Байесовские методы в машинном обучении

Д.П. Ветров

Содержание

Введение

В рамках данного курса мы будем изучать применение байесовских методов к задачам машинного обучения. Нам бы хотелось, чтобы читателю было понятно, как байесовские методы помогают решать конкретные практические задачи. Поэтому по ходу курса мы будем рассматривать как общие инструменты для работы с байесовскими вероятностными моделями (инструменты точного и приближенного байесовского вывода), так и конкретные примеры байесовских моделей машинного обучения. Модели, которые мы будем рассматривать, будут достаточно простые (обобщенная линейная модель регрессии, обобщенная линейная модель классификации, разделение смеси распределений, уменьшение размерности, тематическое моделирование). Однако, после разбора базовых моделей, мы будем говорить о том, какие они допускают расширения и как их можно комбинировать с друг с другом. Более сложные байесовские модели машинного обучения разобраны в курсе "Нейробайесовские методы машинного обучения".

1 Лекция 1. Байесовский подход к теории вероятностей

В этой лекции мы разберем, что такое байесовские методы и чем они отличаются от обычных статистических методов.

1.1 Основные понятия

Машинное обучение является областью математики, которая занимается поиском взаимозависимостей в данных. На вероятностном языке взаимозависимость между величинами можно выразить через условное распределение.

Определение 1. Пусть x и y — две случайные величины. Тогда условным распределением $p(x \mid y)$ (*conditional distribution*) x относительно y называется отношение совместного распределения $p(x, y)$ (*joint distribution*) и маргинального распределения $p(x)$ (*marginal distribution*, оно же безусловное):[1]

$$p(x \mid y) = \frac{p(x, y)}{p(y)}. \tag{1}$$

Смысл этого определения в следующем: условное распределение показывает то, как ведет себя x, если мы уже пронаблюдали y. Заметим, что если величины x и y независимы, т.е. $p(x, y) = p(x)p(y)$, то $p(x \mid y) = p(x)$. Что означает, что никакой информации об x в y не содержится.

Далее из формулы (1), совместное распределение можно выразить через условное и маргинальное:

$$p(x, y) = p(x \mid y)p(y). \tag{2}$$

Такое равенство называют правилом произведения (product rule). Рассуждая по индукции, несложно прийти к его обобщению на n случайных величин:

Теорема 1 (Правило произведения). Пусть $x_1, \ldots, x_n$ — случайные величины. Тогда их совместное распределение можно представить в виде произведения n одномерных условных распределений с постепенно уменьшающейся посылкой:

$$p(x_1, \ldots, x_n) = p(x_n \mid x_1, \ldots, x_{n-1}) \cdots p(x_2 \mid x_1)p(x_1) = p(x_1) \prod_{k=2}^{n} p(x_k \mid x_1, \ldots, x_{k-1}). \tag{3}$$

В дальнейшем мы часто будем сталкиваться с вероятностыми моделями машинного обучения, в которых нужно уметь задавать совместное распределение на все величины, фигурирующие в модели. Работать с одним многомерным распределением, вообще говоря, гораздо сложнее, чем с несколькими одномерными, поэтому для вероятностных моделей машинного обучения совместное распределение очень часто вводится через рассмотренную выше декомпозицию.

Заметим, что при декомпозиции не играет роли порядок выбора величин, для которых мы выписываем условное распределение

$$p(x \mid y)p(y) = p(x, y) = p(y \mid x)p(x). \tag{4}$$

Обобщая это на случай n величин, получаем, что в (3) тоже не важен порядок выбора случайных величин $x_1, \ldots, x_n$ — декомпозиция всё равно будет верна.

Из равенства (4) сразу же получается правило обращения условной вероятности:

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)}. \tag{5}$$

[1] Стоит заметить, что когда пишут $p(x)$, обычно подразумевают плотность в смысле математической статистики. Если случайная величина x дискретна, то $p(x)$ равна вероятности того, что она будет равна какому-то числу x. Если же рассматривается абсолютно непрерывная случайная величина, то $p(x)$ есть плотность в обычном смысле в точке x. Данное обозначение первоначально может казаться очень непривычным, но со временем оно станет интуитивно понятным.

Теперь проинтегрируем обе части равенства (5) по y.[2] Заметим, что слева получится единица, так как интегрируется плотность распределения. Тем самым получаем, что

$$1 = \frac{\int p(x \mid y)p(y)\mathrm{d}y}{p(x)} \quad \Rightarrow \quad p(x) = \int p(x \mid y)p(y)\mathrm{d}y = \int p(x, y)\mathrm{d}y. \tag{6}$$

Данное тождество носит название правила суммирования (sum rule). Оно показывает, как перейти от совместного распределения к маргинальному или же совместному на какое-то подмножество величин: просто интегрируем по всем остальным переменным. Этот процесс называют выинтегрированием (integrate out) или маргинализацией. Поэтому полученное после интегрирования распределение называется маргинальным. Так же, как и с правилом произведения, правило суммирования обобщается по индукции:

Теорема 2 (Правило суммирования). Пусть $x_1, \ldots, x_n$ — случайные величины. Если известно их совместное распределение $p(x_1, \ldots, x_n)$, то совместное распределение подмножества случайных величин $x_1, \ldots, x_k$ будет равно

$$p(x_1, \ldots, x_k) = \int p(x_1, \ldots, x_n)\mathrm{d}x_{k+1} \ldots \mathrm{d}x_n. \tag{7}$$

Теперь посмотрим внимательнее на равенство (6). Можно заметить, что правило суммирования есть не что иное как взятие математического ожидания:

$$p(x) = \int p(x \mid y)p(y)\mathrm{d}y = \mathbb{E}_y[p(x \mid y)].$$

Таким образом, если мы умеем считать $p(x \mid y)$ при всех возможных y, а хотим знать $p(x)$, то нам нужно просто усреднить $p(x \mid y)$ по всем y.

Из правила обращения условной вероятности (5) и правила суммирования (6) получаем широко известную теорему:

Теорема 3 (Байес). Пусть x и y — случайные величины. Тогда

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{\int p(x \mid y)p(y)\mathrm{d}y}. \tag{8}$$

В концептуальной форме это правило звучит так: апостериорное распределение $p(y \mid x)$ (posterior distribution) с точностью до нормировочной константы равно произведению правдоподобия $p(x \mid y)$ (likelihood) и априорного распределения $p(y)$ (prior distribution). Нормировочную константу обычно называют обоснованностью (evidence).

Какой смысл у теоремы Байеса? На самом деле это достаточно простое и элегантное правило, позволяющее уточнять наше незнание о некой величине при поступлении новой информации, косвенно связанной с ней. Пусть $p(y)$ — распределение, которое показывает нашу неопределённость относительно значения y. Теорема Байеса показывает, как наша неопределённость изменилась после наблюдения x (одного или нескольких), который как-то связан с y — то, как именно он связан, задаётся функцией правдоподобия.[3]

Теорема Байеса является частным случаем того, как можно решать обратные задачи: если мы знаем как x влияет на y, то теорема Байеса дает нам возможность узнать, как y влияет на x.

Заметим следующее полезное применение теоремы Байеса. Если задана вероятностная модель (совместное распределение на все переменные), то можно посчитать любое [4] условное распределение. Например, скажем, что на три группы случайных величин x, y и z задана нефакторизуемая вероятностная модель $p(x, y, z)$. Как посчитать $p(x \mid y)$? Достаточно просто:

$$p(x \mid y) = \frac{p(x, y)}{p(y)} = \frac{\int p(x, y, z)\mathrm{d}z}{\iint p(x, y, z)\mathrm{d}x\mathrm{d}z}. \tag{9}$$

[2] Если распределение дискретное, то мысленно заменяйте интеграл на сумму — ситуация не изменится.

[3] Из этой интерпретации и следуют названия распределений: априорное — до эксперимента, апостериорное — после.

[4] На самом деле утверждение о том, что можно посчитать любое условное распределение, верно только в теории: на практике всё упирается в то, получится ли посчитать интегралы.

1.2 Частотный и байесовский подходы

В рамках классических курсов изучался подход, который в англоязычной литературе называют частотным или фреквентистским (frequentist). Вспомним, как в нём решается следующая задача: оценка параметров распределения по выборке из него. Скажем, что есть выборка $X = (x_1, \ldots, x_n)$ из параметрического распределения $p_\theta(x)$. Заметим, что такое распределение вполне можно писать как $p(x \mid \theta)$, т.е. рассматривать параметры θ как случайные величины, — смысл от этого не меняется. Чтобы оценить параметры θ, в классическом частотном подходе используется метод максимального правдоподобия [5] :

$$\theta_{\text{ML}} = \arg\max_\theta p(X \mid \theta) = \arg\max_\theta \prod_{i=1}^{n} p(x_i \mid \theta) = \arg\max_\theta \sum_{i=1}^{n} \log p(x_i \mid \theta). \tag{10}$$

Во многих частных случаях сумма логарифмов правдоподобий будет выпуклой вверх функцией, то есть у неё один максимум, который достаточно легко найти даже в пространствах высокой размерности. Заметим, что θ_{ML} – случайная величина, поскольку она является функцией от выборки.

Оценка максимума правдоподобия (ОМП) обладает очень хорошими свойствами:

- Состоятельность: ОМП сходится к истинному значению параметров по вероятности при $n \to +\infty$ (где n – размер выборки)
- Асимптотическая несмещенность: $\theta_{\text{ML}} = \mathbb{E}[\theta]$ при $n \to +\infty$
- Асимптотическая нормальность: θ_{ML} распределена нормально при $n \to +\infty$
- Асимптотическая эффективность: ОМП обладает наименьшей дисперсией среди всех состоятельных асимптотически нормальных оценок.

Поэтому часто говорят, что лучше ОМП ничего придумать нельзя. Но если всё так хорошо, то зачем вообще нужны другие подходы?

На самом деле всё не так просто. Что мы делаем при оценке максимального правдоподобия? Мы пытаемся найти такие параметры, чтобы вероятность пронаблюдать то, что мы пронаблюдали, была максимальна. Говоря на языке машинного обучения, мы подстраиваем параметры под обучающую выборку. Но мы знаем, что прямая подгонка под данные часто чревата переобучением.

Давайте поймём, какую альтернативу нам дает применение теоремы Байеса. Пусть у нас есть априорное распределение $p(\theta)$, которое отражает некую внешнюю информацию о возможных значениях параметров (если такой информации нет, мы всегда можем ввести неинформативное распределение). Тогда результатом применения теоремы Байеса будет апостериорное распределение на параметры:

$$p(\theta \mid X) = \frac{\prod_{i=1}^{n} p(x_i \mid \theta) \cdot p(\theta)}{\int \prod_{i=1}^{n} p(x_i \mid \theta) \cdot p(\theta) d\theta} \tag{11}$$

Обратите внимание, что теперь ответом является новое распределение на параметры модели, в отличии от метода максимального правдоподобия, где ответом являлось конкретное значение параметров. Сильной стороной данного подхода является то, что при получении апостериорного распределения мы не теряем ни бита информации, которая содержалась в обучающей выборке. В случае же ОМП масса информации теряется (смысл этого утверждения будет показан далее на примерах).

Изобразим таблицу, которая будет показывать различия частотного (классического) и байесовского подходов (см. таблицу 1). Первое и основное отличие состоит в том, как вообще понимать случайность. В частотном подходе предполагается, что случайная величина — это результат некоторого процесса, для которого принципиально невозможно предсказать исход (объективная неопределенность, т.е. у всех одинаковая). В байесовском подходе считается, что процесс на самом деле детерминированный, но часть факторов, которые влияют на этот процесс, неизвестны наблюдателю (субъективное незнание, т.е. у всех разное).

Рассмотрим примеры субъективного незнания.

[5] Напомним, что $p(X \mid \theta)$ – условное распределение на X – называется правдоподобием, если мы рассматриваем его как функцию параметров θ

	Частотный подход	Байесовский подход
Интерпретация случайности	Объективная неопределённость	Субъективное незнание
Виды величин	Случайные и детерминированные	Все величины можно интерпретировать как случайные
Метод вывода	Метод максимального правдоподобия	Теорема Байеса
Виды оценок	Точечная оценка	Апостериорное распределение
Применимость	$n \gg d$	Любое $n \geqslant 0$

Таблица 1: Отличия частотного и байесовского подходов (n — количество элементов в выборке, d — число параметров)

Пример. Допустим, что мы подбрасываем монетку и смотрим, что выпало. В классической теории вероятностей мы привыкли считать, что исход данного эксперимента является объективной неопределенностью, т.е. случайным в частотном смысле. Однако если бы нам были известны все условия эксперимента (переданный импульс, масса монетки, сопротивление воздуха и так далее), то можно было бы с помощью уравнений классической механики точно рассчитать какой стороной упадёт монетка. Мы не можем этого сделать только потому, что нам неизвестны все факторы, влияющие на движение монетки. Таким образом, результат эксперимента является случайной величиной в байесовском смысле.

Пример. Пусть мы каждый день пользуемся автобусом, который по расписанию приходит на остановку в 10:30. Однако в реальности день ото дня автобус то задерживается, то опаздывает, т.е. время его прихода является случайной величиной. Хотя мы не можем сказать, что это объективная неопределенность, так как на время прибытия автобуса в жизни влияет конечный набор факторов (светофоры, пешеходы на переходах и т.д.). И в зависимости от знания этих факторов мы можем точно предсказать время прибытия автобуса. Т.е. это время является случайной величиной в байесовском смысле. Также можно заметить, что в зависимости от степени субъективного незнания наблюдатель может предсказать время прибытия с разной точностью. Например, мы, исходя из наших ежедневных наблюдений, можем сказать, что среднее отклонение от расписания у автобуса ± 7 минут. А наш товарищ пользуется программой, которая отображает в реальном времени положение автобуса. И он может предсказывать время прибытия с точностью ± 3 минуты. Таким образом, с точки зрения обоих наблюдателей время прихода автобуса — случайная величина, но степень субъективного незнания о ней у них разная.

Стоит заметить, что в реальности существуют примеры объективных неопределенностей — это процессы, являющиеся результатом квантово-механических эффектов (например, распады радиоактивных ядер).

Перейдем к видам величин. В байесовском подходе вообще все величины можно считать случайными. Все параметры модели, которые мы не знаем, мы считаем случайными и задаем на них априорные распределения. А если параметр нам известен, то мы можем задать его распределение дельта-функцией и продолжать считать его случайной величиной. В частотном же подходе параметры распределения считаются неизвестными детерминированными величинами. Отсюда вытекает отличие в методе оценивания параметров модели: в байесовском подходе мы уменьшаем наше незнание, получая апостериорное распределение по формуле Байеса, а в частотном — находим конкретные значения параметров с помощью ОМП.

Последнее отличие состоит в том, когда какой подход можно применять. У метода максимального правдоподобия есть одна проблема: все его свойства асимптотические, то есть они выполняются при $n \to +\infty$. В байесовском подходе такого ограничения нет: выводы можно делать при любом $n \geqslant 0$.[6] Таким образом, при малых значениях n гарантии на ОМП не выполняются, и лучше работает байесовский подход. А какой метод лучше применять

[6] Формально их можно сделать даже при $n = 0$ — в таком случае оценкой будет выступать априорное распределение.

при больших n? Оказывается, что при больших размерах выборки один подход переходит в другой: можно показать, что при $n \to +\infty$ апостериорное распределение коллапсирует в дельта-функцию в точке максимума правдоподобия. Поэтому можно не мучиться с байесовским выводом апостериорных распределений и применять частотный подход.

Тут у самых въедливых читателей должен возникнуть вопрос, а зачем мы в век больших данных вообще рассуждаем про малые выборки? Строго говоря, мы должны сделать оговорку, что размер выборки мы должны сравнивать с числом параметров модели. И вот если $n/d \to \infty$ то мы можем использовать ОМП. Но в современных нейросетях часто возникает ситуация, когда $n/d \ll 1$, что ставит под сомнение корректность применения метода максимального правдоподобия.

1.3 Приятные плюсы байесовского подхода

1. Регуляризация: за счёт введения априорного распределения на параметры получается так, что они не слишком «подгоняются» под данные.

2. Композитность: есть возможность постепенно улучшать предсказание на параметры, если предыдущий результат вывода считать априорным распределением при поступлении новых данных. Действительно, если x — имеющиеся данные, y — оцениваемый параметр, а z — это другие данные (предполагается, что они не зависят от x), то

$$p(y \mid x, z) = \frac{p(z \mid y)p(y \mid x)}{\int p(z \mid y)p(y \mid x)\mathrm{d}y}. \tag{12}$$

3. Обработка данных «на лету»: нет необходимости хранить все данные для построения прогноза — достаточно хранить апостериорное распределение и постепенно его пересчитывать: оно будет хранить в себе информацию из всех данных.

4. Построение моделей с скрытыми (латентными) переменными: возможность корректно обрабатывать пропуски в данных (об этом будет рассказано позднее).

5. Масштабируемость: в некоторых случаях байесовский подход переносится на большие данные, при этом оставаясь вычислительно эффективным. Это свойство подробнее будет описываться на курсе нейробайесовских методов.

1.4 Байесовский подход как обобщение булевой логики

Байесовский подход можно рассматривать, в том числе как обобщение булевой логики. В классической логике есть единственное правило для построения рассуждений, а именно modus ponens: если A истинно и из A следует B, то B истинно. Пусть теперь известно, что B истинно и из A следует B. В таком случае про истинность A ничего сказать нельзя. Однако это несколько не соответствует здравому смыслу. Предположим, что днём прошёл матч Италия – Франция, а вечером болельщики с итальянскими флагами радостно пьют пиво в баре. Интуитивно понятно, что в таком случае выиграла Италия, но логика так делать запрещает. Теперь попробуем применить теорему Байеса, но сначала перепишем аналог modus ponens. Если нам известны $p(A)$ и $p(B \mid A)$, то несложно посчитать $p(B)$ по правилу произведения и суммирования

$$p(B) = \sum_A p(B \mid A)p(A) \tag{13}$$

Обратная задача будет звучать так: нам известны $p(B \mid A)$, $p(A)$ и известно то, что B произошло; что можно сказать про A? По теореме Байеса можно сразу же рассчитать $p(A \mid B)$:

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{\sum\limits_A p(B \mid A)p(A)} \tag{14}$$

Тем самым в байесовском подходе можно сделать то, чего нельзя сделать в булевой логике.

1.5 Пример байесовских рассуждений

Предположим, что в квартире установлена сигнализация. Её изготовитель утверждает, что она гарантированно сработает на грабителя, но в 10% случаев бывают ложные срабатывания из-за небольших землетрясений, о которых иногда предупреждают по радио. Попробуем задать это в виде вероятностной модели. Пусть есть четыре случайные величины:

- $a \in \{0, 1\}$ — индикатор того, что сработала сигнализация,
- $t \in \{0, 1\}$ — индикатор того, что грабитель проник в квартиру,
- $e \in \{0, 1\}$ — индикатор того, что произошло небольшое землетрясение,
- $r \in \{0, 1\}$ — индикатор того, что о землетрясении объявили по радио.

Изобразим связи этих величин в виде ориентированного графа, где ребро из b в a означает то, что a зависит от b:

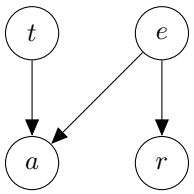


Рис. 1: Граф зависимостей в задаче про сигнализацию.

По такому графу несложно задать совместное распределение на все величины:

$$p(a, e, r, t) = p(a \mid e, t)p(r \mid e)p(t)p(e).$$

Осталось задать эти распределения. Запишем распределения на a и на r в виде таблиц:

$p(a = 1 \mid e, t)$	$t = 0$	$t = 1$
$e = 0$	0	1
$e = 1$	0.1	1

	$p(r = 1 \mid e)$
$e = 0$	0
$e = 1$	0.5

Для распределений на t и на e скажем, что $p(t = 1) = 2 \cdot 10^{-4}$, $p(e = 1) = 10^{-2}$. Теперь можно считать разные вероятности.

Предположим, что пришло уведомление о том, что в квартиру вломились. Нужно ли вызывать полицию или же срабатывание ложное? Другими словами, нужно посчитать вероятность $p(t = 1 \mid a = 1)$. Для этого воспользуемся теоремой Байеса:

$$p(t = 1 \mid a = 1) = \frac{p(a = 1 \mid t = 1)p(t = 1)}{p(a = 1 \mid t = 0)p(t = 0) + p(a = 1 \mid t = 1)p(t = 1)}. \tag{15}$$

Сразу заметим, что $p(a = 1 \mid t = 1) = 1$. Далее, по правилу суммирования

$$\begin{aligned} p(a = 1 \mid t = 0) &= p(a = 1 \mid e = 0, t = 0)p(e = 0) + p(a = 1 \mid e = 1, t = 0)p(e = 1) \\ &= 0 + 0.1 \cdot 10^{-2} = 10^{-3} \end{aligned} \tag{16}$$

Тогда

$$p(t = 1 \mid a = 1) = \frac{1 \cdot 2 \cdot 10^{-4}}{10^{-3} \cdot (1 - 2 \cdot 10^{-4}) + 1 \cdot 2 \cdot 10^{-4}} \approx \frac{1}{6} \tag{17}$$

Тем самым, скорее всего было ложное срабатывание. Но что будет, если квартира расположена в криминальном районе и $p(t = 1) = 2 \cdot 10^{-3}$? В таком случае ситуация кардинально меняется, так как вероятность будет примерно равна $2/3$, т.е. примерно 67%.

Теперь пусть квартира находится в криминальном районе, сработала сигнализация, но при этом по радио было объявлено о землетрясении. Какова вероятность ограбления в

таком случае? Другими словами, нужно найти $p(t = 1 \mid a = 1, r = 1)$. Воспользуемся определением условной вероятности, правилом суммирования и правилом произведения:

$$p(t = 1 \mid a = 1, r = 1) = \frac{p(a = 1, t = 1, r = 1)}{p(a = 1, r = 1)} = \frac{\sum\limits_{e} p(a = 1, e, t = 1, r = 1)}{\sum\limits_{e,t} p(a = 1, e, t, r = 1)} \tag{18}$$

$$= \frac{\sum\limits_{e} p(a = 1 \mid e, t = 1)p(r = 1 \mid e)p(e)p(t = 1)}{\sum\limits_{e,t} p(a = 1 \mid e, t)p(r = 1 \mid e)p(e)p(t)}. \tag{19}$$

Заметим, что достаточно смотреть только на слагаемые с $e = 1$. Тогда

$$p(t = 1 \mid a = 1, r = 1) = \frac{1 \cdot 0.5 \cdot 10^{-2} \cdot 2 \cdot 10^{-3}}{10^{-1} \cdot 0.5 \cdot 10^{-2} \cdot (1 - 2 \cdot 10^{-3}) + 1 \cdot 0.5 \cdot 10^{-2} \cdot 2 \cdot 10^{-3}}. \tag{20}$$

После упрощений получим, что эта вероятность примерно равна $1/51$, то есть около 2%. Обратите внимание, как трансформируются наши предположения о наличии вора в квартире при поступлении новой информации (сравните с предыдущим результатом, когда у нас не было никакой информации о землетрясении).

Узнав это, владелец квартиры спокойно продолжил заниматься своими делами. Вечером он возвращается в квартиру и видит, что она обчищена. Вопрос: что пошло не так? Выкладки верны, но вероятностная модель неправильная. Нужно было учесть то, что грабители тоже могут слушать радио и использовать факт о ложных срабатываниях: $p(t, e) \neq p(t)p(e)$ и $p(t = 1 \mid e = 1) > p(t = 1 \mid e = 0)$.

2 Лекция 2. Сопряженные распределения, экспоненциальный класс распределений

2.1 Сопряжённые распределения

Пусть нам дана выборка из некоторого параметрического семейства $X = \{x_i\}_{i=1}^n, \quad x_i \sim p(x \mid \theta)$, и у нас есть некоторое априорное распределение на параметры $p(\theta)$. Тогда, пользуясь формулой Байеса, мы можем найти апостериорное распределение на θ при условии того, что мы пронаблюдали X.

$$p(\theta \mid X) = \frac{p(X \mid \theta)p(\theta)}{\int p(X \mid \theta)p(\theta)\mathrm{d}\theta} \tag{21}$$

К сожалению, интеграл в числителе берется аналитически в очень редких случаях. Поэтому в дальнейших лекциях курса мы много будем говорить о различных способах оценки апостериорного распределения.

Однако, давайте подумаем, что мы можем сделать, не зная значение интеграла. Например, вполне несложно найти максимум апостериорного распределения. Действительно:

$$\theta_{MP} = \arg\max_\theta p(\theta \mid X) = \arg\max_\theta p(X \mid \theta)p(\theta) = \tag{22}$$

$$= \arg\max_\theta \left(\prod_{i=1}^n p(x_i \mid \theta)p(\theta) \right) = \arg\max_\theta \left(\sum_{i=1}^n \ln p(x_i \mid \theta) + \ln p(\theta) \right) \tag{23}$$

Получили довольно известную регуляризацию на давно знакомую оценку максимального правдоподобия. Так, например, если в качестве априорного распределения мы возьмём нормальное распределение с нулевым математическим ожиданием и некоторой дисперсией λ^{-1}, регуляризация превратится в $\lambda\|\theta\|^2$, то есть L2-регуляризацию.

Однако, хоть мы и получили в каком-то смысле неплохую точечную оценку на θ, у такого метода есть ряд минусов:

- Нет оценки неопределённости. Зачастую в прикладных задачах нам важно не только получить ответ на вопрос, но и понимать, насколько мы в нём уверены. Если у нас есть апостериорное распределение, мы можем построить доверительные интервалы на θ_{MP}, чтобы понимать, в каких пределах может меняться полученное значение. Точечная оценка не дает нам такой возможности.

- Нет возможности объединения информации, полученной из различных источников. Одним из плюсов байесовского подхода является то, что мы можем сложные вероятностные модели строить из простых, как из кирпичиков. Рассчитав апостериорное распределение при условии выборки из одного источника, мы можем подать его в качестве априорного распределения для расчета апостериорного распределения при условии выборки из другого источника. Таким образом, в итоговом апостериорном распределении будет содержаться вся информация от обоих источников. Если же у нас есть только точечная оценка на параметры модели, такого элегантного объединения информации из разных источников у нас сделать не получится.

- Мода распределения может быть нерепрезентативна. Пример такого распределения можно увидеть на Рисунке 2.

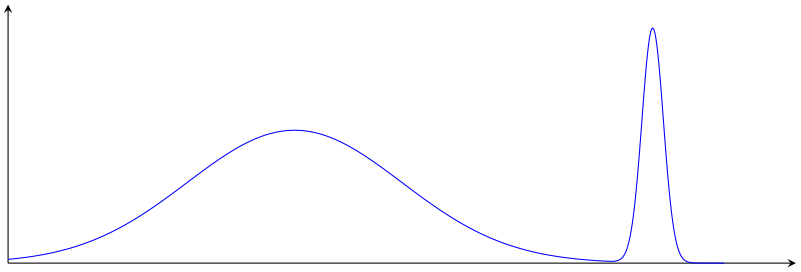


Рис. 2: Пример распределения, у которого мода нерепрезентативна.

Метод замены апостериорного распределения его модой получил название “Байес для бедных” (“Poor man’s Bayes”), как довольно простой вычислительно, но имеющий весомые недостатки. Подробно изучать его мы не будем; предполагается, что он уже достаточно знаком из прочих курсов по машинному обучению. Нас же интересуют более эффективные и интересные подходы к байесовскому выводу.

Начнём с рассмотрения важного частного случая, когда интеграл аналитически вычислить всё-таки возможно: это случай сопряжённых семейств распределений.

Определение 2. Пусть функция правдоподобия и априорное распределение принадлежат некоторым параметрическим семействам распределений: $p(X \mid \theta) \sim \mathcal{A}(\theta)$ и $p(\theta \mid \beta) \sim \mathcal{B}(\beta)$. Семейства $\mathcal{A}$ и $\mathcal{B}$ являются сопряжёнными (conjugate) тогда и только тогда, когда $p(\theta \mid X) \sim \mathcal{B}(\beta')$.

Из этого определения следует, что если функция правдоподобия $p(X \mid \theta)$ и априорное распределение $p(\theta \mid \beta)$ сопряженны, то апостериорное распределение $p(\theta \mid X)$ лежит в том же параметрическом семействе $\mathcal{B}(\beta')$, что и априорное $p(\theta \mid \beta)$. То есть, апостериорное распределение $p(\theta \mid x)$ можно вычислить аналитически. Рассмотрим несколько примеров:

1. Пусть функция правдоподобия $p(x \mid \mu) = \mathcal{N}(x \mid \mu, 1)$. Как будет выглядеть сопряженное ему $p(\mu)$?

$$p(x \mid \mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2} + x\mu - \frac{\mu^2}{2}\right) \tag{24}$$

Нужно подобрать такое $p(\mu)$, чтобы его функциональный вид не изменился при умножении на вышеприведённое выражение (“перевёрнутая парабола под экспонентой”). Легко заметить, что для этого нам подойдёт такой же вид:

$$p(\mu) = \frac{1}{\sqrt{2\pi}s} \exp\left(-\frac{\mu^2}{2s^2} + \frac{\mu m}{s^2} - \frac{m^2}{2s^2}\right) = \mathcal{N}(\mu \mid m, s^2) \tag{25}$$

Теперь проверим:

$$p(x \mid \mu)p(\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2} + x\mu - \frac{\mu^2}{2}\right) \cdot \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{\mu^2}{2s^2} + \frac{\mu m}{s^2} - \frac{m^2}{2s^2}\right) \propto \tag{26}$$

$$\propto \exp\left(-\frac{\mu^2(s^2+1)}{2s^2} + \frac{\mu(m+xs^2)}{s^2} - \frac{x^2s^2+m^2}{2s^2}\right) \propto \exp\left(-\frac{s^2+1}{2s^2}\left(\mu - \frac{m+xs^2}{s^2+1}\right)^2\right) \propto \tag{27}$$

$$\propto \mathcal{N}\left(\mu \mid \frac{m+xs^2}{s^2+1}, \frac{s^2}{s^2+1}\right) \tag{28}$$

Действительно, получили аналитический вид для апостериорного распределения $p(\mu \mid X)$, и оказалось, что $p(\mu \mid X)$ тоже лежит в семействе нормальных распределений.

2. $p(x \mid \gamma) = \mathcal{N}(x \mid 0, \gamma^{-1})$; $p(\gamma)-?$

$$p(x \mid \gamma) = \sqrt{\frac{\gamma}{2\pi}} \exp\Big(-\frac{\gamma}{2}x^2\Big)$$

Получили корень из γ, умноженный на экспоненту линейной функции. Вопрос: какой функциональный вид должно иметь априорное распределение?

$$p(\gamma) = \frac{\beta^\alpha}{\Gamma(\alpha)}\gamma^{\alpha-1}\exp(-\gamma\beta) \sim \mathrm{G}(\gamma \mid \alpha, \beta)$$

3. $p(x \mid \mu, \gamma) \sim \mathcal{N}(x \mid \mu, \gamma^{-1})$; $p(\mu, \gamma)-?$

Сразу хочется сослаться на два предыдущих пункта и записать $p(\mu, \gamma) = p(\mu)p(\gamma)$. Но действительно ли это выполняется?

$$p(x \mid \mu, \gamma) = \sqrt{\frac{\gamma}{2\pi}} \exp\Big(-\frac{\gamma}{2}(x-\mu)^2\Big) = \sqrt{\frac{\gamma}{2\pi}} \exp\left(-\frac{\gamma x^2}{2} + \gamma\mu x - \frac{\gamma x^2}{2}\right)$$

Заметим, что это выражение не факторизуется по μ и γ. Значит, и априорное распределение, если оно сопряжено, факторизоваться не может.

На самом деле сопряжённым распределением является так называемое гамма-нормальное распределение:

$$p(\mu, \gamma) = p(\mu \mid \gamma)p(\gamma) = \mathcal{N}(\mu \mid m, (\lambda\gamma)^{-1})\mathrm{G}(\gamma \mid a, b)$$

Теперь посмотрим, как производить поиск сопряженных распределений не для каждого частного случая, а в некотором общем виде.

2.2 Экспоненциальный класс распределений

До этого мы с вами рассматривали параметрические распределения, подразумевая, что плотность нам известна с точностью до некоторого параметра θ. Такие множества распределений мы называли параметрическими семействами. Теперь мы перейдем к понятию класса распределений, который будем задавать с точностью до функционального вида.

Определение 3. Будем говорить, что распределение $p(x \mid \theta)$ лежит в экспоненциальном классе, если оно может быть представлено в следующем виде

$$p(x \mid \theta) = \frac{f(x)}{g(\theta)} \exp\big(\theta^T u(x)\big), \quad f(\cdot) \geqslant 0, \;\; g(\cdot) > 0, \tag{29}$$

Параметры θ называются естественными параметрами.

