МИНИСТЕРСТВО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

МИНИСТЕРСТВО ПРОМЫШЛЕННОСТИ, НАУКИ И ТЕХНОЛОГИЙ
РОССИЙСКОЙ ФЕДЕРАЦИИ

РОССИЙСКАЯ АССОЦИАЦИЯ НЕЙРОИНФОРМАТИКИ

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ИНЖЕНЕРНО-ФИЗИЧЕСКИЙ ИНСТИТУТ
(ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ)

НАУЧНАЯ СЕССИЯ МИФИ–2002

НЕЙРОИНФОРМАТИКА–2002

**IV ВСЕРОССИЙСКАЯ
НАУЧНО-ТЕХНИЧЕСКАЯ
КОНФЕРЕНЦИЯ**

ЛЕКЦИИ
ПО НЕЙРОИНФОРМАТИКЕ

Часть 2

По материалам Школы-семинара
«Современные проблемы нейроинформатики»

Москва 2002

УДК 004.032.26 (06)
ББК 32.818я5
М82

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

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

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

Ответственный редактор
Ю. В. Тюменцев, кандидат технических наук

ISBN 5–7262–0400–X

Содержание

***С. А. Шумский*. Байесова регуляризация обучения** **30**
Введение 30
Обучение по Байесу 33
Обучение. Основные понятия 33
Регуляризация обучения 35
Предварительное обсуждение 39
Связь с ошибкой обобщения и минимальной длиной описания 43
EM-алгоритм 45
Резюме 47
История и библиография 48
Оценка параметров по Байесу. Семь раз отмерь... 51
Оценка параметра в разных моделях 51
Оценка шума 53
Проверка априорных гипотез 54
Резюме 56
История и библиография 56
Байесова интерполяция функций. Без кросс-валидации 57
Постановка задачи 57
Решение в общем виде 58
Вычисление методом перевала 59
Предварительное обсуждение 61
Итерационное обучение 62
Лапласовский Prior и прореживание модели 63
Оценка ошибок предсказаний 65
Резюме 67
История и библиография 68
Байесова кластеризация. Сколько кластеров «на самом деле» . 70
Постановка задачи 71

УДК 004.032.26 (06) Нейронные сети

Оптимальная гипотеза 71
Сколько кластеров в данных? 73
Оптимальная модель 76
Численные эксперименты 77
Резюме 81
История и библиография 81
Заключение 82
Подробности 83
Бросание монеты (к разделу «Обучение по Байесу») . . . 83
Принцип минимальной длины описания (к разделу «Обучение по Байесу») 84
Проверка априорных гипотез (к разделу «Оценка параметров по Байесу») 86
Bayesian Information Criterion (к разделу «Байесова интерполяция функций») 87
Оптимизация кластерной модели (к разделу «Байесова кластеризация») 89
Литература 90

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

БАЙЕСОВА РЕГУЛЯРИЗАЦИЯ ОБУЧЕНИЯ

Аннотация

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

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

BAYESIAN REGULARIZATION OF LEARNING

Abstract

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

Введение

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

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

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

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

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

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

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

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

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

Обучение по Байесу

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

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

Обучение. Основные понятия

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

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

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

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

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

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

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

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

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

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

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

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

Для фиксации терминологии запишем эту основополагающую формулу в словесном виде:

