания лиц (сцен и т. д.), казавшаяся поначалу более простой, только недавно

стала решаться машиной почти так же хорошо, как человеком (однако и сейчас компьютер не найдет вас на фото, сделанном в начальной школе, тогда как ваши сегодняшние друзья сделают это довольно уверенно). Постепенно пришло осознание того, что надо научить компьютер самому строить алгоритмы из некого их первоначального набора, обучаясь на исходных данных. Это и стало пониматься под термином «машинное обучение» (machine learning).

Возникла (и продолжает развиваться) новая, очень интересная наука, вобравшая в себя и уже хорошо разработанные разделы математики, такие как теория аппроксимации и математическая статистика (с современными добавлениями, в частности размерностью Вапника — Червоненкиса и методом опорных векторов), и совсем новые разделы науки, лежащие где-то между математикой и тем, для чего в русском языке так и не нашлось адекватного перевода — computer science, и, наконец, малопопулярные у нас теоретико-игровые методы в предсказании.

Автор книги Владимир Вячеславович Вьюгин — известный математик и замечательный педагог, специалист с мировым именем в теории алгоритмов, заведующий лабораторией в Институте проблем передачи информации им. А. А. Харкевича Российской академии наук. Основная часть этой книги была «обкатана» в курсах, прочитанных автором студентам МФТИ и ВШЭ.

Мне приятно отметить, что создание этой книги было поддержано лабораторией PreMoLab (http://premolab.ru/) МФТИ, созданной в рамках мегагранта Правительства РФ 11.G.34.31.0073 под руководством профессора В. Г. Спокойного из университета Гумбольдта.

Надеюсь, что эта книга окажется интересной и полезной как для совсем молодых ученых, только начинающих свой путь в науке и ищущих, где и в чем себя реализовать (с этой точки зрения машинное обучение и предсказания весьма перспективны), так и для зрелых исследователей, интересующихся данной тематикой. Наконец, мне бы очень хотелось, чтобы эта книга оказалась полезной для довольно многочисленной армии специалистов, реализующих методы и алгоритмы машинного обучения на практике.

Академик РАН,
директор ИППИ им. А. А. Харкевича
Кулешов А. П.

ВВЕДЕНИЕ

Основная задача науки и реальной жизни — получение правильных предсказаний о будущем поведении сложных систем на основании их прошлого поведения.

Многие задачи, возникающие в практических приложениях, не могут быть решены заранее известными методами или алгоритмами. Это происходит по той причине, что нам заранее не известны механизмы порождения исходных данных или же известная нам информация недостаточна для построения модели источника, генерирующего поступающие к нам данные. Как говорят, мы получаем данные из «черного ящика». В этих условиях ничего не остается, как только изучать доступную нам последовательность исходных данных и пытаться строить предсказания, совершенствуя нашу схему в процессе предсказания. Подход, при котором прошлые данные или примеры используются для первоначального формирования и совершенствования схемы предсказания, называется методом *машинного обучения* (Machine Learning).

Машинное обучение — чрезвычайно широкая и динамически развивающаяся область исследований, использующая огромное число теоретических и практических методов. Данная книга ни в какой мере не претендует на какое-либо исчерпывающее изложение содержания данной области. Наша цель — познакомить читателя с некоторыми современными математическими проблемами данной области и их решениями, основной из которых является проблема построения и оценки *предсказаний* будущих исходов.

С данным подходом тесно связана задача *универсального предсказания*. В том случае, когда мы не имеем достаточной информации, для того чтобы построить модель источника, генерирующего наблюдаемые данные, нам приходится учитывать как можно более широкие классы таких моделей и строить методы, которые предсказывают «не хуже», чем любая модель из данного класса. Понятие универсального предсказания, которое первоначально возникло в теории предсказаний стационарных источников, в настоящее время вышло далеко за рамки этой теории.

Первая часть книги — «Статистическая теория машинного обучения» — использует методы теории вероятностей и математической статистики. В основе данного подхода лежит предположение о том, что наблюдаемые исходы генерируются вероятностным источником, возможно, с неизвестными параметрами.