Несмотря на довольно необычный вид выражения, оказывается, что подавляющее большинство табличных распределений лежит в экспоненциальном классе (нормальное, все дискретные распределения, бета-распределение, гамма-распределение, хи-квадрат распределение и т.д.). То есть большинство распределений, с которыми приходится иметь дело в прикладных задачах, принадлежат экспоненциальному классу распределений.[7] Такие распределения обладают несколькими довольно примечательными свойствами, и мы рассмотрим некоторые из них. Начнем с достаточных статистик.

Для начала вспомним, что же такое достаточная статистика распределения. Неформальное определение можно сформулировать так: достаточная статистика — это функция от выборки, которая содержит всю информацию, необходимую для оценки параметров неизвестного распределения.

Определение несколько размытое. Формализуем его, воспользовавшись критерием факторизации Фишера:

[7] Стоит заметить, что такое популярное в приложениях распределение, как смесь нормальных распределений, не принадлежит экспоненциальному классу

Определение 4. $a(X)$ — достаточна тогда и только тогда, когда $p(X \mid \theta) = f_1(X) f_2(\theta, a(X))$

В общем случае таких статистик может не быть. Однако для экспоненциального класса распределений они существуют. Из функционального вида распределения и критерия Фишера легко следует, что $u(x)$ является достаточной статистикой (можно взять $f_1(X) = f(X), f_2(\theta, u(X)) = \frac{\exp\left(\theta^T u(X)\right)}{g(\theta)}$).

Рассмотрим одно замечательное свойство экспоненциального класса распределений. Заметим, что

$$g(\theta) = \int f(x) \exp\left(\theta^T u(x)\right) \mathrm{d}x, \quad \text{т.к.} \quad \int \frac{f(x)}{g(\theta)} \exp\left(\theta^T u(x)\right) \mathrm{d}x = 1 \tag{30}$$

Прродифференцируем по θ_j

$$\frac{\partial}{\partial \theta_j} g(\theta) = \frac{\partial}{\partial \theta_j} \int f(x) \exp\left(\theta^T u(x)\right) \mathrm{d}x = \int f(x) \exp\left(\theta^T u(x)\right) u_j(x) \mathrm{d}x = \tag{31}$$

$$= g(\theta) \int \frac{f(x)}{g(\theta)} \exp\left(\theta^T u(x)\right) u_j(x) \mathrm{d}x = g(\theta) \int p(x \mid \theta) u_j(x) \mathrm{d}x = g(\theta) \mathbb{E}_{x \sim p(x \mid \theta)} u_j(x) \tag{32}$$

В итоге получаем, что

$$\frac{\partial}{\partial \theta_j} \log g(\theta) = \mathbb{E}_{x \sim p(x \mid \theta)} u_j(x) \tag{33}$$

Таким образом, мы получили простой способ находить математическое ожидание от достаточной статистики для распределения из экспоненциального класса — нужно просто продифференцировать логарифм его нормировочной константы. Аналогично можно показать, что

$$\frac{\partial^2}{\partial \theta_j \partial \theta_k} \log g(\theta) = \operatorname{cov}(u_j(x), u_k(x)) \tag{34}$$

Из выражения 34 следует, что $\nabla^2 \log g(\theta) \succ 0$ (положительно определенная матрица), то есть $\log g(\theta)$ — выпуклая функция.

2.2.1 Оценка параметров распределения из экспоненциального класса

Пусть нам дана выборка из распределения экспоненциального класса:

$$X = \{x_i\}_{i=1}^n, \quad x_i \sim p(x \mid \theta) = \frac{f(x)}{g(\theta)} \exp\left(\theta^T u(x)\right)$$

Оценим параметры распределения методом максимального правдоподобия

$$\theta_{ML} = \arg\max_\theta p(X \mid \theta) = \arg\max_\theta \sum_{i=1}^n \log p(x_i \mid \theta) =$$

$$= \arg\max_\theta \sum_{i=1}^n \Big(\log f(x_i) - \log g(\theta) + \theta^T u(x_i) \Big) \tag{35}$$

Важно заметить, что $\log f(x_i) - \log g(\theta) + \theta^T u(x_i)$ — строго вогнутая по θ функция. Этот факт следует из выпуклости $\log g(\theta)$ и вогнутости линейной функции, то есть максимум $p(X \mid \theta)$ единственный по θ и легко находиться даже численными методами. Продифференцировав по θ_j выражение 2.2.1, приравняем производную к нулю и получим

$$\frac{1}{n} \sum_{i=1}^n u_j(x_i) = \frac{\partial \log g(\theta)}{\partial \theta_j} = \mathbb{E}_{x \sim p(x \mid \theta)} u_j(x)$$

Получается, что мы должны подстроить параметры распределения так, чтобы выборочное среднее достаточных статистик совпало с их математическим ожиданием.

Пример. Рассмотрим в качестве примера нормальное распределение

$$p(x \mid \theta) = \mathcal{N}(x \mid \mu, \gamma^{-1}) = \sqrt{\frac{\gamma}{2\pi}} \exp\Big\{ \frac{\gamma}{2} x^2 + \gamma\mu x - \frac{\gamma}{2}\mu^2 \Big\}$$

Из выражения выше видно, что

$$\theta_1 = \frac{\gamma}{2} \quad u_1(x) = x^2$$

$$\theta_2 = \gamma\mu \quad u_2(x) = x$$

$$g(\theta) = \sqrt{\frac{2\pi}{\gamma}} \exp\Big\{ \frac{\gamma}{2}\mu^2 \Big\}$$

2.2.2 Сопряженное семейство к экспоненциальному классу

Запишем общий вид сопряжённого распределения, исходя из функционального вида распределения из экспоненциального класса:

$$p(\theta \mid \eta, \nu) = \exp\big(\theta^T \eta\big) \frac{1}{g^\nu(\theta)} \frac{1}{h(\eta, \nu)} \tag{36}$$

Всё довольно очевидно, кроме последнего множителя. Может показаться, что нет гарантий на существование нормировочной константы для любых η и ν, так как интеграл может быть невозможно вычислить аналитически. Это не зря — её действительно может не быть, и это будет означать не существование аналитически заданного сопряжённого семейства.

Вычислим апостериорное распределение:

$$p(\theta \mid X) = \frac{1}{Z} \prod_{i=1}^{n} p(x_i \mid \theta) p(\theta \mid \nu, \eta) = \tag{37}$$

$$= \frac{1}{Z} \prod_{i=1}^{n} \Big[f(x_i) \Big] \cdot \frac{1}{g^n(\theta)} \exp\left\{ \theta^T \left(\sum_{i=1}^{n} u(x_i) \right) \right\} \exp\big\{ \theta^T \eta \big\} \frac{1}{g^\nu(\theta)} \frac{1}{h(\eta, \nu)} = \tag{38}$$

$$= \frac{1}{Z'} \exp\left\{ \theta^T \left(\eta + \sum_{i=1}^{n} u(x_i) \right) \right\} \frac{1}{g^{\nu+n}(\theta)} = \frac{1}{h(\eta', \nu')} \exp\left(\theta^T \eta'\right) \frac{1}{g^{\nu'}(\theta)} \tag{39}$$

Легко заметить, что функциональный вид действительно совпадает. Так же видно, как именно мы пересчитываем η и ν при переходе к апостериорному распределению:

$$\eta' = \eta + \sum_{i=1}^{n} u(x_i) \tag{40}$$

$$\nu' = \nu + n \tag{41}$$

Из полученных выражений можно понять физический смысл параметров этого распределения. Параметр ν отвечает количеству проведенных экспериментов, а параметр η — сумме достаточных статистик в этих экспериментах.

3 Лекция 3. Байесовские методы выбора моделей. Принцип наибольшей обоснованности.

В этой лекции мы будем говорить о байесовских критериях выбора модели. Для начала вспомним, какие есть общенаучные принципы для выбора одной теории из нескольких.

3.1 Бритва Оккама. Критерий фальсифицируемости Поппера.

Современная наука пытается находить наиболее простые объяснения наблюдаемым явлениям, следуя бритве Оккама: из нескольких объяснений одного и того же явление выбирается самое простое.

Пример: геоцентрическая система против гелиоцентрической. Геоцентрическая система исходно обладала большей простотой и элегантностью по сравнению с гелиоцентрической. Невооруженным взглядом видно, что Солнце и планеты описывают полуокружности на небесной сфере. И наиболее простым объяснением этого феномена является геоцентрическая система: Солнце и планеты движутся по окружностям, в центре которых находится Земля. Но с появлением все более совершенной оптики выяснилось, что небесные тела описывают не ровные окружности, а с некоторыми колебаниями. Чтобы согласовать теорию с экспериментом, придумали поправку: тела движутся по окружностям вокруг Земли, но при этом ещё описывают маленькую окружность (эпицикл) вокруг центра, движущегося по большой окружности. При дальнейших экспериментальных уточнениях траекторий стали обнаруживаться все новые и новые несоответствия теории с экспериментом, что побудило ученых ввести еще несколько поправок. Таким образом, получая новые данные, люди продолжали увеличивать сложность модели, и в итоге количество эпициклов дошло примерно до 20. В этот момент оказалось, что гелиоцентрическая модель гораздо проще и при этом так же хорошо описывает наблюдаемые данные. Поэтому она и вытеснила геоцентрическую.

Еще одним важным принципом является критерий фальсифицируемости Карла Поппера: чтобы теория считалась научной, должен существовать эксперимент (даже мысленный), при определенном исходе которого можно признать теорию неверной.

Пример ненаучного утверждения: «На всё воля Божья». Этим утверждением можно объяснить любое явление, и опровергнуть его экспериментально невозможно. Пример научного утверждения: «Основной причиной глобального потепления климата является деятельность человека». Это утверждение можно опровергнуть экспериментально, измерив, как на повышение температуры влияют природные процессы (активность Солнца, вулканов, прецессия Земли и т.д.) и антропогенные процессы (промышленность, сельское хозяйство, транспорт и т.д.). И если окажется, что вклад природных процессов в глобальное потепление больше, то утверждение будет опровергнуто.

А теперь посмотрим, как принцип Оккама и критерий фальсифицируемости Поппера могут быть переформулированы с философского на математический язык.

3.2 Вероятностные модели

Для начала определимся, что мы будем называть моделью. В машинном обучении мы обычно имеем дело с тремя видами переменных: x — наблюдаемые переменные, t — целевые переменные, θ — параметры алгоритма прогнозирования. Одна из распространённых постановок задач машинного обучения состоит в следующем. Дана выборка независимых одинаково распределённых объектов. Описание каждого объекта задается парой вида (x, t):

$$(X_{tr}, T_{tr}) = \{x_i, t_i\}_{i=1}^{n} \tag{42}$$

Анализируя обучающую выборку, необходимо подобрать алгоритм (подстроить его параметры θ), который позволил бы по x спрогнозировать значение t. Для решения этой задачи часто вводят модель, описывающую способ порождения данных. На вероятностном языке такой моделью является совместное распределение на переменные x, t и θ. Традиционно выделяют 2 вида моделей:

1. Генеративная модель

$$p(x, t, \theta) = p(x, t \mid \theta)p(\theta) = p(t \mid x, \theta)p(x \mid \theta)p(\theta) \tag{43}$$

Здесь и далее мы используем стандартное предположение о том, что априорные знания о параметрах не зависят от данных.

2. Дискриминативная модель

$$p(t, \theta \mid x) = p(t \mid x, \theta)p(\theta) \tag{44}$$

Генеративная модель более общая, поскольку если нам известно $p(x, t, \theta)$, то мы всегда можем получить $p(t, \theta \mid x)$. Обратное, вообще говоря, неверно. Кроме того, несомненным достоинством генеративной модели является возможность порождать новые x, или же пары (x, t). В рамках дискриминативной модели такое сделать не получится.

Однако, в традиционном машинном обучении чаще рассматривают дискриминативные модели. При этом на практике часто оказывается так, что пространство целевых переменных проще, чем пространство наблюдаемых переменных. Поэтому традиционные дискриминативные модели обычно на порядок проще генеративных, так как они решают гораздо более простую задачу. Например, пусть пространство наблюдаемых переменных — картины известных художников, а пространство целевых переменных — имена этих художников. Тогда определить автора по картине (дискриминативная задача) проще, чем нарисовать картину в стиле автора (генеративная задача). Однако, многие современные дискриминативные модели на практике такие же сложные как и генеративные, потому что пространство целевых переменных не проще пространства наблюдаемых переменных. Например, в задаче машинного перевода с немецкого на французский: x — предложение на немецком, t — предложение на французском.

3.3 Обучение дискриминативных вероятностных моделей

Начнем изучение вероятностных моделей с дискриминативных (хотя, вообще говоря, содержание данного раздела справедливо и для генеративных моделей). Напомним общий вид вероятностной дискриминативной модели

$$p(t, \theta \mid x) = p(t \mid \theta, x)p(\theta) \tag{45}$$

На этапе обучения модели основная задача — оценить ее параметры θ, т.е. найти апостериорное распределение на θ при условии обучающей выборки $(X_{tr}, T_{tr}) = \{(x_i, t_i)\}_{i=1}^n$. На этапе применения необходимо для нового объекта X_{test} предсказать значение целевой переменной T_{test} с учетом извлеченных из обучающей выборки закономерностей, т.е. найти прогнозное распределение на T_{test} при условии X_{test}, X_{tr}, T_{tr}.

Этап	Дано	Неизвестно	Хотим оценить (Freq)	Хотим оценить (Bayes)
Обучение	(X_{tr}, T_{tr})	θ	$\theta_{ML} = arg\max_\theta p(T_{tr} \mid X_{tr}, \theta)$	$p(\theta \mid X_{tr}, T_{tr})$
Тестирование	X_{test}	T_{test}	$p(T_{test} \mid X_{test}, \theta_{ML})$	$p(T_{test} \mid X_{test}, X_{tr}, T_{tr})$

Таблица 2: Схема обучения и применения дискриминативной модели

Апостериорное распределение на параметры θ можно найти, сделав байесовский вывод:

$$p(\theta \mid X_{tr}, T_{tr}) = \frac{p(T_{tr} \mid X_{tr}, \theta)p(\theta)}{\int p(T_{tr} \mid X_{tr}, \theta)p(\theta)d\theta} \tag{46}$$

Прогнозное распределение на значение целевой переменной T_{test} для нового объекта X_{test} можно вычислить по правилу суммирования, с использованием апостериорного распределения на параметры модели θ, полученного на этапе обучения.

$$p(T_{test} \mid X_{test}, X_{tr}, T_{tr}) = \int p(T_{test} \mid X_{test}, \theta) p(\theta \mid X_{tr}, T_{tr}) d\theta \tag{47}$$

На данном этапе мы по сути делаем следующее: применяем все возможные (со всеми возможными значениями θ) алгоритмы прогнозирования $p(T_{test} \mid X_{test}, \theta)$ и усредняем полученные значения с весами, которые задаются нам апостериорным распределением $p(\theta \mid X_{tr}, T_{tr})$. Т.е. интеграл в выражении 47 можно рассматривать как взвешенное усреднение по алгоритмам прогнозирования.[8] Важно отметить, что качество предсказания такого ансамбля моделей оказывается лучше, чем качество предсказания лучшей из этих моделей.

Но что делать, если аналитический байесовский вывод по формуле 46 невозможен, т.е. если интеграл в знаменателе формулы Байеса не берется? В этом случае есть два пути: приближенно оценить апостериорное распределение[9] или перейти к точечной оценке параметров, воспользовавшись уже знакомым нам «Байесом для бедных»:

$$\theta_{MP} = \arg\max_{\theta} p(\theta \mid X_{tr}, T_{tr}) \tag{48}$$

Здесь параметры θ оцениваются только в одной точке, что соответствует замене честного апостериорного распределения 46 на дельта-функцию с центром в точке θ_{MP}

$$p(\theta \mid X_{tr}, T_{tr}) \approx \delta(\theta - \theta_{MP}) \tag{49}$$

Подставив данное приближение в интеграл для прогнозного распределения 47, получим

$$p(T_{test} \mid X_{test}, X_{tr}, T_{tr}) \approx p(T_{test} \mid X_{test}, \theta_{MP}) \tag{50}$$

«Байес для бедных» — вычислительно эффективная и просто реализуемая процедура. Однако она приводит к потерям информации на этапе обучения, что влечет за собой потерю ансамбля на этапе применения. Что в свою очередь ведет к потерям качества.

3.4 Принцип наибольшей обоснованности

Все предыдущие рассуждения делались в предположении о том, что мы уже выбрали и зафиксировали вероятностную модель $p(t, \theta \mid x)$. А что будет если моделей несколько?

Пусть дана обучающая выборка (X_{tr}, T_{tr}). Предположим, что у нас есть три различных варианта задания вероятностной модели:

$$p_j(t, \theta \mid x) = p_j(t \mid x, \theta) p_j(\theta), \quad j = 1, 2, 3 \tag{51}$$

Теперь из этих моделей нужно выбрать ту, которая не только хорошо описывает обучающую выборку, но и обладает наибольшей обобщающей способностью. Как выразить обобщающую способность на математическом языке? С этой проблемой человечество столкнулось уже давно, и на сегодняшний день существует множество различных критериев[10] В нашем курсе мы рассмотрим один из них — принцип наибольшей обоснованности[11]. Как мы увидим далее, этот принцип в некотором смысле является математическим аналогом Бритвы Оккама и критерия фальсифицируемости Поппера.

Теорема 4 (Принцип наибольшей обоснованности). Лучшая модель выбирается по правилу:

$$j^* = \arg\max_j p_j(T_{tr} \mid X_{tr}) = \arg\max_j \int p_j(T_{tr} \mid X_{tr}, \theta) p_j(\theta) d\theta \tag{52}$$

Распределение $p_j(T_{tr} \mid X_{tr})$ называется обоснованностью (evidence). Напомним, что именно эта величина стоит в знаменателе теоремы Байеса (см. выражение 46). Заметим, что по

[8] Типичный пример ансамблирования или взвешенного голосования

[9] О различных способах приближенной оценки апостериорного распределения мы поговорим в следующих лекциях.

[10] Например, теория Вапника-Червоненкиса, принцип минимизации длины описания, информационные критерии Акаике и Байеса-Шварца.

[11] впервые был предложен в 1992 году британским физиком Дэвидом МакКаем

параметрам модели θ мы проводим маргинализацию, поэтому от конкретных значений параметров обоснованность не зависит.

Физический смысл обоснованности модели следующий: насколько вероятно в рамках данной модели пронаблюдать обучающую выборку. Поэтому чем выше обоснованность, тем лучше модель описывает наблюдаемые данные. По сути, принцип наибольшей обоснованности является методом максимума правдоподобия, но не в пространстве параметров модели θ, а в пространстве моделей j.

Давайте теперь убедимся, что приведённый выше критерий можно назвать математической формализацией Бритвы Оккама и критерия фальсифицируемости Поппера. Изобразим для каждой из трех моделей совместное распределение на параметры θ и целевую переменную T при условии X: $p_j(T, \theta \mid X)$. Для иллюстративности будем считать T и θ одномерными (см. Рис. 3).

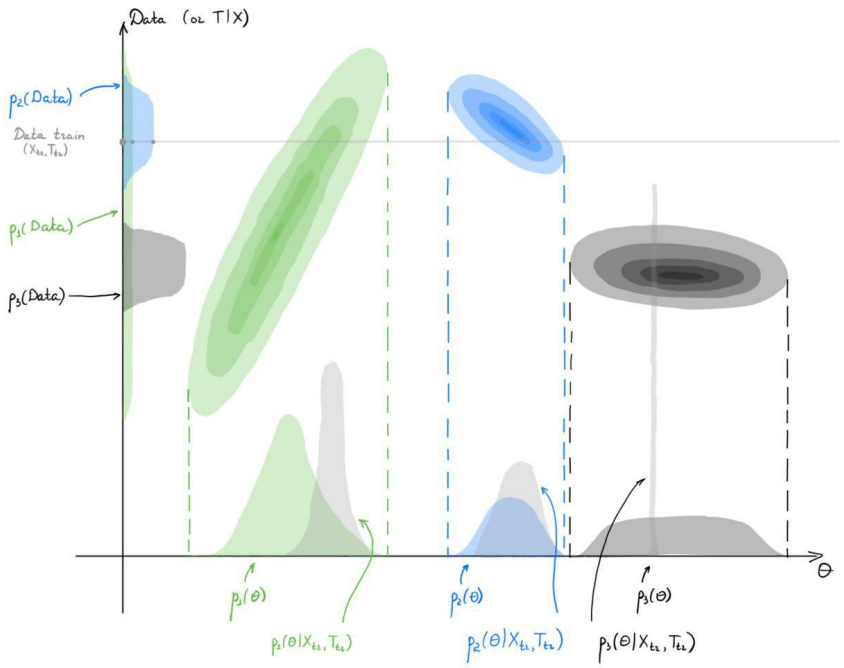


Рис. 3: Совместное распределение $p_j(T, \theta \mid X)$ для трех моделей. По горизонтальной оси отложен параметр θ, по вертикальной — значение целевой переменной T при условии X (для иллюстративности θ и T одномерные). Эллипсы отображают совместное распределение $p_j(T, \theta \mid X)$. Цветные распределения на осях отображают проекции совместного распределения на эти оси. Светло-серые распределения на оси θ показывают апостериорные распределения на параметры моделей после наблюдения данных $\{X_{tr}, T_{tr}\}$

Спроецируем совместное распределение $p_j(T, \theta \mid X)$ на ось θ. Для этого его нужно маргинализовать по T:

$$\int p_j(T, \theta \mid X) dT = \int p_j(T \mid \theta, X) p_j(\theta) dT = p_j(\theta) \tag{53}$$

Таким образом $p(\theta)$ — это проекция совместного распределения на ось θ. Аналогично $p(T \mid X)$ — это проекция совместного распределения на ось $T \mid X$.

Пусть мы пронаблюдали данные $\{X_{tr}, T_{tr}\}$, на картинке их можно изобразить горизонтальной прямой. Теперь в рамках каждой из моделей сделаем байесовский вывод — найдем апостериорное распределение на параметры модели $p(\theta \mid X_{tr}, T_{tr})$. На картинке апостериорному распределению будет соответствовать сечение совместного распределения $p(T, \theta \mid X)$ прямой $T_{tr} \mid X_{tr}$. На рисунке 3 горб $p_2(\theta \mid X_{tr}, T_{tr})$ ниже чем горб $p_1(\theta \mid X_{tr}, T_{tr})$ так как сечение второй совместной плотности шире, а площадь под горбом должна равняться единице (как интеграл от плотности распределения). Плотность распределения $p_3(\theta \mid X_{tr}, T_{tr})$ практически схлопывается в дельта-функцию в точке, где прямая $T_{tr} \mid X_{tr}$ касается линий

уровня совместной плотности распределения $p_3(T, \theta \mid X)$ (считаем, что рассматриваемые совместные распределения определены на всей плоскости, а на рисунке эллипсами показаны только области высокой вероятности). Это происходит из-за того, что в этой точке значение совместной плотности, хотя и очень маленькое, но все же гораздо больше, чем во всех остальных точках, которые пересекает прямая $T_{tr} \mid X_{tr}$[12].

Какая из трех моделей лучше всего описывает наблюдаемые данные? Заметим, что третья модель имеет самый высокий пик апостериорного распределения, однако очень плохо описывает данные. Поэтому по значению пика никаких выводов о качестве модели делать нельзя. А вот первая и вторая модели хорошо объяснят данные, поскольку содержат такие значения θ при которых правдоподобие данных $p(T_{tr} \mid X_{tr}, \theta)$ достаточно высокое. Какая же из этих моделей лучше? Чтобы ответить на этот вопрос рассмотрим небольшой пример.

Пример. Пусть есть 3 кубика со следующими конфигурациями чисел на гранях:

1. 1 2 3 4 5 6
2. 1 2 3 1 2 3
3. 1 2 1 2 1 2

Пусть в эксперименте был наугад подброшен один из кубиков и выпала тройка. Какой из кубиков скорее всего был подброшен? Это точно был не третий кубик, т.к. на его гранях нет тройки, т.е. он не описывает наблюдаемые данные. Первые два кубика описывают наблюдаемые данные, но второй делает это лучше, потому что в рамках этой модели у тройки больше шансов выпасть благодаря тому, что второй кубик может объяснить меньшую совокупность фактов. Действительно второй кубик может объяснить выпадение 1, 2, 3, а выпадение 4, 5, 6 не может, поэтому выпадение тройки при подбрасывании этого кубика оказывается более вероятно, чем выпадение тройки при подбрасывании первого кубика.

Это простой пример в точности отражает принцип наибольшей обоснованности. Посмотрим на проекции совместных распределении на вертикальную ось на рисунке 3. Эти проекции есть $p_i(T \mid X)$, т.е. это обоснованности моделей. Точки, в которых прямая $T_{tr} \mid X_{tr}$ пересекает кривые $p_i(T \mid X)$ равны обоснованностям наблюдаемых данных в рамках рассматриваемых моделей. Больше всего обоснованность данных у второй модели, поскольку ее плотность $p_2(T \mid X)$ выше всех в точке $T_{tr} \mid X_{tr}$. Первая модель тоже объясняет $T_{tr} \mid X_{tr}$, однако она может объяснить и много других значений $T \mid X$, поэтому ее плотность $p_1(T \mid X)$ «размазана» по вертикальной оси и имеет низкое значение в точке $T_{tr} \mid X_{tr}$. То есть чем большую совокупность фактов может объяснить модель, тем меньше у нее обоснованность для конкретных значений $T_{tr} \mid X_{tr}$.

Таким образом принцип наибольшей обоснованности формализует идею бритвы Оккама: «из нескольких возможных объяснений явления выбирай самое простое», где «простое» имеет смысл «то, которое может объяснить меньшую совокупность фактов». Также принцип наибольшей обоснованности находится в согласии с критерием Поппера, т.к. чем меньшую совокупность фактов может объяснить модель, тем больше возможностей ее опровергнуть, пронаблюдав то, что она не может объяснить.

Рассмотрим еще один пример для закрепления изученного принципа.

Пример. Пусть в некоторой стране N за убийство человека присуждается смертная казнь. Кроме того, в N проживают люди двух рас: синей и зеленой. Наша задача понять, используя данные о казнях, есть ли зависимость между расой убийцы, расой жертвы и вердиктом судей. Имеются следующие переменные:

1. m — раса убийцы. 0 — синий, 1 — зеленый.
2. v — раса жертвы. 0 — синий, 1 — зеленый.

[12]Конкретный вид апостериорного распределения зависит от хвостов совместного распределения $p_3(T, \theta \mid X)$. В частности, если совместное распределение имеет квадратичные хвосты в логарифмической шкале (как, например, нормальное распределение), то апостериорное распределение будет становиться все «уже и уже» при удалении от областей высокой плотности совместного распределения, постепенно, коллапсируя в дельта-функцию.

3. d — приговор. 0 — тюрьма, 1 — казнь.

Статистика по казням:

	$m=0$ $d=0$	$m=0$ $d=1$	$m=1$ $d=0$	$m=1$ $d=1$
$v=0$	132	19	52	11
$v=1$	9	0	97	6

Рассмотрим несколько вероятностных моделей, которые могли бы описывать наблюдаемые данные.

1. Приговор не зависит ни от расы убийцы, ни от расы жертвы: $p(d \mid v, m) = p(d) = \theta$
2. Приговор зависит от расы жертвы: $p(d \mid v, m) = p(d \mid v)$. $\quad p(d = 1 \mid v = 0) = \alpha,\ p(d = 1 \mid v = 1) = \beta$.
3. Приговор зависит от расы убийцы: $p(d \mid v, m) = p(d \mid m)$. $\quad p(d = 1 \mid m = 0) = \gamma,\ p(d = 1 \mid m = 1) = \delta$.
4. Приговор зависит и от расы убийцы, и от расы жертвы:

$p(d \mid v, m)$	$m=0$	$m=1$
$v=0$	τ	χ
$v=1$	ν	ξ

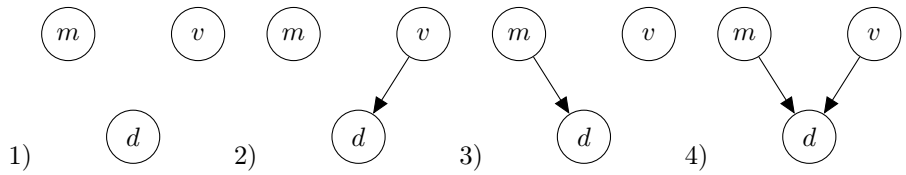


Рис. 4: Предлагаемые модели зависимости приговора d от расы убийцы m и расы жертвы v

Чтобы полностью задать байесовскую модель, необходимо ввести априорные распределения на параметры моделей ($\theta,\ \alpha,\ \beta,\ \gamma,\ \delta,\ \tau,\ \nu,\ \chi,\ \xi$). У нас нет никаких априорных предположений на вероятность казни в каждом случае, поэтому пусть априорное распределение на каждый параметр будет равномерным от нуля до единицы. Теперь посчитаем обоснованность каждой модели. Заметим, что если вероятность смертного приговора q, то вероятность пронаблюдать k смертных приговоров из N уголовных дел описывается распределением Бернулли:

$$p(x = k) = C_N^k q^k (1-q)^{N-k}$$

Тогда обоснованность первой модели:

$$p_1(Data) = \int_0^1 C_{151}^{19}\theta^{19}(1-\theta)^{132} \cdot C_9^0\theta^0(1-\theta)^9 \cdot C_{63}^{11}\theta^{11}(1-\theta)^{52} \cdot C_{103}^6\theta^6(1-\theta)^{97}d\theta =$$
$$= C \cdot C \cdot C \cdot C \cdot B(36, 292) \approx C \cdot C \cdot C \cdot C \cdot 2.8 \cdot 10^{-51}$$

где $B(.,\ .)$ — это бета-функция. Несмотря на то, что в рамках первой модели вероятность казни не зависит от расы, мы не можем сложить числа казней в разных случаях и смотреть на данные как на одну серию испытаний Бернулли. Это было бы ошибкой, поскольку мы знаем, что данные пришли из различных серий (даже если мы предполагаем, что вероятность казни в этих сериях одинакова) и эту информацию также необходимо учитывать.

Аналогично посчитаем обоснованности для остальных моделей:

$$p_2(Data) = \int_0^1 \int_0^1 C_{151}^{19}\alpha^{19}(1-\alpha)^{132} \cdot C_9^0\beta^0(1-\beta)^9 \cdot C_{63}^{11}\alpha^{11}(1-\alpha)^{52} \cdot C_{103}^6\beta^6(1-\beta)^{97}d\alpha d\beta =$$
$$= C \cdot C \cdot C \cdot C \cdot \ldots \approx C \cdot C \cdot C \cdot C \cdot 4.7 \cdot 10^{-51}$$

$$p_3(Data) = \int_0^1 \int_0^1 C_{151}^{19}\gamma^{19}(1-\gamma)^{132} \cdot C_9^0\gamma^0(1-\gamma)^9 \cdot C_{63}^{11}\delta^{11}(1-\delta)^{52} \cdot C_{103}^6\delta^6(1-\delta)^{97}d\gamma d\delta =$$
$$= C \cdot C \cdot C \cdot C \cdot \ldots \approx C \cdot C \cdot C \cdot C \cdot 0.27 \cdot 10^{-51}$$

$$p_4(Data) = \int_0^1 \int_0^1 \int_0^1 \int_0^1 C_{151}^{19}\tau^{19}(1-\tau)^{132} \cdot C_9^0\nu^0(1-\nu)^9 \cdot C_{63}^{11}\chi^{11}(1-\chi)^{52} \cdot C_{103}^6\xi^6(1-\xi)^{97}d\tau d\chi d\nu d\xi =$$
$$= C \cdot C \cdot C \cdot C \cdot \ldots \approx C \cdot C \cdot C \cdot C \cdot 0.18 \cdot 10^{-51}$$

Четвертая модель может идеально подстроиться под каждую из четырех серий испытаний (выставив параметры в частоты казней в каждой серии), поэтому она имеет низкую обоснованность (слишком много всего может хорошо объяснить). Первая модель — самая простая и у нее неплохая обоснованность. Но наблюдаемые данные показывают, что все-таки модели с одним параметром недостаточно и нужно брать вторую модель.

4 Лекция 4. Метод релевантных векторов для задачи регрессии. Автоматическое определение значимости.