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

Регуляризация обучения

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

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

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

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

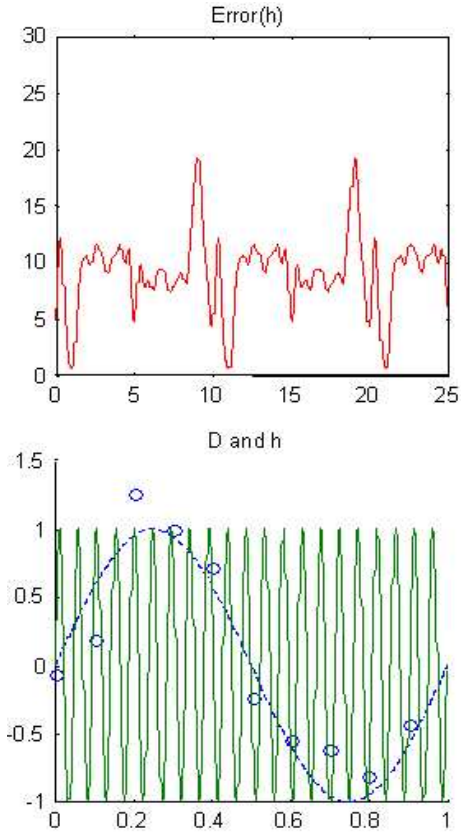

Рис. 1. Определение частоты зашумленного синуса $y = \sin(h\,x) + 0.2\,\eta$. Здесь модель задает характер шума η и вид функции $\sin(h\,x)$, где в роли гипотезы h выступает частота. Функция ошибки (вверху) имеет множество локальных минимумов. Без регуляризации наиболее правдоподобным может оказаться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

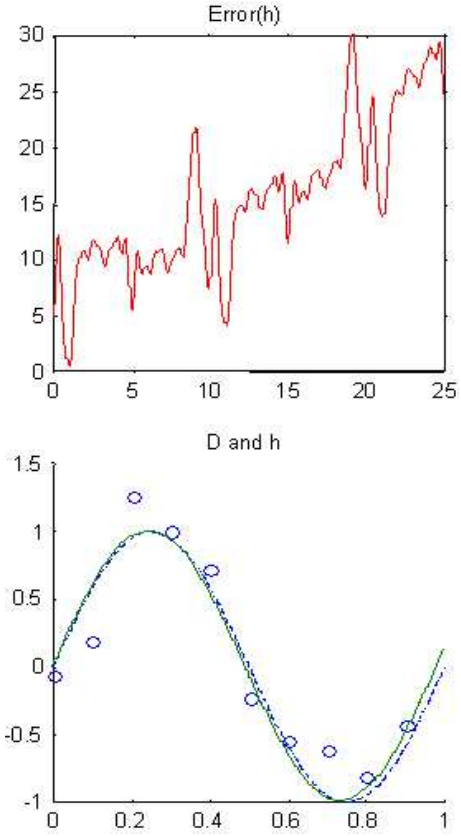

Рис. 2. Регуляризация модели — добавление к ошибке члена, штрафующего большие частоты, помогает выявить наиболее устойчивое к шуму решение, близкое к реальному прототипу.

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

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

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

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

Насколько правдоподобно выглядит объяснение данных моделью определяет знаменатель формулы Байеса