В рамках статистической теории машинного обучения мы рассматриваем задачи классификации и регрессии. Процесс обучения заключается в выборе функции классификации или регрессии из заранее заданного широкого класса таких функций.

Отметим два способа машинного обучения. При первом способе часть совокупности данных — *обучающая выборка* — выделяется только для обучения. После того как метод предсказания определяется по обучающей выборке, более он не изменяется и в дальнейшем используется для решения задачи предсказания.

При втором способе обучение никогда не прекращается, как говорится, оно происходит в *режиме онлайн*, т. е. предсказания и обучение происходят постоянно в процессе поступления новых данных.

Методы машинного обучения первого типа будут рассмотрены в первой части, которая посвящена статистической теории машинного обучения, методы второго типа будут изучаться во второй и третьей частях книги.

После того как схема предсказания определена, нам необходимо оценить ее предсказательные возможности, т. е. качество ее предсказаний. Предварительно напомним, как оцениваются модели предсказания в классической статистической теории. В статистической теории предсказания мы предполагаем, что последовательность исходных данных (или *исходов*) является реализацией некоторого стационарного стохастического процесса. Параметры этого процесса оцениваются на основании прошлых наблюдений, а на основании уточненного стохастического процесса строится правило предсказания. В этом случае *функция риска* данного правила предсказания определяется как среднее значение некоторой функции потерь, измеряющей различие между предсказаниями и исходами. Среднее значение вычисляется по «истинному вероятностному распределению», которое лежит в основе модели генерации данных. Различные правила предсказания сравниваются по значениям своих функций риска.

В статистической теории машинного обучения также используется стохастическая модель генерации данных, а именно, используется

предположение о том, что поступающие данные независимо и одинаково распределены. Вероятность ошибочной классификации или регрессии называется *ошибкой обобщения*. Первый шаг в сторону от классической постановки заключается в том, что распределение, генерирующее данные, нам может быть неизвестно и мы не можем и не будем оценивать его параметры, так как они не используются в оценках ошибок классификации или регрессии. Мы заранее не знаем, какой из методов классификации или регрессии будет построен по наблюдаемой части данных в процессе обучения; нам задан целый класс таких методов — например, это может быть класс разделяющих гиперповерхностей в многомерном пространстве. Оценки ошибки обобщения при классификации или регрессии должны быть равномерными по всем таким вероятностным распределениям и применяемым методам. Иными словами, эти оценки не зависят от распределения, генерирующего данные, а также от функции классификации или регрессии. Впервые данный подход был реализован в работах В. Н. Вапника и А. Я. Червоненкиса (см. [2]).

Для оценки предсказательной способности схемы классификации или регрессии используется *теория обобщения*. В рамках этой теории даются оценки вероятности ошибки классификации будущих данных при условии, что обучение проведено на случайной обучающей выборке достаточно большого размера и в его результате функция классификации (регрессии) согласована с обучающей выборкой. Важнейшим параметром такой оценки является *сложность* (*емкость*) класса функций классификации (регрессии). Обычно в оценке вероятности ошибки конкурируют длина выборки и сложность класса гипотез — при заданной величине ошибки чем больше длина обучающей выборки, тем больший по сложности класс гипотез можно использовать.

Методы вычисления ошибок обобщения и теория размерности классов функций излагаются в гл. 1.

Сложность классов функций будет измеряться тремя способами. Первый из них — функция роста и связанная с ней размерность Вапника — Червоненкиса (VC-размерность), которые известны с середины 60-х годов XX века. Позже были введены числа покрытия и упаковки и связанная с ними пороговая размерность (fat-размерность), которые дают более точные верхние оценки ошибки обобщения в том случае, когда разделение данных производится с заданным порогом. Еще один способ измерения сложности класса функций — средние Радемахера — также изучается в этой главе. Последние два способа

измерения емкости класса функций в отличие от VC-размерности не зависят от размерности пространства объектов.

