Лекции Школы анализа данных Яндекса

А. Я. Червоненкис

Компьютерный анализ данных

Издательство осуществляется
при поддержке компании «Яндекс»

Предисловие

Теория обучения машин в изложении А. Я. Червоненкиса

Перед Вами замечательная, очень нужная сегодня книга.

Она необходима и практику, строящему обучающие машины для реальных проектов, и специалистам, разрабатывающим новые методы обучения машин, и, конечно, студентам, которые решили специализироваться в области анализа данных.

Казалось бы, литература о методах построения прогнозных или обобщающих моделей довольно обширна, программное обеспечение доступно и удобно[1] — чего же еще желать? Однако для построения систем, которые работали бы в реальном мире, этого недостаточно. Нужны еще средства оценки качества моделей — для того, чтобы точность используемых на практике моделей была не ниже некоторого определенного уровня. При массовом производстве конструктор получает гарантии качества методами так называемого разрушающего контроля — прямым тестированием определяют, какой процент деталей выдерживает требуемые испытания. Этот процент и является искомой характеристикой качества. По для конструирования хороших обучающих машин такой подход в принципе не подходит. Здесь, как и в медицинской практике, где все вопросы гарантий решаются применительно к конкретному больному, оценка должна быть сделана для индивидуального экземпляра машины — просто потому, что нет возможности создать различные массовые представительные испытания для многих копий одного типа машины. Это не только дорого и требует больших затрат времени — в подавляющем числе случаев это невозможно осу-

[1] Уточню, что речь идет о литературе на английском языке; на русском языке книга А. Я. Червоненкиса дает первое полное изложение теории обучения машин.

ществить в принципе (обучающий материал часто уникален, а модель должна работать в самых разных условиях).

Вопрос о гарантиях качества обучения машин в литературе не игнорируется, но сводится лишь к упоминанию одной-двух процедур, помогающих находить эти гарантии. Более того, в подавляющем большинстве книг авторы не объясняют, почему эти процедуры работают — исключение составляет книга Ш. Вейца и К. Куликовского [Sholom Weiss and Casimir Kulikowski — Computer systems that learn, 1991] , которая уделяет много внимания оценке качества обучаемых машин. В ней подробно описываются основные процедуры, позволяющие оценить качество конструируемых обучаемых машин, рассматриваются практические примеры их использования, четко описываются выводы из результатов применения этих процедур. Однако и в этой книге не приводится обоснований. В итоге способы создания таких процедур и возможности приложения их к конкретным задачам остаются скрытыми от читателя[2].

Книга А. Червоненкиса — это первая полная монография по теории машинного обучения, центральной задачей которой является именно обоснование оценок качества обучающихся моделей[3]. Для автора проблема обоснования — это источник поиска новых нутей развития всей теории обучающих машин. В книге автор по-новому рассматривает классические принципы Байеса и максимума правдоподобия, детально описывает условия, при которых эти принципы обеспечивают высокие гарантии оценки качества. И он находит новые возможности оценки за счет использования классических принципов совместно с использованием новых оценок обобщающих возможностей обучаемых машин, полученных на основе концепций VC-размерности и структурной минимизации риска. Эти новые возможности не только интересны са-

[2]Мой личный опыт общения с молодыми специалистами по машинному обучению показал, что они в подавляющем большинстве знают базовые процедуры оценки качества моделей, но ситуации, в которых эти оценки надо приспособить к конкретным условиям, вызывают затруднения. Более того, я не встречал ни одного молодого специалиста, который бы знал теоретические принципы, позволяющие обосновать эти процедуры и понимать, какими условиями ограничена область их применения.

[3]А. Червонекису принадлежат наиболее выдающиеся результаты в разработке оценок качества моделей, которые он сделал совместно с В. Вапником еще в 70-е годы прошлого века (VC-размерность оценки и метод структурной минимизации риска).

ми по себе, но и поучительны с общей методологической точки зрения. Они показывают, насколько плодотворны сравнительные исследования. Специалисты в области обучения машин, несомненно, обратят внимание на новые соображения о параллельном анализе оценок качества с позиций Байесова подхода и структурпой минимизации риска. Мпе представляется, что эти соображения могут послужить важным стимулом для нового этапа развития теории обучения машин вообще.

Книга состоит из 22 разделов-лекций. Первая лекция дает общую картину теории, сразу обращая внимание читателя на главную нацеленность книги — на поиск ответа на вопрос о том, как строить прогнозные системы с гарантированной (и по возможности максимально большой) обобщающей способностью. Уже из этой вводной лекции становится ясно, что классическая парадигма разработки методов, позволяющих восстанавливать истинные скрытые зависимости, рассматривается автором как неудовлетворительная идеализация практических потребностей. Эта парадигма объясняет как искать эти истинные зависимости в условиях, когда количество данных для оценки неограниченно. В этих условиях можно строить теорию, которая дает асимптотические оценки качества.

Асимптотические оценки не устраивают Червоненкиса. Во-первых, потому что он исходит из того, что в большинстве практических случаев пользователь или исследователь не располагает возможностями неограниченных наблюдений[4]. Во-вторых, потому что асимптотические гарантии практически не позволяют делать какие-либо оценочные утверждения для моделей, построенных на ограниченных выборочных данных.

А. Червоненкис формулирует главную цель теории обучающихся машин как построение эффективных прогнозных методов и подчеркивает — даже если мы знаем, что лучший прогноз не достигается на аппроксимации истинной модели, надо выбирать его, поскольку при ограниченном объеме наблюдений невозможно приблизить истинную модель так, чтобы вместе с прогнозом получить и строгие гарантии отклонения аппроксимации от истинной модели.

Вторая и третья лекции посвящены основам теории вероятности (понятию случайной величины и закону больших чисел). Автор дози-

[4]Заметим, что законы, управляющие наблюдениями, меняются; иначе говоря, объем наблюдений для обучения принципиальпо ограничен.

ровано выбирает из теории вероятности необходимый для дальнейшего изложения материал. При этом он не забывает, что книгу будут изучать инженеры. Эти две лекции можно рекомендовать для автономного изучения студентам и специалистам инженерных специальностей, которым необходима дополнительная подготовка по теории вероятности, даже если при этом их учеба и работа никак не связаны с теорией обучения машин[5]. Хочу подчеркнуть, что и в этих по существу вспомогательных лекциях автор дает очень важный новый материал, который не рассматривается в современных учебниках по теории вероятности, например, закон о равномерной сходимости частот к вероятностям (обобщение закона больших чисел, открытое им совместно с В. Н. Вапником, составившее теоретическую основу их новых методов оценки качества обучающих машин).

Следующие четыре лекции (4–7) — описание основ современной теории обучения машин. Здесь даются и описания большинства наиболее известных алгоритмов обучения, и основные теоретические положения. Эти четыре лекции также можно читать автономно: даже без чтения лекций 2-3 они дают ясное представление о теории обучения машин, особенно в ее инженерной части, когда наибольший интерес представляют вопросы конструирования обучаемых машин.

С восьмой лекции начинается главная часть книги. Лекции 8 и 9 содержат основные инженерные идеи, предложенные А. Я. Червоненкисом вместе с В. Н. Вапником еще в 70-е годы. Они и в настоящее время занимают центральное место в богатой коллекции методов обучения машин. Анализ именно этих конструкций привел В. Н. Вапника и А. Я. Червоненкиса к созданию общей схемы анализа обобщающей способности обучаемых машин. Ее вероятностно-теоретическая основа излагается в лекциях 10–12.

Остальные лекции (с 13 по 22), хотя и продолжают описание теории Вапника–Червоненкиса, нацелены прежде всего на описание нового общего взгляда на теорию обучения машин. Червоненкис здесь сопоставляет разные методы — и свои, и других авторов, и хорошо известные, и совсем новые. Основная цель автора в этих главах — рассмотреть

[5]Конечно, только этих лекций будет недостаточно, но чтение этих двух глав позволит сделать обучение по стандартным учебникам гораздо более эффективным: читатель будет яснее понимать, зачем в этих учебниках строятся «сложные вспомогательные» конструкции, прежде чем рассматриваются содержательные задачи.

возможности и ограничения этих методов. Его подход воспринимается как совместный с читателем поиск путей усиления этих методов за счет того или иного их интегрирования. Читая эту часть книги, невольно увлекаешься его анализом и начинаешь сам что-то придумывать — ты понимаешь, что автор пригласил тебя в свою творческую лабораторию. Червоненкис часто формулирует новые интересные вопросы, на которые пока не известны ответы, но эти вопросы заставляют думать и лучше нонимать описываемые результаты. Важнейший элемент его объяснений в этой части книги — это тщательное описание исходных предположений для различных методов. Автор систематически проводит параллель между разными исходными предположениями в разных методах. В результате читатель понимает, «что на что обменивается» нри переходе от одного метода к другому, и эта идея очень полезна для создания целостного видения теории машинного обучения.

Червоненкис не пренебрегает детальным анализом частных случаев, например, подробным рассмотрением примера задачи классификации событий из двух гауссовых распределений. Этот пример разбирается практически во всех учебниках по теории обучения машин, но Червоненкис не повторяет этих описаний. Он рассматривает особенности проблемы оценки качества модели и для этого совсем частного случая. Очень удачным, в частности, считаю, включение в описание результатов эксперимента, выполненного Шарунасом Раудисом. Эта иллюстрация убедит любого инженера в том, что задача прогноза при ограниченных данных принципиально отличается от задачи восстановления истинной зависимости.

По объему книга Червоненкиса очень небольшая, и я, читая ее, много раз ловил себя на мысли, что не понимаю, как ему удалось в таком маленьком объеме рассмотреть теорию обучения не только целостно, но с большими подробностями, да еще поместить много интересного дополнительного материала, как, например, метод кригинга[6]. Однако читая книгу несколько раз, начинаешь понимать, что это ему удалось за счет

1. тщательного отбора материала (в частности, за счет подчеркивания особой важности двух общих принципов: ключевой роли оценки способности к генерализации у модели, и необходимости в

[6]Уверен, что этот метод, впервые использованный в геологических науках, в изложении Червоненкиса станет базовым методом анализа данных общей природы.

конкретном приложении максимально учитывать априорную информацию),

2. очень плотного изложения,

3. поиска максимально простого и одновременно точного объяснения всех изложенных утверждений и конструкций (например, Червоненкис для наглядности часто пользуется таким приемом: описывается такой мощный метод, как метод опорных векторов, и сразу же специальным подразделом рассказывается, в каких случаях этот метод не работает; идея обязательно объяснять ограниченность метода идет параллельно с уже отмеченным особым вниманием к предположениям, лежащим в основе каждого метода).

Книга действительно очень «плотная» — и по множеству излагаемых проблем, и по множеству идей, предлагаемых для их решения. Ее нельзя просто читать и тем более — просматривать. Ее надо изучать и продумывать освоенное. Хорошо при этом для сравнения читать и книги других авторов. Не для лучшего понимания материала книги, а для лучшего осознания принципиальной важности общих, я сказал бы — философских принципов теории познания, рассмотренных в книге, где они изложены в рабочем порядке. Поэтому я дополнил список литературы автора некоторыми работами, с которыми стоит познакомиться при изучении этой книги.

Повторю то, с чего начал: книга будет полезна и экспертам в области теории и приложений теории обучения машин, и молодым специалистам, которые только что решили специализироваться в этой области. Возможность же изучения этой книги с помощью одного из авторов, заложивших фундамент теории обучения, делает книгу особенно ценной именно для студентов школы анализа данных Яндекса.

И. Б. Мучник
октябрь 2009.

Содержание

Лекция 1

Задача восстановления зависимостей. Интерпретация в терминах выбора функции из заданного класса. Интерпретация в терминах выбора модели из заданного класса моделей. Интерпретация в терминах имитации одного автомата другим. Критерии выбора.

В этой лекции я дам краткий обзор всего курса, который будет читаться на протяжении ближайших двух семестров.

Проблема восстановления зависимостей по эмпирическим (экспериментальным, статистическим) данным стала весьма актуальной в очень широком круге приложений. Методы решения этой задачи известны под названием «методы машинного обучения» (Machine Learning). Сюда входят методы построения регрессионных зависимостей и решения обратных задач математической физики и статистики, методы машинного обучения распознаванию образов (как зрительных, так и абстрактных — представленных набором признаков) и многие другие. Оказывается, что к этому кругу относятся и многие задачи, возникающие при управлении Интернетом.

Эту задачу можно формально изложить так [4, 5]:

Дано: вход $\mathbf{x}$ — выход $\mathbf{y}$.

Наблюдаем последовательность реализаций (пар): $(\mathbf{x}_1, \mathbf{y}_1)$, $(\mathbf{x}_2, \mathbf{y}_2)$, $(\mathbf{x}_3, \mathbf{y}_3)$, …, $(\mathbf{x}_l, \mathbf{y}_l)$ — *обучающую выборку.*

Хотим найти зависимость $\mathbf{y} = F(\mathbf{x})$ такую, чтобы предсказанные значения $\mathbf{y}^* = F(\mathbf{x})$ как можно точнее аппроксимировали фактическое выходное значение $\mathbf{y}$, соответствующее входу $\mathbf{x}$.

Возможны различные схемы генерации пар $(\mathbf{x}_i, \mathbf{y}_i)$. С теоретической точки зрения удобнее всего считать, что пары генерируются независимо при неизменном (но заранее неизвестном) распределении вероятностей $P(\mathbf{x}, \mathbf{y})$. В практических же задачах наблюдения часто оказываются зависимыми. Бывает и так, что само распределение $P(\mathbf{x}, \mathbf{y})$ меняется со временем. Тогда необходимо применять адаптивную схему обучения, когда восстановленная функция также меняется со временем. Иногда вообще не предполагают, что на множестве пар задано какое-либо распределение вероятностей. Тогда требуют, чтобы функция хорошо аппроксимировала зависимость вход–выход при любых значениях входа в пределах заданной области определения.

Есть и такие задачи, где по наблюдаемым значениям на выходе требуется определить, что же было на входе. Такие задачи называются обратными.

Регрессия

Рассмотрим сначала задачу восстановления функций [1, 2], полагая, что выходное значение — это число или одно из возможных дискретных значений. Входом может в простейшем случае служить просто числовое значение. Тогда мы говорим о восстановлении функции одной переменной. В более общем случае это может быть вектор, имеющий n координат. Тогда мы говорим о восстановлении функции n переменных, имеющих различную природу.

Если зависимость между входными переменными и выходом линейная (или ищется ее линейное приближение), то для нахождения коэффициентов линейной зависимости применяется хорошо известный метод наименьших квадратов, о котором мы будем подробнее говорить на одной из следующих лекций. Однако, если длина выборки мала (не достаточно велика) по сравнению с числом аргументов, то МНК в чистом виде не работает или работает плохо. В этих случаях приходится применять те или иные методы регуляризации. Об этом мы тоже будем говорить позже.

Если связь между аргументами и выходом описывается полиномом конечной степени (или ищется в таком виде), то задачу можно свести к предыдущей, пополнив число аргументов степенями исходных. В случае многих переменных приходится включать и произведения степеней разных аргументов. При этом, конечно, число искомых коэффициентов возрастает и задача усложняется.

Можно использовать не только алгебраические полиномы, но и тригонометрические, и вообще линейные разложения по заранее выбранной системе функций.

Если же ищется зависимость в форме кусочно-линейной функции (или кусочно-полиномиальной), то обычный метод наименьших квадратов уже неприменим, и здесь приходится использовать иные средства, например, настраиваемые нейронные сети.

Существуют методы, позволяющие оценивать по предъявленной обучающей выборке значение функции в заданной точке (или последовательно в заданной сети точек). К их числу относятся гребневая

регрессия (ridge regression) и кригинг, которые мы тоже рассмотрим позже. Одна из важнейших задач, где эти методы применяются — восстановление трехмерного поля концентраций полезных (или вредных) компонентов в недрах земли по измерениям содержания в пробах, взятых в дискретной сети точек. Другой круг приложений связан с предсказанием определенных свойств некоторой новой конструкции по ее заданным параметрам. Например, по описанию профиля крыла, фюзеляжа и другим конструктивным особенностям новой модели самолета предсказать его поведение в воздухе при разных условиях.

Но на самом деле в качестве входных переменных могут служить кривые, изображения, графы, *тексты*, *сообщения* и т.д.

Природа искомой зависимости может быть различной. В одних случаях выходное значение в действительности связано со входом некоторой детерминированной функцией, но наблюдения (как те, которые включены в массив обучения, так и те, которые предстоит прогнозировать) искажены аддитивным, мультипликативным или иным шумом, независимым или зависящим от фактического значения на выходе. В других случаях сама связь между входом и выходом по природе своей оказывается стохастической. Возможно и так, что на выход влияют ненаблюдаемые переменные, так или иначе (детерминировано или случайно) связанные с тем, что доступно наблюдению. В некоторых случаях выходное значение, которое мы пытаемся оценить, в действительности является не следствием, а причиной того, что наблюдается. Например, в задачах медицинской диагностики причиной заболевания будет та или иная инфекция или патологические изменения в организме. Но наблюдаем мы симптомы, которые являются следствием этого. Задача же состоит в том, чтобы по этим наблюдаемым симптомам узнать природу заболевания. Или в сейсмике, по измерениям на сейсмостанциях определить координаты и очага землетрясения и степень его опасности.

Распознавание

Задачи обучения распознаванию образов мы также рассматриваем как частный случай восстановления функциональных зависимостей. Просто здесь значениями функции будут имена классов. Задачи распознавания образов охватывают чрезвычайно широкий круг приложений — от распознавания изображений и звуковых сигналов до распознава-

ния определенных фрагментов в молекуле ДНК, классификации сообщений или распознавания попыток несанкционированного доступа. Во всех случаях в конце концов требуется построить решающее правило, которое по заданному на входе описанию объекта $\mathbf{x}$ относит его к одному из классов $\mathbf{y}$ (правильно или ошибочно, но желательно, чтобы ошибок было мало). Но природа задач распознавания бывает принципиально различной. В одних случаях объективно распознаваемые объекты принадлежат разным классам $Z_1, Z_2, \ldots, Z_k$. Сами объекты даны нам в форме своих описаний $\mathbf{x}$. Это могут быть изображения, звуковые сигналы, наборы симптомов и т.д. Для каждого класса характерно свое распределение вероятностей на множестве описаний $P(\mathbf{x}|Z_i)$. Обычно известны и априорные вероятности принадлежности объекта к тому или иному классу $P(Z_i)$. Так возникает совместное распределение $P(\mathbf{x}, Z_i) = P(Z_i)P(\mathbf{x}|Z_i)$. Теперь по формуле Байеса легко находятся условные вероятности принадлежности объекта к классу по заданному описанию:

$$P(Z_i|\mathbf{x}) = \frac{P(Z_i)P(\mathbf{x}|Z_i)}{C},$$

где C — нормировочная константа. Если носители распределений $P(\mathbf{x}|Z_i)$ не пересекаются, т. е. одно и то же описание не может быть порождено объектами разных классов, то возможно стопроцентное распознавание. В противном случае ошибки неизбежны. Проблема лишь в том, что распределения $P(\mathbf{x}|Z_i)$ неизвестны, и их нужно восстановить по обучающей выборке. Именно таким путем и шли, начиная с 20-х годов прошлого века (дискриминационная функция Фишера, «наивный» Байес, когда признаки считаются условно независимыми). Однако такой путь оказывается эффективным лишь в простых случаях. Восстановление плотности распределения в многомерных пространствах оказывается задачей значительно более сложной, чем прямой поиск решающего правила, обеспечивающего малое число ошибок. Впрочем, идея выделения кластеров в пространстве описаний, да и метод ближайшего соседа, о котором речь еще пойдет, основаны на идее, что носители распределений $P(\mathbf{x}|Z_i)$ образуют компактные непересекающиеся (или слабо пересекающиеся) множества. Часто эти носители хотя и не пересекаются, но в пространстве исходных признаков переплетаются сложным образом, и простые кластерные методы не работают. Приходится или искать достаточно сложную разделяющую поверхность, или

осуществлять сложную предварительную обработку.

В других случаях имеются выходные переменные, связанные со входом, и принадлежность объекта к тому или иному классу зависит от того, попадут ли эти переменные в заданную область. Например, качество продукции зависит от параметров сырья и управляющих переменных (концентрации ингредиентов, температуры, давления). Нас же интересует, при каких значениях параметров $\mathbf{x}$ (входных переменных) продукт удовлетворяет заданным требованиям ($\mathbf{y}$ = «да» или $\mathbf{y}$ = «нет»). В медицине к этому же ряду относится прогноз лечения. Конечно, в этом случае можно было бы спрогнозировать значение выходных переменных и непосредственно поверить, попадают ли они в заданную область. Но такой путь часто оказывается более сложным, чем распознавание, и требует больше данных на обучение. К тому же выходные переменные часто оказываются ненаблюдаемыми. Сообщается лишь, попали они в заданную область или нет.

Функция, по заданному описанию объекта определяющая имя класса, к которому он принадлежит, называется решающим правилом. Решающее правило может искаться непосредственно в пространстве исходных признаков в виде разделяющей поверхности или системы разделяющих поверхностей. Так, например, удалось успешно обучить машину распознавать рукописные цифры в почтовых индексах на конвертах, построив достаточно сложную разделяющую поверхность в пространстве исходных признаков, каковыми были интенсивности отдельных пикселей на изображении (правда, после предварительной центровки и масштабирования). В более сложных случаях необходима предварительная обработка изображения — выделение линий, контуров, точек сопряжения, в результате которой получается описание в терминах преобразованных признаков. В пространстве этих признаков уже строятся разделяющие поверхности. Может использоваться предварительное Фурье преобразование, обеспечивающее инвариантность относительно сдвига, преобразование Фурье—Меллина, которое дает инвариантность относительно поворотов и т.д. Подобного рода преобразования исходных признаков и числовых аргументов применяются и в других задачах распознавания. Правда, эти преобразования обычно фиксируются до начала обучения, а собственно обучение сводится к построению решающего правила в пространстве уже преобразованных признаков. Хотелось бы включить нахождение нужного преобразования в сам процесс обучения. Но это оказывается очень сложной и пока не решенной

задачей. В известной мере это делают нейронные сети, но далеко не так успешно, как хотелось бы.

Широкое поле применения методов распознавания составляет медицинская диагностика от выделения групп риска по анамнезу и обычным клиническим анализам, до диагностики сложных заболеваний с применением современных методов исследования. В последние входит применение микро-эрреев, где измеряется интенсивность экспрессии отдельных генов, протеомика, где диагноз ставится на основании масс–спектрографии сыворотки крови. В качестве входного описания здесь может использоваться не только статическая картина на текущий момент, но и динамика изменений за некоторый промежуток времени.

Другой интересный пример — это распознавание попыток несанкционированного доступа к системам и базам данных. Дело в том, что поведение обычного пользователя существенно отличается от поведения лица, пытающегося осуществить несанкционированный доступ. По ряду признаков, описывающих поведение пользователя, удается решить эту задачу.

Итак, при восстановлении функций мы считаем, что заранее задаются некоторым набором (классом) функций и на основании обучающей выборки выбирают из него такую функцию, которая в том или ином смысле наиболее подходит для описания зависимости между входом и выходом. О критериях, по которым оценивается, насколько функция подходит для описания связи между входом и выходом, мы будем говорить позже. Остается также открытым вопрос, существует ли алгоритм, позволяющий за приемлемое время выбрать из этого набора функций ту, которая достаточно хорошо описывает связь между входом и выходом, даже если таковая заведомо есть.

Модели

В еще более общем случае мы будем рассматривать восстановление моделей по эмпирическим данным. (Как частный случай, восстановление функциональных зависимостей может тоже рассматриваться как восстановление модели).

Например, модель может описываться системой дифференциальных уравнений. Обучающая выборка состоит из последовательности начальных значений $\mathbf{x}_i$ и реализаций $\mathbf{y}_i$ поведения системы при этих начальных условиях. Или же $\mathbf{x}_i$ — это управляющие функции, подава-

емые на вход системы, а $\mathbf{y}_i$ — это реакции системы на эти входные воздействия (входной сигнал и выход могут быть как скалярными функциями, так и векторными). Требуется настроить (или построить) систему уравнений так, чтобы на новые начальные условия или внешние воздействия модель реагировала бы так же, как и реальная система.

Или рассмотрим такой необычный пример. На вход подается текст. Человек умеет по этому тексту составить граф, описывающий содержание этого текста. Последовательность пар $(\mathbf{x}_i, \mathbf{y}_i)$, каждая из которых состоит из текста $\mathbf{x}_i$ и соответствующего графа $\mathbf{y}_i$, составляет обучающую выборку. Требуется построить такую модель (обучить машину), чтобы она строила графы по тексту так же, как это делает человек.

Обратные задачи

К этому же кругу мы отнесем и так называемые обратные задачи. Типичный пример таков. Пусть функция $F(z)$ связана с другой функцией $f(x)$ интегральным уравнением

$$F(z) = \int K(x, z) f(x)\, dx \tag{1}$$

с известным ядром $K(x, z)$. Наблюдаются значения $y_i = F(z_i)$. Требуется же оценить функцию $f(x)$.

Такие задачи возникают при интерпретации физических измерений, но также встречаются в *социологии* и *демографии*.

Целый ряд практических задач состоит в том, что в интегральном уравнении (1) неизвестным считается не функция $f(x)$, а ядро $K(x, z)$. В этом случае реализации функций вполне наблюдаемы (и часто задаются самим исследователем). Но теперь задача оказывается более сложной, поскольку ищется функция двух аргументов (или двух групп аргументов). Поэтому для оценки ядра $K(x, z)$ приходится пронаблюдать реакции $F(z)$ при различных реализациях входной функции $f(x)$.

К этому типу задач относится томография. Тело просвечивают рентгеновскими лучами, проходящими под различными углами, параметры которых описываются функцией $f(x)$. Наблюдается плоская картина, задаваемая как $F(z)$. Требуется оценить ядро $K(x, z)$, описывающее трехмерное распределение оптической плотности (в рентгеновских лучах) внутри тела пациента. Другая подобная задача — это сейсморазведка. Здесь на поверхности земли закладывается серия

взрывов, интенсивность и расположение которых описывается как $f(x)$. Эхо, приходящее из недр земли, опять фиксируется в ряде точек и описывается как $F(z)$. Требуется оценить ядро $K(x,z)$, которое должно определить распределение акустических свойств в недрах.

Задача несколько упрощается, если интегральное уравнение имеет вид свертки:

$$F(z) = \int K(x-z)f(x)\,dx.$$

Здесь ядро и подынтегральная функция оказываются одного порядка сложности. Такая задача возникает при идентификации динамических систем, где ядро имеет смысл переходной функции, или при интерпретации спектров. Здесь $f(x)$ описывает истинный спектр, $F(z)$ — наблюдаемый спектр, а ядро $K(x-z)$ — размытие спектра измерительным прибором или в силу иных физических причин.

Обратные задачи часто относятся к числу так называемых некорректно поставленных, когда при точном решении системы линейных уравнений, соответствующих интегральному уравнению, сколь угодно малые отклонения в наблюдаемой функции приводят к сколь угодно большим отклонениям оценки. Измерения же всегда проводятся с некоторой погрешностью. Поэтому здесь необходимы те или иные методы регуляризации.

Приведем пример обратной задачи в социологии. В декабре 1993 года в России одновременно проходили выборы в Государственную Думу и референдум по проекту Конституции. Каждый избиратель бросал в урну один бюллетень, где он голосовал за одну из партий (избирательных объединений), допущенных к голосованию, а другим бюллетенем он голосовал за или против проекта Конституции. Поэтому прямо установить, как электораты разных партий голосовали по проекту Конституции, было невозможно. Тем не менее, была поставлена задача по имеющимся официальным данным (числу голосов, поданных за различные партии и числу голосов, поданных за проект Конституции в каждом регионе), оценить долю электората каждой партии, проголосовавшую за Конституцию.

Начальная гипотеза при решении этой задачи состояла в том, что доля электората i-той партии p_i, высказавшаяся за Конституцию, не зависела от региона. Отклонения от этого правила рассматривались как «шум». Обозначив за N_{ij} число избирателей, проголосовавших в

j–том регионе за i-тую партию, а за N_j — число избирателей j-го региона, поддержавших проект, получаем переопределенную систему уравнений:

$$\sum_i N_{ij} p_i = N_j.$$

Всего данные имелись по 88 регионам РФ, а в выборах участвовало 12 партий (избирательных объединений). Таким образом, получилась сильно переопределенная система уравнений — 88 уравнений при 12 неизвестных. Применяя обычный метод наименьших квадратов, мы получили совершенно неприемлемое решение — некоторые из величин p_i оказались отрицательными, другие превысили 100%. Тогда была применена регуляризация уравнений, при которой все значения p_i стягивались к общей по России доле избирателей, высказавшихся за проект Конституции. Регуляризатор увеличивался до тех пор, пока все значения p_i не легли в диапазон от нуля до единицы. На первый взгляд решение казалось приемлемым, но при подстановке полученных значений обратно в уравнения оказалось, что отклонения предсказанных значений от фактических чрезмерно велики, в особенности в Автономных республиках и в «красном поясе» — группе регионов, где КПРФ и ЛДПР получили значительное большинство. Тогда было решено решить эту задачу по группам регионов раздельно, выделив центральную группу, Севера и Дальний Восток, Красный Пояс и Автономии. При этом по некоторым группам число уравнений стало даже меньше числа неизвестных. Но, опять-таки, применяя регуляризацию, задачу удалось решить. Оказалось, что основной дисбаланс вносил электорат ЛДПР. На Севере и Дальнем Востоке электорат ЛДПР составляли, видимо, в основном «державники», которые выступали против партий, поддерживавших правительство Ельцина, но считали необходимым принятие новой Конституции для сохранения державы. То же самое относится к автономиям, где за ЛДПР голосовали, видимо, в основном русские, считавшие принятие конституции необходимым для сохранения единства страны. В то же время в «Красном Поясе» за ЛДПР голосовал в основном «протестный» электорат, который выступал как против правительства Ельцина, так и против предложенной им Конституции. В остальных же регионах доля электората ЛДПР, поддержавшая проект Конституции, была такой же, как и в среднем по России — 50 на 50.

Итак, в случае восстановления моделей по эмпирическим данным мы также считаем, что заранее задаются некоторым набором (клас-

сом) моделей и на основании обучающей выборки выбирают из него такую модель, которая в том или ином смысле наиболее подходит для описания связи между входом и выходом.

Автоматы

Задачу восстановления зависимостей можно интерпретировать и в терминах имитации одного автомата другим. В этом случае имеется автомат A, на вход которого подаются входы $\mathbf{x}$, а на выходе наблюдаются реакции $\mathbf{y}$. На основании обучающей выборки, состоящей из пар $(\mathbf{x}_i, \mathbf{y}_i)$, требуется построить (или настроить) автомат B, поведение которого достаточно хорошо имитировало бы поведение исходного автомата A. Опять же мы считаем, что задача сводится к выбору такого автомата из заданного класса (загашника, мешка), который лучше всего подходит для описания поведения автомата A, наблюдаемого в обучающей выборке.

Три общих вопроса возникают применительно к любым алгоритмам обучения.

Первый. Есть ли в том классе (мешке, загашнике), из которого мы выбираем, хорошее решение?

Второй. Позволяет ли критерий, по которому мы считаем, что выбранное решение хорошо описывает связь между входом и выходом на данных обучения, надеяться, что и на новых данных эта модель будет хорошо работать?

Третий. Есть ли эффективный алгоритм выбора нужного решения из заданного класса? Дело в том, что общее число моделей в классе обычно бывает кобинаторно большим и простой перебор, конечно, оказывается неприемлемым.

Критерии

Какими же критериями нужно пользоваться для того, чтобы оценить, насколько хорошо наш выбор подходит для описания наблюдаемой связи между входом и выходом?

Эмпирический риск. Простейшим из них является критерий минимизации эмпирического риска. Пусть $Q(y, y^*)$ — штрафная функция за отклонение предсказания выходного значения y^*от истинного значения y. В случае восстановления функции это может быть квадрат

разности $(y - y^*)^2$, или модуль разности, или иная функция штрафа. В задачах распознавания образов $Q(y, y^*)$ может быть равна нулю в случае правильной классификации объекта и равна 1 при неправильном ответе. Но может использоваться и более сложная функция штрафа, учитывающая различную цену ошибок разного рода. В случае построения моделей $Q(y, y^*)$ может быть мерой отклонения фактической функции на выходе от функции, предсказанной на выходе выбранной моделью, в функциональном пространстве. Целью обучения бывает нахождение такого решения, которое минимизирует истинный риск, т. е. математическое ожидание функции штрафа:

$$\mathsf{E}Q(y, y^*) = \int Q(y, y^*)\, dP(x, y) = \int Q(y, F(x))\, dP(x, y),$$

где $F(x)$ — выбранная зависимость.

Эмпирическим риском называют среднее значение функции штрафа на данных, представленных в материале обучения:

$$\mathsf{E}^*Q(y, y^*) = \frac{1}{l}\sum_i (y_i, F(x_i)),$$

т. е. средний квадрат отклонения или среднее число ошибок и т.д. на обучающей выборке.

Исходя из того, что в силу закона больших чисел средневыборочное значение риска должно сходиться к истинному (математическому ожиданию), считают, что минимизация эмпирического риска — это хороший критерий для выбора искомой зависимости. В следующих лекциях мы расскажем, как этот критерий реализуется в различных алгоритмах обучения, а также покажем, что решение, выбранное по критерию минимума эмпирического риска, может оказаться очень далеким от оптимального. В дальнейшем мы покажем, что близость решения, выбранного по критерию минимума эмпирического риска, к оптимальному (в пределах класса) зависит от соотношения длины выборки и некоторым образом определенного объема класса моделей, из которого мы выбираем. Чем шире класс моделей при фиксированной длине выборки, тем более вероятно, что точка минимума эмпирического риска будет далека от точки, доставляющей минимум истинному риску. И наоборот, при фиксированном классе, чем длиннее выборка, тем ближе минимум эмпирического риска к минимуму истинного риска.

Метод максимального правдоподобия. Другой подход — это метод максимального правдоподобия. Здесь делается три важных допущения. Первое: среди группы моделей есть истинная, т. е. такая, которая правильно описывает связь между входом и выходом (но не в детерминированном виде, а как случайную). Второе: модель строится не в виде функциональной зависимости $\mathbf{y} = F(\mathbf{x})$, а позволяет вычислять условное распределение $P(\mathbf{y}|\mathbf{x})$. Третье: пары $(\mathbf{x}_i, \mathbf{y}_i)$, представленные в материале обучения, получены независимо при неизменном распределении $P(\mathbf{x}, \mathbf{y})$. Описания входа и выхода могут иметь любую природу (величины, векторы, изображения, графы), но в случае, когда $\mathbf{y}$ представляет собой непрерывную величину, $P(\mathbf{y}|\mathbf{x})$ понимается как условная плотность распределения, а в случае дискретных значений $P(\mathbf{y}|\mathbf{x})$ — это просто условное распределение. Тогда условная вероятность (или плотность вероятности) получить на выходе значения $y_1, y_2, \ldots, y_l$ при заданных значениях $x_1, x_2, \ldots, x_l$ на входе и фиксированной модели M будет равна

$$P(y_1, y_2, \ldots, y_l|M, x_1, x_2, \ldots, x_l) = \prod_i P(y_i|x_i),$$

где величины $P(y_i|x_i)$ конечно зависят от выбранной модели.

Функцией правдоподобия называют логарифм этой условной вероятности:

$$W(M) = \sum \log P(y_i|x_i).$$

Метод максимального правдоподобия состоит в том, что ищется такая модель, которая доставляет максимум этой функции, или иными словами такая, при которой упомянутая условная вероятность максимальна.

Можно показать, что если искать максимум математического ожидания функции правдоподобия, то он достигается действительно на истинной модели (кстати, даже на выборках длины 1). Но математическое ожидание нам не дано, и остается надеяться на то, что средневыборочное значение функции правдоподобия сходится к математическому ожиданию и выбираемая по этому критерию модель окажется близка к истинной. Опять же, есть примеры, когда выбираемая по этому критерию модель не приближается к истинной даже при неограниченном увеличении длины выборки, и в дальнейшем мы увидим почему.

Байесов подход. Байесов подход позволяет действительно получить в определенном смысле оптимальное решение. Но при этом помимо условий, которые предполагались для метода максимального правдоподобия (наличие истинной модели в классе и получение на выходе модели условного распределения), требуется задание априорного распределения вероятностей $P_{\text{арг}}(M)$ на множестве всех моделей, из которых мы выбираем. В случае дискретного множества моделей $P_{\text{арг}}(M)$ понимается как априорная вероятность каждой модели, а в случае, когда модель описывается непрерывно меняющимися параметрами, — как априорная плотность. Далее применяется известная формула Байеса для определения апостериорного распределения на множестве моделей:

$$P_{\text{апост}}(M) = \frac{P_{\text{апр}}(M) \cdot P(y_1, y_2, \ldots, y_l | M, x_1, x_2, \ldots, x_l)}{C},$$

где C — нормировочная константа, нормирующая распределение к единице. Теперь для всякого входного значения $\mathbf{x}$ мы можем найти усредненное по всем моделям апостериорное распределение выхода $\mathbf{y}$:

$$P_{\text{апост}}(\mathbf{y}) = \sum P(\mathbf{y}|\mathbf{x}, M) P_{\text{апост}}(M),$$

где сумма берется по всем моделям класса (хотя, в принципе, можно выделить подкласс, апостериорная вероятность которого близка к 1).В случае, когда модель описывается непрерывными параметрами, суммирование заменяется на интегрирование. В зависимости от того, что требуется, с этим апостериорным распределением можно делать разные вещи. Можно вычислить его математическое ожидание, получив апостериорное среднее выходного значения. Можно найти диапазон наиболее вероятных значений по любому порогу. Можно определить вероятность катастрофических значений. Однако есть две неприятности. Неизвестно, откуда взять априорное распределение, и в скольнибудь сложных случаях усреднение по всем моделям приводит к невероятным вычислительным сложностям. Впрочем, в частных случаях получаются красивые решения.

Есть категория статистиков, называемых «байесистами», которые считают, что только байесов подход является по-настоящему научным, а все остальное — импровизация. Но выбор априорного распределения из общих соображений трудно назвать научным.

Упорядочение. Как уже отмечалось, чем у́же класс моделей, из которого мы выбираем, тем ближе оказывается точка, в которой достигается минимум эмпирического риска к точке, доставляющей минимум истинному риску (в пределах этого класса). С другой стороны, чем уже класс, тем меньше шансов, что в нем окажется истинная модель или близкая к ней. Поэтому возникает такая идея: рассмотреть последовательность вложенных классов все возрастающих объемов, и выбирать класс оптимального объема в зависимости от имеющегося материала обучения. Например, можно последовательно увеличивать (или уменьшать) степень аппроксимирующего полинома, или увеличивать число членов в разложении по тригонометрическим (или иным) функциям. В случае поиска линейной зависимости можно заранее упорядочить аргументы и последовательно наращивать их число в этом порядке. Или перебирать все комбинации с уменьшенным числом аргументов. В случае кусочно-линейной аппроксимации — менять число кусков или нейронов в нейронной сети.

Как это сделать, как найти оптимальный объем? Можно пойти простым путем: зарезервировать часть данных обучения в качестве контрольной выборки (validation set), по оставшейся части строить модели в расширяющихся классах и проверять результат на контрольной выборке. Наконец, выбрать ту модель, которая окажется лучшей на контрольной выборке. Но на самом деле обучающих данных почти всегда не хватает, и резервировать ее часть оказывается накладно. Кроме того, при выборе оптимального объема класса контрольная выборка уже участвует в обучении, и результат опять оказывается смещенным (особенно при больших переборах). Можно использовать такие методы контроля, как скользящий контроль или cross-validation, которые не требуют резервирования. При скользящем контроле один объект изымается из обучающей выборки, по остальным строится модель и результат проверяется на этом выделенном объекте. Затем он возвращается в материал обучения, изымается другой объект, и так последовательно проверяются все объекты выборки. Средний риск, полученный на выделяемых объектах, принимается за оценку качества класса, из которого осуществлялся выбор. Но при этом сохраняются значительные погрешности в оценке, и выбор оптимального объема оказывается в значительной мере случайным.

Другой путь состоит в том, чтобы аналитически оценить, насколько эмпирический риск будет отличаться от истинного в зависимости от

имеющихся данных обучения и объема класса, из которого мы выбираем модель. Дело в том, что с расширением класса эмпирический риск всегда уменьшается (не возрастает), тогда как истинный риск, пройдя через минимум, начинает расти. Аналитически получается необходимая добавка в аддитивной или мультипликативной форме. Существуют и другие методы выбора оптимального объема (оптимальной сложности) [28, 32, 26, 2] (Rissanen, Volas, MacKay), но обо всем этом пойдет речь в конце нашего курса. Между прочим, увеличение объема класса не всегда связано с увеличением сложности модели. Например, может осуществляться сложная, но фиксированная предварительная обработка исходных признаков, а решающее правило выбирается в узком классе уже в пространстве признаков, полученных в результате обработки.

Лекция 2

Определение вероятностной меры. Случайные величины, их функции распределения, моменты. Суммы случайных величин. Закон больших чисел. (Стандартные статистические пакеты: вычисление среднего, дисперсии, ковариации, корреляции и т.д. и погрешности их оценивания).

Вероятностная мера

При аксиоматическом подходе к определению теории вероятностей (по Колмогорову) [12, 13] исходят из того, что имеется некоторое вероятностное пространство U, состоящее из *элементарных событий* ω. Это пространство может быть конечным, счетным или иметь мощность континуума, например, образовывать собой конечномерное евклидово пространство. Но этим пространством может быть и множество функций одной или многих переменных, как в случае случайных процессов или случайных полей. Тогда вероятностное пространство будет уже бесконечномерным.

Далее рассматривается некоторая система Q подмножеств вероятностного пространства U, называемых *случайными событиями*. Таким образом, случайными событиями будут множества элементарных событий, принадлежащие системе Q. Например, если пространство U представляет собой прямую, а элементарными событиями будут точки на этой прямой, то случайными событиями будут все множества точек на

прямой, которые объявлены случайными событиями. Предполагается, что для системы событий выполнены следующие аксиомы:

1. Все множество элементарных событий U принадлежит системе Q.

2. Если некоторое множество A принадлежит системе Q, то есть является случайным событием, то и его дополнение в U тоже принадлежит системе Q, то есть тоже будет случайным событием. В частности, пустое множество, будучи дополнением U в U, обязательно является случайным событием.

3. Если два множества A и B принадлежат системе Q, то их объединение и их пересечение тоже принадлежат системе, то есть будут случайными событиями.

4. И, наконец, очень важная аксиома. Если бесконечная последовательность множеств $A_1, A_2, \ldots, A_n, \ldots$ образована случайными событиями, то их (бесконечное) объединение и пересечение тоже будут случайными событиями, то есть принадлежат системе Q. Конечно, эта аксиома существенна только в том случае, когда само множество элементарных событий бесконечно.

Система подмножеств Q, удовлетворяющая этим аксиомам, называется борелевским полем событий, или *σ-алгеброй событий*. Нетрудно видеть, что система Q, состоящая из всех подмножеств пространства U, удовлетворяет всем перечисленным аксиомам. Поэтому по любой системе подмножеств пространства элементарных событий S можно построить минимальную систему, содержащую S и удовлетворяющую всем аксиомам σ-алгебры. Для этого достаточно взять пересечение всех систем, содержащих S и удовлетворяющих всем аксиомам σ-алгебры. Например, на прямой можно взять за основу систему, состоящую из всех интервалов, и далее дополнить до σ-алгебры. Туда войдут все открытые и замкнутые множества, их объединения и пересечения (конечные и счетные) и так далее. Так строится обычная борелевская σ-алгебра на прямой.

Так же строится σ-алгебра на произведении двух случайных пространств U_1 и U_2. Произведением двух пространств называется множество пар элементов, первый из которых принадлежит U_1, а второй — U_2. Пусть теперь на каждом из них уже задана своя σ-алгебра. Тогда за основу берутся прямые произведения множеств из этих двух

σ-алгебр, и дополняется до минимального множества, удовлетворяющего всем аксиомам. Аналогично строится σ-алгебра на произведении любого конечного или счетного числа вероятностных пространств.

Вероятность определяется как неотрицательная мера на множестве событий. Она должна удовлетворять следующим аксиомам:

1. Каждому событию A из поля событий Q поставлено в соответствие неотрицательное действительное число $P(A)$, называемое его вероятностью.

2. Вероятность всего вероятностного пространства равна 1, т. е. $P(U) = 1$.

3. Если события $A_1, A_2, \ldots, A_n$ попарно не пересекаются, то вероятность их объединения равна сумме их вероятностей. Отсюда, в частности, следует, что вероятность дополнения B к событию A равна $P(B) = 1 - P(A)$, а вероятность пустого события равна нулю.

4. Счетная аддитивность (она же расширенная аксиома сложения, она же аксиома непрерывности). Вероятность объединения последовательности попарно непересекающихся событий равна пределу конечных сумм вероятностей этих событий, то есть

$$P(A_1 + A_2 + \ldots + A_n + \ldots) = \lim_{n\to\infty} P(A_1) + P(A_2) + \ldots + P(A_n).$$

Таким образом, вероятностное пространство задается тройкой (U, Q, P) — множеством элементарных событий U, системой событий Q и счетно-аддитивной вероятностной мерой P на множестве этих событий.

Случайные величины

Случайной величиной называется числовая функция элементарного случайного события $X(\omega)$, измеримая относительно заданной σ-алгебры случайных событий Q. Измеримой считается такая функция $X(\omega)$, что для всякого z множество элементарных событий

$$T_z = \{\omega : X(\omega) < z\},$$

является случайным событием, то есть принадлежит системе Q, и следовательно, для него определена вероятность $P(T_z)$. Функция $F(z) = P(T_z) = P(X(\omega) < z)$ называется кумулятивной функцией распределения случайной величины X (или просто функцией распределения случайной величины). Это неубывающая функция, непрерывная слева, такая, что

$$F(-\infty) = 0; \quad F(+\infty) = 1.$$

Если существует такая функция $p(y)$, что $F(z) = \int_{-\infty}^{z} p(y)\,dy$, то $p(y)$ называется плотностью распределения вероятностей случайной величины X.

Можно показать, что если две функции $X(\omega)$ и $Y(\omega)$ измеримы относительно заданной σ-алгебры (то есть являются случайными величинами), то их произведение и их сумма тоже будут измеримыми функциями (случайными величинами). Понятно, что это утверждение далее распространяется на произведения и суммы конечного числа случайных величин.

Существует очень простой способ построить неизмеримую функцию. Рассмотрим сначала обычную борелевскую σ-алгебру на сегменте $R_1 = [0, 1]$. Рассмотрим далее квадрат Z — прямое произведение нашего сегмента R_1 на другой сегмент $R_2 = [0, 1]$. На квадрате определим новую σ-алгебру как множество прямых произведений элементов исходной σ-алгебры на сегмент R_2 — то есть исходные элементы вытягиваются вдоль вертикальной оси от края до края. Можно проверить, что все аксиомы при этом будут выполнены. Теперь любая функция $f(x, y)$ на квадрате, которая меняется вдоль оси Y при фиксированном значении x, будет неизмеримой относительно этой σ-алгебры, так множества вида $T_z = \{(x, y) : f(x, y) \leqslant z\}$ не всегда принадлежат этой алгебре — некоторые из них разрезают вертикальную ось. Функции же, которые зависят только от x, могут быть измеримыми.

На этом принципе построено строгое определение условного математического ожидания, о чем мы будем говорить позже.

Об одном парадоксе. Существует понятие вполне упорядоченного множества. Порядок, заданный на некотором множестве A, считается полным, если любое его непустое подмножество имеет минимальный элемент, то есть A в найдется такой элемент x_0, что любой другой элемент из A будет в этом порядке больше чем x_0. Такой порядок будет линейным, поскольку любые два элемента сравнимы. Примером вполне

упорядоченного множества может служить ряд натуральных чисел в естественном порядке, или множество полиномов с целыми коэффициентами, если коэффициент при старшей степени положителен.

Оказывается, что любые два вполне упорядоченных множества сравнимы, в том смысле, что либо они изоморфны, либо одно из них изоморфно какому то началу другого, то есть подмножеству

$$T_a = \{x : x < a\}.$$

Поэтому любые два полных порядка оказываются сравнимыми в этом смысле, и эти порядки образуют шкалу так называемых трансфинитных чисел. Множество всех трансфинитных чисел, меньших заданного, также будет вполне упорядоченным. Но попытка рассмотреть множество всех трансфинитных чисел немедленно приводит к одному из парадоксов теории множеств. (Это обстоятельство свело с ума создателя теории множеств Кантора).

Известна теорема Цермело, утверждающая, что на любом множестве можно ввести полный порядок. (Правда, эта теорема опирается на так называемую аксиому выбора, которая не всегда принимается при аксиоматизации теории множеств). В частности, если в качестве множества взять обычный сегмент $[0, 1]$, то на нем тоже можно ввести полный порядок. Отсюда следует, что существует минимальное трансфинитное число, обладающее тем свойством, что соответствующее ему вполне упорядоченное множество имеет мощность континуума. Это трансфинитное число называют ω_1. И, более того, на сегменте $[0, 1]$ можно ввести порядок, соответствующий именно этому трансфинитному числу ω_1. (Этот порядок, конечно, не совпадает с обычным порядком по возрастанию чисел). Тогда любое начало такого вполне упорядоченного множества, то есть множество вида $T_a = \{x : x < a\}$, где a любая фиксированная точка сегмента, будет иметь мощность меньше континуума.

Далее, известна знаменитая *континуум-гипотеза*, утверждающая, что между счетной мощностью и мощностью континуума нет других мощностей. Иными словами, всякое подмножество континуума либо пусто, либо конечно, либо счетно, либо имеет мощность континуума. Над этой проблемой долго бились самые сильные математики мира, но в конце концов было доказано, что в определенной аксиоматике теории множеств эта гипотеза не доказуема и не опровержима.

Если все-таки принять континуум-гипотезу, то окажется, что при нашем упорядочении сегмента $[0,1]$ все подмножества вида $T_a = \{x : x < a\}$ будут пустыми, конечными или счетными.

Итак, будем считать, что все точки сегмента $[0,1]$ упорядочены в соответствии с порядковым трансфинитным числом ω_1. Рассмотрим на этом отрезке обычную борелевскую σ-алгебру и равномерное распределение на нем. Все конечные и счетные множества на нем измеримы и имеют вероятность нуль. Поэтому отношение $x < a$ измеримо и имеет вероятность нуль. Но будет ли измеримым отношение $x < y$ на прямом произведении сегментов $[0,1]$? С одной стороны, фиксируя любое значение y, получим $P\{x < y\} = 0$. Но точно так же, фиксируя x, получим и $P\{y < x\} = 0$, что невозможно. Поэтому отношение $x < y$ не может быть измеримым.

Математическое ожидание

Определим сначала математическое ожидание для неотрицательной случайной величины $X(\omega)$. В этом случае функция распределения $F(z)$ будет тождественно равна нулю при $z \leqslant 0$ и далее, не убывая, стремится к 1. Функция же $1 - F(z)$ в области $z \geqslant 0$ монотонно стремится к нулю, начиная с $F(0) = 1$. Если случайная величина ограничена сверху, то функция $1 - F(z)$ обратится в нуль уже при конечном значении z, в противном случае $1 - F(z)$ может (но не обязательно) нигде не обращаться в нуль, а только стремится к нулю при $z \to \infty$. Математическим ожиданием неотрицательной случайной величины X называется интеграл:

$$EX = \int_0^{+\infty} (1 - F(z))\, dz.$$

Этот интеграл понимается как обычный интеграл по Риману. Если случайная величина ограничена, то этот интеграл всегда существует. Если же это не так, то интеграл понимается как предел

$$EX = \lim_{t \to \infty} \int_0^t (1 - F(z))\, dz.$$

Этот предел может быть бесконечным, то есть интеграл может расходиться, и тогда считают, что математическое ожидание случайной величины не существует.

Нетрудно заметить, что наш интеграл выражает площадь под кривой $1 - F(z)$ в области $z \geqslant 0$. Заменяя переменные интегрирования (интегрируя по частям), получим также

$$EX = \int_0^1 z\, dF(z).$$

В общем случае, когда случайная величина может принимать и отрицательные значения, случайную величину представляют в виде разности двух неотрицательных случайных величин:

$$X(\omega) = X^+(\omega) - X^-(\omega),$$

где

$$X^+(\omega) = \begin{cases} X(\omega), & X(\omega) \geqslant 0 \\ 0, & X(\omega) < 0 \end{cases}, \qquad X^-(\omega) = \begin{cases} -X(\omega), & X(\omega) < 0 \\ 0, & X(\omega) \geqslant 0 \end{cases}$$

Нетрудно убедиться, что функции распределения этих случайных величин будут следующими:

для $X^+(\omega)$:

$$F_1(z) = \begin{cases} P\omega :\, X^+(\omega) \leqslant z = F(z), & z \geqslant 0, \\ 0, z < 0. \end{cases}$$

Для $X^-(\omega)$:

$$F_2(z) = \begin{cases} P\omega :\, X^-(\omega) \leqslant z = 1 - F(-z), & z > 0, \\ 0, z \leqslant 0. \end{cases}$$

Поэтому их математические ожидания будут равны:

$$EX^+(\omega) = \int_0^{+\infty} (1 - F(z))\, dz,$$

$$EX^-(\omega) = \int_{-\infty}^0 F(z)\, dz$$

Математическое ожидание самой случайной величины $X(\omega)$ определяется, как разность

$$EX(\omega) = EX^+(\omega) - EX^-(\omega) = \int\limits_0^{+\infty} (1 - F(z))\, dz - \int\limits_{-\infty}^{0} F(z)\, dz.$$

Считается, что математическое ожидание величины $X(\omega)$ существует, если существуют математические ожидания каждой из величин $X^+(\omega)$ и $X^-(\omega)$, т. е. оба соответствующих интеграла сходятся.

Опять, заменяя переменные интегрирования, получим также

$$EX = \int\limits_0^1 z\, dF(z).$$

Математическое ожидание может и не существовать. Например, случайная величина с функцией распределения $F(z) = (1/\pi)\operatorname{arctg} z + 1/2$ не имеет математического ожидания.

Моменты высших порядков

Математическое ожидание случайной величины называется также ее моментом первого порядка.

Могут существовать (или не существовать) также моменты высших порядков:

$$EX^n = M_n = \int\limits_{-\infty}^{+\infty} z^n\, dF(z).$$

Если существует момент порядка n, то существуют и моменты всех низших порядков. Если случайная величина ограничена, то существуют моменты всех порядков.

Рассматриваются также центральные моменты

$$M_n^* = \int\limits_{-\infty}^{+\infty} (z - M_1)^n\, dF(z).$$

Они всегда существуют, если существует соответствующий не центральный момент.

Среди них особую роль играет центральный момент второго порядка, называемый дисперсией:

$$D(X) = M_2^* = \int_{-\infty}^{+\infty} (z - M_1)^2 \, dF(z).$$

Дисперсия характеризует рассеяние случайной величины вокруг ее среднего. Справедливо соотношение

$$D = M_2^* = M_2 - M_1^2,$$

в котором нетрудно убедиться непосредственно.

При добавлении к случайной величине константы a математическое ожидание сдвигается на величину a, а дисперсия не меняется:

$$E(X + a) = E(X) + a, \qquad D(X + a) = D(X).$$

При умножении случайной величины на константу b математическое ожидание также умножается на эту константу, а дисперсия умножается на b^2:

$$E(bX) = bE(X); \qquad D(bX) = b^2 D(X).$$

Поскольку дисперсия меняется нелинейно с изменением масштаба, то рассеяние случайной величины лучше характеризовать не дисперсией, а так называемым среднеквадратичным отклонением $\sigma = \sqrt{D}$. Значение σ при умножении случайной величины на константу также умножается на эту константу.

Если даны две случайны величины $X(\omega)$ и $Y(\omega)$, то их совместная (кумулятивная) функция распределения определяется как

$$F(x, y) = P\{\omega : X(\omega) \leqslant x, Y(\omega) \leqslant y\}.$$

Вероятность любого события A (измеримого множества в пространстве этих двух переменных) определяется как

$$P(A) = \int I(x, y) \, dF(x, y),$$

где $I(x, y)$ — индикаторная функция множества A. В частности, маргинальные распределения $F_1(x)$ и $F_2(y)$ определяются как

$$F_1(x) = P\{X < x\}, \quad F_2(y) = P\{Y < y\}.$$

Если случайные величины независимы, то

$$F(x, y) = F_1(x)F_2(y).$$

Аналогично определяется и совместное распределение n случайных величин (n-мерного вектора).

Большую роль в статистике играют парные центральные моменты двух случайных величин

$$M_2(X, Y) = E(X - M_X)(Y - M_Y).$$

Эти моменты называются ковариацией $\mathbf{cov}_{xy}$ случайных величин X и Y.

Ковариация существует тогда и только тогда, когда существуют дисперсии D_x и D_y. Более того, для нее выполняется неравенство Коши-Буняковского:

$$\mathbf{cov}_{xy}^2 \leqslant D_x D_y.$$

Вообще центрированные случайные величины с конечной дисперсией можно рассматривать как векторы в евклидовом пространстве со скалярным произведением $\mathbf{cov}_{xy}$ и нормой $\sigma = \sqrt{D}$. Косинус угла между этими векторами будет равен

$$\frac{\mathbf{cov}_{xy}}{\sqrt{D_x D_y}}$$

и называется коэффициентом корреляции $\mathbf{cor}_{xy}$ между случайными величинами X и Y. Коэффициент корреляции характеризует степень сходства случайных величин безотосительно к их масштабу.

Матрица, составленная из попарных ковариаций n случайных величин, будет матрицей Грама этих векторов и, потому, всегда оказывается симметрической и положительно полуопределенной, т. е. ее собственные векторы ортогональны, а ее собственные числа неотрицательны. Легко видеть, что ковариация двух независимых случайных величин всегда равна нулю. Обратное, вообще говоря, неверно.

Две величины, корреляция (ковариация) которых равна 0, называются некоррелированными.

Математическое ожидание суммы двух случайных величин всегда равно сумме их математических ожиданий:

$$E(X + Y) = E(x) + E(y),$$

а дисперсия суммы равна

$$D(X+Y) = D(X) + D(Y) + 2\,\mathbf{cov}_{xy}\,.$$

Поэтому дисперсия суммы некоррелированных (в частности, независимых) величин просто равна сумме дисперсий. То же самое относится к сумме n случайных величин. Математическое ожидание суммы равно сумме математических ожиданий каждой из них, а дисперсия суммы некоррелированных (в частности, независимых) величин равна сумме их дисперсий. Если все эти случайные величины имеют одинаковое математическое ожидание M и одинаковую дисперсию D, то матожидание суммы равно nM, а дисперсия (в случае отсутствия корреляции) равна nD.

Закон больших чисел по Чебышеву

Закон больших чисел утверждает, что среднее арифметическое последовательности независимых одинаково распределенных случайных величин с вероятностью 1 стремится к их математическому ожиданию (если оно существует).

В случае, когда известна дисперсия D этих величин, для всякого ε можно оценить сверху вероятность того, что среднее уклонится от математического ожидания по модулю более чем на ε. При этом нет надобности требовать, что эти величины одинаково распределены — достаточно предположить, что они имеют одинаковое среднее и дисперсию. Нет надобности и требовать их независимости — достаточно потребовать некоррелированности.

Действительно, сумма n не коррелированных случайных величин $x_1, x_2, \ldots x_n$, с одинаковым математическим ожиданием M и дисперсией D, будет иметь математическое ожидание nM и дисперсию nD. А их среднее арифметическое $M^* = (1/n)\sum_{i=1}^{n} x_i$ получит математическое ожидание $\mathbf{M} = M$ и дисперсию $\mathbf{D} = D/n$.

Далее для вывода закона больших чисел в форме Чебышева нам понадобится неравенство Чебышева.

Неравенство Чебышева. Пусть дана случайная величина x с математическим ожиданием M и дисперсией D. Тогда для всякого положительного ε вероятность того, что величина $|x - M|$ превысит ε ,

удовлетворяет неравенству

$$P(|x - M| > \varepsilon) \leqslant \frac{D}{\varepsilon^2}.$$

Действительно, дисперсия равна

$$D = \int_{-\infty}^{+\infty} (x - M)^2 \, dP(x),$$

и, ограничивая область интегрирования, получим

$$D \geqslant \int_{-\infty}^{M-\varepsilon} (x - M)^2 \, dP(x) + \int_{M+\varepsilon}^{\infty} (x - M)^2 \, dP(x).$$

В свою очередь подынтегральная функция в диапазонах $(-\infty, M - \varepsilon)$ и $(M + \varepsilon, \infty)$ не меньше чем ε^2. Поэтому

$$D \geqslant \varepsilon^2 \left[\int_{-\infty}^{M-\varepsilon} dP(x) + \int_{M+\varepsilon}^{\infty} dP(x) \right] = \varepsilon^2 P(|x - M| > \varepsilon),$$

откуда

$$P(|x - M| > \varepsilon) \leqslant \frac{D}{\varepsilon^2},$$

что и требуется.

Теперь применим неравенство Чебышева к нашему среднему арифметическому M^* некоррелированных случайных величин. Подставляя вместо $\mathbf{D} = D/n$ и учитывая, что $\mathbf{M} = M$, получим

$$P((M^* - M) > \varepsilon) \leqslant \frac{D}{n\varepsilon^2}.$$

Таким образом, для любого заданного $\varepsilon > 0$ вероятность того, что среднее уклонится по модулю от математического ожидания более чем на ε, стремится к нулю с ростом длины выборки n.

Посмотрим, какова длина выборки, которая с вероятностью P_0 гарантирует, что среднее арифметическое уклонится от математического ожидания не более, чем на ε. Для этого достаточно разрешить неравенство относительно n.

Получаем, что это так при всяком $n > n_0$, где

$$n_0 = \frac{D}{P_0 \varepsilon^2}.$$

Видим, что n_0 растет обратно пропорционально P_0 и обратно пропорционально ε^2.

Неравенство Чебышева принципиально не улучшаемо, если ничего не известно про случайную величину кроме ее дисперсии. Рассмотрим такой пример. Пусть случайная величина x может принимать только три значения: $-a$ и a $(a > 0)$ с одинаковыми вероятностями $p/2$ и ноль с вероятностью $1 - p$. Тогда ее математическое ожидание M равно 0, а дисперсия $D = a^2 p$, откуда $p = D/a^2$. Положим теперь $\varepsilon = a - \delta$, где δ — малая положительная величина. Тогда вероятность того, что случайная величина $|x - M|$ превзойдет ε, равна p (действительно, это возможно только, если $x = -a$ или $x = a$). То есть

$$P(|x - M| > \varepsilon) = p = \frac{D}{a^2} = \frac{D}{(\varepsilon + \delta)^2}.$$

Поскольку величину δ можно выбрать сколь угодно малой, нижняя грань неравенства Чебышева достигается сколь угодно точно.

Лекция 3

Закон больших чисел в форме Линдберга. Сравнение с результатом по Чебышеву. Свойства ковариационной матрицы. Плотность распределения вероятностей случайной величины и группы случайных величин. Метод максимального правдоподобия.

Центральная предельная теорема (в форме Линдберга)

Эта теорема [20] позволяет увидеть, как асимптотически распределено отклонение среднего арифметического от математического ожидания, если среднее вычисляется по последовательности n независимых одинаково распределенных случайных величин, имеющих математическое ожидание M и дисперсию $D = \sigma^2$. Мы уже знаем, что средневыборочное значение имеет то же математическое ожидание M и дисперсию D/n, или $\sigma_{\text{ср}} = \sigma/\sqrt{n}$. Таким образом, распределение средневыборочного стягивается к математическому ожиданию, а отклонение между

ними стягивается к нулю. Но если растянуть последнюю величину, разделив ее на $\sigma_{\text{ср}} = \sigma/\sqrt{n}$, то распределение полученной величины сходится к нормальному распределению с нулевым средним и единичной дисперсией. А именно, справедлива теорема:

$$\lim_{n\to\infty} P\left(\frac{\frac{1}{n}\sum x_i - M}{\sigma_{\text{ср}}} > \varepsilon\right) = \frac{1}{\sqrt{2\pi}} \int\limits_{\varepsilon}^{\infty} \exp\left(-\frac{1}{2}y^2\right)\, dy.$$

Заметим, что речь здесь идет о сходимости кумулятивной функции распределения к кумулятивной функции нормального распределения (а не о плотности распределения). Попробуем, однако, применить эту теорему для оценки вероятности того, что отклонение среднего от математического ожидания превзойдет ε (без нормировки на $\sigma_{\text{ср}}$), применяя асимптотическую формулу.

$$P\left(\frac{1}{n}\sum x_i - M > \varepsilon\right) = P\left[\left(\frac{\frac{1}{n}\sum x_i - M}{\sigma_{\text{ср}}}\right) > \frac{\sqrt{n\varepsilon}}{\sigma}\right] \approx$$

$$\approx \frac{1}{\sqrt{2\pi}} \int\limits_{a}^{\infty} \exp\left(-\frac{1}{2}y^2\right)\, dy,$$

где $a = (\sqrt{n\varepsilon})/\sigma$.

Интеграл в последней части равенства не берется в элементарных функциях, но при больших значениях a может быть достаточно точно оценен. Проведем замену переменных $z = y - a$, $y = z + a$:

$$\frac{1}{\sqrt{2\pi}} \int\limits_{a}^{\infty} \exp\left(-\frac{1}{2}y^2\right)\, dy = \frac{1}{\sqrt{2\pi}} \int\limits_{0}^{\infty} \exp\left(-\frac{1}{2}(z+a)^2\right)\, dz =$$

$$= \frac{1}{\sqrt{2\pi}} \int\limits_{0}^{\infty} \exp\left(-\frac{1}{2}(z^2 + 2az + a^2)\right)\, dz =$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}a^2\right) \int\limits_{0}^{\infty} \exp\left(-\frac{1}{2}(z^2 + 2az)\right)\, dz \approx$$

$$\approx \frac{1}{a\sqrt{2\pi}} \exp\left(-\frac{1}{2}a^2\right).$$

Подставляя сюда значение $a = (\sqrt{n}\varepsilon)/\sigma$, получим

$$P\left(\frac{1}{n}\sum x_i - M > \varepsilon\right) \approx \frac{1}{\sqrt{2\pi n}} \cdot \frac{\sigma}{\varepsilon} \exp\left(-\frac{1}{2} \cdot \frac{n\varepsilon^2}{\sigma^2}\right).$$

Такая же оценка получается для $P\left(((1/n)\sum x_i - M) < -\varepsilon\right)$.

Мы видим, что оценка вероятности уклонения стремится к нулю экспоненциально с ростом n. В то же время согласно неулучшаемой оценке Чебышева $P((M^* - M) > \varepsilon) \leqslant D/(n\varepsilon^2) = \sigma^2/(n\varepsilon^2)$, которая убывает всего лишь обратно пропорционально n. Как разрешить это противоречие?

Оказывается, что при оценке вероятности больших уклонений центральная предельная теорема не работает. Теорема справедлива для любого фиксированного значения y, что до нормировки на $\sigma_{\text{ср}} = \sigma/\sqrt{n}$ соответствует непрерывно убывающему порогу ε. Зачем вообще нужно оценивать вероятности таких больших отклонений? Это мы увидим позже.

Свойства ковариационной матрицы

Пусть даны n случайных величин. Рассмотрим их ковариационную матрицу, то есть матрицу, составленную из их попарных коэффициентов ковариации:

$$a_{ij} = \mathbf{cov}(X_i, X_j) = E(X_i - M_{x_i})(X_j - M_{x_j}).$$

Как уже говорилось, центрированные случайные величины можно рассматривать как векторы в евклидовом пространстве, а коэффициенты их ковариации - как их скалярные произведения. Таким образом, ковариационная матрица будет составлена из скалярных произведений этих векторов, т. е. представляет собой матрицу Грама. Эта матрица будет квадратной и симметрической, так как скалярное произведение симметрично.

Известно, что матрица Грама положительно полуопределена. Мы сейчас покажем это. Но сначала я напомню, что значит, что симметрическая матрица положительно определена или положительно полуопределена. Это будет так, если соответствующая квадратичная форма будет соответственно положительно определена или положительно

полуопределена. Свойства таких матриц и квадратичных форм неоднократно понадобятся нам в дальнейшем.

Функция двух векторов $A(\mathbf{x}, \mathbf{y})$ называется билинейной формой, если она линейна по каждому из аргументов при фиксированном другом аргументе:

$$\begin{aligned} &A(\mathbf{x}_1 + \mathbf{x}_2, \mathbf{y}) = A(\mathbf{x}_1, \mathbf{y}) + A(\mathbf{x}_2, \mathbf{y}), \\ &A(\lambda\mathbf{x}, \mathbf{y}) = \lambda A(\mathbf{x}, \mathbf{y}), \\ &A(\mathbf{x}, \mathbf{y}_1 + \mathbf{y}_2) = A(\mathbf{x}, \mathbf{y}_1) + A(\mathbf{x}, \mathbf{y}_2), \\ &A(\mathbf{x}, \lambda\mathbf{y}) = \lambda A(\mathbf{x}, \mathbf{y}). \end{aligned}$$

Отсюда следует, что $A(\mathbf{x}, 0) = A(0, \mathbf{y}) = 0$. Но, конечно, билинейная форма может обращаться в нуль и в тех случаях, когда ни вектор $\mathbf{x}$, ни вектор $\mathbf{y}$ не равен нулю. Билинейная форма называется симметричной, если $A(\mathbf{x}, \mathbf{y}) = A(\mathbf{y}, \mathbf{x})$.

Квадратичная форма $A(\mathbf{x}, \mathbf{x})$ получатся из билинейной симметрической формы, если положить $\mathbf{x} = \mathbf{y}$. Квадратичная форма $A(\mathbf{x}, \mathbf{x})$ называется *положительно полуопределенной*, если для любого $\mathbf{x}$ ее значение неотрицательно, т. е.

$$A(\mathbf{x}, \mathbf{x}) \geqslant 0.$$

Если же для любого отличного от нуля вектора значение квадратичной формы строго положительно, то эта форма называется *положительно определенной*. В частности, скалярное произведение $(\mathbf{x}, \mathbf{y})$ будет билинейной формой, а соответствующая ему квадратичная форма $(\mathbf{x}, \mathbf{x})$ положительно определена, поскольку скалярное произведение $(\mathbf{x}, \mathbf{x}) = |\mathbf{x}|^2$ отлично от нуля для всех векторов, не равных нулю. Справедливо и обратное утверждение: всякая симметрическая билинейная форма, может служить скалярным произведением, если соответствующая ей квадратичная форма положительно определена.