Погорим о том, как можно использовать метод наибольшей обоснованности для автоматического выбора модели при решении задач машинного обучения. В данной лекции мы сделаем это на примере линейной регрессии. Примечательно, что сформулировав классическую модель на байесовском языке, можно сделать несколько элегантных обобщений, которые придадут старой, хорошо известной модели некоторые новые удивительные свойства. Но для начала вспомним несколько важных понятий, которые потребуются нам в данной лекции.

4.1 Матричное дифференцирование

Пусть $f(A)$ — скалярная функция от матрицы $A \in \mathbb{R}^{n\times n}$, то есть $f : \mathbb{R}^{n\times n} \mapsto \mathbb{R}$. Как записать её градиент? Градиент такой функции записывается, как матрица из частных производных:

$$\frac{\partial f(A)}{\partial A} = \left(\frac{\partial f(A)}{\partial a_{ij}}\right)_{\substack{i=1,\dots,n\\ j=1,\dots,n}}$$

Выпишем некоторые известные градиенты:

1. $\dfrac{\partial A(x)}{\partial x} = \left(\dfrac{\partial a_{ij}(x)}{\partial x}\right)_{\substack{i=1,\dots,n\\ j=1,\dots,n}}$, где $A : \mathbb{R} \mapsto \mathbb{R}^{n\times n}$ - матричная функция;
2. $\dfrac{\partial \det A}{\partial A} = \det A \cdot (A^{-1})^T$;
3. $\dfrac{\partial \log|\det A|}{\partial A} = \dfrac{1}{|\det A|}\dfrac{\partial |\det A|}{\partial A} = \dfrac{1}{|\det A|}|\det A| \cdot (A^{-1})^T = (A^{-1})^T$;
4. $\dfrac{\partial x^T A y}{\partial x} = \dfrac{\partial}{\partial x}\left(\sum\limits_{ij} x_i a_{ij} y_j\right) = Ay, \quad x, y \in \mathbb{R}^n$;
5. $\dfrac{\partial x^T A y}{\partial y} = \dfrac{\partial}{\partial y}\left(\sum\limits_{ij} x_i a_{ij} y_j\right) = A^T x, \quad x, y \in \mathbb{R}^n$;
6. $\dfrac{\partial x^T A x}{\partial x} = \dfrac{\partial}{\partial x}\left(\sum\limits_{ij} x_i a_{ij} x_j\right) = \left(A^T + A\right)x, \quad x, y \in \mathbb{R}^n$.

4.2 Решение системы линейных алгебраических уравнений

Рассмотрим СЛАУ:

$$Ax = b,\ x \in \mathbb{R}^n,\ b \in \mathbb{R}^m,\ A \in \mathbb{R}^{m\times n},\ \mathrm{rk}A = \min(m, n)$$

Как найти x? Напомним, что в зависимости от соотношения между m и n возможны несколько случаев:

1. $m = n,\ x = A^{-1}b$ — единственное решение (A^{-1} существует, так как $\mathrm{rk}A = n$).
2. $m > n$. Система решений не имеет. Тогда найдем точку x^*, которая минимизирует ошибку $\|Ax^* - b\|_2^2$. Почему берем именно $\|\cdot\|_2^2$? Если приравнять градиент функции потерь $\|Ax^* - b\|_2^2$ по x^* к нулю, то получим следующую систему

$$\underbrace{A^T A}_{n\times n} x^* = A^T b,$$

которая легко решается, так как A^TA обратима (т.к. $\mathrm{rk}(A^TA) = \mathrm{rk}(A) = n$). Таким образом, получаем

$$x^* = \left(A^TA\right)^{-1} A^Tb$$

Матрица $\left(A^TA\right)^{-1} A^T$ называется псевдообратной матрицей, а x^* — псевдорешением

3. $n > m$. Решений бесконечно много. В математике используется понятие нормального решения — решения минимальной нормы. Чтобы найти такое решение, рассмотрим выражение

$$x = \left(A^TA + \lambda I\right)^{-1} A^Tb$$

Матрица $A^TA+\lambda I$ невырождена при любом $\lambda > 0$, так как собственные числа матрицы A^TA больше или равны нулю, и при добавлении λ все собственные числа матрицы будут строго больше нуля. Тогда рассмотрим следующий предел

$$x^* = \lim_{\lambda \to 0} \left(A^TA + \lambda I\right)^{-1} A^Tb.$$

Можно строго доказать, что данный предел существует и что x^* будет нормальным решением.

4.3 Вероятностная постановка задачи регрессии. Метод релевантных векторов.

Опишем вероятностную постановку задачи регрессии. Пусть $x \in \mathbb{R}^m$ - объект обучающей выборки, $t \in \mathbb{R}$ - целевая переменная[13], $w \in \mathbb{R}^m$ - веса линейной регрессии. Пусть имеется также $(X, T) = (x_i, t_i)_{i=1}^n$ — обучающая выборка. Введём дискриминативную вероятностную модель

$$p(t, w \,|\, x) = p(t \,|\, w, x)p(w) \tag{54}$$

где $p(t \,|\, w, x)$ - функция правдоподобия, $p(w)$ - априорное распределение на веса. Правдоподобие $p(t \,|\, w, x)$ зададим нормальным распределением по t с математическим ожиданием, равным линейной комбинации признаков w^Tx, и некоторой дисперсией β^{-1}. Такой выбор функции правдоподобия объясняется тем, что при подстановке в нее обучающей выборки и настройке w методом максимального правдоподобия мы получим в точности минимизацию суммы квадратов отклонений t от своих прогнозных значений, т.е. линейную регрессию:

$$\arg\max_w \mathcal{N}\left(\, t \,|\, w^Tx, \beta^{-1}\right) = w_{ML} = \arg\min_w \left\|w^Tx - t\right\|_2^2. \tag{55}$$

Априорное распределение $p(w)$ зададим как нормальное по w с нулевым математическим ожиданием и дисперсией $\alpha^{-1}I$. Смысл очень простой: такое априорное распределение приводит к L_2 регуляризации, штрафуя w за отклонение от нуля. Итоговая вероятностная модель:

$$p(t, w \,|\, x) = p(t \,|\, x, w)p(w) = \mathcal{N}\left(t \,|\, w^Tx, \beta^{-1}\right)\mathcal{N}\left(w \,|\, 0, \alpha^{-1}I\right) \tag{56}$$

Получили вероятностную модель для L_2-регуляризованной линейной регрессии. Посмотрим теперь к чему это приведёт, и даст ли это какие-либо новые свойства.

Как мы будем обучать такую модель? Нужно получить апостериорное распределение на w при условии того, что мы пронаблюдали обучающую выборку: $p(w \,|\, X, T)$. Апостериорное распределение можно получить в явном виде, так как правдоподобие и априорное распределение оказываются сопряженными, поэтому апостериорное распределение лежит в том же параметрическом семействе, что и априорное, то есть является нормальным:

$$p(w \,|\, X, T) = \mathcal{N}(w \,|\, \mu, \Sigma) \tag{57}$$

[13] В общем случае t может быть многомерной, но для простоты выкладок без ограничения общности мы рассмотрим задачу регрессии с одномерной целевой переменной

Чтобы найти μ и Σ воспользуемся формулой Байеса:

$$p(w \mid X, T) = \mathcal{N}(w \mid \mu, \Sigma) = \frac{p(T \mid X, w)\, p(w)}{\int p(T \mid X, w)\, p(w) dw} \tag{58}$$

Знаменатель нам сейчас не очень важен, так как мы знаем какое распределение получится в итоге и, вычислив параметры μ и Σ, легко найдем нормировочную константу. Распишем числитель выражения (58):

$$\begin{aligned} p(T \mid X, w)\, p(w) &= \frac{\beta^{\frac{n}{2}}}{(2\pi)^{\frac{n}{2}}} \exp\left\{\left(-\frac{\beta}{2}\|T - Xw\|^2\right)\right\} \frac{\alpha^{\frac{m}{2}}}{(2\pi)^{\frac{m}{2}}} \exp\left\{\left(-\frac{\alpha}{2} w^T w\right)\right\} = \\ &= \frac{\beta^{\frac{n}{2}} \alpha^{\frac{m}{2}}}{(2\pi)^{\frac{m+n}{2}}} \exp\left\{\left(-\frac{\beta}{2}\left(T^T T - 2w^T X^T T + w^T X^T X w\right) - \frac{\alpha}{2} w^T w\right)\right\} \\ &= \frac{\beta^{\frac{n}{2}} \alpha^{\frac{m}{2}}}{(2\pi)^{\frac{m+n}{2}}} \exp\left\{\left(-\frac{1}{2} w^T \underbrace{\left(\beta X^T X + \alpha I\right)}_{\Sigma^{-1}} w + \beta w^T X^T T - \frac{\beta}{2} T^T T\right)\right\} \end{aligned} \tag{59}$$

Коэффициент при $w^T w$ соответствует обратной ковариационной матрице. Отсюда получаем:

$$\Sigma = \left(\beta X^T X + \alpha I\right)^{-1} \tag{60}$$

Чтобы найти μ можно выделить полный квадрат под экспонентой и провести громоздкие вычисления. Но мы поступим проще и воспользуемся тем, что математическое ожидание нормального распределения совпадает с его модой: $\mu = w_{MP}$. Т.е. надо найти w, максимизирующий (59), то есть максимум следующего выражения:

$$-\frac{\beta}{2}\left(T^T T - 2w^T X^T T + w^T X^T X w\right) - \frac{\alpha}{2} w^T w \tag{61}$$

Найдём производную (61) по w и приравняем её к нулю:

$$\begin{aligned} &\frac{\partial}{\partial w}\left(-\frac{\beta}{2}\left(T^T T - 2w^T X^T T + w^T X^T X w\right) - \frac{\alpha}{2} w^T w\right) = \\ &= \beta X^T T - \beta X^T X w - \alpha w = 0, \\ &\beta X^T T = \left(\beta X^T X + \alpha I\right) w \end{aligned}$$

Отсюда получаем формулу

$$w_{MP} = \beta \underbrace{\left(\beta X^T X + \alpha I\right)^{-1}}_{\Sigma} X^T T \tag{62}$$

Итого, мы получили значения параметров для апостериорного распределения:

$$\Sigma = \left(\beta X^T X + \alpha I\right)^{-1}, \tag{63}$$

$$\mu = w_{MP} = \beta \Sigma X^T T \tag{64}$$

Теперь мы можем сделать предсказание в рамках байесовской линейной регрессии, т.е. найти распределение на значение целевой переменной для нового объекта x_*:[14]

$$p(t_* \mid x_*, X, T) = \int p(t_* \mid x_*, w) p(w \mid X, T) dw \tag{65}$$

Интеграл в формуле (65) всегда имеет такую же сложность как и интеграл в знаменателе формулы Байеса на обучении, то есть либо оба берутся, либо оба не берутся. В нашем

[14] Заметим что в обычной линейной регрессии мы ограничены только нахождением w_{MP} и поэтому можем посчитать только точечную оценку на t_*

случае распределения сопряжены, поэтому можем брать оба интеграла. В результате интегрирования получаем нормальное распределение[15]:

$$p(t_* \mid x_*, X, T) = \int p(t_* \mid x_*, w) p(w \mid X, T) dw = \mathcal{N}\left(t_* \mid x_*^T w_{MP}, \ldots\right) \tag{66}$$

Теперь посмотрим, как наш алгоритм прогнозирования зависит от гиперпараметров α и β. Заметим, что β регулирует величину штрафа за квадрат отклонений, а α — величину L_2 регуляризации, накладываемой на веса w (см. выражение, которое мы максимизировали, когда искали w_{MP} 61).

Теперь зафиксируем β и посмотрим как меняется алгоритм в зависимости от α. Для этого рассмотрим два предельных случая:

1. $\alpha \to 0$

   $$\lim_{\alpha \to 0} w_{MP} = w_{ML}$$

   Так как $w_{MP} = \arg\max_w p\left(T \mid X, w\right) p(w)$ и $p(w)$ становится неинформативным при $\alpha \to 0$, максимум $p\left(T \mid X, w\right) p(w)$ достигается в точке максимального правдоподобия w_{ML}.

2. $\alpha \to \infty$

   $$\lim_{\alpha \to \infty} w_{MP} = 0$$

   Почему так происходит? Первое объяснение: $p(w)$ становится δ-функцией в 0, поэтому апостериорное распределение «схлопывается» туда же. Второе объяснение: α — коэффициент регуляризации, при $\alpha \to \infty$ накладывается слишком большой штраф за отклонение от 0. Третье объяснение: в Σ^{-1} возникает диагональ с бесконечно большими значениями, $\Sigma \to 0$.

В первом случае мы не накладываем никакой регуляризации на веса модели и позволяем ей максимально подстроиться под обучающую выборку, что может привести к переобучению. Во втором случае, наоборот, мы ограничиваем веса максимально строгой регуляризацией, не давая модели ничего выучить про данные. Оптимальное значения параметра α находится где-то посередине между этими предельными случаями, и чтобы его подобрать можно воспользоваться классическим методом кросс-валидации.[16]

Как мы заметили выше, α регулирует способность весов адаптироваться под данные. А что если у нас много признаков и мы подозреваем, что некоторые из них совсем не влияют на значение целевой переменной? Нам бы хотелось, чтобы веса «важных» признаков подстраивались под данные, а веса «неважных» признаков этого не делали, потому что последние могут подстроиться только под шум в данных, что непременно приведет к переобучению. Однако, варьируя α мы не можем этого добиться, потому что она одинаково влияет на все веса.

Попробуем усложнить нашу вероятностную модель так, чтобы веса разных признаков регуляризировались по-разному. Мы можем изменить модель так, чтобы каждому w_i соответствовал свой собственный коэффициент α_i:

$$p(t, w \mid x) = p(t \mid x, w) p(w) = \mathcal{N}\left(t \mid w^T x, \beta^{-1}\right) \mathcal{N}\left(w \mid 0, A^{-1}\right), \quad A = \begin{pmatrix} \alpha_1 & & 0 \\ & \ddots & \\ 0 & & \alpha_m \end{pmatrix} \tag{67}$$

Теперь наша вероятностная модель индексируется параметром β и параметрами $\alpha_1, \ldots, \alpha_m$ на диагонали A, т.е. гиперпараметров стало довольно много и подстраивать их по кросс-валидации уже не очень удобно. Попробуем применить метод наибольшей обоснованности и выбрать наиболее обоснованную модель, учитывая что распределения сопряжены и подсчёт обоснованности должен быть несложным. Заметим, что множество, из которого мы

[15] Точный вид матрицы ковариации предлагаем читателю вывести самостоятельно

[16] Можно ли настраивать α и β с помощью байесовских методов? Теоретически да, но для двух настраиваемых параметров это не очень оправдано и гораздо проще воспользоваться кросс-валидацией.

выбираем модели, не конечно, то есть необходимо посчитать обоснованность от A и β так, чтобы по A и β можно было бы вести оптимизацию.

Рассчитаем обоснованность:

$$p\left(T \mid X, A, \beta\right) = \int p\left(T \mid X, w, \beta\right) p\left(w \mid A\right) dw \tag{68}$$

Интеграл 68 берётся (знаменатель формулы Байеса), но мы можем упростить вычисления. Обозначим $p\left(T \mid X, w, \beta\right) p\left(w, A\right)$ как $Q(w)$. Посмотрим что представляет собой эта функция как функция от w:

$$\begin{aligned} Q(w) &= \frac{\beta^{\frac{n}{2}}}{(2\pi)^{\frac{n}{2}}} \exp\left\{\left(-\frac{\beta}{2}\|T - Xw\|^2\right)\right\} \frac{\sqrt{\det A}}{(2\pi)^{\frac{m}{2}}} \exp\left\{\left(-\frac{1}{2} w^T A w\right)\right\} = \\ &= \frac{\beta^{\frac{n}{2}} \sqrt{\det A}}{(2\pi)^{\frac{m+n}{2}}} \exp\left\{\left(-\frac{1}{2} w^T \left(\beta X^T X + A\right) w + \beta w^T X^T T - \frac{\beta}{2} T^T T\right)\right\} \end{aligned} \tag{69}$$

Обратим внимание на выражение под экспонентой является перевернутой многомерной параболой. Мы знаем точку максимума w_{MP} этой параболы и матрицу при квадратичной форме Σ^{-1}. Разложим функцию под экспонентой в ряд Тейлора в точке w_{MP} до второго порядка. Нулевой член будет присутствовать, первого члена не будет, так как w_{MP} — точка максимума. Тогда (69) расписывается как

$$\begin{aligned} &\frac{\beta^{\frac{n}{2}} \sqrt{\det A}}{(2\pi)^{\frac{m+n}{2}}} \exp\left\{\left(-\frac{1}{2}(w - w_{MP})^T \left(\beta X^T X + A\right)(w - w_{MP})\right)\right\} \cdot \\ &\qquad \cdot \exp\left\{\left(\frac{1}{2} w_{MP}^T \Sigma^{-1} w_{MP} + \beta w_{MP}^T X^T T - \frac{\beta}{2} T^T T\right)\right\} = \\ &= Q(w_{MP}) \exp\left\{\left(-\frac{1}{2}(w - w_{MP})^T \left(\beta X^T X + A\right)(w - w_{MP})\right)\right\} \end{aligned} \tag{70}$$

Вернемся к интегралу (68):

$$\begin{aligned} p\left(T \mid X, A, \beta\right) &= \int p\left(T \mid X, w, \beta\right) p\left(w, A\right) dw = \\ &= \int Q(w_{MP}) \exp\left\{\left(-\frac{1}{2}(w - w_{MP})^T \Sigma^{-1} (w - w_{MP})\right)\right\} dw = \\ &= Q(w_{MP}) \int \exp\left\{\left(-\frac{1}{2}(w - w_{MP})^T \Sigma^{-1} (w - w_{MP})\right)\right\} dw = \\ &= Q(w_{MP}) (2\pi)^{\frac{m}{2}} \sqrt{\det \Sigma} \to \max_{A,\beta} \end{aligned} \tag{71}$$

Теперь применим любопытный приём. Рассмотрим логарифм обоснованности (71):

$$\begin{aligned} \log p\left(T \mid X, A, \beta\right) = \frac{n}{2} \log \beta - \frac{n}{2} \log 2\pi + \frac{1}{2} \log \det A - \frac{\beta}{2} \|T - X w_{MP}\|^2 - \\ - \frac{1}{2} w_{MP}^T A w_{MP} - \frac{1}{2} \log \det \Sigma^{-1} \to \max_{A,\beta} \end{aligned} \tag{72}$$

Насколько сложно промаксимизировать полученное выражение по A и β? β входит под логарифмом, линейно и в Σ^{-1}, A входит в $\log \det A$, линейно и в Σ^{-1}. Но кроме того, w_{MP} зависит от A и β и зависимость эта не очень приятная: нужно обращать матрицу (см. выражение (63)[17]). Вспомним красивый прием из вычислительной математики, которые поможет нам промаксимизировать обоснованность без громоздких вычислений.

Определение 5. Пусть $f(x)$ — некоторая функция действительного переменного. Тогда семейство функций двух переменных $g(x, \xi)$, обладающее свойствами

[17] При переходе от вероятностной модели с ковариационной матрицей априорного распределения $\alpha^{-1} I$ к модели, где эта матрица равна A^{-1}, выражения для параметров апостериорного распределения сохранятся с точностью до замены αI на A

1. $\forall x,\ \forall \xi \quad f(x) \geq g(x, \xi)$
2. $\forall x\ \exists \xi(x)\ :\ f(x) = g(x, \xi(x))$,

называется вариационной нижней оценкой функции f.

Вариационная нижняя оценка является нижней оценкой, и при этом в любой точке x можем так подобрать параметр ξ так, что оценка становится точной. Простейшим примером вариационной нижней оценки служит касательная к выпуклой функции.

Если $g(z, \xi)$ — вариационная нижняя оценка для $f(x)$, то мы можем решить задачу максимизации функции $f(x)$ по x с помощью следующей итеративной процедуры:

$$\begin{cases} x_n = \arg\max\limits_x g(x, \xi_{n-1}), \\ \xi_n = \arg\max\limits_\xi g(x_n, \xi) \end{cases} \tag{73}$$

Можно показать, что такая итеративная процедура сходится в стационарную точку функции $f(x)$. Такая замена оптимизируемой функции может быть удобна, если максимум исходной функции $f(x)$ искать тяжело, а максимизировать вариационную нижнюю оценку $g(x, \xi)$ — просто. Мы еще не раз встретимся с подобными случаями в последующих лекциях.

Возвращаясь к нашей задаче, функционал (72) можно рассмотреть как

$$\begin{aligned} \log p\left(T \mid X, A, \beta\right) = \frac{n}{2}\log\beta - \frac{n}{2}\log 2\pi + \frac{1}{2}\log\det A - \frac{\beta}{2}\|T - Xw_{MP}\|^2 - \\ - \frac{1}{2}w_{MP}^T A w_{MP} - \frac{1}{2}\log\det\Sigma^{-1} \geq \qquad (74) \\ \geq \frac{n}{2}\log\beta - \frac{n}{2}\log 2\pi + \frac{1}{2}\log\det A - \frac{\beta}{2}\|T - Xw\|^2 - \\ - \frac{1}{2}w^T A w - \frac{1}{2}\log\det\Sigma^{-1} \qquad (75) \end{aligned}$$

Оценка (74) верна, поскольку $Q(w_{MP}) \geq Q(w)$ т.к. w_{MP} — точка максимума $Q(w)$. Полученная оценка является вариационной нижней оценкой, потому что для любых A и β существует $w = w_{MP}$, при котором достигается равенство.

Теперь задача оптимизации выглядит как

$$\frac{n}{2}\log\beta - \frac{\beta}{2}\|T - Xw\|^2 + \frac{1}{2}\log\det A - \frac{1}{2}w^T A w - \frac{1}{2}\log\det\Sigma^{-1} \to \max_{A,\beta,w} \tag{76}$$

где мы отбросили константы, не влияющие на оптимизацию. Точку максимума по w мы знаем — это w_{MP}, осталось найти максимум по A, β. Дифференцируем выражение (76) по α_j при $w = w_{MP}$ (считаем, что w_{MP} не зависит от A) и приравниваем к нулю:

$$\begin{aligned} \frac{\partial}{\partial\alpha_j}\left(\frac{n}{2}\log\beta - \frac{\beta}{2}\|T - Xw_{MP}\|^2 + \frac{1}{2}\log\det A - \frac{1}{2}w_{MP}^T A w_{MP} - \frac{1}{2}\log\det\Sigma^{-1}\right) = \\ = 0 - 0 + \frac{1}{2}\frac{\partial}{\partial\alpha_j}\log\det A - \frac{1}{2}\frac{\partial}{\partial\alpha_j}(w_{MP}^T A w_{MP}) - \frac{1}{2}\frac{\partial}{\partial\alpha_j}\log\det\Sigma^{-1} = \\ = \left\{ \frac{\partial}{\partial\alpha_j}\log\det A = \frac{\partial}{\partial\alpha_j}\sum_{i=1}^m \log\alpha_i = \frac{1}{\alpha_j};\ \frac{\partial}{\partial\alpha_j}(w_{MP}^T A w_{MP}) = (w_{MP})_j^2; \right. \\ \left. \frac{\partial}{\partial\alpha_j}\log\det\Sigma^{-1} = \operatorname{tr}\left(\frac{\partial\log\det\Sigma^{-1}}{\partial\Sigma^{-1}}\frac{\partial\Sigma^{-1}}{\partial\alpha_j}\right) = \operatorname{tr}\left(\Sigma^T I_{jj}\right) = \Sigma_{jj} \right\} = \\ = \frac{1}{2\alpha_j} - \frac{1}{2}w_{jMP}^2 - \frac{1}{2}\Sigma_{jj} = 0 \qquad (77) \end{aligned}$$

При вычислении $\dfrac{\partial}{\partial\alpha_j}\log\det\Sigma^{-1}$ мы воспользовались тем, что $\dfrac{\partial\log\det\Sigma^{-1}}{\partial\Sigma^{-1}} = \Sigma^T$ и $\dfrac{\partial\Sigma^{-1}}{\partial\alpha_j} = \dfrac{\partial(\beta X^T X + A)}{\partial\alpha_j} = I_{jj}$. Получаем:

$$\alpha_j = \frac{1}{w_{jMP}^2 + \Sigma_{jj}} \tag{78}$$

Заметим, что в данном выражении Σ_{jj} зависит от A, β. Поскольку мы оптимизируем итеративным методом, для вычисления α_j на следующей итерации мы можем воспользоваться значениями A, β с предыдущей итерации. Эта хитрость не что иное как метод простой итерации, и он не нарушит сходимость процесса. Однако на практике, если мы будем пересчитывать A по формуле (78), то сходиться процесс будет довольно медленно. Почему?

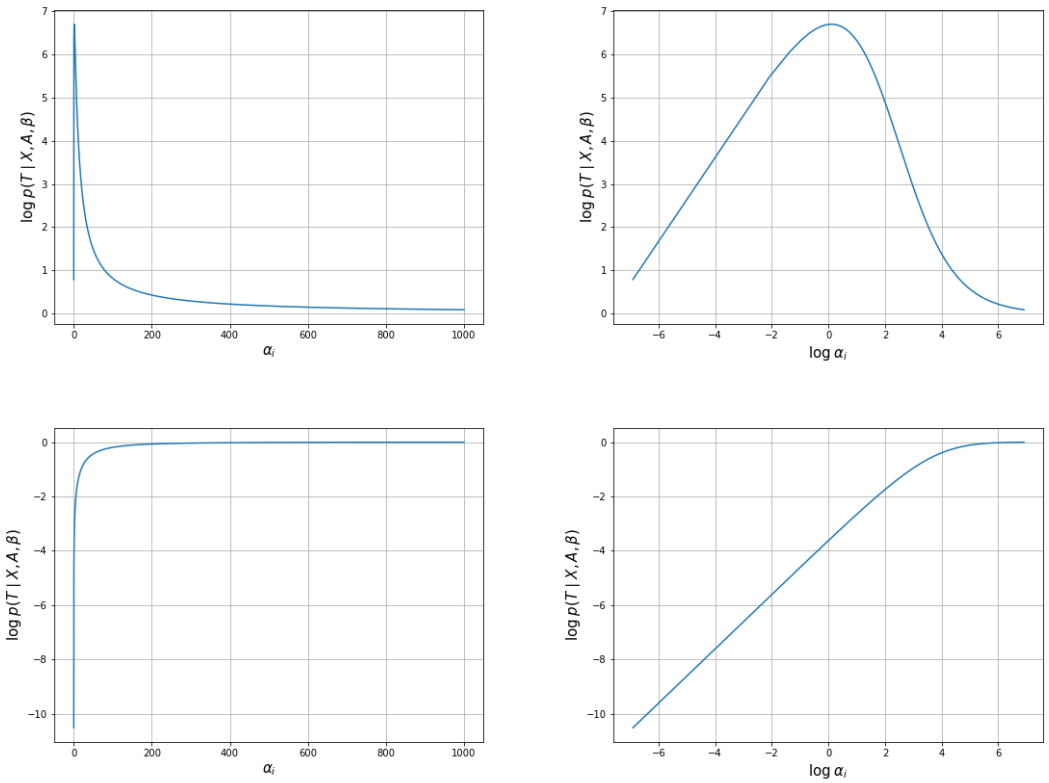


Рис. 5: Возможные виды зависимости оптимизируемой функции от α_j (левые графики) и от $\log \alpha_j$ (правые графики)

Посмотрим на график зависимости оптимизируемой функции от α_j и от от $\log \alpha_j$(Рис 5). Заметим, что функции в правой колонке проще оптимизировать итеративными методами, чем функции в левой колонке, поскольку для функций слева, если начальное приближение оказалось далеко от пика (или от бесконечности), то итеративный метод будет долго сходиться к максимуму (или в бесконечность) по пологому хвосту, т.к. значение производной на нём маленькое. А у функций справа значение производной везде достаточно большое, и поэтому итеративный метод быстро дойдет до максимума (или до достаточно больших значений, чтобы соответствующий вес можно было отбросить без потери точности прогноза) из любого начального приближения.

Но как нам перейти от оптимизации левой функции к оптимизации правой? Нужно перейти к оптимизации по $\log \alpha_j$, т.е. чтобы получить итеративную процедуру, нам нужно взять производную оптимизируемой функции по $\log \alpha_j$. Как перейти от производной по α_j к производной по $\log \alpha_j$? Фактически это эквивалентно тому, что все слагаемые домножаются на α_j. Получаем:

$$1 - \alpha_j w_{jMP}^2 - \alpha_j \Sigma_{jj} = 0 \tag{79}$$

Небольшое замечание: формула позволяет найти α_j при условии фиксированных w_{MP}, β.

В формуле (79) можем дополнительно разделить переменные:

$$1 - \alpha_j^{new} w_{jMP}^2 - \alpha_j^{old} \Sigma_{jj} = 0, \tag{80}$$

откуда получаем

$$\alpha_j^{new} = \frac{1 - \alpha_j^{old}\Sigma_{jj}^{old}}{w_{jMP}^2}. \tag{81}$$

Аналогично выводится формула для β:

$$\beta^{new} = \frac{n - \sum\limits_{j=1}^{m}\left(1 - \alpha_j^{old}\Sigma_{jj}^{old}\right)}{\|T - Xw_{MP}\|^2} \tag{82}$$

Чем хороша полученная процедура на практике? Обычно, она сходится за несколько десятков итераций, и при этом практически сразу многие α_j уходят в бесконечность, что равносильно отбрасыванию лишних признаков. Кроме того, если есть группа скоррелированных между собой признаков, то метод отбросит все признаки из этой группы, кроме одного.

Рассмотренный метод можно сделать нелинейным, перейдя к обобщённой линейной регрессии, когда вместо обычных признаков мы имеем дело с базисными функциями на объектах обучающей выборки. При этом формально количество w равно количеству объектов обучающей выборки, и получается автоматический подбор наиболее релевантных объектов (отсюда и название метода релевантных векторов).

5 Лекция 5. Метод релевантных векторов для задачи классификации

В предыдущей лекции мы рассмотрели вероятностную модель линейной регрессии, задав функции правдоподобия и априорное распределение на параметры модели. Для каждого объекта обучающей выборки x_n мы определили правдоподбие плотностью нормального распределения, где среднее соответствует стандартной модели линейной регрессии: $x_n^T w,\ x_n, w \in \mathbb{R}^d$. Априорное распределение для вектора параметров w выбрали сопряженным к правдоподобию: нормальное распределение с нулевым средним и матрицей ковариации A^{-1}. Сопряжение между функцией правдоподобия и априорным распределением, означает, что апостериорное распределение лежит в том же классе, что и априорное, но с другими параметрами. Такой выбор позволил нам вычислить обоснованность модели (знаменатель в формуле Байеса) и оптимизировать её по матрице ковариации A^{-1}. Специальный выбор пространства оптимизации: $A = \mathrm{diag}(\alpha_1, \ldots, \alpha_d)$ приводит к разряженному решению в пространстве параметров w, где признаки выбираются «автоматически». Можно ли получить аналогичный метод, но для задачи классификации?

В этой лекции мы предложим конструктивный алгоритм в качестве ответа на этот вопрос. Мы переформулируем классическую модель логистической регрессии как вероятностную. Для того чтобы выбирать признаки «автоматически», мы используем такое же априорное распределение, как и для задачи регрессии, но отличную функцию правдоподобия. Она окажется несопряженной с априорным распределением: полноценный «байес для богатых» невозможен. В частности, аналитическое выражение для обоснованности вывести не выйдет. Мы рассмотрим различные способы оценки обоснованности и предложим алгоритм её оптимизации по параметрам априорного распределения $A = \mathrm{diag}(\alpha_1, \ldots, \alpha_d)$.

5.1 Байесовская интерпретация задачи классической логистической регрессии

Мы наблюдаем набор независимых пар $\{(x_n, t_n)\}_{n=1}^N$: вектор признаков $x_n = (1, x_n^2, \ldots, x_n^d)$ и бинарную метку $t_n \in \{-1, 1\}$. Мы ввели фиктивный признак $x_n^1 = 1$, чтобы не писать отдельно свободный член в скалярном произведении $w^T x_n$, где $w \in \mathbb{R}^d$ — параметры модели. Опишем вероятностную модель, определив функции правдоподобия $p(t_n \,|\, w, x_n)$ для каждого объекта и априорное распределение $p(w)$ на параметры модели.