Глава 2 посвящена построению алгоритмов классификации и регрессии. В основном это алгоритмы, использующие метод опорных векторов. Рассматриваются методы распознавания образов на основе построения разделяющих гиперплоскостей или гиперповерхностей в пространствах признаков, построенных с помощью ядерных методов. Излагаются основы теории функциональных гильбертовых пространств, порожденных воспроизводящим ядром (Reproducing Kernel Hilbert Space — RKHS), и их применение для построения разделяющих гиперповерхностей и для получения оценок ошибки классификации.

Во второй части — «Онлайн-методы машинного обучения» — вообще не используются никакие гипотезы о стохастических механизмах, генерирующих данные. Наблюдаемые исходы могут генерироваться совершенно неизвестным нам механизмом, который может быть как детерминированным, так и стохастическим или даже «адаптивно враждебным» к нашим предсказаниям (т. е. может использовать наши прошлые предсказания при генерации очередного исхода).

При этом возникает естественный вопрос: как в этом случае оценивать предсказательную способность метода? В отсутствие вероятностной модели функция риска в виде математического ожидания не может быть определена. В теории последовательного предсказания (гл. 3) используются два подхода для оценки качества предсказаний. В разделе 3.1 рассматриваются максимально широкие классы моделей, описывающих возможные варианты поведения источника, генерирующего данные. Строится метод, собирающий все возможности этих моделей в одну агрегирующую модель. Производится сравнение предсказательной способности этой модели с предсказательными способностями всех моделей класса. В разделе 3.1 в качестве примера агрегирующих моделей рассматривается смешивающий предсказатель и правило Лапласа, а также оптимальный минимаксный предсказатель и его приложения — универсальное кодирование, игра Келли.

В разделе 3.2 приведен метод построения хорошо калибруемых предсказаний. Для оценки качества предсказаний используются тесты, оценивающие рассогласованность между предсказаниями и соответствующими исходами. Эти тесты выбираются исходя из тех задач, для решения которых будут использоваться предсказания.

Один из видов таких тестов — тесты на калибруемость. Цель алгоритма — выдавать такие предсказания, которые выдерживают все тесты на калибруемость.

Тесты на калибруемость строятся в зависимости от того, как мы планируем использовать хорошо калибруемые предсказания. В разделе 3.5 специальные тесты и соответствующие им хорошо калибруемые предсказания будут использованы при построении универсальной алгоритмической стратегии для торговли на финансовом рынке. Алгоритм такой стратегии автоматически покупает и продает акции. Мы докажем, что доход при такой торговле будет не меньше, чем доход любой стационарной алгоритмической стратегии.

Основные принципы сравнительной (или соревновательной) теории предсказания рассматриваются в гл. 4. Эффективность алгоритма предсказания оценивается в форме сравнения с предсказаниями некоторого набора экспертных методов, или просто экспертов. В теории предсказаний с учетом экспертов вводится класс предсказателей, называемых экспертами. Класс экспертов может быть конечным или бесконечным, может иметь мощность континуума. В качестве экспертов могут рассматриваться различные методы предсказания, стохастические теории, методы регрессии и т. д. Эксперты предоставляют свои прогнозы прежде, чем будет представлен соответствующий исход. Когда очередной исход становится известным, каждый эксперт вычисляет свои потери. С ростом числа предсказаний и исходов потери эксперта суммируются и образуют его кумулятивные потери. Наилучшим называется эксперт, несущий минимальные кумулятивные потери. Агрегирующий алгоритм при вычислении своих прогнозов может использовать текущие прогнозы и потери экспертов в прошлом.