Для векторов конечномерного пространства билинейная форма в матричной форме записывается в виде

$$A(\mathbf{x}, \mathbf{y}) = \mathbf{x}\mathbf{A}\mathbf{y}^{\mathrm{T}},$$

где $\mathbf{A}$ — соответствующая матрица. Симметричной билинейной матрице соответствует симметрическая матрица, то есть такая, что $a_{ij} = a_{ji}$.

Квадратичная же форма запишется в виде

$$A(\mathbf{x}, \mathbf{x}) = \mathbf{x}\mathbf{A}\mathbf{x}^{\mathrm{T}},$$

где $\mathbf{A}$ — некоторая симметрическая матрица. Если квадратичная форма положительно определена, то соответствующая ей матрица также считается положительно определенной (то же относится к положительно полуопределенным формам и матрицам).

Покажем теперь, что матрица Грама положительно полуопределена. Возьмем m векторов размерности n. Построим матрицу $\mathbf{B}$ $(m \times n)$, по столбцам которой записаны наши векторы. Тогда матрица Грама $\mathbf{R}$ $(m \times m)$, состоящая из попарных скалярных произведений наших векторов, запишется как

$$\mathbf{R} = \mathbf{B}\mathbf{B}^{\mathrm{T}}.$$

Нам нужно показать, что для любого вектора $\mathbf{x}$ размерности m справедливо $\mathbf{x}\mathbf{R}\mathbf{x}^{\mathrm{T}} \geqslant 0$.

Будем считать теперь, что матрица $\mathbf{B}$ задает некоторое линейное преобразование

$$\mathbf{y} = \mathbf{x}\mathbf{B}$$

векторов евклидова пространства размерности m в векторы пространства размерности n. Для произвольного вектора x исходного пространства рассмотрим скалярное произведение его образа при этом преобразовании на себя:

$$(\mathbf{y}, \mathbf{y}) = (\mathbf{x}\mathbf{B}, \mathbf{x}\mathbf{B}) = \mathbf{x}\mathbf{B}\mathbf{B}^{\mathrm{T}}\mathbf{x}^{\mathrm{T}} = \mathbf{x}\mathbf{R}\mathbf{x}^{\mathrm{T}} \geqslant 0.$$

Утверждение доказано.

Более того, ранг матрицы Грама никогда не выше размерности исходного пространства n.

Для дальнейшего нам понадобятся некоторые дополнительные свойства квадратичных форм. Эти результаты излагаются в курсе линейной алгебры [11], но я позволю себе вам их напомнить.

Всякой квадратичной форме $\mathbf{x}\mathbf{A}\mathbf{x}^{\mathrm{T}}$ соответствует симметрический линейный оператор $\mathbf{y} = \mathbf{x}\mathbf{A}$, задаваемый той же матрицей, что используется при задании формы. Оператор называется симметрическим, если для любых двух векторов $\mathbf{x}$ и $\mathbf{y}$ выполнено условие $\mathbf{x}\mathbf{A}\mathbf{y}^{\mathrm{T}} = \mathbf{y}\mathbf{A}\mathbf{x}^{\mathrm{T}}$. Поскольку матрица симметрична, то ясно, что наш оператор будет симметричным.

Подпространство $\mathbf{R}_1$ исходного линейного пространства называется инвариантным относительно оператора $\mathbf{A}$, если для каждого вектора $\mathbf{x}$ из $\mathbf{R}_1$ вектор $\mathbf{y} = \mathbf{xA}$ также принадлежит $\mathbf{R}_1$. Если инвариантное пространство одномерно, то отличные от нуля векторы этого подпространства называются собственными векторами оператора A. Иными словами, всякий вектор $\mathbf{x} \neq 0$, удовлетворяющий условию $\mathbf{xA} = \lambda\mathbf{x}$, называется собственным вектором, а соответствующие число λ — собственным значением оператора $\mathbf{A}$.

Для симметрических операторов справедливо следующее утверждение: если вектор $\mathbf{z}$ является собственным вектором симметрического оператора A ($\mathbf{zA} = \lambda\mathbf{z}$), то подпространство, ортогональное этому вектору, будет инвариантным подпространством этого оператора. Напомню, что ортогональное к $\mathbf{z}$ подпространство образовано такими векторами $\mathbf{x}$, что $(\mathbf{x}, \mathbf{z}) = \mathbf{xz}^{\mathrm{T}} = 0$. Действительно

$$(\mathbf{xA}, \mathbf{z}) = \mathbf{xAz}^{\mathrm{T}} = \mathbf{zAx}^{\mathrm{T}} = \lambda\mathbf{zx}^{\mathrm{T}} = 0,$$

то есть вектор $\mathbf{xA}$ ортогонален вектору $\mathbf{z}$.

Далее, все собственные векторы симметрического оператора, соответствующие различным собственным числам, взаимно ортогональны. Действительно, пусть

$$\mathbf{xA} = \lambda_1\mathbf{y}, \quad \mathbf{yA} = \lambda_2\mathbf{y}, \qquad \lambda_1 \neq \lambda_2.$$

Тогда

$$\mathbf{xAy}^{\mathrm{T}} = \lambda_1\mathbf{xy}^{\mathrm{T}},$$

и в тоже время в силу симметрии

$$\mathbf{xAy}^{\mathrm{T}} = \mathbf{yAx}^{\mathrm{T}} = \lambda_2\mathbf{yx}^{\mathrm{T}} = \lambda_2\mathbf{xy}^{\mathrm{T}}.$$

Вычитая одно равенство из другого, получим

$$0 = (\mathbf{xAy}^{\mathrm{T}} - \mathbf{xAy}^{\mathrm{T}}) = (\lambda_1 - \lambda_2)\mathbf{xy}^{\mathrm{T}}.$$

Поскольку $\lambda_1 \neq \lambda_2$, это равенство возможно только, если $(\mathbf{x}, \mathbf{y}) = \mathbf{xy}^{\mathrm{T}} = 0$, то есть векторы $\mathbf{x}$ и $\mathbf{y}$ ортогональны.

В то же время, если m собственных векторов $\mathbf{z}_1$, $\mathbf{z}_2$,..., $\mathbf{z}_m$ оператора $\mathbf{A}$ соответствуют одному и тому же собственному числу λ, то и

все подпространство, натянутое на эти векторы, будет состоять из собственных векторов с этим же собственным значением. Действительно, пусть

$$\mathbf{x} = \sum_{i=1}^{m} a_i \mathbf{z}_i.$$

Тогда

$$\mathbf{x}\mathbf{A} = \sum_{i=1}^{m} a_i \mathbf{z}_i \mathbf{A} = \sum_{i=1}^{m} a_i \lambda \mathbf{z}_i = \lambda \sum_{i=1}^{m} a_i z_i = \lambda \mathbf{x}.$$

Теперь, если бы каждый оператор имел хотя бы один собственный вектор, то мы могли бы для симметрического оператора построить ортогональный базис, состоящий из собственных векторов. Для этого во всем пространстве найдем первый единичный собственный вектор. Ортогональное к нему подпространство будет собственным подпространством, и поэтому в нем можно найти второй единичный собственный вектор, и так далее, пока не исчерпается размерность пространства. Так строится базис. В этом базисе матрица нашего симметрического оператора приобретает диагональную форму, где по диагонали стоят собственные числа, а остальные элементы равны нулю. Квадратичная же форма запишется как

$$\mathbf{x}\mathbf{A}\mathbf{x}^{\mathrm{T}} = \sum_{i=1}^{n} \lambda_i x_i^2,$$

где x_i — координаты вектора в новом базисе, а λ_i —собственные числа оператора A.

Если квадратичная форма положительно определена, то ясно, что все собственные числа должны быть положительными. И наоборот, если все собственные числа положительны, то квадратичная форма положительно определена. У оператора, соответствующего положительно полуопределенной форме, все собственные числа должны быть неотрицательными, но некоторые из них могут быть равны нулю. Это значит, что оператор проектирует пространство в подпространство меньшей размерности.

Остается показать, что всякий линейный оператор в конечномерном пространстве имеет хотя бы один собственный вектор. Условие, что вектор $\mathbf{x}$ является собственным вектором оператора $\mathbf{A}$ при некотором собственном значении λ, можно записать как

$$\mathbf{x}\mathbf{A} - \lambda\mathbf{x} = \mathbf{x}(\mathbf{A} - \lambda\mathbf{E}) = 0,$$

где $\mathbf{E}$ — единичная матрица.

Известно, что это уравнение имеет отличное от нуля решение в том и только том случае, когда определитель матрицы $(\mathbf{A} - \lambda\mathbf{E})$ равен нулю. Раскрывая этот определитель, получим полином степени n относительно λ, который называется характеристическим многочленом. Соответствующее уравнение называется характеристическим уравнением. В поле комплексных чисел это уравнение имеет по крайней мере один корень, но этот корень может быть комплексным числом.

Однако оказывается, что если матрица симметрична, то соответствующее характеристическое уравнение может иметь только действительные корни. Доказательство этого я оставляю в качестве упражнения.

Плотность распределения вероятностей

Определим сначала плотность распределения для одномерной случайной величины. Плотностью распределения называется такая функция $p(x)$, что кумулятивная функция распределения $F(z)$ может быть представлена в виде

$$F(z) = \int_{-\infty}^{z} p(x)\,dx.$$

Если функция $F(z)$ дифференцируема, то плотность, очевидно, совпадает с ее производной:

$$p(x) = \frac{dF(x)}{dx}.$$

Но производная по крайней мере в некоторых точках может не существовать (график $F(x)$ может иметь изломы), а определенная выше плотность распределения $p(x)$ все же существует. Однако, если кумулятивная функция распределения имеет скачки, то в этих точках плотность распределения не существует. (Чтобы ее все же определить, придется перейти к обобщенным функциям).

В случае нескольких переменных плотность распределения $p(x_1, x_2, \ldots, x_n)$ определяется из условия

$$F(z_1, z_2, \ldots, z_n) = \int_{-\infty}^{z_1}\int_{-\infty}^{z_2}\ldots\int_{-\infty}^{z_n} p(x_1, x_2, \ldots, x_n)\,dx_1\,dx_2\ldots\,dx_n.$$

Если существуют соответствующие частные производные, то плотность может быть выражена как

$$p(x_1, x_2, \ldots, x_n) = \frac{\partial^n F(x_1, x_2, \ldots, x_n)}{\partial x_1 \partial x_2 \ldots \partial x_n}.$$

Как и в одномерном случае, производная в некоторых точках может не существовать, а определенная выше плотность все же существует.

Метод максимального правдоподобия

Рассмотрим сначала метод максимального правдоподобия в самом общем виде. Пусть на вход объекта подается сигнал $\mathbf{x}$, а на выходе появляется сигнал $\mathbf{y}$ с вероятностью $p_0(\mathbf{y}|\mathbf{x})$. Природа этих сигналов может быть совершенно любой, но в случае, когда выход дискретен, мы будем понимать $p_0(\mathbf{y}|\mathbf{x})$ как условную вероятность, а в случае, когда выход описывается непрерывными величинами, как условную плотность распределения вероятностей.

Условное распределение $p_0(\mathbf{y}|\mathbf{x})$ нам не известно, и, собственно, его требуется найти или оценить. Предполагается, что эта неизвестная функция принадлежит некоторому заданному классу условных распределений, которые могут в параметрической форме быть описаны как $p(\mathbf{y}|\mathbf{x}, \alpha)$, где параметр α, в принципе, может иметь любую природу. Иными словами, существует такое значение α^* параметра, что

$$p_0(\mathbf{y}|\mathbf{x}) = p(\mathbf{y}|\mathbf{x}, \alpha^*).$$

Нам дается конкретная пара $(\mathbf{x}^*, \mathbf{y}^*)$ и по ней требуется найти или оценить значение этого параметра α^*. На самом деле это, конечно, может быть последовательность пар, полученных в ходе независимых (или зависимых) испытаний, пара длинных сигналов, наблюдаемых на входе и выходе, и пр.

Метод состоит в том, что в качестве α^* ищется такое значение параметра α, которое доставляет максимум функции правдоподобия

$$W(\alpha) = \log p(\mathbf{y}|\mathbf{x}, \alpha).$$

Какова же идея, лежащая в основе этого метода? Допустим, что существует распределение вероятностей $P(\mathbf{x})$ на множестве входных сигналов. Тогда математическое ожидание функции правдоподобия будет

равно

$$EW(\alpha) = \int\int \log p(\mathbf{y}|\mathbf{x},\alpha) p_0(\mathbf{y}|\mathbf{x})\, d\mathbf{y}\, dP(\mathbf{x}). \quad (1)$$

(Здесь мы используем $p_0(\mathbf{y}|\mathbf{x})$ как плотность распределения, но совершенно аналогично рассматривается и случай дискретных значений $\mathbf{y}$.) Нетрудно убедиться, что функция

$$\int \log p(\mathbf{y}|\mathbf{x},\alpha) p_0(\mathbf{y}|\mathbf{x})\, d\mathbf{y} \quad (2)$$

достигает максимума при таком значении параметра α^*, что

$$p(\mathbf{y}|\mathbf{x},\alpha^*) = p_0(\mathbf{y}|\mathbf{x}),$$

а такое значение, согласно начальному допущению, есть. Более того, при $\alpha = \alpha^*$ это равенство будет выполняться при всех $\mathbf{x}$, и, следовательно, интеграл (1) достигнет максимума, а с ним и математическое ожидание функции правдоподобия. При любом другом значении α, таком, что плотность $p(\mathbf{y}|\mathbf{x},\alpha)$ не всюду равна $p_0(\mathbf{y}|\mathbf{x})$, значение интеграла (2) будет меньше максимального. Следовательно, если математическое ожидание функции правдоподобия достигнет максимума в какой-то еще точке α^{**}, то плотность$p(\mathbf{y}|\mathbf{x},\alpha^{**})$ может отличаться от $p_0(\mathbf{y}|\mathbf{x})$ лишь на множестве меры нуль.

Все бы было хорошо, но математическое ожидание функции правдоподобия нам не дано. Остается надеяться, что функция правдоподобия на предъявленной паре вход-выход каким то образом служит оценкой ее математического ожидания.

Мы далее ограничимся тем случаем, когда на обучение представлена последовательность входных значений $\mathbf{x} = (x_1, x_2, \ldots, x_l)$ и соответствующая последовательность значений $\mathbf{y} = (y_1, y_2, \ldots, y_l)$ на выходе объекта, полученных в процессе независимых испытаний при неизменном распределении $P(\mathbf{x})$ и условном распределении $P_0(\mathbf{y}|\mathbf{x})$. (На самом деле в более общем случае можно рассматривать и иные формы связи между входом и выходом, обладающие свойством эргодичности, но мы ограничимся этим простым случаем). Тогда условная вероятность получить на выходе конкретную последовательность $\mathbf{y} = (y_1, y_2, \ldots, y_l)$ при заданной последовательности $\mathbf{x} = (x_1, x_2, \ldots, x_l)$ на входе составит

$$P_0(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{l} P_0(y_i|x_i),$$

а ее логарифм

$$W_0 = \sum_{i=1}^{l} \log P_0(y_i|x_i).$$

Согласно методу максимального правдоподобия мы ищем оценку условного распределения в классе условных распределений $P(y|x,\alpha)$, подбирая такое значение параметра α, при котором функция правдоподобия

$$W(\alpha) = \sum_{i=1}^{l} \log P(y_i|x_i,\alpha)$$

достигает максимума. При этом предполагается, что существует такое значение α^*, что $P_0(y|x) = P(y|x,\alpha^*)$, т. е. истинная модель принадлежит тому классу, из которого мы выбираем. Ничего не изменится, если мы разделим функцию $W(\alpha)$ на длину выборки l. Тогда величина

$$\frac{1}{l}W(\alpha) = \frac{1}{l}\sum_{i=1}^{l} \log P(y_i|_i,\alpha)$$

будет средним арифметическим функции правдоподобия в точках выборки, и согласно закону больших чисел при всяком фиксированном значении α сходится к ее математическому ожиданию. Но мы уже показали, что максимум математического ожидания функции правдоподобия достигается при $P(y|x,\alpha) = P_0(y|x)$, т. е. на правильной модели. Такова общая идея, обосновывающая метод максимума правдоподобия.

Она аналогична тем соображениям, которые лежат в основе метода минимума эмпирического риска, и в дальнейшем мы покажем, почему она работает, и почему работает не всегда.

Сравнивая метод максимального правдоподобия с подходом, основанным на минимизации эмпирического риска, отметим, что этот метод гораздо более чувствителен к случаям невозможным или очень маловероятным, потому что при этом уходит в минус бесконечность $\log P(y_i|x_i,\alpha)$, тогда как штрафная функция может оставаться небольшой. С другой стороны, если в нашем классе нет правильной модели, то минимизация эмпирического риска позволяет выбрать модель близкую к истинной, а метод максимума правдоподобия может дать совсем неверный результат.

Оценка дисперсии параметров распределения, найденных методом максимума правдоподобия

Рассмотрим сначала простой случай оценки скалярного параметра плотности распределения вероятностей [20]. Пусть $P_0(y)$ — неизвестная плотность распределения, принадлежащая семейству плотностей $P(y,\alpha)$, где α — скалярный параметр. Это значит, что существует значение такое α_0, что $P_0(y) = P(y,\alpha_0)$. Природа y не имеет значения — это может быть скаляр, вектор и пр. Дана выборка $\mathbf{y} = (y_1, y_2, \ldots, y_l)$, полученная при независимых испытаниях с неизменной плотностью распределения $P_0(y)$. Составляем функцию правдоподобия

$$W(\alpha) = \frac{1}{l}\sum_{i=1}^{l} \log P(y_i,\alpha).$$

Допустим, что точка α^*, доставляющая максимум этой функции, уже близка к истинному значению α_0. Разложим функцию в ряд Тейлора по степеням $\Delta\alpha = \alpha - \alpha_0$, пренебрегая всеми членами выше второго порядка:

$$\begin{aligned} W(\alpha) &\approx W(\alpha_0) + k_1\Delta\alpha + \frac{1}{2}k_2\Delta\alpha^2 = \\ &= W(\alpha_0) + \frac{1}{l}\sum_{i=1}^{l}\frac{\partial \log P(y_i,\alpha_0)}{\partial\alpha}\Delta\alpha + \\ &\quad + \frac{1}{2l}\sum_{i=1}^{l}\frac{\partial^2 \log P(y_i,\alpha_0)}{\partial\alpha^2}\Delta\alpha^2. \end{aligned}$$

Тогда максимум функции достигается (приближенно) при $\Delta\alpha^* = -k_1/k_2$. Величины k_1 и k_2 оказываются случайными, так как они зависят от выборочных значений. Но мы будем считать, что выборка достаточно велика, чтобы значение k_2 сошлось достаточно точно к своему математическому ожиданию:

$$\begin{aligned} k_2 \approx Ek_2 &= \int \frac{\partial^2 \log P(y,\alpha_0)}{\partial\alpha^2} P_0(y)\,dy = \\ &= \int \frac{\partial^2 \log P(y,\alpha_0)}{\partial\alpha^2} P(y,\alpha_0)\,dy. \end{aligned}$$

Поскольку мы ищем максимум (а не минимум), то вторая производная вблизи максимума должна быть отрицательной. И в самом деле, поскольку интеграл от плотности всегда равен 1, имеем $\int P(y,\alpha)\,dy \equiv 1$.

Дифференцируя это тождество по α, получим

$$0 \equiv \int \frac{\partial P(y,\alpha)}{\partial \alpha}\,dy = \int \frac{\partial \log P(y,\alpha)}{\partial \alpha} P(y,\alpha)\,dy,$$

и дифференцируя еще раз по α, имеем

$$0 \equiv \int \left(\frac{\partial \log P(y,\alpha)}{\partial \alpha}\right)^2 P(y,\alpha)\,dy + \int \frac{\partial^2 \log P(y,\alpha)}{\partial \alpha^2} P(y,\alpha)\,dy.$$

Откуда, в частности, при $\alpha = \alpha_0$ получаем

$$\int \frac{\partial^2 \log P(y,\alpha_0)}{\partial \alpha^2} P(y,\alpha_0)\,dy = -\int \left(\frac{\partial \log P(y,\alpha_0)}{\partial \alpha}\right)^2 P(y,\alpha_0)dy.$$

То есть коэффициент k_2 при достаточно большой выборке становится отрицательным.

В отличие от этого, коэффициент k_1 в среднем равен нулю. Дело в том, что математическое ожидание функции правдоподобия в точке α_0 достигает максимума, и, следовательно, его производная по α должна быть равна нулю.

А тогда и математическое ожидание производной в точке равно нулю, т. е.

$$Ek_1 = E\left(\frac{1}{l}\sum_{i=1}^{l} \frac{\partial \log P(y_i,\alpha_0)}{\partial \alpha}\right) = 0.$$

Поэтому погрешность в определении параметра определяется разбросом значений коэффициента k_1 вокруг нуля (а оценка в среднем параметра будет правильной).

Найдем его дисперсию:

$$Dk_1 = E\left(\frac{1}{l}\sum_{i=1}^{l}\frac{\partial \log P(y_i,\alpha_0)}{\partial\alpha}\right)^2 =$$
$$= \frac{1}{l^2}E\left(\sum_{i=1}^{l}\frac{\partial \log P(y_i,\alpha_0)}{\partial\alpha}\right)^2 =$$
$$= \frac{1}{l^2}\sum_{i,j=1}^{l}E\left(\frac{\partial \log P(y_i,\alpha_0)}{\partial\alpha}\cdot\frac{\partial \log P(y_j,\alpha_0)}{\partial\alpha}\right).$$

При $i = j$ имеем

$$E\left(\frac{\partial \log P(y_i,\alpha_0)}{\partial\alpha}\cdot\frac{\partial \log P(y_i,\alpha_0)}{\partial\alpha}\right) =$$
$$= \int\left(\frac{\partial \log P(y,\alpha_0)}{\partial\alpha}\right)^2 P(y,\alpha_0)\,dy,$$

а при $i \neq j$ величина $E\left[(\partial \log P(y_i,\alpha_0)/\partial\alpha)\cdot(\partial \log P(y_j,\alpha_0)/\partial\alpha)\right]$ равна нулю как математическое ожидание двух случайных величин с нулевым средним. Поэтому

$$Dk_1 = \frac{1}{l}\int\left(\frac{\partial \log P(y,\alpha_0)}{\partial\alpha}\right)^2 P(y,\alpha_0)\,dy.$$

Теперь можно оценить дисперсию оценки α (считая коэффициент k_2 постоянным).

$$D\alpha^* = D\Delta\alpha = D\left(\frac{k_1}{k_2^2}\right) = \frac{1}{l\int\left(\frac{\partial \log P(y,\alpha_0)}{\partial\alpha}\right)^2 P(y,\alpha_0)\,dy}.$$

Но нужно помнить, что все эти оценки получены в предположении, что коэффициент при квадратичной части уже очень близок к своему математическому ожиданию, а оценка настолько близка к истинной, что можно ограничиться тремя членами ряда Тейлора в разложении функции правдоподобия.

Лекция 4

Линейные преобразования случайных величин. Метод максимального правдоподобия (случай векторного параметра). Метод наименьших квадратов для оценки регрессии (общий подход). Метод наименьших квадратов для поиска наилучшего линейного приближения. (Стандартные процедуры регрессии и максимума правдоподобия).

Линейные преобразования

Для дальнейшего нам нужно увидеть, как изменятся математическое ожидание и ковариационная матрица n-мерного вектора при заданном линейном преобразовании. Пусть заданы две случайные величины x и y, и по ним строится две новые случайные величины $t = ax + by$ и $u = cx + dy$. Тогда

$$\begin{aligned}
&\mathrm{E}_t = M_t = aM_x + bM_y; \quad \mathrm{E}_u = M_u = cM_x + dM_y;\\
&\mathrm{D}_t = a^2\mathrm{D}_x + b^2\mathrm{D}_y + 2ab\,\mathbf{cov}_{xy}; \quad \mathrm{D}_u = c^2\mathrm{D}_x + d^2\mathrm{D}_y + 2cd\,\mathbf{cov}_{xy};\\
&\mathbf{cov}(t,u) = ac\mathrm{D}_x + bd\mathrm{D}_y + ad\,\mathbf{cov}_{xy} + bc\,\mathbf{cov}_{xy}\,.
\end{aligned}$$

Пусть теперь задано распределение n-мерного вектора $\mathbf{x}$ с координатами $(x_1, \ldots, x_n)$, вектор его математического ожидания $M_\mathbf{x} = (M_{x_1}, \ldots, M_{x_n})$ и ковариационная матрица $\mathbf{K_x}$ с элементами $k_{ij} = \mathbf{cov}(x_i, x_j)$. На этот вектор действует линейное преобразование, задаваемое матицей A:

$$\mathbf{y} = A\mathbf{x}.$$

Тогда математическое ожидание и ковариационная матрица вектора $\mathbf{y}$ будут равны:

$$M_\mathbf{y} = AM_\mathbf{x}; \quad \mathbf{K_y} = A\mathbf{K_x}A^{\mathrm{T}}. \tag{1}$$

В частности, из линейной алгебры известно, что для любой симметрической матрицы (а ковариационная матрица таковой является) существует унитарное (сохраняющее метрику) линейное преобразование Q, которое приводит эту матрицу к диагональному виду. Поэтому существует такая унитарная матрица Q^*, что матрица $\mathbf{K_y} = Q^*\mathbf{K_x}(Q^*)^{\mathrm{T}}$ станет диагональной. Следовательно, координаты вектора $\mathbf{y} = Q\mathbf{x}$ станут некоррелированными. На этом основан метод главных компонент. Правда, если исходные переменные имеют разную природу (разную размерность), то их предварительно нормируют на среднеквадратическое отклонение.

Рассмотрим одно важное применение результата (1), которое нам понадобится как при дальнейшем анализе метода максимального правдоподобия, так и при разборе метода наименьших квадратов. Пусть дана квадратичная функция вектора $\mathbf{x}$:

$$W(\mathbf{x}) = W_0 + \boldsymbol{a}\mathbf{x}^{\mathrm{T}} + \frac{1}{2}\mathbf{x}A\mathbf{x}^{\mathrm{T}},$$

где матрица A квадратичной части считается фиксированной симметрической положительно определенной, а вектор $\boldsymbol{a}$ при линейной части — случайным с нулевым средним и ковариационной матрицей $\mathbf{K}$. Приравнивая к нулю градиент функции W

$$\mathbf{grad}_{\mathbf{x}}\, W(\mathbf{x}) = \boldsymbol{a} + \mathbf{x}A = 0,$$

получаем, что ее минимум достигается при $x^* = -\boldsymbol{a}(A^{-1})^{\mathrm{T}}$. Вектор x^* будет случайным (т.к. вектор $\boldsymbol{a}$ случаен) и согласно (1) видим, что математическое x^* ожидание равно 0, а его ковариационная матрица равна

$$\mathbf{K}^* = A^{-1}\mathbf{K}(A^{-1})^{\mathrm{T}}.$$

Понятно, что, если бы матрица A была бы отрицательно определенной и мы бы искали максимум функции W, то результат был бы тем же самым. Ясно также, что для того, чтобы наш результат имел смысл, матрица A должна иметь обратную, то есть быть невырожденной.

Метод максимума правдоподобия (много параметров)

Теперь мы можем вернуться к методу максимума правдоподобия.

Рассмотрим опять задачу восстановления по выборке неизвестной плотности $p_0(\mathbf{x})$, предполагая, что она принадлежит классу плотностей $p(x, \boldsymbol{\alpha})$, где теперь $\boldsymbol{\alpha}$ уже вектор параметров с координатами $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$. То есть предполагается, что существует такое значение $\boldsymbol{\alpha}_0$, что $p_0(\mathbf{x}) = p(x, \boldsymbol{\alpha}_0)$. (На самом деле существенно, что координаты $\boldsymbol{\alpha}$ — вещественные числа. Значения $\mathbf{x}$ могут быть и дискретными, но тогда нужно понимать не как плотность, а как распределение вероятностей, и интегрирование по в дальнейших формулах заменить на суммирование.)

Как и раньше, предполагается, что дана выборка описаний $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l$, полученная в процессе независимых испытаний при неизменной плотности распределения $p_0(\mathbf{x}) = p(x, \boldsymbol{\alpha}_0)$. Согласно методу максимального правдоподобия, в качестве оценки вектора параметров $\boldsymbol{\alpha}_0$ ищется такое значение $\boldsymbol{\alpha}^*$, которое доставляет максимум функции правдоподобия

$$W(\boldsymbol{\alpha}) = \frac{1}{l}\sum_{i=1}^{l} \log P(x_i, \boldsymbol{\alpha}).$$

Оценим отклонение значения $\boldsymbol{\alpha}^*$, выбранного методом максимального правдоподобия, от истинного значения $\boldsymbol{\alpha}_0$. Вектор этого отклонения с координатами $(\Delta\alpha_1, \ldots, \Delta\alpha_m)$ мы обозначим $\Delta\boldsymbol{\alpha}$. Опять предполагается, что точка $\boldsymbol{\alpha}^*$ достаточно близка к $\boldsymbol{\alpha}_0$, так что функция правдоподобия

$$W(\boldsymbol{\alpha}) = \frac{1}{l}\sum_{i=1}^{l} \log P(x_i, \alpha_i).$$

может быть приближена тремя членами ряда Тейлора в окрестности $\boldsymbol{\alpha}_0$:

$$W(\boldsymbol{\alpha}) \approx W(\boldsymbol{\alpha}_0) + \mathbf{k}\Delta\alpha^{\mathrm{T}} + \frac{1}{2}\Delta\boldsymbol{\alpha}\mathbf{K}\Delta\boldsymbol{\alpha}^{\mathrm{T}},$$

где вектор $\mathbf{k}$ — градиент по функции $W(\boldsymbol{\alpha})$ в точке $\boldsymbol{\alpha}_0$, а отрицательно определенная матрица $\mathbf{K}$ является матрицей вторых производных. Соответственно координаты вектора $\mathbf{k}$ равны

$$k_i = \frac{\partial W(\boldsymbol{\alpha})}{\partial\alpha_i} = \frac{1}{l}\sum_{k=1}^{l} \frac{\partial \log P(xk, \alpha)}{\partial\alpha_i},$$

а элементы матрицы $\mathbf{K}$ равны:

$$k_{ij} = \frac{\partial^2 W(\boldsymbol{\alpha})}{\partial\alpha_i\partial\alpha_j} = \frac{1}{l}\sum_{k=1}^{l} \frac{\partial^2 \log P(x_k, \alpha)}{\partial\alpha_i\partial\alpha_j}.$$

Поскольку мы рассматриваем разложение в ряд Тейлора в окрестности $\boldsymbol{\alpha}_0$, то все производные берутся в точке $\boldsymbol{\alpha}_0$. Тогда в нашем приближении максимум функции правдоподобия будет достигнут при

$$\Delta\boldsymbol{\alpha} = -\mathbf{k}\mathbf{K}^{-1}. \tag{2}$$

Как и в одномерном случае, вектор $\mathbf{k}$ считается случайным (зависящим от выборки) с нулевым математическим ожиданием. Действительно, математическое ожидание функции правдоподобия достигает максимума при $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$, и, значит, его градиент в этой точке равен 0, а тогда и математическое ожидание градиента равно 0. При этом ковариационная матрица координат $\mathbf{R}$ вектора $\mathbf{k}$ состоит из элементов

$$\begin{aligned}\mathbf{cov}(k_i, k_j) &= E\left[\frac{\partial W(\boldsymbol{\alpha})}{\partial \alpha_i} \cdot \frac{\partial W(\boldsymbol{\alpha})}{\partial \alpha_j}\right] = \\ &= E\left[\frac{\partial\left(\dfrac{1}{l}\displaystyle\sum_{k=1}^{l} \log P(x_k, \alpha)\right)}{\partial \alpha_i} \cdot \frac{\partial\left(\dfrac{1}{l}\displaystyle\sum_{k=1}^{l} \log P(x_s, \alpha)\right)}{\partial \alpha_j}\right] = \\ &= \frac{1}{l^2} \sum_{k,1=1}^{l} E\left[\frac{\partial \log P(x_k, \alpha)}{\partial \alpha_i} \cdot \frac{\partial \log P(x_s, \alpha)}{\partial \alpha_j}\right].\end{aligned}$$

Перекрестные члены этой суммы, то есть такие, что $k \neq s$, равны нулю, так как сомножители являются независимыми величинами с нулевым средним, а диагональные равны

$$\int \frac{\partial \log P(x, \alpha)}{\partial \alpha_i} \cdot \frac{\partial \log P(x, \alpha)}{\partial \alpha_j} p(x, \boldsymbol{\alpha}_0)\, dx,$$

и таких членов равно l штук. Поэтому

$$\mathbf{cov}(k_i, k_j) = \frac{1}{l} \int \frac{\partial \log P(x, \alpha)}{\partial \alpha_i} \cdot \frac{\partial \log P(x, \alpha)}{\partial \alpha_j} p(x, \boldsymbol{\alpha}_0)\, dx,$$

причем производные берутся в точке $\boldsymbol{\alpha}_0$. Обозначим эту ковариационную матрицу через $\mathbf{R}$.

Что же касается элементов k_{ts} матрицы $\mathbf{K}$, то считается, что они

уже сошлись к своему математическому ожиданию:

$$k_{ts} \approx E\left(\frac{\partial^2 W(\boldsymbol{\alpha})}{\partial\alpha_t \partial\alpha_s}\right) = E\left[\frac{\partial^2\left(\frac{1}{l}\sum_{i=1}^{l}\log P(x_i,\alpha)\right)}{\partial\alpha_t\partial\alpha_s}\right] =$$

$$= \int \frac{\partial^2 \log P(x_j,\alpha)}{\partial\alpha_t\partial\alpha_s} p(x,\boldsymbol{\alpha}_0)\,dx.$$

Эти элементы могут быть преобразованы так же, как в одномерном случае, путем дифференцирования тождества $\int p(x,\boldsymbol{\alpha})\,dx \equiv 1$:

$$\int \frac{\partial p(x,\boldsymbol{\alpha})}{\partial\alpha_t}\,dx = \int \frac{\partial \log p(x,\boldsymbol{\alpha})}{\partial\alpha_t} p(x,\boldsymbol{\alpha})\,dx \equiv 0;$$

$$\int \frac{\partial^2 p(x,\boldsymbol{\alpha})}{\partial\alpha_t\partial\alpha_s}\,dx = \int \frac{\partial^2 \log p(x,\boldsymbol{\alpha})}{\partial\alpha_t\partial\alpha_s} p(x,\boldsymbol{\alpha})\,dx +$$

$$+ \int \frac{\partial \log P(x,\boldsymbol{\alpha})}{\partial\alpha_t}\frac{\partial \log P(x,\boldsymbol{\alpha})}{\partial\alpha_s} p(x,\boldsymbol{\alpha})\,dx \equiv 0.$$

Откуда

$$E\left[\frac{\partial^2 W(\boldsymbol{\alpha})}{\partial\alpha_t\partial\alpha_s}\right] = -E\left[\frac{\partial W(\boldsymbol{\alpha})}{\partial\alpha_t}\frac{\partial W(\boldsymbol{\alpha})}{\partial\alpha_s}\right].$$

Таким образом, в нашем приближении ковариационная матрица $\mathbf{R}$ и матрица квадратичной части $\mathbf{K}$ связаны соотношением:

$$\mathbf{R} = -\frac{1}{l}\mathbf{K}.$$

Отсюда на основании (1) с учетом соотношения (2) получаем, что ковариационная матрица $\mathbf{C}$ погрешности $\Delta\boldsymbol{\alpha}$ в нашем приближении равна

$$\mathbf{C} = \mathbf{K}^{-1}\mathbf{R}\mathbf{K}^{-1^{\mathrm{T}}} = \mathbf{K}^{-1}\left(-\frac{1}{l}\mathbf{K}\right)\mathbf{K}^{-1^{\mathrm{T}}} = -\frac{1}{l}\mathbf{K}^{-1}.$$

Посмотрим, сколько допущений нам пришлось сделать при выводе этого соотношения. Было предположено, в классе плотностей $p(x,\boldsymbol{\alpha})$ есть истинная плотность $p_0(x)$. Что оценка $\boldsymbol{\alpha}^*$ настолько близка к $\boldsymbol{\alpha}_0$, что можно воспользоваться разложением функции правдоподобия в ряд

Тейлора с точностью до трех членов. Что матрица квадратичной части уже сошлась к своему математическому ожиданию, и, наконец, производные под интегралами должны браться в точке α_0, которая нам неизвестна, и приближенно ее можно заменить на α^*. Все эти классические результаты были получены как асимптотические, но применять то их приходится в случае конечных выборок.

Регрессия

Пусть на вход поступает сигнал произвольной природы $\mathbf{x}$, а на выходе имеем числовое значение $\mathbf{y}$ и для всякого $\mathbf{x}$ существует условное распределение $P(\mathbf{y}|\mathbf{x})$. Регрессией называется числовая функция аргумента $\mathbf{x}$, равная условному математическому ожиданию $\mathbf{y}$ при условии заданного $\mathbf{x}$:

$$R(\mathbf{x}) = E(\mathbf{y}|\mathbf{x}).$$

Математическое ожидание отклонения $\mathbf{y}$ от регрессии равно условной дисперсии при заданном входе $\mathbf{x}$:

$$E(\mathbf{y} - R(\mathbf{x}))^2 = E(\mathbf{y} - E(\mathbf{y}|\mathbf{x}))^2 = D(\mathbf{y}|\mathbf{x}).$$

Если взять произвольную измеримую функцию $F(\mathbf{x})$ аргумента $\mathbf{x}$ и посмотреть, как ведет себя среднеквадратическое отклонение $\mathbf{y}$ от этой функции, то получим

$$\begin{aligned} E(\mathbf{y} - F(\mathbf{x}))^2 &= E\left[(\mathbf{y} - R(\mathbf{x})) + (R(\mathbf{x}) - F(\mathbf{x}))\right]^2 = \\ &= E((\mathbf{y} - R(\mathbf{x}))^2) + (R(\mathbf{x}) - F(\mathbf{x}))^2 + \\ &\quad + 2E\left((\mathbf{y} - R(\mathbf{x}))(R(\mathbf{x}) - F(\mathbf{x}))\right). \end{aligned}$$

Но величина $E((\mathbf{y} - R(\mathbf{x})) = E(\mathbf{y}|\mathbf{x}) - E(\mathbf{y}|\mathbf{x})$ равна нулю, поэтому

$$E(\mathbf{y} - F(\mathbf{x}))^2 = D(\mathbf{y}|\mathbf{x}) + (R(\mathbf{x}) - F(\mathbf{x}))^2.$$

Таким образом, среднее отклонение $\mathbf{y}$ от произвольной функции при заданном $\mathbf{x}$ равно сумме условной дисперсии и квадрата отклонение этой функции от регрессии.

Если также задано распределение вероятностей на множестве значений аргумента $\mathbf{x}$, то усредненное по $\mathbf{x}$ отклонение от произвольной (интегрируемой) функции составит

$$E(\mathbf{y} - F(\mathbf{x}))^2 = \int D(\mathbf{y}|\mathbf{x})\, dP(\mathbf{x}) + \int (R(\mathbf{x}) - F(\mathbf{x}))^2\, dP(\mathbf{x}).$$

Иными словами, оно равно средней условной дисперсии $D(\mathbf{y}|\mathbf{x})$ плюс средний квадрат отклонения этой функции от регрессии. Отсюда ясно, что если мы ищем зависимость $F(\mathbf{x})$, которая обеспечит минимум $E(\mathbf{y} - F(\mathbf{x})^2)$ среднеквадратичного отклонения, то эта задача равносильна поиску регрессии. Если же в классе функций, в котором ведется поиск, нет регрессии, то наилучшей будет такая функция из этого класса, для которой достигается минимум среднего квадрата отклонения от регрессии.

Сравним критерий минимума среднеквадратичной погрешности с критерием максимума функции правдоподобия в следующем частном, но важном случае. Пусть входное значение x связано выходной величиной детерминированной (числовой) зависимостью $Q(\mathbf{x})$, но выходные значения искажены шумом ζ с известной плотностью распределения $p(\zeta)$ и нулевым средним. Шум во всех точках распределен одинаково и независимо. То есть

$$y = Q(\mathbf{x}) + \zeta,$$

и условная плотность распределения $\mathbf{y}$ при заданном значении $\mathbf{x}$ будет равна $p(y|x) = p(y - Q(\mathbf{x}))$. Применяя теперь метод максимума правдоподобия, получаем критерий для поиска функции $Q(\mathbf{x})$:

$$W(Q) = \sum \log p(y_i - Q(\mathbf{x}_i)).$$

Отсюда ясно, что в случае нормально распределенного шума с нулевым средним метод максимума правдоподобия приводит к методу наименьших квадратов. Действительно, в этом случае

$$p(y|\mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - Q(\mathbf{x}))^2}{2\sigma^2}\right),$$

где σ^2 - дисперсия шума. Тогда

$$W(Q) = \text{const} - \sum \frac{(y_i - Q(\mathbf{x}_i))^2}{2\sigma^2}.$$

Заметим, что фактически критерий не зависит от дисперсии шума.

При других распределениях шума (с нулевым средним) метод максимума правдоподобия дает другие критерии. Так, если шум распределен экспоненциально, т. е. его плотность распределения равна

$$p(\zeta) = \frac{1}{2a} \exp(-|a\zeta|),$$

то критерий максимума правдоподобия принимает вид

$$W(Q) = \text{const} - \sum |\mathbf{y}_i - Q(\mathbf{x}_i)|,$$

т. е. ищется зависимость, обеспечивающая минимум модуля невязки. Отметим, что минимум математического ожидания функции правдоподобия в этом случае достигается на функции $Q(\mathbf{x})$, равной условной медиане распределения y при заданном $\mathbf{x}$. Последняя совпадает с регрессией, если шум имеет симметричное относительно нуля распределение, но совсем не обязательно совпадает с ней при произвольном распределении $p(\mathbf{y}|\mathbf{x})$.

Рассмотрим еще такой «экзотический» случай. Пусть плотность распределения шума имеет вид

$$p(\zeta) = \frac{b}{a^2 + \zeta^2}$$

(распределение Коши).

Это распределение, будучи симметричным относительно нуля, не имеет ни математического ожидания, ни дисперсии. В этом случае критерий максимального правдоподобия приобретает вид

$$W(Q) = \text{const} - \sum \log\left(a^2 + (\mathbf{y}_i - Q(\mathbf{x}_i))^2\right).$$

Видно, что в этом случае большие отклонения от регрессии дают значительно меньший вклад в критерий, чем в случае нормально или экспоненциально распределенного шума.

Метод наименьших квадратов для поиска наилучшего линейного приближения

Рассмотрим случай, когда на вход подается вектор $\mathbf{x} = (x_1, \ldots, x_m)$, а выход представляет собой скалярную вещественную величину y. По заданной выборке $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)$ будем искать функцию вида

$$F(\mathbf{x}) = (\boldsymbol{a}\mathbf{x}) + b,$$

где $(\cdot)$ — скалярное произведение, вектор $\boldsymbol{a} = (a_1, \ldots, a_m)$ и константа b — неизвестные параметры, минимизирующие среднеквадратическую

невязку на обучающей выборке:

$$\min I(\boldsymbol{a}, b) = \frac{1}{l}\sum_{i=1}^{l}(y_i - F(\mathbf{x}_i))^2.$$

Этот путь реализует принцип замены истинного риска на эмпирический риск: математическое ожидание среднеквадратического уклонения заменяется на среднее значение невязки по данным обучающей выборки [5].

Наличие свободного члена b вносит некоторую асиметрию в задачу. Поэтому от него желательно избавиться. Это можно сделать двумя путями. Можно добавить к входному вектору еще одну координату, тождественно равную 1, и искать функцию вида $F(\mathbf{x}) = (\boldsymbol{a}\mathbf{x})$. Тогда коэффициент при этой дополнительной координате будет играть ту же роль, что и свободный член. Но с вычислительной точки зрения и более удобно отцентрировать выборку, вычитая из всех векторов их среднее $\mathbf{x}_{\text{ср}} = (1/l)\sum_{i=1}^{l}\mathbf{x}_i$, т. е. будем искать зависимость в виде $F(\mathbf{x}) = (\boldsymbol{a}(\mathbf{x} - \mathbf{x}_{\text{ср}})) + b$. Тогда свободный член находится легко. Действительно, приравнивая к нулю производную по b от выражения $\sum_{i=1}^{l}(y_i - [(\boldsymbol{a}(\mathbf{x}_i - \mathbf{x}_{\text{ср}})) + b])^2$, получим

$$2\sum_{i=1}^{l}(y_i - (\boldsymbol{a}(x_i - \mathbf{x}_{\text{ср}})) - 2lb = 2\sum_{i=1}^{l} y_i + (\boldsymbol{a}\sum(\mathbf{x}_i - \mathbf{x}_{\text{ср}})) - 2lb = 0.$$

Но $\sum(\mathbf{x}_i - \mathbf{x}_{\text{ср}}) = \sum \mathbf{x}_i - l\mathbf{x}_{\text{ср}} = 0$. Поэтому $b_{\text{опт}} = (1/l)\sum_{i=1}^{l} y_i = y_{\text{ср}}$, т. е. средневыборочному значению y (независимо от $\boldsymbol{a}$). В дальнейшем мы будем считать, что значения и уже центрированы (т. е.$\mathbf{x}_{\text{ср}} = 0$ и $y_{\text{ср}} = 0$) и искать зависимость в форме $F(\mathbf{x}) = (\boldsymbol{a}\mathbf{x})$.

Итак, нам нужно найти значение $a_{\text{опт}}$, доставляющее минимум выражению

$$\mathbf{I} = \frac{1}{l}\sum_{i=1}^{l}(y_i - (\boldsymbol{a}\mathbf{x}_i))^2,$$

где входные векторы $\mathbf{x}_i = (x_i^1, x_i^2, \ldots, x_i^m)$ и соответствующие значения y_i заданы в обучающей выборке.

Раскрывая скобки, получим

$$\mathbf{I} = \frac{1}{l}\sum_{i=1}^{l} y_i^2 - \frac{2}{l}\sum_{i=1}^{l} y_i(\boldsymbol{a}\mathbf{x}_i) + \frac{1}{l}\sum_{i=1}^{l}(\boldsymbol{a}\mathbf{x}_i)^2.$$

Далее

$$\frac{1}{l}\sum_{i=1}^{l} y_i(\boldsymbol{a}\mathbf{x}_i) = \left(\boldsymbol{a}, \frac{1}{l}\sum_{i=1}^{l} y_i\mathbf{x}_i\right) = (\boldsymbol{a}, \mathbf{r}),$$

где координаты вектора $\mathbf{r} = (1/l)\sum_{i=1}^{l} y_i\mathbf{x}_i$ равны

$$r^t = \frac{1}{l}\sum_{i=1}^{l} y_i x_i^t \qquad (t = 1, \ldots, m)$$

и имеют смысл эмпирических коэффициентов ковариации между выходной величиной и t-ой координатой входного вектора.

Легко убедиться, что квадратичная часть выражения может быть представлена как

$$\frac{1}{l}\sum_{i=1}^{l}(\boldsymbol{a}\mathbf{x}_i)^2 = \boldsymbol{a}\mathbf{Q}\boldsymbol{a}^{\mathrm{T}},$$

где элементы матрицы $\mathbf{Q}$ равны

$$q_{ts} = \frac{1}{l}\sum_{i=1}^{l} x_i^t x_i^s,$$

то есть имеют смысл эмпирических коэффициентов ковариации между t-ой и s-ой координатами входного вектора. Действительно

$$\frac{1}{l}\sum_{i=1}^{l}(\boldsymbol{a}\mathbf{x}_i)^2 = \frac{1}{l}\sum_{i=1}^{l}\left(\sum_{s=1}^{m} a_s x_s\right)^2 = \frac{1}{l}\sum_{i=1}^{l}\left(\sum_{s,t=1}^{m} a_s x_i^s a_t x_i^t\right) =$$
$$= \sum_{s,t=1}^{m} a_s a_t \left(\frac{1}{l}\sum_{i=1}^{l} x_i^s x_i^t\right) = \boldsymbol{a}\mathbf{Q}\boldsymbol{a}^{\mathrm{T}}.$$

В этих обозначениях

$$\mathbf{I} = \frac{1}{l}\sum_{i=1}^{l} y_i^2 - 2(\boldsymbol{a}\mathbf{r}) + \boldsymbol{a}^{\mathrm{T}}\mathbf{Q}\boldsymbol{a}.$$

Чтобы найти минимум этого выражения, приравняем к нулю его градиент по $\boldsymbol{a}$:

$$\mathbf{grad}(\mathbf{I}) = 2(\boldsymbol{a}\mathbf{Q} - \mathbf{r}) = 0 \quad \text{или} \quad \mathbf{r} = \boldsymbol{a}\mathbf{Q}.$$

(Последнее уравнение в координатной форме получило название системы нормальных уравнений метода наименьших квадратов).

Отсюда

$$\boldsymbol{a}_{\text{опт}} = \mathbf{Q}^{-1}\mathbf{r}. \tag{3}$$

Конечно, это решение имеет смысл только в том случае, когда матрица Q невырождена (имеет обратную). Это матрица будет вырождена в том и только том случае, когда все векторы $\mathbf{x}_i$ обучающей выборки лежат в гиперпространстве размерности меньшей, чем m. Понятно, что в этом случае поведение зависимости в направлениях, ортогональных этому подпространству, по обучающей выборке определить невозможно. Если теперь вернуться к нецентрированным координатам, то искомая зависимость приобретает вид:

$$F(\mathbf{x}) = y_{\text{ср}} + ((\mathbf{x} - \mathbf{x}_{\text{ср}})\mathbf{Q}^{-1}\mathbf{r}),$$

где матрица $\mathbf{Q}$ и вектор $\mathbf{r}$ вычисляются по центрированным значениям.

Теперь попытаемся выяснить, насколько точно мы оцениваем наилучшее линейное приближение регрессии. В общем случае это очень трудно сделать. Однако в следующем частном случае это удается. Пусть истинная регрессия $E(y|\mathbf{x})$ действительно является линейной функцией $(\boldsymbol{a}_0\mathbf{x})$, а наблюдения y_i искажены некоррелированным шумом с нулевым средним и постоянной дисперсией D. То есть

$$y_i = (\boldsymbol{a}_0\mathbf{x}_i) + z_i, \quad Ez_i = 0, \quad Ez_i^2 = D, \quad Ez_iz_i = 0.$$

На этот раз удобнее избавиться от свободного члена путем добавления дополнительной координаты, тождественно равной 1. Все результаты, полученные для центрированных величин, при этом остаются в силе. (Проверить!)

Значения эмпирических коэффициентов ковариации выходной величины $\mathbf{y}$ с координатами вектора $\mathbf{x}$ будут равны

$$r^t = \frac{1}{l}\sum_{i=1}^{l} y_i x_i^t = \frac{1}{l}\sum_{i=1}^{l}[(\boldsymbol{a}_0\mathbf{x}_i) + z_i]x_i^t =$$
$$= \frac{1}{l}\sum_{i=1}^{l}(\boldsymbol{a}_0\mathbf{x}_i)x_i^t + \frac{1}{l}\sum_{i=1}^{l} z_i x_i^t.$$

Соответственно, вектор этих ковариаций представим в виде суммы $\mathbf{r} = \mathbf{r}_0 + \mathbf{r}_1$, с координатами

$$r_0^t = \frac{1}{l}\sum_{i=1}^{l}(\boldsymbol{a}_0\mathbf{x}_i)x_i^t, \quad r_1^t = \frac{1}{l}\sum_{i=1}^{l} z_i x_i^t.$$

Тогда согласно (3) вектор оценки коэффициентов регрессии равен

$$\boldsymbol{a}_{\text{опт}} = \mathbf{Q}^{-1}\mathbf{r} = \mathbf{Q}^{-1}(\mathbf{r}_0 + \mathbf{r}_1) = \mathbf{Q}^{-1}\mathbf{r}_0 + \mathbf{Q}^{-1}\mathbf{r}_1.$$

Но вектор $a^0_{\text{опт}} = \mathbf{Q}^{-1}\mathbf{r}_0$ соответствует оценке коэффициентов регрессии в случае отсутствия шума, потому совпадает с $\boldsymbol{a}_0$, а вектор $a^1_{\text{opt}} = \mathbf{Q}^{-1}\mathbf{r}_1$ представляет собой отклонение оценки коэффициентов регрессии от их истинного значения. Он имеет нулевое среднее (поскольку и $\mathbf{r}_1$ имеет нулевое среднее), а ковариационную матрицу его координат мы сейчас найдем. Но предварительно нам нужно найти ковариационную матрицу $\mathbf{K}$ координат вектора $\mathbf{r}_1$. Ее элементы

$$k_{st} = E\left(\frac{1}{l}\sum_{i=1}^{l} z_i x_i^s \cdot \frac{1}{l}\sum_{i=1}^{l} z_i x_i^t\right) = \frac{1}{l^2}\sum_{i,j=1}^{l} x_i^s x_i^t E(z_i z_j).$$

Но величина $E(z_i z_j)$ отлична от нуля только при $i = j$ и равна в этом случае D. Поэтому

$$k_{st} = \frac{D}{l^2}\sum_{i=1}^{l} x_i^s x_i^t.$$

Таким образом, мы видим, что элементы матрицы $\mathbf{K}$ отличаются от элементов матрицы эмпирических коэффициентов ковариации входных переменных $\mathbf{Q}$ только множителем D/l. Поэтому $\mathbf{K} = (D/l)\mathbf{Q}$. Теперь, используя формулы (1) и (3), мы можем получить ковариационную матрицу $\mathbf{C}$ вектора погрешностей $\boldsymbol{a}^1_{\text{опт}}$:

$$\mathbf{C} = \mathbf{Q}^{-1}\mathbf{K}\mathbf{Q}^{-1} = \frac{D}{l}\mathbf{Q}^{-1}. \tag{4}$$

Итак, мы видим, что ковариации погрешностей (в том числе и дисперсии) в определении коэффициентов регрессии пропорциональны дисперсии шума и обратно пропорциональны длине выборки l. Однако реальный смысл матрицы $\mathbf{Q}^{-1}$ остается неясным (хотя, конечно, ее

можно вычислить). Для того, чтобы пояснить этот смысл, заметим, что поворотом координат матрица $\mathbf{Q}$ всегда может быть приведена к диагональному виду. При этом новые оси координат будут направлены вдоль собственных векторов матрицы $\mathbf{Q}$, а диагональные элементы преобразованной матрицы $\mathbf{Q}^*$ будут равны собственным числам λ_t матрицы $\mathbf{Q}$. То есть элементы этой матрицы

$$q^*_{tt} = \frac{1}{l}\sum_{i=1}^{l} x^{*t}_i x^{*t}_i = \lambda_t,$$

$$q^*_{ts} = \frac{1}{l}\sum_{i=1}^{l} x^t_i x^s_i = 0, \qquad t \neq s.$$

Тогда на основании (4) получаем, что погрешности определения коэффициентов регрессии в новых координатах будут некоррелированы, а дисперсия погрешности t-ого коэффициента равна

$$D_t = \frac{D}{l\lambda_t}.$$

Теперь видно, что дисперсия тем меньше, чем больше соответствующее собственное число λ_t, и становится большой при малых λ_t. Вычислим еще математическое ожидание остаточной невязки

$$E\mathbf{I} = E\left[\frac{1}{l}\sum_{i=1}^{l}(y_i - (\boldsymbol{a}_{\text{опт}}\mathbf{x}_i))^2\right].$$

Опять представим выходной сигнал как сумму истинной регрессии и шума $y_i = (\boldsymbol{a}_0\mathbf{x}_i) + z_i$, а вектор коэффициентов $\boldsymbol{a}_{\text{опт}}$ как $\boldsymbol{a}_{\text{опт}} = \boldsymbol{a}^0_{\text{опт}} + \boldsymbol{a}^1_{\text{опт}}$, где $\boldsymbol{a}^0_{\text{опт}} = \mathbf{Q}^{-1}\mathbf{r}_0$ соответствует оценке коэффициентов регрессии в случае отсутствия шума, потому совпадает с $\boldsymbol{a}_0$, а вектор $\boldsymbol{a}^1_{\text{опт}} = \mathbf{Q}^{-1}\mathbf{r}_1$ представляет собой отклонение оценки коэффициентов регрессии от их истинного значения. Тогда

$$E\mathbf{I} = E\left[\frac{1}{l}\sum_{i=1}^{l}\left(((\boldsymbol{a}_0\mathbf{x}_i) + z_i) - ((\boldsymbol{a}^0_{\text{опт}} + \boldsymbol{a}^1_{\text{опт}})\mathbf{x}_i)\right)^2\right] =$$

$$= E\left[\frac{1}{l}\sum_{i=1}^{l}\left(z_i - (\boldsymbol{a}^1_{\text{опт}}\mathbf{x}_i)\right)^2\right].$$

Проводя преобразования аналогичные тем, что проводились выше, получим

$$\mathbf{I} = \frac{1}{l}\sum_{i=1}^{l} z_i^2 - 2(\boldsymbol{a}^1_{\text{опт}}\mathbf{r}_1) + \boldsymbol{a}^1_{\text{опт}}\mathbf{Q}\boldsymbol{a}^1_{\text{опт}},$$

где $\mathbf{r}_1$ — эмпирический коэффициентов шумовой составляющей z с координатами вектора $\mathbf{x}$.

Подставляя сюда значение $\boldsymbol{a}^1_{\text{опт}} = \mathbf{Q}^{-1}\mathbf{r}_1$, получим

$$\mathbf{I} = \frac{1}{l}\sum_{i=1}^{l} z_i^2 - \mathbf{r}_1\mathbf{Q}^{-1}\mathbf{r}_1.$$

Считая, что координаты уже повернуты так, что матрица $\mathbf{Q}$ приведена к диагональному виду с диагональными элементами λ_t, получим

$$\mathbf{r}_1\mathbf{Q}^{-1}\mathbf{r}_1 = \sum_{t=1}^{m}\frac{r_{1t}^2}{\lambda_t}.$$

Переходя к математическим ожиданиям, получим

$$E\mathbf{I} = D - \sum_{t=1}^{m}\frac{Er_{1t}^2}{\lambda_t}.$$

Но величины $E(r_{1s}r_{1t})$ суть не что иное, как элементы ковариационной матрицы $\mathbf{K}$ координат вектора $\mathbf{r}_1$, а, как было установлено, $\mathbf{K} = (D/l)\mathbf{Q}$. Поэтому $E(r_{1s}r_{1t}) = 0$ при $t \neq s$ и $E(r_{1t}^2) = (D/l)\lambda_t$. Откуда

$$E\mathbf{I} = D - \sum_{i=1}^{m}\frac{(D/l)\lambda_t}{\lambda_t} = D - \frac{mD}{l} = D\frac{l-m}{l}.$$

Разумеется, это верно только если матрица $\mathbf{Q}$ невырождена. Напомним, что число аргументов m равно числу исходных аргументов плюс 1 за счет дополнительного аргумента, соответствующего свободному члену. Заметим, что при $m > l$ оценка оказывается отрицательной, но при этом матрица $\mathbf{Q}$ всегда будет вырождена. При $m = l$ $E\mathbf{I}$ равно нулю, и действительно, в этом случае линейную зависимость всегда можно подогнать так, чтобы невязки не было. Кроме того, отметим, что оценка получена при известной дисперсии шума D. Из полученного результата следует, что величина $\mathbf{I}(l/(l-m))$, где $\mathbf{I}$ — остаточная

невязка МНК, служит несмещенной оценкой дисперсии шума. Поэтому возникает соблазн использовать ее как оценку истинной дисперсии. Но несмещенная не значит близкая к истине, и вопрос о том насколько эта оценка близка к истине, требует дополнительного исследования.

Лекция 5

Задача распознавания образов. Поиск решающего правила, минимизирующего число ошибок или среднее значение функции штрафа на данных обучения, в задачах распознавания образов. Разделение двух нормально распределенных совокупностей. Наивный Байес. Метод ближайшего соседа. (Стандартная процедура распознавания по ближайшему соседу. Процедуры нахождения дискриминантной функции)

Задача обучения распознаванию образов

Теперь мы перейдем к другой проблеме — задаче распознавания образов [4, 5, 9, 24]. Здесь считается, что учитель (человек или природа), получив на вход описание $\mathbf{x}$, относит его к одному из конечного числа классов, то есть дает на выходе одно из конечного множества значений $\mathbf{y}$ — имен классов. Задача состоит в том, чтобы научиться классифицировать входные описания по возможности так же, как это делает учитель. Иными словами, требуется на основании обучающей выборки

$$(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_l, \mathbf{y}_l)$$

построить решающее правило $F(\mathbf{x})$ — функцию, принимающую в качестве свих значений те же имена классов, и минимально отличающуюся от поведения учителя. Входные описания могут иметь любую природу. Это может быть числовой вектор, или последовательность значений дискретных признаков или смесь числовых значений и дискретных признаков. Это могут быть изображения, непрерывные сигналы, графы, текстовые строки (не обязательно одинаковой длины), просто тексты и т.д. Количество классов тоже варьирует в широких пределах — от двух до очень большого числа, например, при распознавании людей по изображению их лиц или радужных оболочек глаза.

Обычно предполагается, что имеется некоторая функция штрафа $Q(\mathbf{y}, \mathbf{y}^*)$ за различие между классификацией учителя $\mathbf{y}$ и классификацией $\mathbf{y}^* = F(\mathbf{x})$, предлагаемой решающим правилом. В простейшем

случае $Q(\mathbf{y}, \mathbf{y}^*)$ равно нулю, если $\mathbf{y}$ и $\mathbf{y}^*$ совпадают, и равно 1 в противном случае, т. е. штраф всегда одинаков в случае ошибки. В общем случае штраф может быть различным за ошибки разного рода, например, за пропуск цели и ложную тревогу. Также в общем случае считается, что учитель может действовать недетерминировано, т. е. поведение учителя описывается (неизвестным) условным распределением $p(\mathbf{y}|\mathbf{x})$.Тогда качество распознавания конкретного описания x может быть записано как

$$R(\mathbf{x}, F) = \sum Q(\mathbf{y}, F(\mathbf{x}))p(\mathbf{y}|\mathbf{x}),$$

где сумма берется по всем возможным значениям $\mathbf{y}$ (по всем возможным классификациям). Понятно, что безошибочная классификация возможна только в том случае, когда учитель ведет себя детерминировано.

Если на множестве описаний также задано распределение вероятностей $P(\mathbf{x})$, то определено математическое ожидание штрафа, называемое истинным риском:

$$R(F)_{\text{ист}} = E_x R(\mathbf{x}, F) = \int \left(\sum Q(\mathbf{y}, F(\mathbf{x}))p(\mathbf{y}|\mathbf{x}) \right) dP(\mathbf{x}) = \\ = \int Q(\mathbf{y}, F(\mathbf{x}))\, dP(\mathbf{x}, \mathbf{y}).$$

Задача состоит в том, чтобы по обучающий выборке найти решающее правило $F(\mathbf{x})$ из заданного класса, минимизирующее истинный риск. Возможный и широко применяемый путь решения задачи состоит в том, что истинный риск, который нам неизвестен, заменяется на эмпирический риск — среднее арифметическое значение риска на обучающей выборке:

$$R(F)_{\text{эмп}} = \frac{1}{l} \sum_{i=1}^{l} Q(\mathbf{y}_i, F(\mathbf{x}_i)),$$

который уже можно вычислить, и ищется решающее правило в заданном классе, минимизирующее эмпирический риск. Интуитивное обоснование такого подхода состоит в том, что если обучающая выборка получена при том же распределении вероятностей $P(\mathbf{x}, \mathbf{y})$ в процессе

независимых испытаний, то в силу закона больших чисел эмпирический риск сходится к истинному, и функция, доставляющая минимум эмпирическому риску, будет близка к функции, доставляющей минимум истинному риску. Конечно, поиск решающего правила, которое минимизирует эмпирический риск, может представлять собой сложную вычислительную задачу, но это уже не имеет отношения к статистике.

Возможен и другой подход к задаче распознавания образов. Предполагается, что объекты некоторого множества «объективно» принадлежат разным классам $\mathbf{y}$. Но даны они нам в виде своих описаний $\mathbf{x}$. Поскольку внутри каждого класса объектов много, и их связь с описанием в общем случае не детерминирована, с каждым классом связано некоторое распределение вероятностей $P_{\mathbf{y}}(\mathbf{x}) = P(\mathbf{x}|\mathbf{y})$ на множестве описаний. Если, кроме того, известны вероятности самих классов $P(\mathbf{y})$, то по формуле Байеса можно найти условную вероятность принадлежности объекта к классу $\mathbf{y}$ при заданном описании $\mathbf{x}$:

$$P(\mathbf{y}|\mathbf{x}) = \frac{P(\mathbf{y})P(\mathbf{x}|\mathbf{y})}{c}, \qquad c = \sum_{\mathbf{y}} P(y)P(\mathbf{x}|\mathbf{y}),$$

где c — нормировочный коэффициент.

Тогда оптимальное решающее правило будет таково:

$$F(\mathbf{x}) = \arg\min_{\mathbf{y}^*} \sum_{\mathbf{y}} P(\mathbf{y}|\mathbf{x})Q(\mathbf{y}, \mathbf{y}^*).$$

А в случае равной платы за ошибку оптимальным будет правило

$$F(\mathbf{x}) = \arg\max_{\mathbf{y}^*} P(\mathbf{y}^*|\mathbf{x}),$$

то есть это правило должно относить объект с описанием $\mathbf{x}$ к тому классу $\mathbf{y}^*$, для которого условная вероятность $P(\mathbf{y}^*|\mathbf{x})$ максимальна.

Проблема только в том, что условное распределение $P(\mathbf{x}|\mathbf{y})$ нам неизвестно. Но в простых случаях его можно оценить (восстановить) по обучающей выборке.

С этого, собственно, и началось решение задач распознавания образов еще в тридцатых годах XX века (хотя этот термин тогда не использовался).

Разделение двух нормально распределенных совокупностей

Рассмотрим случай, когда имеется всего два класса ($\mathbf{y} = 0$ или $\mathbf{y} = 1$), описание $\mathbf{x}$ представляет собой n-мерный вектор, а описания объектов каждого из классов имеют нормальное распределение, т. е. плотности условных распределений имеют вид:

$$P(\mathbf{x}|\mathbf{y} = 0) = \frac{1}{\sqrt{2\pi \det \mathbf{A}_0}} \exp\left[-(\mathbf{x} - M_0)\mathbf{A}_0^{-1}(\mathbf{x} - M_0)^{\mathrm{T}}\right],$$
$$P(\mathbf{x}|\mathbf{y} = 1) = \frac{1}{\sqrt{2\pi \det \mathbf{A}_1}} \exp\left[-(\mathbf{x} - M_1)\mathbf{A}_0^{-1}(\mathbf{x} - M_1)^{\mathrm{T}}\right],$$

где M_0 и M_1 — центры распределений, а $\mathbf{A}_0$ и $\mathbf{A}_1$ — их ковариационные матрицы.

Примем также, что плата за ошибки разного рода одинакова. Тогда для принятия оптимального решения достаточно сравнить величины $P(\mathbf{y} = 0)P(\mathbf{x}|\mathbf{y} = 0)$ и $P(\mathbf{y} = 1)P(\mathbf{x}|\mathbf{y} = 1)$, где $P(\mathbf{y} = 0)$ и $P(\mathbf{y} = 1)$ — априорные вероятности классов, и отнести объект к тому классу, для которых соответствующая величина больше. Переходя к логарифмам, получаем решающее правило:

$$F(\mathbf{x}) = 1, \quad \text{если} \quad Q(\mathbf{x}) > 0, \qquad F(\mathbf{x}) = 0, \quad \text{если} \quad Q(\mathbf{x}) < 0,$$

где «дискриминантная» функция $Q(\mathbf{x})$ равна:

$$\begin{aligned} Q(x) = {} & \log P(\mathbf{y} = 1) - \log P(\mathbf{y} = 0) - \\ & - \big(\log \det \mathbf{A}_1 - \log \det \mathbf{A}_0 + (\mathbf{x} - M_1)\mathbf{A}_1^{-1}(\mathbf{x} - M_1)^{\mathrm{T}} - \\ & - (\mathbf{x} - M_0)\mathbf{A}_0^{-1}(\mathbf{x} - M_0)^{\mathrm{T}}\big). \end{aligned} \qquad (1)$$

Отсюда видно, что дискриминантная функция $Q(\mathbf{x})$ в нашем случае оказывается квадратичной функцией вектора $\mathbf{x}$, а разделяющая поверхность, определяемая уравнением $Q(\mathbf{x}) = 0$, будет поверхностью второго порядка в пространстве описаний $\mathbf{x}$.

Для современной вычислительной машины вычислить дискриминантную функцию не проблема. Проблема состоит в том, что значения средних M_0 и M_1 и ковариационные матрицы $\mathbf{A}_0$ и $\mathbf{A}_1$ мы должны найти по данным обучения. Оказывается, что, даже если распределения действительно будут нормальными, для построения корректного квадратичного решающего правила при сравнительно большом числе признаков правила нужны очень длинные выборки.