Функция правдоподобия должна быть вероятностным распределением относительно $t_n \in \{-1, 1\}$. Соответствующий логистической регрессии выбор — это логистическая функция:

$$p(t_n \,|\, w, x_n) = \frac{1}{1 + \exp(-t_n w^T x_n)}. \tag{83}$$

Проверим, что она является вероятностным распределением относительно $t_n \in \{-1, 1\}$:

$$p(t = -1 \,|\, x, w) + p(t = 1 \,|\, x, w) = \frac{1}{1 + e^{w^T x}} + \frac{1}{1 + e^{-w^T x}} = \frac{1 + e^{-w^T x} + e^{w^T x} + 1}{1 + e^{-w^T x} + e^{w^T x} + e^0} = 1. \tag{84}$$

В качестве априорного распределения возьмем нормальное с нулевым средним и матрицей ковариации A^{-1}:

$$p(w) = \mathcal{N}\left(w \mid 0, A^{-1}\right). \tag{85}$$

Итоговая вероятностная модель имеет вид:

$$p(\{(x_n, t_n)\}_{n=1}^N \,|\, w) = \left[\prod_{n=1}^N \frac{1}{1 + \exp(-t_n w^T x_n)}\right] \mathcal{N}\left(w \mid 0, A^{-1}\right). \tag{86}$$

Покажем, что для такой модели решение задачи w_{MP} «байеса для бедных» соответствует решению задачи оптимизации классической логистической регрессии с l_2-регуляризацией:

$$w_{MP} = \arg\max_w p(w \,|\, X, T) = \arg\max_w \log p(w \,|\, X, T) = \arg\max_w \log[p(T \,|\, w, X) p(w)]. \tag{87}$$

Продолжая (87):

$$= \arg\max_w \left(\log p(T \mid w, X) + \log p(w)\right) = \tag{88}$$

$$= \arg\max_w \left(-\sum_{n=1}^{N} \log\left(1 + \exp\left(-t_n w^T x_n\right)\right) - \frac{1}{2} w^T A w \right) = \tag{89}$$

$$= \arg\min_w \left(\sum_{n=1}^{N} \log\left(1 + \exp\left(-t_n w^T x_n\right)\right) + \frac{1}{2} w^T A w \right). \tag{90}$$

Выбирая матрицу ковариации априорного распределения $A = \alpha I$, получаем:

$$w_{MP} = \arg\min_w \left(\sum_{n=1}^{N} \log\left(1 + \exp\left(-t_n w^T x_n\right)\right) + \frac{\alpha}{2} w^T w \right). \tag{91}$$

Читателю осталось проверить, что задача (91) является классическим функционалом log-loss для логистической регрессии с l_2-регуляризацией. Данный функционал — строго выпуклая функция по w (ведь логарифм сигмоиды выпуклый, а $w^T A w$ положительно определенная квадратичная форма). Задачу поиска единственной точки оптимума можно решать с помощью метода IRLS (Iteratively Reweighted Least Squares), итеративная формула для которого имеет вид:

$$w^{(k+1)} = \underbrace{\left(X^T R(w^{(k)}) X + \alpha I\right)}_{d \times d}^{-1} X^T R(w^{(k)}) z(w^{(k)}), \tag{92}$$

где

$$X = \begin{pmatrix} 1 & x_1^2 & \dots & x_1^d \\ 1 & x_2^2 & \dots & x_2^d \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_N^2 & \dots & x_N^d \end{pmatrix} \text{ — матрица признаков} \tag{93}$$

$$R(w) = \operatorname{diag}\left(s_1(1 - s_1), \dots, s_N(1 - s_N)\right), \quad s_n = \frac{1}{1 + \exp(-t_n w^T x_n)} \tag{94}$$

$$z(w) = Xw + R^{-1}(w) \begin{pmatrix} t_1 & & 0 \\ & \ddots & \\ 0 & & t_N \end{pmatrix} \begin{pmatrix} 1 - s_1 \\ \vdots \\ 1 - s_N \end{pmatrix}. \tag{95}$$

Любопытный читатель может также проверить, что IRLS является ни чем иным, как самым обыкновенным методом Ньютона. Как правило, IRLS метод сходится за достаточно малое количество шагов для любого начального приближения $w^{(0)}$. Стоит учитывать, что в данном методе приходится обращать матрицу $d \times d$, поэтому для задач с большим числом признаков d, стоит рассмотреть метод оптимизации первого порядка, например, градиентный спуск.

Замечание. Матрица: $-\left(X^T R(w_k) X + \alpha I\right)$ — гессиан оптимизируемой функции:

$$\nabla^2 [\log p(T \mid X, w) + \log p(w)] = -\left(X^T R(w) X + \alpha I\right). \tag{96}$$

5.2 Метод релевантных векторов

Мы описали задачу логистической регрессии на «байесовском языке», введя априорное распределение на параметры модели $\mathcal{N}(w \mid 0, A^{-1})$. Затем мы продемонстрировали связь такого выбора априрорного распредления с использованием $l2$-регурелизации в задаче обучения логистической регрессии. Действуя по аналогии с предыдущей лекцией, мы можем выбрать для каждого параметра w_i свой «коэфициент регуляризации»:

$$p(w \mid A) = \mathcal{N}\left(w \mid 0, A^{-1}\right) = \prod_{i=1}^{d} \mathcal{N}\left(w_i \mid 0, \alpha_i^{-1}\right), \qquad A = \operatorname{diag}(\alpha_1, \dots, \alpha_d).$$

Забегая вперёд, скажем, что в данной лекции будет продемонстрирован конструктивный алгоритм оптимизации α_i. Но прежде давайте рассмотрим, что будет происходить, если некоторое $\alpha_i \to +\infty$. Так как i-ый вес $w_i \sim \mathcal{N}(0, \alpha_i^{-1})$, получаем

$$w_i \xrightarrow{d} 0. \tag{97}$$

Таким образом, если мы будем оптимизировать обоснованность модели по параметрам априорного распределения $\text{diag}(\alpha_1, \ldots, \alpha_d)$, то большим значениям α_i будут соответствовать близкие к нулю веса и менее релевантные признаки, а малым α_i — более релевантные. Таким образом, в процессе оптимизации мы получим автоматическое разреживание признаков, как и на предыдущей лекции.

Однако, есть несколько сложностей. В данном случае мы не можем сделать полноценный байесовский вывод в силу того, что распределения $p(t \mid w, x)$ и $p(w \mid A)$ не сопрягаются. А значит, во-первых, мы не сможем найти аналитическое выражения для обоснованности, и непонятно, как ее прооптимизировать по A. Во-вторых, мы не сможем посчитать апостериорное распределение на веса w. Вторую проблему мы можем решить по-бедному: найдем точечную оценку на веса, с помощью максимизации апостериорного распределения. Эта можно сделать с помощью того же самого IRLS, который в данном случае он будет выглядеть так:

$$w_{k+1} = \left(X^T R(w_k) X + A\right)^{-1} X^T R(w_k) z(w_k), \tag{98}$$

где X, $R(w)$ и $z(w)$ определены, соответственно, в (93), (94) и (95). Метод IRLS гарантирует, что $w_k \to w_{MP}$.

Теперь вернемся к самому интересному вопросу, как оптимизировать обоснованность по A? Чтобы решить эту проблему, предлагается пойти по пути «байеса для среднего класса», то есть использовать приближённый байесовский вывод, который носит название вариационный байесовский вывод. Отметим, что вариантов вариационного байесовского вывода существует огромное количество: метод в настоящей лекции лишь один из многих. Однако нужно же с чего-то начинать!

Замечание. Прежде чем мы перейдем к вариационному байесовскому выводу, хочется сказать, что он применим и к так называемой обобщённой логистической регрессии. Пусть у нас есть набор функций (будем называть их базисными функциями) $\{\varphi_i(x)\}_{i=1}^d$. Задача состоит в построении оптимальной линейной комбинации этих функций с весами — параметрами w. При этом, распространена ситуация, при которой число базисных функций совпадает с числом объектов. В качестве примера можно привести радиальные базисные функциии — функции вида

$$\varphi_j(x) = \exp\left(-\gamma \|x - x_j\|^2\right) \tag{99}$$

Радиальные базисные функции применяются для построения существенно нелинейных разделяющих поверхностей. По факту, обобщённая логистическая регрессия — это классическая логистическая регрессия только с преобразованной матрицей признаков. По этой причине мы не будем приводить формулы для обобщённой логистической регрессии, дабы не перегружать обозначения.

5.3 Приближенное вычисление обоснованности методом Лапласа

Мы будем оптимизировать A, решая задачу максимизации обоснованности:

$$p(T \mid X, A) = \int p(T \mid X, w) p(w \mid A)\, dw \longrightarrow \max_A. \tag{100}$$

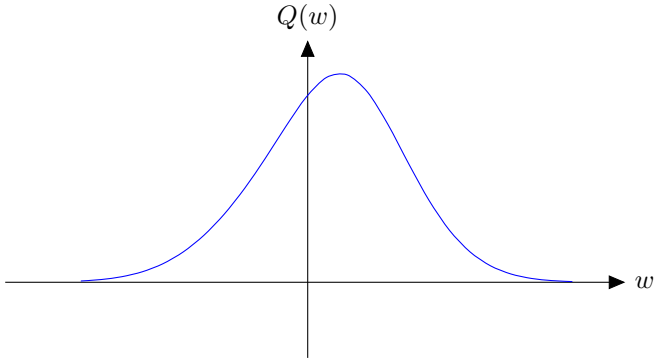
Для решения такой задачи оптимизации нужно уметь вычислять интеграл в (100), который, увы, не берется аналитически. Однако мы можем его оценить для каждого фиксированного значения параметра A! Один из способов приблизить значение интеграла — это заменить его подытгеральную функцию на удобную оценку. По этой причине введём обозначение:

$$Q(w) := p(T \mid X, w) p(w \mid A). \tag{101}$$

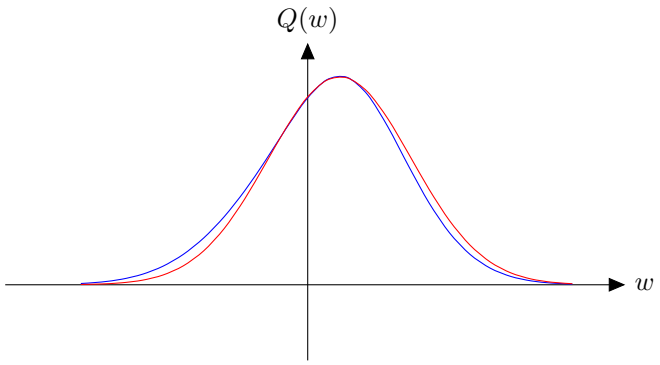
Для того чтобы предложить оценку для $Q(w)$, подумаем, что мы вообще можем сказать об этой функции. Давайте возьмём от неё логарифм:

$$\log Q(w) = -\sum_{n=1}^{N} \log\bigl(1 + \exp\bigl(-t_n w^T x_n\bigr)\bigr) - \sum_{j=1}^{d} \alpha_j w_j^2. \tag{102}$$

Данная функция — строго вогнутая (ведь логарифм сигмоиды строго вогнутый, а $\sum_{j=1}^{d} \alpha_j w_j^2$ — это парабола). Значит, максимум у данной функции единственный, а при больших по норме w функция $\log Q(w)$ будет стремиться к минус бесконечности. Поэтому, взяв от такой функции $\exp(\cdot)$ мы получим колокообразную функцию, которая выглядит примерно так:



Данный колокольчик уж очень напоминает гауссиану, а гауссианы мы успешно умеем интегрировать. Мы воспользуемся данным фактом и попробуем приблизить $Q(w)$ гауссовским колокольчиком. Метод приближения колокообразных функций гауссианами носит название метода Лапласа. Схематично, мы хотим получить такую картину:



где красный колольчик — это гауссиана. Ещё раз подчеркнём, что главной нашей задачей является подсчёт интеграла $\int Q(w)\, dw$. Основной вклад в значение интеграла вносят области носителя с наибольшими значениями подынтегральной функции (в нашем случае $Q(w)$). По этой причине найдём приближение унимодальной $\log Q(w)$ с помощью первых трёх слагаемых в разложении в ряд Тейлора в точке w_{MP} — точке максимума $\log Q(w)$:

$$\log Q(w) \approx \log Q(w_{MP}) + (w - w_{MP})^T \nabla \log Q(w_{MP}) + \\ + \frac{1}{2}(w - w_{MP})^T \underbrace{\nabla^2 \log Q(w_{MP})}_{\text{гессиан}} (w - w_{MP}). \tag{103}$$

Итак, что мы тут можем упростить? Во-первых, $\nabla \log Q(w_{MP}) = 0$, так как w_{MP} точка экстремума. Во-вторых, $\nabla^2 \log Q(w_{MP})$ можно посчитать явно:

$$\nabla^2 \log Q(w_{MP}) = -(X^T R(w_{MP}) X + A), \tag{104}$$

где X, $R(w)$ определены выше (93), (94). Вывод формулы (104) предоставляется читателю в качестве упражнения.

Обозначив $\Sigma := (X^T R(w_{MP})X + A)^{-1}$, положительно определенную из соображений выпуклости, подстановкой получаем приближенное значение обоснованности модели:

$$\int Q(w)dw \approx \int Q(w_{MP}) \exp\left(-\frac{1}{2}(w - w_{MP})^T \Sigma^{-1}(w - w_{MP})\right) dw = Q(w_{MP})(2\pi)^{d/2}\sqrt{\det \Sigma}. \tag{105}$$

Из полученного выражения видно, что мы считаем модель тем более обоснованной, чем, во-первых, шире наш (гауссовский) колокольчик (за так называемую ширину отвечает $\det \Sigma$) и, во-вторых, чем больше значение в точке Maximum Posterior, т.е. $Q(w_{MP})$. Отметим также, что чем шире наш колольчик, тем устойчивее будет модель, ведь $Q(w)$ будет в таком случае слабо изменяться в окрестности значений параметра w_{MP}.

Распишем чуть подробнее (105) как функцию от A:

$$\begin{aligned}&\log p(T \mid X, A) \approx \\ &\approx \frac{d}{2}\log(2\pi) + \log p(T \mid X, w_{MP}) + \log \mathcal{N}\left(w_{MP} \mid 0, A^{-1}\right) - \frac{1}{2}\log\det\left(X^T R(w_{MP})X + A\right).\end{aligned} \tag{106}$$

Полученную функцию уже можно оптимизировать по A. Эффективный подход к этой задаче оптимизации рассмотрен в следующем разделе.

5.4 Оптимизация обоснованности на основе аппроксимации Лапласса

Замечание. Вплоть до этого момента мы обозначали w_{MP} точку максимума $p(T \mid X, w)p(w \mid A)$ при некоторой фиксированной матрице A. В данном разделе нам придётся переобозначить w_{MP} как w^A_{MP}:

$$w^A_{MP} = \arg\max_w p(T \mid X, w)p(w \mid A), \tag{107}$$

для того чтобы подчеркнуть зависимость w_{MP} от матрицы A, по которой мы оптимизируем.

Итак, мы хотим решить задачу

$$\log p(T \mid X, A) \to \max_A. \tag{108}$$

Воспользовавшись приближением (106), оптимальную A можно найти, оптимизируя по A функцию:

$$F(A, w^A_{MP}) := \log p(T \mid X, w^A_{MP}) + \log \mathcal{N}\left(w^A_{MP} \mid 0, A^{-1}\right) - \frac{1}{2}\log\det\left(X^T R(w^A_{MP})X + A\right). \tag{109}$$

Для этого мы решим с помощью метода Ньютона систему уравнений относительно α_j (напомним, что $A = \mathrm{diag}(\alpha_1, \ldots, \alpha_d)$).

$$\frac{\partial F(A, w^A_{MP})}{\partial \log \alpha_j} = \alpha_j \frac{\partial F(A, w^A_{MP})}{\partial \alpha_j} = 0, \qquad j = 1, \ldots, d. \tag{110}$$

Основная проблема заключается в том, что зависимость величины w^A_{MP} от A очень сложна, а при взятии производной (110) без нахождения $\frac{\partial w^A_{MP}}{\partial \alpha_j}$ не обойтись. Однако, можно заметить, что $F(A, w^A_{MP}) \geq F(A, w))$, для любого w, при фиксированной матрице A. Дифференцирование такой оценки аналогично взятию производной, считая $w^A_{MP} = \mathrm{const}$ относительно A.

Давайте распишем $F(A, w^A_{MP})$ подробнее:

$$F(A, w^A_{MP}) = -\sum_{n=1}^{N} \log\left(1 + \exp\left(t_n (w^A_{MP})^T x_n\right)\right) - \frac{1}{2}(w^A_{MP})^T A w^A_{MP} + \tag{111}$$

$$\frac{1}{2}\log\det A - \frac{1}{2}\log\det\left(X^T R(w^A_{MP})X + A\right) + \mathrm{const}. \tag{112}$$

Возьмём логарифмическую производную $F(A, w_{MP}^A)$, считая $w_{MP}^A = \text{const}$. Рассмотрим самое нетривиальное слагаемое подробно:

$$\frac{\partial}{\partial \log \alpha_j} \log\det\left(X^T R(w_{MP}^A)X + A\right) = \alpha_j \frac{\partial}{\partial \alpha_j} \log\det\left(X^T R(w_{MP}^A)X + A\right) = \quad (113)$$

$$= \alpha_j \operatorname{tr}\left(\left(X^T R(w_{MP}^A)X + A\right)^{-1} E_{jj}\right) = \quad (114)$$

$$= \alpha_j \left[\left(X^T R(w_{MP}^A)X + A\right)^{-1}\right]_{jj}. \quad (115)$$

Таким образом, при $w_{MP}^A = \text{const}$:

$$0 = \frac{\partial F(A, w_{MP}^A)}{\partial \log \alpha_j} = -\frac{\alpha_j}{2}\left[\left(w_{MP}^A\right)_j\right]^2 + \frac{1}{2} - \frac{\alpha_j}{2}\left[\left(X^T R(w_{MP}^A)X + A\right)^{-1}\right]_{jj}. \quad (116)$$

Шаг метода оптимизации для такой задачи можно записать так:

$$\alpha_j^{new} = \frac{1 - \alpha_j^{old}\left[\left(X^T R(w_{MP}^{A^{old}})X + A^{old}\right)^{-1}\right]_{jj}}{\left[\left(w_{MP}^{A^{old}}\right)_j\right]^2}. \quad (117)$$

По факту, мы должны делать итеративно следующие два шага[18]:

1. Найти $w_{MP}^{A^{old}}$ для текущей матрицы A^{old}
2. Найти A^{new} по формуле (117)

И это будет работать! Интуитивно это можно представить себе так: мы итеративно шагаем в сторону оптимального значения A, постоянно подкручивая веса w_{MP}^A. На практике такой подход часто работает очень неплохо: довольно быстро α_j, которые соответствуют нерелевантным признакам, начинают стремиться к бесконечности.

Мы рассмотрели, как можно оптимизировать оценку на правдоподобие модели, пользуясь приближением Лапласа для оценки значения интеграла. Этот способ хорошо работает на практике, однако, существуют и другие методы оценить интересующий нас интеграл. Рассмотрим еще один такой способ, чтобы лучше разобраться с техникой вариационных нижних оценок, которая еще не раз пригодится нам в дальнейшем.

5.5 Вариационная нижняя оценка сигмоиды

В предыдущем пункте мы приближали подынтегральную функцию в выражении для обоснованности с помощью гауссианы, после чего интеграл легко брался. Теперь мы будем действовать иначе и построим вариационную нижнюю оценку к подынтегральному выражению, причем такую, чтобы после приближения можно было аналитически посчитать интеграл. Напомним, что функция $g(x, \xi)$ называется вариационной нижней оценкой функции $f(x)$, если

1. $\forall x, \xi \; f(x) \geq g(x, \xi)$
2. $\forall x \; \exists \xi(x) : f(x) = g(x, \xi(x))$

Про вариационную нижнюю оценку можно думать так: у нас есть не одна нижняя оценка, а целый континуум, индексируемый параметром ξ. При этом, для любого x найдется такая функция из этого континуума, значение которой точно совпадает со значением исходной функции в точке x (см. Рис.6). Как обсуждалось ранее, если итеративно максимизировать вариационную нижнюю оценку $g(x, \xi)$ по вариационным параметрам ξ и по

[18] Заметим, что данная итеративная процедура аналогична той, которую мы получили для задачи регрессии на предыдущей лекции

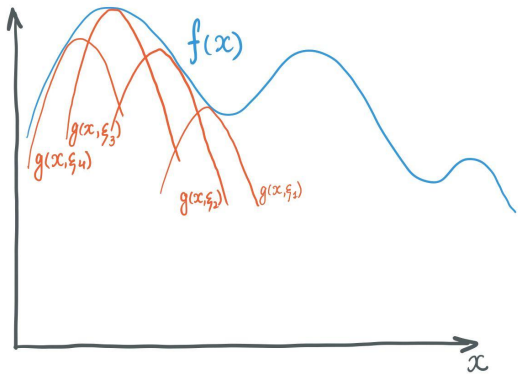


Рис. 6: Возможный вид вариационной нижней оценки при различных значениях вариационного параметра

исходным параметрам x, то такая процедура в итоге сойдется к локальному максимуму исходной функции $f(x)$ (см. выражение 73).[19]

Итак, построим вариационную нижнюю оценку к подынтегральной функции. Подынтегральная функция в выражении для обоснованности есть произведение N сигмоид и нормального распределения:

$$p(T \mid X, A) = \int p(T \mid X, w)p(w \mid A)\ dw = \int \prod_{n=1}^{N} \frac{1}{1+\exp(-t_n w^T x_n)} \mathcal{N}\left(w \mid 0, A^{-1}\right) dw \quad (118)$$

Попробуем оценить произведение сигмоид чем-нибудь хорошим (чтобы интеграл потом взялся аналитически). Забегая вперед, скажем, что это можно сделать ненормированными гауссианами (см Рис. 7). Как мы увидим далее, такая оценка будет и нижней, и вариационной, но насколько такое приближение хорошо описывает исходную функцию? На самом деле, не очень хорошо, слишком уж гауссиана не похожа на сигмоиду. Однако, для нашей задачи такое приближение подходит, поскольку нам нужно оценить не одну сигмоиду, а их произведение, а оно имеет колоколообразный вид и хорошо описывается произведением гауссиан (поэтому каждую отдельную сигмоиду можно оценить гауссианой).

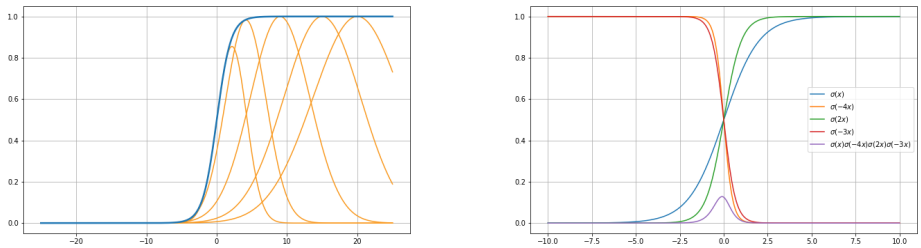


Рис. 7: Вариационная нижняя оценка сигмоиды ненормированными гауссианами (слева). Сигмоиды и их произведение (справа)

Итак, будем искать вариационную нижнюю оценку для сигмоиды. Заметим, что если функция выпуклая, то ее вариационная нижняя оценка есть все ее касательные[20]. Логисти-

[19] Заметим, что мы строим вариационную нижнюю оценку не к самому интегралу, а к подынтегральному выражению, поэтому здесь мы не можем гарантировать сходимость оптимизации нашей оценки к локальному максимуму исходного интеграла.

[20] Действительно, для выпуклой $f(x)$ вариационная нижняя оценка имеет вид:

$$f(x) \geq f'(\xi)(x-\xi) + f(\xi),$$

где точка касания ξ – вариационный параметр.

ческая функция не является выпуклой, поэтому напрямую построить касательные к ней не имеет смысла. Но можно преобразовать её к выпуклой функции, построить касательные в новых координатах, а затем найти их уравнение в исходных координатах.

Применим серию преобразований:

$$\log \sigma(x) = -\log(1 + \exp(-x))$$

— вогнутая функция (а нам нужна выпуклая). Продолжим:

$$\log \sigma(x) = -\log(1 + \exp(-x)) = -\log\Big(\exp\Big(-\frac{x}{2}\Big)(\exp\Big(\frac{x}{2}\Big) + \exp\Big(-\frac{x}{2}\Big))\Big) = \\ = \frac{x}{2} - \log\Big(\exp\Big(-\frac{x}{2}\Big) + \exp\Big(\frac{x}{2}\Big)\Big) \quad (119)$$

Рассмотрим второе слагаемое, являющееся чётной функцией. Сделаем замену $y = x^2$:

$$-\log\big(e^{-\frac{x}{2}} + e^{\frac{x}{2}}\big) = -\log\Big(e^{-\frac{\sqrt{y}}{2}} + e^{\frac{\sqrt{y}}{2}}\Big) \quad (120)$$

Полученная функция является выпуклой и определена на полуинтервале $[0, +\infty)$. Ее вариационную нижнюю оценку можно построить касательной. Выпишем производную по y:

$$\frac{d\left(-\log\left(e^{-\frac{\sqrt{y}}{2}} + e^{\frac{\sqrt{y}}{2}}\right)\right)}{dy} = -\tanh\left(\frac{\sqrt{y}}{2}\right)\frac{1}{4\sqrt{y}}. \quad (121)$$

С учетом общего вида уравнения касательной в точке ξ: $f'(\xi)(x-\xi) + f(\xi)$,, получаем:

$$-\frac{1}{4\sqrt{\xi}}\tanh\left(\frac{\sqrt{\xi}}{2}\right)(y-\xi) - \log\left(e^{-\frac{\sqrt{\xi}}{2}} + e^{\frac{\sqrt{\xi}}{2}}\right) = -\frac{1}{4|\eta|}\tanh\left(\frac{|\eta|}{2}\right)(x^2-\eta^2) - \log\Big(e^{-\frac{|\eta|}{2}} + e^{\frac{|\eta|}{2}}\Big). \quad (122)$$

где мы переопределили вариационный параметр как $|\eta| = \sqrt{\xi}$. Итого, для $\sigma(x)$ получаем следующую нижнюю оценку:

$$\sigma(x) \geq \exp\left(\frac{x}{2} - \frac{1}{4|\eta|}\tanh\left(\frac{|\eta|}{2}\right)(x^2-\eta^2) - \log\Big(e^{-\frac{|\eta|}{2}} + e^{\frac{|\eta|}{2}}\Big)\right) = \quad (123)$$

$$= \exp\left(\frac{x}{2} - \frac{1}{4\eta}\tanh\Big(\frac{\eta}{2}\Big)(x^2-\eta^2) - \log\Big(e^{-\frac{\eta}{2}} + e^{\frac{\eta}{2}}\Big)\right) = \quad (124)$$

$$= \sigma(\eta)\exp\left(\frac{x-\eta}{2}\right)\exp\left(-\frac{1}{4\eta}\tanh\Big(\frac{\eta}{2}\Big)(x^2-\eta^2)\right), \quad (125)$$

где мы убрали модули у второго и третьего слагаемого под экспонентой, т.к. эти функции четные, и воспользовались выражением 119.

Как мы говорили ранее, полученная оценка[21], как функция от x, является ненормированной гауссианой (как экспонента от квадратичной по аргументу функции). Интеграл от произведения гауссиан берется аналитически и итоговое выражение можно промаксимизировать по параметрам матрицы ковариациии A. На практике чаще используется вариант с приближением Лапласа. Однако, альтернативный подход интересен в качестве математического упражнения, которое помогает лучше понять общий принцип использования вариационных оценок.

[21] Эта вариационная оценка именная, получена Джааккола и Джорданом (Tommi S. Jaakkola, Michael Jordan) в 2000 году. Так же заметим, что касание сигмоиды и гауссианы происходит в двух точках, при $x = \eta$ и $x = -\eta$

6 Лекция 6. EM-алгоритм и модели со скрытыми переменными

Это ключевая лекция курса, в которой мы поймём как и зачем нужно строить модель со скрытыми (или латентными) переменными и какими методами можно такие модели обучать.[22] В классическом курсе по машинному обучения EM-алгоритм обычно рассматривается на примере разделения смеси гауссиан. В этом курсе рассмотрим несколько более интересных примеров.

Итак, мы будем решать следующую задачу:

Задача 1. По выборке X восстановить параметры θ распределения методом максимального правдоподобия:

$$p(X \mid \theta) \to \max_{\theta}.$$

Вопрос. В каких параметрических семействах эту задачу можно решить эффективно?

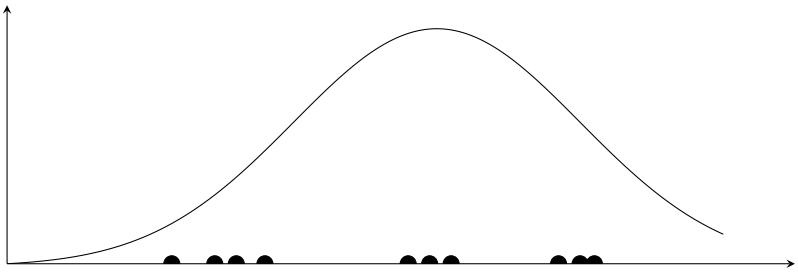
Ответ. Если плотность распределения $p(X \mid \theta)$ лежит в экспоненциальном классе, то мы можем эффективно найти оценку максимального правдоподобия для параметров θ. Иногда это возможно в явном виде (дифференцируем логарифм правдоподобия, приравниваем к нулю, и находим из полученной системы уравнений параметры θ), а в остальных случаях можно построить эффективную численную процедуру оценки (благодаря тому, что логарифм функции правдоподобия — вогнутая функция[23]).

Проблема заключается в том, что экспоненциальный класс не такой широкий, как могло бы показаться. Зачастую на практике наблюдаемые данные имеют гораздо более сложное распределение, которое в экспоненциальный класс никак не вписывается. Возникает дилемма:

- либо пытаться «натянуть ежа на глобус» и вписать распределение из экспоненциального класса в выборку, которая пришла из более сложного распределения (оно будет плохо описывать данные, но зато мы сможем эффективно решить Задачу 1);
- либо переходить к гораздо более сложным семействам распределений, обладающим достаточной гибкостью, чтобы описать данные, но в этом случае процесс нахождения максимума в Задаче 1 может сходиться слишком медленно

Пример. Рассмотрим следующую одномерную выборку:

Можно попытаться восстановить плотность распределения выбрав какое-то параметрическое семейство из экспоненциального класса. Например, нормальные распределения. Получим примерно следующую гауссиану:



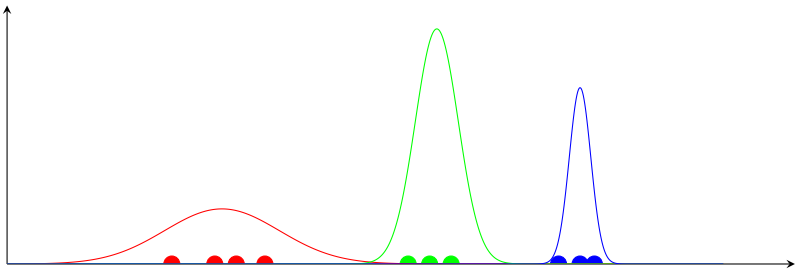
Она наиболее правдоподобно описывает наблюдаемые данные в семействе нормальных распределений. Однако, с точки зрения здравого смысла, модель не очень хорошая. Данные явно пришли не из гауссианы.

[22] Как мы увидим далее, многие методы обучения моделей со скрытыми переменными являются вариациями EM-алгоритма.

[23] Даже в пространствах высокой размерности для вогнутых функций существуют эффективные способы нахождения глобального (он единственный) максимума.

С другой стороны можем предположить, что данные приходят из нескольких гауссиан:

Тогда можем восстановить параметры каждой из гауссиан:



К сожалению, у нас нет информации о цветах, т.е. мы не знаем, какой объект из какой гауссианы пришёл. Однако, если бы у нас была такая информация, то задача восстановления плотности распределения сразу стала бы гораздо проще, потому что свелась бы к решению Задачи 1 для нескольких плотностей из экспоненциального класса.

На этом и строится идеология моделей с латентными переменными: мы говорим, что наши данные пришли из довольно сложного распределения, и имеют сложную природу; но если бы мы знали что-нибудь ещё, какие-то дополнительные переменные, то наше распределение стало бы гораздо более простым.

Можно показать, что для любого распределения не из экспоненциального класса можно ввести дополнительные (латентные) переменные так что совместное распределение на исходные и латентные переменные будет лежать в экспоненциальном классе. Итак, вместо того чтобы решать задачу 1 мы будем решать следующую задачу:

Задача 2. Введём латентные переменные Z так, чтобы совместное распределение $p(X, Z \mid \theta)$ лежало в экспоненциальном классе. Вместо решения исходной задачи (которую мы теперь будем называть задачей максимизации неполного правдоподобия) будем решать задачу

$$p(X, Z \mid \theta) \to \max_{\theta}.$$

Замечание. Помимо того что мы решим исходную задачу, мы также получим информацию о возможных значениях латентных переменных. На практике существует много задач, в которых информация о Z гораздо важнее информации о θ. В дальнейшем мы рассмотрим несколько таких примеров.