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

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

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

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

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

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

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

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

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

Предварительное обсуждение

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

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

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

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

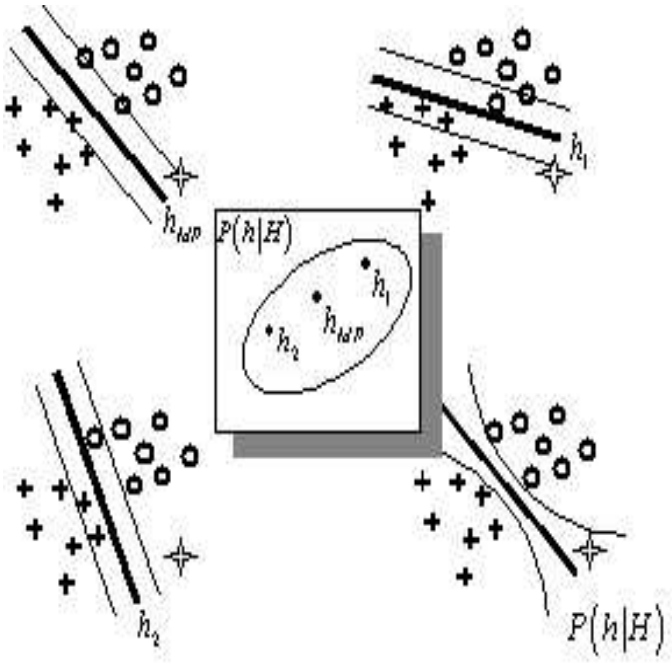

Рис. 3. Иллюстрация байесовского подхода к предсказаниям. Данные представляют собой набор точек из двух классов. Гипотеза классифицирует данные в соответствии с их расположением относительно линии разделения классов, в данном случае — прямой. Звездой отмечена новая точка, отсутствующая в обучающей выборке.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

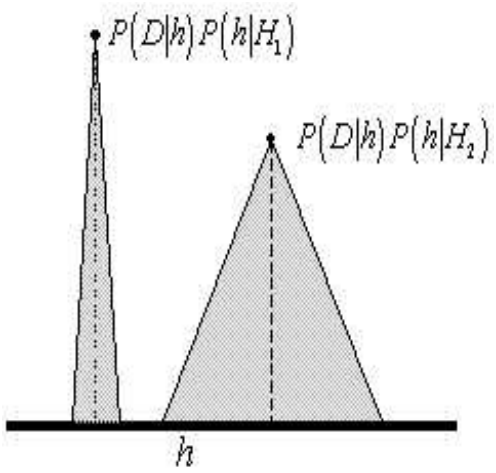

Рис. 4. Из двух моделей, H_1 и H_2, более предпочтительной, по Байесу, является вторая — с большей Evidence (площадь под кривой), несмотря на то, что наилучшая гипотеза в H_1 лучше объясняет данные. Зато H_1 гораздо более чувствительна к вариациям своих параметров, чем H_2.

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

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

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

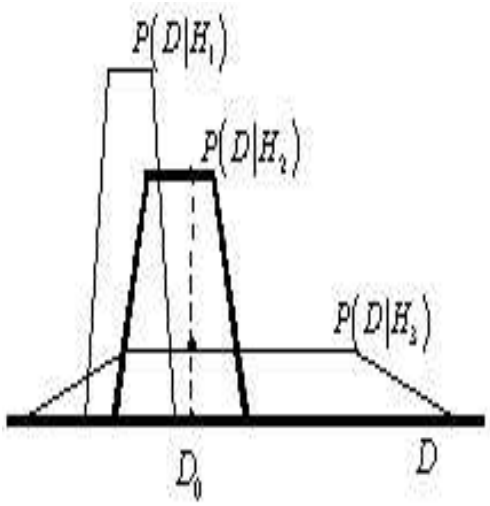

Рис. 5. Максимизация Evidence предполагает выбор наиболее простой модели объяснения данных. Модель H_1 не соответствует данным. Модель H_3 может объяснить не только имеющиеся данные, но и широкий круг других исходов эксперимента. Условие нормировки автоматически понижает ее Evidence. По Байесу, эмпирические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

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

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

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

Связь с ошибкой обобщения и минимальной длиной описания

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

EM-алгоритм

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

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

Оптимизация модели, т. е. минимизация длины описания

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

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

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

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

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

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

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

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

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

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

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

Названия этапов определяется их содержанием.

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

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

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

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

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

Резюме

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

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

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

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

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

История и библиография

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

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

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

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

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

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

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

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

[9] Дядя упомянутого выше Даниила Бернулли.

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

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

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

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

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

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

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

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

Оценка параметров по Байесу. Семь раз отмерь...

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

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

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

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

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

Оценка параметра в разных моделях

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

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

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

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

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

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

По Байесу, вероятностное распределение оценки дается выражением:

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

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

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

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

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

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

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

Покажем как это делается на примере гуассовой модели.