В разделах 4.1 и 4.2 будут построены адаптивные алгоритмы предсказания, которые несут не бóльшие потери, чем потери наилучшего эксперта с точностью до некоторой ошибки обучения, называемой «регретом». В разделе 4.2 приводится алгоритм распределения потерь в режиме онлайн Фройнда и Шапире, который применяется в разделе 4.3 для усиления слабых алгоритмов классификации. Слабый алгоритм классификации делает лишь незначительно меньшее число ошибок, чем простое случайное угадывание. В разделе 4.3 излагается алгоритм усиления слабых классификаторов — бустинг (Boosting). Приводится алгоритм AdaBoost, решающий эту задачу. Алгоритм AdaBoost усиливает слабый алгоритм классификации до алгоритма, который

с некоторого момента в процессе обучения начинает делать как угодно малое число ошибок. Приведены достаточные условия эффективности этого алгоритма.

В разделе 4.4 приводится вероятностный алгоритм Ханнана следования за возмущенным лидером, в котором принимается решение следовать за экспертом, несущим наименьшие потери в прошлом. Прямое следование этой идее может привести к неприемлемым потерям предсказателя, однако введение рандомизации позволяет построить алгоритм с достаточно малым регретом.

В разделе 4.5 мы рассмотрим общие свойства алгоритмов экспоненциального смешивания. При анализе этих алгоритмов основную роль играют так называемые экспоненциально смешанные потери. Кумулятивный вариант этой величины асимптотически близок к потерям каждого из вариантов алгоритмов экспоненциального смешивания. Поэтому анализ свойств этих алгоритмов в основном сводится к анализу экспоненциально смешанных потерь.

В разделе 4.6 рассматривается более детальная постановка, при которой потери экспертов и агрегирующего алгоритма возникают в результате принятых ими решений. В процессе прогнозирования выдаются исходы, а функция потерь сопоставляет решения (прогнозы) с этими исходами и выдает соответствующие потери.

В разделе 4.7 рассматривается алгоритм отслеживания наилучшей комбинации экспертов Fixed-Share. В предыдущих разделах мы сравнивали потери агрегирующего алгоритма с потерями наилучшего эксперта. Другой возможный подход состоит в следующем: серия шагов, на которых производится обучение, делится на сегменты. Каждому сегменту ставится в соответствие свой эксперт, последовательность сегментов и соответствующих экспертов называется составным экспертом. Цель агрегирующего алгоритма также несколько меняется — теперь он должен предсказывать так, чтобы его потери были не больше, чем потери составного эксперта (с точностью до некоторого регрета). Дальнейшие обобщения схем смешивания рассматриваются в разделе 4.8.

В разделе 4.9 рассматривается постановка задачи предсказания с экспертами в условиях частичного мониторинга окружающей среды. В этом случае предсказатель в результате своего действия получает только собственные потери и не имеет доступа к потерям других экспертов. В мировой литературе эта постановка называется «проблема многоруких бандитов» (multi-armed bandit problem).

В гл. 5 мы возвращаемся к задаче предсказания с использованием экспертных стратегий. Рассмотрен агрегирующий алгоритм Вовка, который имеет значительно меньшую ошибку предсказания для логарифмической, квадратичной и некоторых других функций потерь, чем метод экспоненциального смешивания, использованный в гл. 4, и построен соответствующий алгоритм многомерной регрессии в режиме онлайн, основанный на применении агрегирующего алгоритма.

В гл. 6 задача предсказания с использованием экспертных стратегий рассматривается в терминах теории онлайн выпуклой оптимизации. В разделе 6.2 приведен и изучен алгоритм онлайн градиентного спуска, а в разделе 6.3 рассматривается онлайн-оптимизация общего вида, основанная на методе зеркального спуска, частными случаями которого являются онлайн градиентный спуск и алгоритмы оптимального смешивания.

Предсказания в режиме онлайн тесно связаны с теорией игр. Третья часть — «Игры и предсказания» — посвящена изложению теории предсказаний на языке теории игр.

Основные понятия теории игр рассматриваются в гл. 7. Мы рассмотрим матричную игру двух лиц с нулевой суммой и докажем для нее минимаксную теорему Дж. фон Неймана. Доказательство минимаксной теоремы проведено в стиле теории машинного обучения с использованием метода экспоненциального смешивания. В этой главе также вводятся понятия равновесия Нэша и коррелированного равновесия Аумана.