Наивный Байес

Рассмотрим, однако, несколько частных случаев. Допустим, что заведомо известно, что распределения отличаются только средними (их ковариационные матрицы одинаковы), и что координаты векторов независимы и имеют одинаковую дисперсию D. Тогда матрицы $\mathbf{A}_0$ и $\mathbf{A}_1$ будут одинаковыми, диагональными, а их элементы $a_{ii} = D$. Соответственно матрицы $\mathbf{A}_0^{-1}$ и $\mathbf{A}_1^{-1}$ будут тоже диагональными с диагональными элементами $1/D$. Допустим также, что априорные вероятности классов равны. Подставляя эти значения в (1), получим функцию

$$\begin{aligned} Q(x) &= -\frac{1}{D}[(\mathbf{x} - M_1)(\mathbf{x} - M_1)^{\mathrm{T}} - (\mathbf{x} - M_0)(\mathbf{x} - M_0)^{\mathrm{T}}] = \\ &= -\frac{1}{D}[|x - M_1|^2 - |\mathbf{x} - M_0|^2]. \end{aligned}$$

Это значит, что сравниваются расстояния вектора до центров, и вектор $\mathbf{x}$ должен быть отнесен к классу 1, если он ближе к M_1, и к классу 0, если он ближе к M_0. Разделяющей поверхностью будет геометрическое место точек, равноудаленных от M_1 и M_0, то есть гиперплоскость, ортогональная вектору $(M_1 - M_0)$ и проходящая на равных расстояниях от центров. Легко показать, что если априорные вероятности классов не равны, то разделяющая поверхность остается плоской, ортогональной вектору $(M_1 - M_0)$, и просто сдвигается к одному из центров. Рассмотрим другой частный случай. Пусть заведомо известно, что центры распределений совпадают, т. е. $M_1 = M_0 = M$, признаки в каждом классе независимы, имеют одинаковую дисперсию внутри каждого класса, но разную в разных классах, соответственно D_1 и D_0. Тогда матрицы $\mathbf{A}_1$ и $\mathbf{A}_0$ будет диагональными с элементами на диагонали соответственно D_1 и D_0. Подставляя эти значения в (1), получим

$$\begin{aligned} Q(\mathbf{x}) = [\log P(\mathbf{y} = 1) - \log P(\mathbf{y} = 1)] - \Bigg[& n \log D_1 - n \log D_0 + \\ & + \left(\frac{1}{D_1} - \frac{1}{D_0} \right) |(\mathbf{x} - M)|^2 \Bigg]. \end{aligned}$$

Приравнивая это выражение к нулю, получим, что разделяющая поверхность представляет собой сферу некоторого радиуса с центром в точке M. Точки внутри сферы относятся к тому классу, у которого

дисперсия координат меньше, а снаружи — к тому классу, у которого дисперсия больше. Рассмотрим теперь более общий (но все равно частный) случай, когда центры классов M_0 и M_1 различны, а ковариационные матрицы одинаковы, т. е. $\mathbf{A}_0 = \mathbf{A}_1 = \mathbf{A}$. Это именно тот случай, для которого Роберт Фишер получил результат еще в 1938 году. Для простоты будем считать, что $P(\mathbf{y} = 0) = P(\mathbf{y} = 1)$. Формула (1) в этом случае приобретает вид:

$$Q(\mathbf{x}) = -[(\mathbf{x} - M_1)\mathbf{A}^{-1}(\mathbf{x} - M_1)^{\mathrm{T}} - (\mathbf{x} - M_0)\mathbf{A}^{-1}(\mathbf{x} - M_0)^{\mathrm{T}}].$$

Раскрывая скобки, преобразуем ее к виду

$$\begin{aligned} Q(\mathbf{x}) &= 2\left[\mathbf{x}\mathbf{A}^{-1}(M_1 - M_0)^{\mathrm{T}} - \frac{1}{2}(M_1\mathbf{A}^{-1}M_1^{\mathrm{T}} - M_0\mathbf{A}^{-1}M_0^{\mathrm{T}})\right] = \\ &= 2\left[\mathbf{x}\mathbf{A}^{-1}(M_1 - M_0)^{\mathrm{T}} - \frac{1}{2}(M_1 - M_0)\mathbf{A}^{-1}(M_1 + M_0)^{\mathrm{T}}\right]. \end{aligned}$$

Таким образом, видим, что квадратичные члены в дискриминантной функции пропадают, и она становится линейной по x, а решающее правило становится линейным. Функция $Q(\mathbf{x})$ в таком виде получила название «дискриминантной функции Фишера». Соответствующая разделяющая поверхность становится гиперплоскостью в пространстве описаний с уравнением

$$\mathbf{x}\mathbf{A}^{-1}(M_1 - M_0)^{\mathrm{T}} - \frac{1}{2}(M_1 - M_0)\mathbf{A}^{-1}(M_1 + M_0)^{\mathrm{T}} = 0.$$

В силу симметрии задачи она по-прежнему проходит через середину отрезка (M_1, M_0), но уже не ортогонально вектору $(M_1 - M_0)$. Для нахождения решающего правила в этом случае тоже, как и в общем случае, требуется оценить ковариационную матрицу по данным обучения. Но поскольку разделяющая поверхность стала существенно менее сложной, то чувствительность решения к ошибкам оценивания оказывается гораздо более слабой. Поэтому построение линейной дискриминантной функции методом Фишера широко применяется даже в тех случаях, когда нет веских оснований считать ковариационные матрицы классов одинаковыми. Литовский ученый Шарунас Раудис проделал такой эксперимент. Он искусственно сгенерировал две совокупности нормально распределенных векторов сравнительно большой размерности, разделив их на данные обучения и контрольную выборку.

При этом ковариационные матрицы классов были не одинаковы. Далее он строил решающее правило по обучающей выборке и проверял качество его работы на контрольной. Оказалось, что для того, чтобы квадратичное правило стало действительно оптимальным, необходима чрезвычайно большая выборка. При более умеренной длине обучающей выборки оказалось, что линейное решающее правило, построенное при неверном предположении о равенстве ковариационных матиц, дает лучший результат. При коротких выборках лучшим оказалось правило, работающее просто на сравнении расстояний от вектора до центров распределения классов. Напомним, что по эмпирическим данным координаты центров $M_1 = (m_1^1, m_2^1, \ldots, m_n^1)$ и $M_0 = (m_1^0, m_2^0, \ldots, m_n^0)$ считаются как

$$m_j^1 = \frac{1}{l_1}\sum_{i=1}^{l_1} x_i^j, \quad m_j^0 = \frac{1}{l_0}\sum_{i=1}^{l_0} x_i^j,$$

где i — номер вектора, j — номер координаты, l_1 — число векторов класса 1 в обучающей выборке, l_0 — число векторов класса 0, первая сумма берется по векторам класса 1, вторая — по векторам класса 0. Значения эмпирических коэффициентов ковариации (элементов матриц $\mathbf{A}_0$ и $\mathbf{A}_1$) считаются как

$$a_{st}^1 = \frac{1}{l_1}\sum_{i=1}^{l_1}(x_i^s - m_s^1)(x_i^t - m_t^1), \quad a_{st}^0 = \frac{1}{l_0}\sum_{i=1}^{l_0}(x_i^s - m_s^0)(x_i^t - m_t^0),$$

где i — номер вектора, s и t — номера координат, первая сумма берется по векторам класса 1, вторая сумма — по векторам класса 0. Если же считается, что матрицы $\mathbf{A}_1$ и $\mathbf{A}_0$ одинаковы ($\mathbf{A}_1 = \mathbf{A}_0 = \mathbf{A}$), то элементы матрицы $\mathbf{A}$ считаются как

$$a_{st} = \frac{1}{l}\left(\sum_{i=1}^{l_1}(x_i^s - m_s^1)(x_i^t - m_t^1) + \sum_{i=1}^{l_0}(x_i^s - m_s^0)(x_i^t - m_t^0)\right),$$

где l — общая длина обучающей выборки, первая сумма берется по векторам класса 1, вторая сумма — по векторам класса 0.

Рассмотрим еще один случай, когда обучение распознаванию идет через восстановление распределений на множестве описаний. Пусть каждый объект описывается фиксированной последовательностью признаков $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, каждый из которых принимает всего

два значения — 0 или 1. (Можно рассмотреть и вариант, когда число дискретных значений признаков больше двух, но для простоты ограничимся случаем двух значений.) Признаки считаются независимыми. Вероятность того, что i-тый признак примет значение 1 при условии, что объект принадлежит классу 0, обозначим p_0^i, тогда значение 0 он примет с вероятностью $q_0^i = 1 - p_0^i$. Аналогично, для другого класса обозначим эти же величины p_1^i и q_1^i. Тогда вероятность получить вектор $\mathbf{x}$ при условии, что объект принадлежит классу 0, равна

$$P_0(\mathbf{x}) = P(\mathbf{x}|\mathbf{y} = 0) = \prod_{i=1}^{n} t_i^0,$$

где $t_i^0 = p_0^i$, если $x_i = 1$, а $t_i^0 = q_0^i$, если $x_i = 0$, и, аналогично,

$$P_1(\mathbf{x}) = P(\mathbf{x}|\mathbf{y} = 1) = \prod_{i=1}^{n} t_i^1,$$

где $t_i^1 = p_1^i$, если $x_i = 1$, и где $t_i^1 = q_1^i$, если $x_i = 0$. Примем, что плата за ошибки разного рода одинакова. Тогда для принятия оптимального решения достаточно сравнить величины $P(\mathbf{y} = 0)P(\mathbf{x}|\mathbf{y} = 0)$ и $P(\mathbf{y} = 1)P(\mathbf{x}|\mathbf{y} = 1)$, где $P(\mathbf{y} = 0)$ и $P(\mathbf{y} = 1)$ — априорные вероятности классов, и отнести объект к тому классу, для которых соответствующая величина больше. Переходя к логарифмам, получаем решающее правило:

$$F(\mathbf{x}) = 1, \quad \text{если} \quad Q(\mathbf{x}) > 0, \qquad F(x) = 0, \quad \text{если} \quad Q(\mathbf{x}) < 0,$$

где «дискриминантная» функция $Q(x)$ равна:

$$Q(\mathbf{x}) = [\log P(\mathbf{y} = 1) - \log P(\mathbf{y} = 0)] + \sum_{i=1}^{n} [\log t_i^1 - \log t_i^0]. \qquad (2)$$

Это решающее правило получено для известных значений вероятностей p_i^1 и p_i^1. На самом деле их приходится оценивать по обучающей выборке. Для этого подсчитывается число n_i^1 объектов класса 1 в обучающей выборке, для которых i-тый признак принял значение 1, и вероятность оценивается, как $p_i^1 = n_i^1/l_i^1$, где l_i^1 — общее число объектов класса 1 в обучающей выборке. Аналогично, p_i^0 оценивается как

$p_i^0 = n_i^0/l_i^0$. Однако этот путь опасен. Достаточно случиться тому, что ни на одном объекте класса 1 в обучающей выборке i-тый признак не принял значение 1(или 0). Тогда соответствующая оценка вероятности обращается в нуль, а ее логарифм — в минус бесконечность. Поэтому один признак может определить окончательное решение. Более осторожной оказывается оценка

$$p_i^1 = \frac{n_i^1 + 1}{l_i^1 + 2}, \quad p_i^0 = \frac{n_i^0 + 1}{l_i^0 + 2}.$$

Такая оценка получается в результате байесовой процедуры в предположении, что значение априори распределено равномерно между нулем и единицей. Более подробно мы будем говорить при рассмотрении байесовых схем. Этот путь, получивший название «наивный Байес», несмотря на то, что он основан на очень простых предположениях, оказывается в ряде случаев вполне конкурентоспособным по сравнению с более сложными алгоритмами. Особенно часто это встречается в задачах медицинской диагностики. Заметим, что полученное решающее правило также относится к классу линейных, то есть дискриминантная функция линейна относительно вектора $\mathbf{x}$. Действительно величина $[\log t_i^1 - \log t_i^0]$ равна $[\log p_i^1 - \log p_i^0]$, если признак принял значение 1, и равна $[\log q_i^1 - \log q_i^0]$, если признак принял значение 0. Поэтому можно записать

$$\begin{aligned}[\log t_i^1 - \log t_i^0] &= [\log p_i^1 - \log p_i^0]x_i + [\log q_i^1 - \log q_i^0](1 - x_i) = \\ &= [\log q_i^1 - \log q_i^0] + ([\log p_i^1 - \log p_i^0] - [\log q_i^1 - \log q_i^0])x_i = \\ &= a_i + b_i x_i,\end{aligned}$$

где a_i и b_i — настраиваемые величины. Сама же дискриминантная функция приобретает вид

$$Q(\mathbf{x}) = [\log P(\mathbf{y} = 1) - \log P(\mathbf{y} = 0)] + \sum_{i=1}^{n} a_i + \sum_{i=1}^{n} b_i x_i.$$

Метод ближайшего соседа

До сих пор мы задавались вполне конкретными предположениями относительно распределений в пространстве описаний. Более общее предположение приводит нас к так называемому методу ближайшего соседа

(точнее, k ближайших соседей). Предположим, что плотности распределения $P(\mathbf{x}|\mathbf{y})$, где $\mathbf{y}$ — имя класса, меняются не слишком быстро в пространстве описаний. Тогда можно попытаться оценить эти плотности вблизи любой заданной точки $\mathbf{x}$, просто подсчитывая частоты выпадения описаний объектов разных классов в некоторой окрестности этой точки по данным обучения. Точнее, выбирается окрестность минимального размера, такая, что в нее попадает ровно k ближайших соседей из обучающей выборки (не обращая внимания на то, к какому классу они отнесены). Обычно при этом используется евклидова метрика, но, в принципе, можно применить и любую другую. Далее подсчитывается число точек $k_1, k_2, \ldots, k_N$ разных классов, попавших в эту окрестность. Теперь относительные частоты классов

$$\nu_i = \frac{k_i}{k}$$

принимаются за оценки, пропорциональные плотности распределения каждого класса $P(\mathbf{x}|\mathbf{y})$ в точке $\mathbf{x}$. В случае равных априорных вероятностей классов и равной платы за ошибку, точка $\mathbf{x}$ относится к тому классу, для которого величина ν_i максимальна. (Если это не так, следует действовать по общей схеме, описанной выше). Число соседей k является параметром алгоритма. Понятно, что если k мало, то оценки будут очень ненадежными. Если же окрестность выбрать достаточно большой, чтобы в нее попало много точек, то нет оснований считать, что плотности слабо меняются в пределах такой окрестности. Расчет здесь на то, что носители распределений $P(\mathbf{x}|\mathbf{y})$ не пересекаются или слабо пересекаются. Тогда в достаточно малой окрестности почти любой точки x, встречающейся в обучении или при работе, представители одного из классов будут заметно превалировать. В этом случае точность оценки существенной роли не играет. Собственно метод ближайшего соседа получается при $k = 1$. В этом случае просто ищется ближайшая к $\mathbf{x}$ точка обучающей выборки, и описание $\mathbf{x}$ относится к тому же к классу, к которому принадлежит эта точка выборки. Расчет здесь на то, что носители распределений образуют непересекающиеся компактные «острова», разделенные достаточно широкими проливами. Тогда всякая точка, полученная на основании этих распределений, будет близка к одному из островов, и ее следует отнести к тому классу, который породил этот остров. Надо сказать, что в многомерных пространствах такая интуиция часто не работает. Тем не менее, в простых

задачах метод ближайшего соседа часто дает хорошие результаты. Но попробуйте эти методом научиться распознавать выигрышные, проигрышные и ничейные позиции в шахматах, где смещение одной пешки может кардинально изменить оценку позиции.

Лекция 6

Линейные решающие правила. Персептрон. Теорема Новикова. Потенциальные функции.

Линейные решающие правила

Мы видели, что в случае разделения двух нормально распределенных совокупностей с одинаковой матрицей ковариации оптимальным оказывается линейное решающее правило. Также к линейному решающему правилу мы пришли в случае, когда объекты описываются рядом независимых бинарных признаков. Но, например, в первой задаче отклонения от нормального распределения могут приводить к изменению разделяющей поверхности даже в тех случаях, когда отклонения от нормальности происходят далеко от разделяющей плоскости. А во втором случае признаки, сами по себе неинформативные, могут существенно улучшить качество распознавания в совокупности с другими.

Это наводит на мысль, искать оптимальное линейное решающее правило, не связывая его с определенным видом распределения, а следуя принципу замены истинного риска эмпирическим [4]. Предполагается, что объект описывается n числовыми признаками, т. е. вектором размерности n.

Линейным решающим правилом (для случая двух классов 0 и 1) мы называем правило вида

$$F(\mathbf{x}) = 1, \text{ если } Q(x) > 0, \quad F(x) = 0, \text{ если } Q(x) < 0,$$

где функция $Q(\mathbf{x})$ равна $(\boldsymbol{a}\mathbf{x} + b)$. Здесь $\boldsymbol{a}$ — настраиваемый вектор, а b — также настраиваемый порог (скаляр).

Цель — найти такие значения $\boldsymbol{a}$ и b, которые минимизируют истинный риск, а в случае равной платы за ошибку — вероятность ошибки (в дальнейшем мы ограничимся этим случаем).

Тогда критерием для выбора параметров будет

$$I(\boldsymbol{a}, b) = \frac{1}{l} \sum |y_i - F(\mathbf{x}_i)| \qquad (y_i = 0 \text{ или } y_i = 1),$$

где пары $(y_i, \mathbf{x}_i)$ составляют обучающую выборку.

Иными словами мы хотим построить линейное решающее правило, минимизирующее число ошибок на обучающей выборке. В случае, когда точки разных классов в обучающей выборке могут быть безошибочно разделены гиперплоскостью, задача сводится к решению системы линейных неравенств:

$$\begin{aligned} (\boldsymbol{a}\mathbf{x}_i + b) > 0, &\quad \text{если } x_i \text{ принадлежит классу 1,} \\ (\boldsymbol{a}\mathbf{x}_i^* + b) < 0, &\quad \text{если } x_i^* \text{ принадлежит классу 0.} \end{aligned}$$

Эта система совместна, если полное разделение возможно (и наоборот). Обычно эту систему несколько ужесточают и используют систему неравенств вида:

$$\begin{aligned} (\boldsymbol{a}\mathbf{x}_i + b) > \varepsilon, &\quad \text{если } x_i \text{ принадлежит классу 1,} \\ (\boldsymbol{a}\mathbf{x}_i^* + b) < \varepsilon, &\quad \text{если } x_i^* \text{ принадлежит классу 0,} \end{aligned}$$

где величина $\varepsilon > 0$ — параметр алгоритма. Эта система также заведомо совместна, если классы в данных обучения разделимы гиперплоскостью. Для решения системы могут быть применены стандартные процедуры линейного программирования. Другие подходы к их решению мы рассмотрим в следующей лекции.

В случае, когда полной разделимости нет, и система неравенств несовместна, приходится действительно минимизировать число ошибок. Неизвестно процедуры, которая бы без перебора гарантированно находила оптимальное в этом смысле линейное решающее правило. Вместо этого, либо используют те или иные эвристические приемы, либо вводят искусственную штрафную функцию, зависящую от удаленности точки от разделяющей плоскости (в неправильную сторону), и тем сводят задачу к выпуклой, которая уже может быть решена регулярными методами.

Персептрон

В 1957 году американский физиолог Ф. Розенблатт предпринял попытку физически реализовать физиологическую модель восприятия.

Эта схема состояла из сети нейронов. Согласно распространенной тогда и наиболее постой модели нейрона (модели Мак-Калока и Питса) нейрон — это устройство, имеющее несколько входов — дендритов и один выход — аксон. Входы бывают либо возбуждающими, либо тормозящими. Нейрон возбуждается и посылает импульс на выход, если число сигналов, пришедших по возбуждающим входам, превосходит число сигналов, пришедших по тормозящим входам нейрона. Вся модель персептрона состояла из слоя рецепторов S, слоя преобразующих нейронов A и слоя реагирующих нейронов R.

Внешнее раздражение воспринимается рецепторами. Каждый рецептор связан с входом одного или нескольких нейронов преобразующего слоя, а каждый нейрон этого слоя связан с несколькими рецепторами. Выходы преобразующих нейронов в свою очередь связаны с входами нейронов третьего слоя — реагирующих нейронов. Каждый такой нейрон ответственен за распознавание одного из классов внешних сигналов — если этот нейрон возбуждается, то схема распознает входной сигнал, как принадлежащий соответствующему классу (образу). В отличие от нейронов преобразующего слоя, реагирующие нейроны имеют разные по величине коэффициенты суммирования и, возможно, по знаку. Эти то коэффициенты настраиваются в процессе обучения.

Обучение проходит в следующем режиме. Предположим, что на вход реагирующего нейрона пришел некоторый пакет импульсов. Если нейрон среагировал правильно, т. е. его реакция совпадает с указанием учителя, то веса не меняются. Если нейрон сработал, а этого не должно было произойти, потому что учитель указал, что объект не принадлежит данному классу, то веса суммирования по тем каналам, по которым пришли импульсы, уменьшаются не единицу. Если же произошла ошибка другого рода, и нейрон не сработал, хотя учитель указывает, что это нужно было сделать, то веса суммирования по каналам, где прошли импульсы, увеличиваются не единицу.

Для формализации работы персептрона допустим, что каждый вход (дендрит) реагирующего нейрона соединен ровно с одним аксоном одного из преобразующих нейронов. Рассмотрим один реагирующий нейрон, предназначенный для распознавания класса P. Обозначим $\mathbf{r} = (r_1, \ldots, r_n)$ вектор весов суммирования этого нейрона, и пусть вектор $\mathbf{z}(t) = (z_1, \ldots, z_n)$ таков, что $z_i = 1$, если в момент t по i-тому входу прошел импульс, и $z_i = 0$, в противном случае.

Нейрон сработает, если

$$\sum_{i=1}^{n} r_i z_i = (\mathbf{r}, \mathbf{z}(t)) > 0.$$

Таким образом, видим, что нейрон реализует линейное решающее правило в пространстве векторов $\mathbf{z}$, и ему соответствует разделяющая гиперплоскость, проходящая через нуль.

Иногда вводят также отличный от нуля порог срабатывания нейрона, когда чтобы нейрон сработал, должно выполниться условие $(\mathbf{r}, \mathbf{z}(t)) > b$. В этом случае нейрону соответствует линейное решающее правило общего вида. Но этот случай сводится предыдущему, если к координатам вектора $\mathbf{z}(t)$ добавить тождественную единицу. Устройство, реализующее такое правило называется *пороговым элементом*. Обозначив $\mathbf{r}(t)$ состояние весов нейрона на момент t, мы можем формально записать алгоритм обучения так:

$\mathbf{r}(t+1) = \mathbf{r}(t)$, если реакция учителя $P^*(t+1)$ совпадает с назначением нейрона P и нейрон сработал, или реакция не совпадает с назначением и нейрон не сработал,

$\mathbf{r}(t+1) = \mathbf{r}(t) + \mathbf{z}(t+1)$, если реакция учителя $P^*(t+1)$ совпадает с назначением нейрона P, и нейрон не сработал,

$\mathbf{r}(t+1) = \mathbf{r}(t) - \mathbf{z}(t+1)$, если реакция учителя $P^*(t+1)$ не совпадает с назначением нейрона P, но нейрон сработал.

Таким образом, видим, что работа одного реагирующего нейрона соответствует распознаванию двух классов: объект принадлежит образу P ($\mathbf{y} = 1$) или не принадлежит ему ($\mathbf{y} = 0$). Обозначим $\mathbf{y}(t)$ реакцию учителя на объект, представленный в момент t, а $\mathbf{y}^*(t)$ — реакцию нейрона:

$$\mathbf{y}^*(t) = 1, \text{ если } (\mathbf{r}(t), \mathbf{z}(t)) > 0; \quad \mathbf{y}^*(t) = 0, \text{ если } (\mathbf{r}(t), \mathbf{z}(t)) \leqslant 0.$$

В этих терминах алгоритм обучения запишется как:

$$\begin{array}{ll} \mathbf{r}(t+1) = \mathbf{r}(t), & \text{если } \mathbf{y}^*(t+1) = \mathbf{y}(t+1), \\ \mathbf{r}(t+1) = \mathbf{r}(t) + \mathbf{z}(t+1), & \text{если } \mathbf{y}^*(t+1) = 0, \text{ но } \mathbf{y}(t+1) = 1, \\ \mathbf{r}(t+1) = \mathbf{r}(t) - \mathbf{z}(t+1), & \text{если } \mathbf{y}^*(t+1) = 1, \text{ но } \mathbf{y}(t+1) = 0. \end{array}$$

В персептроне, предложенном Розенблаттом, нейроны преобразующего слоя соединялись с рецепторами случайным образом, и веса суммирования сигналов также выбирались случайно. Предполагалось, что

так организована связь нейронов в мозгу, и надеялись, что этим обеспечивается инвариантность к разного рода преобразованиям исходного изображения. В последствии выяснилось, что это неверно. Предлагались различные конкретные способы соединения преобразующих нейронов с рецепторами. Для математического анализа работы персептрона вид преобразования не существенен. Важно, что исходное описание объекта x некоторым образом преобразуется в вектор $\mathbf{z}$, подаваемый на вход реагирующего нейрона. И это преобразование зафиксировано до начала обучения и в ходе его не меняется, а настраиваемые реагирующие нейроны реализуют линейные решающие правила в пространстве этих преобразованных векторов.

Теорема Новикова

Естественно, первый же вопрос, который возник при изучении персептрона, насколько эффективен алгоритм обучения, предложенный Розенблаттом. Для случая разделения на два класса на это в какой-то мере дает ответ теорема А. Новикова, доказанная в 1960 году. Теорема утверждает, что если точки этих двух классов, представленные в обучающей выборке, разделимы гиперплоскостью, то алгоритм сойдется и построит такую гиперплоскость. Точная формулировка теоремы такова:

Пусть дана бесконечная последовательность векторов $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l, \ldots$ и для каждого вектора указано, принадлежит он классу 1 ($\mathbf{y}_i = 1$) или классу 0 ($\mathbf{y}_i = 0$). Все векторы ограничены по модулю: для всех i $|\mathbf{x}_i| \leqslant D$, и существует проходящая через нуль разделяющая гиперплоскость, т. е. существует единичный вектор $\boldsymbol{\varphi}^*$, такой, что

$$\begin{aligned} (\boldsymbol{\varphi}^*, \mathbf{x}_i) &\geqslant \rho \quad \text{для всех векторов класса 1,} \\ (\boldsymbol{\varphi}^*, \mathbf{x}_i) &\leqslant -\rho \quad \text{для всех векторов класса 0,} \end{aligned}$$

где ρ — некоторая положительная константа.

Тогда при использовании «персептронной» процедуры построения разделяющей гиперплоскости с начальными весами, равными нулю, число исправлений ошибок не превзойдет число

$$k_0 = \left[\frac{D^2}{\rho^2}\right].$$

Доказательство. Заменим в нашей последовательности все векторы $\mathbf{x}_i$, принадлежащие классу 0, на $-\mathbf{x}_i$. Векторы новой последовательности обозначим $\mathbf{x}_i^*$. Пусть $\boldsymbol{\varphi}(t)$ — вектор весов R-элемента, настроенный после просмотра t членов последовательности. Тогда ошибка опознания для любого вектора $\mathbf{x}_i^*$ произойдет в том случае, если $(\boldsymbol{\varphi}(t), \mathbf{x}_i^*) \leqslant 0$, а алгоритм настройки может быть записан в следующем виде:

$$\boldsymbol{\varphi}(0) = 0.$$

Если очередной вектор последовательности $\mathbf{x}_{t+1}^*$ опознается правильно, т. е. $(\boldsymbol{\varphi}(t), \mathbf{x}_{t+1}^*) > 0$, то изменения не происходит: $\boldsymbol{\varphi}(t+1) = \boldsymbol{\varphi}(t)$. Если же произошла ошибка, т. е. $(\boldsymbol{\varphi}(t), \mathbf{x}_{t+1}^*) \leqslant 0$, то производится исправление

$$\boldsymbol{\varphi}(t+1) = \boldsymbol{\varphi}(t) + \mathbf{x}_{t+1}^*.$$

Оценим модуль вектора после k исправлений. Если в момент t произошло исправление, то

$$|\boldsymbol{\varphi}(t+1)|^2 = |\boldsymbol{\varphi}(t)|^2 + 2(\boldsymbol{\varphi}(t), \mathbf{x}_{t+1}^*) + |\mathbf{x}_{t+1}^*|^2,$$

и, поскольку $(\boldsymbol{\varphi}(t), \mathbf{x}_{t+1}^*) \leqslant 0$, а $|\mathbf{x}_{t+1}^*|^2 \leqslant D^2$, получаем

$$|\boldsymbol{\varphi}(t+1)|^2 \leqslant |\boldsymbol{\varphi}(t)|^2 + D^2.$$

Значит, после k исправлений $|\boldsymbol{\varphi}(t)|^2 \leqslant kD^2$.

Далее, напомним, что по условию теоремы существует такой единичный вектор $\boldsymbol{\varphi}^*$, что для всех векторов последовательности $(\boldsymbol{\varphi}^*, \mathbf{x}_i^*) \geqslant \rho$. Оценим величину $(\boldsymbol{\varphi}^*, \boldsymbol{\varphi}(t))$ после k исправлений. Если в момент $t+1$ происходит исправление, то

$$(\boldsymbol{\varphi}^*, \boldsymbol{\varphi}(t+1)) = (\boldsymbol{\varphi}^*, \boldsymbol{\varphi}(t)) + (\boldsymbol{\varphi}^*, \mathbf{x}_{t+1}^*) \geqslant (\boldsymbol{\varphi}^*, \boldsymbol{\varphi}(t)) + \rho.$$

Если же исправления не было, то $(\boldsymbol{\varphi}^*, \boldsymbol{\varphi}(t+1)) = (\boldsymbol{\varphi}^*, \boldsymbol{\varphi}(t))$.

В силу неравенства Коши $(\boldsymbol{\varphi}^*, \boldsymbol{\varphi}(t)) \leqslant |\boldsymbol{\varphi}^*||\boldsymbol{\varphi}(t)| = |\boldsymbol{\varphi}(t)|$, и поэтому после k исправлений получим

$$|\boldsymbol{\varphi}(t)| \geqslant k\rho.$$

Сопоставляя это неравенство с неравенством $|\boldsymbol{\varphi}(t)|^2 \leqslant kD^2$, получаем

$$k_0 = \left[\frac{D^2}{\rho^2}\right].$$

Следовательно, число исправлений не превосходит $k_0 = [D^2/\rho^2]$, после чего все остальные члены последовательности будут опознаваться правильно. Теорема доказана. ■

Применительно к реальной работе персептрона теорему Новикова можно трактовать двояко. Имея конечную обучающую выборку, можно предъявлять ее машине циклически (потенциально бесконечное число раз). Тогда теорема утверждает, что если векторы двух классов, представленные в выборке, разделимы гиперплоскостью, то уже после конечного числа итераций разделяющая гиперплоскость будет найдена. В этом смысле работа персептрона не отличается от любого другого способа решения соответствующей системы линейных неравенств. Но можно представлять и так, что машина работает в режиме on line, по мере надобности исправляя решающее правило. Тогда теорема утверждает, что при работе в таком режиме, число исправлений не превзойдет $k_0 = [D^2/\rho^2]$. Если же разделяющей гиперплоскости нет, то теорема не утверждает ничего.

Метод потенциальных функций

Метод потенциальных функций, предложенный применительно к задачам распознавания образов в начале 60-х годов М. Айзерманом, Э. Браверманом и Л. Розоноэром, состоит в следующем. Из каждой точки $\mathbf{x}_0$ обучающей выборки испускается потенциал, который в текущей точке $\mathbf{x}$ принимает значение $K(\mathbf{x}, \mathbf{x}_0)$. *Потенциал* — это симметричная функция двух аргументов, в других публикациях и приложениях называемая также *ядром* (в английской терминологии — *kernel*). В принципе, аргументами потенциала могут быть объекты любой природы, но если в пространстве описаний задана метрика, то потенциал обычно стремится к нулю с увеличением расстояния между $\mathbf{x}$ и $\mathbf{x}_0$. Примерами потенциала могут служить функции

$$K(\mathbf{x}, \mathbf{x}_0) = \frac{1}{1 + ar^2} \quad \text{или} \quad K(\mathbf{x}, \mathbf{x}_0) = \exp(-ar^2),$$

где r — расстояние между $\mathbf{x}$ и $\mathbf{x}_0$.

Для каждого класса P строится потенциальная функция

$$Q^P(\mathbf{x}) = \sum_{(P)} \lambda_i^{(P)} K(\mathbf{x}, \mathbf{x}_i),$$

где сумма берется по объектам обучающей выборки, принадлежащим классу P, а $\lambda_i^{(P)}$ — настраиваемые в ходе обучения веса. Опознание нового объекта $\mathbf{x}$ производится по простому принципу — для каждого класса P вычисляется соответствующее значение $Q^P(\mathbf{x})$, и объект относят к тому классу P^*, для которого это значение максимально. В случае, если максимум достигается одновременно на нескольких значениях P^*, выбор из них делается случайно. Алгоритм обучения здесь очень похож на способ обучения персептрона. В начальный момент все веса полагаются равными нулю. Далее процесс идет рекуррентно. Если новый объект обучающей выборки $\mathbf{x}_i$ опознается правильно по уже построенным потенциальным функциям, то веса не меняются. Если же машина относит его к классу P^s, а учитель отнес его к P^t, то вес $\lambda_i^{(P^s)}$ уменьшается на единицу, а вес $\lambda_i^{(P^t)}$ увеличивается на единицу. В терминах самих потенциальных функций, обозначая $Q_l^P(\mathbf{x})$состояние этих функций к l-тому шагу обучения, получим:

$Q_{l+1}^P(\mathbf{x}) = Q_l^P(\mathbf{x})$, если объект $\mathbf{x}_i$ опознан правильно,

$Q_{l+1}^{P^t}(\mathbf{x}) = Q_l^{P^t}(\mathbf{x}) + K(\mathbf{x}, \mathbf{x}_i)$, $Q_{l+1}^{P^s}(x) = Q_l^{P^s}(\mathbf{x}) - K(\mathbf{x}, \mathbf{x}_i)$, если объект отнесен машиной к классу P^s, а учитель отнес его к P^t.

Сходство с персептроном становится более глубоким, если принять дополнительное допущение, что потенциал (ядро) $K(\mathbf{x}, \mathbf{x}_0)$ является положительно полуопределенным. Это значит, что для любого конечного набора объектов $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ симметрическая матрица K с элементами $k_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ должна быть положительно полуопределенной, т. е. ее собственные числа должны быть неотрицательными. Если матрица обладает этим свойством, то и любая ее подматрица также им обладает.

Из функционального анализа (теорема Рисса) известно, что если описания представлены векторами в векторном пространстве, то положительно полуопределенное ядро при некоторых дополнительных предположениях может быть представлено в виде

$$K(\mathbf{x}, \mathbf{y}) = \sum \lambda_i \boldsymbol{\varphi}_i(\mathbf{x}) \boldsymbol{\varphi}_i(\mathbf{y}),$$

где функции $\boldsymbol{\varphi}_i(\mathbf{x})$ образуют линейно независимую систем (зависящую от ядра), а коэффициенты λ_i положительны. Число функций $\boldsymbol{\varphi}_i(\mathbf{x})$ может быть конечным или счетно бесконечным, но сумма должна сходиться при любых $\mathbf{x}$ и $\mathbf{y}$.

Поставим теперь в соответствие каждому объекту $\mathbf{x}$ вектор $\mathbf{z}_x$ с координатами $z_i(\mathbf{x}) = \sqrt{\lambda_i}\varphi_i(\mathbf{x})$. Тогда

$$K(\mathbf{x}, \mathbf{y}) = \sum z_i(\mathbf{x}) z_i(\mathbf{y}) = (\mathbf{z}_x \mathbf{z}_y),$$

то есть значения потенциала просто равны скалярному произведению соответствующих векторов.

Мы, однако, поступим иначе, для того чтобы иметь дело с исходными описаниями любой природы. Это могут быть графы, текстовые строки, сами тексты. Допустим, что число всех объектов конечно (хотя и может быть очень большим). Составим теперь матрицу K с элементами $k_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$, где $\mathbf{x}_i$ и $\mathbf{x}_j$ пробегают все объекты нашего множества. Эта матрица должна быть положительно полуопределенной. Тогда существует унитарная матрица U, приводящая матрицу K к диагональной форме, т. е.

$$UKU^T = D,$$

где U — диагональная матрица с диагональными элементами $d_i \geqslant 0$. Учитывая, что для унитарной матрицы $U^{-1} = U^{\mathrm{T}}$, матрица K представляется как

$$K = U^{\mathrm{T}} D U,$$

или, в координатной форме,

$$k_{ij} = \sum_s d_s u_{si} u_{sj},$$

где u_{si} и u_{sj} — соответственно элементы i-го и j-го столбцов матрицы U. Поставив теперь в соответствие каждому объекту $\mathbf{x}_i$ вектор $\mathbf{z}_i$ с координатами

$$z_s^i = \sqrt{d_s} u_{si}.$$

Тогда

$$K(\mathbf{x}_i, \mathbf{x}_j) = k_{ij} = \sum_s z_{si} z_{sj} = (\mathbf{z}_i \mathbf{z}_j),$$

то есть значение потенциала, испущенного из точки $\mathbf{x}_i$, в точке $\mathbf{x}_j$ будет равно скалярному произведению соответствующих векторов $\mathbf{z}_i$ и $\mathbf{z}_j$.

Поставим теперь в соответствие потенциальной функции $Q^P(\mathbf{x}) = \sum_{(P)} \lambda_i^{(P)} K(\mathbf{x}, \mathbf{x}_i)$ вектор

$$\mathbf{r} = \sum_{(P)} \lambda_i^{(P)} \mathbf{z}_i.$$

Тогда $Q^P(\mathbf{x}) = (\mathbf{r}, \mathbf{z}(\mathbf{x}))$, и соответствующее решающее правило будет линейным в пространстве векторов $\mathbf{z}$. Поэтому такое пространство назвали «спрямляющим пространством».

В силу тождественности работы метода потенциальных функций в спрямляющем пространстве и алгоритма настройки весов реагирующего нейрона у персептрона, к этому методу применима теорема Новикова, но величины D и ρ нужно считать в спрямляющем пространстве. На самом деле, если матрица $K(\mathbf{x}_i, \mathbf{x}_j)$ строго положительно определена, то, за исключением вырожденных случаев, любая пара конечных множеств будет разделима гиперплоскостью в спрямляющем пространстве. Проблема лишь в том, что величина $k_0 = [D^2/\rho^2]$ может оказаться очень большой.

Здесь возникает такой вопрос. Допустим, что нам дана обучающая выборка, состоящая из представителей двух классов 0 и 1, и предъявляется новый объект $\mathbf{x}$. Если мы отнесем его к классу 1, то расширенные таким образом множества будут разделимы гиперплоскостью. Но и в том случае, если мы присоединим его к классу 0, разделимость сохранится. Что же предпочесть? На самом деле в алгоритме персептрона и метода потенциальных функций неявно (и не в полной мере) заложено предпочтение в пользу такого разделения, при котором расстояние между выпуклыми оболочками классов окажется больше. Явно это сделано в методе обобщенного портрета и его наследнике SVM, о которых речь пойдет в дальнейшем.

Лекция 7

Нейронные сети

Нейронные сети привлекли внимание исследователей в конце прошлого века [21, 14, 10]. Эти сети являются усложненным вариантом персептрона, отличающимся двумя важными чертами. Во-первых, в этих сетях не ограничиваются двумя слоями нейронов — преобразующих и

реагирующих, а рассматривают произвольное число k таких слоев. Во-вторых, веса, с которыми суммируются сигналы, приходящие на вход нейрона, настраиваются в ходе обучения не только у нейронов верхнего слоя, а у всех нейронов.

Принцип обучения нейронных сетей, получивший название «обратное распространение» (back propagation) остается тем же, что и у персептрона. Это принцип поощрения и наказания. В *персептроне* в случае правильного опознания веса не меняются. Если произошла ошибка из-за того, что сумма взвешенных сигналов оказалась слишком велика, то веса каналов, по которым пришли импульсы, уменьшаются (наказываются). И наоборот, если же ошибка произошла из-за того, что сумма оказалась мала, то веса таких каналов увеличиваются (поощряются). В *нейронных сетях* поощряются и наказываются не только каналы, но и нейроны предыдущего уровня. Если сумма оказалась слишком велика, то считается, что нейроны предыдущего уровня, подавшие импульс на канал с положительным весом, «зря сработали», и они наказываются. В этом же случае нейрон предыдущего слоя, подавший импульс на канал с отрицательным весом, поощряется. Если сумма оказалась мала, то поощряются нейроны, подавшие импульс на каналы с положительным весом, и наказываются нейроны предыдущего слоя, подавшие импульс на канал с отрицательным весом. Так волна поощрений и наказаний распространяется от верхнего слоя до самого нижнего.

Формально обучение нейронной сети представляет собой некоторый вариант градиентного спуска. Для его реализации необходимо брать производные от целевой функции по параметрам, но сделать это для пороговых элементов затруднительно. Поэтому здесь в качестве модели нейрона используют не пороговый элемент, а устройство, реализующее «сигмоидальную» функцию. Это монотонно возрастающая дифференцируемая функция $F(s)$, изменяющаяся между нулем и единицей, обращающаяся в нуль, когда аргумент идет к минус бесконечности, и обращающаяся в единицу, когда s идет к плюс бесконечности. В качестве такой функции можно взять

$$F(s) = \frac{1}{2}(1 + \mathrm{th}(as)) \quad \text{или} \quad F(s) = \frac{1}{2}\left(1 + \frac{s}{a + |s|}\right).$$

Теперь уже функция будет иметь непрерывную производную $f(s) = dF/ds$. В частности, если $F(s) = (1/2)(1 + \mathrm{th}(as))$, то $f(s) = CF(1 - F)$,

$C = \text{const}$. Модель нейрона описывается как

$$y = F\left(\sum \lambda_i x_i + b\right),$$

где y — сигнал на выходе нейрона, x_i — сигнал, пришедший на i-й вход нейрона, λ_i — веса суммирования этого нейрона, b — порог срабатывания. Функция F задается одинаковой для всех нейронов сети.

Работа всей нейронной сети описывается так. На нейроны нижнего уровня подаются значения внешнего входного сигнала $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Они преобразуются по указанной схеме в выходные сигналы нейронов первого уровня, которые в свою очередь подаются на входы нейронов второго уровня и так до нейронов высшего уровня. Каждый нейрон высшего уровня ответственен распознавание одного из классов входных описаний. Обозначим значения выхода нейронов высшего уровня вектором $\mathbf{y} = (y_1, \ldots, y_m)$.

По замыслу в идеале ровно один нейрон должен среагировать на предъявления объекта класса P, то есть должно быть $y_P = 1$, а все остальные y_i равны нулю. Таким же вектором опишем указание учителя: $\mathbf{y}^* = (y_1^*, \ldots, y_m^*)$, где $y_P^* = 1$, если учитель относит объект к классу P, а все остальные y_i^* равны нулю. Тогда штраф за различие между реакцией нейронной сети и указанием учителя при предъявлении описания x может быть записан как

$$Q(\mathbf{x}, \boldsymbol{\lambda}) = \sum (y_i - y_i^*)^2,$$

где $\boldsymbol{\lambda}$ — вектор весов суммирования всех нейронов сети.

Понятно, что этот критерий может использоваться не только в задачах распознавания, но и во всех задачах, где реакция учителя описывается вектором $\mathbf{y}^*$, а наша цель — заставить нейронную сеть вести себя та же, как это делает учитель.

Шаги спуска по градиенту функции штрафа делаются поочередно по мере предъявления описаний объектов из обучающей выборки. При этом предъявленный объект считается фиксированным, а градиент берется по изменяемым параметрам сети — весам $\boldsymbol{\lambda}$ и порогам b.

Для того, чтобы найти градиент функции $Q(\mathbf{x}, \boldsymbol{\lambda})$ по $\boldsymbol{\lambda}$, начнем с высшего уровня. Обозначим взвешенную сумму сигналов, пришедших на вход i-го нейрона высшего уровня

$$s_i = \sum (\lambda_i^j x_j + b_i),$$

где x_j — выход j-го нейрона предыдущего уровня, λ_i^j — вес, с которым суммируется сигнал от этого нейрона.

Сначала найдем производную g_i функции штрафа $Q(\mathbf{x}, \boldsymbol{\lambda})$ по значению s_i суммы взвешенных сигналов, пришедших на вход i-го нейрона высшего уровня:

$$g_i = \frac{\partial Q(x, \boldsymbol{\lambda})}{\partial s_i} = (y_i - y_i^*)\frac{dF}{ds_i} = (y_i - y_i^*)f(s_i).$$

(Здесь и далее функция $f(s) = dF/ds$ считается фиксированной и заранее заданной для всех нейронов сети).

Теперь перейдем к частным производным функции штрафа по весам суммирования λ_i^j сигналов, приходящих на вход i-го нейрона высшего уровня от j-го нейрона предыдущего уровня, и по самим сигналам ∂x_j, идущим от этих нейронов:

$$\frac{\partial Q(\mathbf{x}, \boldsymbol{\lambda})}{\partial \lambda_i^j} = \frac{\partial Q(\mathbf{x}, \boldsymbol{\lambda})}{\partial s_i}\frac{\partial s_i}{\partial \lambda_i^j} = g_i x_j.$$

Влияние же самих сигналов x_j на штрафную функцию идет через все нейроны высшего уровня, на которые j-й нейрон передает свой сигнал с весом λ_i^j. Поэтому

$$\frac{\partial Q(\mathbf{x}, \boldsymbol{\lambda})}{\partial x_j} = \sum \frac{\partial Q(\mathbf{x}, \boldsymbol{\lambda})}{\partial s_i}\frac{\partial s_i}{\partial x_j} = \sum g_i \lambda_i^j,$$

где сумма берется по всем нейронам высшего уровня, связанным с j-м нейроном предыдущего уровня.

Пусть теперь уже найдены величины $g_k^* = \partial Q(\mathbf{x}, \boldsymbol{\lambda})/\partial s_k$, $\partial Q(\mathbf{x}, \boldsymbol{\lambda})/\partial \lambda_k^i$, $\partial Q(\mathbf{x}, \boldsymbol{\lambda})/\partial y_i$ для всех нейронов уровня m. (Здесь k — номер нейрона уровня m, i — номера нейронов предыдущего уровня, то есть уровня $m-1$, λ_k^i — вес, с которым k-й нейрон уровня m принимает сигнал от i-го нейрона уровня $m-1$, y_i — выход i-го нейрона уровня $m-1$).

Найдем теперь аналогичные величины для уровня $(m-1)$. Обозначим опять y_i выход i-го нейрона $(m-1)$-го уровня, а s_i — взвешенную сумму сигналов, пришедших на вход этого нейрона:

$$s_i = \sum(\lambda_i^j x_j + b_j),$$

где x_j — выход j-го нейрона предыдущего уровня, λ_i^j — вес, с которым суммируется сигнал от этого нейрона. Тогда

$$g_i = \frac{\partial Q(\mathbf{x}, \boldsymbol{\lambda})}{\partial s_i} = \frac{\partial Q(\mathbf{x}, \boldsymbol{\lambda})}{\partial y_i} \frac{dy_i}{ds_i} = \frac{dF}{ds_i} \frac{\partial Q(\mathbf{x}, \boldsymbol{\lambda})}{\partial y_i} = f(s_i) \frac{\partial Q(\mathbf{x}, \boldsymbol{\lambda})}{\partial y_i}.$$

Но величины $\partial Q(\mathbf{x}, \boldsymbol{\lambda})/\partial y_i$, отражающие влияние выхода y_i на штрафную функцию, были уже найдены нами при анализе нейронов уровня m.

Найдем теперь частные производные функции штрафа по весам суммирования λ_i^j сигналов, приходящих на вход i-го нейрона уровня $m-1$ от j-го нейрона предыдущего уровня, и по самим сигналам ∂x_j, идущим от этих нейронов:

$$\frac{\partial Q(\mathbf{x}, \boldsymbol{\lambda})}{\partial \lambda_i^j} = \frac{\partial Q(\mathbf{x}, \boldsymbol{\lambda})}{\partial s_i} \frac{\partial s_i}{\partial \lambda_i^j} = g_i x_j.$$

Сам же сигнал x_j от j-го нейрона уровня $m-2$ влияет на штрафную функцию через все нейроны уровня $m-1$, с на которые j-й нейрон передает свой сигнал с весом λ_i^j. Поэтому

$$\frac{\partial Q(\mathbf{x}, \boldsymbol{\lambda})}{\partial x_j} = \sum \frac{\partial Q(\mathbf{x}, \boldsymbol{\lambda})}{\partial s_i} \frac{\partial s_i}{\partial x_j} = \sum g_i \lambda_i^j,$$

где суммирование идет по всем нейронам уровня $m-1$, с которыми связан j-й нейрон уровня $m-2$.

Таким образом, опускаясь с высшего уровня вниз, мы можем вычислить все составляющие $\partial Q(\mathbf{x}, \boldsymbol{\lambda})/\partial \lambda_i^j$ градиента функции $Q(\mathbf{x}, \boldsymbol{\lambda})$ по $\boldsymbol{\lambda}$. Это и есть обратное распространение (*back propagation*).

Теперь алгоритм обучения нейронной сети (настройки весов $\boldsymbol{\lambda}$) можно записать так. Пусть в момент t веса имеют значение $\boldsymbol{\lambda}(t)$. В момент $t+1$ предъявляется объект с описанием $\mathbf{x}(t+1)$. Тогда

$$\boldsymbol{\lambda}(t+1) = \boldsymbol{\lambda}(t) - h \, \mathbf{grad}\, Q(\mathbf{x}(t+1), \boldsymbol{\lambda}(t)),$$

где h — шаг настройки.

Аналогично настраиваются пороги b_i.

Как и в случае обычного персептрона, здесь возможны два режима: циклическое повторение конечной обучающей выборки, пока не будет выполнено то или иное условие останова, или режим on-line.

Применительно к режиму on-line интересные результаты получены в теории так называемой «стохастической аппроксимации». Показано, что если функция $E_x Q(\mathbf{x}, \boldsymbol{\lambda})$ имеет один экстремум, то при убывающих значениях шага $h(t)$, таких что их сумма расходится, а сумма квадратов сходится, алгоритм приведет асимптотически к точке $\boldsymbol{\lambda}^*$, доставляющей минимум $E_x Q(\mathbf{x}, \boldsymbol{\lambda})$. Правда, неизвестно, насколько быстро.

В циклическом варианте алгоритм нацелен на нахождение минимума функции $(1/l)\sum Q(\mathbf{x}_i, \boldsymbol{\lambda})$, где $\mathbf{x}_i$ — элементы обучающей выборки, и опять-таки в случае отсутствия локальных минимумов приводит к успеху.

Но на самом деле нейронная сеть реализует сложную кусочно-гладкую разделяющую поверхность, и появление локальных экстремумов почти неизбежно, за исключением простых случаев. Для ускорения работы и «проскока» локальных минимумов предложены различные модификации метода.

Но даже если алгоритм минимизирует штраф на обучающей выборке, остаются открытым вопрос, будет ли решающее правило хорошо работать на новых данных. К этому вопросу мы вернемся в дальнейших лекциях.

Лекция 8

Обобщенный портрет. Двойственная задача. Оптимальная разделяющая гиперплоскость. Машина опорных векторов (SVM) — ядра вместо скалярных произведений (скрещение потенциалов с ОП). Прочие отличия. Виды кернелов, параметры.

Обобщенный портрет

Алгоритм «Машина опорных векторов» (по-английски — Support Vector Machine, сокращенно SVM) в настоящее время широко применяется для решения различных задач машинного обучения распознаванию образов. Это алгоритм разработан В.Н. Вапником в 90-е годы прошлого века. Но в нем получили дальнейшее развитие идеи, реализованные в методе «обобщенного портрета» (ОП), разработанного В.Н. Вапником и мною еще в 60-е годы [4].

Основное различие между ОП и SVM состоит в том, что ОП строит оптимальное в определенном смысле линейное решающее правило

в пространстве исходных признаков (или признаков, полученных в результате фиксированного до начала обучения преобразования). Тогда как SVM строит оптимальное в том же смысле линейное решающее правило, но в спрямляющем пространстве с использованием кернельной (ядерной) техники. Первоначально идея обобщенного портрета была предложена В.Н. Вапником для случая, когда в обучающей выборке представлен только один класс, и следует обучиться распознавать объекты этого класса от всех остальных. Допустим, что объекты представлены нормированными к единице векторами в евклидовом пространстве, то есть располагаются на единичной сфере. Мы ищем линейное решающее правило, отделяющее точки нашего класса от всех остальных, то есть должны отрезать от сферы сегмент, рассчитывая, что вектора нашего класса окажутся внутри этого сегмента, а остальные — вне сегмента.

Поскольку число объектов выделенного класса всегда мало по сравнению с числом всех остальных, естественно было искать сегмент минимального объема, содержащий все точки нашего класса, представленные в данных обучения. Это значит, что угол между направляющим вектором разделяющей гиперплоскости $\boldsymbol{\varphi}$ ($|\boldsymbol{\varphi}| = 1$)и крайними (наиболее удаленными от центра сегмента) векторами должен быть минимальным, а соответственно скалярное произведение этих векторов с направляющим вектором должно быть максимальным. Иными словами следует искать такой единичный вектор $\boldsymbol{\varphi}$, что на нем достигается

$$\max_{\boldsymbol{\varphi}} \min_{i} (\boldsymbol{\varphi}, \mathbf{x}_i),$$

где $\mathbf{x}_i$ пробегают все векторы обучающей выборки. (Напомню, что в этом варианте в выборке присутствуют векторы только одного класса). Обозначим $\boldsymbol{\varphi}_0$ — значение вектора, на котором достигается этот максимум, а через величину $c = \min_i(\boldsymbol{\varphi}_0, \mathbf{x}_i)$. Собственно, вектор $\boldsymbol{\varphi}_0$ и был назван обобщенным портретом класса, а векторы, на которых этот минимум достигается, - крайними (или опорными). Тогда соответствующее решающее правило имеет вид:

Объект $\mathbf{x}$ принадлежит нашему классу, если

$$(\boldsymbol{\varphi}_0, \mathbf{x}) \geqslant c,$$

и не принадлежит в противном случае.

Мы надеемся, конечно, что величина c окажется положительной, — только в этом случае срезаемый сегмент сферы будет мал по сравнению с остатком. При положительных значениях c эта задача оказывается эквивалентна следующей. Найти (уже не нормированный) вектор $\boldsymbol{\psi}$, на котором при ограничениях

$$(\boldsymbol{\psi}, \mathbf{x}_i) \geqslant 1 \qquad (i = 1, \ldots, l) \tag{1}$$

достигается минимум квадрата модуля $(\boldsymbol{\psi}, \boldsymbol{\psi})$. А это задача квадратичного программирования.

Значение вектора $\boldsymbol{\psi}_0$, на котором достигается минимум модуля, связано с $\boldsymbol{\varphi}_0$ и c соотношениями

$$\boldsymbol{\varphi}_0 = \frac{\boldsymbol{\psi}_0}{|\boldsymbol{\psi}_0|}, \quad c = \frac{1}{|\boldsymbol{\psi}_0|}.$$

Оказывается, что ненормированный ОП $\boldsymbol{\psi}_0$ всегда может быть разложен по крайним (опорным) векторам с положительными весами:

$$\boldsymbol{\psi}_0 = \sum \alpha_i \mathbf{x}_i \text{ при условиях } \alpha_i \geqslant 0; \quad \alpha_i > 0 \Leftrightarrow (\boldsymbol{\psi}_0, \mathbf{x}_i) = 1. \tag{2}$$

Справедливо и обратное: всякий вектор $\boldsymbol{\psi}$, удовлетворяющий неравенствам (1) и условиям (2), совпадает с $\boldsymbol{\psi}_0$.

Однако оказалось, что задачи, где в обучающей выборке представлен только один класс, редки и не очень интересны (хотя на практике такие задачи встречаются). Поэтому естественно предположить, что в обучающей выборке даны и объекты, не принадлежащие интересующему нас классу. Но по-прежнему допустим, что классы не равноправны, например, мы распознаем изображения буквы A против всех других букв. Тогда естественно потребовать, чтобы (в случае нормированных векторов) разделяющая гиперплоскость, как и раньше, отсекала от сферы сегмент минимального объема, но такой, что все точки обучающей выборки из нашего класса оказались внутри сегмента, а точки противоположного класса — снаружи и с некоторым зазором.

Формально задача сводится к следующей. Пусть в обучающей выборке представлены l_1 векторов $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{l_1})$ класса 1 и l_0 векторов $(\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_{l_0}^*)$ класса 0. Требуется найти минимальный по модулю

вектор $\boldsymbol{\psi}_0$ такой, что

$$\begin{array}{ll} (\boldsymbol{\psi}_0, \mathbf{x}_i) \geqslant 1 & \text{для всех } i = 1, \ldots, l_1, \\ (\boldsymbol{\psi}_0, x_i^*) \leqslant k & \text{для всех } i = 1, \ldots, l_0, \\ \boldsymbol{\psi}_0 = \min\limits_{\boldsymbol{\psi}} |\boldsymbol{\psi}|^2 & \text{при этих ограничениях,} \end{array} \tag{3}$$

где k — некоторая константа, меньшая единицы.

Опять получаем задачу квадратичного программирования. Если система неравенств (3) совместна, то решение будет единственным, так как ищется минимум строго выпуклой функции при выпуклых ограничениях.

Такой вектор $\boldsymbol{\psi}_0$ также будем называть *обобщенным портретом* класса 1 относительно класса 0. Векторы, для которых неравенства (3) переходят в равенства, будут крайними (опорными) в обоих классах. Обобщенный портрет всегда может быть разложен в сумму взвешенных крайних векторов, причем веса при крайних векторах класса 1 будут положительны, а при крайних векторах класса 0 (если они есть) — отрицательными:

$$\begin{array}{l} \boldsymbol{\psi}_0 = \sum\limits_{i=1}^{l_1} \alpha_i \mathbf{x}_i - \sum\limits_{i=1}^{l_2} \alpha_i^* \mathbf{x}_i^* \quad \text{при условиях} \\ \quad \alpha_i \geqslant 0, \quad \alpha_i > 0 \quad \text{только тогда, когда} \quad (\boldsymbol{\psi}_0, \mathbf{x}_i) = 1, \\ \quad \alpha_i^* \geqslant 0, \quad \alpha_i^* > 0 \quad \text{только тогда, когда} \quad (\boldsymbol{\psi}_0, x_i^*) = k. \end{array} \tag{4}$$

Справедливо и обратное: всякий вектор $\boldsymbol{\psi}$, удовлетворяющий неравенствам (3) и условиям (4), совпадает с $\boldsymbol{\psi}_0$.

Отсюда следует, что если из обучающей выборки удалить любой вектор, не оказавшийся крайним (или все такие векторы), то обобщенный портрет не изменится, и, следовательно, удаленные векторы все равно будут опознаваться правильно.

Распознавание нового вектора x можно вести по правилу:

$$\begin{array}{l} \text{если } (\boldsymbol{\psi}_0, \mathbf{x}) > \dfrac{1+k}{2}, \text{ то объект относится к классу 1,} \\ \text{в противном случае он относится к классу 0.} \end{array} \tag{5}$$

Если система крайних векторов окажется линейно независимой, то их число не превосходит размерности пространства признаков n, а разложение ОП по этим векторам будет единственным. Если же это не так

(что возможно), то в разложении ОП по крайним векторам всегда можно исключить, по крайней мере, один из них (вычитая или прибавляя разложение нуля). Повторяя эту операцию, можно добиться того, что в разложении ОП останутся только линейно независимые векторы, число которых не превышает размерности пространства признаков n. Назовем информативными те из крайних векторов, которые невозможно исключить из разложения ОП (в принципе их может и не быть). Число их $m_{\text{инф}}$, понятно, тоже не превосходит n.

Из этих соображений можно получить простую оценку сверху числа ошибок на новых объектах, не представленных в обучении в предположении, что система неравенств (3) заведомо совместна. Действительно, рассмотрим случайную последовательность векторов $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l, \mathbf{x}_{l+1}$, включающую представителей обоих классов. Построим по ней обобщенный портрет $\boldsymbol{\psi}_{l+1}$. Если последний вектор $\mathbf{x}_{l+1}$ не будет информативным, то существует разложение ОП, в которое этот вектор войдет с нулевым весом. Но тогда ОП $\boldsymbol{\psi}_l$, построенный только по векторам $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l$, совпадет с $\boldsymbol{\psi}_{l+1}$, и последний вектор заведомо будет правильно опознан по ОП $\boldsymbol{\psi}_l$, поскольку он правильно опознается по ОП $\boldsymbol{\psi}_{l+1}$.

Математическое ожидание E_{err} числа ошибок на новых данных по ОП $\boldsymbol{\psi}_l$, построенному по случайной выборке $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l$, равно вероятности того, что случайный вектор $\mathbf{x}_{l+1}$ будет опознан неправильно, а это может произойти, только если он не войдет в число информативных векторов обобщенного портрета $\boldsymbol{\psi}_{l+1}$. При заданном составе выборки $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l, \mathbf{x}_{l+1}$, но случайном порядке ее элементов, вероятность того, что именно последний член последовательности окажется информативным, равна $m_{\text{инф}}/(l+1)$. Откуда

$$E_{\text{err}} \leqslant \frac{E m_{\text{инф}}}{l+1} \leqslant \frac{n}{l+1}.$$

Конечно, на самом деле число информативных векторов может оказаться значительно меньше размерности пространства признаков n. Но в отличие от числа n, количество информативных векторов нам не известно до начала обучения, а фактическое их количество, устанавливаемое в ходе обучения, служит лишь несмещенной, но не точной, оценкой $E m_{\text{инф}}$. Тем не менее, величина $m_{\text{инф}}/(l+1)$ может служить грубой оценкой сверху числа ошибок на новых данных.

Двойственная задача

Оказалось, однако, что более удобно решать двойственную задачу квадратичного программирования, чем прямую. Роль множителей Лагранжа здесь играют как раз коэффициенты α_i и α_i^* разложения вектора

$$\boldsymbol{\psi} = \sum_{i=1}^{l_1} \alpha_i \mathbf{x}_i - \sum_{i=1}^{l_2} \alpha_i^* \mathbf{x}_i^* \tag{6}$$

по векторам обучающей выборки. Двойственная функция (лагранжиан) имеет вид

$$W(\alpha, \alpha^*) = \sum_{i=1}^{l_1} \alpha_i - k \sum_{i=1}^{l_2} \alpha_i^* - \frac{1}{2}(\boldsymbol{\psi}, \boldsymbol{\psi}), \tag{7}$$

где $\boldsymbol{\psi}$ задается выражением (6).

Требуется найти максимум функции $W(\alpha, \alpha^*)$ при ограничениях

$$\alpha_i \geqslant 0, \quad \alpha_i^* \geqslant 0.$$

Величина $\boldsymbol{\psi}$, рассматриваемая как функция коэффициентов α_i и α_i^*, представляет собой положительно полуопределенную квадратичную форму.

Если система неравенств (3) совместна, то функция W имеет максимум, и, как легко убедиться, условия достижения этого максимума совпадают с условиями (3) и (4). Если же система (3) несовместна, то функция $W(\alpha, \alpha^*)$ неограниченно возрастает в положительном квадранте коэффициентов α_i и α_i^*.

Заметим, что все вычисления, как при решении двойственной задачи, так и при распознавании новых объектов, могут производиться без использования векторов $\mathbf{x}$ в координатной форме, а только на основе их скалярных произведений. Действительно, решающая функция (5) может быть записана как

$$\begin{aligned}(\boldsymbol{\psi}_0, x) &= \left(\sum_{i=1}^{l_1} \alpha_i \mathbf{x}_i - \sum_{i=1}^{l_2} \alpha_i^* \mathbf{x}_i^*, \mathbf{x} \right) = \\ &= \sum_{i=1}^{l_1} \alpha_i (\mathbf{x}_i, \mathbf{x}) - \sum_{i=1}^{l_2} \alpha_i^* (\mathbf{x}_i^*, \mathbf{x}),\end{aligned}$$

а коэффициенты квадратичной формы $(\boldsymbol{\psi}, \boldsymbol{\psi})$, рассматриваемой как функция α_i и α_i^*, образованы скалярными произведениями векторов обучающей выборки.

Переход к двойственной задаче имеет ряд преимуществ. Во-первых, существенно упрощается вид ограничений. Во-вторых, путем итераций вычисления можно вести в пространстве размерности, равной числу (текущих) крайних векторов, которое обычно значительно меньше размерности исходного пространства признаков. И, наконец, остановимся на еще одном важном обстоятельстве.

До сих пор мы предполагали, что множества разделимы гиперплоскостью, т. е. система неравенств (4) совместна. Но на практике это часто бывает не так. Как быть? Можно, конечно, отказаться решать задачу в данном пространстве признаков и попытаться увеличить их число, например, вводя в качестве новых признаков какие-то функции от исходных. Но во многих случаях ошибки неизбежны, например, в случае, когда признаки искажены шумом, или связь между признаками и классификацией будет стохастической. Тогда естественно согласиться на некоторое число ошибок на обучающей выборке.

Строго обоснованные методы минимизации числа ошибок при построении линейного решающего правила (если они неизбежны) сейчас неизвестны. В качестве эвристического приема предлагается удалять из выборки объекты «наиболее мешающие» разделению. Как уже упоминалось, в случае отсутствия разделения функция $W(\alpha, \alpha^*)$ неограниченно возрастает в положительном квадранте. При поиске ее максимума методом восхождения по градиенту (или методом сопряженных градиентов) определяются те векторы обучающей выборки, у которых соответствующие значения α_i или α_i^* (или их инкременты) первыми достигают заданного критического значения. Эти объекты и удаляются из обучающей выборки. Затем операция повторяется до тех пор, пока оставшаяся часть не сможет быть разделена линейным решающим правилом.

Именно этот алгоритм был реализован в комплексе программ построения обобщенного портрета.

Оптимальная разделяющая гиперплоскость

Рассматривая метод обобщенного портрета, мы считали, что классы, подлежащие распознаванию, неравноправны — один класс узкий, а

другой содержит все прочие объекты. Но существует множество задач, где классы скорее равноправны. Тогда более естественно не минимизировать объем сегмента, содержащего векторы первого класса, а искать разделяющую гиперплоскость, которая наиболее удалена от ближайших векторов обоих классов. Такую плоскость будем называть оптимальной разделяющей гиперплоскостью.

Опять считаем, что в обучающей выборке представлено l_1 векторов $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{l_1})$ класса 1 и l_0 векторов $(\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, x_{l_0}^*)$ класса 0. Пусть $\boldsymbol{\varphi}$ — единичный направляющий вектор гиперплоскости. Обозначим

$$C_1(\boldsymbol{\varphi}) = \min_{i=1,\ldots,l_1} (\boldsymbol{\varphi}, \mathbf{x}_i), \quad C_2(\boldsymbol{\varphi}) = \max_{i=1,\ldots,l_2} (\boldsymbol{\varphi}, \mathbf{x}_i^*).$$

Тогда, если множества векторов $\mathbf{x}_i$ и $\mathbf{x}_i^*$ разделимы гиперплоскостью, то существует такое направление $\boldsymbol{\varphi}$, что $C_1(\boldsymbol{\varphi}) > C_2(\boldsymbol{\varphi})$. Обозначив

$$D(\boldsymbol{\varphi}) = C_1(\boldsymbol{\varphi}) - C_2\boldsymbol{\varphi}),$$

будем искать такое направление $\boldsymbol{\varphi}_0$, на котором достигается максимум функции $D(\boldsymbol{\varphi})$. Уравнение разделяющей гиперплоскости определим как

$$(\boldsymbol{\varphi}, \mathbf{x}) = \frac{1}{2}(C_1(\boldsymbol{\varphi}) + C_2(\boldsymbol{\varphi})).$$

Тогда при $D(\boldsymbol{\varphi}) > 0$ величина $D(\boldsymbol{\varphi})/2$ и будет означать расстояние от ближайших векторов классов до разделяющей гиперплоскости. Эта задача оказывается эквивалентной следующей задаче квадратичного программирования:

Найти (ненормированный) вектор $\boldsymbol{\psi}_0$ и скаляр c, удовлетворяющие неравенствам

$$\begin{array}{ll} (\boldsymbol{\psi}_0, \mathbf{x}_i) \geqslant 1 + c & \text{для всех } i = 1, \ldots, l_1, \\ (\boldsymbol{\psi}_0, \mathbf{x}_i^*) \leqslant -1 + c & \text{для всех } i = 1, \ldots, l_0, \end{array} \tag{8}$$

и доставляющие минимум функции $|\boldsymbol{\psi}|^2 = (\boldsymbol{\psi}, \boldsymbol{\psi})$ при этих ограничениях. Тогда $\boldsymbol{\varphi}_0 = \boldsymbol{\psi}_0/|\boldsymbol{\psi}_0|$, $D(\boldsymbol{\varphi}) = 2/|\boldsymbol{\psi}_0|$, уравнение оптимальной разделяющей гиперплоскости будет

$$(\boldsymbol{\psi}_0, \mathbf{x}) = c,$$

а соответствующее решающее правило примет вид

$$R(\mathbf{x}) = \operatorname{sign}((\boldsymbol{\psi}_0, \mathbf{x}) - c).$$

Система неравенств (8) будет совместной тогда и только тогда, когда векторы, представленные в обучающей выборке, разделимы гиперплоскостью, а ближайшими к оптимальной гиперплоскости будут крайние (опорные) векторы, для которых неравенства (8) переходят в равенства:

$$(\boldsymbol{\psi}_0, \mathbf{x}_i) = 1 + c, \quad (\boldsymbol{\psi}_0, \mathbf{x}_i^*) = -1 + c.$$

Как и в случае ОП, вектор $\boldsymbol{\psi}_0$ может быть разложен по крайним векторам:

$$\boldsymbol{\psi}_0 = \sum_{i=1}^{l_1} \alpha_i \mathbf{x}_i - \sum_{i=1}^{l_2} \alpha_i^* \mathbf{x}_i^*, \tag{9}$$

причем $\alpha_i \geqslant 0$ и $\alpha_i^* \geqslant 0$.

Здесь также удобно перейти к двойственной задаче. Роль множителей Лагранжа здесь играют как раз коэффициенты α_i и α_i^* разложения вектора

$$\boldsymbol{\psi} = \sum_{i=1}^{l_1} \alpha_i \mathbf{x}_i - \sum_{i=1}^{l_2} \alpha_i^* \mathbf{x}_i^* \tag{10}$$

по векторам обучающей выборки.

Двойственная функция (лагранжиан) имеет вид

$$W(\alpha, \alpha^*) = \sum_{i=1}^{l_1} \alpha_i + \sum_{i=1}^{l_2} \alpha_i^* - \frac{1}{2}(\boldsymbol{\psi}, \boldsymbol{\psi}), \tag{11}$$

где $\boldsymbol{\psi}$ задается выражением (9). Требуется найти максимум функции $W(\alpha, \alpha^*)$ при ограничениях

$$\alpha_i \geqslant 0, \quad \alpha_i^* \geqslant 0,$$

и дополнительном ограничении $\sum_{i=1}^{l_1} \alpha_i - \sum_{i=1}^{l_2} \alpha_i^* = 0$, которое соответствует оптимизации по скаляру c.

После вычисления оптимальных значений α_i^0 и α_i^{*0} вектор $\boldsymbol{\psi}_0$ определяется как

$$\boldsymbol{\psi}_0 = \sum_{i=1}^{l_1} \alpha_i^0 \mathbf{x}_i - \sum_{i=1}^{l_2} \alpha_i^{*0} \mathbf{x}_i^*, \tag{12}$$

а порог c — как

$$c = \frac{1}{2}(\min(\mathbf{x}_i, \boldsymbol{\psi}_0) - \max(x_i^*, \boldsymbol{\psi}_0)).$$

Решающее правило имеет вид

$$R(\mathbf{x}) = \text{sign}((\boldsymbol{\psi}_0, \mathbf{x}) - c).$$

Как и в случае построения обобщенного портрета, все вычисления могут производиться без представления векторов в координатной форме, а лишь с использованием скалярных произведений векторов $\mathbf{x}$ между собой.

Если крайние (опорные) векторы не образуют линейно независимую систему, то разложение (12) не будет однозначным. Но из числа крайних векторов всегда можно выделить подсистему информативных векторов, которые в любое разложение вектора по опорным векторам (с неотрицательными коэффициентами) входят с положительным весом. Они заведомо образуют линейно независимую систему, и, потому, их количество не превышает размерность пространства признаков n.