6.1 Вывод EM-алгоритма

Итак, пусть мы смогли ввести такие дополнительные переменные Z, что совместное распределение $p(X, Z \mid \theta)$ стало лежать в экспоненциальном классе. Таким образом, мы можем сравнительно легко найти оценку максимального правдоподобия на параметры θ. В частности, функция $\log p(X, Z \mid \theta)$ вогнута по θ при фиксированных X и Z.

Записываем цепочку тождеств:[24]

$$\begin{aligned}\log p(X \mid \theta) &= \int q(Z) \log p(X \mid \theta) dZ = \\ &= \int q(Z) \log \frac{p(X, Z \mid \theta)}{p(Z \mid X, \theta)} dZ = \\ &= \int q(z) \log \left[\frac{p(X, Z \mid \theta)}{p(Z \mid X, \theta)} \frac{q(Z)}{q(Z)} \right] dZ = \\ &= \int q(Z) \log \frac{p(X, Z \mid \theta)}{q(Z)} dZ + \int q(Z) \log \frac{q(Z)}{p(Z \mid X, \theta)} dZ.\end{aligned} \tag{126}$$

Здесь $q(Z)$ — произвольноое распределение в пространстве латентных переменных. Рассмотрим повнимательнее получившиеся слагаемые. Для этого вспомним определение и некоторые свойства дивергенции Кульбака-Лейблера.

Определение 6. Дивергенция Кульбака—Лейблера между двумя распределениями p и q определяется следующим образом:

$$KL(p(x) \parallel q(x)) = \int p(x) \log \frac{p(x)}{q(x)} dx.$$

Свойство (KL-дивергенции). $KL(p \parallel q) \geq 0$, причём $KL(p \parallel q) = 0$ если и только если эти распределения почти всюду (везде кроме множества меры ноль) совпадают.

Упражнение. Докажите это свойство при помощи неравенства Йенсена.

Замечание. У KL-дивергенции есть теоретико-информационный смысл. Если мы работаем с дискретными случайными величинами, KL-дивергенция показывает, на сколько дополнительных бит длиннее будет сообщение при не оптимальном кодировании: если символы приходят из распределения p, а кодируем мы их как будто они приходят из распределения q.

Вернемся к (126). Заметим, что первое слагаемое не является KL-дивергенцией, поскольку у него под логарифмом стоит отношение совместного распределения $p(X, Z \mid \theta)$ и $q(Z)$, а эти распределения лежат в разных пространствах. А вот второе слагаемое является KL-дивергенцией распределений $q(Z)$ и $p(Z \mid X, \theta)$. Тогда, вследствие неотрицательности KL-дивергенции можем записать следующее неравенство:

$$\log p(X \mid \theta) \geq \int q(Z) \log \frac{p(X, Z \mid \theta)}{q(Z)} dZ \tag{127}$$

Идея EM-алгоритма заключается в том, чтобы вместо оптимизации логарифма неполного правдоподобия оптимизировать полученную нижнюю оценку, но теперь уже как по θ так и по распределению q.

Определение 7. Правая часть выражения 127 называется нижней границей на обоснованность (ELBO, evidence lower bound) и обозначается $\mathcal{L}(q, \theta)$.

Вопрос. Является ли нижняя граница на обоснованность вариационной нижней оценкой. Почему?

Ответ. Напомним, что вариационная нижняя оценка требует выполнения двух свойств:

- она всегда не превосходит выражения которое она оценивает (этому как раз удовлетворяет (127));
- для любого аргумента исходной функции (θ) найдутся такие значения вариационных (q), для которых неравенство превращается в равенство. В самом деле, если взять $q(Z) = p(Z \mid X, \theta)$, тогда их KL-дивергенция обратится в ноль, а неравенство — в равенство.

[24] Поскольку $\log p(X \mid \theta)$ от Z не зависит, то первый переход является домножением на единицу: $\int q(Z) dZ = 1$. Во втором переходе используется правило для условной вероятности. Третий переход — опять же умножение на единицу. Наконец в последнем переходе мы сгруппировали числители и знаменатели под логарифмом «накрест» и разбили логарифм произведения на сумму двух логарифмов.

Благодаря этому свойству мы можем перейти от оптимизации неполного правдоподобия к оптимизации нижней оценки на обоснованность. Будем решать полученную задачу итерационно:

- оптимизировать по q при фиксированном θ (E-шаг):

$$\mathcal{L}(q, \theta_0) \to \max_q \implies q(Z) = p(Z \mid X, \theta); \tag{128}$$

- оптимизировать по θ при фиксированном q (M-шаг):

$$\mathcal{L}(q_0, \theta) \to \max_\theta \iff \int q(Z) \log p(X, Z \mid \theta) dZ \to \max_\theta . \tag{129}$$

На E-шаге у нас задача функциональной оптимизации. В общем случае такие задачи решать эффективно невозможно. Однако, есть одно обстоятельство, которое позволяет легко с этим разобраться. А именно, сумма в (126) не зависит от q, а потому максимизация по q первого слагаемого эквивалентна минимизации по q второго слагаемого, а второе слагаемое — KL-дивергенция. Мы знаем, где она достигает минимума, потому и приравниваем $q(Z) = p(Z \mid X, \theta)$. Таким образом, если модель позволяет выполнить аналитический байесовский вывод на латентные переменные (т.е. найти апостериорное распределение $p(Z \mid X, \theta)$), то E-шаг проделывается в явном виде.

6.2 Обсуждение EM-алгоритма и примеры

Вопрос. Что делать, если невозможно аналитически посчитать апостериорное распределение $p(Z \mid X, \theta))$ на E-шаге?

Ответ. В этом случае E-шаг выполняется приближенно[25]. Как было замечено выше, максимизация нижней оценки на обоснованность по q эквивалентна минимизации KL-дивергенции между $q(Z)$ и апостериорным распределением $p(Z \mid X, \theta)$. Поэтому приближенный E-шаг будет выглядеть так:

$$q(Z) = \arg\min_{q \in Q} KL(q(Z) \parallel p(Z \mid X, \theta)),$$

где Q — некоторое параметрическое семейство, в котором мы ищем оптимальное $q(Z)$.

Вопрос. Что будет, если $Q = \Delta$, где Δ – семейство всевозможных δ-функций?

Ответ. Получим «байесовский вывод для бедных», т.е. точечную оценку на параметры θ, максимизирующую апостериорное распределение $p(Z \mid X, \theta)$:

$$\int \delta(Z - Z_0) \log \frac{\delta(Z - Z_0)}{p(Z \mid X, \theta)} dZ \to \min_{Z_0},$$

что эквивалентно

$$C - \int \delta(Z - Z_0) \log p(Z \mid X, \theta) dZ \to \min_{Z_0},$$

или же

$$\log p(Z_0 \mid X, \theta) \to \max_{Z_0} .$$

Т.е. «байесовский вывод для бедных» является приближением апостериорного распределения с помощью дельта-функции.

Таким образом, E-шаг всегда можно проделать приближённо, даже если сопряженных распределений нет и аналитический вывод невозможен.

[25] EM-алгоритм с таким E-шагом называется приближенным. Тем не менее, он будет обладать рядом приятных свойств. В частности, вариационная нижняя оценка будет монотонно расти, а потому алгоритм будет гарантированно сходиться. Но, вообще говоря, необязательно к точке локального максимума неполного правдоподобия

Замечание. В дальнейших лекциях будет рассмотрен «промежуточный» случай когда аналитический байесовский вывод невозможен, но семейство Q более широкое чем семейство дельта-функций. Оказывается, что в некоторых случаях такие задачи можно эффективно решать. Эти идеи лежат в основе современных нейробайесовских методов, когда эта парадигма применяется к нейронным сетям.

Вопрос. Что можно сказать про М-шаг? На первый взгляд, максимизируется какое-то матожидание, какой-то интеграл, который может даже не взяться. Что делать?

Ответ. Напомним, что $\log p(X, Z \mid \theta)$ вогнутая по θ функция (мы так вводили скрытые переменные Z). Однако мы оптимизируем не её саму а её матожидание. К счастью, матожидание вогнутой функции — функция также вогнутая.[26] Таким образом, даже если аналитическая оптимизация такого выражения невозможна, то численная всегда возможна и эффективна, даже в пространствах высокой размерности.

Итак, резюмируя, Е-шаг иногда можно выполнить аналитически или хотя бы сделать «байеса для бедных», а на М-шаге у нас задача оптимизации вогнутой функции. Так выглядит классический EM-алгоритм (формулы 128 129). Существуют разные экспериментальные постановки при которых эти процессы необходимо модифицировать. Например, есть стохастический EM-алгоритм, ME-алгоритм, вариационный EM-алгоритм и множество других модификаций. Все они так или иначе опираются на эту базовую схему, немного её модифицируя.

Пример. Предположим мы попали в следующую ситуацию: на Е-шаге возможно аналитически рассчитать распределение $p(Z \mid X, \theta_0)$, а на М-шаге неберущийся интеграл и для нахождения максимума приходится выполнять большое количество итераций численного метода оптимизации.

Вопрос. Как можно тогда оптимизировать (сделать более эффективной) процедуру?

Ответ. Например, на М-шаге необязательно дожидаться сходимости, можно выполнить одну либо небольшое количество итераций численного метода оптимизации.

Замечание. Бывает и наоборот, когда М-шаг быстрый, а на Е-шаге приходится численно оптимизировать KL-дивергенцию. Тогда опять же можно останавливать численный метод оптимизации раньше.

В любом из таких случаев у нас всё равно будет выполняться свойство монотонного роста вариационной нижней оценки, а потому сходимость нам гарантирована.

Пример. Наиболее известным примером EM-алгоритма, безусловно, выступает разделение смеси гауссиан.

Вопрос. Что выступает в роли латентных переменных для задачи разделения смеси гауссиан?

Ответ. Номера гауссиан из которых пришли объекты.

Если вспомнить, то алгоритм разделения смеси гауссиан представляет собой как раз-таки (128) и (129): на Е-шаге мы для каждого объекта рассчитываем вероятность того что он пришёл из каждой из гауссиан, а на М-шаге, хотя нигде явно и не записываем интеграл, но пользуемся взвешенными оценками максимального правдоподобия, которые как раз и являются аргмаксимумами (129).

6.3 Байесовский метод главных компонент

В этой части лекции мы посмотрим на классический метод главных компонент с байесовской точки зрения. Конкретнее, мы сформулируем эту модель на языке моделей с латентными переменными и обсудим какие преимущества нам это даёт.

[26] Поскольку выпуклая комбинация вогнутых функций – вогнутая функция, а любое матожидание – это выпуклая комбинация из бесконечного числа слагаемых.

Напомним, что метод главных компонент решает задачу уменьшения размерности признакового пространства. Итак, пусть мы наблюдаем данные $x \in \mathbb{R}^D$ и хотим найти линейное подпространство заданной размерности d в котором содержится наибольшая часть дисперсии наблюдаемых данных. Задача решается в явном виде: строим ковариационную матрицу размера $D \times D$ по нашим объектам X, приводим её к главным осям и проецируем её на d собственных векторов, отвечающих наибольшим собственным значениям.

Оказывается, то же самое можно сделать на вероятностном языке. Вводим модель с латентными переменными:

$$p(x, z \mid \theta) = p(x \mid z, \theta)p(z) = \mathcal{N}(x \mid \mu + Wz, \sigma^2 I)\,\mathcal{N}(z \mid 0, I), \tag{130}$$

где $z \in \mathbb{R}^d$ и играет роль сжатого представления исходного вектора $x \in \mathbb{R}^D$. В роли параметров модели θ выступают вектор $\mu \in \mathbb{R}^D$, линейный оператор $W \in \mathbb{R}^{D \times d}$ и скаляр σ. Эта вероятностная модель говорит, что у каждого x размерности D есть некоторое латентное представление z размерности d такое, что x является результатом действия линейного оператора W на z плюс какой-то сдвиг μ и плюс какой-то шум.

Поскольку мы наблюдаем только $X = (x_1, \ldots, x_n)$, в модели 130 переменные z являются скрытыми. Соответственно, исходная задача подбора параметров модели θ ставится как

$$\theta_{ML} = \arg\max_{\theta} P(X \mid \theta). \tag{131}$$

Представим неполное правдоподобие как интеграл от совместной плотности:

$$\theta_{ML} = \arg\max_{\theta} p(X \mid \theta) = \arg\max_{\theta} \int p(X \mid Z, \theta)p(Z)dZ. \tag{132}$$

Замечание. Если устремить $\sigma \to 0$ то полученное θ_{ML} для остальных параметров будет стремится к классической оценке из метода главных компонент.

Интеграл в выражении 132 берется аналитически, поскольку априорное распределение и правдоподобие сопряжены. Предположим однако, что мы не умеем брать такой интеграл, и выпишем для этой же задачи EM-алгоритм. Во-первых это полезное упражнение, а во-вторых в некоторых ситуациях применять EM-алгоритм оказывается более эффективно чем решать задачу аналитически. Итак,

- E-шаг:

$$\begin{aligned} q(Z) = p(Z \mid X, \theta) = \\ = \frac{p(X \mid Z, \theta)p(Z)}{\int p(X \mid Z, \theta)p(Z)dZ} = \\ = \frac{\prod_{i=1}^n p(x_i \mid z_i, \theta)p(z_i)}{\int \prod_{i=1}^n p(x_i \mid z_i, \theta)p(z_i)dz_i} = \\ = \prod_{i=1}^n \frac{p(x_i \mid z_i, \theta)p(z_i)}{\int p(x_i \mid z_i, \theta)p(z_i)dz_i} = \\ = \prod_{i=1}^n p(z_i \mid x_i, \theta). \end{aligned} \tag{133}$$

Если аккуратно расписать распределения в последних двух строчках 133, получим:

$$z_i \sim \mathcal{N}\left((\sigma^2 I + W^T W)^{-1} W^T (x_i - \mu), (I + \sigma^{-2} W^T W)^{-1}\right).$$

- M-шаг:

$$\begin{aligned}
&\mathbb{E}_Z \log p(X, Z \mid \theta) = \\
&= \mathbb{E}_Z \left(\sum_{i=1}^{n} \log p(x_i \mid z_i, \theta) + \log p(z_i) \right) = \\
&= C + \sum_{i=1}^{n} \mathbb{E}_{z_i} \left[-\tfrac{D}{2} \log 2\pi - D \log \sigma - \tfrac{1}{2\sigma^2} (x_i - \mu - W z_i)^T (x_i - \mu - W z_i) \right] = \\
&= C + \sum_{i=1}^{n} \left(-\tfrac{D}{2} \log \sigma - \tfrac{1}{2\sigma^2} \mathbb{E}_{z_i} \left((x_i - \mu)^T (x_i - \mu) - 2(x_i - \mu)^T W z_i + z_i^T W^T W z_i \right) \right) = \\
&= C + \sum_{i=1}^{n} \left(-\tfrac{D}{2} \log \sigma - \tfrac{1}{2\sigma^2} \left((x_i - \mu)^T (x_i - \mu) - 2(x_i - \mu)^T W \, \mathbb{E}\, z_i + \operatorname{tr} \left[W^T W \, \mathbb{E}[z_i z_i^T] \right] \right) \right)
\end{aligned} \tag{134}$$

Осталось понять что происходит с матожиданиями. На самом деле $\mathbb{E}\, z_i$ мы уже выписывали на E-шаге, а для $x \sim \mathcal{N}(X \mid \mu, \Sigma)$ имеет место $\mathbb{E}\, xx^T = \Sigma + \mu\mu^T$, что является матричным (многомерным) обобщением того факта, что матожидание квадрата есть дисперсия плюс квадрат матожидания.

Теперь это выражение необходимо прооптимизировать по σ, μ, и W. Проделаем это для W, дифференцируем полученное выражение для W и приравниваем производную к нулю:

$$\sum_{i=1}^{n} \frac{1}{2\sigma^2} \left[-2(x_i - \mu) \, \mathbb{E}\, z_i^T + 2W \, \mathbb{E}[z_i z_i^T] \right] = 0,$$

откуда

$$\sum_{i=1}^{n} (x_i - \mu) \, \mathbb{E}\, z_i^T - W \sum_{i=1}^{n} \mathbb{E}\, z_i z_i^T = 0,$$

и наконец

$$W = \left(\sum_{i=1}^{n} (x_i - \mu) \, \mathbb{E}\, z_i^T \right) \left(\sum_{i=1}^{n} \mathbb{E}\, z_i z_i^T \right)^{-1}. \tag{135}$$

Упражнение. Выведите формулы для μ и σ.

6.3.1 Вычислительная сложность

Формула 135 представляет практический интерес, потому что её вычислительная сложность составляет $O(nDd + nd^2 + d^3)$. На практике часто $n > D > d$, т.е. вычислительная сложность 135 равна $O(nDd)$. В то же время сложность метода главных компонент есть $O(nD^2 + D^3) = O(nD^2)$. Если $D \gg d$, то за время выполнения аналитических расчётов можно успеть сделать достаточно много итераций EM-алгоритма для сходимости.

Пример. Если $D = 10'000$ и $d = 5$, а EM-алгоритм сходится за 200 итераций, то он будет работать в 10 раз быстрее классического метода главных компонент. Похожая ситуация наблюдается и с другими классическими методами которые в пространствах большой размерности работают медленнее чем итерационные процессы.

6.3.2 Пропуски в данных

Предположим, что в наших данных есть пропуски.

Вопрос. Что делает базовый метод главных компонент, если в данных есть пропуски?

Ответ. В общем случае — ничего, он не умеет работать с такими данными. Если пропусков мало, или они относятся к малому числу признаков или к малому числу объектов, то можно заполнять средними значениями или просто выбрасывать объекты или признаки в которых есть пропуски, но в общем случае это не работает.

С точки зрения EM-алгоритма, однако, можно считать пропущенные значения дополнительными латентными переменными. Это в каком-то смысле стирает грань между X и Z: часть исходных признаков может быть неизвестна, а часть результирующих признаков может быть известна, или по крайней мере мы можем располагать какой-то дополнительной информацией о них.

X			Z
1	2	3	?
1	3	2	?
2	1	3	?
2	3	1	?
3	1	2	?
3	2	1	?

Таблица 3: Данные для базового PCA.

X			Z
?	?	3	0
1	3	2	?
2	?	?	1
2	3	1	?
?	1	?	2
3	2	1	?

Таблица 4: Данные для EM-алгоритма.

Пример. Предположим, что нас интересуют всё те же векторные представления слов, но на сей раз мы явно требуем, чтобы первая компонента отображала эмоциональную окраску слова. Тогда у слов «дурак», «сволочь», «негодяй» первая компонента должна быть отрицательной, а у слов «умница», «хорошист», «молодец» — положительной. Можно заложить такую информацию в модель, после чего остальные слова непременно растянутся по эмоциональной шкале.

Это свойство байесовского метода главных компонент представляет практический интерес, потому что ручная разметка данных обычно стоит дорого. Поэтому крайне важно иметь модели, способные обучаться по частично размеченной выборке, т.е. по данным, в которых для части объектов какие-то признаки неизвестны.

6.3.3 Расширения

1. Байесовское расширение метода главных компонент.

   Пусть теперь $W = (W_1, ..., W_D)$, где $W_j \in \mathbb{R}^{D\times 1}$, $A = diag(\alpha_1, ..., \alpha_D)$ и

   $$p(W \mid A) = \prod_{j=1}^{D} \left(\frac{\alpha_j}{2\pi}\right)^{D/2} \exp\left(-\frac{\alpha_j}{2}\|W_j\|^2\right) \tag{136}$$

   Определим модель следующим образом:

   $$p(X, Z, W \mid \mu, \sigma, A) = p(X, Z \mid W, \mu, \sigma)p(W \mid A) \tag{137}$$

   Если как и в методе релевантных векторов выписать выражение для обоснованности и максимизировать его по коэффициентам α_j, то оказывается, что значительная часть этих коэффициентов аналогично будет уходить в $+\infty$. Это соответствует отбрасыванию соответствующих столбцов матрицы W. Таким образом, введённое расширение решает вопрос выбора параметра d в методе главных компонент.

2. Смесь методов главных компонент.

   Расширим номенклатуру латентных переменных: введём дополнительную дискретную латентную переменную T:

   $$p(x, z, t \mid \theta) = p(x \mid z, t, \theta)p(z)p(t \mid \theta) = \mathcal{N}(x \mid \mu_t + W_t z, \sigma_t^2 I)\mathcal{N}(z \mid 0, I)\mathrm{Cat}(t \mid \theta). \tag{138}$$

   Такая модель говорит, что данные лежат в одном из нескольких линейных подпространств низкой размерности:

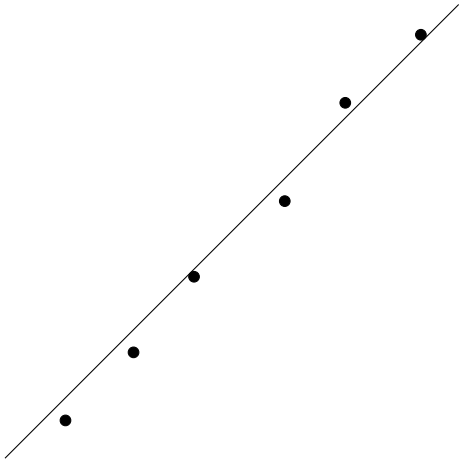


Рис. 8: Данные для базового PCA.

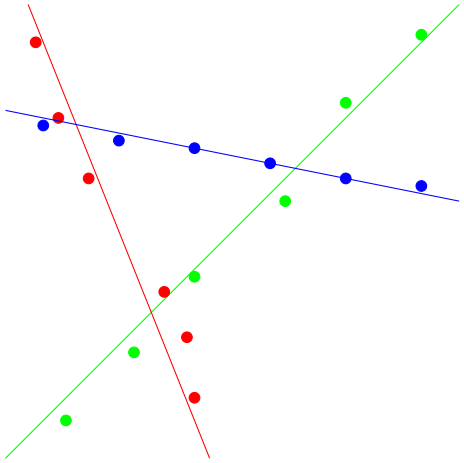


Рис. 9: Данные для смеси PCA.

3. Нелинейные методы.

   Вместо того чтобы считать что x получается из z посредством линейного преобразования можно взять любую другую настраиваемую нелинейную функцию, например, нейросеть. Применение нейросети в качестве нелинейной функции в байесовском методе главных компонент приведёт нас к вариационному автокодировщику (VAE, variational autoencoder).

   Вопрос. Какие проблемы могут возникнуть при таком подходе и как их решать?

   Ответ. Распределения перестают сопрягаться, а потому на E-шаге невозможно выполнить аналитический байесовский вывод. Тем не менее, как мы увидим далее в курсе, E-шаг можно сделать приближенно, подобрав вариационное распределение из заданного семейства, лежащее максимально близко к настоящему апостериорному распределению.

   Вопрос. Какова геометрическая интерпретация такого нелинейного подхода?

   Ответ. Многомерные данные зачастую лежат в (нелинейных) многообразиях более низких размерностей, и основная проблема состоит в том чтобы эти многообразия находить. Эту задачу и решает такой подход.

6.4 Пример применения EM-алгоритма на практике

Рассмотрим пример совмещения EM-алгоритма в модели word2vec. Данная модель позволяет строить векторные представления слов естественного языка, при этом полученные векторные представления сохраняют семантический смысл слов: алгебраические операции над векторными представлениями соответствуют семантическим операциям над словами (пример: «король» - «мужчина» + «женщина» = «королева»). Однако в зависимости от контекста слово может иметь различные значения, а векторное word2vec-представление этого слова останется неизменным. Например, слово «bank» может означать как «банк», так и «побережье».

Идея — построить векторные представления не для слов, а для их смыслов. Пусть дан корпус текстов — последовательность вхождений слов в предложения. При этом нам не дана разметка смыслов слов — заранее неизвестно, означает ли в текущем контексте слово «bank» «банк» или «побережье». Естественным образом в задаче возникают латентные переменные — для каждого вхождения слова заводим дискретную латентную переменную, которая показывает индекс значения слова в конкретном контексте. Количество возможных смыслов заранее не фиксируем, автоматически определяем структуру пространства латентных переменных (непараметрические Байесовские методы, будут рассмотрены в конце

курса). Полученную задачу можно решить с помощью EM-процедуры и теперь для каждого многозначного слова можно определить, какое значение слово имело в конкретном контексте, — этой информации не было в исходных данных (разметки смыслов слов нет)!

В результате для слова «bank» было обнаружено целых 5 смыслов:

1. Побережье: «The bank of the river».
2. Банк как здание: «Turn right at the bank».
3. Банк как место работы: «Yesterday, I started working in a bank».
4. Микрофинансовый смысл — банк как место, где люди хранят деньги.
5. Макрофинансовый смысл — банк как элемент финансовой системы государства.

7 Лекция 7. Вариационный Байесовский вывод

7.1 Вывод формул

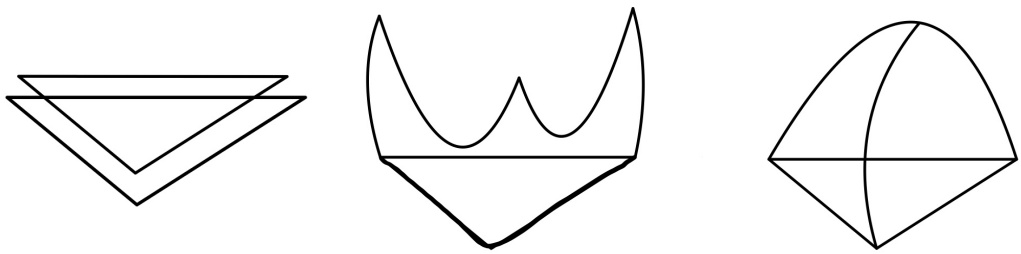
Пусть $x \in 1, ..., k$ с вероятностями $\pi_1, ..., \pi_k$. Будем говорить, что вектор вероятности лежит на k-значном вероятностном симплексе $\pi \in S_k$??, если $\pi_k \geq 0$ и $\sum_k \pi_k = 1$. Тогда x имеет категориальное распределение:

$$x \approx Cat(x \mid \pi) = \prod_{k=1}^{K} \pi_k^{[x=k]}$$

Тогда сопряженное априорное семейство распределений — распределение Дирихле:

$$p(\pi) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}, \quad \alpha_k > 0$$

- $\alpha_k = 1$ - равномерное распределение 10a
- $\alpha_k < 1$ - U-образное распределение 10b
- $\alpha_k > 1$ - колокообразное распределение 10c, причем $\mathbb{E}\,\pi_k = \frac{\alpha_k}{\sum_l^K \alpha_l}$



(a) Вероятностный симплекс при $K = 3$, $\alpha_k = 1$ - равномерное распределение

(b) Вероятностный симплекс при $K = 3$, $\alpha_k < 1$ - U-образное распределение

(c) Вероятностный симплекс при $K = 3$, $\alpha_k > 1$ - колокообразное распределение

Рис. 10: Описания для всех изображений

Сопряженное к многомерному нормальному распределению
Многомерное нормальное распределение:

$$\mathcal{N}(x \mid \mu, \Lambda^{-1}) = \frac{\sqrt{\det(\Lambda)}}{2\pi^{\frac{d}{2}}} \exp\left\{-\frac{1}{2}\operatorname{tr}(x-\mu)(x-\mu)^T\Lambda\right\}$$

где d — размерность x
Сопряженное распределение:

$$p(\Lambda) = \mathcal{W}(\Lambda \mid W, \nu) = \frac{1}{Const(w,\nu)}(\det\Lambda)^{\frac{\nu-d-1}{2}} \exp\left\{-\frac{1}{2}\operatorname{tr}\left(W^{-1}\Lambda\right)\right\}$$

где $\nu > d - 1$, $W = W^T \geq 0$
Причем $\mathbb{E}\,\lambda = \nu W$ и чем больше ν, тем меньше отклонение от мат. ожидания.

$$p(\mu, \Lambda) = \mathcal{N}(\mu \mid m, (\beta\Lambda)^{-1})\mathcal{W}(\Lambda \mid w, \nu)$$

7.2 EM-алгоритм

7.2.1 Классический EM-алгоритм

Для начала вспомним классический EM-алгоритм, который мы рассматривали на прошлой лекции. Пусть дана модель с наблюдаемыми переменными X и латентными переменными Z, параметризованная вектором θ:

$$p(X, Z \mid \theta). \tag{139}$$

Мы бы хотели оценить вектор параметров θ по методу максимального правдоподобия, но в качестве выборки нам даны только X, а Z мы не знаем. Таким образом, мы пытаемся получить оценку максимального правдоподобия по наблюдаемым данным, то есть решить задачу максимизации неполного правдоподобия:

$$\theta_{ML} = \arg\max_{\theta} p(X \mid \theta) = \arg\max_{\theta} \int p(X, Z \mid \theta) dZ. \tag{140}$$

Неполного потому, что мы не наблюдаем Z; если бы наблюдали и X, и Z, то у нас была бы стандартная задача максимизации (полного) правдоподобия. При этом на практике часто возможно посчитать только значение совместной плотности (139) при известных X и Z, но невозможно посчитать неполное правдоподобие в данной точке X (т.е. не можем посчитать интеграл в правой части (140)).

Пример. Латентные переменные естественно возникают в случае, когда плотность наблюдаемых переменных $p(X \mid \theta)$ имеет очень сложный характер. Тогда один из способов упрощения задачи — добавление латентных переменных до тех пор, пока совместное распределение (139) не станет принадлежать экспоненциальному классу распределений. У экспоненциального класса распределений функция правдоподобия является логарифмически вогнутой, в этом случае легко решать задачу её максимизации.

Возникает идея свести невыпуклую задачу (140) к выпуклой путём добавления латентных переменных. Перейдём к логарифму:

$$\theta_{ML} = \arg\max_{\theta} p(X \mid \theta) = \arg\max_{\theta} \log p(X \mid \theta). \tag{141}$$

Логарифм неполного правдоподобия можно разложить на вариационную нижнюю оценку и KL-дивергенцию между вариационным распределением $q(Z)$ и апостериорным распределением $p(Z \mid X, \theta)$:

$$\log p(X \mid \theta) = \mathcal{L}(q, \theta) + KL(q(Z) \parallel p(Z \mid X, \theta)), \ \forall q(Z). \tag{142}$$

Далее заменяем задачу максимизации левой части по θ на задачу максимизации вариационной нижней оценки $\mathcal{L}(q, \theta)$ по θ и по q. Распределение q в данном случае является вариационным параметром:

- $\forall q, \theta \ \mathcal{L}(q, \theta) \geq \log p(X \mid \theta)$, потому что $KL \geq 0$
- $\forall \theta \ \exists q(Z) = p(Z \mid X, \theta) : \log p(X \mid \theta) = \mathcal{L}(q, \theta)$, потому что $KL(p \parallel p) = 0$

Отсюда возникает итерационный EM-алгоритм:
E-step

$$q_n(Z) = \arg\max_{q} \mathcal{L}(q, \theta_n) = p(Z \mid X, \theta_n) \tag{143}$$

M-step

$$\theta_{n+1} = \arg\max_{\theta} \mathcal{L}(q_n, \theta) = \arg\max_{\theta} \mathbb{E}_{q_n(Z)} \log p(X, Z \mid \theta) \tag{144}$$

В последнем равенстве мы воспользовались определением вариационной нижней оценки:

$$\mathcal{L}(q, \theta) = \int q(Z) \log \frac{p(X, Z \mid \theta)}{q(Z)} dZ = \int q(Z) \log p(X, Z \mid \theta) dZ - \int q(Z) \log q(Z) dZ. \tag{145}$$

Второе слагаемое можно отбросить, потому что энтропия q не зависит от θ.

Преимущества такой процедуры:

1. На E-шаге можем выполнить пересчёт в явном виде (если умеем считать апостериорное распределение на Z).

2. На M-шаге возникает задача оптимизации $\mathbb{E}_{q_n(Z)} \log p(X, Z \mid \theta)$ — вогнутой функции по θ, так как $\log p(X, Z \mid \theta)$ вогнута, а матожидание, как выпуклая комбинация выпуклых функций, тоже является вогнутой функцией от θ

3. Итак, задача максимизации вогнутой функции. Если повезёт, то можно решить в явном виде. Если нет, то её можно хотя бы эффективно решать.