Оценка шума

Логарифм Evidence для гауссовой модели равен:

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

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

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

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

Логарифм Evidence оптимальной гауссовой модели равен, таким образом:

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

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

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

Проверка априорных гипотез

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

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

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

Если для шума также выбрана гауссова модель

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

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

то апостериорная вероятность оценки будет иметь вид:

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

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

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

Наиболее вероятная оценка, максимизирующая апостериорное распределение, есть:

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

А значение Evidence

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

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

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

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

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

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

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

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

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

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

Резюме

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

История и библиография

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

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

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

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

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

Байесова интерполяция функций. Без кросс-валидации

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

Постановка задачи

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

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

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

Запишем это выражение в более компактной форме:

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

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

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

Решение в общем виде

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

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

причем оптимальные значения параметров α,β максимизируют значение Evidence:

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

выраженное через статсуммы:

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

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

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

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

Вычисление методом перевала

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

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

Аналогично, для гауссовой величины Prior:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Длина описания данных наилучшей моделью равна:

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

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

Предварительное обсуждение

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

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

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

Итерационное обучение

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

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

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

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

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

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

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

Лапласовский Prior и прореживание модели

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

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

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

Для построения таких моделей можно использовать лапласовский Prior:

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

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

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

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

откуда:

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

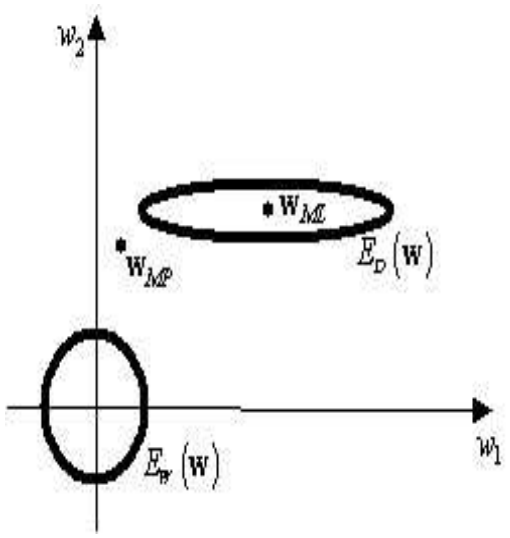

Рис. 6. Пространство параметров после перехода в систему главных осей матрицы вторых производных функции ошибки. Обозначены контуры гауссова Prior E_W и эмпирического Likelihood E_D. Горизонтальному направлению соответствует малое собственное значение $\lambda_1 \ll \alpha$. Соответственно, комбинация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

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

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

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

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

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

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

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

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

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

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

Оценка ошибок предсказаний

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

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

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

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

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

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

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

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

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

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

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

И разброс предсказаний характеризуется соответствующей дисперсией [Bishop 1995]:

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

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

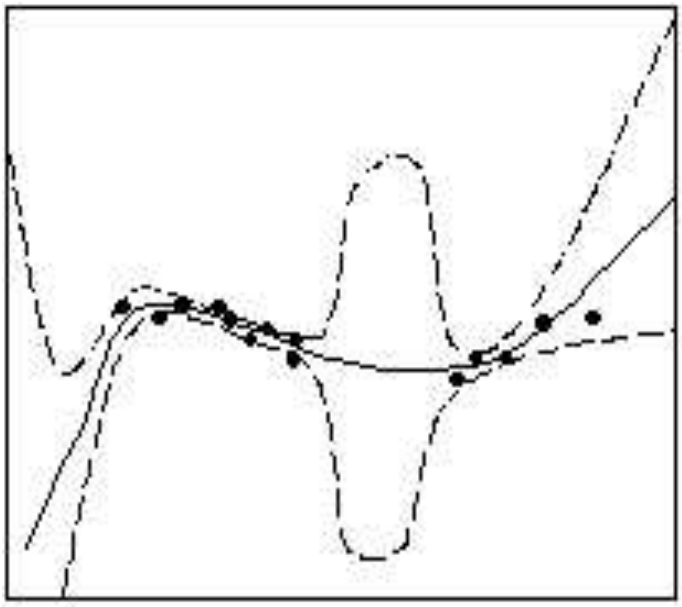

Рис. 7. Предсказания оптимальной модели и оценка разброса предсказаний. Последний возрастает вдали от области эмпирических данных.