Однако во многих случаях число опорных векторов оказывается значительно меньше размерности пространства признаков. Дело тут вот в чем. При линейном программировании минимум линейной функции в общем случае достигается в точке пересечения n ограничений, где n — размерность пространства. В отличие от этого квадратичная функция может достигать минимума на многообразиях большей размерности, где пересекается меньшее число ограничений. Точно так же, как для ОП, здесь можно получить оценку математического ожидания числа ошибок E_{err} на новых объектах:

$$E_{\text{err}} \leqslant \frac{Em_{\text{инф}}}{l+1} \leqslant \frac{n}{l+1},$$

а вместо математического ожидания $Em_{\text{инф}}$ числа информативных векторов использовать его грубую оценку — фактическое число опорных векторов.

В случае построения оптимальной разделяющей гиперплоскости также возникает вопрос, что делать, если предъявленные в обучении векторы не разделимы гиперплоскостью. И здесь одно из возможных решений состоит в том, чтобы удалять из обучающей выборки векторы, наиболее «мешающие разделению». Определить их можно по тому, какие значения α_i или α_i^* при восхождении по функции $W(\alpha, \alpha^*)$ первыми достигают критически больших значений.

Алгоритм построения оптимальной разделяющей гиперплоскости с возможным удалением векторов «наиболее мешающих разделению»

также был реализован в системе программ обобщенного портрета. Правда, представление векторов в координатной форме там все же использовалось.

Машина опорных векторов (Support Vector Machine)

Машина опорных векторов (SVM) возникла на основе соединения идей метода потенциальных функций и построения оптимальной разделяющей гиперплоскости [29, 30]. Как уже говорилось в одной из предыдущих лекций, если задано положительно полуопределенное ядро $K(z_1, z_2)$, определенное на множестве объектов любой природы, то существует отображение $\mathbf{x}(z)$ объектов в евклидово пространство, такое, что значения ядра для любой пары объектов z_1 и z_2 равно скалярному произведению соответствующих векторов:

$$K(z_1, z_2) = (\mathbf{x}(z_1), \mathbf{x}(z_2)).$$

Евклидово пространство, в которое производится это отображение, называется *спрямляющим пространством*. Напомним, что положительно полуопределенным ядром может быть любая симметрическая функция, такая, что для любой конечной совокупности объектов $z_1, z_2, \ldots, z_l$ матрица K, составленная из элементов $k_{ij} = K(z_i, z_j)$, будет положительно полуопределенной, т. е. все ее собственные числа будут неотрицательными.

Всякому линейному решающему правилу в спрямляющем пространстве вида

$$R(\mathbf{x}) = \operatorname{sign}((\boldsymbol{a}, \mathbf{x}) + b),$$

где b — скаляр, а вектор $\boldsymbol{a}$ представим в виде

$$\boldsymbol{a} = \sum_{i=1}^{l} \lambda_i \mathbf{x}(z_i)$$

соответствует решающее правило в исходном пространстве

$$R(z) = \operatorname{sign}\left(\sum_{i=1}^{l} \lambda_i K(z, z_i) + b\right).$$

Следовательно, оптимальной разделяющей гиперплоскости в спрямляющем пространстве соответствует определенное решающее правило в

исходном пространстве (которое, вообще говоря, будет уже нелинейным). Как уже отмечалось, все вычисления при построении оптимальной разделяющей гиперплоскости могут проводиться без представления векторов в координатной форме, а только с использованием их скалярных произведений, которым соответствуют значения ядра для пар объектов в исходном пространстве. В этом и состоит основная идея SVM — строить решающее правило, соответствующее оптимальной разделяющей гиперплоскости в спрямляющем пространстве, без фактического перехода в это пространство. Такой переход практически невозможен, так как размерность спрямляющего пространства в большинстве случаев оказывается бесконечной или очень большой.

Другое существенное отличие SVM от алгоритмов группы обобщенного портрета состоит в отношении к векторам, «мешающим разделению». В первом случае они просто удалялись по определенному критерию из обучающей выборки. В SVM вводится штрафная функция за ошибочно опознаваемые объекты из обучающей выборки, пропорциональная удаленности соответствующего вектора от разделяющей гиперплоскости (в неправильную сторону). С одной стороны, это позволяет свести задачу к единой задаче квадратичного программирования без дополнительных операций по удалению объектов из выборки. С другой стороны, ошибочно распознаваемые объекты влияют на положение разделяющей гиперплоскости (хотя все равно опознаются неправильно).

Перейдем теперь к формальному описанию машины опорных векторов. Пусть задана обучающая выборка $z_1, z_2, \ldots, z_l$ и соответствующая последовательность реакций учителя $y_1, y_2, \ldots, y_l$, где $y_i = 1$, если объект z_i принадлежит заданному классу, и $y_i = -1$, если объект z_i ему не принадлежит.

Требуется найти в спрямляющем пространстве вектор w, скаляр b и величины ξ_i, удовлетворяющие условиям

$$\begin{aligned} &(w, \mathbf{x}(z_i)) + b \geqslant 1 - \xi_i, \quad y_i = 1, \\ &(w, \mathbf{x}(z_i)) + b \leqslant 1 + \xi_i, \quad y_i = -1, \\ &\xi_i \geqslant 0, \end{aligned}$$

доставляющие при этих условиях минимум функции

$$Q(w, b, \xi) = \frac{1}{2}(w, w) + C \sum_{i=1}^{l} \xi_i.$$

Член $C\sum_{i=1}^{l}\xi_i$ ответственен за возможные ошибки на данных обучения. Величина C является параметром алгоритма и определяет компромисс между желанием уменьшить число ошибок на данных обучения и желанием увеличить зазор между выпуклыми оболочками безошибочно разделяемых векторов.

При переходе к двойственной задаче получаем следующее. Требуется максимизировать лагранжиан

$$W(\boldsymbol{\lambda}) = \sum \lambda_i - \frac{1}{2}(w,w), \qquad w = \sum \lambda_i y_i \mathbf{x}_i,$$

при ограничениях

$$0 \leqslant \lambda_i \leqslant C, \quad \sum \lambda_i y_i = 0.$$

Квадратичная форма (w,w) в исходном пространстве имеет вид

$$(w,w) = \sum_{i,j} \lambda_i \lambda_j y_i y_j K(z_i, z_j).$$

Таким образом, задача может решаться без фактического перехода к спрямляющему пространству. Векторы $\mathbf{x}_i$ обучающей выборки, для которых соответствующие коэффициенты λ_i не равны нулю (положительны), образуют систему опорных векторов. В их число войдут и ошибочно опознаваемые объекты. Оптимальное значение порога определяется как

$$b^0 = \frac{1}{2}[(w^0, \mathbf{x}) + (w^0, \mathbf{x}^*)],$$

где w^0 — оптимальное значение вектора w, $\mathbf{x}$ — любой безошибочный опорный вектор из класса $y = 1$, x^* — любой безошибочный опорный вектор из класса $y = -1$.

Обозначив λ_i^0 оптимальные значения λ_i, получаем результирующее решающее правило:

$$R(z) = \operatorname{sign}\left(\sum_{i=1}^{l} \lambda_i^0 K(z, z_i) + b^0\right).$$

К сожалению, переход к кернельной (ядерной) технике реализован только для случая оптимальной разделяющей гиперплоскости, но не для обобщенного портрета в классическом варианте.

Ядра (Kernels)

Перечислим наиболее широко используемые ядерные функции.

1. Простое скалярное произведение

$$K(\mathbf{x},\mathbf{y}) = (\mathbf{x},\mathbf{y}).$$

   В этом случае алгоритм SVM в общем совпадает с алгоритмом построения оптимальной разделяющей гиперплоскости, за исключением иного обращения с объектами «мешающими разделению».

2. Простое полиномиальное ядро

$$K(\mathbf{x},\mathbf{y}) = (1 + (\mathbf{x},\mathbf{y}))^d \quad \text{или} \quad K(\mathbf{x},\mathbf{y}) = (\mathbf{x},\mathbf{y})^d.$$

   Здесь d — параметр, задаваемый пользователем. Начиная с этого ядра, решающие правила в исходном пространстве будут нелинейными.

3. Дробно-рациональное

$$K(\mathbf{x},\mathbf{y}) = \frac{1 - (\mathbf{x},\mathbf{y})^d}{1 - (\mathbf{x},\mathbf{y})}$$

   Для случая нормированных векторов, когда $-1 \leqslant (\mathbf{x},\mathbf{y}) \leqslant 1$:

4. Гауссово ядро

$$K(\mathbf{x},\mathbf{y}) = \exp(-a|\mathbf{x}-\mathbf{y}|^2).$$

   При больших значениях результат оказывается близким к тому, что дает метод ближайших соседей.

5. Полное полиномиальное ядро

$$K(\mathbf{x},\mathbf{y}) = \left(\frac{(\mathbf{x},\mathbf{y})}{a} + b\right)^d.$$

6. Ядро «АНОВА»

$$K(\mathbf{x},\mathbf{y}) = \sum \left[\exp\left(-a(x_k - y_k)^2\right)\right]^d,$$

   сумма берется по координатам x_k и y_k векторов $\mathbf{x}$ и $\mathbf{y}$.

Существуют еще ядра для построения сплайнов, для сравнения текстовых строк и др. Но рассказ об этом требует отдельной лекции.

Лекция 9

Критика подхода. Примеры, когда он не работает. Проблема равномерной сходимости эмпирического риска к истинному (или частот вероятностям, или средних к математическим ожиданиям). (Примеры задач, когда использование рассмотренных методов не приводит к успеху).

Общий принцип перечисленных методов

Все перечисленные в предыдущих лекциях методы восстановления зависимостей — метод максимального правдоподобия, метод наименьших квадратов, методы поиска оптимального решающего правила — имели в основе своей общий принцип. Это вера в то, что эмпирический риск в силу закона больших чисел с ростом обучающей выборки сойдется к истинному риску, и, поэтому, находя зависимость, для которой эмпирический риск минимален, мы будем стремиться к зависимости, для которой и истинный риск минимален. А это и есть истинная зависимость, или ее наилучшее приближение в заданном классе.

Правда, при конкретном рассмотрении оказывалось, что хорошие оценки близости полученной в результате обучения модели к истинной возникают только тогда, когда выполняется определенное соотношение между числом неизвестных параметров и длинной обучающей выборки. Но в случае метода максимума правдоподобия при этом предполагалось, что функция правдоподобия может быть приближена своим разложением в ряд Тейлора с точностью до трех членов, и квадратичная часть уже сошлась к своему математическому ожиданию. В случае наименьших квадратов — что ковариационная матрица тоже сошлась к ее математическому ожиданию. Для нейронных сетей вообще кроме общего принципа минимизации эмпирического риска ничего не было доказано. Для метода обобщенного портрета и SVM оценки были получены только в случае безошибочного разделения обучающей выборки. Так что вера в корректность замены математического ожидания средним арифметическим во всех случаях оставалась.

Примеры, где он не работает

С другой стороны, есть ряд совершенно очевидных примеров, когда это не так. Даже в простейшем случае восстановления зависимости от одной переменной, когда истинная зависимость является линейной,

а наблюдения искажены шумом, всегда можно подобрать полином достаточно высокой степени, который точно пройдет через все точки обучающей выборки. При этом остаточная невязка (эмпирический риск) будет равна нулю, а полученная зависимость далека от истинной. И это будет так при любой длине выборки. Вместо полинома можно воспользоваться кусочно-линейным приближением с тем же эффектом. К тому же результату мы придем, если будем пользоваться методом максимального правдоподобия.

В задачах распознавания образов можно просто запомнить все точки обучающей выборки и их классификацию, данную учителем. А все иные объекты отнести к одному из классов. Материал обучения будет распознаваться стопроцентно правильно, но на новых объектах, как правило, будет много ошибок.

Кстати, на одном из первых конгрессов ИФАКа на секции методов машинного обучения выступал металлург. Он именно так предлагал обучать машину ведению плавки. Для каждой ситуации (состав сырья, требуемая марка стали) предлагалось просто запоминать действия человека-мастера. На вопрос, а где же здесь экстраполяция на новые ситуации, докладчик отвечал, что все это теоретические штучки, а на практике нужно действовать просто.

Это наиболее явные и очевидные примеры, когда принцип замены истинного риска эмпирическим не работает. Но и реально применяемые алгоритмы часто страдают тем же недостатком, например, алгоритм SVM или потенциальных функций при неудачном выборе ядра (или его параметров). Нейронные сети при избыточном числе нейронов позволяют получить прекрасный результат на данных обучения, но очень плохо работают на новых данных, т. е. не обладают способностью к правильной экстраполяции. По-русски это называется подгонка, по-английски overfitting. На этом, кстати, основаны все предрассудки.

В чем же дело?

В чем же дело? Мы ведь доказали закон больших чисел в форме неравенства Чебышева. Во многих упомянутых случаях можно считать, что дисперсия случайной величины — риска — известна (в задачах распознавания, в задаче восстановления регрессии при известной дисперсии шума), и тем не менее значение эмпирического риска в точке минимума оказывается далеким от значения истинного риска даже

при неограниченном увеличении длины обучающей выборки. В случае ограниченной случайной величины известны даже более сильные, чем неравенство Чебышева, оценки. Показано, что вероятность уклонения по модулю среднего арифметического от математического ожидания более чем на любую заданную положительную величину стремится к нулю экспоненциально с ростом выборки. И это, безусловно, верно, если зафиксировать решающее правило (зависимость, модель) до получения выборки, и для него сравнивать близость эмпирического риска к истинному. Но как только решающее правило разрешается выбирать на основании обучающей последовательности (то есть после ее получения), мы сталкиваемся с упомянутым явлением подгонки.

В 60-е годы прошлого века появлялись даже публикации в научных журналах, где замену истинного риска эмпирическим пытались обосновать простой ссылкой на закон больших чисел. Но тогда «обосновывались» бы и упомянутые методы с простым запоминанием данных обучения. Так, все-таки, в чем же дело?

Лотерея

Для того, чтобы понять природу этого явления, рассмотрим игру в лотерею. Допустим, что вероятность выигрыша отельного игрока составляет малую величину p (скажем, $p = 10^{-6}$). Но в игре участвует большое число игроков N (скажем, $N = 10^7$), и вероятности их выигрыша независимы. Шанс выигрыша одного конкретного лица пренебрежимо мал. Но какова вероятность P, что ни один игрок не выиграет? Вероятность того, что один игрок не выиграет, равна $(1 - p)$, а вероятность того, что это событие не произойдет в серии из N независимых испытаний, равна $(1 - p)^N$. То есть $P = (1 - p)^N$.

При малых значениях p следующие преобразования будут достаточно точными:

$$P = (1 - p)^N = \exp(\ln((1 - p)^N)) = \exp(N \ln(1 - p)) \approx \exp(-Np).$$

Подставляя сюда наши значения ($p = 10^{-6}$ и $N = 10^7$), получим вероятность того, что ни один игрок не выиграет равна

$$P \approx \exp(-Np) = \exp(-10),$$

то есть получаем опять пренебрежимо малую величину. Таким образом, несмотря на то, что шанс выиграть у отдельного лица очень мал,

вероятность того, что ни один из них не выиграет в нашем случае, тоже очень мала. Заметим, что если наоборот, $p = 10^{-7}$, а $N = 10^6$, то вероятность

$$P \approx \exp(-Np) = \exp(-0.1)$$

близка к единице.

Таким образом, результат существенно зависит от соотношения малой вероятности p выигрыша одного игрока и большого числа игроков N.

Сходная ситуация возникает и в алгоритмах обучения, основанных на минимизации эмпирического риска. Здесь мы, конечно, стремимся выбрать правило (зависимость, модель), для которого математическое ожидание риска минимально, но одновременно отбираем правило, для которого случайное отклонение от математического ожидания оказалось большим (в нужную сторону). И, поскольку выбор идет из большого числа вариантов, может случиться так, что отберется совсем не то правило, для которого истинный риск минимален (или близок к этому), а то, для которого случайное отклонение оказалось большим. И, как в лотерее, результат существенно зависит от соотношения длины выборки (с увеличением длины выборки вероятность большого случайного отклонения для конкретного правила падает) и объема класса, из которого этот выбор производится.

Конечное число решающих правил

Попробуем применить ту же схему, что для лотереи, чтобы оценить вероятность правильного выбора в задаче распознавания, когда число решающих правил, из которого мы выбираем, конечно и равно N.

Рассмотрим такой простой вариант: среди заданных N решающих правил есть такое, которое безошибочно разделяет всю генеральную совокупность объектов (и заведомо безошибочно разделяет любую обучающую выборку). Правило выбора таково: отсеиваются все решающие правила, допустившие ошибку хотя бы на одном элементе выборки. Из оставшихся выбирается одно, неважно по какому принципу. Мы будем считать результат *правильным*, если вероятность ошибки для выбранного правила не превосходит заданную малую положительную величину ε. Тогда наш выбор заведомо будет правильным, если в ходе обучения отсеются все решающие правила, у которых вероят-

ность ошибки больше ε. Такие правила будем называть «нехорошими», а число их не больше N.

Следовательно, вероятность получения правильного результата $P > P_0$, где P_0 —вероятность того, что отсеются все правила с вероятностью ошибки, большей ε.

Вероятность того, что решающее правило не сделает ни одной ошибки на случайной выборке длины l, равна

$$p_n = (1-p)^l,$$

где p — вероятность ошибки. Для «нехороших» правил получим $pn < (1-\varepsilon)^l$. Тогда вероятность P_n того, что хотя бы одно из нехороших правил не сделает ни одной ошибки, оценивается как

$$P_n < N(1-\varepsilon)^l.$$

Соответственно,

$$P_0 = 1 - P_n > 1 - N(1-\varepsilon)^l,$$

и вероятность правильного результата

$$P > P_0 > 1 - N(1-\varepsilon)^l. \tag{1}$$

Для того, чтобы наша оценка (снизу) вероятности правильного результата стала близка к единице, необходимо, чтобы величина $N(1-\varepsilon)^l$ стала близка к нулю. А это произойдет, если длина выборки станет существенно больше, чем

$$l_0 = -\frac{\ln N}{\ln(1-\varepsilon)} \approx \frac{\ln N}{\varepsilon}. \tag{2}$$

Так мы получили оценку снизу для вероятности правильного результата. Если бы решающие правила «ошибались» независимо друг от друга, то, действуя аналогично случаю лотереи, можно получить достаточно точную оценку этой вероятности, которая будет очень близка к оценке (1). Но на самом деле решающие правила «ошибаются» не независимо, и оценка (1) во многих случаях оказывается существенно заниженной (особенно, если число решающих правил бесконечно).

Посмотрим, однако, что можно получить из оценки (2) в конкретных случаях.

Конечное число объектов и все возможные разделения на два класса

Допустим, что полное число объектов M, подлежащих распознаванию, конечно, и наш класс решающих правил состоит из всех возможных разбиений этого множества на два класса. Тогда число $N = 2^M$. Подставляя это число в формулу (2), получим, что $l_0 \approx (\ln N)/\varepsilon = M/\varepsilon$, то есть нужно показать все объекты и даже больше. Последнее связано с тем, что распределение вероятностей на множестве объектов может быть не равномерным, и одни объекты в случайной выборке могут появиться уже много раз, тогда как другие еще не встретились ни разу. Но это оценка сверху (для длины обучающей выборки). Эту задачу можно рассмотреть как игровую, когда один игрок выбирает распределение вероятностей на множестве объектов и одно из решающих правил, а другой на основании обучающей выборки находит его или близкое к нему (в смысле допустимого числа ошибок). Седловая точка в этой игре дает оценку, очень близкую к (2).

В случае описания объектов n бинарными признаками число различных описаний $M = 2^n$, то есть комбинаторно большое число, и, если наше множество решающих правил допускает любую классификацию, то нет никаких гарантий, что мы получим правильный результат, пока не покажем их (почти) все и даже больше.

Конечное число объектов и линейные решающие правила

Пусть конечное число объектов m задано своими описаниями в виде точек в n-мерном евклидовом пространстве. Ищется линейное решающее правило в этом пространстве вида

$$R(\mathbf{x}) = \operatorname{sign}\left(\sum \lambda_i x_i\right),$$

где $(x_1, \ldots, x_n)$ - координаты вектора $\mathbf{x}$, а λ_i — искомые веса решающего правила. Каждому такому решающему правилу соответствует разделяющая гиперплоскость $\sum \lambda_i x_i = 0$, проходящая через 0.

Количество решающих правил такого вида бесконечно (континуум), но они могут быть разбиты на группы, внутри каждой из которых классификация точек будет одинаковой. Число таких групп бу-

дет конечно. Оценим это число N. При непрерывном изменении коэффициентов смена классификации будет происходить только тогда, одна из точек $\mathbf{x}_j$ сменит свое положение относительно разделяющей гиперплоскости, а это произойдет в момент, когда выполняется равенство $(\lambda, \mathbf{x}_j) = 0$. Таким образом, число различных классификаций N не больше, чем число компонент связности N_1, на которые пространство вектора коэффициентов разбивается гиперплоскостями вида $(\lambda, \mathbf{x}_j) = 0$. Оценим это число как функцию $\Phi(n, m)$ размерности пространства n и числа объектов m.

В одномерном случае существует единственная разделяющая гиперплоскость вида $(\lambda, x_j) = 0$ — это нуль. Поэтому $\Phi(1, m) = 2$.

Одна гиперплоскость делит пространство любой размерности на две части. Поэтому $\Phi(n, 1) = 2$.

Пусть теперь известно, что $(m-1)$ гиперплоскость делит n-мерное пространство не более чем на $\Phi(n, m-1)$ компонент. Добавим новую гиперплоскость Γ_m. Если эта гиперплоскость проходит через одну из «старых» компонент, то она дробит ее на две. В противном случае старая компонента сохраняется. Таким образом, добавление новой гиперплоскости увеличивает число компонент на столько, сколько «старых» компонент разделилось на две. В свою очередь, каждая такая компонента K_i оставляет на Γ_m след $\Gamma_m \cap K_i$. Число таких следов в точности равно числу компонент, на которые «старые» гиперплоскости делят новую гиперплоскость Γ_m. Поскольку размерность разделяющей гиперплоскости равна $n-1$, то число следов не превосходит $\Phi(n-1, m-1)$.

Таким образом, получаем следующее рекуррентное уравнение:

$$\begin{aligned}
&\Phi(n, m) = \Phi(n, m-1) + \Phi(n-1, m-1),\\
&\Phi(1, m) = 2,\\
&\Phi(n, 1) = 2.
\end{aligned}$$

Решая это уравнение, получим

$$\Phi(n, m) = \begin{cases} 2\sum\limits_{i=0}^{n-1} C_{m-1}^i, & n \leqslant m, \\ 2^m, & n > m. \end{cases}$$

Эту оценку можно заменить более грубой:

$$\Phi(n, m) \leqslant m^n,$$

откуда число различимых между собой решающих правил $N \leqslant m^n$. Подставляя это значение в (2), получим оценку

$$l_0 \approx \frac{\ln N}{\varepsilon} < \frac{n \ln m}{\varepsilon}. \tag{3}$$

Таким образом, оценка достаточной длины обучающей выборки растет линейно с размерностью пространства n и логарифмически с ростом числа объектов m.

В случае описания объектов бинарными признаками число (потенциально) различных объектов $m = 2^n$. Тогда

$$l_0 \approx \frac{\ln N}{\varepsilon} < \frac{n \ln m}{\varepsilon} < \frac{n^2}{\varepsilon}.$$

Оценка не очень хорошая, но все же не 2^n.

Мы уже получали лучшую оценку для метода обобщенного портрета или SVM, но данная оценка получена в более общих условиях.

Рассмотрим еще случай, когда признаки в описании разбиты не на две, а на k градаций. Тогда число различимых описаний тоже будет конечным и равно $m = k^n$. Подставляя это число в (3), получим

$$l_0 \approx \frac{\ln N}{\varepsilon} < \frac{n \ln m}{\varepsilon} < \frac{n^2 \ln k}{\varepsilon}.$$

Отсюда видно, что наша оценка хотя и медленно, но неограниченно растет с числом градаций, на которые разбиты признаки.

Оценку числа всех способов разбиения конечного множества объектов на два класса можно подсчитать и для случая кусочно-линейных разделяющих поверхностей с заданным числом кусков. Например, если разделяющая поверхность состоит из двух линейных кусков, то общее число способов разбиения N не превосходит $4N_1^2$, где N_1 — число способов разбиения одной гиперплоскостью.

Н. Нильсон вообще предлагал оценивать мощность обучающейся машины, функцией, выражающей число способов разбиения в зависимости от числа различимых объектов.

Когда были получены эти оценки для достаточной длины обучающей выборки, то обстоятельство, что оценка уходит в бесконечность с ростом числа градаций, послужило основанием для критики вообще методов минимизации эмпирического риска со стороны сторонников

рекуррентных алгоритмов. У вас, мол, все должно быть дискретно, а мы легко справляемся и со случаем непрерывных параметров описания.

Однако оказалось, что средствами, очень близкими к изложенным выше, можно разобраться и со случаем, когда и число объектов, и число решающих правил потенциально бесконечны. Но об этом я расскажу в следующей лекции.

Любопытная деталь. Когда начали заниматься задачами обучения распознаванию, то имели дело с дискретными объектами, множество, которых было конечным, хотя и очень большим (порядка $m = k^n$). Обучиться значило придти к правильному результату после показа значительно меньшего числа объектов (даже n^2 — это слишком много). При переходе к непрерывным описаниям стали считать, что обучиться — это значит придти к правильному результату после показа конечного числа объектов (неважно какого). Если, однако, разобраться с доказательством многих теорем (в том числе и появляющихся ныне), то оказывается, что проблема решается огрублением (в том или ином смысле) описаний, и сведению их к конечному числу. Но это конечное число оказывается опять порядка k^n. То есть, в старом понимании никакого обучения нет.

Посмотрим, что же мы доказали, и что мы хотим доказать.

Мы имеем с одной стороны истинный (средний) риск, который мы хотим минимизировать:

$$R_{\text{true}}(\alpha) = \int Q(\mathbf{y}, F(\mathbf{x}, \alpha))\, dP_{xy},$$

где $Q(\mathbf{y}, \mathbf{y}^*)$ — функция штрафа за предсказание $\mathbf{y}^*$ при истинном значении $\mathbf{y}$, $F(\mathbf{x}, \alpha)$ — искомое решающее правило (зависимость, модель), α — абстрактный параметр, определяющий этот выбор, а с другой стороны эмпирический риск на обучающей выборке:

$$R_{\text{emp}}(\alpha) = \frac{1}{l} \sum Q(\mathbf{y}_i, F(\mathbf{x}_i, \alpha)).$$

Мы увидели, что максимальное по множеству значений параметра α уклонение эмпирического риска от среднего может оставаться большим, несмотря на то, что в силу закона больших чисел при любом фиксированном значении параметра это уклонение стремится по

вероятности к нулю и достаточно быстро. Это, в свою очередь, вызывает (или может вызвать) тот эффект, что при значении параметра, доставляющем минимум эмпирическому риску, истинный риск будет большим, даже при неограниченном увеличении выборки.

Этого бы заведомо не произошло, если бы максимальное по множеству всех значений параметра уклонение эмпирического риска от истинного тоже стремилось к нулю.

Мы доказали (по крайней мере, для задач распознавания, в которых существует безошибочное решающее правило), что если параметр принимает лишь конечное число значений, то вероятность того, что это уклонение превзойдет любую заданную положительную величину, действительно стремится к нулю, и оценили, насколько быстро.

Теперь нам предстоит выяснить в общем случае, при каких условиях вероятность

$$P\left(\sup_{\alpha}|R_{\text{true}}(\alpha)-R_{\text{emp}}(\alpha)|>\varepsilon\right),\qquad \varepsilon>0,$$

стремится к нулю, и, если так, то с какой скоростью.

Этой проблеме соответствует вопрос в чистой статистике. При каких условиях имеет место равномерная сходимость (по вероятности) средних к математическим ожиданиям? Пусть случайная величина $F(\mathbf{x},\alpha)$ зависит от случайного события $\mathbf{x}$ и абстрактного параметра α. Ее математическое ожидание равно

$$M(\alpha)=\mathrm{E}F(\mathbf{x},\alpha)=\int F(\mathbf{x},\alpha)\,dP_x,$$

а среднее арифметическое по независимой выборке с неизменным распределением равно

$$S(\alpha)=\frac{1}{l}\sum F(\mathbf{x}_i,\alpha).$$

При каких условиях имеет место равномерная сходимость:

$$P\left(\sup_{\alpha}|M(\alpha)-S(\alpha)|>\varepsilon\right)\to 0,\qquad \varepsilon>0,$$

и если да, то с какой скоростью.

В случае, когда $F(\mathbf{x}, \alpha)$ являются индикаторными функциями некоторой системы событий Λ, та же проблема может быть сформулирована, как обобщенная теорема Бернулли: при каких условиях

$$P\left(\sup_{A\in\Lambda}|P(A)-\nu(A)|>\varepsilon\right)\to 0,\quad \varepsilon>0,\quad \text{с ростом длины выборки,}$$

где $P(A)$ — вероятность события A, $\nu(A)$ — частота выпадения события A на выборке длины l, а супремум берется по всем событиям системы Λ.

Эта проблема рассматривалась (получила положительное решение — теорема Гливенко) еще в 30-е годы прошлого столетия применительно к очень частному случаю для системы событий вида

$$A : (\xi \leqslant a),$$

где ξ — случайная величина, а a — скалярный параметр. В этом случае $P(A)$ — это кумулятивная функция распределения, а $\nu(A)$ — ее эмпирическое приближение. Получены были и очень точные оценки вероятности максимального уклонения эмпирического приближения от истинной функции распределения (критерий Колмогорова–Смирнова). Но в наше время интересен и случай, когда события системы зависят от большого (или даже бесконечного) набора параметров. И в этом случае ответ может быть как положительным, так и отрицательным. Проблема получила название обобщенная проблема Гливенко–Кантелли.

Лекция 10

Критерии равномерной сходимости частот к вероятностям. Функция роста. VC-размерность.

На прошлой лекции я обещал вам рассказать условия равномерной сходимости средних к математическим ожиданиям или частот к вероятностям (равномерной по классу событий). Но прежде чем перейти к этой общей проблеме, я хочу вывести условия успешной работы одного класса алгоритмов обучения распознаванию образов. Техника, используемая при решении этой задачи, оказывается применима и для решения общей проблемы равномерной сходимости [4, 5].

Алгоритмы с полной памятью

Собственно, этот класс включает многие алгоритмы из тех, что были уже рассказаны. Допустим, что задан некоторый класс решающих правил, среди которого заведомо есть безошибочное решающее правило, то есть любую обучающую выборку оно делит без ошибок. Алгоритмами с полной памятью называют такие алгоритмы обучения, которые в этих условиях выбирают правило из этого класса, не делающее ошибок на обучающей выборке. Разумеется, выбранное правило может и не совпадать с тем, которое без ошибок распознает всю генеральную совокупность, ведь оно не делает ошибок только на материале обучения. К этому классу относятся алгоритмы персептрона и потенциальных функций в режиме циклического повторения обучающей выборки до полного разделения, а также алгоритмы обобщенного портрета, оптимальной разделяющей гиперплоскости и SVM, но, разумеется, в предположении, что соответствующая разделяющая поверхность, полностью отделяющая точки разных классов в обучающей выборке, заведомо существует. На предыдущей лекции мы уже имели дело с этим видом алгоритмов обучения, но применительно лишь к случаю конечного класса решающих правил, из которого осуществляется выбор. Применительно к потенциально бесконечным классам те рассуждения уже не имеют силы. Обратим внимание, однако, на то, что на практике мы проверяем качество выбранного решающего правила не на всей генеральной совокупности объектов, а на конечной, хотя и достаточно длинной, экзаменационной выборке. При этом не возникает никаких проблем, связанных с подгонкой (overfitting), поскольку экзаменационная выборка применяется к фиксированному решающему правилу уже после того, как оно выбрано.

Если вероятность ошибки для выбранного решающего правила составляет p, то распределение числа ошибок на экзаменационной выборке будет биномиальным распределением с параметром p, и частота ошибок на экзамене достаточно быстро сходится (по вероятности) к вероятности p. Мы не будем рассматривать сейчас этот вопрос в деталях, а поставим вопрос иначе.

Допустим, что длина обучающей и экзаменационной выборки одинакова и равна l. Какова вероятность того, что решающее правило, выбранное алгоритмом с полной памятью по обучающей выборке и безошибочно ее разделяющее, на экзамене покажет частоту ошибок,

превышающую заданное положительное число ε? Во всяком случае, попытаемся оценить эту вероятность сверху.

Заметим, что два следующих варианта одинаковы с точки зрения оценки этой вероятности.

Вариант первый. Случайным образом получаем независимую выборку длины l, и на ней строим решающее безошибочное правило. Затем таким же образом получаем экзаменационную выборку длины l, и на ней проводим экзамен (подсчитываем частоту ошибок).

Вариант второй. Случайным образом получаем полную выборку длины $2l$. Случайным образом делим ее на две полувыборки длины l, и после этого одну полувыборку используем для обучения, а вторую — для экзамена. Все варианты такого разбиения считаем равновероятными.

Во втором случае, после того как полная выборка длины $2l$ получена (но до того как она разделена на две полувыборки), мы уже имеем дело с конечным числом объектов. Раз так, то и число групп решающих правил, одинаково классифицирующих все объекты полной выборки, будет конечным. Назовем, вообще, число решающих правил класса S, по-разному классифицирующих объекты заданной выборки $X_l = (x_1, \ldots, x_l)$, *индексом* $\Delta^S(x_1, \ldots, x_l)$ *системы* S *относительно выборки* X_l. Дальше мы можем работать, опираясь на то, что число различных решающих правил конечно и равно $\Delta^S(x_1, \ldots, x_{2l})$.

Оценим вероятность того, что при заданной полной выборке и ее случайном разбиении на обучающую и экзаменационную, алгоритм с полной памятью выберет такое решающее правило, которое на экзамене дает частоту ошибок, превышающую ε . Такой исход обучения будем называть *неправильным*, в противном случае будем считать его *правильным*. Назовем, как на прошлой лекции, «нехорошими» такие решающие правила, которые на полной выборки дают число ошибок, превышающее εl. Тогда исход обучения будет заведомо правильным, если в ходе обучения отсеются все «нехорошие» решающие правила.

Рассмотрим одно из них, и оценим вероятность p_H того, что оно не сделает ни одной ошибки на обучающей выборке. Пусть число ошибок этого решающего правила на полной выборке составляет $l_1 > \varepsilon l$. Тогда вероятность того, что оно не ошибется на первом элементе обучающей выборки будет $(2l - l_1)/2l = 1 - l_1/2l$, на втором — $(2l - 1 - l_1)/(2l - 1) = 1 - l_1/(2l - 1)$ и т.д. Вероятность же того, что не будет ни одной ошибки,

составит

$$p_H = \prod_{k=0}^{l-1} \left(1 - \frac{l_1}{2l-k}\right) \leqslant \left(1 - \frac{l_1}{2l}\right)^l < \left(1 - \frac{\varepsilon}{2}\right)^l.$$

А вероятность того, то хотя бы одно из «нехороших» решающих правил не сделает ни одной ошибки на обучающей выборке, оценивается как

$$P_H(\varepsilon) < \Delta^S(x_1, \ldots, x_{2l}) \left(1 - \frac{\varepsilon}{2}\right)^l \tag{1}$$

Это и есть оценка сверху вероятности того, что алгоритм с полной памятью даст «неправильный» результат при заданной полной выборке $(x_1, \ldots, x_{2l})$.

Мы видим, что второй сомножитель в этом выражении экспоненциально стремится к нулю с ростом длины выборки. Первый же сомножитель растет, и в зависимости от того, насколько быстро он растет, наша оценка будет или не будет стремиться к нулю.

Функция роста

Величина $\Delta^S(x_1, \ldots, x_{2l})$ является случайной и зависит от конкретной выборки. Поэтому, чтобы получить надежную оценку сверху, введем функцию

$$M^S(l) = \max \Delta^S(x_1, \ldots, x_l),$$

где максимум берется по всем последовательностям $(x_1, \ldots, x_l)$ длины l. Эта функция получила название *функция роста*. Она характеризует максимальное по всем возможным выборкам заданной длины число различных разбиений ее с помощью решающих правил системы S. Функция роста не зависит ни от конкретной выборки, ни даже от распределения вероятностей на множестве объектов. Подставляя функцию роста в (1) качестве оценки сверху для индекса $\Delta^S(x_1, \ldots, x_l)$, получим

$$P_H(\varepsilon) < M^S(2l) \left(1 - \frac{\varepsilon}{2}\right)^l. \tag{2}$$

Теперь ясно, что если функция роста растет полиномиально (или, по крайней мере, медленнее любой экспоненты), то вероятность $P_H(\varepsilon)$ стремится к нулю при любом положительном значении ε. Мы фактически уже вывели на прошлой лекции функцию роста для класса

линейных решающих правил в пространстве размерности n, и получили

$$M^S(l) = \Phi(n, l) \leqslant l^n,$$

то есть в случае линейных решающих правил функция роста мажорируется степенной функцией с показателем, равным размерности n. Значит, в этом случае вероятность получения «неправильного» результата всегда стремится к нулю, но тем медленнее, чем выше размерность пространства.

Напомню, как я говорил на прошлой лекции, Н. Нильсон предлагал использовать функцию, аналогичную функции роста, как показатель мощности распознающей машины. С этой точки зрения, чем быстрее растет эта функция с числом объектов, тем лучше — тем больше возможности машины. Но из наших оценок мы видим, что есть и обратная сторона медали: чем больше мощность машины, тем больше нужно данных, чтобы ее обучить.

VC-размерность

При дальнейшем исследовании оказалось, что функция роста обладает следующим замечательным свойством. Функция роста $M^S(l)$ либо тождественно равна 2^l, либо, если это не так, мажорируется функцией $\sum_{i=0}^{n-1} C_l^i$, где n — минимальное число l, при котором

$$M^S(l) \neq 2^l.$$

Иначе говоря,

$$M^S(l) \equiv 2^l \quad \text{либо} \quad M^S(l) \leqslant \sum_{i=0}^{n-1} C_l^i.$$

Для доказательства этого утверждения нам понадобится следующая лемма.

Лемма. Если для некоторой последовательности $x_1, \ldots, x_l$ и некоторого $n \geqslant 1$

$$\Delta^S(x_1, \ldots, x_l) > \sum_{i=0}^{n-1} C_l^i,$$

то существует подпоследовательность X_n длины n, такая, что

$$\Delta^S(X_n) = 2^n,$$

то есть подпоследовательность X_n может быть произвольно разбита на два класса решающими правилами из системы S.

Доказательство. Обозначим

$$\Phi(n,l) = \sum_{i=0}^{n-1} C_l^i.$$

(Здесь и дальше считаем, что $C_l^i = 0$ при $i > l$). Для этой функции, как легко убедиться, выполняются соотношения

$$\begin{aligned} &\Phi(1,l) = 1, \\ &\Phi(n,l) = 2^l, && l \leqslant n-1, \\ &\Phi(n,l) = \Phi(n,l-1) + \Phi(n-1,l-1), && n > 1, \quad l > 1. \end{aligned} \tag{3}$$

Эти соотношения, в свою очередь, однозначно определяют функцию $\Phi(n,l)$ при $n > 0$, $l > 0$.

Будем доказывать лемму индукцией по n и l. Для $n = 1$ и любого $l > 0$ утверждение леммы очевидно. Действительно, из $\Delta^S(x_1,\ldots,x_l) > 1$ следует, что существует по крайней мере один элемент x_i этой последовательности, который по-разному классифицируется правилами из системы S, и значит $\Delta^S(x_i) = 2$.

Для $l < n$ утверждение леммы тривиально верно ввиду ложности посылки. В этом случае посылка есть

$$\Delta^S(x_1,\ldots,x_l) > 2^l,$$

что невозможно, так как выборку длины l невозможно разбить более чем 2^l способами.

Наконец, допустим, что лемма верна для всех $n \leqslant n_0$ ($n_0 \geqslant 1$) при всех $l > 0$. Зафиксируем $n = n_0 + 1$ и проведем индукцию по l. Для $l < n_0 + 1$, как указывалось, лемма верна. Предположим, что она верна при $l \leqslant l_0$ и покажем, что она верна для $l = l_0 + 1$.

Пусть для некоторой последовательности $x_1,\ldots,x_{l_0},x_{l_0+1}$ справедливо условие леммы:

$$\Delta^S(x_1,\ldots,x_{l_0},x_{l_0+1}) > \Phi(n_0+1,l_0+1).$$

Рассмотрим укороченную на один элемент последовательность $x_1, \ldots, x_{l_0}$. Для нее возможны два варианта:

$$\text{а)}\ \Delta^S(x_1, \ldots, x_{l_0}) > \Phi(n_0 + 1, l_0), \tag{4}$$

$$\text{б)}\ \Delta^S(x_1, \ldots, x_{l_0}) \leqslant \Phi(n_0 + 1, l_0). \tag{5}$$

В случае а) в силу предположения индукции существует подпоследовательность X_{n_0+1} длины $n_0 + 1$ последовательности X_{l_0} такая, что $\Delta^S(X_{n_0+1}) = 2^{n_0+1}$. Это и будет искомая подпоследовательность.

Исследуем подробнее случай б). Все решающие правила у нас разбиты на группы правил одинаково классифицирующие объекты последовательности $x_1, \ldots, x_{l_0}$. Их общее число и есть $\Delta^S(x_1, \ldots, x_{l_0})$. Но среди этих групп есть такие, в которых найдутся правила, по-разному классифицирующие объект x_{l_0+1}, и такие группы, правила которых однозначно классифицируют объект x_{l_0+1}. Пусть число групп первого рода равно K_1, а число групп второго рода равно K_2. Тогда

$$\Delta^S(x_1, \ldots, x_{l_0}) = K_1 + K_2, \quad \Delta^S(x_1, \ldots, x_{l_0}, x_{l_0+1}) = 2K_1 + K_2,$$

откуда

$$\Delta^S(x_1, \ldots, x_{l_0}, x_{l_0+1}) = \Delta^S(x_1, \ldots, x_{l_0}) + K_1. \tag{6}$$

Рассмотрим теперь только те решающие правила, которые входят в группы первого рода. Обозначим их систему S'. Тогда $K_1 = \Delta^{S'}(x_1, \ldots, x_{l_0})$. Если

$$K_1 = \Delta^{S'}(x_1, \ldots, x_{l_0}) > \Phi(n_0, l_0),$$

то по предположению индукции существует подпоследовательность X_{n_0} длины n_0 последовательности X_{l_0} такая, что $\Delta^{S'}(X_{n_0}) = 2^{n_0}$, то есть она может быть разбита всеми возможными способами с помощью правил из S'. Но в каждой из групп этих правил, одинаково разбивающих X_{n_0} , найдется правило, относящее объект x_{l_0+1} к первому классу, и правило, относящее его ко второму классу. Поэтому, если мы дополним последовательность X_{n_0} объектом x_{l_0+1}, то эту расширенную последовательность также можно расклассифицировать всеми возможными способами правилами из S' (тем более из S). Это и будет искомая подпоследовательность: ее длина равна $n_0 + 1$. Если же $K_1 = \Delta^{S'}(x_1, \ldots, x_{l_0}) \leqslant \Phi(n_0, l_0)$, то получим в силу (6)

$$\Delta^S(x_1, \ldots, x_{l_0}, x_{l_0+1}) = \Delta^S(x_1, \ldots, x_{l_0}) + K_1 \leqslant \Phi(n_0 + 1, l_0) + \Phi(n_0, l_0).$$

А тогда в силу свойства (3) функции $\Phi(n, l)$

$$\Delta^S(x_1, \ldots, x_{l_0}, x_{l_0+1}) \leqslant \Phi(n_0 + 1, l_0 + 1)$$

в противоречии с допущением. Лемма доказана. ■

Теперь в качестве почти очевидного следствия леммы получим упомянутое замечательное свойство функции роста $M^S(l)$. Действительно, всегда

$$M^S(l) \leqslant 2^l,$$

так как никакую выборку длины l нельзя разбить на два класса более чем 2^l способами. Пусть теперь n — первое значение l, при котором $M^S(l) \neq 2^l$. Это значит, что никакую выборку длины n невозможно разбить всеми возможными способами с помощью правил системы S. Тогда, если бы при каком-то $l > n$ оказалось бы, что $\Delta^S(x_1, \ldots, x_l) > \sum_{i=0}^{n-1} C_l^i = \Phi(n, l)$, то согласно лемме нашлась бы подвыборка X_n длины n, у которой индекс

$$\Delta^S(X_n) = 2^n,$$

чего не может быть.

Можно непосредственно проверить, что функция $\Phi(n, l)$ удовлетворяет неравенству

$$\Phi(n, l) \leqslant \frac{3l^{n-1}}{2(n-1)!},$$

и, таким образом, мажорируется степенной по l функцией с показателем, равным $n - 1$.

Таким образом, для любой системы решающих правил S мы получили результат, что функция роста либо тождественно равна 2^l, либо мажорируется фиксированной степенной функцией с показателем, равным максимальной длине последовательности n_0, которую еще можно разбить на классы всеми возможными способами с помощью решающих правил из S.

В мировой литературе эта величина n_0 получила название VC-dimension — размерность Вапника–Червоненкиса (в случае, когда $M^S(l) = 2^l$, эта размерность считается бесконечной). Но выяснилось, что еще до нас подобный результат был получен в алгебраических исследованиях, не имевших отношения ни к распознаванию, ни к статистике. Поэтому я предлагаю называть ее комбинаторной размерностью системы S.

Можно заметить значительные аналогии между приведенным выше доказательством и рассказанным на прошлой лекции способом оценки того, каково максимальное число вариантов разделения конечного числа точек с помощью линейных решающих правил. И полученные функции $\Phi(n, l)$ оказались очень близки. Однако в данном доказательстве не использовались никакие линейные операции и вообще не предполагалось параметрическое описание объектов или решающих правил.

Возвращаясь к линейным решающим правилам, отметим, что максимальное число точек в n-мерном пространстве, которое может быть разбито на два класса всеми возможными способами, равно $n+1$. То есть $n_0 = n+1$. Во многих случаях также оказывается, что значение n_0 равно числу параметров, описывающих решающее правило. Но в общем случае это, конечно, неверно. Расположим, например, $n+1$ точку в n-мерном пространстве так, чтобы их можно было разделить всеми возможными способами с помощью линейных решающих правил. Это можно сделать, поставив их в вершины n-мерного симплекса. Каждому разбиению этих точек на два класса соответствует некоторое линейное правило и определенная точка в пространстве его коэффициентов. Соединяя эти точки непрерывной (или даже дифференцируемой) кривой, мы получим однопараметрическое семейство решающих правил, но оно по-прежнему содержит все способы разбиения наших объектов.

Попытаемся теперь, воспользовавшись неравенством (2) ($P_H(\varepsilon) < M^S(2l)[1-\varepsilon/2]^l$), оценить достаточную длину обучающей выборки, если известна только комбинаторная размерность n_0, заданы точность и требуемая вероятность успеха $1-\eta$ ($P_H(\varepsilon) = \eta$). Разрешая неравенство (2) относительно l, можно получить оценку

$$l_{\text{дост}} = \frac{4n_0}{\varepsilon}\left(1 - \ln\left(\frac{\varepsilon}{4}\right) - \ln\left(\frac{\eta}{3}\right)\right). \qquad (7)$$

Таким образом, достаточная длина выборки при фиксированных значениях ε и η растет линейно с увеличением комбинаторной размерности n_0.

Покажем, что, если комбинаторная размерность системы S бесконечна, то невозможно получить никакой нетривиальной оценки возможности обучения, если заданы только система решающих правил и длина выборки l. Рассмотрим такую игру. Судья задает систему решающих правил S и длину обучающей выборки. Один игрок имеет право

выбрать распределение вероятностей на множестве объектов и одно из решающих правил, предложенных судьей. Выбор может быть детерминированным или случайным. После этого второму игроку предъявляется обучающая выборка длины l, полученная в серии независимых испытаний при распределении, заданном первым игроком, и классификация объектов выборки в соответствии с выбранным первым игроком решающим правилом. Второй игрок может воспользоваться любым алгоритмом обучения, и выбрать свое решающее правило (детерминированное или допускающее случайность), не обязательно принадлежащее системе S. Это игра с нулевой суммой, оцениваемая как вероятность ошибки для решающего правила, выбранного вторым игроком, на множестве всех объектов при распределении, выбранном первым игроком.

Второй игрок имеет тривиальную стратегию — относить любой объект к первому или второму классу случайно с вероятностью 0.5. В этом случае вероятность ошибки будет 0.5 при любой стратегии первого игрока. Покажем, что есть стратегии первого игрока, такие, что при любой стратегии второго игрока вероятность ошибки будет сколь угодно близка к 0.5.

Поскольку комбинаторная размерность бесконечна, первый игрок может задаться большим числом N, и выбрать Nl объектов так, чтобы они могли быть разделены всеми возможными способами с помощью правил из S. Далее он сосредотачивает распределение на этих объектах и объявляет их равновероятными. Затем этот игрок случайно назначает классификацию этих объектов, с вероятностью $1/2$ относя их к первому или второму классу. Полученная классификация, может быть реализована одним из решающих правил системы S, потому что все классификации допустимы.

Из отобранных Nl объектов не более l (возможны повторы) будут представлены в обучающей выборке. Классификацию остальных невозможно определить никаким алгоритмом обучения, поскольку она выбиралась случайно, и, значит, вероятность ошибки на таком объекте составит 0.5. Вероятность встретить объект, не представленный в обучающей выборке, не меньше $(Nl - l)/(Nl) = 1 - l/N$. Поэтому вероятность ошибки при любой стратегии второго игрока не меньше $0.5(1 - l/N)$. Выбирая N достаточно большим, первый игрок может сколь угодно приблизить ее к 0.5, сделав обучение фактически невозможным.

Это, конечно, не значит, что невозможно получить разумную

оценку, если будут привлечены какие-либо дополнительные сведения о распределении вероятностей на множестве объектов. Например, мы видели, что для метода обучения машины средствами SVM и ряда других получается оценка достаточной длины обучающей последовательности порядка $[D/\rho]^2$, где D — диаметр множества, ρ — расстояние между выпуклыми оболочками классов в спрямляющем пространстве, хотя комбинаторная размерность в этом случае может быть бесконечной. В случае, когда множество объектов счетно, алгоритм с полной памятью всегда рано или поздно сходится, но дать оценку достаточной длины обучающей выборки при бесконечной комбинаторной размерности невозможно. В дальнейшем мы увидим, что алгоритмы с полной памятью, основанные на оптимизации сложности решающего правила, также рано или поздно сходятся к правильному решению, несмотря на то, что комбинаторная размерность системы может быть бесконечной. Но дать оценку длины обучающей выборки, при которой с высокой вероятностью гарантировалась бы заданная точность решения, невозможно, если нет никаких дополнительных сведений, кроме класса решающих правил.

Выше мы получили оценку достаточной длины обучающей последовательности, когда известна комбинаторная размерность системы решающих правил S. Попробуем получить оценку необходимой длины обучения в той же игровой постановке, когда судья задает класс решающих правил, один игрок выбирает распределение вероятностей на множестве объектов и одно из допустимых правил, стараясь по возможности затруднить обучение, а другой игрок может использовать любой алгоритм обучения. Значение выигрыша-проигрыша в этой игре равно вероятности ошибки для решающего правила, найденного вторым игроком, при распределении заданном первым игроком. Поскольку стратегии игроков могут содержать случайность (и конкретная обучающая выборка тоже случайна), то значение выигрыша-проигрыша будет случайной величиной, и мы будем оценивать качество стратегий, как математическое ожидание вероятности ошибки.

Пусть n_0 — комбинаторная размерность системы S. Первый игрок выбирает n_0 объектов так, чтобы их можно было произвольно классифицировать с помощью решающих правил из S. Распределение сосредотачивается на этом множестве объектов. Одному из них присваивается вероятность $1 - p$, а остальные считаются равновероятными с вероятностью $p/(n_0 - 1)$. Параметр p подберем позже, чтобы макси-

мально затруднить обучение. Затем этот игрок случайно и независимо классифицирует эти объекты, с вероятностью $1/2$ относя их к первому или второму классу. Любая такая классификация допустима в силу выбора множества объектов, на которых сосредоточена вероятность. Второму игроку предъявляется обучающая выборка длины l, случайно построенная в соответствии с указанным распределением. Классификация объектов, не представленных в обучающей выборке, не может быть предсказана никаким алгоритмом обучения, поскольку первый игрок классифицировал их случайно и независимо от объектов, встретившихся в обучении. Поэтому вероятность ошибки на таком объекте равна 0.5. Математическое ожидание вероятности ошибки после обучения будет

$$EP_{\text{ош}} \geqslant 0.5p\,Q(l),$$

где Q — математическое ожидание доли объектов, не представленных в обучении, если обучающая выборка длины l получена случайно в соответствии с распределением вероятностей, выбранным первым игроком.

Это математическое ожидание достаточно хорошо оценивается формулой

$$Q(l) \approx \exp\left(-\frac{lp}{n_0}\right), \quad \text{откуда } EP_{\text{ош}} \geqslant 0.5p\exp\left(-\frac{lp}{n_0}\right). \tag{8}$$

Подберем теперь p так, чтобы максимизировать эту оценку. Максимум выражения $p\exp(-lp/n_0)$ достигается при $p = n_0/l$. (Нас будут интересовать только те случаи, когда $l > n_0$). Подставляя это значение в (8), получим

$$EP_{\text{ош}} \geqslant 0.5e^{-1}\frac{n_0}{l}.$$

Отсюда получаем, что, если мы хотим, чтобы математическое ожидание вероятности ошибки не превысило ε, необходима длина обучающей выборки

$$l_{\text{необ}} = 0.5e^{-1}\frac{n_0}{\varepsilon}. \tag{9}$$

Сравнивая эту оценку с оценкой достаточной длины выборки (7), видим, что они отличаются лишь множителем при выражении n_0/ε.

Заметим, что когда мы говорим о необходимой длине обучающей выборки, это не значит, что для любой задачи обучения распознаванию, для которой известно, что безошибочное решающее правило принадлежит классу S, такая длина выборки необходима. Утверждается только

то, что существует такая задача, где это необходимо. Поэтому нельзя найти оценку, лучшую чем (9), иначе она была бы приложима ко всем задачам, включая приведенную выше, что неверно.

В этой лекции мы рассмотрели алгоритмы с полной памятью, когда предполагалось, что в заданном классе есть безошибочное решающее правило, и обучающаяся машина выбирает из них такое, которое не делает ошибок на данных обучения. Общий случай минимизации эмпирического риска мы рассмотрим в следующих лекциях. Но оказывается, что техника, предложенная здесь, почти без изменений работает и в общем случае.

Лекция 11

Критерии равномерной сходимости частот к вероятностям

В этой лекции мы рассмотрим условия равномерной сходимости частот к вероятностям по классу событий [4]. Согласно теореме Бернулли, частота выпадения некоторого события A сходится (по вероятности) в последовательности независимых испытаний к вероятности этого события. Мы убедились, однако, что возникает необходимость судить одновременно о вероятностях событий целого класса S по одной и той же выборке. При этом требуется, чтобы частота событий сходилась к вероятности равномерно по всем событиям класса S. Точнее, требуется, чтобы вероятность того, что максимальное по классу уклонение частоты от вероятности превзойдет сколь угодно малую положительную константу, стремилась к нулю при неограниченном увеличении числа испытаний.

Оказывается, даже в простейших примерах такая равномерная сходимость может не иметь места. Поэтому хотелось бы найти критерий, по которому можно было бы судить, есть ли такая сходимость или же ее нет.

Строгая постановка задачи

Пусть X — множество элементарных событий, на котором задана вероятностная мера $P(x)$. Пусть S — некоторая совокупность случай-

ных событий, т. е. измеримых подмножеств пространства X. Пространство выборок длины l, полученных в процессе независимых испытаний при неизменном распределении, обозначим $X(l)$. Вероятностная мера в этом пространстве определяется из условия

$$P[A_1 \cdot A_2 \cdot \ldots \cdot A_l] = P(A_1)P(A_2)\ldots P(A_l),$$

где A_i — события в пространстве, $P(A_i)$ — их вероятность. Для каждой выборки $\mathbf{X}^l = x_1, \ldots, x_l$ и события A определена частота выпадения события A, равная отношению числа $n(A)$ элементов выборки, принадлежащих A, к общей длине выборки:

$$\nu(A; x_1, \ldots, x_l) = \frac{n(A)}{l}.$$

Теорема Бернулли утверждает, что при фиксированном событии A уклонение частоты от вероятности стремится к нулю (по вероятности) с ростом объема выборки. Нас же будет интересовать максимальное по классу S уклонение частоты от вероятности

$$\pi^S(x_1, \ldots, x_l) = \sup_{A \in S} |\nu(A; x_1, \ldots, x_l) - P(A)|.$$

Величина $\pi^S(x_1, \ldots, x_l)$ является функцией выборки. Будем предполагать, что эта функция измерима в пространстве выборок $X(l)$, т. е. является случайной величиной.

Если величина $\pi^S(x_1, \ldots, x_l)$ стремится по вероятности (или с вероятностью 1) к нулю при неограниченном увеличении выборки, то мы говорим, что частота событий $A \in S$ стремится по вероятности (или вероятностью 1) к их вероятности равномерно по классу событий S.

В отличие от закона больших чисел равномерная сходимость частот к вероятностям может иметь или не иметь место в зависимости от того, как выбрана система событий S и задана вероятностная мера $P(x)$.

Приведем несколько примеров, когда равномерной сходимости нет. Пусть на сегменте [0,1] задано равномерное распределение. Система S состоит из всевозможных множеств, являющихся объединением конечного (но любого) числа интервалов. Пусть задана произвольная выборка $x_1, \ldots, x_l$. Ее можно покрыть конечным числом интервалов. Множество A, образованное объединением этих интервалов, будет содержать все точки выборки, и значит, частота выпадения события A

будет равна 1. В то же время, выбрав длину интервалов достаточно малой, можно добиться того, что величина $P(A)$ будет сколь угодно близка к нулю. Следовательно, в этом примере

$$\pi^S(x_1, \ldots, x_l) = \sup_{A \in S} |\nu(A; x_1, \ldots, x_l) - P(A)| \equiv 1$$

при любой длине выборки.

Другой пример. Пусть система S состоит из всех замкнутых выпуклых множеств в евклидовом пространстве размерности n не меньше двух. Тогда, если вероятностная мера сосредоточена на некоторой строго выпуклой поверхности C размерности $n-1$, и вероятность одноточечных множеств равна нулю, то равномерной сходимости нет. Действительно, в этом случае с вероятностью 1 все точки выборки будут сосредоточены на этой поверхности. Натянем на точки выборки замкнутую выпуклую оболочку A. Выпуклая оболочка A принадлежит системе S и содержит все точки выборки. Поэтому частота выпадения события равна 1. В то же время пересечение оболочки со строго выпуклой поверхностью C будет состоять только из конечного множества точек. Поэтому вероятность $P(A)$ равна нулю. Следовательно, при любой длине выборки найдется событие из системы S, для которого частота отклоняется от вероятности на единицу. Если же распределение $P(x)$ имеет плотность, то оказывается, что равномерная сходимость имеет место. В этом примере факт равномерной сходимости зависит от распределения вероятностей.

Достаточно очевидно, что для конечных систем, содержащих N событий, равномерная сходимость всегда имеет место. Основная идея выводимых ниже условий равномерной сходимости связана с тем, что и в случае бесконечной системы, лишь конечное число событий различимо на конечной выборке. Правда это число не постоянно, а зависит от длины выборки. Если это число возрастает с длиной выборки достаточно медленно (медленнее любой показательной функции), то оказывается, что равномерная сходимость есть. В противном случае ее нет.

Связь с задачами обучения распознаванию образов

Как же проблема равномерной сходимости частот к вероятностям связана с задачей обучения распознаванию образов? Пусть совместное рас-

пределение входных сигналов $\mathbf{x}$ и реакций учителя $\mathbf{y}$ определено как $P(\mathbf{x}, \mathbf{y})$ и задана система S решающих правил $F(\mathbf{x}, \alpha)$. Рассмотрим систему событий $A = \{F(\mathbf{x}, \alpha) \neq \mathbf{y}\}$. Тогда вероятность этих событий равна вероятности ошибки для решающего правила $F(\mathbf{x})$, а частота выпадений на выборке — частоте ошибок на данных обучения. Таким образом, равномерная сходимость эмпирического риска к истинному равносильна равномерной сходимости частот к вероятностям по этому классу событий.

Пусть минимум истинного риска $R_{\text{ист}}(\alpha)$ достигается в точке α_0, а минимум эмпирического риска $R_{\text{эмп}}(\alpha)$ достигается в точке α^*. Если значения эмпирического риска для всех решающих правил заданного класса отличаются от значений истинного риска не более чем на ε, то значение истинного риска в точке α^* будет отличаться от минимума истинного риска $R_0 = R_{\text{ист}}(\alpha_0)$ не более чем на 2ε. Действительно, в этом случае

$$R_{\text{эмп}}(\alpha_0) \leqslant R_0 + \varepsilon.$$
$$R_{\text{эмп}}(\alpha_0) \geqslant R_{\text{эмп}}(\alpha^*) \qquad \text{(поскольку при } \alpha^* \text{ достигается минимум } R_{\text{эмп}}(\alpha)).$$

Откуда

$$R_{\text{эмп}}(\alpha^*) \leqslant R_0 + \varepsilon.$$

В то же время из условия равномерной близости истинного и эмпирического риска имеем

$$R_{\text{ист}}(\alpha^*) \leqslant R_{\text{эмп}}(\alpha*) + \varepsilon.$$

Откуда

$$R_{\text{ист}}(\alpha^*) \leqslant R_0 + 2\varepsilon.$$

С другой стороны, каждому событию A системы S можно поставить в соответствие решающее правило $R_A(\mathbf{x}) = \mathbf{I}_A(\mathbf{x})$, где $\mathbf{I}_A$ — индикаторная функция множества A. Тогда, как и на прошлой лекции, для системы S можно определить индекс системы относительно выборки $x_1, \ldots, x_l$, функцию роста и комбинаторную размерность. Индекс системы $\Delta^S(x_1, \ldots, x_l)$ определяется как число групп событий (множеств

A), внутри каждой из которых события неразличимы, т. е. если элемент x_i выборки принадлежит одному из множеств группы, то он принадлежит и любому другому множеству этой группы. Функция роста $M^S(l)$ определяется как максимальное значение индекса по всем последовательностям $x_1, \ldots, x_l$ длины l, а комбинаторная размерность — как максимальное значение l, при котором еще $M^S(l) = 2^l$.

Основная лемма

Как было уже сказано, основная идея, на которой строятся условия равномерной сходимости частот к вероятностям, состоит в том, что бесконечная система событий S заменяется конечной подсистемой, состоящей из событий, различимых на конечной выборке. Для того, чтобы сделать такой переход корректным, оказывается необходимым заменить исходную проблему равномерной близости частот событий к их вероятностям проблемой равномерной близости частот событий в двух следующих друг за другом выборках одинаковой длины.

Оказывается, что равномерная сходимость к нулю разности частот в двух полувыборках является необходимой и достаточной для равномерной сходимости частот к вероятностям, и из оценок скорости одной сходимости следуют оценки для другой.

Итак, пусть дана выборка длины $2l$:

$$\mathbf{X}^{2l} = x_1, \ldots, x_l, \ldots, x_{2l},$$

и подсчитаны частоты выпадения события $A \in S$ на первой полувыборке $\mathbf{X}_1^l = x_1, \ldots, x_l$ и второй полувыборке $\mathbf{X}_2^l = x_{l+1}, \ldots, x_{2l}$. Обозначим соответственно эти частоты как $\nu'(A)$ и $\nu''(A)$ и рассмотрим отклонение этих величин

$$\rho_A(x_1, \ldots, x_l, \ldots, x_{2l}) = |\nu'(A) - \nu''(A)|.$$

Нас будет интересовать максимальное отклонение этих частот по всем событиям системы S:

$$\rho^S(x_1, \ldots, x_l, \ldots, x_{2l}) = \sup_{A \in S} \rho_A(x_1, \ldots, x_l, \ldots, x_{2l}).$$

Напомним, что через $\pi^S(x_1, \ldots, x_l)$ мы обозначили

$$\pi^S(x_1, \ldots, x_l) = \sup_{A \in S} |\nu(A) - P(A)|.$$

Далее будем полагать, что как $\pi^S(x_1, \ldots, x_l)$, так и $\rho^S(x_1, \ldots, x_l, \ldots, x_{2l})$ — измеримые функции выборки.

Основная лемма. Распределения величин $\pi^S(x_1, \ldots, x_l)$ и $\rho^S(x_1, \ldots, x_l, \ldots, x_{2l})$ связаны соотношением

$$P(\pi^S(x_1, \ldots, x_l) > \varepsilon) \leqslant 2P(\rho^S(x_1, \ldots, x_l, \ldots, x_{2l}) > \varepsilon/2),$$

если только $l \geqslant 2/\varepsilon$.

Доказательство. Доказательство этого утверждения построено по следующей схеме. Полувыборки $x_1, \ldots, x_l$ и $x_{l+1}, \ldots, x_{2l}$ берутся последовательно и независимо. Допустим, что первая полувыборка оказалась такой, что

$$\sup_{A \in S} |\nu'(A; x_1, \ldots, x_l) - P(A)| > \varepsilon.$$

Это значит, что в классе S нашлось такое событие A^*, что

$$|\nu'(A^*; x_1, \ldots, x_l) - P(A^*)| > \varepsilon.$$

На второй полувыборке будем следить за отклонением частоты от вероятности лишь для этого фиксированного события A^*. Так как нас интересует одно событие, то можно воспользоваться обычным законом больших чисел. Поэтому при достаточно большом значении l с достаточно высокой вероятностью

$$|\nu''(A^*; x_{l+1}, \ldots, x_{2l}) - P(A^*)| < \frac{\varepsilon}{2}$$

и, следовательно,

$$|\nu'(A) - \nu''(A)| > \frac{\varepsilon}{2} \quad \text{и} \quad \rho^S(x_1, \ldots, x_l, \ldots, x_{2l}) > \frac{\varepsilon}{2}.$$

Перейдем к формальному доказательству.

По определению

$$P\left(\rho^S(x_1, \ldots, x_l, \ldots, x_{2l}) > \frac{\varepsilon}{2}\right) = \\ = \int \theta\left(\rho^S(x_1, \ldots, x_l, \ldots, x_{2l}) - \frac{\varepsilon}{2}\right) dp(\mathbf{X}^{2l}),$$

где $\theta(z)$ равно 1 при $z > 0$, и равно 0 при $z \leqslant 0$.

Учитывая, что пространство выборок $\mathbf{X}^{2l}$ длины $2l$ есть прямое произведение $\mathbf{X}_1^l$ и $\mathbf{X}_2^l$ полувыборок длины l, представим наш интеграл в виде

$$\int \theta\left(\rho^S(x_1,\dots x_{2l}) - \frac{\varepsilon}{2}\right)\, dp(\mathbf{X}^{2l}) = \\ = \int \left[\int \theta\left(\rho^S(x_1,\dots x_{2l}) - \frac{\varepsilon}{2}\right)\, dp(\mathbf{X}_2^l)\right]\, dp(\mathbf{X}_1^l)$$

(во внутреннем интеграле первая полувыборка фиксирована).

Обозначим через Q событие в пространстве выборок $\mathbf{X}_1^l$

$$Q = \pi^S(x_1,\dots,x_l) > \varepsilon,$$

и, ограничивая интегрирование, получим

$$P\left(\rho^S(\mathbf{X}^{2l}) > \frac{\varepsilon}{2}\right) \geqslant \int\limits_{\mathbf{X}_1^l \in Q} \left[\int \theta\left(\rho^S(x_1,\dots x_{2l}) - \frac{\varepsilon}{2}\right)\, dp(\mathbf{X}_2^l)\right]\, dp(\mathbf{X}_1^l).$$

Оценим внутренний интеграл правой части неравенства, обозначив его через $\mathbf{I}$. Здесь последовательность $x_1,\dots x_l$ фиксирована и такова, что $\pi^S(x_1,\dots,x_l) > \varepsilon$. Следовательно, существует $A^* \in S$ такое, что $|\nu'(A^*; x_1,\dots,x_l) - P(A^*)| > \varepsilon$. Тогда

$$\mathbf{I} = \int \theta\left(\sup_{A\in S} \rho_A(x_1,\dots,x_{2l}) - \frac{\varepsilon}{2}\right)\, dp(\mathbf{X}_2^l) \geqslant \\ \geqslant \int \theta\left(\rho_{A^*}(x_1,\dots x_{2l}) - \frac{\varepsilon}{2}\right)\, dp(\mathbf{X}_2^l).$$

Пусть, например,

$$\nu'(A^*; x_1,\dots,x_l) < P(A^*) - \varepsilon$$

(аналогично рассматривается случай $\nu'(A^*; x_1,\dots,x_l) > P(A^*) + \varepsilon$).