7.2.2 Модификация EM-алгоритма: априорное распределение на веса (EM' алгоритм)

Рассмотрим пример небольшой модификации EM-алгоритма, который пригодится нам в дальнейшем. Предположим, что вероятностная модель полностью задана, то есть известно совместное распределение на X, Z, θ:

$$p(X, Z, \theta) = p(X, Z \mid \theta)p(\theta). \tag{146}$$

Пусть теперь мы хотим найти максимум не у оценки максимального неполного правдоподобия (140), а у апостериорного распределения:

$$\theta_{MP} = \arg\max_{\theta} p(\theta \mid X) = \arg\max_{\theta} \log p(\theta \mid X) = \arg\max_{\theta} \big[\log p(X \mid \theta) + \log p(\theta)\big]. \tag{147}$$

В выкладках выше мы применили теорему Байеса, знаменатель не зависит от θ, поэтому максимум апостериорной плотности эквивалентен максимуму числителя. В числителе — логарифм неполного правдоподобия плюс логарифм априорного распределения.

Как изменится EM-алгоритм при такой постановке задачи?

Выражение (142) примет вид:

$$\log p(X \mid \theta) + \log p(\theta) = \mathcal{L}(q, \theta) + KL(q(Z) \parallel p(Z \mid X, \theta)) + \log p(\theta). \tag{148}$$

На E-шаге мы максимизируем по q при фиксированном θ. Добавленное слагаемое не зависит от q, поэтому E-шаг (143) не изменится.

$$q(Z) = \arg\min_{q} KL(q(Z))||p(Z \mid X, \Theta) = p(Z \mid X, \Theta) \tag{149}$$

На M-шаге (144) возникнет ещё одно аддитивное слагаемое:

$$\theta_{n+1} = \arg\max_{\theta} \mathcal{L}(q_n, \theta) + \log p(\theta) = \arg\max_{\theta} \big[\mathbb{E}_{q_n(Z)} \log p(X, Z \mid \theta) + \log p(\theta)\big]. \tag{150}$$

Таким образом, EM-алгоритм практически не меняется при замене оценки максимума правдоподобия на максимизацию апостериорного распределения. Это наблюдение пригодится нам в дальнейшем.

7.2.3 От EM-алгоритма к вариационному выводу

Что будет, если на E-шаге распределения не сопрягаются и мы не можем точно выполнить Байесовский вывод? Придётся выполнять его приближённо. Заметим, что E-шаг (143) эквивалентен минимизации KL-дивергенции:

$$q_n(Z) = \arg\max_{q} \mathcal{L}(q, \theta_n) = \arg\min_{q} KL(q(Z) \parallel p(Z \mid X, \theta_n)). \tag{151}$$

Проблема: для минимизации KL-дивергенции (151) мы должны уметь её считать, но мы не знаем $p(Z \mid X, \theta_n)$. Тем не менее, эту задачу можно решить приближённо. Будем для простоты минимизировать KL-дивергенцию не по всевозможным распределениям q, а по распределениям q из какого-то ограниченного семейства (например, из параметрического или функционального) — то есть будем искать вариационную аппроксимацию истинного апостериорного распределения.

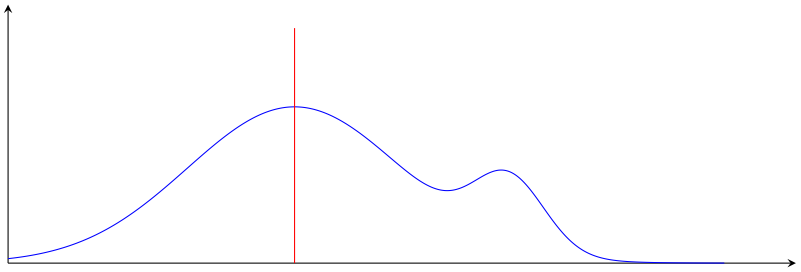


Рис. 11: Пример аппроксимации распределения в семействе дельта-функций.

Пример. Что будет, если мы ограничим семейство распределений q, к примеру, множеством дельта-функций? То есть захотим аппроксимировать $p(Z \mid X, \theta_n)$ в классе дельта-функций, минимизируя KL-дивергенцию между аппроксимацией и исходным распределением. «Байес для бедных»:

$$\begin{aligned}
&\arg\min_{q\in\Delta} KL(q(Z)||p(Z \mid X, \Theta)) \\
&= \arg\min_{q\in\Delta} \int q(Z) \log \frac{q(z)}{p(Z \mid X, \Theta)} dZ \\
&= \{q(z) = \delta(Z - Z_0)\} \\
&= \arg\min_{Z_0} \int \delta(Z - Z_0)[\log \delta(Z - Z_0) - \log p(Z \mid X, \Theta)] dZ \\
&= \arg\min_{Z_0} [Const - \int \delta(Z - Z_0) \log p(Z \mid X, \Theta) dZ] \\
&= \arg\max_{Z_0} \log p(Z_0 \mid X, \Theta)
\end{aligned}$$

Ранее мы уже выяснили, что для этого нужно взять точку в моде этого распределения (рис. 11). С точки зрения KL-дивергенции это самая репрезентативная точка. Если рассматривать другие дивергенции, ответ может поменяться.

7.3 Вариационный Байесовский вывод: mean-field аппроксимация

Mean-field аппроксимация (теория среднего поля) была разработана физиками для решения задач теории поля. Является частным случаем более общего подхода, который носит название вариационный Байесовский вывод (так сказать, «Байес для среднего класса»).

Пусть у нас есть сложное апостериорное распределение, которое мы бы хотели приблизить каким-то распределением, для которого знаем (умеем считать) нормировочную константу. Мы не хотим применять Байес для бедных, так как при этом теряется существенное количество информации, а значит, и ухудшается качество.

Пусть модель состоит из наблюдаемых и латентных переменных:

$$p(X, Z). \tag{152}$$

При этом для апостериорного распределения $p(Z \mid X)$ мы можем посчитать числитель в формуле Байеса, а знаменатель — нет (интеграл не берется). Давайте попробуем приблизить $p(Z \mid X)$ распределением $q(Z)$ из некоторого ограниченного семейства распределений, для которого мы знаем, как считать нормировочные константы. Приближаем, минимизируя KL-дивергенцию:

$$q(Z) = \arg\min_{q\in Q} KL(q(Z) \parallel p(Z \mid X)). \tag{153}$$

Для простоты здесь мы не предполагаем зависимости от дополнительных параметров θ, но и на этот случай все текущие рассуждения тривиально обобщаются.

Какое семейство Q нам взять? Обычно ограничиваются параметрическим семейством, например, классом нормальных распределений. Однако давайте рассмотрим не параметрическое, а функциональное mean-field ограничение. Разобьем множество переменных Z на

непересекающиеся подмножества (факторизируем) и будем рассматривать лишь факторизованные распределения q:

$$Z = \bigsqcup_{i=1}^{l} z_i; \quad z_i \cap z_j = \varnothing; \quad q(Z) = \prod_{i=1}^{l} q_i(z_i). \tag{154}$$

В связи с тем, что мы ввели ограничение на множество рассматриваемых распределений, KL-дивергенцию, как правило, мы уже не сможем сделать нулевой. Как уже упоминалось, KL-дивергенция зависит от апостериорного распределения, которое мы не умеем считать. Заменим (153) на эквивалентную задачу максимизации вариационной нижней оценки:

$$q(Z) = \arg\min_{q \in Q} KL(q(Z) \parallel p(Z \mid X)) = \arg\max_{q \in Q} \int q(Z) \log \frac{p(X, Z)}{q(Z)} dZ. \tag{155}$$

Апостериорное распределение здесь нигде не фигурирует, и мы можем посчитать все составляющие интеграла. Будем решать задачу блочно-координатно: зафиксируем все группы латентных переменных z_i, кроме одной — z_j, для которой в явном виде получим уравнения для обновления.

Подставим в правую часть (155) факторизацию (154):

$$\int \prod_{i=1}^{l} q_i(z_i) \log \frac{p(X, Z)}{\prod_{i=1}^{l} q_i(z_i)} \prod_{i=1}^{l} dz_i = \int \prod_{i=1}^{l} q_i(z_i) \log p(X, Z) dZ - \int \prod_{i=1}^{l} q_i(z_i) \Big[\sum_{k=1}^{l} \log q_k(z_k) \Big] dZ \tag{156}$$

Во втором слагаемом вынесем сумму по k за знак интеграла (матожидание суммы равно сумме матожиданий). Получили сумму матожиданий, в которой для каждого матожидания подынтегральная функция зависит только от одной z_k, то есть по всем $i \neq k$ мы получим интеграл по плотности, т.е. 1:

$$= \int \prod_{i=1}^{l} q_i(z_i) \log p(X, Z) dZ - \sum_{k=1}^{l} \int q_k(z_k) \log q_k(z_k) dz_k = \tag{157}$$

Фиксируем все z_i, кроме z_j. Распишем выражение как функцию от q_j. В первом слагаемом вынесем её наружу. Во втором — от z_j зависит только 1 член, остальные выносим в константу:

$$= \int q_j(z_j) \left(\int \prod_{i \neq j} q_i(z_i) \log p(X, Z) dZ_{\neq j} \right) dz_j - \int q_j(z_j) \log q_j(z_j) dz_j + Const. \tag{158}$$

Посмотрим на выражение $\int \prod_{i \neq j} q_i(z_i) \log p(X, Z) dZ_{\neq j}$. Обозначим

$$\hat{p}(z_j) \equiv \exp \left(\int \prod_{i \neq j} q_i(z_i) \log p(X, Z) dZ_{\neq j} \right). \tag{159}$$

То есть исходное выражение — это логарифм ненормированной плотности $\hat{p}(z_j)$:

$$\int \prod_{i \neq j} q_i(z_i) \log p(X, Z) dZ_{\neq j} = \log \hat{p}(z_j) \tag{160}$$

$$p(z_j) = \frac{\hat{p}(z_j)}{\int \hat{p}(z_j) dz_j} \equiv \frac{\hat{p}(z_j)}{A}; \quad \hat{p}(z_j) = A \cdot p(z_j) \tag{161}$$

После перенормировки (A — нормировочная константа) $p(z_j)$ можно рассматривать как плотность вероятности. Подставим её в (158) и объединим интегралы, при этом составляющая интеграла с константой A будет вынесена в новую константу:

$$\int q(Z) \log \frac{p(X, Z)}{q(Z)} dZ = \ldots = \int q_j(z_j) \log(A p(z_j)) dz_j - \int q_j(z_j) \log q_j(z_j) dz_j + Const$$
$$= \int q_j(z_j) \log \frac{p(z_j)}{q_j(z_j)} dz_j + Const'. \tag{162}$$

Напомним, что в соответствии с (155) мы хотим максимизировать это выражение по q_j. Заметим, что если поменять числитель и знаменатель под логарифмом местами, то получим KL-дивергенцию:

$$\int q(Z) \log \frac{p(X,Z)}{q(Z)} dZ = \ldots = -KL(q_j(z_j) \parallel p(z_j)) + Const'. \tag{163}$$

Наша задача максимизация по q_j эквивалентна минимизации $KL(q_j(z_j) \parallel p(z_j))$. Решение — положить $q_j(z_j) = p(z_j)$. Подставим выражение для $p(z_j)$ (159) с учётом нормировки и получим финальное выражение для обновления $q_j(z_j)$:

$$q_j(z_j) = \frac{\exp\big(\mathbb{E}_{q(Z_{\neq j})} \log p(X,Z)\big)}{\int \exp\big(\mathbb{E}_{q(Z_{\neq j})} \log p(X,Z)\big) dz_j} \tag{164}$$

Обычно эту формулу применяют в более удобном виде. Возьмем логарифм от обоих частей:

Основная формула mean-field аппроксимации:

$$\log q_j(z_j) = \mathbb{E}_{q(Z_{\neq j})} \log p(X,Z) + Const. \tag{165}$$

В таком виде, вообще говоря, формула не выглядит очень конструктивной по двум причинам:

1. Нужно считать матожидание, неизвестно, берется ли оно аналитически.
2. Откуда брать константу?

Однако существуют условия, при которых мы гарантированно можем аналитически рассчитать и матожидание, и константу — об этом далее.

Итак, как работает mean-field? Стартуем с некоторого начального приближения $q = \prod_i q_i$ и начинаем итеративно обновлять его компоненты: обновили q_1, зафиксировали, обновили q_2, зафиксировали, …, обновили q_l. И так по кругу: следующая итерация — обновляем q_1, фиксируем, обновляем q_2 и так до сходимости. Процесс гарантированно сходится, потому что на каждой итерации мы увеличиваем вариационную нижнюю оценку (каждый раз обнуляем KL-дивергенцию в (163)). Поэтому процесс монотонный и гарантированно сходится из любого начального приближения, но, вообще говоря, к разным локальным экстремумам.

Заметим, что если $l = 1$, то есть мы не разбиваем Z, то из mean-field получается обычный Байесовский вывод. То есть в каком-то смысле mean-field можно рассматривать как обобщение Байесовского вывода.

7.3.1 Условная сопряженность (conditional conjugate).

Определение 8. Условная сопряженность есть соблюдение двух условий:

- $p(X, Z)$ — лежит в экспоненциальном классе
- $\forall z_j$ $p(X \mid z_j, Z_{\neq j})$ и $p(z_j \mid Z_{\neq j})$ сопрягаются (относительно z_j).

То есть если мы зафиксируем все z_i кроме z_j, то получающиеся априорное распределение $p(z_j \mid Z_{\neq j})$ и функция правдоподобия $p(X \mid z_j, Z_{\neq j})$ сопрягаются друг с другом. Иными словами, мы можем совершить аналитический Байесовский вывод на $p(z_j \mid X, Z_{\neq j})$.

Видно, что условная сопряженность является обобщением понятия обычной сопряженности. В обычной сопряженности никакие латентные переменные фиксировать не приходилось — мы получали апостериорное распределение сразу на все переменные. На практике, однако, часто оказывается, что полного сопряжения нет, но есть условное, то есть множество неизвестных переменных можно разбить на непересекающиеся группы так, что для каждой отдельной группы (при фиксированных остальных) есть сопряженность.

Теорема 5. Если есть условная сопряженность, то можно получить аналитические формулы для итеративного пересчета всех q на основе уравнений (164, 165).

Примем без доказательства, но далее рассмотрим на конкретных примерах.

7.3.2 Связь mean-field аппроксимации и EM'-алгоритма

Покажем, что EM' алгоритм есть частный случай MF-аппроксимации. Рассмотрим ограничение $p(Z, \theta \mid X) \approx q(Z)q(\theta)$, где $q(\theta) = \delta(\theta - \theta_{\text{MP}})$.

Применяем основную формулу MF для подсчёта $q(Z)$:

$$\begin{aligned}\log q(Z) &= \mathbb{E}_{q(\theta)} \log p(X, Z, \theta) + \text{Const} \\ &= \{\text{матожидание по распределению с дельта-функцией}\} \\ &= \log p(X, Z, \theta_{MP}) + \text{Const}\end{aligned}$$

Далее «делаем софтмакс», чтобы найти нормированное распределение:

$$q(Z) = \frac{p(X, Z, \theta_{MP})}{\int p(X, Z, \theta_{MP})\, dZ} = \frac{p(X, Z, \theta_{MP})}{p(X, \theta_{MP})} = p(Z \mid X, \theta_{MP}) \tag{166}$$

Получили апостериорное распределение на латентные переменные, прямо как на E-шаге.

Применяем основную формулу MF для подсчёта $q(\theta)$:

$$\begin{aligned}\log q(\theta) &= \mathbb{E}_{q(Z)} \log p(X, Z, \theta) + \text{Const} \\ &= \mathbb{E}_{q(Z)} \log p(X, Z \mid \theta) + \log p(\theta) + \text{Const} \\ &= \mathcal{L}(q, \theta) + \log p(\theta) + \text{Const},\end{aligned}$$

где $\mathcal{L}(q, \theta)$ — это ELBO. По свойству дельта-функции: $\theta_{\text{MP}} = \arg\max_\theta \log q(\theta)$. Получили в точности M'-шаг, т.к. результатом M'-шага является MAP-оценка на параметры θ.

7.3.3 Связь mean-field аппроксимации и EM-алгоритма

Посмотрим на формулу для mean-field аппроксимации и попробуем в ней увидеть EM-алгоритм. Разобьем Z на два непересекающихся подмножества:

$$Z = z_1 \sqcup z_2. \tag{167}$$

Аппроксимируем апостериорное распределение факторизованным и будем q_2 рассматривать только из семейства дельта-функций:

$$p(Z \mid X) \approx q_1(z_1)\delta(z_2 - z_2^*). \tag{168}$$

Тогда формула для пересчета q_1 (на основе (165)):

$$\log q_1(z_1) = \mathbb{E}_{z_2} \log p(X, Z) + Const = \log p(X, z_1, z_2^*) + Const \tag{169}$$

Возьмем экспоненту от обеих частей, при этом z_2 и X фиксированы, а значит, получаем перенормированное апостериорное распределение на z_1 при данных $z_2 = z_2^*$ и X:

$$q_1(z_1) = \frac{p(X, z_1, z_2^*)}{\int p(X, z_1, z_2^*) dz_1} = p(z_1 \mid X, z_2^*) \tag{170}$$

Мы получили E-шаг EM-алгоритма.

Далее мы оптимизируем по q_2, но заметим, что это не произвольное распределение, а дельта-функция. Поэтому, оптимизируя KL-дивергенцию, мы должны поставить дельта-функцию в точку максимума:

$$\mathbb{E}_{z_1} \log p(X, Z)) \to \max_{z_2}. \tag{171}$$

Тогда получим:

$$z_2^* = \arg\max_{z_2} \mathbb{E}_{z_1} \log p(X, z_1 \mid z_2)) + \log p(z_2) \tag{172}$$

Если мысленно подставить θ вместо z_2, а вместо z_1 — старый Z, то становится очевидно, что мы получили M-шаг EM-алгоритма. То есть mean-field аппроксимация является также обобщением EM-алгоритма (с условием принадлежности q_2 к семейству дельта-функций).

Пример. Переходим к практике. Классический пример применения EM-алгоритма — разделение смеси гауссиан. В этом случае и E-шаг и M-шаг можно выполнить аналитически.

Введем вероятностную модель:

- X — наблюдаемые переменные, Z — индексы компонент смеси
- π — априорная вероятность компоненты смеси
- μ — мат.ожидания каждой из K гауссиан
- Λ — обратные ковариационные матрицы каждой из K гауссиан
- $p(X, Z, \pi, \mu, \Lambda)$ — вероятностная модель

π — вектор размера K, сумма всех компонент равна 1, все компоненты неотрицательны. Значит априорное распределение можем взять в виде распределения Дирихле:

$$Dir(\pi \mid \alpha) = \frac{1}{C(\alpha)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1},\ \alpha_k > 0. \tag{173}$$

Зависимость формы плотности вероятности распределения Дирихле от α:

- $\alpha_k > 1$ — плотность распределения имеет форму «колокола». Изменением соотношения между разными компонентами α можно изменять форму получаемого «колокола».
- $\alpha_k = 1$ — равномерная плотность на вероятностном симплексе.
- $\alpha_k < 1$ — U-образная форма плотности. Значения внутри симплекса — почти нулевые, на гранях значения выше, максимальные значения — в углах. Благодаря этому свойству, распределение Дирихле с $\alpha_k < 1$ удобно брать в качестве априорного, если важно занулить большинство компонент смеси.

Введем распределение Дирихле на π, с одной $\alpha_0 = 10^{-3}$ на все компоненты. Распишем вероятностную модель $p(X, Z, \pi, \mu, \Lambda)$:

$$p(X, Z, \pi, \mu, \Lambda) = p(X \mid Z, \mu, \Lambda) p(Z \mid \pi) p(\mu, \Lambda) p(\pi) = \tag{174}$$

$$= \prod_{i=1}^{n} p(x_i \mid z_i, \mu, \Lambda) p(z_i \mid \pi) \prod_{k=1}^{K} p(\mu_k, \Lambda_k) p(\pi) = \tag{175}$$

$$= \prod_{i=1}^{n} \left[\mathcal{N}(x_i \mid \mu_{z_i}, \Lambda_{z_i}^{-1}) \prod_{k=1}^{K} \pi_k^{[z_i = k]} \right] \prod_{k=1}^{K} p(\mu_k, \Lambda_k) Dir(\pi \mid \alpha_0). \tag{176}$$

Проблемы с базовым EM-алгоритмом для разделения смеси гауссиан:

1. Не позволяет автоматически определять количество гауссиан в смеси. В рассматриваемой модели можно задать число компонент K избыточным, тогда априорное распределение будет поощрять зануление значительной их части.
2. Бесконечно большое правдоподобие достигается, когда одна гауссиана бесконечно узкая (гауссиана покроет одну точку). Значение плотности в этой точке при этом будет бесконечным, а значит, и правдоподобие будет бесконечно большим. Если в рассматриваемой модели ввести априорное распределение на дисперсию, она будет лишена этих проблем.

Осталось ввести распределение на μ, Λ. Выберем распределение таким, чтобы оно сопрягалось с многомерным нормальным распределением (это будет Уишарт-нормальное распределение):

$$p(\mu_k, \Lambda_k) = p(\mu_k \mid \Lambda_k) p(\Lambda_k) = \mathcal{N}(\mu_k, \mid m_0, (\beta \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k \mid \nu, W), \tag{177}$$

где $\mathcal{W}(\Lambda_k \,|\, \nu, W)$ — распределение Уишарта, многомерное обобщение гамма-распределения:

$$\mathcal{W}(\Lambda_k \,|\, \nu, W) = \frac{1}{C(\nu, W)} \big(\det \Lambda_k\big)^{\frac{\nu-d-1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left(W^{-1}\Lambda_k\right)\right),\ \nu > d-1,\ W = W^T \succ 0. \tag{178}$$

где $C(\nu, W)$ – нормировочная константа. Матожидание распределения Уишарта:

$$\mathbb{E}\,\Lambda_k = \nu W \tag{179}$$

Чем больше ν, тем «уже» распределение — меньше отклонение от матожидания.

Пример. Применение распределение Уишарта на практике.

Трекинг мышей: мышки бегают в тазике с опилками, сверху их снимает камера. Качество съемки не очень хорошее, каждая мышка — серый комочек. По отдельности трекать мышек еще получается, но когда мышки сбиваются в группы — уже нет.

Первоначальная модель: отдельных мышек параметризовывали эллипсом. Упрощая, можно сказать, что вписывали гауссиану. Когда мышки собирались вместе получалась смесь гауссиан, которую можно было разделить классическим методом. Но в этом случае получались неправдоподобные размеры каждой из мыши.

Улучшение модели: задав распределение на размер мыши с помощью распределения Уишарта, а матожидание характерным размером мыши (ν подобрали за несколько итераций), удалось качественно разделять мышей даже в «слипшихся» группках.

Теперь подставим $p(\mu_k, \Lambda_k)$ в вероятностную модель (174):

$$\prod_{i=1}^{n}\left[\mathcal{N}(x_i \,|\, \mu_{z_i}, \Lambda_{z_i}^{-1}) \prod_{k=1}^{K} \pi_k^{[z_i=k]}\right] \prod_{k=1}^{K} \mathcal{N}(\mu_k, \,|\, m_0, (\beta\Lambda_k)^{-1}) \mathcal{W}(\Lambda_k \,|\, \nu, W) Dir(\pi \,|\, \alpha_0). \tag{180}$$

Модель полностью Байесовская, а значит, мы хотели бы сделать Байесовский вывод на все переменные (μ, Λ, Z, π), то есть получить апостериорное распределение:

$$p(Z, \pi, \mu, \Lambda \,|\, X). \tag{181}$$

Проверим, есть ли сопряженность на все переменные, т.е. сопряжены ли $p(X \,|\, Z, \pi, \mu, \Lambda)$ и $p(Z, \pi, \mu, \Lambda)$. Можно показать, что полного сопряжения не будет. Но есть условное сопряжение! Покажем, что условное сопряжение имеется на Z и (μ, Λ, π), а именно, если зафиксируем (μ, Λ, π), то будет сопряженность на Z, и наоборот. Действительно, рассмотрим $p(Z, \pi, \mu, \Lambda)$.

$$p(Z, \pi, \mu, \Lambda) = p(Z \,|\, \pi) p(\pi) p(\mu, \Lambda) \propto \Big(\prod_{i=1}^{N}\prod_{k=1}^{K} \pi_k^{[z_i=k]}\Big)\Big(\prod_{k=1}^{K} \pi_k^{\alpha_k - 1}\Big) \times$$
$$\times \Big(\prod_{k=1}^{K} (\det \beta\Lambda_k)^{\frac{1}{2}} \exp\big(-\frac{1}{2}(m_0 - \mu_k)^T \beta\Lambda_k (m_0 - \mu_k)\big)(\det \Lambda_k)^{\frac{\nu-d-1}{2}} \exp\big(-\frac{1}{2}\operatorname{tr}\big(W^{-1}\Lambda_k\big)\big)\Big) \tag{182}$$

Теперь распишем правдоподобие в модели (180):

$$p(X \,|\, Z, \pi, \mu, \Lambda) \ \propto\ \Big(\prod_{i=1}^{N}\prod_{k=1}^{K}\Big[(\det \Lambda_k)^{\frac{1}{2}} \exp\big(-\frac{1}{2}(x_i - \mu_k)^T \Lambda_k (x_i - \mu_k)\big)\pi_k\Big]^{[z_i=k]}\Big) \tag{183}$$

Если сравнить выражения 183 и 182 то можно заметить, что они имеют разный функциональный вид относительно $z_i, \mu_k, \Lambda_k, \pi_k$. Действительно, в правдоподобии 183 есть, например, члены вида $\Big((\det \Lambda_k)^{\frac{1}{2}}\Big)^{[z_i=k]}$, аналогов которым нет в априорном распределении 182 [27]. При этом, если мы зафиксируем Z и посмотрим на выражения 183 и 182 как на

[27] Проблема в том, что z_i – не константа, а переменная, поэтому мы не можем сказать, что $(\det \Lambda_k)^{\frac{[z_i=k]}{2}}$ – это $(\det \Lambda_k)^{\alpha}$, где α – какая-то константа

функции от μ_k, Λ_k, π_k, то увидим, что они имеют один и тот же функциональный вид. Действительно:

Относительно π:

- Априорное распределение: $\propto \prod_{k=1}^{K} \pi_k^{\alpha_0 - 1}$
- Правдоподобие: $\propto \prod_{k=1}^{K} \pi_k^{[z_i = k]}$

Они принадлежат одному функциональному классу, следовательно, на π будет сопряжение (при фиксированном Z).

Если зафиксировали все Z, то что можно сказать про (μ, Λ)?

- Априорное распределение:

$$\propto \prod_{k=1}^{K} (\det \beta\Lambda_k)^{\frac{1}{2}} \exp\big(-\frac{1}{2}(m_0 - \mu_k)^T \beta\Lambda_k (m_0 - \mu_k)\big) (\det \Lambda_k)^{\frac{\nu - d - 1}{2}} \exp\big(-\frac{1}{2}\operatorname{tr}\big(W^{-1}\Lambda_k\big)\big) \tag{184}$$

- Правдоподобие:

$$\propto \prod_{i=1}^{N} \prod_{k=1}^{K} \Big[(\det \Lambda_k)^{\frac{1}{2}} \exp\big(-\frac{1}{2}(x_i - \mu_k)^T \Lambda_k (x_i - \mu_k)\big)\Big]^{[z_i = k]} \tag{185}$$

Выражения 184 и 185 относительно Λ_k и μ_k имеют одинаковый функциональный вид. Оба выражения можно представить в виде $\prod_k (\det \Lambda_k)^{\gamma_{1,k}} \exp(t(\mu_k, \Lambda_k))^{\gamma_{2,k}}$, где $t(\mu_k, \Lambda_k) = \zeta_{1,k} \operatorname{tr}(A_k \Lambda_k) + \zeta_{2,k} \operatorname{tr}\big((b_k \mu_k^T + \mu_k b_k^T)\Lambda_k\big) + \zeta_{3,k} \operatorname{tr}\big(\mu_k \mu_k^T \Lambda_k\big)$, с параметрами $\gamma_{i,k}$ и $\zeta_{i,k}$ (скаляры), A_k (матрица), b_k (вектор), не зависящими от μ_k, Λ_k.

Таким образом, правдоподобие $p(X \mid \mu, \Lambda, \pi, Z)$ и априорное распределение $p(\mu, \Lambda, \pi \mid Z)$ при фиксированном Z сопряжены[28].

Теперь рассмотрим правдоподобие и априорное распределение как функции от Z при фиксированных (μ, Λ, π):

- Априорное распределение: $\propto \prod_{k=1}^{K} \pi_k^{[z_i = k]}$
- Правдоподобие: $\propto \prod_{k=1}^{K} \mathcal{N}(x_i \mid \mu_k, \Lambda_k^{-1})^{[z_i = k]}$

Оба распределения имеют вид произведения каких-то элементов в степени индикатора от z, то есть на z тоже есть сопряженность (при фиксированных (μ, Λ, π)).

Таким образом, у нас есть условная сопряженность на параметры (μ, Λ, π) и Z, т.е. мы можем эффективно приблизить апостериорное распределение (181) mean-field аппрокимацией:

$$p(Z, \pi, \mu, \Lambda \mid X) \approx q_1(Z) q_2(\mu, \Lambda, \pi). \tag{186}$$

Далее необходимо просто применить формулу (165): расписать логарифм правдоподобия и проматожидать по Z, чтобы найти q_2, и по (μ, Λ, π), чтобы найти q_1.

7.4 Концептуальная схема

Рассмотрим вероятностную модель, где X — наблюдаемые переменные, Z — скрытые или латентные переменные, а θ — параметры модели:

$$p(X, Z, \theta) = p(X, Z \mid \theta) p(\theta). \tag{187}$$

Разберем, какой метод вывода необходимо применять в зависимости от свойств, которыми обладает вероятностная модель. Чем ниже опускаемся по таблице, тем меньше требований накладываем на свойства вероятностной модели, а следовательно, тем шире область применимости.

Пояснения к таблице:

[28] Внимательный читатель может заметить, что мы как будто бы рассматривали распределения $p(X \mid Z, \pi, \mu, \Lambda)$ и $p(Z, \pi, \mu, \Lambda)$, а в итоге делаем вывод о сопряженности $p(X \mid \mu, \Lambda, \pi, Z)$ и $p(\mu, \Lambda, \pi \mid Z)$. Здесь нет ошибки, поскольку $p(\mu, \Lambda, \pi \mid Z) \propto p(Z, \pi, \mu, \Lambda)$ как функция от μ, Λ, π

Свойства	Метод вывода	Вид аппроксимации/Ограничения
(Z, θ) — полная сопряженность (conjugacy)	Байесовский вывод (Bayes Theorem), «Байес для богатых»	$p(Z, \theta \mid X)$
Условная сопряженность (Conditional conjugacy) Z & θ	Mean-Field	$q(Z)q(\theta)$
Z — сопряженность (conjugacy) при фикс. θ	EM' алгоритм для оптимизации $p(\theta \mid X)$	$\delta(\theta - \theta_{MP})p(Z \mid X, \theta_{MP})$
θ — сопряженность (conjugacy) при фикс. Z	M'E для оптимизации $p(Z \mid X)$	$\delta(Z - Z_{MP})p(\theta \mid X, Z_{MP})$
Условная сопряженность (conditional conjugacy) по z_j при фикс. θ	Mean-Field EM' (variational EM')	$\delta(\theta - \theta_{MP}) \prod_{i=1}^{l} q_i(z_i)$
Нет сопряженности (No conjugacy)	Poor man's Bayes	$\delta(\theta - \theta_{MP})\delta(Z - Z_{MP})$

Таблица 5: Методы Байесовского вывода.

1. Есть полная сопряженность на (Z, θ). Если расписать $p(X, Z, \theta) = p(X \mid Z, \theta)p(Z, \theta)$, то у нас имеется полная сопряженность между функцией правдоподобия того, что мы наблюдаем, $p(X \mid Z, \theta)$ и априорным распределением $p(Z, \theta)$ на то, что мы не наблюдаем. Есть полная сопряженность, значит, можем получить апостериорное распределение $p(Z, \theta \mid X)$ с помощью теоремы Байеса.

2. Условная сопряженность на Z и θ. Фиксируем θ, оказывается, что есть сопряжение на Z, и наоборот. Следовательно, применяем mean-field аппроксимацию. Получим приближение на чистое апостериорное в виде $q(Z)q(\theta) = \arg\min KL(q(Z)q(\theta) \parallel p(Z, \theta \mid X))$.