Резюме

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

весов до необходимого минимума.

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

История и библиография

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

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

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

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

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

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

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

Байесова кластеризация. Сколько кластеров «на самом деле»?

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

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

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

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

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

Постановка задачи

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

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

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

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

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

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

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

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

Оптимальная гипотеза

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

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

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

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

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

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

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

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

Отсюда легко получить:

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

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

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

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

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

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

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

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

Повторяем эти итерации до гарантированной сходимости.

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

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

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

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

Гауссовы смеси с конечным β осуществляют мягкую или нечеткую кластеризацию.

Сколько кластеров в данных?

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

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

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

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

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

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

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

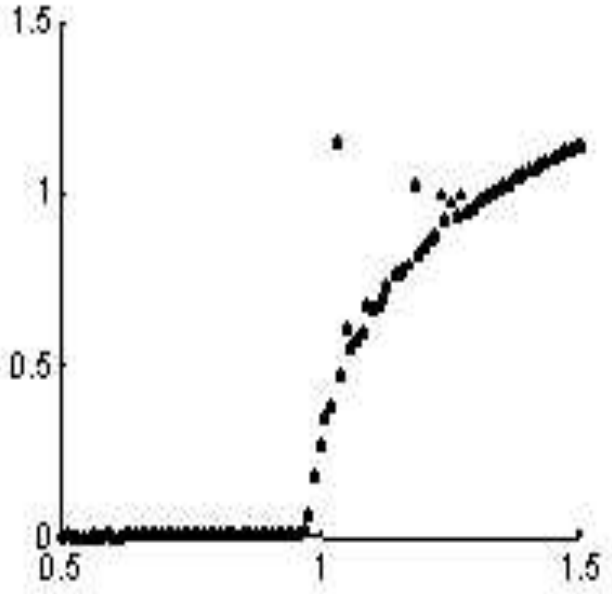

Рис. 8. Зависимость конечного расстояния между двумя центроидами от параметра β. Данные представляли собой 1000 случайных точек с двумерным гауссовым распределением единичной дисперсии. В качестве начального положения центроидов выбирались координаты двух случайно выбранных точек данных. При $\beta 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

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

Оптимальная модель

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

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

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

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

С учетом этого факта, приравнивая нулю производную

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

получим следующее выражение для оптимальной β:

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

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

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

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

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

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

Численные эксперименты

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

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

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

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

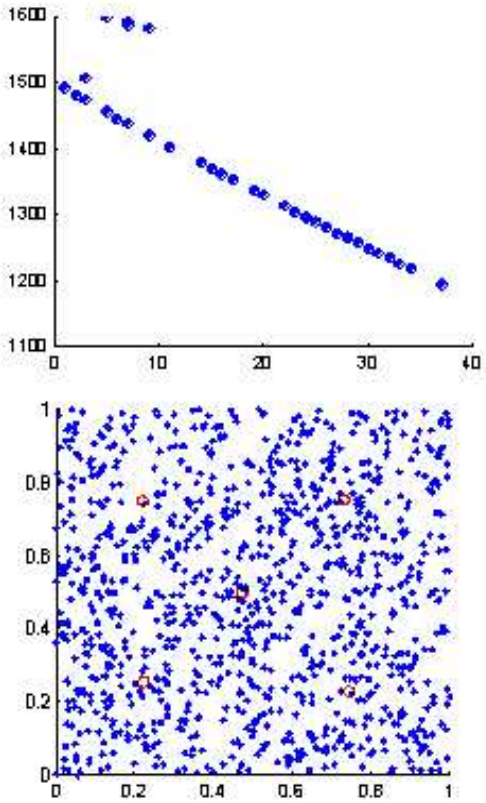

Рис. 9. Верхний график: значения Evidence для точек с однородным распределением в единичном квадрате как функция от конечного числа источников. Максимумы соответствуют моделям с однородным распределением кластеров. Внизу — пример кластеризации с наибольшей Evidence.