Тогда для выполнения условия

$$|\nu'(A^*; x_1,\dots,x_l) - \nu''(A^*; x_{l+1},\dots,x_{2l})| > \frac{\varepsilon}{2}$$

достаточно потребовать, чтобы выполнялось

$$\nu''(A^*; x_{l+1},\dots,x_{2l}) > P(A^*) - \frac{\varepsilon}{2},$$

откуда на основании биномиального распределения получаем

$$\mathbf{I} \geqslant \int \theta \left[\nu''(A^*; x_{l+1}, \ldots, x_{2l}) - \left(P(A^*) - \frac{\varepsilon}{2}\right)\right] dp(\mathbf{X}_2^l) = \\ = \sum_{\frac{k}{l} > P(A^*) - \frac{\varepsilon}{2}} C_l^k P(A^*)^k (1 - P(A^*))^{l-k}.$$

Как известно, последняя сумма превосходит $1/2$, если только $l > 2/\varepsilon$. Поэтому при $l > 2/\varepsilon$

$$P(\rho^S\left(\mathbf{X}^{2l}\right) > \frac{\varepsilon}{2}\Big) \geqslant \frac{1}{2} \int (\mathbf{X}_1^l \in Q)\, dp(\mathbf{X}_1^l) = \frac{1}{2} P(\pi^S(x_1, \ldots, x_l) > \varepsilon),$$

что и требуется. ■

Достаточное условие равномерной сходимости и оценка

Итак, задача может быть сведена к оценке равномерной близости частот выпадения событий в двух последующих полувыборках. Схему сравнения частот выпадения событий в двух полувыборках можно представить себе так. Берется выборка двойной длины $\mathbf{X}^{2l}$ и затем делится случайным образом на две полувыборки равной длины. Будем считать, что выборка $\mathbf{X}^{2l}$ зафиксирована. Если два события A_1 и A_2 неразличимы на выборке $\mathbf{X}^{2l}$, то есть всякий элемент этой выборки, принадлежащий A_1, принадлежит и A_2, и наоборот, то частоты выпадений этих событий на всякой подвыборке одинаковы. Поэтому для оценки максимального уклонения частот достаточно из всякой группы неразличимых событий взять по одному. Число таких событий будет конечно и равно индексу $\Delta^S(x_1, \ldots, x_{2l})$ системы S относительно выборки $x_1, \ldots, x_{2l}$. Рассмотрим одно из таких событий A и, по-прежнему считая выборку $\mathbf{X}^{2l}$ фиксированной, разобьем ее случайно на две полувыборки равной длины и оценим уклонение частот этого события в двух полувыборках. Это схема равноценна схеме с невозвращаемыми шарами, а потому, как известно из литературы, уклонение частот подчиняется гипергеометрическому распределению. Тогда получим

$$P|\nu'(A; x_1, \ldots, x_l) - \nu''(A; x_{l+1}, \ldots, x_{2l})| > \varepsilon = \sum \frac{C_m^k C_{2l-m}^{l-k}}{C_{2l}^l},$$

где m — число выпадений события A на полной выборке $\mathbf{X}^{2l}$, k — число выпадений этого события на первой полувыборке (тогда $(m-k)$ будет числом выпадений этого события на второй полувыборке), а суммирование производится по тем значениям k, для которых

$$\left|\frac{k}{l} - \frac{m-k}{l}\right| > \varepsilon.$$

Путем несложных преобразований сумма в правой части этого равенства может быть оценена сверху при любых $\varepsilon > 0$ и $l > 0$ как

$$\sum \frac{C_m^k C_{2l-m}^{l-k}}{C_{2l}^l} \leqslant 3\exp[-\varepsilon^2(l-1)].$$

Формально переход к схеме испытаний с невозвращаемыми шарами осуществляется таким путем. Пусть требуется оценить интеграл вида $I = \int \varphi(\mathbf{X}^{2l})\, dp(\mathbf{X}^{2l})$ при условии, что вероятностная мера инвариантна относительно перестановки элементов выборки $\mathbf{X}^{2l}$. Это условие в нашем случае выполнено. Тогда

$$I = \int \left[\frac{1}{(2l)!}\sum \varphi(T_i\mathbf{X}^{2l})\right] dp(\mathbf{X}^{2l}),$$

где сумма берется по всем $(2l)!$ перестановкам T_i последовательности $\mathbf{X}^{2l}$.

Теперь оценивать можно подынтегральное выражение, заключенное в квадратные скобки. Применительно к разбиению на две полувыборки это приводит к схеме с невозвращаемыми шарами. Таким образом,

$$P(|\nu'(A; x_1, \ldots, x_l) - \nu''(A; x_{l+1}, \ldots, x_{2l})| > \varepsilon) \leqslant 3\exp[-\varepsilon^2(l-1)].$$

Вероятность того, что хотя бы для одного из этих событий уклонение частот превысит ε, по теореме о сложении вероятностей оценивается как

$$P\left(\sup_{A\in S} |\nu'(A; x_1, \ldots, x_l) - \nu''(A; x_{l+1}, \ldots, x_{2l})| > \varepsilon\right) \leqslant$$
$$\leqslant 3\Delta^S(x_1, \ldots, x_{2l})\exp[-\varepsilon^2(l-1)].$$

В свою очередь, по определению функции роста

$$\Delta^S(x_1, \ldots, x_{2l}) \leqslant M^S(2l),$$

и поэтому

$$P\left(\sup_{A \in S} \rho_A(x_1, \ldots, x_l, \ldots, x_{2l}) > \varepsilon\right) \leqslant 3M^S(2l) \exp[-\varepsilon^2(l-1)].$$

Соединяя этот результат с утверждением основной леммы, получим

$$P\left(\sup_{A \in S} |\nu(A; x_1, \ldots, x_l) - P(A)| > \varepsilon\right) \leqslant 6M^S(2l) \exp\left(-\frac{1}{4}\varepsilon^2(l-1)\right).$$

Очевидно, если комбинаторная размерность системы S конечна, и, значит, $M^S(l)$ растет лишь степенным образом, то правая часть неравенства стремится к нулю при $l \to \infty$. Это и дает достаточное условие равномерной сходимости (по вероятности). В этих условиях нетрудно показать, что равномерная сходимость имеет место и с вероятностью 1 (почти наверное).

Если же комбинаторная размерность бесконечна, то правая часть неравенства будет больше 1 при любых l и ε, и оценка становится тривиальной.

Заметим, что полученное условие и оценка вероятности того, что максимальное уклонение частоты события от его вероятности превзойдет заданную положительную величину, не зависят от распределения вероятностей на множестве объектов X.

Строя конструкции, аналогичные тем, которые приводились на предыдущей лекции, можно показать, что, если комбинаторная размерность системы S бесконечна, то без привлечения сведений о распределении $P(x)$ невозможно получить никакой нетривиальной оценки равномерной близости частот к вероятностям ни для какого конечного значения длины выборки l.

Энтропийный критерий

Но верно ли, что если комбинаторная размерность системы S бесконечна, то равномерной сходимости нет? Ведь, заменив индекс $\Delta^S(x_1, \ldots, x_l)$ на его максимум по всем выборкам длины l, мы слишком жестко оцениваем ситуацию. Оказывается, это действительно так.

Необходимые и достаточные условия равномерной сходимости частот к вероятностям получаются с использованием математического ожидания логарифма индекса (конечно, в предположении измеримости его как функции выборки).

Определим энтропию системы событий S относительно выборок длины l как математическое ожидание логарифма индекса:

$$H^S(l) = E \log \Delta^S(x_1, \ldots, x_l).$$

В отличие от функции роста энтропия зависит от распределения $P(x)$. Очевидно, что всегда выполняются неравенства:

$$0 \leqslant H^S(l) \leqslant l,$$
$$H^S(l) \leqslant \log M^S(l).$$

Всегда существует предел энтропии, деленной на длину выборки:

$$\lim_{l\to\infty} \frac{H^S(l)}{l} = C,$$

причем величина C лежит в пределах $0 \leqslant C \leqslant 1$. Более того, оказывается, что случайная величина $\log \Delta^S(x_1, \ldots, x_l)/l$ тоже стремится по вероятности к тому же значению C. Будем называть эту величину *предельной энтропией на символ*.

Если значение $C > 0$, то индекс системы растет в среднем экспоненциально. Если же $C = 0$, то индекс в среднем растет медленнее любой экспоненты. Легко видеть, что если комбинаторная размерность системы конечна и функция роста мажорируется показательной функцией, то величина C всегда равна нулю. Если же комбинаторная размерность бесконечна, то значение C может быть отлично от нуля. Есть примеры систем, у которых комбинаторная размерность бесконечна, но величина C все равно равна нулю при любом распределении $P(x)$. Но во многих случаях при бесконечной комбинаторной размерности величина C может принимать любые значения между нулем и единицей в зависимости от конкретного распределения.

Для того, чтобы выполнялась равномерная сходимость частот к вероятностям, т. е.

$$P\left(\sup_{A\in S} |\nu(A; x_1, \ldots, x_l) - P(A)| > \varepsilon\right) \to 0$$

при $l \to \infty$, необходимо и достаточно, чтобы величина C равнялась нулю, т. е.

$$\lim_{l \to 0} \frac{H^S(l)}{l} = 0.$$

Достаточность этого условия доказывается аналогично выводу достаточных условий с использованием функции роста, опираясь на то, что значение индекса растет с увеличением длины выборки медленнее любой экспоненты. Разница, однако, в том, что в случае конечной комбинаторной размерности мы смогли получить оценку приближения $P(\pi^S(x_1, \ldots, x_l) > \varepsilon)$ к нулю с ростом l, не зависящую от распределения, если же нам известно только то , что предельная энтропия на символ $C = 0$, и комбинаторная размерность системы бесконечна, то такую оценку получить невозможно. (Как отмечалось, есть примеры, когда, несмотря на то, что комбинаторная размерность системы бесконечна, величина C равна нулю при любом распределении. В частности, это так, если множество объектов бесконечно, но счетно).

Если же предельная энтропия на символ больше нуля, т. е. $\lim(H^S(l)/l) = C > 0$, то оказывается справедливым такое утверждение. Существует подмножество X_0, мера которого в точности равна C, обладающее следующим свойством: пусть дана выборка $x_1, \ldots, x_l$, полученная в процессе независимых испытаний при неизменном распределении $P(x)$. Пусть $x_{i_1}, \ldots, x_{i_n}$ — подвыборка длины n, состоящая из всех элементов исходной выборки, попавших в множество X_0. Длина этой подвыборки n при больших l будет близка к Cl. Тогда с вероятностью 1 индекс системы относительно этой подвыбрки равен 2^n, т. е.

$$\Delta^S(x_{i_1}, \ldots, x_{i_n}) = 2^n.$$

Иными словами, с вероятностью 1 эта подвыборка может быть разбита всеми возможными способами с помощью множеств системы S. Доказывается это утверждение довольно сложно, и я не буду приводить доказательство.

Далее при использовании этого утверждения с помощью техники двух полувыборок можно показать, что с высокой вероятностью максимальное по классу событий уклонение частот в двух полувыборках остается большим даже при неограниченном увеличении длины всей выборки. А этого не может быть, если имеет место равномерная сходимость частот к вероятностям.

Лекция 12

Критерии равномерной сходимости средних к мат. ожиданиям. Проблема выбора оптимальной сложности модели.

В общем случае вопрос о равномерной близости эмпирического и истинного риска сводится у равномерной по параметру сходимости средних к математическим ожиданиям.

Сформулируем в точных терминах эту проблему. Пусть x — элементарное событие из пространства X, $P(x)$ — вероятностная мера в нем, α — некоторый абстрактный параметр, $F(x, \alpha)$ — некоторая функция, измеримая при всех α относительно меры $P(x)$ в пространстве X.

Предположим, что существует математическое ожидание этой функции при всех α:

$$M(\alpha) = EF(x, \alpha) = \int F(x, \alpha)\, dP(x).$$

Пусть далее задана выборка $\mathbf{X}^l = x_1, \ldots, x_l$, полученная в ходе независимых испытаний при неизменном распределении $P(x)$. Тогда для каждого значения α по этой выборке можно определить среднее значение

$$M_{\text{эмп}}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} F(x_i, \alpha).$$

Если бы величина α была постоянной, то сходимость среднего к математическому ожиданию обеспечивалась бы законом больших чисел. Но если абстрактный параметр α может изменяться в пределах некоторого множества Ω, то возникает вопрос о равномерности по параметру α оценки математического ожидания средним значением. Равномерность близости средних к математическим ожиданиям может быть оценена величиной

$$P_\varepsilon(\Omega, l) = P\left(\sup_{\alpha \in \Omega} |M(\alpha) - M_{\text{эмп}}(\alpha)| > \varepsilon\right),$$

т. е. вероятностью того, что максимальное уклонение средневыборочного значения от математического ожидания превзойдет $\varepsilon > 0$. Говорят, что имеет место равномерная по параметру сходимость средних к математическим ожиданиям, если случайная величина $\sup_{\alpha \in \Omega} |M(\alpha) -$

$M_{\text{эмп}}(\alpha)|$ стремится к нулю соответственно по вероятности или почти наверное при $l \to \infty$.

Приводимые далее условия такой сходимости сводят этот вопрос к исследованной в предыдущей лекции проблеме равномерной сходимости частот к вероятностям по классу событий.

Достаточные условия

Пусть $F(x, \alpha)$ $(\alpha \in \Omega)$ — семейство измеримых на X функций, причем выполнено условие $0 \leqslant F(x, \alpha) \leqslant a$ (число a не зависит ни от x, ни от α). Рассмотрим систему событий S вида

$$A(\alpha, c) = \{x : F(x, \alpha) \geqslant c\}$$

для всевозможных значений α и c.

Тогда равномерная сходимость частот к вероятностям по классу событий S является достаточным условием для равномерной сходимости средних к математическим ожиданиям. При этом выполняется соотношение

$$\sup_{\alpha \in \Omega} |M(\alpha) - M_{\text{эмп}}(\alpha)| \leqslant a \sup_{A \in S} |\nu(A; x1, \ldots, xl) - P(A)|.$$

Доказательство. Согласно определению интеграла Лебега

$$M(\alpha) = \int F(x, \alpha)\, dP(x) = \lim_{n \to \infty} \sum_{i=0}^{n} \frac{a}{n} P\left\{x : F(x, \alpha) \geqslant \frac{ia}{n}\right\}.$$

Аналогично

$$M_{\text{эмп}}(\alpha) = \lim_{n \to \infty} \sum_{i=0}^{n} \frac{a}{n} \nu\left(\left\{x : F(x, \alpha) \geqslant \frac{ia}{n}\right\}, \mathbf{X}^l\right).$$

Обозначим событие $\{x : F(x, \alpha) \geqslant (ia)/n\}$ через $A_{in} \in S$. Тогда

$$|M(\alpha) - M_{\text{эмп}}(\alpha)| \leqslant \lim_{n \to \infty} \sum_{i=0}^{n} \frac{a}{n} |P(A_{in}) - \nu^l(A_{in})| \leqslant$$
$$\leqslant a \sup_{A \in S} |\nu(A; x_1, \ldots, x_l) - P(A)|.$$

Этим и доказывается наше утверждение. ■

Тем самым из оценок равномерной сходимости частот к вероятностям можно всегда получить оценки равномерной сходимости средних к математическим ожиданиям для классов равномерно ограниченных функций $F(x,\alpha)$.

В частности, на основании результата, выведенного на предыдущей лекции, получаем

$$P\left\{\sup_{\alpha\in\Omega}|M(\alpha)-M_{\text{эмп}}(\alpha)|>a\varepsilon\right\}\leqslant 6M^S(2l)\exp\left[-\frac{1}{4}\varepsilon^2(l-1)\right],$$

где $M^S(l)$ — функция роста системы событий $A(\alpha,c)=\{x:F(x,\alpha)\geqslant c\}$.

Вместо функции роста можно использовать введенную на прошлой лекции энтропию системы событий S на выборках длины l. Но, во-первых, для вычисления этой энтропии требуется знать распределение вероятностей на множестве X, которое обычно заранее неизвестно, а во-вторых, необходимые и достаточные условия равномерной сходимости по классу переходят здесь лишь в достаточные условия равномерной сходимости средних к математическим ожиданиям.

Требование $0\leqslant F(x,\alpha)\leqslant a$ можно, конечно, заменить более слабым

$$C_1\leqslant F(x,\alpha)\leqslant C_2,$$

что достигается путем соответствующего сдвига и изменения масштаба. Но равномерная ограниченность функций $F(x,\alpha)$ здесь существенна. В противном случае можно построить примеры, где равномерная сходимость по классу S частот к вероятностям имеет место, тогда как равномерной сходимости средних к математическим ожиданиям нет.

Рассмотрим один такой пример. Пусть задана система функций

$$F(x;a,\sigma)=\frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right),$$

где x — аргумент, а величины a и $\sigma>0$ — параметры. Эта система функций не будет равномерно ограниченной, поскольку при $x=a$ и значении σ, стремящемся к нулю, значение функции уходит в бесконечность. Пусть теперь на сегменте [0,1] задано равномерное распределение $P(x)$. Тогда всегда

$$M(a,\sigma)=\int F(x;a,\sigma)\,dP(x)<1.$$

Для произвольной выборки $x_1, \ldots, x_l$ положим $a = x_i$, где x_i — один из элементов выборки, и устремим к нулю σ. Значение функции $F(x; a, \sigma)$ в точке $x = x_i$ при этом идет к бесконечности, тогда как ее значение в остальных точках выборки остается положительным. Поэтому

$$M_{\text{эмп}}(a, \sigma) = \frac{1}{l} \sum_{i=1}^{l} F(x_i, a, \sigma) \to \infty.$$

Поэтому $\sup_{(a,\sigma)} |M(a, \sigma) - M_{\text{эмп}}(a, \sigma)| = \infty$ для любой выборки и для любой длины выборки. Равномерной сходимости нет. В то же время комбинаторная размерность для системы событий

$$A(a, \sigma, c) = \{x : F(x, a, \sigma) \geqslant c\}$$

равна всего лишь двум, и поэтому равномерная сходимость частот к вероятностям по этому классу событий имеет место.

Требование равномерной ограниченности можно несколько ослабить. Пусть существует функция $K(x)$, не зависящая от α, такая что

$$K(x) \geqslant 0; \qquad \int K(x)\, dx < \infty;$$
$$|F(x, \alpha)| < K(x) \quad \text{при всех} \quad \alpha \in \Omega.$$

Тогда для равномерной сходимости средних к математическим ожиданиям (почти наверное) достаточна равномерная сходимость частот к вероятностям по классу S событий вида $A(\alpha, c) = \{x : F(x, \alpha) \geqslant c\}$.

Еще в 1956 году Ле Кам показал, что равномерная сходимость средних к математическим ожиданиям имеет место, если

1. Параметр α пробегает метрический компакт,

2. Функции $F(x, \alpha)$ непрерывны по α почти при всех $x \in X$,

3. $|F(x, \alpha)| < K(x)$ при всех $\alpha \in \Omega$, где $K(x) \geqslant 0$; $\int K(x)\, dx < \infty$.

Этот результат, однако, неприменим к задачам распознавания, где функция $F(x, \alpha)$ разрывна, и из него не следуют никакие оценки.

Энтропийный критерий

Выше мы показали, как достаточные условия равномерной сходимости средних к математическим ожиданиям могут быть получены путем исследования равномерной сходимости частот к вероятностям по классу событий S вида $A(\alpha, c) = \{x : F(x, \alpha) \geqslant c\}$. Но даже энтропийный критерий сходимости частот к вероятностям дает при этом лишь достаточные условия.

Проиллюстрируем это на следующем примере.

Пример. В качестве системы функций $F(x, \alpha)$ рассмотрим всевозможные числовые функции, разложимые в ряд Фурье (по косинусам) на сегменте $[-\pi, \pi]$ с достаточно быстро убывающими коэффициентами:

$$F(x, \alpha) = \sum a_n \cos(nx), \qquad |a_n| \leqslant C2^{-n},$$

где C — некоторая константа, а вектор коэффициентов a_n соответствует абстрактному параметру α. На сегменте $[-\pi, \pi]$ задано равномерное распределение.

Равномерная сходимость средних к математическим ожиданиям в этом примере имеет место. Это показывается так. Найдем достаточно большое значение n, чтобы сумма коэффициентов ряда с номерами, превышающими n, стала заведомо малой величиной при всех значениях параметра. Тогда хвост ряда будет вносить заведомо малый вклад как в среднее, так и в математическое ожидание. Оставшаяся часть ряда имеет конечную размерность, откуда следует равномерная сходимость.

В то же время комбинаторная размерность системы S событий $A(\alpha, c) = \{x : F(x, \alpha) \geqslant c\}$ в этом примере, как не трудно убедиться, бесконечна. Более того, предельная энтропия на символ для событий системы S равна 1. Поэтому равномерной сходимости частот к вероятностям для системы S нет.

Необходимые и достаточные условия равномерной сходимости средних к математическим ожиданиям, как и в случае сходимости частот к вероятностям, будут, вообще говоря, зависеть от распределения вероятностей на множестве X, которое обычно заранее не известно. Поэтому эти условия носят больше теоретический интерес. Все же целесообразно их привести [6] (для классов равномерно ограниченных функций).

Для этого приходится определить энтропию системы функций $F(x,\alpha)$ прямо, без перехода к системе событий $A(\alpha,c)=\{x : F(x,\alpha)\geqslant c\}$.

Пусть $\mathbf{x}^l=(x_1,x_2,\ldots,x_l)$ — повторная выборка из X длины l при неизменном распределении P. Задана система функций $0\leqslant F(x,\alpha)\leqslant 1$, где $\alpha\in\Omega$ — абстрактный параметр. Для заданного $\varepsilon>0$ рассмотрим множество Y векторов $\mathbf{y}=(y_1,y_2,\ldots,y_l)$ из E^l:

$$Y^\Omega(\mathbf{x}^l,\varepsilon)=\left\{\mathbf{y} : \exists\alpha\in\Lambda\ \forall i\ |F(x_i,\alpha)-y_i|<\frac{\varepsilon}{2}\right\}.$$

Множество $Y^\Omega(\mathbf{x}^l,\varepsilon)$ представляет собой объединение открытых кубов из E^l, ориентированных вдоль осей координат, с ребром длиной ε и центрами, расположенными в концах векторов $(F(x_1,\alpha),\ F(x_2,\alpha),\ldots,$ $F(x_l,\alpha))$, при всевозможных значениях $\alpha\in\Omega$.

Пусть $V^\Omega(\mathbf{x}^l,\varepsilon)$ —объем множества $Y^\Omega(\mathbf{x}^l,\varepsilon)$. Очевидно, что

$$\varepsilon^l\leqslant V^\Omega(\mathbf{x}^l,\varepsilon)\leqslant(1+\varepsilon)^l.$$

Обозначим

$$H^\Omega(l,\varepsilon)=E\ln V^\Omega(\mathbf{x}^l,\varepsilon),$$

где математическое ожидание берется по всем повторным выборкам длины l. Эту величину мы будем называть ε-энтропией системы функций $F(x,\alpha)$ на выборках длины l.

Всегда существует предел

$$C^\Omega(\varepsilon)=\lim_{l\to\infty}\frac{H^\Omega(l,\varepsilon)}{l}.$$

Этот предел мы будем называть предельной ε-энтропией системы функций $F(x,\alpha)$ на символ.

Более того, величина $(\ln V^\Omega(\mathbf{x}^l,\varepsilon))/l$ сходится к $C^\Omega(\varepsilon)$ по вероятности.

Оказывается, что необходимым и достаточным для равномерной сходимости средних к математическим ожиданиям по классу функций $F(x,\alpha)$ будет условие

$$C^\Lambda(\varepsilon)=\ln\varepsilon.$$

То есть асимптотически объем $V^\Omega(\mathbf{x}^l,\varepsilon)$ должен вести себя так же, как объем одного кубика с ребром ε.

Более того, оказывается, что если это условие выполняется для некоторого значения $\varepsilon > 0$, то оно выполняется и при всех $\varepsilon > 0$.

Если для системы функций $F(x, \alpha)$ окажется, что $C^{\Omega}(\varepsilon) > \ln \varepsilon$, то она обладает следующим интересным свойством. Пусть $C^{\Omega}(\varepsilon) = \ln \varepsilon + \eta$ $(\eta > 0)$. Тогда существуют две функции $\psi_0(x)$ и $\psi_1(x)$, определенные в X, такие, что

1. $\psi_1(x) \geqslant \psi_0(x)$;
2. $\int (\psi_1(x) - \psi_0(x))\, dP \geqslant \varepsilon(e^{\eta} - 1)$
3. Для любого целого $n > 0$, почти для любой последовательности $x_1, x_2, \ldots, x_n$ $(x_i \in X)$, для любой бинарной последовательности $\omega_1, \omega_2, \ldots, \omega_n$ $(\omega_i = 0, 1)$ и любого $\delta > 0$ найдется значение $\alpha \in \Omega$ такое, что

$$\forall i \ |F(x_i, \alpha) - \psi_{\omega_i}(x)| < d \qquad (i = 1, \ldots, n).$$

(Порядок кванторов здесь существенен).

Иными словами, существует коридор ненулевой толщины, и почти для любой последовательности $x_1, x_2, \ldots, x_n$ можно потребовать, чтобы в одних точках выборки функция прошла сколь угодно близко от верхней грани коридора, а в остальных точках выборки она была сколь угодно близка к нижней грани коридора (выбор верхних и нижних точек произволен). И найдется такое значение $\alpha \in \Omega$, что это требование будет выполнено.

Отсюда достаточно ясно, что в этом случае равномерной сходимости нет. Действительно можно потребовать, чтобы в точках первой полувыборки функция была близка к верхней грани, а в точках второй полувыборки — к нижней. Тогда при неограниченном увеличении длины полной выборки среднее по первой полувыборке будет сходиться к $E\psi_1(x)$, а по второй — к $E\psi_0(x)$, а эти математические ожидания различны.

Проблема выбора оптимальной сложности модели

На протяжении всех лекций, прочитанных в этом семестре, мы постоянно убеждались, что надежный результат обучения можно получить

только при определенном соотношении длины обучающей выборки и объема класса моделей, из которого осуществлялся выбор. Мерой этого объема служило число настраиваемых параметров для методов максимального правдоподобия и восстановления регрессии, комбинаторная размерность или отношение расстояния между классами к их диаметру в задачах распознавания образов, или, наконец, энтропия системы функций относительно выборки. На практике количество данных, предъявляемых для обучения, всегда ограничено. Следовательно, естественно желание выбирать из наиболее простого класса моделей, чтобы обеспечить максимальную надежность. Но при этом в узком классе может не оказаться истинной модели или достаточно близкой к ней. Тогда приходится загрублять модель — идти на большее значение остаточной невязки или большее количество ошибок на данных обучения.

Поэтому возникает проблема выбора оптимальной сложности модели — поиск компромисса между ошибками, вызванными недостатком данных для обучения, и ошибками, связанными с грубостью модели.

Рассмотрению методов решения этой проблемы будет посвящена серия лекций следующего семестра. Здесь важно отметить два момента. Во-первых, встает вопрос, как упорядочить модели по их сложности. Ведь для оптимизации обучения важна не сложность в буквальном смысле слова, а объем класса, из которого мы выбираем. Например, данные могут подвергаться сложному преобразованию, зафиксированному до начала обучения, и требуется лишь настроить небольшое число параметров модели. Так, человек умеет узнавать лицо после предъявления небольшого числа фотографий. Во-вторых, стоит проблема собственно выбора оптимальной сложности.

Байесов подход в принципе дает ответ на оба эти вопроса: задание априорного распределения позволяет упорядочить модели, а усреднение по апостериорному распределению приводит к необходимому огрублению модели. Мы расскажем, как на основании байесовой процедуры можно выбрать оптимальное значение регуляризатора в задачах регрессии и при решении обратных задач, к каким алгоритмам она приводит при восстановлении многомерных полей. Но главными недостатками этого подхода остаются произвольность выбора априорного распределения и большие вычислительные трудности, за исключением сравнительно простых случаев.

Интересные результаты получаются при попытке скрестить метод максимального правдоподобия и байесов подход. «Априорное» распре-

деление при этом задается с точностью до небольшого числа параметров. Последние оцениваются методом максимального правдоподобия, а далее с этими параметрами реализуется байесова процедура.

Другой путь основан на прямом поиске оптимальной сложности модели. Простейший вариант здесь состоит в том, что резервируется часть данных для проверки. Сложность модели последовательно увеличивается, а результат проверяется на той части данных, которая не участвовала в построении модели. Вместо простого экзамена здесь могут использоваться такие процедуры, как cross validation или скользящий контроль, которые дают несмещенную оценку результата.

Наконец, можно оценить истинный риск, добавляя к значению эмпирического риска оценку того, насколько эмпирический риск, в точке своего минимума, занижен по сравнению с истинным.

Но обо всем этом мы расскажем уже в следующем семестре.

Лекция 13

Выбор модели. Байесов подход к проблеме. Общая постановка задачи. Формула Байеса. Байесова стратегия в теории игр. (Простейшие байесовы процедуры).

Лекции этого семестра будут посвящены выбору оптимальной сложности модели (а шире — оптимальной структуры модели). В прошлом семестре мы убедились, что все методы настройки моделей, основанные на принципе минимизации эмпирического риска, (метод максимума правдоподобия, метод наименьших квадратов, методы поиска решающего правила, минимизирующего число ошибок) требуют определенного соотношения между сложностью модели и длиной обучающей выборки. Только при достаточно большом объеме данных обучения, зависящем от сложности модели, можно добиться приемлемого результата.

При этом сложность модели понимается не как число блоков или число команд в программе, а как некоторым образом определенный объем класса моделей, из которого осуществляется выбор. Это может быть число настраиваемых параметров, функция роста, информация по Фишеру или другая характеристика класса моделей, но всякий раз она связана с разнообразием конкретных реализаций модели в пределах класса, из которого осуществляется выбор.

Если же количество данных обучения заранее ограничено, то для получения приемлемого результата приходится ограничивать и сложность модели — огрублять модель и переходить к более тонкой структуре только по мере поступления новых данных на обучение.

Одним из самых старых методов на этом пути оказывается подход, основанный на принципе Байеса. Согласно этому принципу задаются априорными вероятностями всех моделей в пределах некоторого класса, далее по данным обучения вычисляются апостериорные вероятности этих моделей и, наконец, находится такое решение, которое в среднем обеспечивает минимальный риск. Это среднее вычисляется на основе полученных апостериорных вероятностей.

Представим себе, что все модели достаточно широкого класса объединены в систему вложенных классов, причем каждый следующий класс содержит много больше конкретных моделей, чем предыдущий, т. е. имеет большую сложность. Допустим, что априорные вероятности того, что модель принадлежит одному из этих классов, приблизительно равны, а в пределах класса модели равновероятны. Тогда индивидуальная априорная вероятность каждой модели резко падает с увеличением ее сложности. Для того, того чтобы апостериорная вероятность определенной модели (или группы близких моделей) выделилась на фоне остальных, нужно, чтобы по данным обучения были отвергнуты все более простые модели и все далекие модели из данного класса. Кроме того, усреднение по апостериорным вероятностям также приводит к огрублению модели. Как это происходит в конкретных задачах, мы увидим в дальнейших лекциях.

С другой стороны, байесов подход подвергается критике с двух позиций. Во-первых, требуется точное задание априорной вероятности, которую неизвестно откуда взять. Во-вторых, в сложных задачах байесова процедура приводит к непреодолимым трудностям интегрирования в многомерных пространствах. Поэтому бурно развиваются альтернативные методы выбора оптимальной сложности и структуры модели, которые мы тоже рассмотрим в дальнейшем. Но многие приемы, такие как регуляризация и др., наследуются ими из байесовых процедур.

Формула Байеса

Напомним, что такое формула Байеса. Пусть $\mathbf{A}$ и $\mathbf{B}$ — два события, т. е. два множества элементарных событий, $\mathbf{AB} = \mathbf{A} \cap \mathbf{B}$ — их пересечение. Пусть далее $P(\mathbf{A})$ и $P(\mathbf{B})$ — вероятности событий $\mathbf{A}$ и $\mathbf{B}$ соответственно, $P(\mathbf{AB})$ — вероятность их одновременного выполнения, $P(\mathbf{A}|\mathbf{B})$ — вероятность выполнения $\mathbf{A}$ при условии $\mathbf{B}$, а $P(\mathbf{B}|\mathbf{A})$ — вероятность выполнения $\mathbf{B}$ при условии $\mathbf{A}$. Тогда по определению условной вероятности

$$P(\mathbf{AB}) = P(\mathbf{A}|\mathbf{B})P(\mathbf{B}) = P(\mathbf{B}|\mathbf{A})P(\mathbf{A}) \tag{1}$$

Откуда

$$P(\mathbf{B}|\mathbf{A}) = \frac{P(\mathbf{A}|\mathbf{B})P(\mathbf{B})}{P(\mathbf{A})}, \quad P(\mathbf{A}|\mathbf{B}) = \frac{P(\mathbf{B}|\mathbf{A})P(\mathbf{A})}{P(\mathbf{B})}. \tag{2}$$

Это и есть формула Байеса в простейшем случае.

Если же имеется целый ряд событий $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{A}_3, \ldots$ и $\mathbf{B}_1$, $\mathbf{B}_2$, $\mathbf{B}_3, \ldots$ с вероятностями $P(\mathbf{A}_i)$ и $P(\mathbf{B}_j)$ соответственно, то формула Байеса приобретает вид:

$$P(\mathbf{A}_i|\mathbf{B}_j) = \frac{P(\mathbf{B}_j|\mathbf{A}_i)P(\mathbf{A}_i)}{P(\mathbf{B}_j)} \tag{3}$$

Этот ряд событий может быть конечным или счетным.

Если же при этом события $\mathbf{A}_i$ не пересекаются и в целом покрывают все вероятностное пространство, то

$$\sum_i P(\mathbf{A}_i|\mathbf{B}_j) = 1.$$

Тогда значение $P(\mathbf{A}_i|\mathbf{B}_j)$ может быть вычислено и в том случае, когда мы не знаем значений $P(\mathbf{B}_j)$, как

$$P(\mathbf{A}_i|\mathbf{B}_j) = \frac{P(\mathbf{B}_j|\mathbf{A}_i)P(\mathbf{A}_i)}{\sum\limits_k P(\mathbf{B}_j|\mathbf{A}_k)P(\mathbf{A}_k)} \tag{4}$$

Обычно формула Байеса применяется, когда события $\mathbf{A}_i$ являются ненаблюдаемыми, а события $\mathbf{B}_j$ — наблюдаемы. Тогда получив наблюдение $\mathbf{B}_j$ и зная априорные вероятности $P(\mathbf{A}_i)$ и условные вероятности

$P(\mathbf{B}_j|\mathbf{A}_i)$ по этой формуле можно вычислить апостериорные вероятности событий $\mathbf{A}_i$ при условии наблюдения $\mathbf{B}_j$. Та же схема переносится на непрерывный случай. Пусть $\mathbf{X}$ и $\mathbf{Y}$ — два векторных пространства, $\mathbf{XY}$ — их прямое произведение. Пусть задана плотность распределения $p(x)$ в пространстве $\mathbf{X}$ и условная плотность распределения $p(y|x)$ в пространстве Y. Тогда по формуле Байеса условная плотность распределения

$$p(x|y) = \frac{p(x)p(y|x)}{\int p(z)p(y|z)\,dz}. \tag{5}$$

Обычно каждый элемент x из $\mathbf{X}$ трактуется как определенная гипотеза, а плотность $p(x)$ — как плотность априорного распределения на множестве гипотез. Элемент y из $\mathbf{Y}$ трактуется как наблюдение. Условная плотность $p(y|x)$ понимается как плотность распределения в пространстве $\mathbf{Y}$ при условии справедливости гипотезы x. Тогда условная плотность $p(x|y)$ выражает плотность апостериорного распределения на множестве гипотез.

Приведем типичный пример применения формулы (5). Пусть ищется неизвестная плотность распределения в пространстве $\mathbf{Z}$, про которую предполагается, что она принадлежит параметрическому классу $p(z,\alpha)$, где α — вектор параметров распределения. (Например, в одномерном случае это может быть плотность нормального распределения с неизвестными средним и дисперсией, а в многомерном пространстве — плотность нормального распределения с неизвестным вектором среднего и ковариационной матрицей). Плотность $p(\alpha)$ — априорная плотность распределения на множестве значений вектора параметров. Наблюдением является конечная последовательность векторов $y = z_1, z_2, \ldots, z_l$, полученная в серии независимых испытаний с неизменной плотностью $p(z,\alpha)$.

Тогда условная плотность распределения на множестве выборок длины l будет

$$p(y|\alpha) = \prod_{i=1}^{l} p(z_i,\alpha).$$

Полагая теперь в формуле (5) $x = \alpha$, получим апостериорную плот-

ность распределения вектора параметров α

$$p(\alpha|y) = \frac{p(\alpha)\prod\limits_{i=1}^{l} p(z_i,\alpha)}{\int p(\beta)\prod\limits_{i=1}^{l} p(z_i,\beta)\,d\beta}. \tag{6}$$

Это, конечно, не единственный случай применения формулы Байеса в форме (5). В дальнейшем мы увидим и много других применений.

Большой интерес для дальнейшего представляет случай, в котором смешиваются дискретный и непрерывный вариант формулы Байеса. Пусть имеется ряд (макро)гипотез Γ_1, Γ_2, Γ_3,... (конечный или счетный) и заданы априорные вероятности этих гипотез P_1, P_2, P_3,... Каждой такой гипотезе Γ_i соответствует определенная плотность распределения $p_i(x)$ на множестве $\mathbf{X}$ и условная плотность $p_i(y|x)$ на множестве $\mathbf{Y}$. (Вообще говоря, множества можно считать различными для разных гипотез Γ_i, но множество наблюдений $\mathbf{Y}$ должно быть одно для всех). Покажем, как вычисляется апостериорная вероятность гипотез Γ_i при заданном наблюдении y.

Условная плотность вероятности $p(y|\Gamma_i)$ получить наблюдение y при заданной гипотезе Γ_i равна:

$$p(y|\Gamma_i) = \int p_i(x)p_i(y|x)\,dx.$$

Поэтому по формуле Байеса получаем:

$$p(\Gamma_i|y) = \frac{P_i\int p_i(x)p_i(y|x)\,dx}{\sum\limits_j\left(P_j\int p_j(x)p_j(y|x)\,dx\right)}. \tag{7}$$

Эта формула в дальнейшем нам очень пригодится при выборе оптимальной сложности модели.

Связь с теорией игр и статистических решений

В теории игр используется термин «байесова стратегия», хотя он напрямую не связан с формулой Байеса. Пусть задана игра с нулевой

суммой, характеризуемая матрицей a_{ij}, где i — номер стратегии первого игрока, j — номер стратегии второго игрока, а a_{ij} — выигрыш второго игрока, если выбраны стратегии соответственно с номерами i и j. Тогда, если второму игроку известны вероятности P_i, с которыми первый игрок выбирает свою стратегию, то байесовой стратегией второго игрока называется такая его стратегия, которая доставляет максимум математического ожидания выигрыша:

$$\max_j \sum_i P_i a_{ij}. \tag{8}$$

Набор вероятностей P_i, с которыми первый игрок выбирает свои чистые стратегии, называется смешанной стратегией.

Связь с формулой Байеса появляется (в теории статистических решений), если эту схему усложнить следующим образом. Пусть теперь первый игрок после выбора своей стратегии передает второму игроку одно из сообщений (наблюдений) $s_1, s_2, s_3, \ldots$ с вероятностью $P(s_k) = P_{ik}$, где i — номер выбранной стратегии, а k — номер сообщения. Теперь, если второму игроку известны вероятности P_i и P_{ik} и само сообщение s_k, то он может вычислить по формуле Байеса апостериорные вероятности выбора стратегии первым игроком:

$$P_{i\,\text{апост}} = \frac{P_i P_{ik}}{\sum\limits_j P_j P_{jk}}. \tag{9}$$

Тогда второй игрок имеет возможность максимизировать условное математическое ожидание своего выигрыша при условии известного ему сообщения (наблюдения) s_k, подставив в формулу (9) вместо исходного распределения P_i апостериорные вероятности стратегий первого игрока:

$$\max_j \sum_j P_{i\,\text{апост}} a_{ij}. \tag{10}$$

Это и будет собственно байесовой стратегией.

Рассмотрим, в частности, случай, когда чистой стратегией первого игрока является выбор определенного вектора параметров α, характеризующего некоторую плотность распределения $p(z, \alpha)$, а его смешанная стратегия задается плотностью распределения $p(\alpha)$ на множестве

значений α. Чистой стратегией второго игрока будет вектор t из пространства Z, а функция штрафа $Q(\alpha, t)$ равна

$$Q(\alpha, t) = \int |t - z|^2 p(z, \alpha)\, dz.$$

Наблюдением является последовательность векторов z_1, z_2, …, z_l, полученная в серии независимых испытаний с неизменной плотностью $p(z, \alpha)$. Тогда апостериорная вероятность стратегий первого игрока $p_{\text{апост}}(\alpha)$ может быть найдена по формуле (6). Теперь оптимальной байесовой стратегией второго игрока будет такое значение t_0, которое доставляет минимум среднему риску:

$$t_0 = \arg\min \int Q(\alpha, t) p_{\text{апост}}(\alpha)\, d\alpha =$$
$$= \arg\min \int \left[\int |t - z|^2 p(z, \alpha)\, dz \right] p_{\text{апост}}(\alpha)\, d\alpha.$$

Нетрудно убедиться, что

$$\int |t - z|^2 p(z, \alpha)\, dz = |t - M(\alpha)|^2 + D(\alpha),$$

где $M(\alpha)$ и $D(\alpha)$ — соответственно математическое ожидание и дисперсия распределения $p(z, \alpha)$, причем второй член суммы от t не зависит. Поэтому достаточно найти минимум $\int |t - M(\alpha)|^2 p_{\text{апост}}(\alpha)\, d\alpha$. В свою очередь он достигается при

$$t_0 = \int M(\alpha) p_{\text{апост}}(\alpha)\, d\alpha,$$

то есть на апостериорном среднем математических ожиданий $M(\alpha)$.

Оценка математического ожидания нормального распределения

Одной из простейших задач, где может быть применена байесова стратегия, является оценка математического ожидания нормального распределения с известной дисперсией. Пусть случайная величина Z распределена согласно нормальному закону с плотностью

$$p(z) = \frac{1}{\sqrt{2\pi D}} \exp\left(-\frac{(z - M)^2}{2D} \right)$$

с неизвестным математическим ожиданием M и заданной дисперсией D. Пусть задано априорное распределение значений с плотностью $P_a(M)$. Если задана повторная выборка $\mathbf{z} = (z_1, \ldots, z_n)$ с этим распределением, то условная вероятность

$$P(\mathbf{z}|M) = \prod_{i=1}^{n} p(z_i) = \left(\frac{1}{\sqrt{2\pi D}}\right)^n \exp\left(-\frac{1}{2}D\sum_{i=1}^{n}(z_i - M)^2\right).$$

Обозначим средневыборочное значение $z_{\text{ср}} = (1/n)\sum_{i=1}^{n} z_i$, и положим $\Delta z_i = z_i - z_{\text{ср}}$. Заметим, что $\sum_{i=1}^{n} \Delta z_i = 0$. Тогда

$$\begin{aligned}\sum_{i=1}^{n}(z_i - M)^2 &= \sum_{i=1}^{n}(\Delta z_i + z_{\text{ср}} - M)^2 = \\ &= \sum_{i=1}^{n}\Delta z_i^2 + 2(z_{\text{ср}} - M)\sum_{i=1}^{n}\Delta z_i + n(z_{\text{ср}} - M)^2 = \\ &= \sum_{i=1}^{n}\Delta z_i^2 + n(z_{\text{ср}} - M)^2.\end{aligned}$$

$$P(\mathbf{z}|M) = \left(\frac{1}{\sqrt{2\pi D}}\right)^n \exp\left(-\frac{1}{2D}\sum_{i=1}^{n}\Delta z_i^2\right)\exp\left(-\frac{1}{2D/n}(z_{\text{ср}} - M)^2\right).$$

Заметим, что первые два сомножителя в этой формуле от величины M не зависят. Поэтому согласно формуле (6)

$$P_{\text{апост}}(M|\mathbf{z}) = \frac{1}{c}p_{\text{апр}}(M)\exp\left(-\frac{1}{2D/n}(z_{\text{ср}} - M)^2\right), \qquad (11)$$

где $c = \int p_{\text{апр}}(M)\exp\left(-\frac{1}{2D/n}(z_{\text{ср}} - M)^2\right)\,dM$ — нормировочный коэффициент.

По формуле (11) для произвольного априорного распределения можно вычислить апостериорное распределение величины M.

Рассмотрим более подробно тот случай, когда априорное распределение величины M также является нормальным и имеет плотность

$$p_{\text{апр}}(M) = \frac{1}{\sqrt{2\pi d}}\exp\left(-\frac{(M - M_0)^2}{2d}\right).$$

Здесь M_0 — среднее априорного распределения, а d — его дисперсия. Тогда

$$P_{\text{апост}}(M|\mathbf{z}) = \frac{1}{c_1}\exp\left[-\frac{1}{2}\left(\frac{(z_{\text{ср}}-M)^2}{D/n}+\frac{(M-M_0)^2}{d}\right)\right],$$

где c_1 — нормировочная константа.

Показатель в экспоненте представляет собой квадратичную функцию от M, и потому апостериорное распределение будет снова нормальным. Его дисперсия определяется коэффициентом при квадрате M в показателе, который равен $[1/(D/n)+1/d]$. Поэтому дисперсия апостериорного распределения равна

$$D_{\text{апост}} = \frac{1}{1/(D/n)+1/d} = \frac{Dd}{nd+D}.$$

При $n=0$ апостериорная дисперсия, естественно, совпадает с априорной, а при $n\to\infty$ она стремится к нулю.

Математическое ожидание $M_{\text{апост}}$ апостериорного распределения величины M определим как точку, в которой показатель экспоненты достигает максимума, приравняв к нулю производную от функции $(z_{\text{ср}}-M)^2/(D/n)+(M-M_0)^2/d$):

$$\frac{M-z_{\text{ср}}}{D/n}+\frac{M-M_0}{d}=0.$$

Отсюда

$$M_{\text{апост}} = \frac{ndz_{\text{ср}}+DM_0}{nd+D}.$$

Теперь видно, что апостериорное среднее с ростом n плавно перетекает от априорного среднего M_0 к средневыборочному значению $z_{\text{ср}}$. При $nd=D$ величина $M_{\text{апост}}$ попадет точно на середину между M_0 и $z_{\text{ср}}$.

Выбор уровня разбиения на кластеры

Рассмотрим следующую задачу, близкую по своему характеру к задачам Яндекса. Пусть задано разбиение пространства признаков на кластеры. Считаем, что это разбиение уже зафиксировано и задано правило, по которому каждый объект может быть отнесен к тому или иному кластеру. Это разбиение многоуровневое, в том смысле, что на каждом

следующем уровне кластеры предыдущего уровня разбиваются на более мелкие, а на каждом заданном уровне кластеры не пересекаются. Всего количество уровней равно N.

Пусть далее ставится задача распознавания двух классов и эта задача прямо не связана с разбиением на кластеры. Дается обучающая выборка $x_1, x_2, \ldots, x_l$ с указанием $y_1, y_2, \ldots, y_l$ принадлежности этих объектов к одному из двух классов. Эти объекты как-то распределятся по кластерам. Для каждого кластера подсчитывается число $k_1(i)$ объектов обучающей выборки, отнесенных к первому классу, и число $k_2(i)$ объектов, отнесенных ко второму классу, где i — номер кластера. Сумма $(k_1(i) + k_2(i))$ по каждому уровню фиксирована и равна l.

Далее предлагается такой путь нахождения решающего правила. Выбирается определенный уровень кластеризации, и на этом уровне все объекты определенного кластера относятся к одному классу, а именно к тому классу, представителей которого из обучающей выборки в этом кластере оказалось больше. Иными словами, этот кластер целиком относится к первому классу, если $k_1(i) > k_2(i)$, или целиком относится ко второму классу, если $k_1(i) < k_2(i)$. В случае равенства выбор осуществляется случайно.

Остается вопрос, на каком уровне остановиться. Чем ниже мы опустимся, тем на большее число кластеров разделится пространство признаков, тем более дифференцированным будет распознавание. С другой стороны, чем ниже уровень, тем больше кластеров, и, соответственно, тем меньше точек из обучающей выборки попадет в каждый отдельный кластер, и тем менее надежной будет его классификация. Попробуем получить критерий для выбора глубины уровня, основываясь на байесовой схеме. Рассмотрим N (макро)гипотез Γ_1, $\Gamma_2, \ldots$, Γ_N, каждая из которых состоит в следующем: все объекты, принадлежащие одному кластеру данного уровня, имеют равную априорную вероятность $p(i)$ принадлежать первому классу и, соответственно, $1-p(i)$ — принадлежать второму классу, где i — номер кластера данного уровня. Объекты, принадлежащие разным кластерам, могут иметь разные вероятности $p(i)$. Априорные вероятности этих гипотез положим равными P_1, $P_2, \ldots$, P_N .

Сами значения $p(i)$ нам неизвестны, и их конкретные значения образуют (микро)гипотезы. Примем поэтому, что эти значения априори могут принимать любые значения между нулем и единицей, априори равномерно распределены и независимы от кластера к кластеру. (За-

метим, что хотя гипотезы Γ_j пересекаются, но вероятностная мера их пересечения в наших предположениях равна нулю).

В этих условиях вероятность получить заданную в обучающей выборке классификацию y_1, y_2, …, y_l при заданной гипотезе Γ_j и фиксированных значениях $p(i)$ будет равна

$$\prod_{i=1}^{L}(p(i))^{k_1(i)}(1-p(i))^{k_2(i)},$$

где произведение берется по всем кластерам j-го уровня, а L — их число.

Тогда условная вероятность получить эту классификацию при условии макрогипотезы Γ_i будет равна

$$P(y|\Gamma_j) = \int\dots\int \prod_{i=1}^{L} p(i)^{k_1(i)}(1-p(i))^{k_2(i)}\,dp(1)\dots dp(L),$$

где интеграл берется по всему единичному кубу размерности L.

Произведение в нашем случае можно вынести за интеграл, поэтому получим

$$P(y|\Gamma_j) = \prod_{i=1}^{l}\int_0^1 p(i)^{k_1(i)}(1-p(i))^{k_2(i)}\,dp(i).$$

Но последний интеграл известен как бета-функция

$$B(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1}\,dt = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)},$$

где $\Gamma(z)$ — гамма-функция, для которой в свою очередь справедливо

$$\Gamma(n+1) = n!$$

Полагая теперь $p = k_1(i)+1$, $q = k_2(i)+1$, $t = p(i)$, получим

$$P(y|\Gamma_j) = \prod_{i=1}^{L} B(k_1(i)+1, k_2(i)+1) = \prod_{i=1}^{L}\frac{k_1(i)!k_2(i)!}{(k_1(i)+k_2(i))!}.$$

И теперь, согласно формуле (7), получаем

$$P(\Gamma_j|y) = \frac{P_j \prod_{i=1}^{L} \frac{k_1(i)!k_2(i)!}{(k_1(i)+k_2(i))!}}{\sum_t P_t \prod_{i=1}^{L} \frac{k_1(i)!k_2(i)!}{(k_1(i)+k_2(i))!}}.$$

Апостериорные вероятности гипотез Γ_j позволяют ранжировать их по степени правдоподобия.

Лекция 14

Регуляризация метода наименьших квадратов на основе байесова подхода. Асимптотика. Случай единичной матрицы. Обусловленность и псевдо-обратные матрицы. Общность единичной матрицы. Оптимальность для квадратичной штрафной функции (процедуры метода наименьших квадратов с регуляризацией).

Применим теперь байесов подход к задаче восстановления регрессионной зависимости по данным обучения. Ищется числовая зависимость, определенная на некотором пространстве X. Допустим, что эта зависимость имеет вид

$$\sum_{i=1}^{n} a_i \varphi_i(x),$$

где $\varphi_i(x)$ — известная система базовых функций, a_i — неизвестные коэффициенты. Для нас сейчас будут безразличны как природа пространства X, так и вид базовых функций $\varphi_i(x)$. Эта зависимость наблюдается (измеряется) с независимой аддитивной помехой ξ, имеющей нулевое среднее и фиксированную (известную) дисперсию D_ξ, распределенной по нормальному закону $N(0, D_\xi)$. Тогда наблюдаемая величина будет иметь вид

$$c(x) = \sum_{i=1}^{n} a_i \varphi_i(x) + \xi. \tag{1}$$

Ее математическое ожидание

$$M(x) = \sum_{i=1}^{n} a_i \varphi_i(x).$$

Это и есть искомая зависимость (регрессия).

Пусть теперь дана обучающая выборка длины l — в точках $x_1, x_2, \ldots, x_l$ измерены (наблюдались) значения $c(x_1)$, $c(x_2), \ldots$, $c(x_l)$. По этим наблюдениям (измерениям) и требуется оценить (восстановить) регрессию $M(x)$.

Теперь для того, чтобы применить байесов подход, необходимо задать априорное распределение на множестве коэффициентов a_i. Допустим, что это априорное распределение n-мерного вектора $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)$ также является нормальным с нулевым средним и ковариационной матрицей $\mathbf{R}$, т. е.

$$P_{\text{апр}}(\boldsymbol{a}) = N(0, \mathbf{R}).$$

Соответствующая плотность распределения будет равна

$$p_{\text{апр}}(\boldsymbol{a}) = \frac{1}{q_0} \exp\left(-\frac{1}{2}\boldsymbol{a}^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{a}\right), \tag{2}$$

где q_0 — нормировочная константа.

На самом деле то требование, что априори коэффициенты регрессии имеют нулевое среднее, не является принципиальным ограничением. Если известно $M_{\text{апр}}(\boldsymbol{a}) = \boldsymbol{a}_0 = (a_1^0, a_2^0, \ldots, a_n^0)$, то мы можем построить функцию

$$c^0(x) = \sum_{i=1}^{n} a_i^0 \varphi_i(x)$$

и вычесть ее из всех наблюдений. Тогда остатки будут иметь уже нулевое среднее. Получив оценку регрессии для остатков, прибавим к ней функцию $c^0(x)$ и получим оценку регрессии для исходной функции.

Обозначим теперь вектор значений, измеренных в точках x_1, $x_2, \ldots, x_l$, как

$$\mathbf{c} = (c(x_1), c(x_2), \ldots, c(x_l)),$$

а матрицу значений функций $\varphi_i(x)$ в этих точках — как $\mathbf{F}$ с элементами

$$f_{ij} = \varphi_i(x_j).$$

Найдем теперь условное распределение $P(\mathbf{c}|\boldsymbol{a})$ вектора $\mathbf{c}$ при заданных значениях вектора $\boldsymbol{a}$ и матрицы $\mathbf{F}$. Обозначим $\mathbf{c}_0$ вектор с координатами $c_{0j} = \sum_{i=1}^{n} a_i \varphi_i(x)$, где j меняется от 1 до l. Тогда

$$\mathbf{c}_0 = \mathbf{F}^{\mathrm{T}}\boldsymbol{a}.$$

Вектор представляет собой вектор условных математических ожиданий наблюдений $c(x_1), c(x_2), \ldots, c(x_l)$ в точках $x_1, x_2, \ldots x_l$ при условии, что вектор коэффициентов регрессии равен $\boldsymbol{a}$. В этих условиях отклонение фактически наблюдаемых значений от математических ожиданий объясняется только шумом, поэтому условное распределение величин $\Delta c(x_i) = (c(x_i) - c_{0j})$ при заданном значении вектора $\boldsymbol{a}$ будет таким же, как распределение шума ξ:

$$p(\Delta c(x_i)|a) = \frac{1}{\sqrt{2\pi D_\xi}} \exp\left(-\frac{(\Delta c(x_i))^2}{2D_\xi}\right).$$

Поскольку шум мы считаем независимым, то плотность совместного условного распределения величин $\Delta c(xi)$ будет равна:

$$\begin{aligned} p(\Delta\mathbf{c}|a) &= \prod_{j=1}^{l} p(\Delta c(x_i)|a) = \\ &= \left(\frac{1}{\sqrt{2\pi D_\xi}}\right)^l \exp\left(-\frac{1}{2D_\xi}\sum_{j=1}^{l}(c(x_i) - c_{0j})^2\right) = \\ &= \left(\frac{1}{\sqrt{2\pi D_\xi}}\right)^l \exp\left(-\frac{1}{2D_\xi}(\Delta\mathbf{c}^{\mathrm{T}}\Delta\mathbf{c})\right), \end{aligned}$$

где через $\Delta\mathbf{c} = \mathbf{c} - \mathbf{c}_0$ обозначен вектор с координатами $\Delta c(x_i)$. В свою очередь

$$\Delta\mathbf{c}^{\mathrm{T}}\Delta\mathbf{c} = (\mathbf{c} - \mathbf{c}_0)^{\mathrm{T}}(\mathbf{c} - \mathbf{c}_0) = \mathbf{c}^{\mathrm{T}}\mathbf{c} - 2\mathbf{c}_0^{\mathrm{T}}\mathbf{c} + \mathbf{c}_0^{\mathrm{T}}\mathbf{c}_0.$$

Далее, поскольку $\mathbf{c}_0 = \mathbf{F}^{\mathrm{T}}\boldsymbol{a}$, имеем

$$\mathbf{c}_0^{\mathrm{T}}\mathbf{c} = \boldsymbol{a}^{\mathrm{T}}\mathbf{F}\mathbf{c}, \quad \mathbf{c}_0^{\mathrm{T}}\mathbf{c}_0 = \boldsymbol{a}^{\mathrm{T}}\mathbf{F}\mathbf{F}^{\mathrm{T}}\boldsymbol{a}.$$

Поэтому

$$p(\Delta\mathbf{c}|a) = \left(\frac{1}{\sqrt{2\pi D_\xi}}\right)^l \exp\left(-\frac{1}{2D_\xi}\left(\mathbf{c}^{\mathrm{T}}\mathbf{c} - 2\boldsymbol{a}^{\mathrm{T}}\mathbf{F}\mathbf{c} + \boldsymbol{a}^{\mathrm{T}}\mathbf{F}\mathbf{F}^{\mathrm{T}}\boldsymbol{a}\right)\right).$$

Это и будет плотностью условного распределения вектора $\mathbf{c}$ при заданном векторе коэффициентов регрессии $\boldsymbol{a}$, если рассматривать послед-

нее выражение как функцию $\mathbf{c}$:

$$p(\mathbf{c}|\boldsymbol{a}) = \left(\frac{1}{\sqrt{2\pi D_\xi}}\right)^l \exp\left(-\frac{1}{2D_\xi}\left(\mathbf{c}^{\mathrm{T}}\mathbf{c} - 2\boldsymbol{a}^{\mathrm{T}}\mathbf{F}\mathbf{c} + \boldsymbol{a}^{\mathrm{T}}\mathbf{F}\mathbf{F}^{\mathrm{T}}\boldsymbol{a}\right)\right). \quad (3)$$

Теперь по формуле Байеса мы можем получить апостериорное распределение вектора коэффициентов при заданном векторе наблюдений $\mathbf{c}$:

$$p_{\text{апост}}(\boldsymbol{a}|\ c) = \frac{1}{q} p_{\text{апр}}(\boldsymbol{a}) p(\mathbf{c}|\boldsymbol{a}),$$

где q — нормировочная константа: $q = \int p_{\text{апр}}(\boldsymbol{a}) p(\mathbf{c}|\boldsymbol{a})\, d\boldsymbol{a}$.

В соответствии с формулами (2) и (3) имеем

$$\begin{aligned} p_{\text{апост}}(\boldsymbol{a}|\mathbf{c}) &= \frac{1}{q_1}\exp\left(-\frac{1}{2}\boldsymbol{a}^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{a}\right)\times \\ &\qquad \times \exp\left(-\frac{1}{2D_\xi}\left(\mathbf{c}^{\mathrm{T}}\mathbf{c} - 2\boldsymbol{a}^{\mathrm{T}}\mathbf{F}\mathbf{c} + \boldsymbol{a}^{\mathrm{T}}\mathbf{F}\mathbf{F}^{\mathrm{T}}\boldsymbol{a}\right)\right) = \\ &= \frac{1}{q_1}\exp\left(-\frac{1}{2D_\xi}(\mathbf{c}^{\mathrm{T}}\mathbf{c})\right)\times \\ &\qquad \times \exp\left(-\frac{1}{2D_\xi}\left(\boldsymbol{a}^{\mathrm{T}} D_\xi \mathbf{R}^{-1}\boldsymbol{a} - 2\boldsymbol{a}^{\mathrm{T}}\mathbf{F}\mathbf{c} + \boldsymbol{a}^{\mathrm{T}}\mathbf{F}\mathbf{F}^{\mathrm{T}}\boldsymbol{a}\right)\right), \end{aligned} \quad (4)$$

где q_1 — некоторая другая нормировочная константа.

В этом выражении только вторая экспонента зависит от $\boldsymbol{a}$, и показатель этой экспоненты представляет собой квадратичную функцию от $\boldsymbol{a}$:

$$-\frac{1}{2D_\xi}\left[-2\boldsymbol{a}^{\mathrm{T}}\mathbf{F}\mathbf{c} + \boldsymbol{a}^{\mathrm{T}}(\mathbf{F}\mathbf{F}^{\mathrm{T}} + D_\xi\mathbf{R}^{-1})\boldsymbol{a}\right]. \quad (5)$$

Поэтому апостериорное $p_{\text{апост}}(\boldsymbol{a}|\mathbf{c})$ распределение вектора a снова будет нормальным. Его математическое ожидание совпадает с точкой, в которой достигается экстремум этой квадратичной функции, а ковариационная матрица определяется матрицей соответствующей квадратичной формы.

Найдем теперь апостериорное математическое ожидание $M_{\text{апост}}(\boldsymbol{a})$ вектора коэффициентов $\boldsymbol{a}$. Для этого, как было сказано,

нам достаточно найти экстремум выражения (5), приравняв к нулю его градиент по $\boldsymbol{a}$:

$$\begin{aligned}\mathbf{grad}_a \left[-2\boldsymbol{a}^{\mathrm{T}}\mathbf{F}\mathbf{c} + \boldsymbol{a}^{\mathrm{T}}(\mathbf{F}\mathbf{F}^{\mathrm{T}} + D_\xi\mathbf{R}^{-1})\boldsymbol{a}\right] = \\ = 2\left[-\mathbf{F}\mathbf{c} + (\mathbf{F}\mathbf{F}^{\mathrm{T}} + D_\xi\mathbf{R}^{-1})\boldsymbol{a}\right] = 0,\end{aligned}$$

Откуда

$$M_{\text{апост}}(\boldsymbol{a}) = (\mathbf{F}\mathbf{F}^{\mathrm{T}} + D_\xi\mathbf{R}^{-1})^{-1}\mathbf{F}\mathbf{c}. \tag{6}$$

Ковариационная матрица апостериорного распределения вектора коэффициентов $\boldsymbol{a}$, согласно сказанному, будет равна

$$\mathbf{R}_{\text{апост}}(\boldsymbol{a}) = D_\xi(\mathbf{F}\mathbf{F}^{\mathrm{T}} + D_\xi\mathbf{R}^{-1})^{-1}. \tag{7}$$

Заметим, что при $\mathbf{R}^{-1} = 0$ (точнее, стремящемся к нулю), что соответствует бесконечной размытости априорного распределения, решение (6) в точности совпадает с оценкой регрессии методом наименьших квадратов. Добавка же члена $D_\xi\mathbf{R}^{-1}$ при $\mathbf{R}^{-1} \neq 0$ приводит к смещению результата относительно оценки метода наименьших квадратов (и, вообще, оценка оказывается смещенной). Эта добавочная матрица $D_\xi\mathbf{R}^{-1}$ называется регуляризующей матрицей. Ее роль мы увидим ниже.

Исследуем более подробно, как ведет себя решение (6) с ростом длины обучающей выборки l. Для этого разделим на l выражение внутри скобок и снаружи (это законно, так как выражение внутри скобок обращается в степень -1).

$$M_{\text{апост}}\,(\boldsymbol{a}) = \left(\frac{1}{l}(\mathbf{F}\mathbf{F}^{\mathrm{T}}) + \frac{1}{l}D_\xi\mathbf{R}^{-1}\right)^{-1}\frac{1}{l}\mathbf{F}\ c. \tag{8}$$

Матрица $\mathbf{K} = (1/l)\mathbf{F}\mathbf{F}^{\mathrm{T}}$, состоящая из элементов

$$k_{ij} = \frac{1}{l}\sum_s \varphi_i(x_s)\varphi_j(x_s),$$

представляет собой эмпирическую ковариационную матрицу переменных $\varphi_i(x)$ и $\varphi_j(x)$. Аналогично, вектор $(1/l)\mathbf{F}\mathbf{c}$ будет иметь координаты

$$r_i = \frac{1}{l}\sum_{j=1}^{l}\varphi_i(x_j)c_j,$$

то есть эмпирические коэффициенты ковариации наблюдаемой величины $\mathbf{c}$ и значений базисных функций $\varphi_i(x)$ в точках x_j.

Допустим, что на множестве X задано распределение вероятностей и существует математическое ожидание произведений $\varphi_i(x)$ и $\varphi_j(x)$ (истинные коэффициенты их корреляции) $k^*_{ij} = E\varphi_i(x)\varphi_j(x)$ и математическое ожидание $r^*_i = E\varphi_i(x)c$ (истинные коэффициенты корреляции выходной величины c и базисных функций $\varphi_i(x)$).

Тогда в силу закона больших чисел эмпирические коэффициенты корреляции k_{ij} и r_i стремятся при $l \to \infty$ к истинным значениям k^*_{ij} и r^*_i, а матрица $(1/l)\mathbf{F}\mathbf{F}^{\mathrm{T}}$ стремится к матрице $\mathbf{K}$ с элементами $k^*_{ij} = E\varphi_i(x)\varphi_j(x)$. В то же время регуляризующая матрица $(1/l)D_\xi\mathbf{R}^{-1}$ стремится к нулю с ростом l, и ее роль в решении становится пренебрежимой, если только матица $\mathbf{K}$ не вырождена. Таким образом, мы видим, что за исключением случая вырожденной матрицы $\mathbf{K}$, решение по Байесу стремится к оценке метода наименьших квадратов при неограниченном росте обучающей выборки.

Однако в общем случае не обязательно предполагать, что на множестве задано какое-то распределение вероятностей, или что оно не меняется по ходу получения выборки, или что точки x_i получаются независимо. Например, точки могут задаваться экспериментатором. При удачном расположении этих точек матрица $(1/l)\mathbf{F}\mathbf{F}^{\mathrm{T}}$ может стать большой по сравнению с регуляризующей матрицей, а при неудачном — матрица $(1/l)\mathbf{F}\mathbf{F}^{\mathrm{T}}$ может оставаться вырожденной или плохо обусловленной и при неограниченном росте выборки. Нахождению удачного расположения экспериментальных точек посвящена теория планирования эксперимента (которой мы касаться не будем). Однако для байесовой схемы неважно хорошо или плохо расположились экспериментальные точки, велика или мала длина обучающей выборки —во всех случаях метод дает точное значение условного математического ожидания коэффициентов регрессии.

Для того, чтобы более ясно понять роль регуляризации по Байесу, рассмотрим частный случай (к которому, как мы увидим, сводится и общий случай). Допустим, что априори коэффициенты a_i независимы и имеют одинаковую дисперсию D. Это значит, что априорная ковариационная матрица $\mathbf{R}$ имеет вид $\mathbf{R} = D\mathbf{E}$, где $\mathbf{E}$ — единичная матрица. Иными словами, все элементы матрицы $\mathbf{R}$ за исключением диагональных равны нулю, а диагональные равны D. Тогда $R^{-1} = (1/D)\mathbf{E}$. Тогда

формула (8) принимает вид:

$$M_{\text{апост}}(\boldsymbol{a}) = \left(\frac{1}{l}\mathbf{F}\mathbf{F}^{\mathrm{T}} + \frac{D_\xi}{lD}\mathbf{E}\right)^{-1}\frac{1}{l}\mathbf{F}\mathbf{c}. \tag{9}$$

Но матрица $(1/l)\mathbf{F}\mathbf{F}^{\mathrm{T}}$, будучи эмпирической матрицей коэффициентов ковариации, является симметрической положительно полуопределенной матрицей. Поэтому простым поворотом координат она может быть приведена к диагональному виду, то есть после поворота все ее элементы за исключением диагональных будут равны нулю, а на диагонали будет стоять собственные числа λ_i матрицы $(1/l)\mathbf{F}\mathbf{F}^{\mathrm{T}}$. При этом все собственные числа неотрицательны ($\lambda_i \geqslant 0$) и могут обращаться в нуль только в случае вырождения этой матрицы.

Повороту координат соответствует некоторое унитарное линейное преобразование системы базисных функций:

$$\varphi_i^*(x) = \sum_{j=1}^{n} b_{ij}\varphi_j(x),$$

где матрица с элементами b_{ij} сохраняет скалярное произведение. При этом элементы преобразованной матрицы $(1/l)\mathbf{F}\mathbf{F}^{\mathrm{T}}$ будут равны

$$\begin{aligned} k_{i,j} &= \frac{1}{l}\sum_s \varphi_i^*(x_s)\varphi_j^*(x_s) = 0 \qquad (i \neq j) \\ k_{i,i} &= \frac{1}{l}\sum_s \varphi_i^*(x_s)\varphi_i^*(x_s) = \lambda_i. \end{aligned} \tag{$*$}$$

Элементы вектора $(1/l)\mathbf{F}\mathbf{c}$ после преобразования координат будут равны

$$r_i = \frac{1}{l}\sum_{j=1}^{l} \varphi_i^*(x_j)c_j.$$

(Разумеется, при вычислениях это преобразование проводить не обязательно. Мы делаем его для того, чтобы увидеть, как параметр регуляризации (априорная дисперсия коэффициентов) влияет на решение). Поэтому матрица $(1/l)\mathbf{F}\mathbf{F}^{\mathrm{T}} + (D_\xi/(Dl))\mathbf{E}$ будет тоже диагональной, а ее диагональные элементы равны

$$\lambda_i + \frac{D_\xi}{Dl}.$$

Обратная матрица $((1/l)\mathbf{F}\mathbf{F}^{\mathrm{T}} + (D_\xi/(Dl))\mathbf{E})^{-1}$ также будет диагональной с диагональными элементами, равными

$$\frac{1}{\lambda i + \dfrac{D_\xi}{Dl}},$$

Соответственно, решение в преобразованных координатах принимает вид

$$M_{\text{апост}}(a_i) = \frac{r_i}{\lambda_i + \dfrac{D_\xi}{Dl}}. \tag{10}$$

Вырождение или плохая обусловленность матрицы $(1/l)\mathbf{F}\mathbf{F}^{\mathrm{T}}$ связана с тем, что собственные числа λ_i обращаются в нуль или становятся малыми. Теперь мы видим, что добавка $D_\xi/(lD)$ не дает собственным числам регуляризованной матрицы обращаться в нуль или становиться слишком малыми. Пока собственное число λ_i мало по сравнению с величиной $D_\xi/(Dl)$, соответствующий элемент обратной матрицы определяется преимущественно этой последней величиной. Когда же значение $D_\xi/(Dl)$ становится малым в сравнении с λ_i, соответствующий элемент обратной матрицы становится близким к $1/\lambda_i$. Когда все λ_i будут велики по сравнению с регуляризующей добавкой, решение приблизится к решению метода наименьших квадратов.

В то же время мы видим, что решение существенно зависит от того, какой дисперсией коэффициентов D мы задались априори. В зависимости от этого те или иные собственные числа будут подавлены регуляризатором.

Посмотрим, как меняется результат при изменении величины D от плюс бесконечности до нуля в соответствии с (10). При $D = +\infty$ для всех i

$$M_{\text{апост}}(a_i) = \frac{r_i}{\lambda_i},$$

что соответствует решению методом наименьших квадратов. При уменьшении D, как только $D_\xi/(Dl)$ становится соизмеримо или больше λ_i, соответствующее значение $M_{\text{апост}}(a_i)$ начинает уменьшаться и стремится к нулю, когда значение D стремится к нулю. Так одна за другой оценки $M_{\text{апост}}(a_i)$ начинают уменьшаться в порядке возрастания λ_i, то есть обращаются в нуль или становятся малыми оценки коэффициентов a_i сначала для малых λ_i, а затем для все больших значений λ_i. В

этом и состоит эффект регуляризации. При $D = 0$ все оценки $M_{\text{апост}}(a_i)$ обращаются в нуль, то есть апостериорное среднее становится равным априорному и данные наблюдения (измерения) игнорируются. Мы видим, что с уменьшением величины фактически уменьшается размерность пространства поиска — упрощается модель.