   Обратим внимание, что в первых двух строчках (с математической точки зрения) пропадают различия между латентными переменными Z и параметрами θ.

3. Сопряжение только на Z. Для θ остается сделать только Байес для бедных: $\delta(\theta - \theta_{MP})$. Это и есть EM'-алгоритм: на E-шаге мы честно оцениваем плотность для Z, на M'-шаге находим θ_{MP}.

4. Сопряжение только на θ. Исходя из симметрии Z и θ, это будет M'E-алгоритм: E-шаг на θ, M'-шаг на Z. Обратите внимание, что здесь у нас явным образом стирается грань между параметрами и скрытыми переменными: по-сути это одно и тоже, а различия в вычислении соответствующих распределений обусловлены наличием сопряженности.

5. Условная z_j сопряженность. Мы можем разбить Z на непересекающиеся группы так, что для каждой отдельной группы z_j при фиксировании остальных $z_{i \neq j}$ и θ будет сопряженность. На θ при этом нет сопряженности — её аппроксимируем δ-функцией (ищем максимимум апостериорной плотности). На Z — mean-field аппроксимация. Таким образом, мы по сути применяем EM'-алгоритм, где на M'-шаге пересчитываем точку максимума по θ, а на E-шаге вместо честного Байесовского вывода применяем mean-field аппроксимацию. В литературе эту версию обычно называют EM'-алгоритм.

   Аналогично можем расписать условную θ-сопряженность и получить вариационный M'E-алгоритм.

6. Нет никакой сопряженности. Приближаем все распределения δ-функцией (Байес для бедных).

8 Лекция 8. Методы Монте-Карло с Марковскими цепями (MCMC)

8.1 Общие предпосылки метода Монте-Карло

Часто в задачах машинного обучения возникает задача оценивания математического ожидания функции $f(x)$ по распределению $p(x)$: $\mathbb{E}_{p(x)}f(x) = \int f(x)p(x)dx$. Где x, чаще всего, является вектором высокой размерности. Примерами подсчёта такого интеграла является нахождение следующих распределений:

1. Обоснованность: $p(T \mid X) = \int p(T \mid X, \omega)p(\omega)d\omega$, где (X, T) – обучающая выборка, ω – параметры распределения.
2. Прогнозное распределение: $p(t \mid x, X, T) = \int p(t \mid x, \omega)p(\omega \mid X, T)d\omega$, где x – объект, для которого мы хотим получить прогнозное распределение на предсказание t.

На практике этот интеграл чаще всего не получается посчитать аналитически, поэтому применяют численные методы. Однако сложность классических численных методов интегрирования возрастают экспоненциально при увеличении размерности пространства.

В предыдущих лекциях мы уже имели дело с неберущимися интегралами, которые мы приближали с помощью EM-алгоритма и/или Mean Field аппроксимации. Однако, в этих случаях приближенная оценка получалась смещенной. Можно ли как-то решить задачу интегрирования в высокоразмерном пространстве с использованием несмещённой оценки? В этом нам поможет метод Монте-Карло:

Заметим, что мы имеем дело не с произвольным интегралом, а с математическим ожиданием функции $f(x)$ по распределению $p(x)$. Приблизим его, произведя сэмплирование из распределения $p(x)$:

$$I = \int f(x)p(x)dx \approx \frac{1}{K}\Sigma_{k=1}^{K} f(x_k) = J, \quad x_k \sim p(x) \tag{188}$$

Поскольку все x_i случайные величины, то и J – случайная величина. Посчитаем её математическое ожидание:

$$\mathbb{E}J = \mathbb{E}\frac{1}{K}\sum_i f(x_i) = \frac{1}{K}\sum_i \mathbb{E}f(x_i) = \frac{K}{K}\,\mathbb{E}\,f(x) = \mathbb{E}f(x) = I, \tag{189}$$

Следовательно, J является несмещенной оценкой интеграла $\mathbb{E}_{p(x)}f(x)$. Теперь найдем дисперсию J:

$$\mathbb{D}J = \mathbb{D}\frac{1}{K}\sum_i f(x_i) = \frac{1}{K^2}\sum_i \mathbb{D}f(x_i) = \frac{1}{K}\mathbb{D}f(x) = \tag{190}$$

$$= \frac{1}{K}\int p(x)(f(x) - \mathbb{E}f(x))^2 dx = \frac{1}{K}\int p(x)(f(x) - I)^2 dx \tag{191}$$

И, как следствие из ЦПТ, $J \sim \mathcal{N}(J \mid I, \frac{1}{K}\mathbb{D}f(x))$. Следовательно, чем меньше дисперсия оценки, тем ближе J приближает исходное $\mathbb{E}_{p(x)}f(x)$. Этого можно достичь путём увеличения количества сэмплов.

Также стоит отметить, что дисперсия $\mathbb{D}f(x)$ зависит от поведения функции $f(x)$ и не зависит от размерности пространства. Т.е. если $f(x)$ меняется плавно в области носителя, то можно не брать много точек для приближения [29]. Если же $f(x)$ сильно флуктуирует, то для точной оценки понадобится много сэмплов.

Таким образом, задача подсчёта интеграла сводится к задаче генерации выборки из заданного распределения. Эта задача не такая тривиальная, как может показаться на первый взгляд и для ее решения нам как раз и пригодятся марковские цепи. Но сперва мы рассмотрим классические методы генерации сэмплов из распределений.

[29] Предельный случай - $f(x) \equiv const \Rightarrow \mathbb{D}f(x) = 0$, для приближения можно взять одну точку

8.2 Общие методы генерации выборок из одномерных распределений

8.2.1 Простейшие методы

В большинство популярных языков программирования заложена возможность генерации случайной величины, равномерно распределённой на $[0,1]$, этому будем отталкиваться от того, что это мы делать умеем. Пусть $\xi \sim \mathcal{U}[0,1]$; $\mathbb{E}\xi = \frac{1}{2}$; $\mathbb{D}\xi = \frac{1}{12}$

Если же мы хотим получить случайную величину из произвольного отрезка $\eta \sim \mathcal{U}[a,b]$, то можно сначала просэмлировать $\xi \sim \mathcal{U}[0,1]$, а потом линейно ее преобразовать: $\eta = \xi(b-a)+a$

Теперь рассмотрим сэмплирование из стандартного нормального распределения. Для этого можно сгенерировать n случайных величин из $\mathcal{U}[0,1]$ и воспользоваться ЦПТ:

$$S_n = \sum_{i=1}^{n} \eta_i$$

$$\mathbb{D}S_n = \frac{n}{12}$$

$$\mathbb{E}S_n = \frac{n}{2}$$

$$\eta = \frac{S_n - \frac{n}{2}}{\frac{n}{12}}$$

При достаточно больших значениях n, $\eta \sim \mathcal{N}(0,1)$. [30]. Получив стандартно-нормальную величину, легко получить нормально распределённую случайную величину, умножив на дисперсию и добавив математическое ожидание. Если возникла необходимость генерации случайной величины из многомерного нормального семейства, то можно применить разложение Холецкого и свести задачу к одномерным генерациям.

Общим подходом для генерации случайной величины с заданной обратимой функцией распределения является следующий метод:

Пусть $F_X(x) = \mathbb{P}\{X < x\}$. Какой вид имеет функция распределения $F_X(X)$?

Распишем

$$\mathbb{P}\{F_X(X) < \xi\} = \mathbb{P}\{X < F_X^{-1}\xi\} = F_X(F_X^{-1}(\xi)) = \xi$$

Из этого следует, что

$$X = F_X^{-1}(\xi);\ \xi \sim \mathcal{U}[0,1]$$

т.к. функция распределения равномерно распределённой случайной величины равна сумме значений аргументов на $[0,1]$.

Однако не для каждой функции можно выписать обратную функцию в явном виде. Приведём наиболее значимые распределения, для которых это возможно:

1. Показательное распределение:

$$p(x \mid \lambda) = \lambda e^{-\lambda x}, x \geq 0, \lambda > 0 \tag{192}$$

$$F(x) = 1 - e^{-\lambda x} = \xi \Rightarrow x = -\frac{1}{\lambda}\ln(1-\xi) \tag{193}$$

2. Распределение Коши – имеет тяжёлые хвосты, поэтому не существует моментов [31]. Возможно только определение моментов в смысле главного значения.

$$p(x) = \frac{1}{\pi}\frac{1}{1+x^2} \tag{194}$$

$$F(x) = \frac{1}{\pi}\operatorname{arctg}(x) + \frac{1}{2} = \xi \Rightarrow x = \operatorname{tg}(\pi(\xi - \frac{1}{2})) \tag{195}$$

[30] В некоторых задачах используется генерация следующего вида: $\sum_{i=1}^{12} \xi_i - 6 \approx \mathcal{N}(0,1)$; $\xi_i \sim \mathcal{U}[0,1]$. Если посмотреть на дисперсию генерируемой по такой формуле величины, то она примет значение, равное 1, поскольку $\mathbb{D}\xi_i = \frac{1}{12}$. Качество приближения гарантировано, в какой-то мере, ЦПТ.

[31] Этот факт имеет интересную реализацию в жизни: Если у всех людей в аудитории время на часах распределено по Коши относительно настоящего времени, то сбор всей информации и усреднение не поможет определить точное время

8.2.2 Метод **Rejection Sampling**

Итак, не умеем генерировать явно из плотности $p(x)$. Идея – давайте промажорируем плотностью, из которой умеем генерировать

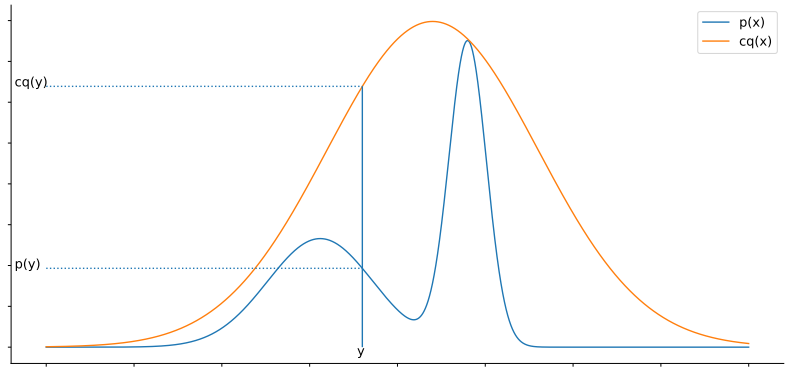


Рис. 12: Пример мажорирования плотностью в Rejection Sampling

Т.е. если $\forall x\ p(x) \leq cq(x)$, где c – нормировочная константа (т.к. не каждую плотность можно промажорировать без нормировки), и мы умеем генерировать из плотности $q(x)$ (см. рис. 12). Тогда будем принимать точку, сгенерированную из $q(x)$, только с определённой вероятностью, отражающей приближение.

Запишем формально: $y \sim q(x);\ \xi \sim \mathcal{U}[0, cq(y)] \Rightarrow accept\ y\ if\ \xi < p(y)$

Докажем, что схема работает:

Фактически, мы утверждаем, что если x_{n+1} – следующая точка выборки, а $y \sim q(x)$, то мы примем y в качестве следующего сэмпла с вероятностью $\frac{p(y)}{cq(y)}$.

Запишем это как:

$$\mathbb{P}\{accept\} = \mathbb{E}_y \mathbb{P}\{accept \,|\, y\} = \int \frac{p(y)}{cq(y)} q(y) dy = \frac{1}{c} \tag{196}$$

Причем, чем ближе распределение $q(y)$ к распределению $p(x)$, тем меньше мажорирующая константа c, тем эффективнее работает схема – будет приниматься больше семплов.

Следовательно,

$$r(y \,|\, accept) dy = \frac{r(accept \,|\, y) r(y)}{r(accept)} dy = \frac{\frac{p(y)}{cq(y)} q(y)}{\frac{1}{c}} dy = p(y) dy \tag{197}$$

Мы показали, что те сэмплы, которые мы приняли, будут иметь нужное распределение $p(y)$. ■

Как видно из формул, качество приближения определяется отношением площадей под плотностями, которое равно $\frac{1}{c}$ (см. формулу 196). Метод будет работать хорошо, если будет мало отклонений, и если c будет малым, то необходимо будет больше точек для приближения.

Проблемные примеры:

Пример. Пусть мы не умеем сэмплировать распределение Коши, хотим приблизить его нормальным распределением, из которого умеем (рис. 13)

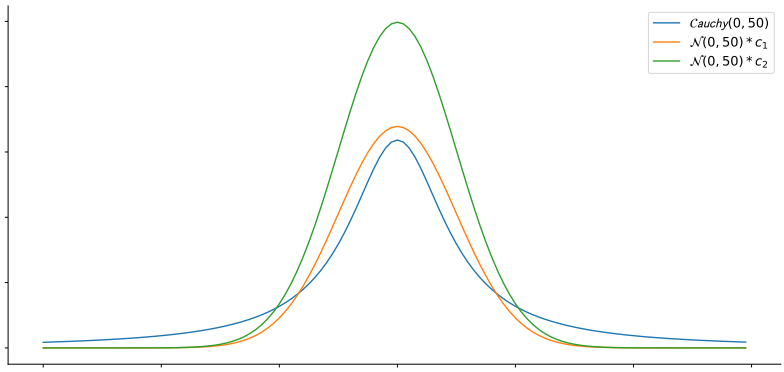


Рис. 13: Пример плохого приближения методом Rejection Sampling

В данном случае, мы сколь угодно точно можем приблизить центральную часть распределения, однако, из-за того, что распределение Коши имеет тяжёлые хвосты, нам понадобится $c \to \infty$. В свою очередь, обратная схема приближения (нормальное - Коши) будет работать хорошо, т.к. гарантировано будет хорошее приближение на хвостах.

Пример. Если распределение стремится к виду δ-функции (см. рис.14), то в приближении мы будем генерировать много точек, хотя распределение задано, фактически, только в одной точке.

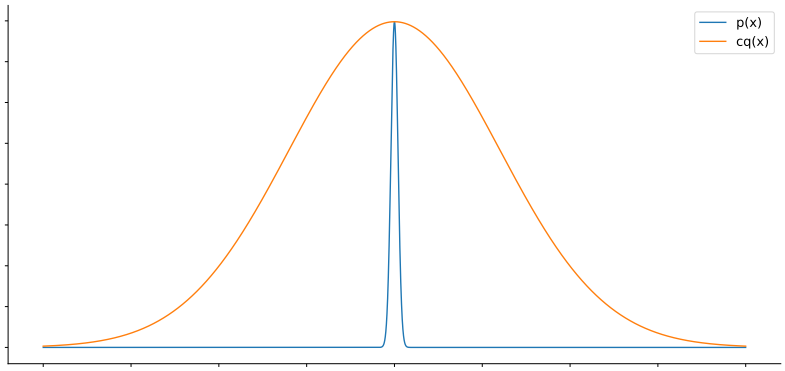


Рис. 14: Пример приближения в одной точке Rejection Sampling

В связи с этим, хотелось бы внести информацию о важности сэмплированной точки. Это призван осуществить метод *Importance sampling*.

Как же улучшить метод *Reject sampling* на случай произвольного распределения, чтобы хорошее приближение было почти всегда? Т.к. эффективность схемы определяется отношением площадей под распределениями, то давайте попробуем построить решётку и приближать на каждом сегменте, исходное распределение, например, равномерным. А в хвостах, например, распределением Коши (можно взять так, чтобы площадь под хвостами была равна $\frac{1}{2}$). Тогда при использовании такого приближения и наличии умения считать площадь на каждом сегменте схема будет иметь следующий вид:

1) Реализуем дискретную случайную величину, равную количеству «бинов» , чтобы понять, из какого сегмента будем сэмплировать дальше ($\mathbb{P}(bin) \propto S_{bin}$). Если выпал

определённый «бин» , реализуем на нём выборку из равномерного распределения.

2) Применяем *Reject sampling* на этом «бине».

Таким образом, улучшаем качество метода. В общем случае, если носитель конечен, то приближаем равномерным распределением, а на хвостах – полубесконечным (например, показательным, т.к. умеем генерировать его и считать площадь под плотностью).

8.2.3 Метод **Importance sampling**

Метод призван уменьшить отклонения сэмплов. Это реализуется при помощи внедрения весов сэмпла:

$$\mathbb{E}_{p(x)} f(x) = \mathbb{E}_{q(x)} \frac{p(x)}{q(x)} f(x) \approx \frac{1}{K} \sum_{k=1}^{K} \frac{p(x_k)}{q(x_k)} f(x_k) = \sum_{i=1}^{N} v_i f(x_i) \tag{198}$$

где $v_i = \frac{1}{K} \frac{p(x_k)}{q(x_k)}$ и задаёт важность объекта $x_k \sim q(x)$. Снова получаем несмещённую оценку, но теперь берём все точки. Однако если v_i задаются сильно неравномерно (см. рис. 15), то большие значения распределения будут вносить бóльший вклад.

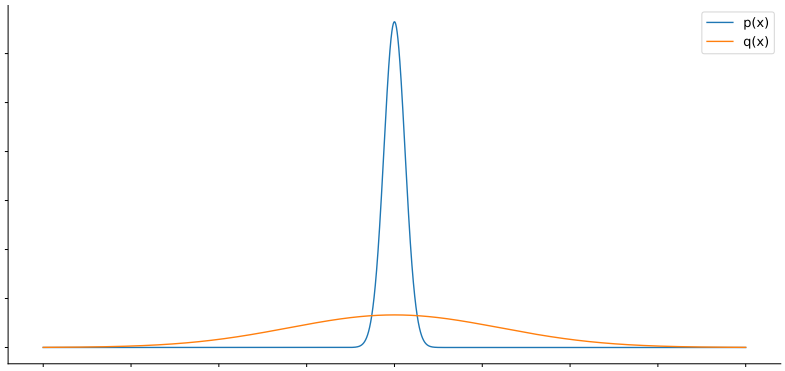


Рис. 15: Пример плохого выбора $q(x)$ в Importance sampling

Когда же важности будут распределены неравномерно? Когда распределения $p(x)$ и $q(x)$ сильно различаются. Но в таком виде *Importance sampling* запускать бессмысленно. Мы снова неявно предполагаем качественное приближение $p(x)$ распределением $q(x)$, и только в этом предположении схема становится эффективной (то есть v_i распределены равномерно), что важно для любого метода Монте-Карло, т.к. это медленно сходящиеся методы, в отличии от вариационной аппроксимации [32].

Плотность распределения $p(x)$ обычно представима в виде: $\frac{\hat{p}(x)}{B}$. И в случае, когда мы явно можем посчитать знаменатель Z, проинтегрировав плотность, то получаем в нашей схеме явный байесовский вывод с апостериорной плотностью. Если же не можем посчитать

[32] Имеем некоторую проблему tradeoff: Монте-Карло методы дают несмещённую оценку, однако долго приближают из-за высокой дисперсии, для сокращения которой необходимо большое по объёму сэмплирование, а вариационные методы дают смещённую оценку, меняя статистики от $f(x)$, но достаточно быстрые. Компромисс пытается найти современный подход – неявное вероятностное моделирование

нормировочную константу, то имеем дело только с $\hat{p}(x)$. Проверим, усложняет ли это схему:

$$\int p(x)f(x)dx = \frac{1}{B}\int \hat{p}(x)f(x)dx = \frac{\int \hat{p}(x)f(x)dx}{\int \hat{p}(x)dx} =$$

$$= \frac{\int q(x)\frac{\hat{p}(x)}{q(x)}f(x)dx}{\int q(x)\frac{\hat{p}(x)}{q(x)}dx} \approx \frac{\frac{1}{K}\sum\limits_{k=1}^{K}\frac{\hat{p}(x_k)}{q(x_k)}f(x_k)}{\frac{1}{K}\sum\limits_{k=1}^{K}\frac{\hat{p}(x_k)}{q(x_k)}} = \sum_{i=1}^{K} w_i f(x_k), \quad x_k \sim q(x) \tag{199}$$

Если распределение $q(x)$ сильно не похоже на $p(x)$ (см. рис. 15), то многие w_i будут близки к 0, но будет много сэмплов, точно не попавших в $p(x)$. В случае, когда мода распределения $q(x)$ находится далеко от моды $p(x)$, будет также происходить доминирование одной точки, попавшей в $p(x)$, над большим количеством точек с меньшими весами.

Итак, были рассмотрены основные методы сэмплирования и методы сэмплирования из одномерных плотностей. Но последние методы плохо переносятся на случай высокой размерности. Оказалось, что нам поможет аппарат Марковских цепей.

8.3 Метод Метрополиса-Хастингса

Основной идеей метода Метрополиса-Хастингса является отказ от сэмплирования из равномерного распределения. Вместо этого предлагается сэмплировать последовательно из марковской цепи.

Определение 9. Марковской цепью назовём бесконечную последовательность случайных величин $\{x_n\}_{n=1}^{\infty}$ такую, что $\forall n$ верно разложение:

$$p(x_1, ..., x_n) = p_1(x_1)p_2(x_2 \,|\, x_1)p_3(x_3 \,|\, x_2)\dots p_n(x_n \,|\, x_{n-1}),$$

где функции $p_i(\cdot \,|\, \cdot)$ называют функциями перехода.

Важными видами марковских процессов являются те, у которых функция перехода не зависит от номера перехода.

Определение 10. Марковская цепь называется гомогенной (или однородной), если

$$\forall n \quad p_n(x_n \,|\, x_{n-1}) = q(x_n \,|\, x_{n-1}).$$

В этом случае Марковская цепь определяется стартовым распределением $p_1(x_1)$ и вероятностью перехода $q(x' \,|\, x)$.

Пусть сгенерировали много переходов марковской цепи.

Какое распределение будет иметь x_1? Очевидно, это $p_1(x_1)$.

А чему равна функция $p_2(x_2)$? Для её нахождения достаточно провести процедуру маргинализации совместного распределения:

$$p_2(x_2) = \int p_2(x_2 \,|\, x_1)p_1(x_1)dx_1$$

Эта закономерность повторяется и для дальнейших плотностей, например,

$$p_3(x_3) = \int p_3(x_3 \,|\, x_2)p_2(x_2)dx_2 = \int\int p_3(x_3 \,|\, x_2)p_2(x_2 \,|\, x_1)p_1(x_1)dx_1dx_2$$

В дискретном случае – это выражение означало бы умножение матрицы на вектор, в непрерывном – скалярное произведение, т.е. действие линейного оператора в $\mathcal{L}_2$

Повторяется вид действия линейного оператора, который, в общем случае, для $f \in \mathcal{L}_2$ и K – лин. оператора, имеет вид: $g = Kf$, $g(y) = \int K(x, y)f(x)dx$.

Итак, мы поняли, что маргинальные плотности получаются действием линейного оператора. Обычно мы встречали такой вид в линейной алгебре. Может быть, можем получить какую-то сходимость последовательности плотностей?

Определение 11. $p_*(x)$ называется инвариантным распределением (или стационарным) для гомогенной марковской цепи, если $p_*(x') = \int q(x' \mid x) p_*(x) dx$.

Т.е. если мы были в распределении $p_*(x')$, то подействовав на него вероятностью перехода $q(x' \mid x)$, мы останемся в распределении $p_*(x')$.

Рассмотрим примеры функций перехода, ведь, фактически, от неё зависит наличие стационарных точек в марковской цепи.

Пример. Пусть $q(x' \mid x) = \delta(x' - x)$. Тогда любое распределение $p_*(x)$ является стационарным. Если же $q(x' \mid x) = \delta(x' - x - a)$, то не существует инвариантных распределений.

Следовательно, есть целый спектр функций перехода. Когда же гарантировано существование и единственность инвариантного распределения.

Определение 12. Цепи, для которых существует единственное инвариантное распределение $p_*(x)$ будем называть эргодическими, если

$$\forall p_1(x) \quad \lim_{n \to \infty} p_n(x) = p_*(x).$$

Поэтому в задаче генерации выборки мы можем построить одну эргодическую марковскую цепь и будем генерировать точки по ней. Тогда гарантировано, что мы сойдёмся к стационарному распределению. Но как сделать так, чтобы оно имело вид исходного распределения?

Определение 13. Если $\forall x', x'' \in \mathcal{X}$, где $\mathcal{X}$ — носитель марковской цепи, справедливо, что $q(x'' \mid x') > 0$, следовательно, марковская цепь с $q(x'' \mid x')$ эргодическая.

Что же это значит для нас? Пусть мы захотим генерировать выборку из $p(x)$. Если мы сможем построить однородную марковскую цепь с этим свойством такую, что её стационарное распределение будет совпадать с $p(x)$, то из любого начального приближения марковская цепь, начиная с какого-то момента, начнёт генерировать сэмплы из целевого распределения $p(x)$.

Теорема 6. (Уравнение детального баланса *Detailed Balance*) Если для эргодичной однородной Марковской цепи и для $p_*(x)$ верно

$$\forall x, x' \in \mathcal{X} \quad p_*(x) q(x' \mid x) = p_*(x') q(x \mid x'),$$

то p_* – инвариантное распределение.

Доказательство. Т.е. если в какой-то момент времени мы находимся в распределении $p_*(x)$ и в следующий момент времени мы также будем находиться в этом распределении, то докажем, что p_* – инвариантное(стационарное).

p_*:

$$\int q(x' \mid x) p_*(x) dx = \{\text{Уравнение DB}\} = \int p_*(x') q(x \mid x') dx = p_*(x') \underbrace{\int q(x \mid x') dx}_{=1} = p_*(x') \quad \blacksquare$$

Почему это уравнение называется уравнением «баланса»? Пусть x, x' – страны, p_* – ВВП страны, а $q(x' \mid x)$ – то, какую долю ВВП страна x инвестирует в x'. Т.е. $q(x' \mid x) p_*(x)$ – сумма (напр. в \$), которую x тратит на x', а $q(x \mid x') p_*(x')$ – сумма, которую x' тратит на x. И если здесь получается равенство, то финансовые потоки сбалансированы, т.е. ВВП стран меняться не будет. Следовательно, если изначально было одно ВВП страны, но потом страна фиксировала долю затрат на другие страны, как и другие, то всё сбалансируется. Причём если какие-то страны находятся в изоляции, то инвариантное распределение не единственно.

Теорема 7. (Метод Метрополиса-Хастингса) Пусть имеется распределение $p(x) = \hat{p}(x)/B$, известное с точностью до нормировочной константы, и функция $r(x' \mid x) > 0 \ \forall x', x \in \mathcal{X}$. Определим вероятность $A(x, x')$ принятия перехода из x' в x как

$$A(x, x') = \min\left(1, \frac{\hat{p}(x') r(x \mid x')}{\hat{p}(x) r(x' \mid x)}\right) = \min\left(1, \frac{p(x') r(x \mid x')}{p(x) r(x' \mid x)}\right).$$

Тогда для $\forall x_0$ и марковской цепи с правилом перехода:

$$x_{n+1} = \begin{cases} x' \sim r(x' \mid x_n), & \text{with probability } A(x_n, x'), \\ x_n, & \text{otherwise,} \end{cases} \tag{200}$$

гарантируется, что $\exists N : \forall n \geqslant N \Rightarrow x_n \sim \hat{p}(x_n)$.

Доказательство. Посмотрим, для начала, на формулу под min: Если рассмотреть симметричное[33] $r(x' \mid x) = r(x \mid x')$ [34]. Тогда берём x' с вероятностью $A(x, x') = \min(1, p(x')/p(x))$. То есть описанный процесс — это блуждание в пространстве иксов, при котором каждая следующая точка гарантированно принимается (если в ней плотность больше) либо принимается с некоторой вероятностью (пропорциональной отношению плотностей в текущей и предыдущей точках). Заметим, что если бы мы принимали только точки, увеличивающие плотность, то получился бы метод оптимизации, который ищет моду распределения $p(x)$.

Итак, описанная процедура сэмплирования задает марковскую цепь с функцией перехода

$$q(x' \mid x) = \underbrace{r(x' \mid x)}_{\text{Распределение новой точки } x'.} \times \underbrace{A(x, x')}_{\text{Вероятность принятия } x'} + \\ + \underbrace{\delta(x - x')}_{\text{Распределение точки}} \times \underbrace{(1 - \int A(x, y) r(y \mid x) dy)}_{\text{Вероятность репликации}}.$$

Во-первых, описанная марковская цепь эргодическая, поскольку свойство $q(x' \mid x) > 0, \forall x, x'$ наследуется от $r(x' \mid x) > 0$. Тогда если распределение $p(x)$ инвариантно для марковской цепи с функцией перехода $q(x' \mid x)$, то «прогретая» марковская цепь будет генерировать сэмплы из $p(x)$. Инвариантность $p(x)$ покажем с помощью уравнения детального баланса:

$$p(x) q(x' \mid x) \stackrel{?}{=} p(x') q(x \mid x') \tag{201}$$

Рассмотрим левую часть уравнения 201:

$$p(x) q(x' \mid x) = p(x) r(x' \mid x) \min\left(1, \frac{p(x') r(x \mid x')}{p(x) r(x' \mid x)}\right) + \tag{202}$$

$$+ p(x) \delta(x - x') \left(1 - \int A(x, y) r(y \mid x) dy\right) \tag{203}$$

второе слагаемое $\neq 0 \iff x' = x$, тогда заменим все x на x'; в первом слагаемом занесём $p(x) r(x' \mid x)$ в минимум:

$$\tag{204}$$

$$= \min\{p(x) r(x' \mid x), p(x') r(x \mid x')\} + \tag{205}$$

$$+ p(x') \delta(x - x') \left(1 - \int A(x', y) r(y \mid x') dy\right) \tag{206}$$

в первом слагаемом вынесем $p(x') r(x \mid x')$ из минимума:

$$= p(x') r(x \mid x') \underbrace{\min\left(\frac{p(x) r(x' \mid x)}{p(x') r(x \mid x')}, 1\right)}_{= A(x', x)} + \tag{207}$$

$$+ p(x') \delta(x - x') \left(1 - \int A(x', y) r(y \mid x') dy\right) \tag{208}$$

$$= p(x') q(x \mid x'). \tag{209}$$

[33] Именно таковыми были предпосылки Метрополиса. Хастингс вывел для несимметричного случая

[34] Например, нормальное распределение с $\mathbb{E}y = x$ и дисперсией – любой изотропной ковариационной матрицей

Что и требовалось доказать. ■

Схема считается эффективной, если по ходу процесса точки-кандидаты x' принимаются с высокой частотой. Если же большая часть сгенерированных точек будет отвергаться, то в схеме будет много повторов, что можно считать недостатком метода. Еще одним недостатком является корреляция между последовательными элементами В принципе, это недостаток MCMC методов – высокая скоррелированность соседних случайных величин, а хотели бы выборку (H.O.P.C.B.).

Частичное решение этой проблемы – брать каждый n-ый элемент. Но, в общем случае, корреляция сохраняется, следовательно, возникает задача подбора q такой, чтобы была высокая вероятность перехода.

Рассмотрим частный случай схемы Метрополиса-Хастингса, который называется схема Гиббса.

8.4 Схема Гиббса

Имеем многомерное распределение $p(X)$. Идея состоит в сэмплировании каждой координаты по очереди, для этого используются условные вероятности. Схема такая:

$$\begin{aligned}
x_1^{(k+1)} &\sim p(x_1 \,|\, x_2^{(k)}, x_3^{(k)}, \ldots, x_d^{(k)}),\\
x_2^{(k+1)} &\sim p(x_2 \,|\, x_1^{(k+1)}, x_3^{(k)}, \ldots, x_d^{(k)}),\\
&\ldots\\
x_d^{(k+1)} &\sim p(x_d \,|\, x_1^{(k+1)}, x_3^{(k+1)}, \ldots, x_{d-1}^{(k+1)}).
\end{aligned}$$

Очевидным преимуществом схемы Гиббса является тот факт, что сэмплирование происходит из одномерного распределения, а для этого существуют сравнительно эффективные подходы, например, упоминаемый выше Rejection Sampling. Кроме того, в схеме Гиббса нет репликаций и вычисления $A(x, x')$ (часто подсчет корректировки является вычислительно трудным) в отличии от схемы Метрополиса-Хастингса.

К недостаткам можно отнести невозможность распараллеливания вычислений из-за того, что каждый сампл получается от непосредственно предыдущего. Ещё одной проблемой является сложность сэмплирования из распределений с несколькими модами, поскольку одна марковская цепь с высокой вероятностью будет блуждать около одной модой и лишь изредка будет «перепрыгивать» с моды на моду.

9 Лекция 9. Гамильтоновы методы Монте-Карло

Продолжая разбирать методы Монте-Карло с марковскими цепями, мы рассмотрим более продвинутые методы MCMC, а именно Гамильтонову и Ланжевенову динамику. Из названия можно заметить, что методы имеют некоторые параллели с физикой, поэтому придется немного вспомнить физику.