В гл. 8 рассматривается новый теоретико-игровой подход к теории вероятностей, предложенный В. Г. Вовком и Г. Р. Шейфером [49]. В рамках этого подхода формулируются игры с предсказаниями, на траекториях которых при определенных условиях выполнены различные законы теории вероятностей. Примеры таких законов — закон больших чисел, закон повторного логарифма, центральная предельная теорема и т. д. Вводится понятие теоретико-игровой вероятности, которое определяется для подобных игр.

В рамках этого подхода также наиболее естественным образом формулируется задача построения универсальных предсказаний, рассмотренная в гл. 3.

В гл. 9 рассматриваются более сложные вопросы теории игр. В основе излагаемой теории находится знаменитая теорема Блекуэлла о достижимости (Blackwell approachability theorem). Эта теорема яв-

ляется обобщением минимаксной теоремы для игр двух лиц с произвольными векторнозначными функциями выигрыша. Теорема Блекуэлла служит основой для построения калибруемых предсказаний для случая произвольного конечного числа исходов.

В свою очередь, в этой же главе будет показано, что использование калибруемых предсказаний позволяет построить стратегии, при которых совместное частотное распределение ходов всех игроков сходится к коррелированному равновесию Аумана.

Данная книга представляет собой краткий обзор идей и математических методов современной теории машинного обучения и тесно связанных с ней теории предсказания с использованием экспертных стратегий, нестохастических теоретико-игровых методов предсказания, теоретико-игровых основ теории вероятностей и теории универсальных предсказаний. По мнению автора, все эти темы представляют собой необходимый минимум теоретических знаний для студентов и аспирантов, специализирующихся в области машинного обучения и искусственного интеллекта.

Работа над этой книгой частично проводилась в лаборатории структурных методов анализа данных в предсказательном моделировании МФТИ. Материал данной книги использовался в качестве основы курсов лекций, прочитанных автором в 2008–2013 годах на базовых кафедрах ИППИ РАН, на факультете управления и прикладной математики МФТИ и на факультете компьютерных наук НИУ ВШЭ. Данная книга является существенным расширением учебного пособия [3].

Главы 1 и 2 могут послужить основой для курса лекций «Статистическая теория машинного обучения». Главы 3, 4 и 5 могут послужить основой для курса лекций «Универсальные предсказания», и, наконец, на основе гл. 3, 7, 8 и 9 можно составить курс «Онлайн методы машинного обучения»

С рядом идей и постановок задач, представленных в данной книге, автор познакомился во время краткосрочных визитов в отделение компьютерных наук колледжа Royal Holloway Лондонского университета. Автор с благодарностью вспоминает многолетнее общение с сотрудниками и аспирантами этого отделения: Александром Гаммерманом, Владимиром Вовком, Юрием Калнишканом, Михаилом Вьюгиным, Ильей Нуретдиновым, Алексеем Черновым, Федором Ждановым, Дмитрием Адамским и Лео Гордоном.

Автор благодарен Юрию Калнишкану и Алексею Чернову за предоставленные материалы.

Автор особенно благодарен Владимиру Вовку за многочисленные беседы и разъяснения по многим вопросам, затронутым в данной работе.

Второе издание содержит следующие дополнительные разделы. В гл. 3 добавлен раздел 3.1. В гл. 4 добавлены разделы 4.5–4.9. В гл. 5 добавлен раздел 5.8. Новой также является гл. 6.

В третье издание добавлены разделы 6.1 и 9.5. Дополнены разделы 1.2.1, 2.5, 4.2 и 4.3.

Автор благодарен А. В. Гасникову, Г. А. Кабатянскому, С. М. Карпенко, А. А. Коротину, А. П. Кулешову, А. В. Назину, Ю. М. Ольховской за внимание к данной работе и ценные замечания, а также студентам МФТИ и ВШЭ, указавшим на опечатки и неточности первого и второго изданий этой книги.

Работа над дополнениями к третьему изданию книги частично поддержана грантом РФФИ: проект 20-01-00203.