Интересно посмотреть, что происходит с оценкой коэффициента при тех функциях $\varphi_i^*(x)$, для которых оказалось, что $\lambda_i = 0$. Дело в том, что согласно (*)

$$\lambda_i = \frac{1}{l}\sum_{j=1}^{l}\varphi_i^*(x_j)\varphi_i^*(x_j) = \frac{1}{l}\sum_{j=1}^{l}\left(\varphi_i^*(x_j)\right)^2.$$

Поэтому λ_i может обратиться в нуль только в том случае, когда все значения $\varphi_i^*(x_j)$ при всех j равны нулю. Но тогда и значение в числителе выражения (10) будет равно нулю:

$$r_i = \frac{1}{l}\sum_{j=1}^{l}\varphi_i^*(x_j)c_j = 0.$$

Поэтому в соответствии с (10) $M_{\text{апост}}(a_i) = r_i/(\lambda_i + D_\xi/(Dl)) = 0$ при всех значениях D. Это значит что вектор $M_{\text{апост}}(\boldsymbol{a})$ всегда лежит в координатном подпространстве, оси которого соответствуют значениям λ_i, отличным от нуля.

На самом деле мы видим, что поведение каждого коэффициента a_i зависит от безразмерной величины $\lambda_i D/(D_\xi/l)$ — отношения $\lambda_i D$ априори ожидаемого вклада функции $\varphi_i^*(x)$ в выходную величину и дисперсии усредненного по выборке длины l шума D_ξ/l. Чем больше эта величина, тем больше мы доверяем экспериментальным данным, чем меньше — тем больше верим априорному среднему.

Насколько общим является рассмотренный случай, когда коэффициенты априори независимы и имеют одинаковую дисперсию? Оказывается, линейным преобразованием системы исходных базисных функций всегда можно привести задачу к этому случаю. Действительно, ковариационная матрица априорного распределения $\mathbf{R}$ заведомо является симметрической и положительно определенной. Поэтому поворотом координат (унитарным линейным преобразованием системы базисных функций) ее можно привести к диагональному виду, где по диагонали

будут стоять неотрицательные собственные числа t_i исходной матрицы $\mathbf{R}$. Те новые базисные функции, которым соответствуют нулевые значения t_i, можно просто удалить из базиса, так как коэффициенты при них будут заведомо равны нулю. Остальные же можно пронормировать, разделив на $\sqrt{t_i}$. Тогда ковариационная матрица априорного распределения в новом базисе станет просто единичной.

Посмотрим теперь, как может быть использована оценка $M_{\text{апост}}(\boldsymbol{a})$ для предсказания значения функции $c(x)$ в новых точках x. Оказывается, что для квадратичной функции штрафа наилучшей функцией предсказания $t(x)$ при всех x будет

$$t(x) = \sum_{i=1}^{n} M_{\text{апост}}(a_i)\varphi_i(x).$$

Действительно, в этом случае при любом x средний риск равен математическому ожиданию функции штрафа:

$$R = E(c(x) - t(x))^2,$$

где математическое ожидание определяется усреднением по апостериорному распределению коэффициентов регрессии и распределению шума. Но, как было установлено, минимум среднего риска в этом случае достигается при $t(x)$, равном математическому ожиданию функции $c(x)$:

$$t(x) = E(c(x)) = E\left(\sum_{i=1}^{n} a_i\varphi_i(x) + \xi\right) = \sum_{i=1}^{n} M_{\text{апост}}(a_i)\varphi_i(x),$$

что и требуется.

Обозначив через $\mathbf{f}$ вектор с координатами $f_i = \varphi_i(x)$, получим в матричной форме

$$t(x) = \mathbf{f}^{\mathrm{T}}(\mathbf{F}\mathbf{F}^{\mathrm{T}} + D_\xi\mathbf{R}^{-1})^{-1}\mathbf{F}\mathbf{c}.$$

Если коэффициенты априори независимы и имеют единичную дисперсию ($\mathbf{R} = \mathbf{E}$), то имеем

$$t(x) = \mathbf{f}^{\mathrm{T}}(\mathbf{F}\mathbf{F}^{\mathrm{T}} + D_\xi\mathbf{E})^{-1}\mathbf{F}\mathbf{c}.$$

Если же функция штрафа отлична от квадратичной, то, вообще говоря, этот результат неверен. В этом случае нужно учитывать не только апостериорное математическое ожидание коэффициентов, но и все их апостериорное распределение.

Итак, на рассмотренном примере мы хорошо видим как недостатки, так и достоинства байесова подхода. С одной стороны, результат существенно зависит от априорных значений дисперсии коэффициентов (в общем случае — от априорной ковариационной матрицы), которые неизвестно откуда взять, и дисперсии шума, которая иногда бывает известна. Кроме того, нам удалось избежать проблем интегрирования в многомерном пространстве благодаря предположению о нормальности априорного распределения коэффициентов. В этом случае апостериорное среднее совпадает с модой распределения — максимумом его плотности. В общем же случае для нахождения апостериорного среднего пришлось бы брать интеграл по пространству всех возможных значений коэффициентов a_i размерности n. При больших значениях n это в общем случае приводит к очень существенным вычислительным трудностям. С другой стороны, мы видим, как регуляризация по Байесу позволяет справиться с проблемами плохой обусловленности или вырожденности эмпирической ковариационной матрицы, которые возникают при простом применении метода наименьших квадратов.

Лекция 15

Обратные задачи и их решение с использованием байесовой стратегии. Постановка задачи. Примеры. Природа некорректности. Решение. Обсуждение. Ограничение по норме.

Следующая задача, которую мы рассмотрим как пример применения байесова подхода — это решения обратных задач математической физики и статистики. Решение оказывается очень близким к тому, которое мы получили на предыдущей лекции применительно к задаче оценки регрессии, но роль регуляризации здесь видна еще более ярко.

Постановка задачи

Ищется оценка числовой функции $u(x)$, определенной в общем случае на произвольном множестве X. Но эта функция считается ненаблюдаемой. Вместо этого наблюдаемой оказывается другая функция $v(t)$,

связанная с искомой линейным оператором $\mathbf{A}$:

$$v(t) = \mathbf{A}[u(x)]. \tag{1}$$

Пространство T, в котором определена функция $v(t)$ может совпадать с исходным пространством X, но может быть и совсем иным.

Далее предполагается, что функция $v(t)$ наблюдается (измеряется) в конечном числе l точек $t_1, t_2, \ldots, t_l$ с независимой помехой ξ с нулевым средним

$$v_i = v(t_i) + \xi_i, \tag{2}$$

и по этим данным требуется оценить функцию $u(x)$.

Чаще всего линейный оператор $\mathbf{A}$ задается интегральным оператором

$$v(t) = \int K(x,t)u(x)\,dx, \tag{3}$$

где интеграл берется по области определения функции $u(x)$ в пространстве X, а ядро $K(x,t)$ считается известным. Но в принципе это не обязательно так, и оператор $\mathbf{A}$ можно считать произвольным линейным оператором. Приведем несколько примеров, где возникает такая задача.

- **Размытое изображение**

  Пусть исходное изображение задается функцией $R(x,y)$, выражающей яркость изображения в точке (x,y). Изображение размывается, так что дельта-функция в точке (x,y) превращается в функцию $v(s,t) = K(x-s, y-t)$. Тогда все изображение превратится в размытое:

  $$V(s,t) = \iint K(x-s, y-t)R(x,y)\,dx\,dy.$$

  Наблюдается размытое изображение в дискретном (конечном) множестве точек. По этим данным требуется восстановить исходное изображение.

- **Дифференцирование.**

  Простое нахождение производной от экспериментально измеренной зависимости при наличии шума также относится к нашему

кругу задач. Действительно, пусть $u(x)$ и $v(t)$ — функции одной действительной переменной и

$$v(t) = \int_0^t u(x)\,dx.$$

Измерив значения функции $v(t)$ в точках $t_1, t_2, \ldots, t_l$ с помехой, мы хотим оценить ее производную $u(x)$. Оператор интегрирования — это линейный оператор. Его можно записать также в форме (3):

$$v(t) = \int_0^\infty I(x,t)u(x)\,dx,$$

где $I(x,t) = 1$ при $x \leqslant t$, и $I(x,t) = 0$ при $x > t$.

Близкими к этой, хотя и более сложными, оказываются многие задачи интерпретации физических экспериментов.

- **Анализ линейных динамических систем.**

  Пусть реакция линейной динамической системы на дельта- функцию, поданную на вход, будет $k(t)$, считая, что $k(t) = 0$ при $t < 0$. Тогда реакция $v(t)$ на произвольное входное воздействие $u(t)$ будет

$$v(t) = \int I(x,t)k(t-x)u(x)\,dx, \tag{4}$$

  где функция $I(x,t)$ определена так же, как и в предыдущем примере. Если принять, что переходная функция $k(t)$ нам известна, можно поставить задачу — по измерениям выходного сигнала $v(t)$ с помехой оценить значения входного воздействия $u(t)$.

  Но можно поставить задачу и иначе. Считая, что входное воздействие измерено точно (или задано), по измерениям выходного сигнала (с помехой) оценить переходную функцию $k(t)$. Выражение (4) в этом случае также задает линейный оператор, отображающий $k(t)$ в $v(t)$. В такой постановке задача оказывается актуальной при идентификации динамических систем в режиме on-line.

- **Восстановление трехмерной модели по проекциям.**

  В медицине (томография) и ряде технических приложений возникает такая задача. Имеется трехмерное тело, описываемое функцией $C(x, y, z)$, которая задает его плотность (например, оптическую) в каждой точке (x, y, z) некоторого объема. Непосредственно измерить эту плотность нельзя, но можно просветить это тело, например, рентгеновскими лучами, идущими в определенном направлении, и получить проекцию тела на плоскость, перпендикулярную потоку лучей. Имея ряд таких проекций можно попытаться восстановить трехмерную картину.

  Предположим, что тело просвечивается потоком параллельных горизонтальных лучей, идущих под углом φ к оси x. Обозначим через s горизонтальную ось проекции, а через t —ось, идущую в направлении лучей. Тогда проекция может быть описана функцией

$$R(s, z, \varphi) = \int C(s \sin\varphi + t \cos\varphi, s \cos\varphi - t \sin\varphi, z)\, dt.$$

  Заметим, что этот оператор будет линейным, хотя и не записан в форме (3). По измеренным значениям яркости $R(s, z, \varphi)$ проекции в конечном множестве точек (s, z) проекции и при конечном дискретном наборе углов φ требуется оценить функцию $C(x, y, z)$.

- **Социологическая задача.**

  Допустим, что известно число $N(i)$ пользователей интернета по населенным пунктам некоторого региона (i — номер населенного пункта) и известно распределение $M(i, j)$ населения этих пунктов по социальным группам ($M(i, j)$ — число лиц j-той группы в i-том населенном пункте). Нас интересует доля пользователей интернета $p(j)$ в каждой из групп. Если допустить, что доля пользователей зависит только от номера группы, но не от населенного пункта, получим соотношение

$$N(i) = \sum_j p(j) M(i, j).$$

  Это соотношение также задает линейный оператор, отображающий вектор с координатами $p(j)$ на вектор с координатами $N(i)$.

Помехой в данном случае служат неучтенные факторы — особенности каждого населенного пункта, отличные от распределения населения по социальным группам.

О природе некорректности

Обратные задачи обычно считаются некорректными или некорректно поставленными [18, 15]. Что это значит? Допустим, множество функций $u(x)$, на котором определен линейный оператор $\mathbf{A}$, представляет собой нормированное линейное (функциональное) пространство с нормой $|u(x)|$. Тогда множество функций $v(t) = \mathbf{A}[u(x)]$ также образует линейное пространство, на котором может быть определена, вообще говоря, другая норма $|v(t)|$. Рассмотрим величину

$$c = \inf_{|u(x)|=1} |\mathbf{A}[u(x)]|,$$

где inf берется по всем функциям $u(x)$, норма которых равна 1. Если величина c равна нулю, то это значит, что для всякого $\varepsilon > 0$ найдется функция $u(x)$ ($|u(x)| = 1$) такая, что норма $|A[u(x)]|$ будет меньше ε.

Обычно обратные задачи бывают именно такими, что величина c оказывается равной нулю. Допустим далее, что функция $v(t)$ наблюдается (измеряется) с помехой $s(t)$, то есть наблюдаемая функция есть

$$v^*(t) = v(t) + s(t).$$

Тогда формальное применение обратного оператора $\mathbf{A}^{-1}$ к функции $v^*(t)$ может быть некорректным в следующих двух смыслах.

Во-первых, помеха $s(t)$ может не принадлежать множеству значений оператора $\mathbf{A}$, и тогда оператор $\mathbf{A}^{-1}$ будет просто не определен на функции $v^*(t)$. Например, оператор $\mathbf{A}$ может быть сглаживающим, то есть, будучи определен на множестве непрерывных функций, имеет в качестве значений дифференцируемые функции. Тогда, если помеха $s(t)$ не дифференцируема, то обратный оператор на ней не определен, и, значит, не определен и на функции $v^*(t)$.

Во-вторых, если даже помеха $s(t)$ заведомо принадлежит множеству значений оператора $\mathbf{A}$, найдется сколь угодно малая по норме функция s(t), такая что функция $\mathbf{A}^{-1}s(t)$ будет сколь угодно большой по своей норме. Тогда формальное решение

$$u^*(x) = \mathbf{A}^{-1}[v(t) + s(t)] = u(x) + \mathbf{A}^{-1}s(t)$$

может оказаться сколь угодно далеким (по норме) от истинного решения $u(x)$.

Чтобы более ясно понять этот эффект, рассмотрим частный случай. Пусть самосопряженный оператор $\mathbf{A}$ задан в гильбертовом пространстве, $\varphi_i(x)$ — ортогональная нормированная система его собственных функций, λ_i — соответствующие собственные числа. Допустим, что эти числа отличны от нуля, упорядочены в порядке убывания по модулю и $\sum \lambda_i^2 < \infty$. В частности, отсюда следует, что $|\lambda_i| \to 0$ при $i \to \infty$.

Пусть функция $u(x)$ имеет разложение в этом базисе:

$$u(x) = \sum a_i \varphi_i(x).$$

Норма $|u(x)| = \sqrt{\sum a_i^2} < \infty$. Тогда

$$v(t) = \mathbf{A}[u(x)] = \sum \lambda_i a_i \varphi_i(x).$$

Обратный оператор $\mathbf{A}^{-1}$ будет иметь те же собственные функции и собственные числа $1/\lambda_i$.

Пусть теперь помеха $s(x)$ также принадлежит этому пространству и имеет в нашем базисе разложение

$$s(x) = \sum b_i \varphi_i(x). \qquad (*)$$

Тогда формально

$$\mathbf{A}^{-1} s(x) = \sum \frac{b_i}{\lambda_i} \varphi_i(x).$$

Но ряд $\sum (b_i/\lambda_i)^2$ вполне может расходиться, и значит $\mathbf{A}^{-1}s(x)$ не принадлежит нашему Гильбертову пространству и оператор $\mathbf{A}^{-1}$ на самом деле не определен на функции $s(x)$. Но даже если ряд $\sum (b_i/\lambda_i)^2$ заведомо сходится, норма функции $\mathbf{A}^{-1}s(x)$ может быть сколь угодно большой при сколь угодно малой норме функции $s(x)$. Действительно, допустим, что ряд (*) состоит всего из одного члена, т. е. $s(x) = b_i \varphi_i(x)$ при некотором значении i. Тогда

$$|s(x)| = |b_i|, \quad |\mathbf{A}^{-1} s(x)| = \frac{|b_i|}{|\lambda_i|}.$$

Положив теперь $|b_i| = \sqrt{|\lambda_i|}$, получим при $i \to \infty$

$$|s(x)| \to 0, \quad \text{тогда как} \quad |\mathbf{A}^{-1}s(x)| \to \infty.$$

Таким образом, сколь угодно малая по норме помеха может приводить к сколь угодно большим отклонениям решения нашего операторного уравнения.

Решение методом Байеса

Для решения поставленной задачи в общем случае, следуя байсову подходу, нужно задать априорное распределение вероятностей на множестве функций $u(x)$, найти условное распределение вероятности вектора наблюдений $(v_1, v_2, \ldots, v_l)$ в точках $t_1, t_2, \ldots, t_l$ для каждой функции $u(x)$ при заданном распределении шума, и получить апостериорное распределение на множестве функций $u(x)$. После этого апостериорное среднее значение $u(x)$ задаст наилучшую оценку при квадратичной функции штрафа. Но работать в функциональном пространстве затруднительно.

Поэтому мы опять предположим, что искомая функция может быть представлена конечным разложением в ряд по заданным базисным функциям с неизвестными коэффициентами:

$$u(x) = \sum_{i=1}^{n} a_i \varphi_i(x),$$

и предположим, что коэффициенты этого ряда априори распределены нормально с нулевым средним и ковариационной матрицей $\mathbf{R}$. Независимая помеха ξ также распределена нормально с нулевым средним и известной дисперсией D_ξ. Число членов разложения можно выбрать произвольно большим, поскольку, как мы увидим, решение все равно будет принадлежать подпространству размерности не превышающей длину выборки. Однако при очень большом числе членов разложения с вычислительной точки зрения более эффективными оказываются другие методы, которые мы рассмотрим на следующих лекциях.

Тогда при заданном векторе коэффициентов $\boldsymbol{a} = (a_1, a_2, \ldots a_n)$ и, соответственно, функции $u(x)$ зависимость $v(t) = \mathbf{A}[u(x)]$ примет вид

$$v(t) = \sum_{i=1}^{n} a_i \mathbf{A}\varphi_i(x).$$

Обозначив далее $\psi_i(t) = \mathbf{A}\varphi_i(x)$, получим

$$v(t) = \sum_{i=1}^{n} a_i \psi_i(t). \tag{5}$$

Но функции $\psi_i(t)$ тоже фактически заданы заранее (до получения данных обучения), и таким образом мы свели задачу к рассмотренной на предыдущей лекции: найти апостериорное распределение вектора коэффициентов $\boldsymbol{a} = (a_1, a_2, \dots a_n)$ по наблюдениям при наличии помехи функции $v(t)$ задаваемой выражением (5) в точках $t_1, t_2, \dots, t_l$.

Разница состоит лишь в том, что вместо матрицы $\mathbf{F}$ с элементами $f_{ij} = \varphi_i(x_j)$ должна использоваться матрица $\mathbf{F}^*$ с элементами $f^*_{ij} = \psi_i(t_j)$. Обозначим (при заданном векторе коэффициентов $\boldsymbol{a}$) вектор $(v(t_1), v(t_2), \dots v(t_l))$ через $\mathbf{v}_0$, вектор измеренных значений $(v_1, \dots v_l)$ через $\mathbf{v}$, а их разность $\mathbf{v} - \mathbf{v}_0$ — через $\Delta\mathbf{v}$. Тогда, следуя тем же путем, что на прошлой лекции, получим

$$p(\Delta\mathbf{v}|\boldsymbol{a}) = \left(\frac{1}{\sqrt{2\pi D_\xi}}\right)^l \exp\left(-\frac{1}{2} D_\xi (\Delta\mathbf{v}^{\mathrm{T}} \Delta\mathbf{v})\right).$$

И далее

$$\begin{aligned} p(\mathbf{v}|\boldsymbol{a}) &= \left(\frac{1}{\sqrt{2\pi D_\xi}}\right)^l \exp\left(-\frac{1}{2} D_\xi (\Delta\mathbf{v}^{\mathrm{T}} \Delta\mathbf{v})\right), \\ p_{\text{апост}}(\boldsymbol{a}|\mathbf{v}) &= \frac{1}{q} p_{\text{апр}}(\boldsymbol{a}) p(\mathbf{v}|\boldsymbol{a}) = \\ &= \frac{1}{q_1} \exp\left(-\frac{1}{2}\left[(D_\xi^{-1}(\Delta\mathbf{v}^{\mathrm{T}} \Delta\mathbf{v}) + \boldsymbol{a}^{\mathrm{T}} \mathbf{R}^{-1} \boldsymbol{a})\right]\right), \end{aligned} \tag{6}$$

где q и q_1 — нормирующие константы.

Как и раньше, для того, чтобы найти апостериорное математическое ожидание вектора $\boldsymbol{a}$, достаточно найти минимум квадратичной (относительно вектора $\boldsymbol{a}$) функции $(D_\xi^{-1}(\Delta\mathbf{v}^T \Delta\mathbf{v}) + \boldsymbol{a}^{\mathrm{T}} \mathbf{R}^{-1} \boldsymbol{a})$ или, с тем же результатом, функции

$$(\Delta\mathbf{v}^{\mathrm{T}} \Delta\mathbf{v}) + D_\xi \boldsymbol{a}^{\mathrm{T}} \mathbf{R}^{-1} \boldsymbol{a}. \tag{7}$$

Проделав необходимые преобразования, получим

$$M_{\text{апост}}(\boldsymbol{a}) = \left(\mathbf{F}^* \mathbf{F}^{*\mathrm{T}} + D_\xi \mathbf{R}^{-1}\right)^{-1} F^*\, v.$$

Это и дает наилучшую по Байесу оценку вектора коэффициентов разложения $\boldsymbol{a}$ при квадратичном критерии штрафа.

Решение $M_{\text{апост}}(\boldsymbol{a}) = (\mathbf{F}^*\mathbf{F}^{*\mathrm{T}} + D_\xi \mathbf{R}^{-1})^{-1}\mathbf{F}^*\mathbf{v}$ всегда лежит в некотором подпространстве пространства всех возможных значений вектора $\boldsymbol{a}$. Размерность этого подпространства не превосходит длину выборки, а его положение определяется матрицами $\mathbf{F}^*$ и $\mathbf{R}$ и не зависит от вектора измеренных значений $\mathbf{v}$. Более того, как мы видели на прошлой лекции, регуляризатор $D_\xi \mathbf{R}^{-1}$ подавляет степени свободы соответствующие слишком малым значениям собственных чисел матрицы $\mathbf{F}^*\mathbf{F}^{*\mathrm{T}}$.

Заметим, что функция (7) представляет собой сумму остаточной невязки и квадрата нормы вектора $\boldsymbol{a}$, взвешенного с коэффициентом D_ξ, если за норму принять величину $\sqrt{\boldsymbol{a}^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{a}}$.

Если же коэффициенты $\boldsymbol{a}$ априори независимы и имеют одинаковую дисперсию D, то матрица $\mathbf{R}$ будет иметь вид $D\mathbf{E}$. Тогда выражение (7) приобретает вид

$$(\Delta\mathbf{v}^{\mathrm{T}}\Delta\mathbf{v}) + \frac{D_\xi}{D}\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a},$$

то есть сумму остаточной невязки и взятого с весом (D_ξ/D) квадрата обычной евклидовой нормы вектора $\boldsymbol{a}$.

Если теперь устремить к нулю величину D_ξ, то роль добавки становится сколь угодно малой, и точка, доставляющая минимум функционалу (7), должна стремиться к точке, обеспечивающей минимум квадрата невязки, то есть к решению метода наименьших квадратов.

Допустим теперь, что система уравнений

$$\boldsymbol{a}^{\mathrm{T}}\mathbf{F}^* = \mathbf{v}$$

совместна (хотя может быть недоопределена). Тогда точное решение этой системы уравнений, очевидно, обеспечивает минимальную (нулевую) невязку. Но если система недоопределена (а это всегда так, когда число членов разложения n больше длины выборки l), то таких решений будет много: они образуют целое гиперпространство размерности $n - l$. Введение же в функционал взвешенного квадрата нормы позволяет выделить при $D_\xi \to 0$ решение с минимальной нормой $\sqrt{\boldsymbol{a}^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{a}}$, а оно уже будет единственным.

Сравнение с результатом из функционального анализа

В функциональном анализе для решения обратных некорректно поставленных задач предлагается следующий путь. Вместо решения операторного уравнения

$$v(t) = \mathbf{A}[u(x)],$$

где $u(x)$ — неизвестная функция, а функция $v(t)$ задана (но с возмущением), предлагается искать минимум по $u(x)$ функционала

$$L(u(x)) = |(v(t) + s(t)) - \mathbf{A}[u(x)]|^2 + k|u(x)|^2, \tag{8}$$

где $v(t) = \mathbf{A}[u_0(x)]$, $s(t)$ — возмущающая помеха, $|\cdot|$ — норма в функциональном пространстве, а k — некоторый положительный параметр регуляризации. Здесь $u_0(x)$ — точное решение операторного уравнения без помехи.

Для того чтобы понять, почему и в каком смысле этот путь приводит к успеху, рассмотрим частный случай. Пусть $u(x)$ принадлежит гильбертову пространству, линейный самосопряженный оператор $\mathbf{A}$ отображает это пространство в себя, и возмущение $s(x)$ также принадлежит этому пространству. Пусть далее $\varphi_i(x)$ — ортонормированный базис собственных функций самосопряженного оператора $\mathbf{A}$, а λ_i — соответствующие собственные числа, причем все λ_i отличны от нуля.

В разложении по этому базису наши функции примут вид

$$u_0(x) = \sum a_i^0 \varphi_i(x); \quad u(x) = \sum a_i \varphi_i(x); \quad s(x) = \sum b_i \varphi_i(x).$$

Обозначим уклонение функции $u(x)$ от истинного решения $u_0(x)$ $\Delta u(x)$:

$$\Delta u(x) = u(x) - u_0(x) = \sum (a_i - a_i^0)\varphi_i(x) = \sum \Delta a_i \varphi_i(x).$$

Квадраты норм нужных нам функций будут равны:

$$|u(x)|^2 = \sum a_i^2; \quad |s(x)|^2 = \sum b_i^2; \quad |\Delta u(x)|^2 = \sum (\Delta a_i)2.$$

В этих терминах функционал $L(u(x))$ запишется как

$$\begin{aligned} L(u(x)) &= |(v(x)+s(x)) - \mathbf{A}\left[u_0(x)+\Delta u(x)\right]|^2 + k|u(x)|^2 = \\ &= |(\mathbf{A}[u_0(x)]+s(x)) - \mathbf{A}[u_0(x)+\Delta u(x)]|^2 + k|u(x)|^2 = \\ &= |(s(x)) - \mathbf{A}[\Delta u(x)]|^2 + k|u(x)|^2 = \\ &= \sum (b_i - \lambda_i \Delta a_i)^2 + k\sum (a_i^0 + \Delta a_i)^2. \end{aligned}$$

Минимум этого выражения по значениям Δa_i будет достигнут при

$$\Delta a_i^* = \frac{\lambda_i b_i - k a_i^0}{\lambda_i^2 + k} = \frac{\lambda_i b_i}{\lambda_i^2 + k} - \frac{k a_i^0}{\lambda_i^2 + k}.$$

Из этого выражения ясно, что при значениях b_i и k, стремящихся к нулю, величина Δa_i^* тоже стремится к нулю, поскольку λ_i по условию отлично от нуля. Но этого недостаточно, чтобы норма уклонения $\Delta u(x)$ стремилась к нулю. Покажем, при каких условиях это будет так.

Разложим уклонение полученного минимизацией функционала (8) решения от истинного на две составляющие:

$$\Delta u(x) = \Delta u_1(x) - \Delta u_2(x) = \sum \frac{\lambda_i b_i}{\lambda_i^2 + k}\varphi_i(x) - \sum \frac{k a_i^0}{\lambda_i^2 + k}\varphi_i(x).$$

Первая из этих составляющих вызвана наличием помехи, вторая — фактом регуляризации. Далее,

$$|\Delta u(x)| \leqslant |\Delta u_1(x)| + |\Delta u_2(x)|.$$

Исследуем два этих слагаемых по отдельности:

$$|\Delta u_1(x)|^2 = \sum \left(\frac{\lambda_i b_i}{\lambda_i^2 + k}\right)^2 = \sum \frac{\lambda_i^2 b_i^2}{(\lambda_i^2 + k)^2}.$$

Но выражение $z^2 b^2/(z^2+k)^2$ достигает максимума по z при $z = \sqrt{k}$ и этот максимум равен $b^2/(4k)$. Поэтому

$$|\Delta u_1(x)|^2 \leqslant \sum \frac{b_i^2}{4k} = \frac{|s(x)|^2}{4k}.$$

Поэтому, если квадрат нормы возмущения $|s(x)|^2$ и параметр регуляризации k стремятся к нулю так, что и отношение $|s(x)|^2/k$ стремится к нулю, то и норма $|\Delta u_1(x)|$ стремится к нулю.

Что касается второго слагаемого $|\Delta u_2(x)|$, то для того, чтобы оно шло к нулю, достаточно, чтобы регуляризатор k стремился к нулю. Покажем это.

Заметим, что, как уже было сказано, для любого i величина $[ka_i^0/(\lambda_i^2 + k)]$ стремится к нулю при $k \to 0$, поскольку $\lambda_i^2 > 0$ при всех i. Кроме того, справедливо

$$\frac{ka_i^0}{\lambda_i^2 + k} < a_i^0.$$

Теперь для произвольного $\varepsilon > 0$ найдем такой номер n, что

$$0 < \sum_{i=n}^{\infty} \left(a_i^0\right)^2 < \varepsilon.$$

Далее

$$0 \leqslant |\Delta u_2(x)|^2 = \sum_{i=1}^{n-1} \left[\frac{ka_i^0}{\lambda_i^2 + k}\right]^2 + \sum_{i=1}^{\infty} \left[\frac{ka_i^0}{\lambda i2 + k}\right]^2 <$$
$$< \sum_{i=1}^{n-1} \left[\frac{ka_i^0}{\lambda_i^2 + k}\right]^2 + \sum_{i=n}^{\infty} (a_i^0)^2.$$

Но первая сумма стремится к нулю при $k \to 0$ как сума конечного числа слагаемых, каждое из которых стремится к нулю, а вторая сумма, будучи положительной, меньше ε. Ввиду произвольной малости числа ε, получаем, что $|\Delta u_2(x)| \to 0$ при $k \to 0$. Заметим, однако, что при фиксированном значении $k > 0$ норма $|\Delta u_2(x)|$ будет положительной, если только $|u(x)| > 0$.

Таким образом, если норма возмущения $|s(x)|$ и параметр регуляризации k стремятся к нулю так, что и отношение $|s(x)|^2/k$ стремится к нулю, то и норма уклонения $|\Delta u(x)|$ стремится к нулю.

Но на самом деле все измерения или наблюдения всегда происходят с помехой, норма которой не равна нулю. Поэтому при решении практических задач параметр ругуляризации k приходится оставлять положительным, и эта теория не дает ответа на вопрос, как его выбрать. И действительно, при слишком малых значениях k решение, как говорят, разваливается, и в большей мере отражает влияние помехи, чем входной функции $u(x)$. При положительном же значении

параметра k регуляризация вносит искажение, которое не устраняется даже, если норма помехи стремится к нулю.

Заметим также, что вместо добавления к остаточной невязке $|v(t) + s(t)) - \mathbf{A}[u(x)]|^2$ дополнительного члена, равного норме решения с некоторым весом, можно искать условный минимум остаточной невязки при ограничении

$$|u(x)|^2 \leqslant C,$$

где C — некоторая положительная константа. Тогда, применяя метод множителей Лагранжа, приходим к тому же функционалу. Просто вместо неизвестной константы регуляризации k, здесь имеем неизвестную константу C в ограничении.

Итак, мы видим, что как теория Байеса, так и подход, предлагаемый в функциональном анализе, приводят к одинаковому результату: следует минимизировать функционал, представляющий собой сумму остаточной невязки и взятого с некоторым весом квадрата нормы решения. Функциональный анализ дает лишь асимптотический результат, и не предлагает никакого критерия для выбора этого веса в случае конечной нормы помехи. Байесов подход дает оптимальное решение, но дорогой ценой. Помимо предположения о нормальности распределений и о гауссовой природе независимого шума нужно заранее (априори) знать дисперсию коэффициентов разложения и дисперсию шума (или их отношение). Это для случая априори независимых коэффициентов с одинаковой дисперсией. В общем случае, требуется гораздо больше априорных сведений.

Фактически отношение дисперсии шума к априорной дисперсии коэффициентов разложения играет ту же роль, что и параметр регуляризации в функциональном анализе, и значение этого отношения, в большинстве случаев, также непонятно откуда взять. Этот параметр на самом деле отражает ограничение сложности модели. При больших значениях параметра подавляются степени свободы соответствующие малым значениям собственных чисел оператора. По мере снижения величины параметра проявляются все более тонкие детали, но и возрастает чувствительность к помехе. К выбору наилучшего значения этого параметра мы вернемся в одной из следующих лекций.

Лекция 16

Метод кригинга. Сравнение с методом разложения по базисным функциям. (Стандартные процедуры кригинга).

В случае, когда для удовлетворительного представления искомой функции $u(x)$ требуется очень большое число базисных функций, более удобен другой путь, нежели тот, который был изложен в предыдущих лекциях. В частности, к таким задачам относится задача восстановления случайных полей по измерениям (наблюдениям) в конечном множестве точек.

Случайные поля

Случайным полем мы называем случайную функцию, определенную на некотором подмножестве n-мерного евклидова пространства, то есть функцию $u(\mathbf{x}, \omega)$, где $\mathbf{x}$ — вектор координат этого пространства, а $\omega \in \Omega$ — элементарное случайное событие в вероятностном пространстве Ω, на котором задана вероятностная мера $P(\omega)$. При фиксированном значении ω функция $u(\mathbf{x}, \omega)$ будет просто некоторой функцией вектора $\mathbf{x}$, называемой конкретной реализацией поля. Предполагается, что для любого набора точек $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l$ существует совместное распределение величин $u(\mathbf{x}_1), u(\mathbf{x}_2), \ldots, u(\mathbf{x}_l)$ и существуют первые и вторые моменты этих величин:

$$M(\mathbf{x}) = Eu(\mathbf{x}), \quad R(\mathbf{x}_1, \mathbf{x}_2) = Eu(\mathbf{x}_1)u(\mathbf{x}_2),$$

а также условное математическое ожидание значения поля в любой точке $\mathbf{x}$ при условии заданных значений в точках $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l$:

$$E(u(\mathbf{x})|u(\mathbf{x}_1), u(\mathbf{x}_2), \ldots, u(\mathbf{x}_l)).$$

Задача состоит в том, чтобы по заданным значения поля в точках $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l$ оценить значение поля $u(\mathbf{x})$ в произвольной текущей точке $\mathbf{x}$ (в пределах области определения поля [23]). Если критерием служит средний квадрат отклонения оценки от фактического значения поля, то наилучшей оценкой будет как раз условное математическое ожидание $E(u(\mathbf{x})|u(\mathbf{x}_1), u(\mathbf{x}_2), .., u(\mathbf{x}_l))$. Именно такую оценку мы будем искать, исходя из заданных характеристик поля.

Заметим, что наше определение случайного поля в одномерном случае ($n = 1$) совпадает с определением случайного процесса. Задача, в частности, может интерпретироваться как предсказание значения процесса $u(t)$ в какой-то момент t по заданным значениям в моменты $t_1, t_2, \ldots, t_l$. В двумерном случае это может быть изображение, рассматриваемое как случайное, или поле загрязнения некоторой поверхности, когда требуется оценить значения этого поля в достаточно густой сетке точек по заданным измерениям в небольшом числе точек. В трехмерном случае такая задача возникает при оценке поля концентрации полезных компонентов в некотором объеме земной коры по измерениям этой концентрации в пробах, полученных при разведке. При разведке нефтяных и газовых месторождений часто требуется восстановить функцию, выражающую проницаемость и пористость пород в объеме месторождения, по анализам отдельных проб. В общем случае — это может быть функция многих переменных различной природы.

Рассмотрим теперь случай гауссовых случайных полей. Это значит, что совместное распределение значений $u(\mathbf{x}_1), u(\mathbf{x}_2), .., u(\mathbf{x}_l)$ в произвольных точках $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l$ будет нормальным. Но нормальное распределение полностью характеризуется математическим ожиданием $M(\mathbf{x}) = Eu(\mathbf{x})$ и ковариационной функцией $R(\mathbf{x}_1, \mathbf{x}_2) = Eu(\mathbf{x}_1)u(\mathbf{x}_2)$. Математическое ожидание, не ограничивая общности, можно принять равным нулю. Действительно, в противном случае функция $M(\mathbf{x})$ считается известной, ее можно вычесть из всех экспериментальных значений $u(x_i)$, оценить остаточное поле, и, наконец, прибавить $M(\mathbf{x})$ к полученному результату.

Что же касается ковариационной функции, то в общем случае она будет функцией двух n-мерных векторов, и оценить ее, а тем более назначить ее априори, оказывается затруднительно. Поэтому чаще всего принимают гипотезу стационарности. Для случайного поля это значит, что вероятностная мера инвариантна относительно вектора сдвига. (В одномерном случае это определение соответствует определению стационарного случайного процесса).

В случае гауссова поля это значит, что математическое ожидание постоянно, а функция $R(\mathbf{x}, \mathbf{y})$ зависит только от вектора сдвига $(\mathbf{x}-\mathbf{y})$:

$$R(\mathbf{x}, \mathbf{y}) = R(\mathbf{x} - \mathbf{y}).$$

При этом она, как всякая ковариационная функция, будет симметрич-

ной: $R(\mathbf{x}-\mathbf{y}) = R(\mathbf{y}-\mathbf{x})$. Отсюда следует, что дисперсия процесса также будет постоянной:

$$D(\mathbf{x}) = R(\mathbf{x},\mathbf{x}) = R(\mathbf{x}-\mathbf{x}) = R(0) = \text{const}.$$

Достаточно часто в ковариационной функции выделяют регулярную составляющую $R_0(\mathbf{x},\mathbf{y})$, непрерывную при $\mathbf{x}=\mathbf{y}$, и некоррелированную составляющую $D_0\Delta(\mathbf{y}-\mathbf{x})$, где $\Delta(\mathbf{y}-\mathbf{x}) = 1$ при $\mathbf{x}=\mathbf{y}$ и $\Delta(\mathbf{y}-\mathbf{x}) = 0$ в остальных случаях:

$$R(\mathbf{x},\mathbf{y}) = R_0(\mathbf{x},\mathbf{y}) + D_0\Delta(\mathbf{y}-\mathbf{x}).$$

Некоррелированной составляющей может соответствовать просто погрешность измерения или же такие составляющие поля, корреляцией которых можно пренебречь на практически интересных расстояниях.

Решение (метод кригинга)

Найдем выражение для наилучшей оценки поля — условное математическое ожидание $E(u(\mathbf{x})|u(\mathbf{x}_1),u(\mathbf{x}_2),\ldots,u(\mathbf{x}_l))$ в текущей точке $\mathbf{x}$ при условии заданных значений $u(\mathbf{x}_1),u(\mathbf{x}_2),\ldots,u(\mathbf{x}_l)$ в точках $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_l$ для гауссова поля. Математическое ожидание $M(\mathbf{x})$ будем считать равным нулю, что не ограничивает общности, но значительно упрощает выкладки, а ковариационную функцию $R(\mathbf{x},\mathbf{y})$ будем считать заданной.

Для удобства обозначим текущую точку $\mathbf{x}$, в которой мы оцениваем значение поля, через $\mathbf{x}_0$, а значения поля в точках $\mathbf{x}_0,\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_l$ вектором u с координатами $u_i = u(\mathbf{x}_i)$, $(i=0,\ldots,l)$.

Тогда совместное распределение величин u_i будет нормальным с нулевым средним и ковариационной матрицей $\mathbf{R}^*$, элементы которой равны:

$$r_{ij} = R(\mathbf{x}_i,\mathbf{x}_j).$$

Плотность этого совместного распределения имеет вид

$$p(\mathbf{u}) = \frac{1}{c}\exp\left(-\frac{1}{2}\mathbf{u}\mathbf{R}^{*-1}\mathbf{u}\right), \tag{1}$$

где c — нормирующая константа.

Теперь для нахождения условной плотности $p(u_0|u_1,\ldots,u_l)$ распределения величины $u_0 = u(\mathbf{x}_0)$ при заданных значениях $u_i = u(\mathbf{x}_i)$

$(i = 1, \ldots, l)$ достаточно в формуле (1) зафиксировать все значения $u_1, \ldots, u_l$ и рассматривать выражение (1) как функцию только u_0, изменив нормировочную константу.

Сделаем это. Обозначим через b_{ij} элементы матрицы $\mathbf{R}^{*-1}$. Тогда

$$p(u_0|u_1, \ldots, u_l) = \frac{1}{c_1} \exp\left(-\frac{1}{2} \sum_{i,j=0}^{l} b_{ij} u_i, u_j\right) =$$

$$= \frac{1}{c_1} \exp\left(-\frac{1}{2}\left[\sum_{i,j=1}^{l} b_{ij} u_i u_j + 2\sum_{j=1}^{l} b_{0j} u_0 u_j + b_{00}(u_0)^2\right]\right). \quad (2)$$

Отсюда видно, что показатель в экспоненте последнего выражения является квадратичной функцией величины u_0, и значит, условное распределение $p(u_0|u_1, \ldots, u_l)$ снова будет нормальным, а его математическое ожидание равно такому значению u^* величины u_0, при котором квадратичная функция $2\sum_{j=1}^{l} b_{0j} u_0 u_j + b_{00}(u_0)^2$ достигает минимума. Приравняв к нулю ее производную по u_0, получим

$$\sum_{j=1}^{l} b_{0j} u_j + b_{00} u^* = 0,$$

откуда

$$E(u_0|u_1, \ldots, u_l) = u^* = -\sum_{j=1}^{l} \frac{b_{0j} uj}{b_{00}}.$$

Таким образом, условное математическое ожидание значения поля $u(\mathbf{x})$ в текущей точке $\mathbf{x}$ будет линейной функцией от измеренных значений $u(\mathbf{x}_1), \ldots, u(\mathbf{x}_l)$.

Но результат, выраженный через элементы b_{ij} обратной матрицы $(R^*)^{-1}$ не столь удобен и не столь очевиден. Поэтому учтем, что математическое ожидание остаточной невязки $E(u^* - u(\mathbf{x}))^2$ достигает минимума при $u^* = E(u_0|u_1, \ldots, u_l)$ и будем искать такую линейную функцию $u^* = \sum_{j=1}^{l} a_j u_j$, которая доставляет минимум остаточной невязке:

$$\min_{a_1,\ldots,a_l} E\left(\sum_{j=1}^{l} a_j u_j - u(\mathbf{x})\right)^2.$$

Далее,

$$E\left(\sum_{j=1}^{l} a_j u_j - u(\mathbf{x})\right)^2 = \sum_{i,j=1}^{l} a_i a_j E(u_i u_j) - \\ - 2\sum_{j=1}^{l} a_j E(u_j u(\mathbf{x})) + Eu(\mathbf{x})^2 = \\ = \sum_{i,j=1}^{l} a_i a_j R(\mathbf{x}_i, \mathbf{x}_j) + 2\sum_{j=1}^{l} a_j R(\mathbf{x}_j, \mathbf{x}) + R(\mathbf{x}, \mathbf{x})^2.$$

Запишем это соотношение в матричной форме, обозначив через $\boldsymbol{a}$ вектор $\boldsymbol{a} = (a_1, \ldots, a_l)$, через $\mathbf{u}$ вектор $\mathbf{u} = (u_1, \ldots, u_l)$, через $\mathbf{r}$ вектор $\mathbf{r} = (R(\mathbf{x}_1, x), \ldots, R(\mathbf{x}_l, \mathbf{x}))$ и через $\mathbf{R}$ матрицу размерности $l \times l$ с элементами $r_{ij} = R(\mathbf{x}_i, x_j)$:

$$E\left(\boldsymbol{a}^{\mathrm{T}}\mathbf{u} - u(\mathbf{x})\right)^2 = \boldsymbol{a}^{\mathrm{T}}\mathbf{R}\boldsymbol{a} - 2\boldsymbol{a}^{\mathrm{T}}\mathbf{r} + R(\mathbf{x}, \mathbf{x}). \tag{3}$$

Приравняв к нулю градиент по $\boldsymbol{a}$ от этой функции, получим систему уравнений

$$\boldsymbol{a}^{\mathrm{T}}\mathbf{R} - \mathbf{r} = 0, \tag{4}$$

и ее решение

$$\boldsymbol{a}_{\text{опт}} = \mathbf{r}^{\mathrm{T}}\mathbf{R}^{-1}. \tag{5}$$

Это решение и называют методом кригинга.

Подставляя это значение в выражение (3), получим формулу для условной дисперсии $u(\mathbf{x})$

$$E(u^* - u(\mathbf{x}))^2 = E(\boldsymbol{a}_{\text{опт}}^{\mathrm{T}}\mathbf{u} - u(\mathbf{x}))^2 = R(\mathbf{x}, \mathbf{x}) - \mathbf{r}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{r}.$$

Здесь $R(\mathbf{x}, \mathbf{x})$ равно безусловной дисперсии поля в точке $\mathbf{x}$, а член $\mathbf{r}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{r}$ выражает ее снижение благодаря информации, полученной из окрестных точек. Заметим, что как вектор коэффициентов $\boldsymbol{a}$, так и остаточная (условная) дисперсия зависят только от расположения точек измерения и положения текущей точки относительно них, но не от самих измеренных значений u_j.

Система уравнений (4) совершенно аналогична системе нормальных уравнений метода наименьших квадратов с той разницей, что в ней

используются не эмпирические коэффициенты ковариации, а «теоретические», полученные подстановкой координат в известную заранее ковариационную функцию $R(\mathbf{x}, \mathbf{y})$.

Окончательно, наилучшей оценкой поля в текущей точке x будет

$$E(u_0|u_1, \ldots, u_l) = \boldsymbol{a}_{\text{опт}}^{\mathrm{T}}\mathbf{u} = \mathbf{r}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{u} = \sum_{j=1}^{l} a_{j\,\text{опт}}\, u_j. \tag{6}$$

Поскольку в последнем рассуждении не использовалась гипотеза нормальности, то заключаем, что эта оценка будет наилучшей линейной оценкой поля в текущей точке $\mathbf{x}$ по измерениям в точках $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l$. А для гауссова поля она будет абсолютно лучшей (для квадратичного критерия).

Вспомним теперь, что обычно ковариационную функцию $R(\mathbf{x}, \mathbf{y})$ представляют в виде суммы непрерывной составляющей $R_0(\mathbf{x}, \mathbf{y})$ и некоррелированной составляющей $D_0\delta(\mathbf{y} - \mathbf{x})$. Тогда матрица $\mathbf{R}$ может быть представлена в виде

$$\mathbf{R} = \mathbf{R}_0 + D_0\mathbf{I},$$

где $\mathbf{R}_0$ — матрица с элементами $r_{ij} = R_0(\mathbf{x}_i, \mathbf{x}_j)$, а $\mathbf{I}$ — единичная матрица $(l \times\ l)$. Тогда решение (5) примет вид

$$\boldsymbol{a}_{\text{опт}} = \mathbf{r}^{\mathrm{T}}(\mathbf{R}_0 + D_0\mathbf{I})^{-1},$$

а наилучшее предсказание значения поля в точке $\mathbf{x}$ —

$$u^*(\mathbf{x}) = \mathbf{r}^{\mathrm{T}}(\mathbf{R}_0 + D_0\mathbf{I})^{-1}\mathbf{u}.$$

Добавка $D_0\mathbf{I}$ играет роль регуляризатора.

В теории кригинга ее называют эффектом самородков, так как первоначально в горных задачах некоррелированная составляющая ассоциировалась с попаданием самородков в пробы. На самом деле, эта составляющая может соответствовать погрешностям измерения или очень быстро меняющимся составляющим поля.

Можно показать, что если функция $R_0(\mathbf{x}, \mathbf{y})$ непрерывна (а тем более дифференцируема) при $\mathbf{x} = \mathbf{y}$, то наша оценка поля $E(u(\mathbf{x})|u_1, \ldots, u_l)$ в текущей точке $\mathbf{x}$ будет непрерывной функцией $\mathbf{x}$ за исключением, быть может, точек измерения $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l$, где она

скачком отклоняется от непрерывной составляющей, принимая фактически измеренные значения $u_1, \ldots, u_l$. Если же $D_0 = 0$, то она и в точках измерения останется непрерывной (а в случае дифференцируемости $R_0(\mathbf{x}, \mathbf{y})$ — и дифференцируемой). Это значит, что если погрешность измерения на самом деле есть, а мы примем $D_0 = 0$, то мы искусственно протянем гладкую оценку поля через все фактически измеренные точки, включая погрешность. Это, как правило, приводит к развалу решения. Если же мы правильно учтем некоррелированную составляющую, то гладкая составляющая решения пройдет на некотором расстоянии от фактически измеренных значений, то есть сгладит поле.

Обсуждение

Понятно, что наше решение в существенной мере зависит от того, насколько правильно мы знаем корреляционную функцию $R_0(\mathbf{x}, \mathbf{y})$. В нашей постановке задачи она считается заданной априори. Но при решении практических задач ее нужно откуда-то взять.

Во-первых, приходится делать предположение о стационарности поля (в нашем смысле), то есть что ковариационная функция зависит только от вектора сдвига (по крайней мере, в пределах достаточно больших областей):

$$R(\mathbf{x}, \mathbf{y}) = R(\mathbf{x} - \mathbf{y}).$$

В этом случае ковариационная функция будет зависеть только от одной переменной (правда, векторной), а безусловная дисперсия будет постоянной: $R(\mathbf{x}, \mathbf{x}) = D$. Поэтому вместо ковариационной функции можно пользоваться безразмерной корреляционной:

$$K(\mathbf{x} - \mathbf{y}) = \frac{1}{D} R(\mathbf{x} - \mathbf{y}).$$

Действительно,

$$a_{\text{опт}} = \mathbf{r}\mathbf{R}^{-1} = \boldsymbol{a}_{\text{опт}} = \mathbf{k}\mathbf{K}^{-1},$$

где $\mathbf{k}$ — вектор с координатами $(1/D)R(\mathbf{x} - \mathbf{x}_j) = K(\mathbf{x} - \mathbf{x}_j)$, а $\mathbf{K}$ — матрица с элементами $k_{ij} = (1/D)R(\mathbf{x}_i - \mathbf{x}_j) = K(\mathbf{x}_i - \mathbf{x}_j)$. Во-вторых, при невысокой размерности пространства ($n = 1, 2, 3$) корреляционную функцию можно пытаться оценить по тем же экспериментальным данным (об этом речь пойдет ниже). Можно и дальше упрощать модель,

считая, что корреляции зависят только от расстояния между точками, но не от направления сдвига. Тогда корреляционная функция будет зависеть от скалярного аргумента и ее легче оценить. Но результат будет более грубым.

В случае же высокой размерности ковариационной или корреляционной функцией приходится задаться априори. Но, как мы увидим, априорное задание ковариационной функции (в гауссовом случае) полностью эквивалентно априорному заданию распределения коэффициентов и выбору базисных функций при представлении поля $u(\mathbf{x})$ в виде разложения по базисным функциям $\varphi_i(\mathbf{x})$.

Метод кригинга впервые предложил для решения задач рудничной геологии горный инженер из ЮАР Дэниэл Криге (Крихе) в начале 50-х годов прошлого века. Позже математики дали обоснование метода и предложили многочисленные его модификации — логнормальный кригинг, индикаторный кригинг, дизъюнктивный кригинг. Но тот вариант, который изложен выше, предельно близок к исходному.

Лично я применил этот метод для планирования горных работ на крупнейшем в СССР золоторудном месторождении Мурун-Тау, разрабатываемом открытым способом (карьером). При отработке карьера на каждом открывшемся горизонте проводится эксплуатационная разведка — по регулярной сети (5×5 метров) бурятся короткие скважины на глубину уступа (10 метров). Порода, извлеченная из этих скважин, опробуется на содержание золота. Далее по этим данным следует оценить поле концентрации золота в опробованной части, и зоне влияния каждой скважины (прямоугольному параллелепипеду 5×5×10 метров) приписать среднее содержание золота в нем. Это содержание сравнивается с порогом (бортовым содержанием), и если среднее содержание превосходит бортовое, то порода из этой зоны влияния считается рудой и отправляется на переработку. В противном случае содержимое блока считается пустой породой и отправляется в отвал.

Традиционно за среднее содержание в зоне влияния скважины принималось содержание, измеренное в самой этой скважине. Но, во-первых, измерения проводятся с довольно большой погрешностью, а во-вторых, содержание в узкой (20 см) скважине совсем не обязательно отражает содержание во всей зоне ее влияния. Значит, нужно учесть и данные опробования соседних скважин.

Поэтому было предложено методом кригинга дать оценку содержаний золота в достаточно густой сетке точек, и содержание в зоне

влияния каждой скважины оценить, как среднее содержание в точках, попавших в зону.

Правда, распределение концентрации золота в породе гораздо ближе к логнормальному, чем к нормальному (т. е. распределение логарифмов содержаний близко к нормальному). Поэтому кригинг применялся к логарифмам содержаний, а результат экспонировался. Но при этом возникает регулярное смещение, так как среднее геометрическое положительных величин меньше среднего арифметического. Это требовало дополнительных поправок. Но самое трудное было оценить пространственную (трехмерную) корреляционную функцию, которая строилась каждый раз для нового участка. Дело облегчалось тем, что скважины были расположены по регулярной сети. Благодаря этому можно было вычислить эмпирические коэффициенты корреляции по ряду выделенных направлений, а затем аппроксимировать корреляционную функцию выражением

$$K(\mathbf{r}) = (1-\varepsilon)\exp(-\mathbf{r}\mathbf{A}\mathbf{r}) + \varepsilon\delta(\mathbf{r}),$$

где $\mathbf{r}$ — вектор сдвига, а $\mathbf{A}$ — матрица, характеризующая анизоторпию поля. Собственные векторы матрицы соответствуют осям эллипсоида рассеяния, а собственные числа обратно пропорциональны соответствующим радиусам корреляции. Эффект самородков (в логарифмической шкале) принимался постоянным для всего месторождения и оценивался путем сравнения анализов проб, полученных из одной и той же скважины.

Эксперименты на реальных данных показали, что от выбора корреляционной функции результат меняется существенно, и правильный выбор этой функции является принципиальным.

Сравнение с методом разложения по базисным функциям

На позапрошлой лекции мы рассмотрели представление функции $u(\mathbf{x})$ в виде разложения ее по базисным функциям $\varphi_i(\mathbf{x})$ со случайными коэффициентами a_i и случайной некоррелированной помехой ξ:

$$u(\mathbf{x}) = \sum_{i=1}^{n} a_i\varphi_i(x) + \xi.$$

Будем считать, что каждому выбору коэффициентов a_i и реализации помехи ξ соответствует определенная реализация случайного поля. Тогда задание распределения на множестве коэффициентов и распределения помехи полностью определяет задание вероятностной меры на множестве реализаций поля.

Тогда, если априорное распределение коэффициентов будет нормальным и независимая помеха также распределена нормально, то и соответствующее случайное поле тоже будет гауссовым. Действительно, для любого заданного набора точек $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l$ вектор $\mathbf{u} = u(\mathbf{x}_1), u(\mathbf{x}_2), \ldots, u(\mathbf{x}_l)$ имеет вид

$$\mathbf{u} = \mathbf{F}^{\mathrm{T}} \boldsymbol{a} + \xi,$$

где $\mathbf{F}$ — матрица с элементами $f_{ij} = \varphi_i(\mathbf{x}_j)$, а ξ — вектор значений помехи в точках $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l$. Поскольку матрица $\mathbf{F}$ не считается случайной, то отсюда следует, что, если векторы $\boldsymbol{a}$ и ξ распределены нормально, то и вектор u будет распределен нормально.

Если априорное математическое ожидание коэффициентов равно нулю и помеха идет с нулевым средним, то и математическое ожидание поля будет равно нулю. А какова будет его ковариационная функция $R(\mathbf{x}, \mathbf{y})$? Обозначим ковариационную матрицу априорного распределения коэффициетов $\mathbf{S}$, дисперсию помехи $D\xi$. Тогда получим

$$R(\mathbf{x}, \mathbf{y}) = E(u(\mathbf{x})u(\mathbf{y})) = E\left(\left(\sum_{i=1}^{n} a_i \varphi_i(\mathbf{x})\right)\left(\sum_{i=1}^{n} a_i \varphi_i(\mathbf{y})\right)\right) + \\ + D_\xi \delta(\mathbf{y} - \mathbf{x}) = \\ = \sum_{i,j=1}^{n} \varphi_i(\mathbf{x})\varphi_j(\mathbf{y}) E(a_i a_j) + D_\xi \delta(\mathbf{y} - \mathbf{x}), \tag{7}$$

Как уже отмечалось на позапрошлой лекции, не ограничивая общности можно считать, что коэффициенты a_i априори независимы и имеют единичную дисперсию. (Этого всегда можно добиться линейным преобразованием исходного базиса.) Тогда матрица $\mathbf{S}$ будет просто единичной: $\mathbf{S} = \mathbf{E}_{n \times n}$. Иными словами $E(a_i a_i) = 1$ и $E(a_i a_j) = 0$ при $i \neq j$. В этом случае формула (7) примет вид

$$R(\mathbf{x},\ y) = E(u(\mathbf{x})u(\mathbf{y})) = \sum_{i=1}^{n} \varphi_i(\mathbf{x})\varphi_i(\mathbf{y}) + D_\xi \delta(\mathbf{y} - \mathbf{x}). \tag{8}$$

Обозначим, как в лекции 14, через $\mathbf{F}$ матрицу с элементами $f_{ij} = \varphi_i(x_j)$, а через $\mathbf{f}$ вектор с координатами $f_i = \varphi_i(x_j)$. Тогда согласно (8) матрица $\mathbf{R}$ коэффициентов корреляции поля между точками измерения будет равна

$$\mathbf{R} = \mathbf{F}^{\mathrm{T}}\mathbf{F} + D_\xi \mathbf{I},$$

а вектор коэффициентов корреляции значений поля в текущей точке и точках измерения

$$\mathbf{r} = \mathbf{F}^{\mathrm{T}}\mathbf{f}.$$

Теперь согласно формуле (5) наша оценка поля в текущей точке будет

$$u^*(x) = \mathbf{r}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{u} = \mathbf{f}^{\mathrm{T}}\mathbf{F}(\mathbf{F}^{\mathrm{T}}\mathbf{F} + D_\xi \mathbf{I})^{-1}\mathbf{u}. \tag{9}$$

С другой стороны, согласно результату лекции 14 байесов метод дает оценку

$$u^*(x) = \mathbf{f}^{\mathrm{T}}(\mathbf{F}\mathbf{F}^{\mathrm{T}} + D_\xi \mathbf{E})^{-1}\mathbf{F}\mathbf{u}. \tag{10}$$

Формулы (9) и (10), соответствующие методу кригинга и методу Байеса, выглядят по-разному, но должны давать одинаковый результат. Ведь обе формулы дают оптимальную (для квадратичного критерия) оценку поля в текущей точке. Оказывается, можно показать, что и в самом деле справедливо матричное тождество

$$\mathbf{F}(\mathbf{F}^{\mathrm{T}}\mathbf{F} + D_\xi \mathbf{I})^{-1} \equiv (\mathbf{F}\mathbf{F}^{\mathrm{T}} + D_\xi \mathbf{E})^{-1}\mathbf{F}.$$

Итак, мы показали, что в гауссовом случае задание априорного распределения коэффициентов разложения функции по заранее фиксированной системе базисных функций влечет задание корреляционной функции поля, и задача может быть решена методом кригинга. Справедливо и обратное. В гауссовом случае каждой корреляционной функции поля $R(\mathbf{x}, \mathbf{y})$ соответствует нормальное распределение коэффициентов разложения по некоторой системе базисных функций (правда их число будет, вообще говоря, счетно бесконечной).

Действительно, корреляционная функция $R(\mathbf{x}, \mathbf{y})$, будучи симметричной и положительно определенной, может рассматриваться как ядро линейного, самосопряженного, положительно определенного оператора:

$$u(\mathbf{y}) = \int R(\mathbf{x}, \mathbf{y})u(\mathbf{x})\, d\mathbf{x}. \tag{11}$$

Согласно известным результатам функционального анализа, каждому такому оператору соответствует полная система ортогональных нормированных собственных функций $\varphi_i(x)$, таких что

$$\int R(\mathbf{x},\mathbf{y})\varphi_i(\mathbf{x})\,d\mathbf{x} = \lambda_i\varphi_i(\mathbf{y}),$$
$$\int \varphi_i^2(\mathbf{x})\,d\mathbf{x} = 1,$$
$$\int \varphi_i(\mathbf{x})\varphi_j(\mathbf{x})\,d\mathbf{x} \ = 0 \text{ при } i \neq j,$$

где $\lambda_i \geqslant 0$ — собственные числа этого оператора. Тогда произвольное поле $u(\mathbf{x})$ может быть разложено по этой системе собственных функций

$$u(\mathbf{x}) = \sum a_i\varphi_i(\mathbf{x}), \quad a_i = \int u(\mathbf{x})\varphi_i(\mathbf{x})\,d\mathbf{x}.$$

Поскольку само поле является гауссовым, то и коэффициенты разложения a_i будут распределены нормально, как в отдельности, так и по совокупности. Если поле имеет нулевое среднее, то и коэффициенты a_i будут иметь нулевое математическое ожидание.

Посмотрим, чему в этом случае равна дисперсия коэффициентов и коэффициент их ковариации.

$$E(a_ia_j) = E\left(\int u(\mathbf{x})\varphi_i(x)\,d\mathbf{x}\int u(\mathbf{y})\varphi_j(\mathbf{y})\,d\mathbf{y}\right) =$$
$$= \int \varphi_j(\mathbf{y})\left[\int E(u(\mathbf{x})u(\mathbf{y}))\varphi_i(\mathbf{x})\,d\mathbf{x}\right]\,d\mathbf{y} =$$
$$= \int \varphi_j(\mathbf{y})\left[\int R(\mathbf{x},\mathbf{y})\varphi_i(\mathbf{x})\,d\mathbf{x}\right]\,d\mathbf{y}.$$

И, поскольку $\varphi_j(\mathbf{x})$ есть собственная функция оператора (11), получаем

$$E(a_ia_j) = \lambda_i\int \varphi_i(\mathbf{y})\varphi_j(\mathbf{y})\,d\mathbf{y}.$$

Но интеграл $\int \varphi_i(\mathbf{y})\varphi_j(\mathbf{y})\,d\mathbf{y}$ равен нулю при $i \neq j$ и равен 1 при $i = j$. Поэтому

$$E(a_ia_j) = 0 \text{ при } i \neq j, \qquad E(a_ia_i) = D_i = \lambda_i \text{ при } i = j.$$

То есть в нашем случае, когда в качестве базисных функций взяты нормированные собственные функции оператора (11), коэффициенты разложения будут некоррелированы, а их дисперсия равна соответствующим собственным числам этого оператора. Разумеется, при этом нужно выделить некоррелированную составляющую поля, и рассматривать ее как помеху.

Итак, мы установили, что в гауссовом случае метод кригинга и байесов метод поиска наилучших коэффициентов разложения по базисным функциям формально полностью эквивалентны. Каким же из них пользоваться?

При восстановлении поля в многомерном пространстве, для удовлетворительного его описания обычно требуется использовать очень большое число базисных функций и обращать матрицы высокой размерности. Поэтому поиск оптимальных коэффициентов разложения оказывается вычислительно неэффективным. В то же время при оценке значения поля методом кригинга можно обращать не всю матрицу $l \times l$, а лишь ее часть, относящуюся к точкам измерения из некоторой окрестности текущей точки. Дело в том, что, во-первых, функция ковариации в правой части нормальных уравнений обычно быстро затухает на больших расстояниях от текущей точки, и веса суммирования от далеких точек становятся пренебрежимыми. Во-вторых, наблюдается так называемый эффект экранирования. Если текущая точка окружена достаточно большим числом близких точек измерения, то влияние более далеких точек становится очень малым. Эффект экранирования тем сильнее, чем меньше эффект самородков — при высоком соотношении сигнал/шум нет надобности в усреднении по большому числу экспериментальных точек.

С другой стороны, если заранее известно, что для удовлетворительного описания зависимости достаточно сравнительно небольшого числа базисных функций, то байесов метод поиска коэффициентов разложения оказывается предпочтительным.

Лекция 17

Гребневая регрессия. Критика байесова подхода. Регуляризация как приближенная реализация байесовой стратегии. Проблема выбора констант регуляризации и системы функций разложения.

Гребневая регрессия

В последних трех лекциях мы видели, как байесова стратегия приводит к идее введения регуляризации. Эта регуляризация состоит в том, что в критерий, подлежащий минимизации, наряду с членом, характеризующим остаточную невязку, вводится с некоторым весом дополнительный член, характеризующий норму решения.

Однако исторически идея регуляризации возникла без всякого отношения к методу Байеса при решении недоопределенных или плохо обусловленных систем уравнений. Такая ситуация возникала прежде всего при решении обратных задач математической физики.

Об этом речь уже шла. Напомню постановку задачи. Задан оператор $\mathbf{A}$, связывающий искомую функцию $v(y)$ с наблюдаемой функцией $u(x)$ уравнением:

$$\mathbf{A}v(y) = u(x). \tag{1}$$

Обычно оператор $\mathbf{A}$ выражается в форме

$$\mathbf{A}v(y) = \int K(x,y)v(y)\,dy,$$

а уравнение (1) принимает вид

$$\int K(x,y)v(y)\,dy = u(x). \tag{2}$$

Переменные x и y могут быть как скалярными, так и векторными.

Наблюдаемая функция $u(x)$ измеряется в точках $x_1, x_2, \ldots, x_l$ и при этом получаются значения $u_i = u(x_i)$. По этим данным требуется оценить функцию $v(y)$. Это можно попытаться сделать не путем разложения функции $v(y)$ в ряд по системе базисных функций, а непосредственно искать значения этой функции в достаточно густой сети точек $y_1, y_2, \ldots, y_n$. Обозначим искомые значения $v_j = v(y_j)$. Заменив интегрирование приближающим его суммированием, получим систему уравнений:

$$\sum_{j=1}^{n} K(x_i, y_j)v_j\Delta y_j = u_i \qquad (i = 1, \ldots l). \tag{3}$$

Здесь Δy_j — объем зоны влияния точки y_j. В одномерном случае это может быть просто длина интервала между соседними точками, а в многомерном — соответствующий объем векторного пространства. Переходя к матричной форме, введем следующие обозначения:

$\mathbf{u} = (u_1, u_2, \ldots, u_l)$, $\mathbf{v} = (v_1, v_2, \ldots, v_l)$ $\mathbf{K}$ — матрица с элементами $k_{ij} = K(x_i, y_j)\Delta y_j$. Тогда уравнение (3) запишется в форме

$$\mathbf{v}^{\mathrm{T}}\mathbf{K} = \mathbf{u}^{\mathrm{T}}. \tag{4}$$

Но это уравнение может не иметь решения или иметь много решений (целое гиперпространство), поскольку n, вообще говоря, не равно l, то есть матрица $\mathbf{K}$ может быть прямоугольной. Поэтому вместо решения уравнения (4) применяли метод наименьших квадратов. Для этого достаточно правую и левую части уравнения (4) умножить на K^{T}. Получим

$$\mathbf{v}^{\mathrm{T}}(\mathbf{K}\mathbf{K}^{\mathrm{T}}) = \mathbf{u}^{\mathrm{T}}\mathbf{K}^{\mathrm{T}}. \tag{5}$$

Теперь уже матрица $\mathbf{K}\mathbf{K}^{\mathrm{T}}$ будет заведомо квадратной, но она все рвано может быть вырожденной или слабо обусловленной. Для того, чтобы все же получить приемлемое решение (и притом единственное), было предложено к матрице $\mathbf{K}\mathbf{K}^{\mathrm{T}}$ добавить с некоторым весом ε единичную матрицу и решать уравнение

$$\mathbf{v}^{\mathrm{T}}(\mathbf{K}\mathbf{K}^{\mathrm{T}} + \varepsilon\mathbf{E}) = \mathbf{u}^{\mathrm{T}}\mathbf{K}^{\mathrm{T}}. \tag{6}$$

Тогда решением будет

$$\mathbf{v} = (\mathbf{K}\mathbf{K}^{\mathrm{T}} + \varepsilon\mathbf{E})^{-1}\mathbf{K}\mathbf{u}. \tag{7}$$

Это решение совершенно аналогично полученному в лекциях 14 и 15, но было предложено без всякой связи с байесовым подходом. Фактически оно получается, если минимизировать критерий, представляющий собой сумму остаточной невязки и взятой с весом ε нормы вектора $\mathbf{v}$. Действительно, рассмотрим критерий

$$R = (\mathbf{v}^{\mathrm{T}}\mathbf{K} - \mathbf{u})^2 + \varepsilon\mathbf{v}^{\mathrm{T}}\mathbf{E}\mathbf{v} = (\mathbf{v}^{\mathrm{T}}\mathbf{K}\mathbf{K}^{\mathrm{T}}\mathbf{v} - 2\mathbf{v}^{\mathrm{T}}\mathbf{K}\mathbf{u} + \mathbf{u}\mathbf{u}^{\mathrm{T}}) + \varepsilon\mathbf{v}^{\mathrm{T}}\mathbf{E}\mathbf{v}. \tag{8}$$

Найдем значение вектора $\mathbf{v}$, доставляющее минимум этому критерию, приравняв к нулю градиент критерия по $\mathbf{v}$:

$$\mathbf{v}^{\mathrm{T}}(\mathbf{K}\mathbf{K}^{\mathrm{T}} + \varepsilon\mathbf{E}) - \mathbf{K}\mathbf{u} = 0,$$

то есть получаем уравнение (7).

Этот прием — добавление единичной матрицы с некоторым весом к матрице $\mathbf{K}\mathbf{K}^{\mathrm{T}}$ — получил название *гребневая регрессия* (Ridge

regression). Название происходит оттого, что при этом добавляется как бы гребень на диагонали матрицы. Метод гребневой регрессии получил применение и при решении любых задач оценки регрессии, когда матрица коэффициентов ковариации оказывается плохо обусловленной. Как мы видели, добавление регуляризатора $\varepsilon\mathbf{E}$ приводит к подавлению тех составляющих регрессии, которые соответствуют малым собственным числам матрицы $\mathbf{K}\mathbf{K}^{\mathrm{T}}$. С увеличением константы регуляризации ε подавляется все больше таких составляющих. При этом решение становится все менее чувствительным к помехе, но все более огрубляется. Однако конкретный выбор значения параметра ε остается открытым, и мы обратимся к этому вопросу в следующих лекциях.

При решении обратных задач оказалось, что добавление квадрата нормы вектора $\mathbf{v}$ часто оказывается недостаточным, и нужно ограничить еще и производные. Тогда к критерию добавляют с некоторым весом еще и член, отражающий норму производной (или градиента в многомерном случае).