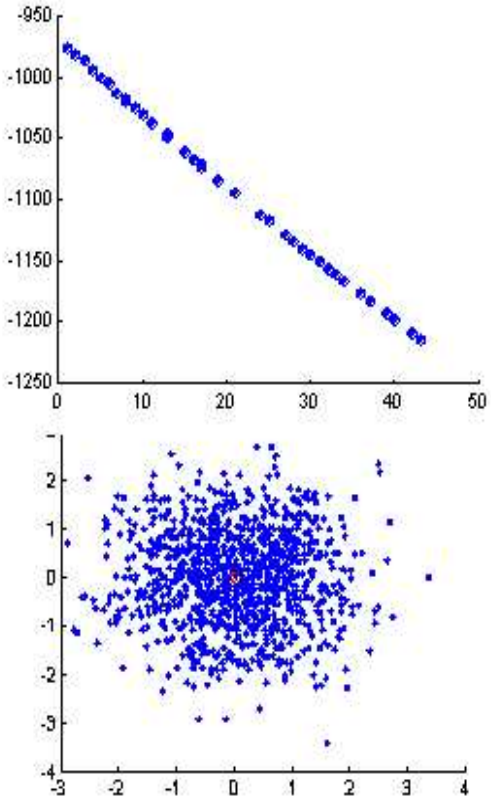

Рис. 10. Верхний график: значения Evidence для точек с гауссовым распределением как функция от конечного числа источников. Максимум соответствует модели с одним кластером. Внизу — пример кластеризации с наибольшей Evidence.

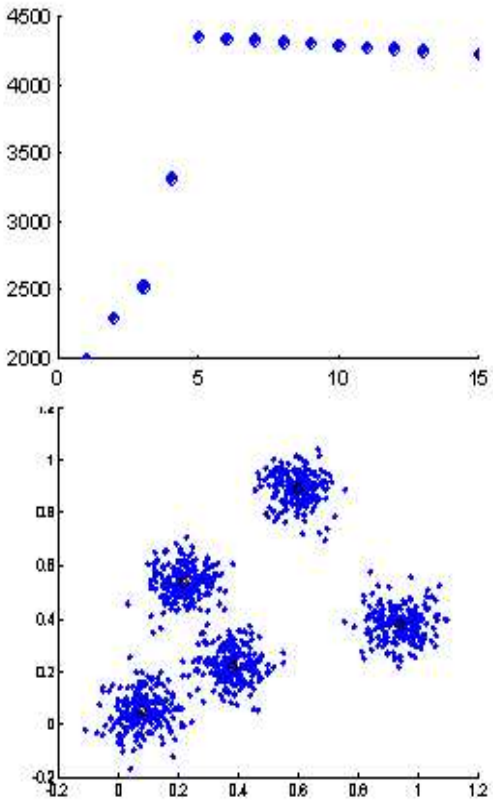

Рис. 11. Верхний график: значения Evidence для точек, порожденных пятью гауссовыми источниками. Максимум соответствует модели с пятью кластерами. Внизу — пример кластеризации с наибольшей Evidence.

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

Резюме

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

История и библиография

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

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

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

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

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

Заключение

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

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

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

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

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

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

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

Подробности

Бросание монеты (к разделу «Обучение по Байесу»)

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

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

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

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

с ожидаемым значением кривизны:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Именно суммарная длина описания дает правильную оценку ожидаемого риска

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Логарифм Evidence в этом случае равен:

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

и достигает максимума при следующем значении α:

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

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

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

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

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

Аналогичным образом находим оптимальную оценку β:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Действительно, последнее неравенство можно переписать в виде:

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

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

Оптимизация кластерной модели (к разделу «Байесова кластеризация»)

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

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

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

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

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

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

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

Литература

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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