Аналогичным образом трансформируется и метод кригинга. Мы уже отмечали, что в многомерном случае ковариационную функцию $R(\mathbf{x}, \mathbf{y})$ на самом деле приходится считать заданной априори, и взять ее неоткуда. Поэтому просто задаются некоторым положительно определенным ядром $R(\mathbf{x}, \mathbf{y})$ (обычно в форме $R(\mathbf{x} - \mathbf{y})$) и ищут оценку значения поля в текущей точке в виде

$$u(\mathbf{x}) = \sum a_i R(\mathbf{x}, \mathbf{x}_i) u_i,$$

где a_i — неизвестные коэффициенты, $\mathbf{x}_i$ — точки, в которых поле было измерено, u_i — измеренные значения поля ($i = 1, \ldots, l$). Считая далее, что поле было измерено с помехой, дисперсия которой равна D_0, применяют формулу метода кригинга:

$$\boldsymbol{a}_{\text{опт}} = \mathbf{r}\mathrm{T}(\mathbf{R}_0 + D_0\ I)^{-1}, \quad u(x) = \mathbf{r}^{\mathrm{T}}(\mathbf{R}_0 + D_0\ I)^{-1}\mathbf{u},$$

где $\mathbf{r}$ — вектор с координатами $r_i = R(\mathbf{x}, \mathbf{x}_i)$, $\mathbf{R}_0$ — матрица с элементами $r_{ij} = R(\mathbf{x}_i, \mathbf{x}_j)$, а $\mathbf{u}$ — вектор измеренных значений с координатами u_i ($i = 1, \ldots, l$). Этот метод получил название ядерная гребневая регрессия (Kernel Ridge Regression).

Посмотрим, какой норме (в сравнении с методом обычной гребневой регрессии) соответствует это решение. Как было показано на

прошлой лекции, если в качестве базисных функций выбрать нормированные собственные функции оператора $\mathbf{A}$: $u(\mathbf{y}) = \int R(\mathbf{x},\mathbf{y})u(\mathbf{x})\,d\mathbf{x}$, где $R(\mathbf{x},\mathbf{y})$ — ковариационная функция, то этой функции соответствует априорное распределение коэффициентов разложения с диагональной корреляционной матрицей, по диагонали которой стоят собственные числа λ_i этого оператора. (Напомним, что $\lambda_i \geqslant 0$). В свою очередь, в лекции 13 мы показали, что регуляризующая добавка $D_\xi \boldsymbol{a}^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{a}$ представляет собой взятый с весом D_ξ квадрат нормы вектора весов $\boldsymbol{a}$, если за норму принять $\sqrt{\boldsymbol{a}^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{a}}$. Но в нашем случае матрица R^{-1} будет тоже диагональной с элементами на диагонали, равными $1/\lambda_i$. Таким образом, квадрат нормы весов будет равен $\sum(1/\lambda_i)a_i^2$. Рассмотрим оператор, описываемый в нашем базисе диагональной матрицей с элементами на диагонали, равными $1/\sqrt{\lambda_i}$. Обозначим этот оператор $\mathbf{B} = \mathbf{A}^{-1/2}$. Тогда

$$\boldsymbol{a}^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{a} = |\mathbf{B}\boldsymbol{a}|^2.$$

Итак, метод ядерной гребневой регрессии минимизирует критерий, равный сумме остаточной невязки и взятой с весом D_0 величины $|\mathbf{A}^{-1/2}u(\mathbf{x})|^2$, где $\mathbf{A}$ — оператор $u(\mathbf{y}) = \int R(\mathbf{x},\mathbf{y})u(\mathbf{x})\,d\mathbf{x}$. Когда величина D_0 стремится к нулю, решение стремится к такому, которое, проходя точно через все экспериментальные точки, доставляет минимум норме $|\mathbf{A}^{-1/2}u(\mathbf{x})|$.

Но помеха всегда присутствует, и обращать в нуль (или делать слишком малыми) константы регуляризации (ε или D_0) нельзя. В противном случае решение пыталось бы «объяснить» и помеху.

Итак, мы видим, что метод гребневой регрессии содержит произвол в двух отношениях. Во-первых, это выбор константы регуляризации, а во-вторых, это выбор конкретной нормы, в случае обычной гребневой регрессии, или конкретного ядра, в случае ядерной гребневой регрессии. Все теоретические утверждения, что решение сходится к истинному при помехе, стремящейся к нулю, или при неограниченном росте длины выборки, не снимают вопроса о том, как выбрать эти параметры, потому что помеха всегда есть, а выборка всегда конечна.

Заметим, что байесов подход и эквивалентный ему метод кригинга дают (в гауссовом случае) оптимальный ответ, но требуют задания априорного распределения или предварительного задания корреляционной функции. Про методы гребневой регрессии не утверждают, что решение оптимально, но оно также требует априорного знания при вы-

боре параметров метода. Отметим также, что байесов подход дает хорошее эвристическое основание для выбора метода, прямо не связанного с заданием априорного распределения.

Достоинства и недостатки байесова подхода

Подведем итог нашего рассмотрения методов, основанных на байесовой стратегии. Прежде всего, мы видим, что решения, основанные на этой стратегии (в тех случаях, когда нам удалось их получить) в существенной мере зависят от априорных сведений. В случае представления искомой зависимости разложением по базисным функциям это было априорное распределение коэффициентов разложения, а для метода кригинга это была априори заданная ковариационная функция. Байесов подход как таковой не дает никаких рекомендаций, откуда эти сведения взять.

Далее, мы недаром ограничились гауссовым случаем. Здесь нам удалось вместо интегрирования по всему пространству возможных значений коэффициентов просто найти минимум критерия, который и совпадал с условным математическим ожиданием. В общем же случае, получив апостериорное распределение $P_{\text{апост}}(\boldsymbol{a})$ на множестве всех возможных моделей $\boldsymbol{a}$ и задавшись функцией штрафа $S(t, \boldsymbol{a})$, где t — решение, следовало бы найти функцию риска

$$R(t) = \int S(t, \boldsymbol{a})\, dP_{\text{апост}}(\boldsymbol{a}),$$

и далее найти решение t^*, доставляющее минимум этой функции.

Модели, как правило, описываются большим набором параметров, и операция интегрирования должна проводиться в многомерном пространстве этих параметров. Аналитически этот интеграл в большинстве случаев не берется, а численные методы интегрирования в многомерных пространствах крайне неэффективны. Иногда предлагают использовать для этого метод Монте-Карло. Но, даже отвлекаясь от того, что интегрирование нужно проводить для каждого значения решения t, обратим внимание на следующее обстоятельство. Распределение $P_{\text{апост}}(\boldsymbol{a})$ обычно сосредоточено на множестве сравнительно правдоподобных моделей, которое составляет ничтожную долю от всех возможных моделей. Даже задача найти какую-то из этих правдоподобных моделей довольно сложна. Если методом Монте-Карло генериро-

вать модели с равной вероятностью, то вероятность случайно попасть на правдоподобную модель будет ничтожно мала. Можно, конечно, попытаться создать такой генератор, который будет сразу генерировать модели с вероятностью $P_{\text{апост}}(\boldsymbol{a})$, но это непросто сделать.

Это одна из причин того, почему я не рассказывал применение байесовой стратегии к задаче распознавания образов в общем виде. Рассмотрим такой пример. Пусть заведомо известно, что два класса объектов, представленных n-мерными векторами $\mathbf{x}$, полностью разделимы решающими правилами вида

$$\begin{aligned} &\text{если } (\varphi, \mathbf{x}) > 0, \quad \text{то объект принадлежит первому классу,} \\ &\text{если } (\varphi, \mathbf{x}) \leqslant 0, \quad \text{то объект принадлежит второму классу,} \end{aligned}$$

где $(\cdot, \cdot)$ — скалярное произведение, а φ — единичный вектор. Допустим, что априори значения вектора φ равномерно распределены на единичной сфере.

Пусть теперь дана обучающая выборка $(\mathbf{x}_i, y_i)$ длины l, где $y_i = \text{I}$, если объект принадлежит первому классу, и $y_i = \text{II}$, если объект принадлежит второму классу. Тогда при построении апостериорного распределения все векторы φ, у которых соответствующее решающее правило допустило хотя бы одну ошибку на обучающей выборке, будут признаны невероятными, а остальные останутся равновероятными. Это значит, что апостериорное распределение вектора φ будет равномерно распределено на сферическом многограннике, ограниченном неравенствами:

$$\begin{aligned} &(\varphi, \mathbf{x}_i) > 0, \qquad \text{если } y_i = \text{I}, \\ &(\varphi, \mathbf{x}_i) \leqslant 0, \qquad \text{если } y_i = \text{II}. \end{aligned} \tag{$*$}$$

Для того, чтобы найти хотя бы один вектор φ, удовлетворяющий условию (*), нужно решить систему линейных неравенств. Для того же, чтобы найти апостериорное среднее вектора φ, нужно проинтегрировать его значения по апостериорному распределению, т. е. по сферическому многограннику, ограниченному неравенствами (*). Понятно, что аналитически это сделать затруднительно. Если же воспользоваться методом Монте-Карло, задав равномерное распределение на многомерной сфере, то вероятность случайного попадания точки в наш многогранник будет крайне мала. Если же создать такой генератор, который генерировал бы точки только внутри многогранника, равномерно их распределяя, то задача решалась бы. Но как придумать такой генератор, я не знаю. Этот пример наиболее ярко иллюстрирует проблему, но

она возникает и при решении других задач с использованием байесова подхода.

Таким образом, мы видим, что прямое применение байесовой стратегии имеет весьма ограниченное применение в задачах восстановления зависимостей. Но байесов подход дает основание для построения методов выбора оптимальной структуры модели, которые можно было бы назвать «квази-байесовыми». Мы видели, как методы регуляризации, гребневой регрессии, ядерной гребневой регрессии, прямо не связанные с методом Байеса, основываются, тем не менее, на тех же формулах. По Байесу решение строилось как оптимальное при заданных параметрах априорного распределения (которые неясно откуда взять), тогда как в прямых методах регуляризации те же параметры просто назначались (и их значения тоже неизвестно откуда взять).

В то же время мы видели, что с увеличением коэффициента регуляризации одна за другой подавляются составляющие регрессии, соответствующие малым собственным числам корреляционной матрицы аргументов. То есть в определенном смысле происходит упрощение модели, но с другой стороны нивелируется влияние помехи. Чем хуже соотношение сигнал/шум, тем более простую модель нужно использовать.

Это наводит на мысль, что и вообще модели нужно упорядочить по сложности, и переходить к более сложным моделям только с увеличением длины выборки. Мерой сложности может быть число настраиваемых параметров модели, норма решения в какой-либо метрике или иной более сложный критерий.

Посмотрим на качественном уровне, за счет чего при применении метода Байеса происходит огрубление модели.

Представим себе, что все возможные модели упорядочены по уровню сложности. Число возможных моделей с переходом на каждый более сложный уровень многократно возрастает. Примем, грубо говоря, что все модели одного уровня сложности априори равновероятны. Тогда индивидуальная априорная вероятность каждой модели будет равна априорной вероятности уровня, деленной на число моделей этого уровня. Следовательно, индивидуальная априорная вероятность модели будет резко падать с увеличением уровня сложности. Апостериорная вероятность модели пропорциональна произведению меры ее правдоподобия, применительно к данным обучения, и ее априорной вероятности. Значит, при равном правдоподобии более простая модель

будет иметь значительно большую апостериорную вероятность. Например, полином высокой степени лучше аппроксимирует зашумленные экспериментальные данные, чем полином низкой степени, но его индивидуальная априорная вероятность мала, и апостериорная вероятность окажется меньшей. Кроме того, если данных недостаточно, чтобы среди моделей высокого уровня сложности выбрать одну конкретную или несколько близких, то при усреднении (при нахождении условного математического ожидания) автоматически происходит огрубление модели. Мы видели, что когда число коэффициентов регрессии превосходит число экспериментальных точек, то целое гиперпространство в пространстве коэффициентов отвечает моделям, обеспечивающим нулевую невязку. После усреднения по этому множеству выбирается модель с минимальной нормой, т. е. наиболее простая.

И все же мы видели, что даже в тех случаях, когда считается, что коэффициенты регрессии априори независимы и имеют одинаковую дисперсию, мы получаем целый спектр решений в зависимости от того, какую долю дисперсии шума в общей дисперсии данных мы примем априори. Грубо говоря, в зависимости от этого и будет выбрана оптимальная сложность модели. Можно ли средствами байесова подхода по экспериментальным данным определить или оценить и этот параметр — соотношение сигнал/шум? Мы вернемся к этому вопросу в одной из следующих лекций.

В теории игр и статистических решений кроме байесова критерия известны и другие. Это, например, минимаксный критерий или критерий минимакса потерь. Но при слабых ограничениях на стратегию природы оказывается, что природа может сыграть настолько каверзно, что любой алгоритм обучения окажется плохим. При более жестких ограничениях оптимальная стратегия обучения опять таки зависит от произвола выбора параметров этих ограничений.

В последующих лекциях мы рассмотрим пути выбора оптимальной сложности (а шире — структуры) моделей, не связанные напрямую с методом Байеса. Мы видели на лекциях прошлого семестра, что методы минимизации эмпирического риска дают оценку риска, заниженную по сравнению с истинным риском. Это занижение тем больше, чем хуже соотношение длины выборки и сложности модели в том или ином смысле. Основная идея методов, которые будут рассмотрены, состоит в том, чтобы предсказать значение истинного риска для модели, выбранной по критерию минимума эмпирического риска. С ростом

сложности значение минимума эмпирического риска всегда снижается, но добавка, характеризующая степень его занижения по сравнению с истинным риском, всегда растет. Оптимальная сложность соответствует тому уровню, для которого достигается минимум оценки истинного риска.

Лекция 18

Структурная минимизация эмпирического риска, общий подход. Прямые средства выбора оптимальной сложности модели. Learning set, validation set, control set. Скользящий контроль (cross validation). Конформные предикторы.

На лекциях прошлого семестра мы показали, что методы минимизации эмпирического риска — метод максимального правдоподобия, метод наименьших квадратов, методы выбора решающего правила, минимизирующие число ошибок на обучении — обладают следующим недостатком. Значение эмпирического риска в точке, доставляющей минимум этому критерию, оказывается всегда меньше значения истинного риска в этой точке. Причем это занижение тем больше, чем выше сложность модели. С другой стороны, снижение сложности приводит к огрублению модели по сравнению с истинной — все больше сложных деталей приходится сглаживать, пренебрегать тонкими эффектами.

На примере применения байесовой стратегии мы видели, что оптимальное по Байесу решение требует введения регуляризации, которая с одной стороны подавляет влияние помехи, а с другой стороны заставляет пренебречь составляющими, соответствующими малым собственным числам ковариационной матрицы.

Кривые изменения значений эмпирического и истинного риска в точке, доставляющей минимум эмпирическому риску, в зависимости от сложности модели можно представить следующим образом. Значение эмпирического риска монотонно убывает с ростом сложности. Это происходит потому, что класс более сложных моделей включает в себя и простые как частный случай. В то же время занижение эмпирического риска по сравнению с истинным монотонно растет с увеличением сложности. Поэтому значение истинного риска, пройдя через минимум, начинает расти, когда сложность чересчур велика. Проблема состоит в том, что значение эмпирического риска мы видим, тогда как значение истинного риска приходится оценивать тем или иным путем.

Заметим, что подход, в котором ищется истинная сложность модели, неправилен. Например, если истинная зависимость описывается полиномом пятой степени, то неверно, что наилучший результат мы получим, если будем искать зависимость в форме полинома пятой степени. Присутствие помехи и ограниченность длины выборки приводят к тому, что при малой выборке мы вынуждены будем огрубить модель и искать полином меньшей степени. То же самое относится к обучению распознаванию образов. При малой выборке мы бываем вынуждены исключить ряд признаков, хотя в принципе для распознавания они нужны.

Упорядочение можно вести по разным критериям, в различных отношениях отражающих сложность модели [5, 7, 8].

Простейший критерий — это число входных параметров моделей (число признаков в задачах распознавания, число аргументов или фиксированных базисных функций в задачах регрессии). Признаки или аргументы могут быть заранее упорядочены, и тогда выбор оптимальной сложности может идти путем включения (или исключения) признаков в порядке их априорной важности. Но аргументы могут и не быть заранее упорядочены. Тогда остается только перебор их комбинаций. Но полный перебор всех комбинаций очень затратен как в вычислительном отношении, так и в смысле занижения риска из-за большого количества просматриваемых вариантов. Тогда применяют различные процедуры последовательного отбора признаков путем последовательного включения и/или исключения признаков, в наибольшей степени улучшающих оценку истинного риска.

В нейронных сетях параметрами сложности модели будут число слоев сети и число нейронов в каждом слое. В некоторых алгоритмах исходные аргументы агрегируются путем построения вторичных признаков как нелинейных функций небольшого числа исходных признаков. Тогда мерой сложности будет число вторичных признаков и сложность их построения.

Параметрами, характеризующими сложность модели, может быть величина коэффициента регуляризации, радиус корреляции при кригинге или соответствующая «ширина» ядра при ядерной гребневой оценке. Мы видели также, что в случае априори независимых коэффициентов регрессии решение существенно зависит от того, насколько быстро убывает априорная дисперсия этих коэффициентов (если упорядочить их в порядке убывания дисперсии). Скорость убывания

априорной дисперсии также определяет сложность модели.

Оптимизация может вестись сразу по нескольким критериям. Может быть включен перебор по дискретным характеристикам, таким как выбор вида ядра или назначение вида нормы решения, используемой при регуляризации.

Простой экзамен

Самый верный способ оценить истинный риск для выбранной модели — это протестировать ее на новом материале, никак не использованном при обучении. Эта оценка оказывается довольно точной в силу обычного закона больших чисел. Но такую оценку можно сделать только один раз. Если же использовать материал экзамена для настройки параметров алгоритма или выбора сложности модели, то фактически эти данные будут входить в обучение, и оценка риска окажется заниженной.

Действительно, пусть, например, вероятность того, что риск, оцененный по материалу экзамена (при однократном использовании), окажется ниже истинного на величину a, составляет одну тысячную. Если же мы проведем тысячу независимых испытаний, то вероятность p того, что ни в одном испытании риск не уклонится в лучшую сторону более чем на a, составит

$$p = (1 - 0.001)^{1000} \approx e^{-1} \approx 40\%,$$

то есть с вероятностью 60% хотя бы в одном испытании риск будет занижен более чем на величину a. А так как мы выбираем лучший вариант, то именно этот вариант будет признан оптимальным, хотя результат может быть совершенно случайным. На самом деле испытания при подборе параметров алгоритма не являются независимыми, но, тем не менее, выдавать полученный просмотром многих вариантов результат как окончательный крайне рискованно.

Поэтому весь имеющийся материал с известным выходным значением (классификацией объекта, числовым или векторным значением на выходе) обычно делят на три части: собственно данные обучения (Learning set), данные, используемые для настройки параметров алгоритма (Validation set), и честный экзамен (Test set). Данные обучения используют непосредственно в алгоритме обучения, например, при настройке нейронной сети с фиксированной структурой, в алгоритме

SVM или при построении регрессии с заданными параметрами регуляризации. Результат оценивается на настроечных данных (Validation set) и подбираются лучшие параметры алгоритма. Окончательная оценка качества обучения получается тестированием наилучшего варианта на независимом материале, никак не использованном при обучении.

В моей практике был такой случай. В задаче распознавания двух классов весь материал использовался для разбивки признаков, представленных действительными числами, на градации, для того чтобы превратить их в дискретные признаки. Часть преобразованных таким путем данных использовалась далее для обучения, а другая часть — для теста. В результате получилось примерно 75% правильных ответов. Но когда разбиение на градации было проведено только по данным обучения, без использования теста, то на экзамене получилось только 50% правильных ответов (два класса были представлены в одинаковой пропорции). То есть результат оказался полностью отрицательным — никакого распознавания не получилось.

Поскольку обычно количество данных, оставленных на экзамен, бывает не слишком большим, то в этом случае дисперсия результата экзамена оказывается значительной. Тогда случайное разбиение данных на обучение и экзамен повторяют несколько раз и в качестве оценки качества обучения берут среднее значение результата.

Скользящий контроль (Cross validation)

Данных с известным выходным значением обычно бывает мало, и, если их разделить на три части, то обучающая выборка оказывается слишком короткой. Поэтому был придуман метод, позволяющий экономить данные, резервируемые для настройки алгоритма или для экзамена. Этот метод в русской литературе получил название *скользящий контроль*, а в англоязычной — метод *Leave-one-out* или *Cross Validation*.

Метод состоит в следующем. Из обучающей выборки удаляется один объект $\mathbf{x}_i$ и соответствующее выходное значение y_i. По оставшейся части выборки проводится обучение и получается решение t_{-i} (решающее правило, регрессионная зависимость, модель). Решение применяется к выделенному объекту $\mathbf{x}_i$ и вычисляется предсказание $y_i^*(t_{-i}, \mathbf{x}_i)$ в точке $\mathbf{x}_i$. Далее сравнивается это предсказание с фактическим значением y_i и по заданной функции штрафа S находится штраф $S_i = S(y_i, y_i^*)$ за расхождение предсказания и фактического значения. Это может

быть штраф за ошибки разного рода в задачах распознавания, квадрат уклонения или иной критерий в задачах регрессии.

Затем выделенный объект возвращается в обучающую выборку, удаляется другой объект, и операция повторяется применительно ко всем объектам выборки. Для каждого выделенного объекта вычисляется штраф S_i и средний штраф

$$S^* = \frac{1}{l}\sum_{i=1}^{l} S_i$$

по всем объектам выборки длины l.

Поскольку выделенные объекты не участвуют в процессе обучения, то величина S^* оказывается несмещенной оценкой качества обучения на выборках длины $l-1$. Это, конечно, верно только в том случае, когда все объекты обучающей выборки получены случайно, независимо и при неизменном распределении вероятностей. Если же точки измерения размещаются искусственно, несмещенность оценки может и не выполняться. Так, например, в задаче оценки поля содержаний золота, о которой я рассказывал, точки измерения размещались на плоскости по сети 5×5 м, и прогноз нужно было дать в точках, расположенных не дальше $2.5\sqrt{2}$ м от ближайшей точки измерения. Если же мы удаляем одну из точек опробования, то прогноз в ней придется дать по измерениям, ближайшее из которых расположено не ближе $5\sqrt{2}$ м от нее.

Если же метод скользящего контроля применяется для выбора лучших параметров алгоритма путем перебора вариантов, то оценка будет смещенной, и для окончательной оценки качества обучения необходимо резервировать часть данных на чистый экзамен. Сложнее оказывается оценить дисперсию или среднее уклонение результата скользящего контроля от математического ожидания. Дело в том, что величины S_i не являются независимыми, ведь большая часть выборки при их вычислении сохраняется. Поэтому оценок дисперсии скользящего контроля в общем случае получено не было. (Может быть, у вас получится?)

Метод скользящего контроля в принципе можно применять в любых задачах обучения. Но в общем случае метод оказывается довольно трудоемким в вычислительном отношении: для каждого удаленного объекта выборки нужно заново проводить процесс обучения.

Однако в частных случаях дело упрощается. При применении метода обобщенного портрета или SVM при удалении вектора, не являющегося опорным и не выброшенного из обучения, как мешающего разделению, ошибки на нем заведомо не будет. Ошибка вероятнее всего будет, если удалить вектор, который выбрасывается из обучения. Поэтому можно получить оценку сверху результата скользящего контроля, проведя однократно процесс обучения по полной выборке:

$$S^* \leqslant \frac{1}{l}(n_{\text{опор}} + n_{\text{выбр}}),$$

где $n_{\text{опор}}$ — число опорных векторов, $n_{\text{выбр}}$ — число выброшенных из обучающей выборки векторов (как мешающих разделению).

Также удается избежать повторных вычислений при нахождении оценки скользящего контроля в задаче оценки регрессии с регуляризацией по формуле

$$t(\mathbf{x}) = \mathbf{f}^{\mathrm{T}}(\mathbf{F}\mathbf{F}^{\mathrm{T}} + D_\xi \mathbf{E})^{-1}\mathbf{F}\mathbf{c}.$$

Здесь $\mathbf{c} = (c(\mathbf{x}_1), \ldots, c(\mathbf{x}_l))$ — вектор значений, измеренных в точках $\mathbf{x}_1, \ldots, \mathbf{x}_l$, $\mathbf{F}$ — матрица с элементами $f_{ij} = \varphi_i(\mathbf{x}_j)$, где $\varphi_i(\mathbf{x})$ — базисные функции, $\mathbf{f}$ — вектор значений базисных функций в текущей точке $\mathbf{x}$: $\mathbf{f}(\mathbf{x}) = (\varphi_1(\mathbf{x}), \ldots, \varphi_n(\mathbf{x}))$, D_ξ — параметр регуляризации, $t(\mathbf{x})$ — прогноз значения функции в точке $\mathbf{x}$.

Если мы хотим получить прогноз $t(\mathbf{x}_j)$ в одной из точек измерения, то в качестве вектора $\mathbf{f}$ следует взять вектор $\mathbf{f}(\mathbf{x}_j) = (\varphi_1(\mathbf{x}_j), \ldots, \varphi_n(\mathbf{x}_j))$. Получим

$$t(\mathbf{x}_j) = \mathbf{f}^{\mathrm{T}}(\mathbf{x}_j)(\mathbf{F}\mathbf{F}^{\mathrm{T}} + D_\xi \mathbf{E})^{-1}\mathbf{F}\mathbf{c}.$$

Обозначим вектор прогноза во всех точках измерения $\mathbf{t} = (t(\mathbf{x}_j), \ldots, t(\mathbf{x}_l))$. Тогда

$$\mathbf{t} = \mathbf{F}^{\mathrm{T}}(\mathbf{F}\mathbf{F}^{\mathrm{T}} + D_\xi \mathbf{E})^{-1}\mathbf{F}\mathbf{c},$$

поскольку j-тый столбец матрицы $\mathbf{F}$ совпадает с вектором $\mathbf{f}(\mathbf{x}_j)$. Обозначим через $\mathbf{H}$ матрицу $\mathbf{F}^{\mathrm{T}}(\mathbf{F}\mathbf{F}^{\mathrm{T}} + D_\xi \mathbf{E})^{-1}\mathbf{F}$, а через h_{jj} — диагональные элементы, стоящие на пересечении j-й строки j-го столбца матрицы $\mathbf{H}$.

Обозначим $t^*(\mathbf{x}_j)$ предсказание значения функции в точке $\mathbf{x}_j$, если этот элемент удален из обучающей последовательности, а функция

оценивается тем же путем, что и раньше. Оказывается, что уклонения предсказаний $t(\mathbf{x}_j)$ и $t^*(\mathbf{x}_j)$ от истинного (измеренного) значения $c(\mathbf{x}_j)$ связаны следующим соотношением:

$$c(\mathbf{x}_j) - t^*(\mathbf{x}_j) = \frac{c(\mathbf{x}_j) - t(\mathbf{x}_j)}{1 - h_{jj}}.$$

Поэтому для квадратичного критерия невязки оценка качества обучения методом скользящего контроля в нашем случае может быть подсчитана по формуле:

$$S* = \frac{1}{l}\sum_{j=1}^{l}\left(\frac{c(\mathbf{x}_j) - t(\mathbf{x}_j)}{1 - h_{jj}}\right)^2.$$

При этом для вычисления значений $t(\mathbf{x}_j)$ и диагональных элементов h_{jj} матрицу $(\mathbf{F}\mathbf{F}^{\mathrm{T}} + D_\xi \mathbf{E})$ достаточно обратить всего один раз, что, конечно, резко сокращает объем вычислений.

Метод конформных предикторов

Теперь я расскажу об одном методе, не связанном прямо с нашей основной темой — выбором оптимальной сложности, но близком по своей идее к методу скользящего контроля.

До сих пор мы считали, что результатом восстановления зависимости является правило, по которому оценивается значение зависимости в точках, соответствующих новым объектам, не представленных в данных обучения. Но во многих приложениях требуется также оценить точность предсказания, для конкретного объекта. Применительно к задачам регрессии это будет доверительный интервал для заданного уровня толерантности — вероятности того, что истинное значение лежит в пределах интервала. В задачах распознавания это будет вероятность того, что та или иная классификация объекта окажется правильной.

Метод конформных предикторов, предложенный в Лондонском университете А. Гаммерманом и В. Вовком, предназначен для ответа на этот вопрос в широком диапазоне задач восстановления зависимостей [31, 25]. Метод состоит в следующем. Предполагается, что алгоритм обучения позволяет определенным (разумным) образом линейно

упорядочить точки обучающей выборки по степени их «странности». В задачах регрессии это может быть упорядочение объектов выборки по величине квадрата остаточной невязки. Здесь объект считается тем более «странным», чем выше остаточная невязка. При применении машины опорных векторов (SVM) упорядочение может идти по величине веса опорных векторов в решающем правиле. Более «странными» считаются векторы с большим весом, а векторы, не вошедшие в число опорных, считаются наименее странными, но могут быть упорядочены по степени близости к разделяющей поверхности. Векторы, удаленные из обучающей выборки как мешающие разделению считаются наиболее «странными». Кроме того, предполагается, что это упорядочение не зависит от последовательности, в которой элементы выборки задаются алгоритму обучения.

Итак, пусть дана обучающая выборка $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_l, y_l)$, где $\mathbf{x}_i$ — описание объекта, y_i — выходное значение, и дан новый объект, который мы обозначим $\mathbf{x}_{l+1}$, для которого требуется оценить неизвестное выходное значение y_{l+1}. Далее пробегаются все возможные значения выхода y_{l+1}. В задачах распознавания это будут все возможные значения классификации объекта $\mathbf{x}_{l+1}$. В задачах регрессии, конечно, приходится перебирать значения выходной величины y_{l+1} с достаточно малым шагом и в каком-то правдоподобном диапазоне. Для каждого значения y_{l+1} строится расширенная обучающая выборка путем пополнения ее парой $(\mathbf{x}_{l+1}, y_{l+1})$, и все объекты расширенной выборки упорядочиваются по их «странности» (как было сказано выше). При этом пара $(\mathbf{x}_{l+1}, y_{l+1})$ займет какое-то место $m(y_{l+1})$ в этом порядке.

Теперь задают некоторый уровень толерантности p. Полагается, что значение y_{l+1} лежит в пределах доверительного интервала, соответствующего уровню p, если величина $m(y_{l+1})/(l+1)$ меньше или равна p (здесь $l+1$ — длина расширенной выборки). Все значения y_{l+1}, для которых выполнилось условие

$$\frac{m(y_{l+1})}{l+1} \leqslant p, \tag{1}$$

и образуют «доверительный интервал». Слова «доверительный интервал» мы берем в кавычки, потому что здесь мы используем традиционный термин. В нашем случае это просто множество значений y, про которое мы утверждаем, что с вероятностью не меньшей p значение y_{l+1} на выходе принадлежит этому множеству. В задачах регрессии

этот интервал действительно обычно оказывается связным. В задачах распознавания все значения классификации, для которых выполнилось условие (1), считаются правдоподобными на доверительном уровне p. Заметим, что при определении доверительного интервала используются только обучающая выборка $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_l, y_l)$ и описание объекта $\mathbf{x}_{l+1}$, но не фактическое значение y_{l+1}.

Утверждается, что, если все пары $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_l, y_l)$ и $(\mathbf{x}_{l+1}, y_{l+1})$ получены независимо при одинаковом распределении вероятностей, то вероятность того, что фактическое значение окажется в пределах доверительного интервала, не меньше $p - 1/(l+1)$. А с точностью до аппроксимации величины p отношением целых чисел $m/(l+1)$ просто равна p. При этом вероятность считается по всем возможным реализациям обучающей выборки и пары $(\mathbf{x}_{l+1}, y_{l+1})$.

На качественном уровне доказательство этого утверждения таково. Возьмем случайную выборку пар $(\mathbf{x}_i, y_i)$ длины $l+1$ и проведем их упорядочение по «странности» в соответствии с данным алгоритмом обучения. Тогда вероятность того, что именно последняя пара $(\mathbf{x}_{l+1}, y_{l+1})$ окажется на месте m в этом порядке, будет одинаковой для всех m и равна $1/(l+1)$. Вероятность же того, что для нее выполняется условие $m/(l+1) \leqslant p$, равна $1/(l+1)$ умножить на число позиций, для которых это условие выполнено. С точностью до величины $1/(l+1)$ эта вероятность просто равна p.

Но при построении доверительного интервала по выборке $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)$ и описанию объекта $\mathbf{x}_{l+1}$ фактическое значение y_{l+1} просматривалось в числе прочих. При этом оно попадет в доверительный интервал в том и только том случае, когда пара $(\mathbf{x}_{l+1}, y_{l+1})$ окажется при упорядочении по «странности» на месте m, удовлетворяющем условию $m/(l+1) \leqslant p$. А это происходит, как мы видели, как раз с вероятностью p (с точностью до $1/(l+1)$).

Формальное же доказательство таково. Обозначим через $\mathbf{R}((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l), \mathbf{x}_{l+1})$ доверительный интервал, построенный по выборке $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l)$ и описанию объекта $\mathbf{x}_{l+1}$, а событием $A((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{l+1}, y_{l+1}))$ назовем выполнение условия $y_{l+1} \in \mathbf{R}$. Тогда вероятность выполнения A равна

$$P(A) = \int I((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{l+1}, y_{l+1}))\, dP(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{l+1}, y_{l+1}),$$

где $I((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{l+1}, y_{l+1}))$ — индикаторная функция события A. Но

поскольку все пары $(\mathbf{x}_i, y_i)$ получаются независимо при одинаковом распределении, то величина интеграла не должна меняться при любой перестановке этих пар. Поэтому

$$P(A) = \int \sum \frac{1}{l+1} I\big(T_k\big((\mathbf{x}_1, y_1), \ldots \\ \ldots (\mathbf{x}_{l+1}, y_{l+1})\big)\big)\, dP((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{l+1}, y_{l+1})),$$

где сумма берется по всем $(l+1)$ циклическим перестановкам T_k последовательности пар $(\mathbf{x}_i, y_i)$. При этом каждая пара окажется ровно один раз на последнем месте в этой последовательности.

Допустим, что все пары расширенной выборки $((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{l+1}, y_{l+1}))$ упорядочены по «странности» (в порядке возрастания «странности») в соответствии с данным алгоритмом обучения, и каждая пара $(\mathbf{x}_i, y_i)$ оказалась на месте $m(\mathbf{x}_i, y_i)$ в этом порядке. Понятно, что каждое место m $(1 \leqslant m \leqslant l+1)$ будет занято какой-то парой и при том только одной.

Найдем теперь значение индикаторной функции $I(T_k((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{l+1}, y_{l+1})))$ для фиксированной перестановки T_k. Допустим, что при этой перестановке на последнем месте в последовательности оказалась пара $(\mathbf{x}_i, y_i)$. Тогда значение индикаторной функции будет равно 1 в том и только том случае, когда выполнено неравенство

$$m(\mathbf{x}_i, y_i) \leqslant p(l+1). \tag{2}$$

Действительно, по нашему правилу все пары, кроме последней, упорядочиваются в соответствии с их фактическим значением, а для последней пары просматриваются все возможные значения выхода y, и значение y включается в доверительный интервал в том и только том случае, когда выполнено неравенство $m(\mathbf{x}_i, y) \leqslant p(l+1)$. При этом просматривается и фактическое значение y_i, и оно попадет в доверительный интервал только в том случае, когда выполнено неравенство (2). В этом случае значение индикаторной функции будет равно 1, во всех остальных оно будет равно 0.

Теперь сумма индикаторных функций по всем циклическим перестановкам T_k будет равна числу случаев, когда для пары $(\mathbf{x}_i, y_i)$, попавшей при перестановке T_k на последнее место, выполнится неравенство (2). А так как все места m от 1 до $l+1$ заняты ровно одной

парой, то число таких случаев будет равно $[p(l+1)]$, где $[\cdot]$ означает целую часть. Таким образом

$$\sum \frac{1}{l+1} I(T_k((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{l+1}, y_{l+1}))) = \frac{[p(l+1)]}{l+1}.$$

Мы видим, что эта сумма не зависит от конкретной последовательности пар. Отсюда

$$P(A) = \int \sum \frac{1}{l+1} I\big(T_k((\mathbf{x}_1, y_1), \ldots \\ \ldots (\mathbf{x}_{l+1}, y_{l+1}))\big)\, dP(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{l+1}, y_{l+1}) = \frac{[p(l+1)]}{l+1}.$$

Утверждение доказано.

Этот метод отличается своей универсальностью. Он может быть применен к любым алгоритмам обучения, лишь бы с этим алгоритмом естественным образом связывался некоторый способ упорядочения объектов, представленных в обучающей выборке, по степени их «странности».

Конечно, полученный доверительный интервал будет зависеть от способа упорядочения объектов по «странности», связанного с алгоритмом обучения. Как уже отмечалось, в случае построения регрессии критерием упорядочения может быть остаточная невязка, а в случае применения SVM — вес опорного вектора в направляющем векторе разделяющей гиперплоскости и/или штраф за ошибку. Но в принципе доверительный интервал и должен зависеть от применяемого алгоритма обучения.

Отметим еще, что метод конформных предикторов требует, чтобы все объекты расширенной выборки строго упорядочивались по странности, т. е. чтобы не было нескольких объектов, имеющих одинаковую степень «странности». В случае упорядочения по остаточной невязке это, как правило, так и будет. В случае упорядочения по весу опорных векторов объекты, не вошедшие в число опорных или одинаково ошибочно распознанные, будут иметь одну и ту же степень «странности», то есть слипнутся при упорядочении. Тогда приходится различать их по какому-либо дополнительному критерию, например, по расстоянию до разделяющей поверхности. Это вносит, конечно, дополнительный произвол.

В задачах регрессии «доверительный интервал» в нашем смысле получается, как правило, действительно связанным отрезком. В задачах распознавания возможны несколько вариантов. 1) Только одна классификация нового объекта попадает в интервал. Тогда с надежностью p эта классификация может быть признана правильной. 2) Несколько классификаций попадают в доверительный интервал. Это значит, что данных обучения недостаточно для того, чтобы их различить. 3) Ни одна из классификаций не попадает в доверительный интервал. Это значит, что новый объект попал в зону неопределенности — любая из классификаций не полностью согласуется с данными обучающей выборки.

Мы видели, изучая байесову стратегию, что апостериорное распределение может в существенной мере зависеть от априорного распределения — от априорных данных меняется среднее и дисперсия апостериорного распределения. Поэтому никакого способа построить объективный доверительный интервал без использования априорных данных быть не может. Тем не менее, экспериментальная проверка показала, что при достаточно большой обучающей выборке метод конформных предикторов дает результат очень близкий к тому, который получается по Байесу.

Лекция 19

Структурная минимизация эмпирического риска на базе оценок равномерной сходимости. Общий подход.

Другой подход к выбору оптимальной сложности модели состоит в том, чтобы оценить для каждого уровня сложности степень занижения эмпирического риска в сравнении с истинным риском, исходя из теоретических результатов, полученных нами ранее [7, 8].

В прошлом семестре мы выводили условия равномерной сходимости частот к вероятностям по системе событий S, и получили соответствующие оценки. Сейчас мы применим этот результат для того, чтобы оценить, насколько значение минимума эмпирического риска занижено по сравнению со значением истинного риска в точке, доставляющей этот минимум, в задаче обучения распознаванию образов.

Рассмотрим вероятностное пространство, состоящее из пар $(\mathbf{x}, y)$, где $\mathbf{x}$ — описание объекта, y — его фактическая классификация, и будем

считать, что в этом пространстве задана вероятностная мера $P(\mathbf{x}, y)$. Пусть теперь задан класс $\mathbf{K}$ решающих правил $R(\mathbf{x})$, в котором путем минимизации эмпирического риска (числа ошибок на обучающей выборке) ищется оптимальное правило. Рассмотрим систему S событий

$$A(R) = \{(\mathbf{x}, y) : y \neq R(\mathbf{x})\},$$

то есть множеств пар $(\mathbf{x}, y)$, на которых решающее правило R совершает ошибку. Предполагается, конечно, что все эти события измеримы относительно меры P. Тогда частота выпадения события $A(R)$ на данных обучения будет просто равна частоте ошибок на обучающей выборке для решающего правила R, а вероятность события $A(R)$ равна вероятности ошибок, совершаемых этим решающим правилом на всей генеральной совокупности, то есть истинному риску. Поэтому уклонение частоты события $A(R)$ от его вероятности как раз и равно уклонению значения эмпирического риска от истинного риска для этого решающего правила.

Если мы говорим, что частота выпадения событий A равномерно по системе S близка к их вероятности, то есть

$$\sup_{A \in S} |P(A) - \nu(A)| < \varepsilon,$$

то это значит, что для всех решающих правил класса $\mathbf{K}$ эмпирический риск уклонится от истинного не более чем на ε ($\varepsilon > 0$). В частности, это будет верно и для решающего правила, на котором достигается минимум эмпирического риска в классе $\mathbf{K}$, то есть для решающего правила, обеспечивающего минимум числа ошибок на обучающей выборке. Следовательно, значение истинного риска для этого решающего правила гарантированно не превысит значения эмпирического риска плюс ε.

В свое время мы получили следующую оценку

$$P\left(\sup_{A \in S} |P(A) - \nu(A)| > \varepsilon\right) < 3M^S(2l) \exp\left(-\frac{\varepsilon^2(l-1)}{4}\right), \qquad (1)$$

где l — длина выборки, а $M^S(l)$ —функция роста, которая равна максимальному числу различных способов разделения выборки на классы с помощью решающих правил класса $\mathbf{K}$. Если для всякого l найдется выборка длины l такая, что ее можно разбить на классы всеми возможными способами с помощью решающих правил вида $R(\mathbf{x})$, то $M^S(l) \equiv 2^l$,

и оценка (1) становится тривиальной, так как правая часть неравенства при этом будет больше 1. Если же это не так, то существует максимальная длина выборки, которую еще можно разбить всеми возможными способами. Обозначим эту длину $r-1$. Как было показано, в таком случае функция роста мажорируется степенной функцией длины выборки:

$$M^S(l) < \frac{3l^{r-1}}{2(r-1)!}.$$

Подставляя эту оценку в (1), получим

$$P\left(\sup_{A\in S}|P(A)-\nu(A)|<\varepsilon\right) < 4.5\frac{(2l)^{r-1}}{(r-1)!}\exp\left(-\frac{\varepsilon^2(l-1)}{4}\right). \qquad (2)$$

В определенном смысле величина r определяет емкость (сложность) класса решающих правил $\mathbf{K}$. В большинстве случаев (но не всегда) она равна числу настраиваемых параметров решающего правила. Правая часть неравенства (2) стремится к нулю с ростом длины выборки l, и тем быстрее, чем меньше величина r.

Потребуем, чтобы вероятность $P(\sup_{A\in S}|P(A)-\nu(A)|<\varepsilon)$ не превышала заданное значение $\eta>0$. Это произойдет, если добиться выполнения равенства

$$4.5\frac{(2l)^{r-1}}{(r-1)!}\exp\left(-\frac{\varepsilon^2(l-1)}{4}\right)=\eta. \qquad (3)$$

Это равенство можно разрешить относительно ε. Таким образом, справедливо утверждение: с вероятностью, превышающей $1-\eta$, максимальное по классу S уклонение частоты от вероятности не превосходит величину

$$\varepsilon = 2\sqrt{\frac{r(\ln(2l/r)+1)-\ln(\eta/5)}{l-1}}. \qquad (4)$$

Отсюда следует, что с вероятностью, превышающей $1-\eta$, значение истинного риска для решающего правила $\mathbf{R}^*$, минимизирующего эмпирический риск, не превосходит величину

$$R_{\text{эмп}}(\mathbf{R}^*)+2\sqrt{\frac{r(\ln(2l/r)+1)-\ln(\eta/5)}{l-1}}.$$

Эта величина является гарантированной оценкой истинного риска с надежностью $1-\eta$.

Прибавляя добавку $2\sqrt{(r(\ln(2l/r)+1)-\ln(\eta/5))/(l-1)}$ к полученному значению эмпирического риска для различных уровней сложности, определяемой величиной r, можно найти тот уровень, который обеспечивает минимум гарантированной оценки истинного риска.

Относительные уклонения

Полученные нами оценки скорости равномерной сходимости в действительности оказываются заниженными. Уклонение истинного риска от эмпирического оценено только сверху. Это связано с тем, что пришлось пойти на завышение оценок во избежание чрезмерной громоздскости самих оценок и технических сложностей при их выводе. Но в еще большей степени это вызвано тем, что, желая получить общий результат, пришлось ориентироваться на наихудший случай (с точки зрения занижения эмпирического риска в сравнении с истинным) по тем параметрам, которые не входят явно в оценку.

В частности, для того, чтобы уложиться в заданное абсолютное уклонение частоты от вероятности некоторого события A, приходится взять бóльшую выборку, если вероятность A близка к $1/2$, и меньшую, если $P(A)$ близка к 0 или 1. В самом деле, для $P(A)=0.5$ и допустимого абсолютного уклонения в 1% необходимо $\approx 10^4$ показов, тогда как при $P(A)=0.05$ достаточна длина выборки порядка 100–200 показов. Если же необходимо получить оценку сверху, не зависящую от $P(A)$, то приходится ориентироваться на наихудший случай, то есть $P(A)=0.5$.

Вообще известно, что для фиксированного события A отклонение частоты от вероятности имеет прядок ε, если среднеквадратичное уклонение частоты σ имеет тот же порядок. В свою очередь

$$\sigma = \sqrt{\frac{1}{l}} \cdot \sqrt{P(A)(1-P(A))},$$

т. е. при фиксированной длине выборки l отклонение пропорционально $\sqrt{P(A)(1-P(A))}$.

Поэтому естественно было бы и равномерное уклонение измерять в относительных единицах, т. е. оценивать величину

$$\sup_{A\in S} \frac{|P(A)-\nu(A)|}{\sqrt{P(A)(1-P(A))}}.$$

Однако такого рода оценку при разумных предположениях удается получить только для равномерного уклонения частот в двух полувыборках, нормированного к эмпирической оценке величины $\sqrt{P(A)(1-P(A))}$ по всей выборке. А именно, поучена оценка

$$P\left\{\sup_{A\in S}\frac{|\nu'(A)-\nu''(A)|}{\sqrt{(\nu(A)+0.5l)(1-\nu(A)+0.5l)}}>\varepsilon\right\}< \\ <4M^S(2l)\exp\left(-\frac{\varepsilon^2 l}{4}\right), \tag{5}$$

где $\nu'(A)$ и $\nu''(A)$ — частоты выпадения события A соответственно в первой и второй полувыборках длины l, $\nu(A)=(\nu'(A)+\nu''(A))/2$ — частота выпадения события на полной выборке длины $2l$, $M^S(l)$ — функция роста для событий системы S.

При этом достигается определенное «равноправие» событий системы S. Что же касается равномерного относительного уклонения частот от вероятностей, то здесь удается получить одностороннюю оценку:

$$P\left\{\sup_{A\in S}\frac{P(A)-\nu(A)}{\sqrt{P(A)}}>\varepsilon\right\}<16M^S(2l)\exp\left(-\frac{\varepsilon^2 l}{4}\right).$$

Нормирующий делитель $\sqrt{P(A)}$ при малых значениях $P(A)$ близок к величине $\sqrt{P(A)(1-P(A))}$.

Это неравенство также можно разрешить относительно ε, приравняв его правую часть к заданной величине $\eta>0$. Но мы не будем этого сейчас делать ввиду громоздкости выкладок, а отложим до следующей лекции, когда рассмотрим применение этого результата к выбору оптимальной сложности в регрессионных задачах.

Упорядочение по оценкам относительного расстояния между классами и задача минимизации суммарного риска

Теперь мы рассмотрим постановку задачи обучения распознаванию образов, которая отличается от поставленной ранее, но для многих случаев, встречающихся на практике, может оказаться более естественной.

Особенность этой постановки состоит в том, что наряду с обучающей выборкой длины l предъявляется выборка векторов

$$x_1^*, x_2^*, \ldots, x_p^*$$

без указания их классификации, которую мы будем называть *рабочей выборкой*. Предполагается, что эта выборка и фактическая классификация представленных в ней векторов получены при том же распределении вероятностей, что и обучающая выборка. Задача состоит в том, чтобы используя обучающую и рабочую выборку найти такое линейное решающее правило, которое обеспечивает минимальное число ошибок на рабочей выборке. Таким образом, наша цель состоит не в том, чтобы найти общее решающее правило, минимизирующее число ошибок на всей генеральной совокупности, а лишь такое, которое обеспечивает минимум ошибок на рабочей выборке. Эту задачу мы будем называть задачей минимизации суммарного риска.

Мы будем различать постановки двух видов — детерминистскую и стохастическую. Детерминистская постановка предполагает, что искомое решающее правило должно безошибочно классифицировать обучающую выборку. Если это невозможно — следует просто отказ от задачи. Но при большом числе первичных или вторичных признаков это всегда возможно. Стохастическая постановка предполагает, что выбранное решающее правило может допускать ошибки на материале обучения.

Замечательная особенность задачи минимизации суммарного риска состоит в том, что уже до начала обучения множество всех решающих правил распадается на конечное число эквивалентных с точки зрения обучающей и рабочей выборок правил. (Эквивалентными мы считаем правила, одинаково разделяющие все элементы обучающей и рабочей выборок).

Линейные решающие правила можно упорядочить по числу m используемых в них признаков. Тогда число эквивалентных решающих гиперплоскостей, как мы видели, оценивается величиной

$$M^S(l+p) < 1,5\frac{(l+p)^m+1}{(m+1)!}, \tag{6}$$

где l — длина обучающей выборки, а p — длина рабочей выборки.

Но линейные решающие правила можно упорядочить и по другому принципу. Обозначим D диаметр полной выборки, состоящей из

обучающей и рабочей выборки:

$$D = \max_{i,j} |x_i - x_j|.$$

Далее, для каждого линейного решающего правила R, разделяющего полную выборку на два класса, можно построить выпуклые оболочки T_1 и T_2 точек полной выборки, относимых к первому и ко второму классам, и найти расстояние $\rho(R)$ между ними. Теперь все линейные решающие правила можно упорядочить по величине

$$t = \frac{D^2}{4\rho^2}.$$

Оказывается, что, если ограничить класс решающих правил условием $(D^2/4\rho^2) \leqslant t$, то максимальное число точек k, которые можно разделить всеми возможными способами с помощью этих решающих правил равно $[t+2]$. Но если так, то число всех возможных различных способов разделить полную выборку длины $l+p$ с помощью решающих правил, подчиненных условию $(D^2/4\rho^2) < t$, не превышает

$$M^S(l+p) < 1.5\frac{(l+p)^{[t+2]}}{[t+2]!}. \tag{7}$$

Для этого нужно показать, что максимальное число точек k, которые еще можно разделить всеми возможными способами с помощью линейных решающих правил, удовлетворяющих условию $D^2/4\rho^2 < t$, оценивается как $[t+2]$. Тогда на основании результата, полученного нами в прошлом семестре, можно оценить функцию роста $M^S(l)$. Из соображений симметрии ясно, что для того, чтобы точки можно было разделить гиперплоскостями всеми возможными способами и при этом минимальное расстояние $\rho(R)$ было как можно большим, их нужно разместить в вершинах правильного симплекса, вписанного в сферу диаметра D . При этом оказывается, что минимальное расстояние между выпуклыми оболочками равно $D/\sqrt{k-1}$ для четных k и $(Dk/(k-1))/\sqrt{k-1}$ для нечетных k (здесь k — число точек). Отсюда следует, что если для всех способов разделения $D^2/4\rho^2 < t$, то должно выполняться условие $k < t+2$.

Объединим теперь оценки (6) и (7). Обозначим $d = \min(t{+}2, m{+}1)$, где t — величина ограничения по соотношению $D^2/4\rho^2$, а m — число используемых признаков, и обозначим через $S(t,m)$ множество линейных

решающих правил, подчиненных этим условиям:

$$M^{S(t,m)}(l+p) < 1.5\frac{(l+p)^{[d]}}{[d]!}.$$

Для упрощения выкладок положим далее, что длина обучающей выборки равна длине рабочей выборки: $l = p$. Тогда вероятность того, хотя бы для одного решающего правила из $S(t,m)$ частоты ошибок на обучающей и рабочей выборках отклонятся более чем на ε, не превосходит

$$P < 4.5\frac{(2l)^d}{d!}\exp(-\varepsilon^2(l-1)). \tag{8}$$

А вероятность того, что найдется решающее правило из $S(t,m)$, безошибочно разделяющее обучающую выборку и ошибающееся на рабочей с частотой, превосходящей ε, меньше

$$P < 1.5\frac{(2l)^d}{d!}\exp\left(-\frac{\varepsilon l}{2}\right). \tag{9}$$

Потребуем, чтобы вероятность $P(\sup_{A\in S}|P(A)-\nu(A)| < \varepsilon)$ не превышала заданное значение $\eta > 0$. Для этого достаточно приравнять к η правые части неравенств (8) и (9) и разрешить их относительно ε. Таким путем получим

$$\varepsilon = 2\sqrt{\frac{d(\ln(2l) - \ln d + 1) - \ln(\eta/5)}{l}}$$

для общего случая и

$$\varepsilon = 2\frac{d(\ln(2l) - \ln d + 1) - \ln(\eta/2)}{l}$$

для детерминистского случая.

Таким образом, с вероятностью $1-\eta$ для всех решающих правил системы $S(t,m)$ справедливы соотношения

$$R_{\text{раб}} \leqslant R_{\text{эмп}} + 2\sqrt{\frac{d(\ln(2l) - \ln d + 1) - \ln(\eta/5)}{l}} \tag{10}$$

при решении задач в общей постановке, и

$$R_{\text{раб}} \leqslant 2\frac{d(\ln(2l) - \ln d + 1) - \ln(\eta/2)}{l} \tag{11}$$

в детерминистской постановке. Здесь $R_{\text{эмп}}$ — доля ошибок на обучающей выборке, а $R_{\text{раб}}$ — доля ошибок на рабочей выборке.

Полученные оценки позволяют построить упорядоченную процедуру минимизации риска. Для детерминистской постановки процедура сводится к тому, чтобы так отнести элементы рабочей выборки к первому или второму классу, чтобы, во-первых, разделение гиперплоскостью векторов этих классов по совокупности обучающей и рабочей выборок было возможно, а, во-вторых, расстояние между их выпуклыми оболочками было максимальным. При этом качество полученного решающего правила оценивается по критерию (11).

В общем случае проводится индексация не только рабочей, но и переиндексация элементов обучающей выборки. При этом количество переиндексированных векторов задает число ошибок на материале обучения. Требуется так индексировать рабочую выборку и переиндексировать обучающую последовательность, чтобы минимизировать критерий (10). Понятно, что полный перебор трудно осуществить, но можно применять те или иные формы сокращения перебора.

Оценки (10) и (11) удалось строго обосновать только в описанной схеме минимизации суммарного риска. Однако как эвристические они могут применяться и в обычной схеме обучения, когда представлена только обучающая выборка. Алгоритмы обобщенного портрета и SVM позволяют одновременно с построением линейного решающего правила (в случае ОП — непосредственно в пространстве признаков, а в случае SVM — в спрямляющем пространстве) находить и расстояние между выпуклыми оболочками классов, представленных в обучающей выборке. Эту величину и можно использовать как оценку ρ, по ней вычислить значения $t = D^2/4\rho^2$ и $d = \min(t+2, m+1)$, где m — размерность пространства, и далее применять формулы (10) или (11) в качестве критериев для оптимизации сложности решающего правила.

Вообще, полученные нами добавки, которые мы предлагаем прибавлять к значению эмпирического риска, чтобы оценить истинный риск, оказываются значительно завышенными. Но это не сильно влияет на положение оптимума при поиске оптимальной сложности. Это можно проверить при большом количестве данных, тестируя результат на чистом экзамене. Более точные критерии обычно связаны с теми или иными предположениями о виде распределения на множестве пар (x, y), и если они не выполняются, то результат оказывается хуже. Приведенные же критерии оказываются более консервативными — они

чаще занижают сложность, чем завышают, но, не будучи связаны никакими дополнительными предположениями, не приводят к большим ошибкам.

Лекция 20

Применение структурной минимизации к задачам восстановления действительных функций. Относительные оценки равномерной близости средних к математическим ожиданиям. Их применение к структурной минимизации риска.

В общем случае проблема уклонения истинного риска от значения эмпирического риска в точке, доставляющей минимум последнему, связана с оценкой равномерного уклонения средних от математических ожиданий по системе случайных величин.

Пусть x — элементарное событие из пространства X, $P(x)$ — вероятностная мера в нем, а α — абстрактный параметр. Пусть далее задан класс функций $F(x, \alpha)$ $(\alpha \in \Omega)$, причем при каждом значении α функция $F(x, \alpha)$ измерима по x, и существует математическое ожидание

$$M(\alpha) = \int F(x, \alpha)\, dP(x).$$

Положим теперь, что задана выборка $X^l = x_1, x_2, \ldots x_l$, полученная в процессе независимых испытаний при неизменном распределении $P(x)$. Для каждого значения по этой выборке можно посчитать среднее арифметическое

$$M_{\text{эмп}}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} F(x_i, \alpha).$$

Если бы параметр α принимал единственное значение, то сходимость среднего к математическому ожиданию обеспечивалась бы законом больших чисел. Но, если параметр α может изменяться в пределах некоторого множества Ω, то возникает вопрос о равномерности по параметру α оценки математического ожидания средним значением. Вероятность того, что среднее значение равномерно приближает математическое ожидание с точностью ε, выражается следующей формулой:

$$P\left\{ \sup_{\alpha \in \Omega} |M(\alpha) - M_{\text{эмп}}(\alpha)| \right\} < \varepsilon.$$

Оценка абсолютного уклонения для равномерно ограниченных функций

Для равномерно ограниченных функций задача оценки этой вероятности легко сводится к рассмотренной ранее задаче оценки равномерной близости частот к вероятностям по классу событий. Класс функций $F(x,\alpha)$ мы считаем равномерно ограниченным, если для всех выполняется неравенство

$$a_1 \leqslant F(x,\alpha) \leqslant a_2,$$

где a_1 и a_2 не зависят от α и x.

Рассмотрим множество S событий A вида

$$A(\alpha,c) = \{x:\ F(x,\alpha) \geqslant c\}$$

для всевозможных значений параметра α и скаляра c, и обозначим $\Delta a = a_2 - a_1$. Тогда оказывается справедливым следующее соотношение:

$$\sup_{\alpha\in\Omega} |M(\alpha) - M_{\text{эмп}}(\alpha)| \leqslant \Delta a \sup_{A\in S} |P(A) - \nu(A)|, \tag{1}$$

где $M_{\text{эмп}}(\alpha)$ — среднее значение функции $F(x,\alpha)$ на выборке $X^l = x_1, x_2, \ldots, x_l$ длины l, а $\nu(A)$ — частота выпадения соответствующего события A на этой выборке. Действительно, по Лебегу, для того, чтобы вычислить интеграл $\int F(x,\alpha)\,dP(x)$, нужно разбить область изменения функции $F(x,\alpha)$ (а она у нас конечна) на n равных интервалов и просуммировать по всем этим интервалам меры множеств $\{x:\ F(x,\alpha) \geqslant a_1 + i\Delta a/n\}$, где i — номер интервала, с весом $\Delta a/n$, а далее устремить значение к бесконечности. Поэтому

$$M(\alpha) = \int F(x,\alpha)\,dP(x) = \lim_{n\to\infty} \frac{\Delta a}{n} \sum_{i=1}^{n} P\left\{F(x,\alpha) \geqslant a_1 + i\frac{\Delta a}{n}\right\}.$$

Аналогично,

$$M_{\text{эмп}}(\alpha) = \sum_{i=1}^{l} F(x_i,\alpha) = \lim_{n\to\infty} \frac{\Delta a}{n} \sum_{i=1}^{n} \nu\left\{F(x,\alpha) \geqslant a_1 + i\frac{\Delta a}{n}\right\}.$$

Обозначим $A_{in} \in S$ события вида $\{F(x,\alpha) \geqslant a_1 + i\Delta a/n\}$. Тогда

$$|M(\alpha) - M_{\text{эмп}}(\alpha)| \leqslant \lim_{n\to\infty} \frac{\Delta a}{n} \sum_{i=1}^{n} |P(A_{in}) - \nu(A_{in})| \leqslant$$
$$\leqslant \Delta a \sup_{A\in S} |P(A) - \nu(A)|,$$

откуда непосредственно следует утверждение (1).

Таким образом, получаем оценку:

$$P\left\{\sup_{\alpha\in\Omega} |M(\alpha) - M_{\text{эмп}}(\alpha)| > \Delta a\varepsilon\right\} \leqslant P\left\{\sup_{A\in S} |P(A) - \nu(A)| > \varepsilon\right\}.$$

Но мы уже знаем из прошлых лекций, что для величины $P\{\sup_{A\in S} |P(A) - \nu(A)|\}$ справедлива оценка

$$P\left\{\sup_{A\in S} |P(A) - \nu(A)| > \varepsilon\right\} < 3M^S(2l)\exp\left(-\frac{\varepsilon^2(l-1)}{4}\right),$$

где $M^S(l)$ — функция роста для системы событий S. Поэтому получаем:

$$P\left\{\sup_{\alpha\in\Omega} |M(\alpha) - M_{\text{эмп}}(\alpha)| > \Delta a\varepsilon\right\} < 3M^S(2l)\exp\left(-\frac{\varepsilon^2(l-1)}{4}\right).$$

Для случая конечной VC-размерности, равной r, справедливо неравенство

$$M^s(l) < 1.5\frac{l^{r-1}}{(r-1)!},$$

откуда

$$P\{\sup_{\alpha\in\Omega} |M(\alpha) - M_{\text{эмп}}(\alpha)| > \Delta a\varepsilon\} < 4.5\frac{(2l)^{r-1}}{(r-1)!}\exp\left(-\frac{\varepsilon^2(l-1)}{4}\right).$$

Теперь, как на прошлой лекции, мы можем задаться некоторой положительной величиной η и, приравняв правую часть неравенства к этой величине, разрешить уравнение

$$4.5\frac{(2l)^{r-1}}{(r-1)!}\exp\left(-\frac{\varepsilon^2(l-1)}{4}\right) = \eta$$

относительно ε. Получим:

$$\varepsilon = 2\sqrt{\frac{r(\ln(2l/r+1) - \ln(\eta/5))}{l-1}}.$$

Отсюда следует, что с вероятностью, превышающей $1-\eta$, значение $M(\alpha)$ не превосходит величину

$$M_{\text{эмп}}(\alpha) + 2\Delta a\sqrt{\frac{r(\ln(2l/r+1) - \ln(\eta/5))}{l-1}}$$

при всех значениях $\alpha \in \Omega$.

Применяя этот результат для значения $\alpha^* \in \Omega$, доставляющего минимум эмпирическому риску, получим

$$R_{\text{ист}}(\alpha^*) \leqslant R_{\text{эмп}}(\alpha^*) + 2\Delta a\sqrt{\frac{r(\ln(2l/r+1) - \ln(\eta/5))}{l-1}}. \qquad (2)$$

Здесь $R_{\text{ист}}(\alpha) = \int Q(x,\alpha)\,dP(x)$, а $R_{\text{эмп}}(\alpha^*) = (1/l)\sum_{i=1}^{l} Q(x_i,\alpha)$, где $Q(x,\alpha)$ — функция штрафа, причем на эту функцию наложено ограничение $a_1 \leqslant Q(x,\alpha) \leqslant a_2$, и $\Delta a = a_2 - a_1$.

Полученная оценка существенно зависит от величины Δa, характеризующей допустимый диапазон изменения функций $Q(x,\alpha)$.

Оценка относительного уклонения

Требование равномерной ограниченности может быть ослаблено. В ряде случаев существенно не абсолютное, а относительное уклонение истинного риска от эмпирического, или же в других терминах — относительное уклонение среднего от математического ожидания. Допустим, что

$$F(x,\alpha) \geqslant 0, \qquad \frac{F(x,\alpha)}{M(\alpha)} \leqslant k, \qquad M(\alpha) > 0,$$

где величина k не зависит ни от α, ни от x. Тогда аналогично приведенному выше рассуждению получим:

$$\sup_{\alpha\in\Omega} \frac{|M(\alpha) - M_{\text{эмп}}(\alpha)|}{M(\alpha)} \leqslant k \sup_{A\in S} |P(A) - \nu(A)|,$$

где система событий S определена как раньше. Тогда

$$P\left\{\sup_{\alpha\in\omega}\frac{|M(\alpha)-M_{\text{эмп}}(\alpha)|}{M(\alpha)}>k\varepsilon\right\}<3M^S(2l)\exp\left(-\varepsilon^2\frac{l-1}{4}\right).$$

Снова воспользовавшись оценкой $M^S(l)<1.5l^{r-1}/(r-1)!$, где r — VC-размерность, получим

$$P\left\{\sup_{\alpha\in\Omega}\frac{|M(\alpha)-M_{\text{эмп}}(\alpha)|}{M(\alpha)}>k\varepsilon\right\}<4.5\frac{(2l)^{r-1}}{(r-1)!}\exp\left(-\frac{\varepsilon^2(l-1)}{4}\right).$$

Опять же, задавшись малой положительной величиной η , приравняв к ней правую часть неравенства и разрешив его относительно ε, получим

$$\varepsilon=2\sqrt{\frac{r(\ln(2l/r+1)-\ln(\eta/5))}{l-1}}.$$

Теперь можно утверждать, что с вероятностью, превышающей $1-\eta$, выполняется неравенство

$$\frac{|M(\alpha)-M_{\text{эмп}}(\alpha)|}{kM(\alpha)}<2\sqrt{\frac{r(\ln(2l/r+1)-\ln(\eta/5))}{l-1}}.$$

Обозначим $\beta=2\sqrt{r(\ln(2l/r+1)-\ln(\eta/5))/(l-1)}$. Тогда

$$\frac{M(\alpha)-M_{\text{эмп}}(\alpha)}{M(\alpha)}<k\beta.$$

Откуда, учитывая, что $M(\alpha)>0$,

$$M(\alpha)(1-k\beta)<M_{\text{эмп}}(\alpha),$$

и можно утверждать, что с вероятностью, превышающей $1-\eta$,

$$M(\alpha)<\frac{M_{\text{эмп}}(\alpha)}{1-k\beta},$$

если, конечно, $(1-k\beta)>0$.

Применим теперь этот результат к задаче оценки истинного риска в точке α^*, доставляющей минимум эмпирическому риску, при следующих допущениях: истинный риск $R(\alpha)=\int Q(x,\alpha)\,dP(x)$ положителен

при всех значениях α. Отношение функции штрафа $Q(x,\alpha)$ к истинному риску ограничено константой k, не зависящей ни от α, ни от x:

$$\frac{Q(x,\alpha)}{R(\alpha)} \leqslant k.$$

Величина $1-\beta = 1-2\sqrt{r(\ln(2l/r+1)-\ln(\eta/5))/(l-1)}$ положительна.

Тогда истинный риск $R(\alpha^*)$ в точке α^*, доставляющей минимум эмпирическому риску, с вероятностью, не меньшей $1-\eta$ оценивается сверху как

$$R(\alpha^*) < \frac{R_{\text{эмп}}(\alpha^*)}{1-k\beta}. \tag{3}$$

Оценка относительного уклонения при ограничении на моменты

В обоих рассмотренных случаях ограничение накладывалось непосредственно на функцию $F(x,\alpha)$. Но на самом деле в задачах регрессии функция штрафа может с некоторой вероятностью принимать сколь угодно большие значения, даже когда ее математическое ожидание (риск) ограничено. И действительно, оценки равномерного относительного уклонения можно получить, если наложить ограничение только на связь математического ожидания и моментов более высокого порядка.

Допустим, что функция $F(x,\alpha)$ неотрицательна, и рассмотрим следующий функционал от ее распределения при любом фиксированном значении параметра α:

$$T(\alpha) = \int \sqrt{1-P\{F(x,\alpha)\leqslant t\}}\,dt.$$

Оказывается, что этот функционал будет заведомо конечным (интеграл сходится), если конечен любой момент распределения выше второго (хотя бы и не целого).

Справедливо следующее утверждение:

$$P\left\{\sup_{\alpha\in\Omega}\frac{M(\alpha)-M_{\text{эмп}}(\alpha)}{T(\alpha)}>\varepsilon\right\} < 8M^S(2l)\exp\left(-\frac{\varepsilon^2(l-1)}{4}\right), \tag{4}$$

где S — система событий $A(\alpha,c)=\{x:\ F(x,\alpha)\geqslant c\}$, а $M^S(l)$ — ее функция роста.

Для доказательства воспользуемся формулой, приведенной на прошлой лекции:

$$P\left\{\sup_{A\in S}\left(P(A)-\frac{\nu(A)}{\sqrt{P(A)}}\right)>\varepsilon\right\}<8M^S(2l)\exp\left(-\frac{\varepsilon^2 l}{4}\right),$$

дающей одностороннюю оценку равномерного относительного уклонения частот от вероятностей по классу событий S. Это неравенство также можно записать в форме

$$P\left\{\sup_{A\in S}(P(A)-\nu(A))>\varepsilon\sqrt{P(A)}\right\}<8M^S(2l)\exp\left(-\frac{\varepsilon^2 l}{4}\right). \qquad (5)$$

Опять воспользуемся определением интеграла по Лебегу:

$$M(\alpha)=\int_0^t F(x,\alpha)\,dP(x)=$$
$$=\lim_{t\to\infty}\lim_{n\to\infty}\frac{t}{n}\sum_{i=1}^{n}P\left\{F(x,\alpha)>\frac{it}{n}\right\}.$$

Аналогично,

$$M_{\text{эмп}}(\alpha)=\sum_{i=1}^{l}F(x_i,\alpha)=\lim_{t\to\infty}\lim_{n\to\infty}\frac{t}{n}\sum_{i=1}^{n}\nu\left\{F(x,\alpha)>\frac{it}{n}\right\}.$$

Вычитая второе равенство из первого, получим

$$M(\alpha)-M_{\text{эмп}}(\alpha)=\lim_{t\to\infty}\lim_{n\to\infty}\frac{t}{n}\sum_{i=1}^{n}\left[P\left\{F(x,\alpha)\geqslant\frac{it}{n}\right\}\right.$$
$$\left.-\nu\left\{F(x,\alpha)>\frac{it}{n}\right\}\right].$$

Но, если выполнено неравенство

$$\sup_{A\in S}[P(A)-\nu(A)]\leqslant\varepsilon\sqrt{P(A)}, \qquad (6)$$

то применяя его к событиям вида $F(x,\alpha) \geqslant it/n$, получим для всех α

$$M(\alpha) - M_{\text{эмп}}(\alpha) \leqslant \lim_{t\to\infty}\lim_{n\to\infty} \varepsilon\frac{t}{n}\sum_{i=1}^{n}\sqrt{P\left(F(x,\alpha) > \frac{it}{n}\right)} = \\ = \varepsilon\int\sqrt{P(F(x,\alpha) > t)}\,dx.$$

В свою очередь

$$\int\sqrt{P(F(x,\alpha) > t)}\,dx = \int\sqrt{1 - P\{F(x,\alpha) \leqslant t\}}\,dt = T(\alpha).$$

Поэтому, если выполнено неравенство (6), то

$$\sup_{\alpha\in\Omega}(M(\alpha) - M_{\text{эмп}}(\alpha)) \leqslant \varepsilon T(\alpha). \tag{7}$$

Но вероятность выполнения неравенства, противоположного (6) оценивается формулой (5). Поэтому получаем

$$P\left\{\sup_{\alpha\in\Omega}\frac{(M(\alpha) - M_{\text{эмп}}(\alpha)}{T(\alpha)} > \varepsilon\right\} < 8M^S(2l)\exp\left(-\frac{\varepsilon^2 l}{4}\right). \tag{8}$$

Теперь, приравнивая правую часть неравенства (8) к заданной малой величине η и разрешая это равенство относительно ε, можно получить оценки, подобные приведенным выше.

Но хотелось бы функционал $T(\alpha)$ заменить на более привычные моменты распределения. Оказывается, что значение $T(\alpha)$ можно оценить сверху, если известен какой-либо момент распределения случайной величины $F(x,\alpha)$ выше второго. Обозначим его $M_p(\alpha)$, где p — порядок момента. Если момент $M_p(\alpha)$ порядка $p > 2$ существует, то справедливо неравенство

$$T(\alpha) \leqslant s(p)(M_p(\alpha))^{1/p}, \qquad s(p) = \left[\frac{(p-1)^{p-1}}{2(p-2)^{p-1}}\right]^{1/p}. \tag{9}$$

Этот результат может быть получен путем решения методом множителей Лагранжа такой задачи: найти распределение, при котором достигается максимум функционала $T(\alpha)$ при заданном значении $M_p(\alpha)$. Такое распределение находится, и для него неравенство (9) переходит

в равенство. Для случая, когда существует или известен только второй момент $M_2(\alpha)$, справедлива оценка, которую мы приведем без доказательства:

$$P\left\{\sup_{\alpha\in\Omega}\frac{M(\alpha)-M_{\text{эмп}}(\alpha)}{\sqrt{M_2(\alpha)}}>\varepsilon\sqrt{1-\frac{1}{2}\ln\varepsilon}\right\}<8M^S(2l)\exp\left(-\frac{\varepsilon^2 l}{4}\right).$$

Вернемся, однако, к оценке (8), где уклонение математического ожидания от среднего нормируется на функционал $T(\alpha)$. Заметим, что величина $T(\alpha)$ размерная и ее размерность совпадает с размерностью $M(\alpha)$. Допустим, что для всех значений α выполняется соотношение

$$\frac{T(\alpha)}{M(\alpha)}\leqslant r,$$

где r — некоторая фиксированная положительная константа, не зависящая от α. Тогда из неравенства (8) следует, что

$$P\left\{\sup_{\alpha\in\Omega}\frac{M(\alpha)-M_{\text{эмп}}(\alpha)}{rM(\alpha)}>\varepsilon\right\}<8M^S(2l)\exp\left(-\frac{\varepsilon^2 l}{4}\right). \qquad (10)$$

Опять, приравнивая правую часть этого неравенства к некоторой малой положительной величине η и разрешая полученное равенство относительно ε, получим

$$\varepsilon=2\sqrt{\frac{\ln M^S(2l)-\ln(\eta/8)}{l}}. \qquad (11)$$

Теперь мы можем утверждать, что при наших условиях с вероятностью, превышающей $1-\eta$, для всех $\alpha\in\Omega$ выполняется неравенство

$$\frac{M(\alpha)-M_{\text{эмп}}(\alpha)}{rM(\alpha)}\leqslant\varepsilon,$$

где ε задается равенством (11). Отсюда

$$M(\alpha)\leqslant\frac{M_{\text{эмп}}(\alpha)}{1-r\varepsilon},$$

если только величина $(1-r\varepsilon)$ положительна.

Для конечных значений n VC-размерности функция роста оценивается как

$$M^S(l) < 1,5\frac{l^n}{n!} \approx \left(\frac{le}{n}\right)^n.$$

Подставляя эту оценку в (11), получим

$$\varepsilon = 2\sqrt{\frac{n\ln(2l/(n+1)) - \ln(\eta/8)}{l}}.$$

Откуда с вероятностью, превышающей $1-\eta$, для всех $\alpha \in \Omega$

$$M(\alpha) \leqslant \frac{M_{\text{эмп}}(\alpha)}{1 - 2r\sqrt{\dfrac{n\ln(2l/(n+1)) - \ln(\eta/8)}{l}}}. \tag{12}$$

Применим теперь это соотношение для оценки истинного риска в точке α^*, доставляющей минимум эмпирическому риску. Поскольку неравенство (12) выполняется для всех $\alpha \in \Omega$, то, в частности, оно выполняется и для α^*. Поэтому

$$R(\alpha^*) \leqslant \frac{R_{\text{эмп}}(\alpha^*)}{1 - 2r\sqrt{\dfrac{n\ln(2l/(n+1)) - \ln(\eta/8)}{l}}}. \tag{13}$$

Конечно, при этом должны выполнятся для всех $\alpha \in \Omega$ условия $Q(x,\alpha) \geqslant 0$, где $Q(x,\alpha)$ — функция штрафа, и $T(\alpha)/R(\alpha) \leqslant r$, где r — фиксированная константа, а

$$T(\alpha) = \int \sqrt{1 - P\{Q(x,\alpha) \leqslant t\}}\, dt.$$

Последнее условие выражает отсутствие так называемых «тяжелых хвостов» в распределении функции штрафа при всех значениях α. В той или иной форме оно все равно должно быть наложено.

Если величина в знаменателе правой части выражения (13) становится отрицательной, мы считаем, что оценка истинного риска уходит в плюс бесконечность.

Допустим теперь, что классы возможных решений упорядочены по сложности, и в качестве критерия сложности принята VC-размерность. Тогда оценка (13) дает гарантированную с вероятностью $1-\eta$ оценку сверху истинного риска точке α^*, доставляющей минимум эмпирическому риску. Значение $R_{\text{эмп}}(\alpha^*)$ убывает с

ростом сложности, тогда как мультипликативная добавка $1/(1 - 2r\sqrt{(n\ln(2l/(n+1)) - \ln(\eta/8))/l}$ растет, и это позволяет искать оптимальную сложность модели, доставляющую минимум гарантированной оценке истинного риска.

Лекция 21

Комбинированный подход: максимум правдоподобия — Байес.

Итак, мы видели, что применение метода Байеса при выборе оптимальной структуры модели связано с заданием априорного распределения параметров модели. Даже в тех случаях, когда удавалось значительно упростить априорную информацию, решение все равно существенно зависело от ряда параметров: соотношение сигнал/шум, число членов разложения искомой функции по базисным функциям и/или порядок убывания априорной дисперсии коэффициентов разложения и т.п.

С другой стороны, оценку истинного риска для модели, выбранной по критерию минимума эмпирического риска, удается дать только сверху и притом очень грубо. Использование независимого экзамена, скользящего контроля и т.п. требует большого количества данных с известным выходным значением. Поэтому возникает желание, хотя бы в частных случаях, получить более точные и экономные критерии выбора оптимальной структуры модели.

В то же время мы видели на одной из первых наших лекций, что метод максимального правдоподобия хорошо работает при условии, что число настраиваемых параметров много меньше длины выборки, по которой они настраиваются. Поэтому, возникает идея соединить метод максимального правдоподобия и метод Байеса следующим образом. Допустим, что оптимальное байесово решение зависит от небольшого числа параметров априорного распределения. Построим функцию правдоподобия для этих параметров и найдем их значение, доставляющее максимум этой функции. Теперь подставим это значение в байесово решение и получим модель. Эта идея, видимо, впервые была предложена в 1970 году в работе В.Ф. Турчина, В.П. Козлова и М.С. Малкевича «Использование методов математической статистики для решения некорректных задач» [19]. Здесь дисперсия шума предполагалась известной, а дисперсия априори нормально распределенных

коэффициентов разложения считалась неизвестным параметром априорного распределения. Затем, уже в 90-е годы этот подход был широко развит в ряде работ Давида МакКая и в некоторых наших работах [26, 22].

Формально этот путь можно сформулировать так. Пусть модель описывается вектором параметров $\boldsymbol{a}$. Вектор наблюдений x связан с вектором параметров плотностью условного распределения $p(\mathbf{x}|\boldsymbol{a})$ и задана априорная плотность распределения $p_{\text{апр}}(\boldsymbol{a})$. Тогда безусловная плотность распределения вектора наблюдений $\mathbf{x}$ будет равна

$$p(\mathbf{x}) = \int p(\mathbf{x}|\boldsymbol{a}) p_{\text{апр}}(\boldsymbol{a})\, d\boldsymbol{a},$$

где интеграл берется по всему множеству возможных значений вектора параметров $\boldsymbol{a}$.

По формуле Байеса апостериорная плотность распределение вектора параметров $\boldsymbol{a}$ будет равна

$$p_{\text{апост}}(\boldsymbol{a}) = \frac{p(\mathbf{x}|\boldsymbol{a}) p_{\text{апр}}(\boldsymbol{a})}{p(\mathbf{x})}.$$

Пусть теперь плотность $p(\mathbf{x}|\boldsymbol{a})$ известна нам с точностью до (абстрактного) параметра α:

$$p(\mathbf{x}|\boldsymbol{a}) = p(\mathbf{x}|\boldsymbol{a}, \alpha),$$

а априорная плотность $p_{\text{апр}}(\boldsymbol{a})$ — с точностью до параметра β:

$$p_{\text{апр}}(\boldsymbol{a}) = p_{\text{апр}}(\boldsymbol{a}, \beta).$$

Тогда и безусловная плотность $p(\mathbf{x})$ будет зависеть от этих параметров:

$$p(\mathbf{x}; \alpha, \beta) = \int p(\mathbf{x}|\boldsymbol{a}, \alpha) p_{\text{апр}}(\boldsymbol{a}, \beta)\, d\boldsymbol{a}. \tag{1}$$

От них же будет зависеть и апостериорное распределение вектора $\boldsymbol{a}$:

$$P_{\text{апост}}(\boldsymbol{a}; \alpha, \beta) = \frac{p(\mathbf{x}|\boldsymbol{a}, \alpha) p_{\text{апр}}(\boldsymbol{a}, \beta)}{p(\mathbf{x})}. \tag{2}$$

Но функция правдоподобия для параметров α и β при заданном векторе наблюдений $\mathbf{x}$ есть

$$W(\alpha, \beta) = \log p(\mathbf{x}; \alpha, \beta) = \log \left[\int p(\mathbf{x}|\boldsymbol{a}, \alpha) p_{\text{апр}}(\boldsymbol{a}, \beta)\, d\boldsymbol{a} \right].$$

Теперь можно найти наиболее правдоподобное значение (α^*, β^*) этих параметров, то есть такое значение, которое доставляет максимум функции правдоподобия:

$$(\alpha^*, \beta^*) = \arg\max W(\alpha, \beta).$$

Подставляя значения α^* и β^* в (2), мы получим наиболее правдоподобное апостериорное распределение вектора параметров $\boldsymbol{a}$.

Если число переменных, описывающих параметры α и β, невелико, или они (один из них) принимает небольшое число дискретных значений, то можно надеяться, что метод максимума правдоподобия даст результат, близкий к истинному. В то же время вектор параметров $\boldsymbol{a}$ может иметь большую или даже бесконечную размерность — байесова стратегия приведет к необходимой регуляризации.

Применим предложенный подход к задаче восстановления регрессии путем оценки коэффициентов разложения искомой функции по заданной системе базисных функций. Итак, пусть ищется числовая зависимость, определенная на некотором пространстве X. Допустим, что эта зависимость имеет вид

$$\sum_{i=1}^{n} a_i \varphi_i(x),$$

где $\varphi_i(x)$ — известная система базовых функций, a_i — неизвестные коэффициенты. Эта зависимость наблюдается (измеряется) с независимой аддитивной помехой ξ, имеющей нулевое среднее, фиксированную (известную) дисперсию D_ξ и распределенной по нормальному закону $N(0, D_\xi)$.

Тогда наблюдаемая величина будет иметь вид

$$c(x) = \sum_{i=1}^{n} a_i \varphi_i(x) + \xi.$$

Пусть теперь дана обучающая выборка длины l — в точках $x_1, x_2, \ldots, x_l$ измерены (наблюдались) значения $c(x_1), c(x_2), \ldots, c(x_l)$. Обозначим теперь вектор измеренных значений c в точках $x_1, x_2, \ldots, x_l$ как

$$\mathbf{c} = (c(x_1), c(x_2), \ldots c(x_l)),$$

а матрицу значений функций $\varphi_i(x)$ в этих точках — как $\mathbf{F}$ с элементами

$$f_{ij} = \varphi_i(x_j).$$

Найдем теперь $p(\mathbf{c}|\boldsymbol{a})$ — плотность условного распределения вектора $\mathbf{c}$ при заданных значениях вектора $\boldsymbol{a}$ и матрицы $\mathbf{F}$. Обозначим $\mathbf{c}_0 = \mathbf{F}^{\mathrm{T}}\boldsymbol{a}$ вектор с координатами $c_{0j} = \sum_{i=1}^{n} a_i\varphi_i(x_j)$, где j меняется от 1 до l. Тогда

$$p(\mathbf{c}|\boldsymbol{a}) = \left((2\pi D_\xi)^l\right)^{-1/2} \exp\left(-\frac{1}{2D_\xi}(\mathbf{c}-\mathbf{c}_0)(\mathbf{c}-\mathbf{c}_0)^{\mathrm{T}}\right) =$$
$$= \left((2\pi D_\xi)^l\right)^{-1/2} \exp\left(-\frac{1}{2D_\xi}(\mathbf{c}-\mathbf{F}^{\mathrm{T}}\boldsymbol{a})(c-\mathbf{F}^{\mathrm{T}}\boldsymbol{a})^{\mathrm{T}}\right).$$

Теперь для того, чтобы применить байесов подход, необходимо задать априорное распределение на множестве коэффициентов a_i. Допустим, что это априорное распределение n-мерного вектора $\boldsymbol{a} = (a_1, a_2, \ldots, a_n)$ также является нормальным с нулевым средним и ковариационной матрицей $\mathbf{R}$, т. е.

$$P_{\text{апр}}(\boldsymbol{a}) = N(0, \mathbf{R}).$$

Соответствующая плотность распределения будет равна

$$p_{\text{апр}}(\boldsymbol{a}) = \left((2\pi)^n \det(\mathbf{R})\right)^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{a}^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{a}\right).$$

Как мы установили в лекции 12, при квадратичном критерии невязки оптимальной оценкой вектора коэффициентов разложения будет апостериорное математическое ожидание

$$M_{\text{апост}}(\boldsymbol{a}) = (\mathbf{F}\mathbf{F}^{\mathrm{T}} + D_\xi\mathbf{R}^{-1})^{-1}\mathbf{F}\mathbf{c}. \tag{3}$$

Этот результат зависит от дисперсии шума D_ξ и от априорной ковариационной матрицы коэффициентов $\mathbf{R}$, которые в общем случае неизвестно откуда взять. Но согласно предлагаемой методике мы должны построить функцию правдоподобия $W(D_\xi, \mathbf{R})$, зависящую от этих факторов, найти значения D_ξ и $\mathbf{R}$, доставляющие максимум этой функции, и использовать их в выражении (3). Для этого нужно найти безусловную плотность распределения вероятности получить вектор наблюдений $\mathbf{c}$ при заданных значениях D_ξ и $\mathbf{R}$, которая будет равна

$$p(\mathbf{c}; D_\xi, \mathbf{R}) = \int p(\mathbf{c}|\boldsymbol{a})p_{\text{апр}}(\boldsymbol{a})\,d\boldsymbol{a}.$$

В свою очередь

$$p(\mathbf{c}|\boldsymbol{a})p_{\text{апр}}\,(\boldsymbol{a}) = \frac{\exp\left(-0.5(1/D_\xi)(\mathbf{c}-\mathbf{F}^{\mathrm{T}}\boldsymbol{a})(\mathbf{c}-\mathbf{F}^{\mathrm{T}}\boldsymbol{a})^{\mathrm{T}} + \boldsymbol{a}^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{a}\right)}{\sqrt{(2\pi)^{l+n}\det\mathbf{R}\cdot D_\xi^l}}.$$

После интегрирования этого выражения по всему пространству значений вектора $\boldsymbol{a}$ получается:

$$p(\mathbf{c};D_\xi,\mathbf{R}) = \frac{\exp\left(-0.5(1/D_\xi)(\mathbf{c}-\mathbf{F}^{\mathrm{T}}\boldsymbol{a}^*)(\mathbf{c}-\mathbf{F}^{\mathrm{T}}\boldsymbol{a}^*)^{\mathrm{T}} + (\boldsymbol{a}^*)^{\mathrm{T}}\mathbf{R}^{-1}\boldsymbol{a}^*\right)}{\sqrt{(2\pi)^n\det(\mathbf{R}\mathbf{F}^{\mathrm{T}}\mathbf{F}+D_\xi\mathbf{E})D_\xi^{l-n}}},$$

где $a^* = (\mathbf{F}\mathbf{F}^{\mathrm{T}} + D_\xi\mathbf{R}^{-1})^{-1}\mathbf{F}\mathbf{c}$ — апостериорное математическое ожидание вектора $\boldsymbol{a}$, а $\mathbf{E}$ — единичная матрица $(n\times n)$.

Теперь функция правдоподобия просто получается логарифмированием выражения (4)

$$W(D_\xi,\mathbf{R}) = \ln p(\mathbf{c};D_\xi,\mathbf{R}),$$

где $\mathbf{c}$ — вектор измеренных значений выходной величины.

Однако в общем случае мы не получили ничего хорошего. Вместо поиска n значений коэффициентов $\boldsymbol{a}$ нам приходится искать $n^2/2$ элементов матрицы $\mathbf{R}$, да еще значение D_ξ. Поэтому приходится упростить задачу. Примем, что коэффициенты разложения априори независимы и имеют одинаковую дисперсию D. Тогда ковариационная матрица априорного распределения будет равна $\mathbf{R} = D\mathbf{E}$, а функция правдоподобия будет зависеть только от двух скалярных переменных D_ξ и D:

$$W(D_\xi,D) = -\frac{1}{2}\Bigg(n\ln(2\pi) + \ln\det(D\mathbf{F}^{\mathrm{T}}\mathbf{F}+D_\xi\mathbf{E}) + (l-n)\ln D_\xi + \\ + \frac{1}{D_\xi}(\mathbf{c}-\mathbf{F}^{\mathrm{T}}\boldsymbol{a}^*)(\mathbf{c}-\mathbf{F}^{\mathrm{T}}\boldsymbol{a}^*)^{\mathrm{T}} + \frac{1}{D}(a^*)^{\mathrm{T}}\boldsymbol{a}^*\Bigg), \qquad (5)$$

где $\boldsymbol{a}^* = (\mathbf{F}\mathbf{F}^{\mathrm{T}} + (D_\xi/D)\mathbf{E})^{-1}\mathbf{F}\mathbf{c}$. Параметр $t = D_\xi/D$ служит здесь коэффициентом регуляризации. Однако сама функция правдоподобия зависит от двух переменных D_ξ и D, а не только от их отношения.

Поскольку в этом частном случае метод максимума правдоподобия применяется для поиска всего двух неизвестных параметров, можно надеяться, что уже при сравнительно небольшой длине выборки эти параметры будут оценены достаточно точно.

Формально теперь нужно перебрать по некоторой сети всевозможные разумные значения пар D_ξ и D, вычислить для каждой пары по формуле (5) значение функции правдоподобия и взять ту пару, для которой это значение будет максимальным. Для нее получим значение коэффициента регуляризации, и с его помощью найдем оптимальную по Байесу оценку коэффициентов разложения a^*. Но с вычислительной точки зрения этот путь не эффективен, так как для каждой пары пришлось бы обращать матрицу (или решать систему уравнений) и вычислять детерминант. Более эффективным оказывается путь, связанный с переходом к новому базису — ортонормированному базису собственных векторов матрицы $\mathbf{F}\mathbf{F}^T$. Дело в том, что собственные векторы матрицы $(D\mathbf{F}\mathbf{F}^{\mathrm{T}} + D_\xi\mathbf{E})$, как нетрудно убедиться, совпадают с собственными векторами матрицы $\mathbf{F}\mathbf{F}^{\mathrm{T}}$. Поэтому операцию нахождения собственных векторов можно выполнить только один раз.

Итак, пусть $(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \ldots, \boldsymbol{\varphi}_k)$ — ортонормированный базис собственных векторов положительно полуопределенной симметрической матрицы $\mathbf{F}\mathbf{F}^{\mathrm{T}}$ размерности $(l \times l)$, а $(\lambda_1, \lambda_2, \ldots, \lambda_k)$ — соответствующие собственные числа. Собственные векторы симметрической матрицы заведомо будут ортогональны между собой, а собственные числа — неотрицательны. Но мы считаем, что все $\lambda_i > 0$ — векторы, соответствующие нулевым собственным числам, просто не включаем в базис. Тогда число k векторов в базисе будет равно рангу матрицы $\mathbf{F}\mathbf{F}^{\mathrm{T}}$. Поэтому $k \leqslant l$ и $k \leqslant n$, где l — длина выборки, а n —число используемых базисных функций. Таким образом, этот базис оснащает подпространство размерности k в пространстве наблюдений размерности l. Если размерность k меньше, чем l, то вектор наблюдений нельзя полностью представить в новом базисе, но можно представить его проекцию на соответствующее подпространство. Координаты C_i этой проекции будут равны

$$C_i = (\boldsymbol{\varphi}_i, \mathbf{c}),$$

а вектор $\mathbf{C} = \sum C_i\boldsymbol{\varphi}_i$ выражает эту проекцию в исходном базисе. Тогда величина

$$C^2_{\mathrm{res}} = \|\mathbf{c} - \mathbf{C}\|^2$$

будет равна квадрату остаточной невязки метода наименьших квадратов. Если же $n \geqslant l$ и матрица $\mathbf{F}\mathbf{F}^{\mathrm{T}}$ невырождена (т.е $k = l$), то $C^2_{\mathrm{res}} = 0$.

Тогда формула (5) после несложных преобразований выражается

через введенные величины следующим образом:

$$W(D_\xi, D) = -\frac{1}{2}\left(l\ln(2\pi) + (l-k)\ln D_\xi + \frac{C_{\text{res}}^2}{D_\xi} + \right.$$
$$\left. + \sum_{i=1}^{k}\left(\frac{C_i^2}{D\lambda_i + D_\xi} + \ln(D\lambda_i + D_\xi) \right) \right). \qquad (6)$$

Таким образом, значения функции правдоподобия можно вычислять по аналитической формуле, не прибегая к многократному обращению матрицы и вычислению детерминанта, но после однократного нахождения системы собственных векторов матрицы $\mathbf{F}\mathbf{F}^{\mathrm{T}}$.

Более того, если от переменных D и D_ξ перейти к переменным $t = D_\xi/D$ и D_ξ, то максимум функции $W(D_\xi, t)$ по D_ξ при фиксированном значении t удается найти аналитически. Действительно

$$W(D_\xi, t) = -\frac{1}{2}\left(l\ln(2\pi) + (l-k)\ln D_\xi + \frac{C_{\text{res}}^2}{D_\xi} + \right.$$
$$\left. + \sum_{i=1}^{k}\left[\frac{C_i^2}{D_\xi\big((\lambda_i/t)+1\big)} + \ln\left(D_\xi\left(\frac{\lambda_i}{t}+1\right)\right)\right] \right). \qquad (7)$$

Приравнивая к нулю производную от этого выражения по D_ξ при фиксированном значении t, получим

$$D_{\xi\,\text{опт}} = \frac{1}{l}\left(C_{\text{res}}^2 + \sum_{i=1}^{k}\frac{C_i^2}{((\lambda_i/t)+1)} \right). \qquad (8)$$

Таким образом, задача сводится к перебору всего по одной переменной — параметру регуляризации t.

На качественном уровне, варьируя значения D_ξ и D, мы меняем априорные дисперсии величин C_i и стараемся подогнать их так, чтобы дать наиболее правдоподобное объяснение фактически наблюдавшимся значениям. Дело в том, что дисперсия шума равномерно распределяется по степеням свободы, тогда как дисперсия сигнала в нашем случае пропорциональна собственным значениям λ_i. Если бы все собственные числа были одинаковы, то сигнал невозможно было бы отличить от шума, но, как правило, это не так, и максимум функции правдоподобия оказывается довольно острым. В то же время мы ограничены

всего двумя переменными, и поэтому чрезмерная подгонка (overfitting) оказывается невозможной.

Интересно отметить, что хотя мы стремимся найти истинные значения D_ξ и D, но регуляризация окончательного решения все равно остается. А это приводит к огрублению модели, степень которой как раз и зависит от соотношения D_ξ/D.

Может показаться, что условие равенства априорных дисперсий D_i у коэффициентов разложения является слишком тяжелым. Действительно, более естественно предположить, что они как-то убывают с ростом номера базисной функции. Но если бы мы искали методом максимума правдоподобия все n дисперсий, то опять бы столкнулись с той же проблемой нехватки данных. Вместо этого можно предположить, что они убывают по какому-то закону вида $D_i = Df(i)$, например,

1. $D_i = D\exp(-di)$,
2. $D_i = D\exp(-di^2)$,
3. $D_i = Di^{-d}$,

где d — дополнительный параметр (но всего один), характеризующий убывание дисперсии коэффициентов разложения в зависимости от номера базисной функции.

Если закон убывания дисперсии и его параметр фиксированы, то есть если задана функция $f(i)$, то задача легко сводится к рассмотренной выше. Для этого достаточно заменить исходные базисные функции $\varphi_i(x)$ на

$$\varphi_i^*(x) = \sqrt{f(i)}\varphi_i(x).$$

Теперь априорная дисперсия коэффициентов при новых базисных функциях будет одинаковой и равной D.

Процедура

Теперь процедура выбора параметров модели выглядит так.

1. Выбрать систему базисных функций $\varphi_i(x)$, считая, что коэффициенты при них априори независимы, а их дисперсия убывает по закону $D_i = Df(i)$, где $f(i)$ — заданная функция, а D — неизвестный параметр.

2. Пронормировать базисные функции, заменив их на $\varphi_i^*(x) = \sqrt{f(i)}\varphi_i(x)$.

3. Построить матрицу $\mathbf{F}$ с элементами $f_{ij} = \varphi_i^*(x_j)$, где x_j — описание объектов обучающей выборки.

4. Найти собственные (нормированные) векторы $\boldsymbol{\varphi}_i$ и собственные числа λ_i матрицы $\mathbf{F}\mathbf{F}^{\mathrm{T}}$, оставив из них только те, которым соответствуют положительные значения собственных чисел (всего k векторов).

5. Найти значения $C_i = (\boldsymbol{\varphi}_i, \mathbf{c})$ и $C_{\mathrm{res}}^2 = \|\mathbf{c} - \mathbf{C}\|^2$, где $\mathbf{c} = (c(x_1), c(x_2), \ldots, c(x_l))$ — вектор наблюдений в точках обучающей выборки, а $C = \sum C_i \boldsymbol{\varphi}_i$.

6. Перебирая с некоторым шагом значения параметра регуляризации t, найти для каждого значения t значение D_ξ по формуле (8) и значение функции правдоподобия по формуле (7).

7. Найти значение t и соответствующее значение D_ξ, для которых достигается максимум функции правдоподобия и запомнить значение этого максимума.

8. Если теперь мы хотим попробовать другую систему базисных функций, другой закон убывания априорных дисперсий или его параметры, или хотим изменить число используемых базисных функций, то нужно вернуться к пункту 1 и заново повторить операции 1-7.

9. Выбрать тот вариант, для которого достигается максимум максиморум значения функции правдоподобия.

10. Найти оценку коэффициентов регрессии для этого варианта, используя в качестве коэффициента регуляризации соответствующее оптимальное значение параметра t.

Выводы

Важно отметить, что сравнение различных вариантов по критерию максимума функции правдоподобия допустимо. Дело в том, что функ-

ция правдоподобия

$$W(\alpha,\beta) = \log p(\mathbf{x};\alpha,\beta) = \log\left[\int p(\mathbf{x}|\boldsymbol{a},\alpha)p_{\text{апр}}(\boldsymbol{a},\beta)\,d\boldsymbol{a}\right]$$

оценивает плотность распределения в пространстве X, которая не связана с природой параметров α и β. Выбираются те значения этих параметров, для которых эта плотность в точке $\mathbf{x}$, соответствующей экспериментальным данным, достигает максимума. Конечно, перебор по множеству значений этих параметров не должен быть большим, так как в противном случае метод максимума правдоподобия приводит к значительным ошибкам. В частности, в нашем примере метод был бы более эффективным, если бы мы заранее знали дисперсию шума D_ξ, закон и параметр убывания априорной дисперсии коэффициентов. (Именно этот случай рассматривался в пионерской работе Турчина и др.) Но, к сожалению, так бывает не всегда.

Может показаться, что более эффективно было бы и при оценке влияния параметров α и β оставаться в рамках байесовой стратегии. Можно было бы задаться априорным распределением $p_{\text{апр}}(\alpha,\beta)$ этих параметров, на основании выборки построить их апостерирорное распределение $p_{\text{апост}}(\alpha,\beta) = (1/q)p_{\text{апр}}(\alpha,\beta)p(\mathbf{x};\alpha,\beta)$ и далее провести соответствующее усреднение. Но если распределение $p_{\text{апр}}(\alpha,\beta)$ достаточно размыто, что естественно, когда мы заранее не знаем параметры α и β, и функция $p(\mathbf{x};\alpha,\beta)$ также размыта, то это значит, что эти параметры невозможно определить по выборке $\mathbf{x}$. Если же функция $p(\mathbf{x};\alpha,\beta)$ сосредоточена в некоторой узкой области значений этих параметров, то метод максимума правдоподобия и метод Байеса дают близкий результат, и можно не мучиться с проблемами интегрирования и выбора конкретного априорного распределения $p_{\text{апр}}(\alpha,\beta)$.

Предлагаемый подход дает сравнительно точную оценку качества структуры модели, но связан (как и все методы, опирающиеся на байесову стратегию) с априорными представлениями о распределении помехи, виде априорного распределения параметров модели и т.п. Если эти предположения действительно выполняются, то метод работает хорошо, в противном же случае ошибки могут быть очень большими.

Лекция 22

Применение метода максимума правдоподобия при восстановлении зависимости методом кригинга. Информационный критерий Акаике.

Применение метода максимума правдоподобия при восстановлении зависимости методом кригинга

В методе кригинга, изложенном нами ранее, параметрами метода были отношение дисперсии некоррелированной составляющей к дисперсии регулярной составляющей (эффект самородков) и параметры корреляционной функции $R(\mathbf{x}, \mathbf{y})$.

Зависимость корреляционной функции от параметров α запишем как $R(\mathbf{x}, \mathbf{y}; \alpha)$. Тогда плотность совместного распределения значений поля $u(x_i)$ в точках $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l$ можно в гауссовом случае записать как

$$p(u_1, u_2, \ldots, u_l; \alpha) = \frac{\exp\left(-\frac{1}{2}\mathbf{u}\mathbf{K}^{-1}(\alpha)\mathbf{u}^{\mathrm{T}}\right)}{\sqrt{(2\pi)^l \det \mathbf{K}(\alpha)}},$$

где $\mathbf{K}$ — матрица с элементами $k_{ij} = R(\mathbf{x}_i, \mathbf{x}_j; \alpha)$, а вектор $\mathbf{u} = (u_1, u_2, \ldots, u_l)$. Тогда функция правдоподобия с точностью до постоянных членов примет вид

$$W(\alpha) = -\frac{1}{2}\big(\log\det \mathbf{K}(\alpha) + \mathbf{u}\mathbf{K}^{-1}(\alpha)\mathbf{u}^{\mathrm{T}}\big).$$

Но непосредственное вычисление этой функции затруднительно. Представим, однако, плотность $p(u_1, u_2, \ldots, u_l; \alpha)$ в последовательной форме:

$$\begin{aligned} p(u_1, u_2, \ldots, u_l; \alpha) = p(u_1; \alpha)p(u_2|u_1; \alpha)p(u_3|u_1, u_2; \alpha)\ldots \times \\ \times\, p(u_l|u_1, u_2, \ldots, u_{l-1}; \alpha). \end{aligned}$$

Тогда функция правдоподобия примет вид

$$\begin{aligned} W(\alpha) = -\frac{1}{2}\big(\log p(u_1; \alpha) + \log p(u_2|u_1; \alpha) + \ldots + \\ + \log p(u_l|u_1, u_2, \ldots, u_{l-1}; \alpha)\big). \end{aligned}$$

Поскольку условные распределения в нашем случае тоже будут гауссовыми, то они полностью характеризуются своим математическим ожиданием и дисперсией. А формулы для них мы получили, когда исследовали метод кригинга.

Разумеется, при вычислении условных вероятностей $p(u_i|u_1, u_2, \ldots, u_{i-1}; \alpha)$ можно учитывать не все точки $\mathbf{x}_1$, $\mathbf{x}_2, \ldots, \mathbf{x}_{i-1}$, а только те из них, которые находятся в некоторой окрестности точки $\mathbf{x}_i$.

Здесь не видно другого пути, кроме перебора возможных значений эффекта самородков и параметров корреляционной функции для регулярной составляющей. Для каждого значения вычисляется функция правдоподобия $W(\alpha)$ и выбирается то значение, для которого эта функция достигает максимума.

Помимо эффекта самородков наиболее интересными параметрам метода кригинга оказываются оси анизотропии и радиусы корреляции вдоль них. Эти оси соответствуют направлениям максимальной, минимальной и промежуточной изменчивости поля. К сожалению, применение метода максимального правдоподобия здесь ограничено случаями поля в пространстве размерности 2–3. При более высокой размерности число параметров становится слишком большим и метод максимального правдоподобия может приводить к существенным ошибкам, да и с вычислительной точки зрения перебор становится невыполнимым. В случае большой размерности приходится ограничиться оценкой эффекта самородков и радиуса корреляции без учета анизотропии поля или пользоваться другими средствами.

Информационный критерий Акаике

Рассмотрим еще один критерий, использующий метод максимального правдоподобия для выбора оптимальной структуры (сложности) модели [17, 3, 16]. Допустим, что каждая структура модели задает ее с точностью до n неизвестных параметров $(a_1, a_2, \ldots, a_n)$. Число может меняться при изменении структуры. Для каждой такой структуры по экспериментальным данным удается построить функцию правдоподобия $W(a_1, a_2, \ldots, a_n)$, найти вектор параметров $\boldsymbol{a}^*$, на котором достигается максимум функции правдоподобия, и значение функции $W(a_1, a_2, \ldots, a_n)$ в точке максимума. Напомним, что функцией прав-

доподобия мы считаем

$$W(a_1, a_2, \ldots, a_n) = \frac{1}{l} \sum_{i=1}^{l} \log P(x_i, a_1, a_2, \ldots a_n),$$

где $\mathbf{x} = (x_1, x_2, \ldots, x_l)$ — описания объектов, представленных в обучающей выборке, $P(x, a_1, a_2, \ldots, a_n)$ — плотность распределения в пространстве X при заданных значениях параметров $(a_1, a_2, \ldots, a_n)$.

Пусть $\boldsymbol{a}_0$ есть то значение вектора параметров, которое соответствует истинной плотности распределения, а $W^*(\boldsymbol{a})$ — математическое ожидание функции правдоподобия, вычисленное при этой плотности.

$$W^*(\boldsymbol{a}) = \int \left(\frac{1}{l} \sum_{i=1}^{l} \log P(x_i, a_1, a_2, \ldots, a_n) \right) dP(x_1, \boldsymbol{a}_0) \ldots dP(x_l, \boldsymbol{a}_0) =$$
$$= \int \log P(\mathbf{x}, \boldsymbol{a})\, dP(\mathbf{x}, \boldsymbol{a}_0).$$

Значение функции $W^*(\boldsymbol{a})$ в точке $\boldsymbol{a}^*$ (если математическое ожидание берется после того, как значение $\boldsymbol{a}^*$ уже найдено по экспериментальным данным) могло бы служить критерием качества структуры, а за его оценку можно было бы взять фактическое значение $W(\boldsymbol{a}^*)$. Но оказывается, что фактическое значение $W(\boldsymbol{a}^*)$ бывает практически всегда завышено по сравнению с математическим ожиданием функции правдоподобия в точке $\boldsymbol{a}^*$, и это завышение зависит от числа неизвестных параметров n. Попробуем оценить это завышение при некоторых дополнительных предположениях.

Истинный максимум математического ожидания $W^*(\boldsymbol{a})$ достигается в точке $\boldsymbol{a}_0$. Обозначим значение функции W в этой точке через W_0: $W_0 = W(\boldsymbol{a}_0)$. Разложим функцию правдоподобия $W(\boldsymbol{a})$ в ряд Тейлора в окрестности этой точки с точностью до второго члена, обозначив $\Delta\boldsymbol{a} = \boldsymbol{a} - \boldsymbol{a}_0$:

$$W(\boldsymbol{a}) \approx W_0 + \mathbf{k}\Delta\boldsymbol{a} + \frac{1}{2}\Delta\boldsymbol{a}^{\mathrm{T}}\mathbf{K}\Delta\boldsymbol{a}, \tag{1}$$

где вектор $\mathbf{k}$ — градиент функции $W(\boldsymbol{a})$ в точке $\boldsymbol{a}_0$, а отрицательно определенная матрица $\mathbf{K}$ является матрицей вторых производных. Соответственно, координаты вектора $\mathbf{k}$ равны

$$k_i = \frac{\partial W(\boldsymbol{a})}{\partial a_i} = \frac{1}{l} \sum_{j=1}^{l} \frac{\partial \log P(x_j, \boldsymbol{a})}{\partial a_i},$$

а элементы матрицы равны:

$$k_{ts} = \frac{\partial^2 W(\boldsymbol{a})}{\partial a_t \partial a_s} = \frac{1}{l}\sum_{j=1}^{l} \frac{\partial^2 \log P(x_j, \boldsymbol{a})}{\partial a_t \partial a_s}$$

(производные берутся в точке $\boldsymbol{a}_0$).

Тогда в нашем приближении максимум функции правдоподобия будет достигнут при

$$\Delta \boldsymbol{a} = \boldsymbol{a}^* - \boldsymbol{a}_0 = -\mathbf{K}^{-1}\mathbf{k}.$$

Поворотом системы координат в пространстве параметров матрицу $\mathbf{K}$ всегда можно привести к диагональному виду. В новой системе вектор $\Delta \boldsymbol{a}$ будет иметь координаты

$$\Delta a_i = -\frac{\dfrac{1}{l}\displaystyle\sum_{j=1}^{l} \frac{\partial \log P(x_j, \boldsymbol{a})}{\partial a_i}}{\dfrac{1}{l}\displaystyle\sum_{j=1}^{l} \frac{\partial^2 \log P(x_j, \boldsymbol{a})}{\partial a_i^2}},$$

где производные берутся в точке $\boldsymbol{a}_0$.

Допустим теперь, что выборка настолько велика, что матрица вторых производных уже сошлась к своему математическому ожиданию:

$$\frac{1}{l}\sum_{j=1}^{l} \frac{\partial^2 \log P(x_j, \boldsymbol{a})}{\partial a_i^2} \approx \frac{\partial^2 W^*(\boldsymbol{a})}{\partial a_i^2}.$$

Заметим, что в точке максимума $\boldsymbol{a}_0$ вторые производные отрицательны. В этом приближении получим:

$$\Delta a_i = -\frac{\dfrac{1}{l}\displaystyle\sum_{j=1}^{l} \frac{\partial \log P(x_j, \boldsymbol{a})}{\partial a_i}}{\dfrac{\partial^2 W^*(\boldsymbol{a})}{\partial a_i^2}}. \qquad (2)$$

Величины $\partial \log P(x_j, \boldsymbol{a})/\partial a_i$ в точке максимума функции $W^*(\boldsymbol{a})$ имеют нулевое математическое ожидание, и мы продолжаем считать их

случайными, зависящими от конкретной выборки. Соответственно и значение функции $W(\boldsymbol{a})$ в точке ее максимума $\boldsymbol{a}^*$ будет случайным. Подставляя оценку (2) в приближение (1), получим

$$W(\boldsymbol{a}^*) \approx W_0 - \frac{1}{2}\mathbf{k}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{k} \approx$$

$$\approx W_0 - \frac{1}{2}\sum_{i=1}^{n}\frac{\dfrac{1}{l}\displaystyle\sum_{j=1}^{l}\left(\dfrac{\partial \log P(x_j, \boldsymbol{a})}{\partial a_i}\right)^2}{\dfrac{\partial^2 W^*(\boldsymbol{a})}{\partial a_i^2}}.$$

Оценим теперь математическое ожидание величины $W(\boldsymbol{a}^*)$. Поскольку случайные величины $\partial \log P(x_j, \boldsymbol{a})/\partial a_i$ независимы, имеют нулевое математическое ожидание и одинаковую дисперсию, то заменяя математическое ожидание суммы на сумму математических ожиданий, получим

$$EW(\boldsymbol{a}^*) \approx W^*(\boldsymbol{a}_0) - \frac{1}{2}\sum_{i=1}^{n}\frac{1}{l}E\left\{\frac{\left(\dfrac{\partial \log P(x, \boldsymbol{a})}{\partial a_i}\right)^2}{\dfrac{\partial^2 W^*(\boldsymbol{a})}{\partial a_i^2}}\right\}.$$

Но, как мы показали еще на лекции 3, справедливо соотношение

$$E\left(\frac{\partial \log P(x, \boldsymbol{a})}{\partial a_i}\right)^2 = -E\frac{\partial^2 \log P(x, \boldsymbol{a})}{\partial a_i^2} = -\frac{\partial^2 W^*(\boldsymbol{a})}{\partial a_i^2}. \qquad (3)$$

Поэтому дробь внутри фигурных скобок сокращается, и получаем

$$EW(\boldsymbol{a}^*) \approx W^*(\boldsymbol{a}_0) + \frac{n}{2l}.$$

(здесь математическое ожидание вычисляется усреднением по тем же данным, по которым находится $\boldsymbol{a}^*$).

Таким образом, значение функции правдоподобия в точке ее максимума оказывается в среднем завышено по сравнению с ее значением в точке максимума функции $W^*(\boldsymbol{a}_0)$ на величину $n/2l$. Но нам интересно оценить значение функции правдоподобия $W^*(\boldsymbol{a})$, вычисленной

по всей генеральной совокупности, в точке $\boldsymbol{a}^*$, найденной только по экспериментальным данным.

Точно тем же путем показывается, что значение функции $W^*(\boldsymbol{a})$ в точке $\boldsymbol{a}^*$ будет в среднем меньше чем $W^*(\boldsymbol{a}_0)$ также на величину $n/2l$. Поэтому оказывается, что значение $W(\boldsymbol{a}^*)$ в среднем завышено по отношению к $W^*(\boldsymbol{a}^*)$ на величину n/l.

Поэтому величина $W(\boldsymbol{a}^*) - n/l$ будет (приблизительно) несмещенной оценкой среднего (по всей генеральной совокупности) значения функции правдоподобия в точке $\boldsymbol{a}^*$, найденной по экспериментальным данным, и может служить оценкой качества структуры модели. Это и есть *информационный критерий Акаике*. Перебирая возможные структуры, можно по этому критерию выбрать лучшую из них.

Как видим, этот критерий очень прост — из значения максимума функции правдоподобия нужно вычесть отношение числа неизвестных параметров модели n к длине обучающей выборки l. Однако его применение ограничено рядом обстоятельств.

Прежде всего, для применения метода максимального правдоподобия модель должна быть представлена настолько точно, чтобы можно было вычислить для каждого объекта x_j значение плотности распределения $p(x_j, \alpha)$. Недостаточно просто получить решающее правило или регрессионную зависимость: в описание модели должно входить точное описание стохастической связи между входным описанием и выходным значением, то есть плотность $p(y|\mathbf{x}; \boldsymbol{a})$, зависящую от ряда неизвестных параметров $\boldsymbol{a}$.

Далее, предполагается, что точка $\boldsymbol{a}^*$, в которой достигается максимум функции правдоподобия, настолько близка к $\boldsymbol{a}_0$, что достаточно ограничиться двумя членами разложения в ряд Тейлора. Но сравнению подлежат и такие структуры модели, для которых это условие не выполнено. Предполагается также, что каждая из сравниваемых структур содержит истинную модель, иначе неправомочно использовать тождество (3). А это исключает сравнение структур, упрощающих модель.

Более того, выборка должна быть настолько большой, чтобы квадратичная часть этого разложения сошлась в силу закона больших чисел к ее математическому ожиданию. Для этого необходимо, чтобы число неизвестных параметров было много меньше длины выборки, а тогда и поправка Акаике будет малой. В противном случае вместо

отношения

$$E\frac{\left(\dfrac{1}{l}\displaystyle\sum_{j=1}^{l}\frac{\partial \log P(x_j, \boldsymbol{a})}{\partial a_i}\right)^2}{\dfrac{\partial^2 \log P(x, \boldsymbol{a})}{\partial a_i^2}}$$

пришлось бы оценивать величину

$$E\frac{\left(\dfrac{1}{l}\displaystyle\sum_{j=1}^{l}\frac{\partial \log P(x_j, \boldsymbol{a})}{\partial a_i}\right)^2}{\dfrac{1}{l}\displaystyle\sum_{j=1}^{l}\frac{\partial^2 \log P(x_j, \boldsymbol{a})}{\partial a_i^2}}.$$

Но математическое ожидание отношения не равно отношению математических ожиданий, и при малых значениях знаменателя расхождение может быть очень большим.

Это становится особенно ясно видно на следующем примере. Рассмотрим задачу восстановления линейной регрессии

$$y = \sum_{i=1}^{n} a_i x_i$$

по выборке $\mathbf{x}_1, \ldots, \mathbf{x}_l$, если данные представлены с независимой нормально распределенной помехой, имеющей нулевое среднее и известную дисперсию D_ξ. Метод наименьших квадратов здесь совпадает с методом максимального правдоподобия, а функция правдоподобия имеет вид

$$W(a) = -\frac{1}{2l}\left[\log(2\pi D_\xi)^l + \frac{(\mathbf{c} - \mathbf{c}(\boldsymbol{a}))(\mathbf{c} - \mathbf{c}(\boldsymbol{a}))^{\mathrm{T}}}{D_\xi}\right],$$

где $\mathbf{c}$ — вектор наблюдавшихся выходных значений, $\mathbf{c}(\boldsymbol{a}) = \boldsymbol{a}\mathbf{X}$ — вектор предсказываемых значений, $\mathbf{X}$ — матрица, составленная из векторов $\mathbf{x}_i$.

Обозначим через $\boldsymbol{a}_0$ истинный вектор коэффициентов регрессии, а через $\boldsymbol{a}^*$ — вектор коэффициентов, полученных при помощи МНК по данным обучения. Сравним теперь значение функции правдоподобия

для произвольного входного вектора $\mathbf{x}$ и соответствующего выходного значения c при $\boldsymbol{a} = \boldsymbol{a}_0$ и $\boldsymbol{a} = \boldsymbol{a}^*$:

$$W_0 = W(\boldsymbol{a}_0, \mathbf{x}, c) = -\frac{1}{2l}\left[\log(2\pi D_\xi) + \frac{(c - \boldsymbol{a}_0^{\mathrm{T}}\mathbf{x})^2}{D_\xi}\right],$$

$$W(\boldsymbol{a}^*, \mathbf{x}, c) = -\frac{1}{2l}\left[\log(2\pi D_\xi) + \frac{(c - \boldsymbol{a}_0^{\mathrm{T}}\mathbf{x})^2}{D_\xi} + \frac{((\boldsymbol{a}_0 - \boldsymbol{a}^*)^{\mathrm{T}}\mathbf{x})^2}{D_\xi}\right].$$

Обозначим через $\Delta\boldsymbol{a} = \boldsymbol{a}^* - \boldsymbol{a}_0$ погрешность оценки коэффициентов регрессии, а через $\Delta W = W(\boldsymbol{a}^*, \mathbf{x}, c) - W(\boldsymbol{a}_0, \mathbf{x}, c)$ отклонение значений функции правдоподобия. Считая теперь фиксированным значение $\boldsymbol{a}^*$, а случайными только вектор $\mathbf{x}$ и шум ξ, оценим математическое ожидание ΔW:

$$E_1\Delta W = -\frac{1}{2}\cdot\frac{((\boldsymbol{a}_0 - \boldsymbol{a}^*)^{\mathrm{T}}\mathbf{x})^2}{D_\xi} = -\frac{1}{2D_\xi}\left[\Delta\mathbf{a}^{\mathrm{T}}E(\mathbf{x}\mathbf{x}^{\mathrm{T}})\Delta\boldsymbol{a}\right].$$

Но матрица $E(\mathbf{x}\mathbf{x}^{\mathrm{T}})$ — это просто ковариационная матрица вектора $\mathbf{x}$. Линейным преобразованием координат эта матрица может быть приведена к единичной (что не меняет значений функции правдоподобия). Тогда получаем

$$E_1\Delta W = -\frac{1}{2D_\xi}[\Delta\boldsymbol{a}^{\mathrm{T}}\Delta\boldsymbol{a}], \tag{4}$$

где вектор вычисляется в преобразованных координатах.

На самом деле вектор $\Delta\boldsymbol{a}$ тоже является случайным, поскольку данные обучения выбирались случайно и в них также присутствовал шум. Примем, однако, что матрица $\mathbf{X}$ векторов, представленных в обучении, фиксирована, и только шум при получении выходных значений был случайным. Усредним теперь величину ΔW по возможным значениям помехи. В этом случае, как было показано еще в лекции 3, погрешность в определении коэффициентов регрессии определяется как

$$E[\Delta\boldsymbol{a}^{\mathrm{T}}\Delta\boldsymbol{a}] = \sum_{i=1}^{n}\frac{D_\xi}{\lambda_i l},$$

где λ_i — собственные числа матрицы $(1/l)\mathbf{X}\mathbf{X}^{\mathrm{T}}$. Подставляя эту оценку в формулу (4), получим

$$E_2\Delta W = E\left\{-\frac{1}{2D_\xi}[\Delta\boldsymbol{a}^T\Delta\boldsymbol{a}]\right\} = -\frac{1}{2}\sum_{i=1}^{n}\frac{1}{\lambda_i l}.$$

Таким образом, значение функции правдоподобия на новых данных для регрессии, построенной с помощью МНК, оказывается в среднем меньше такого же значения для истинной регрессии на величину $(1/2)\sum_{i=1}^{n} 1/(\lambda_i l)$.

Тем же путем можно показать, что значение функции правдоподобия для регрессии по МНК, вычисленное на данных обучения, оказывается в среднем больше W_0 на ту же величину $(1/2)\sum_{i=1}^{n} 1/(\lambda_i l)$. Следовательно, значение функции правдоподобия в точке $\boldsymbol{a}^*$, вычисленное по данным обучения, будет завышено по сравнению со значением, вычисленным по новым данным в среднем на величину $(1/l)\sum_{i=1}^{n}(1/\lambda_i)$.

В силу закона больших чисел матрица $(1/l)\mathbf{X}\mathbf{X}^{\mathrm{T}}$ сходится к истинной ковариационной матрице, т.е. в нашем случае к единичной. Поэтому и ее собственные числа λ_i должны сходиться к 1. Тогда асимптотически

$$\frac{1}{l}\sum_{i=1}^{n}\frac{1}{\lambda_i} = \frac{n}{l},$$

что совпадает с оценкой Акаике.

Но пока асимптотика не наступила, малые значения λ_i играют большую роль и $\sum_{i=1}^{n} 1/\lambda_i$ оказывается существенно большей, чем n.

Можно было бы попытаться вычислить математическое ожидание $E(\sum_{i=1}^{n}(1/\lambda_i))$ и внести соответствующую поправку в критерий (что и было сделано для нормального распределения в пространстве аргументов). Но оказывается, что эта поправка существенно зависит от распределения вероятностей на множестве аргументов, которого, как правило, мы заранее не знаем.

В нашем выводе критерия Акаике мы предполагали, что обучающая последовательность $\mathbf{x} = (x_1, x_2, \ldots, x_l)$ получена в ходе независимых испытаний при неизменном распределении. Иногда критерий Акаике предлагают использовать и в тех случаях, когда входные данные рассматривают как один объект. Например, это может быть пара, состоящая из достаточно длинной реализации входного и выходного сигналов. Но в этом случае совершенно непонятно, на каком основании можно заменить матрицу вторых производных ее математическим ожиданием.

Тем не менее, информационный критерий Акаике получил широкое распространение при выборе оптимальной сложности модели. Вообще заметим, что использование теоретических оценок при выбо-

ре оптимальной сложности в сравнении с такими методами, как простой экзамен или скользящий контроль (cross validation), оказывается наиболее эффективным при решении обратных задач математической физики или статистики. Дело в том, что сглаживающий оператор в прямой задаче приводит к тому, что избыточная сложность модели слабо сказывается на критерии остаточной невязки даже при независимом экзамене. Случайные отклонения могут привести к тому, что по критерию невязки на экзамене будет выбрано излишне сложное решение, сильно отличающееся от истинной зависимости. Теоретические же оценки прямо вводят штраф за сложность и не дают без надобности увеличивать сложность модели.

Заключение

Мы рассмотрели ряд методов выбора оптимальной сложности модели. Каждый из них обладал теми или иными достоинствами и недостатками. Но возникает вопрос, зачем вообще упрощать модель. Мы знаем ответ: потому, что иначе нам не хватит данных обучения, и, даже получив хороший результат на обучении, мы получим модель, которая неудовлетворительно ведет себя на новых данных. С другой стороны, чрезмерно упрощая модель, мы получим плохой результат и на данных обучения, и на новых данных. Поэтому нашей целью был выбор оптимальной сложности — путем проверки результата на новых данных, или методом скользящего контроля, или на основании теоретических оценок.

Но есть задачи, где сложная модель нужна принципиально. В последнее время появились примеры (в задачах распознавания), когда после обучения на выборках длиной порядка 10000 доля ошибок на новом материале составляла 20–30%. Но когда длина выборки доходила до миллиона и больше, доля ошибок снижалась до 5% и меньше. Это особенно чувствуется при применении кернельных методов, которые при достаточно большой выборке позволяют строить сколь угодно сложные модели. Видимо, в этих случаях разделяющая поверхность оказывается достаточно сложной, и найти ее по малой выборке невозможно.

В то же время нужно помнить, что сложность мы понимаем не совсем в обыденном смысле слова. Для нас сложность — это число на-

страиваемых параметров. Существуют очень сложные схемы в обычном смысле, но число настраиваемых параметров в них невелико — например, если исходные данные заранее преобразованы так, что расстояние между классами оказывается велико по сравнению с их диаметром, или заранее выделена система вторичных признаков, пространство которых имеет существенно меньшую размерность, чем пространство исходных признаков.

В частности, человек может научиться узнавать лицо другого человека по нескольким фотографиям. Ясно, что схема, по которой работает система глаз–мозг, очень сложна, но она заранее приспособлена для того, чтобы быстро обучаться. Число настраиваемых параметров здесь не может быть очень большим, иначе не хватило бы информации в данных обучения. Человеческая система узнавания заранее настроена так, чтобы можно было обучиться по небольшому числу показов. Эта «настройка» выработана как в процессе филогенеза, так и при онтогенезе, то есть частично передается генетически, а частично приобретается в течение всего жизненного опыта. Видимо, к числу «запаянных» относятся такие возможности, как выделение контуров, определение инвариантов к тем или иным преобразованиям и пр.

В то же время существующие алгоритмы обучения, за исключением разработанных специально для решения узко конкретных задач, работают «с чистого листа». Есть данные, про которые мы заранее ничего не знаем, и есть алгоритм обучения, который должен по этим данным найти правильное решение. И, если это решение принципиально должно быть сложным, то и нужна выборка длиной в миллион показов.

В программировании мы знаем, что большие и сложные программы не могут писаться просто команда за командой. Такая программа строится иерархически из модулей, каждый из которых решает более простую и специальную задачу. Может быть, так и должны строиться системы обучения?

Позвольте немного пофантазировать. Представим себе, что есть группа взаимосвязанных задач в некоторой области, среди которых есть более простые и более сложные. Система обучения ищет среди них те, которые удается решить по сравнительно небольшим выборкам, строя простые решающие правила, несложные регрессионные зависимости, простые модели. А далее полученные решения используются как готовые модули при решении более сложных задач. Так иерархически

строятся сложные модели. При этом, когда сложная модель строится из уже готовых модулей более низкого уровня, не приходится работать «с чистого листа», как это происходит сегодня, — настройке подлежит сравнительно небольшое число неизвестных параметров, и не нужны будут выборки объемом в миллион примеров.

Конечно, тут можно сказать — если бы знал как, сделал бы. Таких систем обучения в настоящее время нет. Но дело еще и в том, что такая система обучения должна быть нацелена сразу на целый комплекс взаимосвязанных задач. К этому сейчас приближаются системы распознавания зрительных образов. В задачах геологии построение поверхностей раздела пластов различной геологической природы и других геологических структур может служить основой для нахождения осей анизотропии или границ блоков при оценке полей содержания компонентов методом кригинга. Такого рода система может существовать и в задаче классификации сайтов или страниц для их разбраковки и поиска оптимальных ответов на запросы. Мы видим на наших семинарах, как переплетаются эти задачи, но каждая из них решается «с чистого листа», без использования результатов, полученных при решении других задач. Создание именно системы было бы здесь очень интересно.

Список литературы

[1] Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Исследование зависимостей. В книге «Прикладная статистика» / под ред. С. А. Айвазяна. *Москва, Финансы и статистика, 1985.*

[2] Айвазян С. А., Бухштабер В. М., Енюков И. С., Л. Д. Мешалкин. Классификация и снижение размерности. В книге «Прикладная статистика» / под ред. С. А. Айвазяна. *Москва, Финансы и статистика, 1989.*

[3] Бэстенс Д.— Э., ван Ден Берг В.— М., Вуд Д. Нейронные сети и финансовые рынки. Принятие решений в торговых операциях.

[4] Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. *Москва, Наука 1974.*

[5] Вапник В. Н. Восстановление зависимостей по эмпирическим данным. *Москва, Наука 1979.*

[6] Вапник В. Н., Червоненкис А. Я. Необходимые и достаточные условия равномерной сходимости средних к математическим ожиданиям. *Теория вероятностей и ее приложения. 1981, т. 26, № 3, стр. 543-564.*

[7] Вапник В. Н., Червоненкис А.Я̇. О методе упорядоченной минимизации риска. I. *Автоматика и телемеханика, №8, 1974, стр.21-–*30.*

[8] Вапник В. Н., Червоненкис А.Я̇. О методе упорядоченной минимизации риска. II. *Автоматика и телемеханика, №9, 1974, стр. 29— 39.*

[9] Алгоритмы и программы восстановления зависимостей. / Под редакцией В. Н. Вапника. *Москва, Наука 1984.*

[10] Галушкин А. И. Теория нейронных сетей. Т. 1. Нейрокомпьютеры и их применение. *Изд-во ИПРЖР, 2000.*

[11] Гельфанд И. М. Лекции по линейной алгебре. *Добросвет, 2007.*

[12] Гнеденко Б. В. Курс теории вероятностей. *М.: Наука, Главная редакция физико-математической литературы, 1965.*

[13] Колмогоров А. Н., Фомин С. В. Элементы теории функций и функционального анализа. *М.: Физматлит, Издание 7-е, 2006.*

[14] Круглов В. В., Борисов В. В. Искусственные нейронные сети. Теория и практика. *Изд-во Горячая Линия — Телеком*

[15] Лаврентьев М. М., Савельев Л. Я. Линейные операторы и некорректные задачи. *М.: Наука, 1991.*

[16] Льюнг Л. Идентификация систем. Теория для пользователя.

[17] Носко В. П. Эконометрика для начинающих. *М.: Институт экономики переходного периода, 2000.*

[18] Тихонов А. Н., Арсенин В. Я. Методы решения некорректных задач. *М.: Наука, 1979.*

[19] Турчин В. Ф., Козлов В. П., Малкевич М. С. Использование методов математической статистики для решения некорректных задач. *УФН 1970, 102: 345-386.*

[20] Уилкс С. Математическая статистика. *М.: Наука 1967.*

[21] Саймон Хайкин. Нейронные сети. Полный курс. *Изд-во Вильямс, 2006.*

[22] A. Ja. Chervonenkis. A combined Bayes – Maximum likelihood method for regression. // В книге «Data Fusion and Perception» / edited by Giacomo Della Riccia, Hanz-Joachim Lenz, Rudolf Kruse. *Springer, Wien, New York, 2001.*

[23] Jean Paul Chiles, Pierre Delfiner. Geostatistics. Modeling Spatial Uncertainty. *Wiley, Series in Probability and Statistics. 1999.*

[24] Devroye Luc, Gorfi Laslo, Lugosi Gabor A probabilistic theory of pattern recognition. *Springer, 1996.*

[25] A. Gammerman, V. Vovk. Hedging prediction in Machine Learning. *Computer Journal 50, 151 –157. 2007.*

[26] J. D. C. MacKay. Bayesian interpolation. *Neural Computation, 4(3): 415-447, 1992.*

[27] H. Mohamad. Hassoun Fundamentals of Artificial Neural Networks.

[28] J. Rissanen. Stochactic complexity. *Journal of the Royal Statistical Society series B, 49: 223-239. 1987.*

[29] V. N. Vapnik. Statistical Learning Theory. *Wiley, New York, 1998.*

[30] V. N. Vapnik. The Nature of Statistical Learning Theory. *Springer, New York, 2000.*

[31] V. Vovk, A. Gammerman and G. Shafer. Algorithmic learning in a random world. *Springer, 2004.*

[32] C. S. Wallace and P. R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society series B, 49: 240-265. 1987.*

От редактора:

дополнительная литература

[1] Christopher M. Bishop. Pattern Recognition and Machine Learning. *Springer, 2006.* (ISBN 0-387-31073-8)

[2] Bing Liu. Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. *Springer, 2007.* (ISBN 3-540-37881-2)

[3] P. M. Bhagat. Pattern Recognition in Industry, *Elsevier, 2005.* (ISBN 0-08-044538-1)

[4] T.–M. Huang, V. Kecman, I. Kopriva. Kernel Based Algorithms for Mining Huge Data Sets, Supervised, Semi-supervised, and Unsupervised Learning. *Springer-Verlag, Berlin, Heidelberg, 2006.*

[5] D. J. C. MacKay. Information Theory, Inference, and Learning Algorithms. *Cambridge University Press, 2003.* (ISBN 0-521-64298-1)

[6] Sholom Weiss and Casimir Kulikowski. Computer Systems That Learn. *Morgan Kaufmann, 1991.* (ISBN 1-55860-065-5)

[7] A. Chernov and V. Vovk. Prediction with expert evaluator's advice. `http://arxiv.org/PS_cache/arxiv/pdf/0902/0902.4127v2.pdf`

[8] E. Agichtein, E. Gabrilovich, H. Zha. The Social Future of Web Search: Modeling, Exploiting, and Searching Collaboratively Generated Content. *IEEE Data Engineering Bulletin, 06/2009, Volume 32, Issue 2, p.52-61, 2009.*

[9] Unsupervised search-based prediction. `http://www.cs.utah.edu/~hal/docs/daume09unsearn.pdf`

[10] One class SVMs for document classification.
http://jmlr.csail.mit.edu/papers/volume2/manevitz01a/manevitz01a.pdf

[11] Improving one class SVM detection.
http://www.it.iitb.ac.in/~deepak/deepak/courses/mtp/papers/svm%20anomaly%20detection.pdf

[12] Discriminating against new classes: one class versus multi-class classification.
http://www.cs.waikato.ac.nz/~eibe/pubs/hf08-domainsForOCC.pdf

[13] Probabilistic classification.
http://genie.weizmann.ac.il/pubs/conference/ijcai01.pdf

[14] Nikolik and Gurfs Madevska. Probabilistic SVM. *Neurocomputing, vol. 62, p. 293–303, 2003.*

Послесловие

Книга, которую вы держите в руках, тесно связана с курсом, читаемым автором в Школе Анализа Данных Яндекса.

ШАД — не совсем обычное учебное заведение. Школа создана и поддерживается российской компанией, благополучие и успех которой во многом основаны на решении сложных алгоритмических и математических задач, возникающих в ходе создания и развития интернет-сервисов.

В 90-е годы, когда Яндекс только стартовал, эталонная коллекция, по которой мы настраивали поиск, содержала несколько тысяч документов, а в ранжировании использовались лишь несколько признаков. Сейчас, осенью 2009 года, коллекция запросов с оценками экспертов, использующаяся для настройки алгоритмов поиска, составляет 75 тысяч, документов, по которым проходит поиск — 5 миллиардов, а признаков, анализируемых для каждого найденного документа в каждом запросе, больше 350. Очень многие из этих признаков являются обученными на отдельных данных и отдельных экспертных оценках нетривиальными классификаторами, построенными из более элементарных признаков.

Такие масштабы и уровень сложности — не наша прихоть. Они диктуются высокими требованиями к нашему основному программному продукту — веб-поиску для миллионов пользователей сети Интернет.

Путь, который пройден нами от малого числа признаков и обучающих примеров, от «короткой модели», к большому числу признаков, оптимизируемых метрик, обучающих примеров, был бы невозможен без машинного обучения.

Можно без преувеличения сказать, что машинное обучение является ключом и основой функционирования всего Яндекса — не толь-

ко веб-поиска с его своеобразными лингвистическими моделями, но и рекламных технологий, поиска изображений, Яндекс.Новостей, Яндекс.Маркета и многих других сервисов. Именно на основе машинного обучения мы можем ожидать интегрирования специфических экспертных знаний, использующихся в лингвистических словарях, грамматических правилах, в методах расшифровки и интерпретации космических и медицинских изображений, в процедурах, синтезирующих виртуальные реальности. Умение автоматически обрабатывать огромные массивы информации, используя одновременно логический (комбинаторный) и статистический анализ, делает эту науку особенно важной для развития Интернета. Именно машинное обучение и его мощные, надежно работающие модели и техники дают возможность инженерам и командам, независимо работающим над отдельными факторами и группами факторов, решать самые разнообразные задачи.

Несколько лет назад мы столкнулись с кадровой проблемой: хотя студенты, обучающиеся в МГУ, МФТИ, МВТУ, получают хорошее образование общего уровня, их математическая подготовка недостаточно сфокусирована на использующихся на практике дисциплинах, таких как статистика, теория вероятностей, случайные процессы. Машинное обучение до недавнего времени, как правило, отсутствовало в списке обязательных дисциплин в ведущих вузах страны, да и сейчас ему отводят лишь вспомогательную роль в системе подготовки современных инженеров компьютерных специальностей.

Это, а также отсутствие фокусировки на методах обработки и анализа текстов и изображений, вдохновило нас на создание Школы Анализа Данных. Уверенность придавали нам история и традиции отечественных ученых 60–70-х годов XX века (и более ранних лет) в этой области. Мы счастливы, что А. Червоненкис, стоящий у основ невероятно популярных во всем мире методов и теории машинного обучения, стал одним из тех, кто помогает нам строить школу.

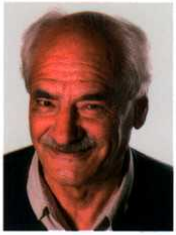
Ситуация с учебниками в области машинного обучения также не кажется удовлетворительной. Перед нами первый оригинальный непереводной учебник-монография по машинному обучению, написанный одним из основоположников этой дисциплины. Следует отметить, что машинное обучение — синтетическая дисциплина, использующая в своих разработках совершенно различные базовые методы алгебры, функционального анализа, теории вероятности и информации, теории оптимизации и игр, а также многих других предметов, преподавание кото-

рых на математических и компьютерных факультетах зачастую ограничено не связанными друг с другом полугодовыми спецкурсами. Поэтому большинство существующих учебников (на английском языке) по машинному обучению отражают приверженность авторов к определенным направлениям теории машинного обучения. Учебник Червоненкиса выгодно отличается от них. Излагаемый предмет рассмотрен всеобъемлюще. Казалось бы, автор важного направления этой теории также должен был сделать акцент па своих разработках, но Червоненкис поставил перед собой совершенно другую задачу — описать предмет всесторонне и целостно, чтобы читатель пе только увидел всё огромное разнообразие базовых методов, используемых в этой области, но и почувствовал их взаимосвязь.

Книга дает глубокий взгляд на дисциплину. Она написана очень доступно и при этом — с соблюдением строгих принципов математического обоснования основных моделей и средств их анализа.

Я очень рад, что эта книга вышла, и надеюсь что она оказалась для вас интересной и полезной.

Илья Сегалович,
директор по технологиям,
Яндекс



Алексей Яковлевич Червоненкис родился в Москве в 1938 году. Закончив ФИЗТЕХ в 1961 году, он поступил на работу в Институт Проблем Управления РАН в г. Москве, где работает до сих пор.
С 2000 года по настоящее время профессор Royal Holloway University of London. В школе Яндекса (ШАД) работает с самого ее основания, совмещая преподавательскую работу с научной деятельностью, участвует в проектных разработках Яндекса.

Наиболее известные работы Алексея Яковлевича посвящены условиям равномерной сходимости частот к вероятностям по классу событий и равномерной сходимости средних к математическим ожиданиям по классу функций. Наряду с этими фундаментальными теоретическими исследованиями, А.Я. Червоненкис вел крупные практические разработки, используя их как полевые исследования для настоящей проверки своих теоретических результатов. Среди этих его работ следует выделить цикл комплексных исследований с применением методов управления, математической статистики и теории вероятностей для разработки крупных рудных месторождений. Этими исследованиями он занимался более 35 лет с 1970 года. В частности, совместно с сотрудниками им была разработана и внедрена система, реализующая эти методы на крупнейшем в мире золотоносном карьере Мурун-Тау (Узбекистан). Эта работа была отмечена Государственной премией СССР за 1987 год.

В последние годы Алексей Яковлевич работает над методами распознавания образов для прогноза редких событий, а также применительно к новым объектам, например, представляемым символьными последовательностями с символами из конечного алфавита. Как и раньше в этих работах он ищет характерные особенности новых задач на конкретных практических примерах, и, в частности, на примерах, которые в большом количестве предоставляет ему работа над интернетовскими проектами в Яндексе.

ISBN 590469601-9
9 785904 696016

Лекции школы
анализа данных Яндекса