Пусть у нас есть потенциальное поле $\Pi(x)$, $x \in \mathbb{R}^d$, тогда как известно из курса физики уравнение движение выглядит следующим образом (считаем массу единичной):

$$\ddot{x} = -\frac{\partial \Pi}{\partial x} \tag{210}$$

Следовательно, если потенциальное поле не однородно, то тело приобретает такое ускорение, чтобы достигнуть минимума потенциальной энергии. Уравнение 210, можно записать иначе, например так:

$$\begin{cases} \dot{x} = v \\ \dot{v} = -\frac{\partial \Pi}{\partial x} \end{cases} \tag{211}$$

Теперь если мы введем кинетическую энергию $K(v) = \frac{\|v\|^2}{2}$ (все также считаем массу единичной), тогда следуя физике мы можем ввести гамильтониан системы $H(x, v) = \Pi(x) + K(v)$. Введенные определения позволят сформулировать систему 211 в единообразном виде:

$$\begin{cases} \dot{x} = \frac{\partial H}{\partial v} \\ \dot{v} = -\frac{\partial H}{\partial x} \end{cases} \tag{212}$$

Для исследования поведения системы во времени следует рассмотреть производную $\frac{dH}{dt}$, которая по определению равна:

$$\frac{dH}{dt} = \frac{\partial H}{\partial x}\frac{dx}{dt} + \frac{\partial H}{\partial v}\frac{dv}{dt} = \{\text{используем 212}\} = \frac{\partial H}{\partial x}\frac{\partial H}{\partial v} - \frac{\partial H}{\partial v}\frac{\partial H}{\partial x} = 0 \tag{213}$$

Получение тождество 213 означает закон сохранения энергии для тела, на которое не действует никаких внешних сил. Поскольку производная гамильтониана равна нулю, его значение вдоль траектории решения системы остается постоянным.

Другой важной особенностью уравнений Гамильтона является симметрия относительно обращения времени. Если кривая $(x(t), v(t))$ удовлетворяет уравнениям Гамильтона, то и кривая $(x(-t), -v(-t))$ будет удовлетворять этим уравнениям, что можно проверить подстановкой.

Нашей задачей, как и в прошлой лекции, является построение сэмплера из распределения $p(x) = \frac{\hat{p}(x)}{B}$. Введем $\Pi(x) = -\log \hat{p}(x)$, мотивацией служит желание сэмплировать те x, для которых $p(x)$ выше, то есть где $\Pi(x)$ меньше.

9.1 Гамильтонов метод Монте-Карло

Генерируем цепочку 214 по алгоритму 1

$$[x_1, v_1] \to \ldots \to [x_n, v_n] \to \ldots \tag{214}$$

Пример работы алгоритма представлен на Рис. 16, важно заметить, что для каждой итерации $H(x_n, v_n) = H(x', v')$ (движение по линии уровня). При этом при сэмплировании новой скорости мы перепрыгиваем с одной линии уровня, на другую, то есть при большой скорости перейти от одной моды к другой.

Algorithm 1 Процесс получения цепочки

for $n = \overline{1, \infty}$ do
  $v_n \sim \mathcal{N}(v_n \,|\, 0, I)$
  Получаем $[x', v']$ как решение динамики 212 для $[x_n, v_n]$
  $x_{n+1} = x'$
end for

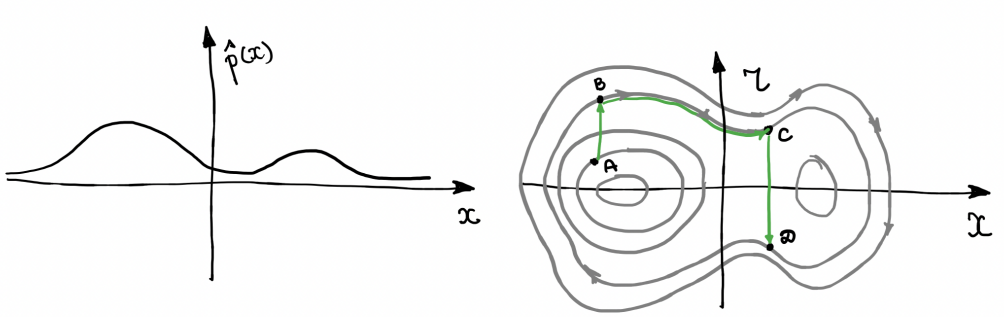


Рис. 16: Слева: Пример бимодального распределения, для которого может быть применен гибридный Монте-Карло. Справа: Решения уравнений Гамильтона на плотности (x, r) (фазовом пространстве) (здесь $r \equiv v$). При этом в силу закона сохранения энергии $H(x, r) = H(x', r') = const$, то есть конечная точка решения динамики Гамильтона находится на той же линии уровня, что и начальная.

В алгоритме 1 нам необходимо решать динамику 212, то есть нам нужно знать $\frac{\partial \log \hat{p}(x)}{\partial x}$. Стоит отметить, что $\frac{\partial \log \hat{p}(x)}{\partial x} \equiv \frac{\partial \log p(x)}{\partial x}$, то есть нам достаточно знать градиент ненормированной плотности, который найти намного проще. При этом все точки принимаются и нет коррекции Метрополиса–Хастингса.

Докажем корректность алгоритма 1, для этого нам нужно показать эргодичность получаемой цепочки и показать, что распределение $p(x)$ для неё инвариантна. Начнём с эргодичности. Функция перехода имеет вид

$$q(x', v' \,|\, x, v) = \delta([x', v'] - f(x, v)), \quad f(\cdot, \cdot) - \text{решение гамильтоновой динамики} \tag{215}$$

Если функция перехода является дельта-функцией, то говорить об эргодичности не приходится. Однако нас интересует эргодичность в пространстве x, а не $[x, v]$. Функция $q(x' \,|\, x)$ получается по правилу суммирования:

$$q(x' \,|\, x) = \iint q(x', v' \,|\, x, v) p(v) dv dv'.$$

Видим, что $q(x' \,|\, x) > 0$, $\forall x', x$, т.е. марковская цепь $\{x_n\}_{n=1}^{\infty}$ эргодическая.

Проверим инвариантность $p(x)$ для этой функции перехода по определению:

$$\int p(x) q(x' \,|\, x) dx = \iiint p(x) p(v) q(x', v' \,|\, x, v) dv dv' dx$$

подставим $p(v) = \mathcal{N}(v \,|\, 0, I)$, и $p(x) = \exp \log p(x)$:

$$= \iiint e^{\log p(x)} \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{\|v\|^2}{2}} q(x', v' \,|\, x, v) dv dv' dx$$

по определению $H(x, v) = -\log p(x) + \|v\|^2/2$:

$$= \iiint \frac{e^{-H(x,v)}}{(2\pi)^{d/2}} q(x', v' \,|\, x, v) dv dv' dx$$

обозначим $p_H(x,v) = e^{-H(x',v')}/(2\pi)^{d/2}$:

$$= \iiint p_H(x,v)q(x',v' \mid x,v)dvdv'dx$$

в силу того, что (x,v) и (x',v') на одной линии уровня:

$$= \iiint p_H(x',v')q(x',v' \mid x,v)dvdv'dx$$

в силу обратимости гамильтоновой динамики во времени:

$$= \iiint p_H(x',v')q(x,-v \mid x',-v')dvdv'dx$$
$$= \{u := -v,\ u' := -v'\}$$

переменные x, u выинтегрировались:

$$= \iiint p_H(x',u')q(x,u \mid x',u')dudu'dx$$
$$= \int p_H(x',u')du'$$

в силу независимости $p_H(x,v) = e^{-\log p(x)}e^{\|v\|^2/2} = p(x)p(v)$:

$$= \int p(x')p(v')dv' = p(x')$$

Получили $\int p(x)q(x' \mid x)dx = p(x')$ — определение инвариантности $p(x)$ для рассматриваемой марковской цепи. Корректность схемы доказана, самплы действительно будут из распределения $p(x)$.

С одной стороны, схема крайне эффективная, поскольку теперь нет корректировки Метрополиса–Хастингса, принимаются все точки. С другой стороны, теперь необходимо каждый раз считать градиент $\nabla \log \hat{p}(x)$. Можно провести аналогию с методами оптимизации: метод МХ — это метод нулевого порядка, а гамильтонов метод Монте-Карло — первого.

Ещё пространство сэмплирования увеличилось в 2 раза, что может быть критично при больших d.

К сожалению, на практики мы не умеем решать систему дифференциальных уравнений аналитически (за редким исключением), поэтому для решения гамильтоновой динамики (системы 212) нужно использовать численные методы.

9.2 Leap-frog Integration.

В качестве начальных условий для системы гамильтоновых уравнений 212 возьмем точку x_n и просэмплируем $v_n \sim N(v_n \mid 0, I)$.

Разобьем интервал времени $[0,T]$ по сетке. В начальный момент времени положим: $[x_0, v_0] = [x_n, v_n]$

Будем использовать следующую численную схему:

$$\begin{cases} v_{t+\frac{1}{2}} = v_t - \frac{\varepsilon}{2}\frac{\partial H(x_t,v_t)}{\partial x} = v_t + \frac{\varepsilon}{2}\frac{\partial \log \hat{p}(x_t)}{\partial x}, \\ x_{t+1} = x_t + \varepsilon v_{t+\frac{1}{2}}, \\ v_t = v_{t+\frac{1}{2}} + \frac{\epsilon}{2}\frac{\partial \log \hat{p}(x_t)}{\partial x}. \end{cases} \tag{216}$$

Отметим ряд свойств этой схемы:

- схема является симметричной относительно перехода $t \to t+1$, так как явно можно выразить $[x_t, v_t]$ через $[x_{t+1}, v_{t+1}]$;

- свойство сохранения объема в фазовом пространстве:

$$\det\left(\frac{\partial[x_{t+1}, v_{t+1}]}{\partial[x_t, v_t]}\right) = 1 + o(\varepsilon^2);$$

- устойчивость, благодаря которой схема не накапливает ошибку (подробнее в следующей секции)

9.3 Leap-frog Sampling.

Вообще говоря, при переходе от аналитического решения к численной схеме, нарушается условие $p_H(x, v) = p_H(x', v')$ и корректность всей схемы ставится под большое сомнение. Например, если использовать самую простую схему интегрирования — схема Эйлера, — то с каждым шагом будет накапливаться ошибка, уводящая далеко от исходной линии уровня. Однако leap-frog достаточно устойчив, чтобы его ошибку можно было компенсировать с помощью корректировки Метрополиса–Хастингса, см. Алгоритм 2. В нём в качестве функции f использована динамика leap-frog с инверсией скорости:

$$f : [x, v] \xrightarrow{LF} [x', \hat{v}] \xrightarrow{v'=-\hat{v}} [x', v']\,. \tag{217}$$

Вероятность корректировки:

$$A(x, x') = \min\left\{1, \frac{p_H(x', v')}{p_H(x, v)}\right\}. \tag{218}$$

Трюк с инверсией скорости нужен для обратимости во времени:

$$z = f(z') = f(f(z)) = f(f^{-1}(z)) = f(z) \Rightarrow f(z) = f(z'). \tag{219}$$

Algorithm 2 Leap-frog Sampling

Set x_1
for $l = 1, L$ do
  Sample $v_l \sim \mathcal{N}(v_l \,|\, 0, I)$
  $[x', v'] := f([x_l, v_l])$
  $x_{l+1} = \begin{cases} x', & \text{с вероятностью } A(x_l, x'), \\ x_l, & \text{с вероятностью } 1 - A(x_l, x'). \end{cases}$
end for

9.4 Динамика Ланжевена.

Рассмотрим частный случай leap-frog integration с $T = 1$, который также известен под названием динамика Ланжевена.

В схеме 216 пусть $T = 1$. Тогда процесс генерации точек описывается уравнениями:

$$v_{\frac{1}{2}} = v_0 + \frac{\varepsilon}{2}\frac{\partial \log \hat{p}(x_t)}{\partial x}, \tag{220}$$

$$x_1 = x_0 + \varepsilon v_{\frac{1}{2}} = x_0 + \varepsilon v_0 + \frac{\varepsilon^2}{2}\frac{\partial \log \hat{p}(x_0)}{\partial x}, \tag{221}$$

где $v_0 \sim N(v_0 \,|\, 0, I)$.

Ранее уже упоминалась связь методов сэмплирования MCMC с методами оптимизации. Для динамики Ланжевена эту связь особенно легко заметить: если взять $v_0 = 0$, то эта схема эквивалентна градиентному подъему, который ищет локальную моду распределения $p(x)$. Слагаемое $v_0 \sim \mathcal{N}(v_0 \,|\, 0, I)$ по сути случайный шум. Поэтому динамику Ланжевена можно трактовать так: градиентный подъём к моде распределения с впрыскиванием шума. Причём заметим, что шум не абы какой, а строго завязанный на лернинг рейте подъёма: шум с отклонением ε, а лернинг рейт $\varepsilon^2/2$.

Также можно проследить параллель с SGD:

- Оба метода обновляют параметры системы (веса модели в случае SGD и координаты и импульсы в случае динамики Ланжевена) на основе градиента логарифма правдоподобия.
- Оба метода также используют масштабирование параметров, чтобы учесть скорость обучения (learning rate в SGD и дисперсия шума в динамике Ланжевена).
- После сходимости к глобальному оптимуму SGD начинает блуждать вокруг его окрестности, аналогично в динамике Ланжевена происходит сэмплирование в окрестности моды максимума правдоподобия.

9.5 Обоснование динамики Leap-frog

Как было сказано, после перехода к численной схеме условие $p_H(x, v) = p_H(x', v')$ нарушается, так как решения системы 216 колеблются вокруг линий уровня, поэтому приходится делать корректировку Метрополиса–Хастингса и в некоторых ситуациях реплицировать предыдущие сэмплы.

Пусть $z = [x, v]$ — начальная точка, $f(z) = z'$ — результат динамики leap-frog с инверсией.

Корректировка MX, введённая в 2 является частным случаем корректировки

$$A(z, z') = \min\left(1, \frac{p(z')}{p(z)}\left|\frac{dz'}{dz}\right|\right) \tag{222}$$

для случая сохранения объёма $|dz'/dz| = 1$.

Итак, описанный алгоритм, очевидно, задает следующую функцию перехода:

$$q(z^{new} \mid z) = \underbrace{\delta(z^{new} - f(z))}_{\text{распределение новой точки } z'} \times \underbrace{A(z, f(z))}_{\text{вероятность принятия } z'} + \\ + \underbrace{\delta(z^{new} - z)}_{\text{распределение старой точки}} \times \underbrace{(1 - A(z, f(z)))}_{\text{вероятность репликации}}.$$

Как и в общем случае гамильтоновой динамики, такая функция перехода не дает эргодичность в пространстве $[x, v]$, но дает в пространстве x.

Докажем инвариантность $p(z) := p_H(x, v)$ для такой цепочки по определению

$$\int p(z)q(z^{new} \mid z)dz = \int p(z)\delta(z^{new} - f(z))A(z, f(z))dz + \\ + \int p(z)\delta(z^{new} - z)(1 - A(z, f(z)))dz.$$

Работаем с первым слагаемым:

$$\int p(z)\delta(z^{new} - f(z))A(z, f(z))dz = \left\{ y := f(z),\ dz := \left|\frac{df}{dz}\right|^{-1} dy = dy \right\}$$
$$= \int p(f^{-1}(y)\delta(z^{new} - y)A(z, f(z))dy$$
$$= \int p(f^{-1}(y)\delta(z^{new} - y)\min\left\{1, \frac{p(y)}{p(f^{-1}(y))}\right\} dy$$
$$= p(f^{-1}(z^{new}))\min\left\{1, \frac{p(z^{new})}{p(f^{-1}(z^{new}))}\right\}$$
$$= \min\left\{p(f^{-1}(z^{new})), p(z^{new})\right\}$$
$$= p(z^{new})\min\left\{\frac{p(f^{-1}(z^{new}))}{p(z^{new})}, 1\right\}$$
$$= \{f^{-1}(z^{new}) = f(z^{new}) \text{ — обратимость по времени}\}$$
$$= p(z^{new})\min\left\{\frac{p(f(z^{new}))}{p(z^{new})}, 1\right\}$$

Второе слагаемое:

$$\int p(z)\delta(z^{new} - z)(1 - A(z, f(z)))dz$$
$$= p(z_{new})(1 - A(z^{new}, f(z^{new})))$$
$$= p(z^{new}) - p(z^{new})A(z^{new}, f(z^{new})$$
$$= p(z^{new}) - p(z^{new})\min\left\{1, \frac{p(f(z^{new}))}{p(z^{new})}\right\}$$

Складываем и видим, что остается только $p(z^{new})$. Инвариантность доказана.

10 Лекция 10. Латентное размещение Дирихле

10.1 Тематическая модель LDA

Напоминание: говорим, что $\boldsymbol{\theta}$ лежит на K-мерном вероятностном симплексе ($\theta \in S_K$), если $\forall k\ \theta_k \geq 0$ и $\sum_{k=1}^{K} \theta_k = 1$.

Определение 14. Говорят, что случайная величина $\boldsymbol{\theta}$ имеет распределение Дирихле с параметрами α, если ее плотность можно выразить в следующем виде:

$$\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\theta} \,|\, \boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_j)}{\prod_k \Gamma(\alpha_j)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1},\ \ \alpha_k > 0$$

Ранее обсуждалось, что если $\forall k\ \ \alpha_k < 1$, то распределение имеет U-образную форму на симплексе, то есть поощряется ситуация, когда часть компонент θ имеют близкие к нулю или нулевые значения и максимальная плотность будет достигаться, когда одна компонента равна 1, остальные 0. Если параметры $\forall k\ \ \alpha_k = 1$, то это равномерное распределение на симплексе, а если $\forall k\ \ \alpha_k > 1$, то колоколообразное распределение.

Можно посчитать математическое ожидание:

$$\mathbb{E}\,\theta_k = \frac{\alpha_k}{\sum_{j=1}^{K} \alpha_j}$$

Распределение Дирихле относится к экспоненциальному классу распределений с набором параметров $[\alpha_1 - 1, \ldots, \alpha_K - 1]$ и достаточными статистиками $\boldsymbol{u}(\boldsymbol{\theta}) = [\log\theta_1, \ldots, \log\theta_K]$. Можно рассмотреть математическое ожидание логарифма, получится следующее выражение:

$$\mathbb{E}\log\theta_k = \psi(\alpha_k) - \psi(\alpha_0), \quad \text{где } \psi(x) \stackrel{def}{=} \frac{d\log\Gamma(x)}{dx} \text{ — дигамма-функция.}$$

Считаем, что $\mathbb{E}\log\theta_k$ можно аналитически посчитать.

Введем одну из вариаций на тему латентного размещения Дирихле.

Модель латентного размещения Дирихле была разработана, чтобы выделять в текстах тематические профили. Предполагается, что каждый текст – это набор слов, каждое слово соответствует одной из тем, каждая тема задает распределение на множестве слов языка. Постановка задачи выглядит следующим образом: есть корпус текстов, требуется восстановить распределение слов для каждой из тем и какие темы присутствуют в каждом документе. Предполагается, что всего тем много, но в каждом тексте содержится небольшое количество тем.

Введем ряд обозначений:

W	— размер словаря
$w_{dn} \in \{1, \ldots, W\}$	— слово на позиции n в документе d, категориальная величина
T	— гиперпараметр, количество тем
$z_{dn} \in \{1, \ldots, T\}$	— из какой темы пришло слово на позиции n в документе d
$\boldsymbol{\theta}_d = [\theta_{d,1}, \ldots, \theta_{d,T}] \in S_T$	— вероятности тем в документе d, $\theta_{d,t} \geqslant 0$, $\sum_{t=1}^{T} \theta_{d,t} = 1$
$\Phi = [\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_T]^T \in \mathbb{R}^{T \times W}$	— матрица, задает профиль каждой темы, $\phi_t \in S_W$

Вероятностная модель LDA задаётся следующим образом:

$$p(W, Z, \Theta \,|\, \Phi) = \prod_{d=1}^{D} \left(p\left(\boldsymbol{\theta}_d\right) \prod_{n=1}^{N_d} p\left(w_{dn} \,|\, z_{dn}, \Phi\right) p\left(z_{dn} \,|\, \boldsymbol{\theta}_d\right) \right) =$$

$$= \prod_{d=1}^{D} \left(\mathrm{Dir}\left(\boldsymbol{\theta}_d \,|\, \boldsymbol{\alpha_0}\right) \prod_{n=1}^{N_d} \varphi_{z_{dn}, w_{dn}} \theta_{d, z_{dn}} \right) =$$

$$= \left(\prod_{d=1}^{D} \frac{1}{\mathrm{B}(\boldsymbol{\alpha_0})} \prod_{t=1}^{T} \theta_{dt}^{\alpha_0 - 1} \right) \left(\prod_{d=1}^{D} \prod_{n=1}^{N_d} \prod_{t=1}^{T} (\varphi_{t w_{dn}} \theta_{dt})^{[z_{dn} = t]} \right).$$

α_0 - небольшой параметр, порядка $10^{-3} - 10^{-6}$, чтобы поощрить разреженность получаемых тематических профилей каждого документа.

10.2 EM-алгоритм для модели LDA

В рассматриваемой задаче дано W, требуется найти Φ. Рассмотрим решение задачи обучения модели LDA с помощью метода максимального правдоподобия:

$$p(W \mid \Phi) \to \max_{\Phi}$$

Величина правдоподобия $p(\mathcal{W} \mid \Phi)$ не может быть вычислена аналитически даже для небольших T и объемов документов в корпусе, так как требует, в частности, суммирования по всем Z, что соответствует суммированию по $T^{\sum_d N_d}$ слагаемым.

Воспользуемся вариационным EM-алгоритмом:

$$\text{E-шаг: } q(Z, \Theta) \simeq p(Z, \Theta \mid W, \Phi),$$
$$\text{M-шаг: } \mathbb{E}_{z,\theta} \log p(W, Z, \Theta \mid \Phi) \to \max_{\Phi}$$

Нюанс для рефлексирования на экзамене: в явном виде вычислить $p(Z, \Theta \mid W, \Phi)$ можно если только если возможно в аналитическом виде рассчитать $p(W \mid \Phi)$, так как по теореме Байеса $p(Z, \Theta \mid W, \Phi) = \frac{p(W \mid Z, \Theta W, \Phi) p(Z, \Theta)}{p(W \mid \Phi)}$.

10.2.1 E-шаг

На E-шаге EM-алгоритма необходимо найти апостериорное распределение $p(Z, \Theta, \mid W, \Phi)$.

Распределения $p(Z \mid \Theta)p(\Theta)$ и $p(W \mid Z, \Theta, \Phi)$ не сопряжены, следовательно, нельзя аналитически вывести E-шаг. Однако, есть условная сопряженность. Относительно Θ правдоподобие $p(Z \mid \Theta)$ и априорное распределение $p(\Theta)$ сопряжены. Аналогично по параметру Z правдоподобие $p(W \mid Z, \Phi)$ и априорное распределение $p(Z \mid \Theta)$ сопряжены. И значит, можно использовать метод Mean Field Approximation.

Будем искать аппроксимирующее распределение q в семействе факторизованных распределений:

$$q(Z, \Theta) \approx q(Z) q(\Theta)$$

Напомним, что в методе Mean Field Approximation приближенное апостериорное распределение ищется в форме факторизованного распределения, и чтобы найти каждый фактор, нужно взять математическое ожидание по всем остальным факторам от логарифма полного правдоподобия. Сначала распишем $\log q(\Theta)$, группируя все члены, не зависящие от Θ, в $\text{const}(\Theta)$:

$$\log q(\Theta) = \mathbb{E}_{q(Z)} \log p(W, Z, \Theta \mid \Phi) + \text{const}(\Theta) =$$
$$= \mathbb{E}_{q(Z)} \left(\sum_{d=1}^{D} \sum_{t=1}^{T} (\alpha_0 - 1) \log \theta_{dt} + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{t=1}^{T} [z_{dn} = t] \left(\log \varphi_{t w_{dn}} + \log \theta_{dt} \right) \right) + \text{const}(\Theta) =$$
$$= \sum_{d=1}^{D} \sum_{t=1}^{T} (\alpha_0 - 1) \log \theta_{dt} + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{t=1}^{T} \log \theta_{dt} \mathbb{E}_{q(Z)} [z_{dn} = t] + \text{const}(\Theta) =$$
$$= \left[\mu_{dnt} = \mathbb{E}_{q(Z)} [z_{d,n} = t] = P[z_{dn} = t] \right] =$$
$$= \sum_{d=1}^{D} \sum_{t=1}^{T} \log \theta_{dt} \left(\alpha_0 - 1 + \sum_{n=1}^{N_d} \mu_{dnt} \right) + \text{const}(\Theta).$$

Тогда можем записать

$$q(\Theta) = \prod_{d=1}^{D} q(\boldsymbol{\theta}_d) = \prod_{d=1}^{D} \frac{\prod_{t=1}^{T} \theta_{dt}^{\alpha_0 - 1 + \sum_{n=1}^{N_d} \mu_{dnt}}}{B_d}$$

$$\Rightarrow q(\theta_d) \sim \mathrm{Dir}(\boldsymbol{\theta}_d \,|\, \beta_d), \quad \beta_{dt} = \alpha_0 + \sum_n \mu_{dnt}$$

Теперь распишем $\log q(Z)$, группируя все члены, не зависящие от Z, в $\mathrm{const}(Z)$:

$$\log q(Z) = \mathbb{E}_{q(\Theta)} \log p(W, Z, \Theta \,|\, \Phi) + \mathrm{const}(Z) =$$
$$= \mathbb{E}_{q(\Theta)} \sum_{d=1}^{D} \left(\sum_{t=1}^{T} (\alpha_0 - 1) \log \theta_{dt} + \sum_{n=1}^{N_d} \sum_{t=1}^{T} [z_{dn} = t] \left(\log \varphi_{tw_{dn}} + \log \theta_{dt} \right) \right) + \mathrm{const}(Z) =$$
$$= \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{t=1}^{T} [z_{dn} = t] \left(\mathbb{E}_{q(\Theta)} \log \theta_{dt} + \log \varphi_{tw_{dn}} \right) + \mathrm{const}(Z).$$

Величина $\mathbb{E}_{q(\Theta)} \log \theta_{dt}$ аналитически считается, $\log \varphi_{tw_{dn}}$ известно.
Следовательно,

$$q(Z) = \prod_{d=1}^{D} \prod_{n=1}^{N_d} q(z_{dn}),$$

где

$$q(z_{dn}) = \frac{\prod_{t=1}^{T} \left(\varphi_{tw_{dn}} \exp\left(\mathbb{E}_{q(\Theta)} \log \theta_{dt} \right) \right)^{[z_{dn}=t]}}{B_d} = \sum_{t=1}^{T} \left(\frac{\varphi_{tw_{dn}} \exp\left(\mathbb{E}_{q(\Theta)} \log \theta_{dt} \right)}{\sum_s^T \varphi_{sw_{dn}} \exp\left(\mathbb{E}_{q(\Theta)} \log \theta_{ds} \right)} \right)^{[z_{dn}=t]}$$

$$\Rightarrow \mu_{dnt} = \frac{\varphi_{tw_{dn}} \exp\left(\mathbb{E}_{q(\Theta)} \log \theta_{dt} \right)}{\sum_s^T \varphi_{sw_{dn}} \exp\left(\mathbb{E}_{q(\Theta)} \log \theta_{ds} \right)}$$

10.2.2 M-шаг

На M-шаге EM-алгоритма необходимо оптимизировать вариационную нижнюю оценку $\mathbb{E}_{q(\mathcal{Z},\Theta)} \log p(\mathcal{W}, \mathcal{Z}, \Theta \,|\, \Phi)$ по параметрам модели Φ. Распишем данную оптимизационную задачу:

$$\mathbb{E}_{q(Z)q(\Theta)} \sum_{d=1}^{D} \left(\sum_{t=1}^{T} (\alpha_0 - 1) \log \theta_{dt} + \sum_{n=1}^{N_d} \sum_{t=1}^{T} [z_{dn} = t] \left(\log \varphi_{tw_{dn}} + \log \theta_{dt} \right) \right) =$$
$$= \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{t=1}^{T} \mathbb{E}_{q(Z)q(\Theta)} [z_{dn} = t] \varphi_{tw_{dn}} + \mathrm{const}(\Theta, Z) \to \max_{\Phi}$$
$$\Rightarrow \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{t=1}^{T} \left(\mu_{dnt} \log \varphi_{tw_{dn}} \right) \to \max_{\Phi}$$

Также знаем, что

$$\sum_{v=1}^{W} \varphi_{tv} = 1 \quad \forall t = \overline{1, T}$$

Составим функцию Лагранжа:

$$\mathcal{L} = \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{t=1}^{T} \left(\mu_{dnt} \log \varphi_{tw_{dn}} \right) + \sum_{t=1}^{T} \lambda_t \left(\sum_{v=1}^{W} \varphi_{tv} - 1 \right)$$

Для нахождения φ_{tv} найдем экстремум функции Лагранжа по этим переменным:

$$\frac{\partial \mathcal{L}}{\partial \varphi_{tv}} = \sum_{d=1}^{D} \sum_{n=1}^{N_d} \frac{\mu_{dnt} [w_{dn} = v]}{\varphi_{tv}} + \lambda_t = 0$$
$$\Rightarrow \varphi_{tv} = -\frac{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \mu_{dnt} [w_{dn} = v]}{\lambda_t}$$

Суммируя это равенство по v и зная, что $\sum_{v=1}^{W} \varphi_{tv} = 1$, найдем λ_t:

$$1 = \sum_{v=1}^{W} \varphi_{tv} = \sum_{v=1}^{W} -\frac{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \mu_{dnt}[w_{dn} = v]}{\lambda_t}$$

$$\Rightarrow \lambda_t = -\sum_{d=1}^{D} \sum_{n=1}^{N_d} \mu_{dnt}$$

Таким образом, получим выражение для φ_{tv}:

$$\varphi_{tv} = \frac{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \mu_{dnt}[w_{dn} = v]}{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \mu_{dnt}}$$

E- и M-шаги алгоритма повторяются до сходимости алгоритма. В итоге мы получим профили тем — матрицу вероятностей слов в темах. Каждую тему можно охарактеризовать ее наиболее вероятными словами.

11 Лекция 11. Гауссовские процессы для регрессии и классификации

11.1 Описание байесовских непараметрических моделей

Из курса классического машинного обучения мы знаем, что методы бывают параметрические и непараметрические. До этого в байесовском подходе мы смотрели только на параметрические методы, сейчас же мы будем изучать непараметрические байесовские модели.

Для нас главная особенность параметрических методов заключается в том, что у нас есть заранее определенный набор параметров, который полностью определяет сложность модели, и она не зависит от объёма выборки.

В непараметрических методах количество «параметров» растёт с увеличением объёма выборки, и сложность модели предварительно не фиксируется. Например, в ядровых методах, когда мы переходим к ядрам, то чем больше объектов у нас есть, тем больше признаков, а, значит, тем больше сложность модели. Неформально можно сказать, что «параметрами» непараметрической модели является сама выборка данных. Такие модели бывает полезно использовать, например, при динамическом поступлении данных, либо при неизвестном заранее объёме исследуемых данных.

11.2 Гауссовские случайные процессы.

Изложение теории случайных процессов в данном курсе не является математически строгим, но дается в достаточном объеме для понимания происходящего.

Определение 15. Случайный процесс – совокупность счетного или континуального числа случайных величин. Случайные величины индексируются параметром, который может принимать континуальное или счетное число значений. Будем обозначать эти случайные величины как $f(x)$, где x – индексирующий параметр. $X \in R^d$.

Определение 16. Зафиксируем n индексирующих параметров. Тогда n-мерной проекцией случайного процесса называется случайный вектор $(f(x_1), \dots, f(x_n))$. Если все конечномерные проекции имеют нормальное распределение, то такая бесконечная совокупность случайных величин называется гауссовским случайным процессом:

$$\forall x_1, \dots, x_n : x_i \in \mathbb{R}^d \;\; p(f(x_1), \dots, f(x_n)) \sim \mathcal{N}(f(x_1), \dots, f(x_n) \,|\, \mu, \Sigma) \tag{223}$$

Отметим, что определение не гарантирует существование гауссовских процессов, но мы примем на веру, что они существуют.

Аналогично, как гауссовская случайная величина полностью определяется средним и дисперсией, гауссовский случайный процесс полностью определяется двумя своими параметрами: функцией среднего значения и ковариационной функцией.
По определению функция среднего значения есть:

$$m(x) \stackrel{\text{def}}{=} \mathbb{E} f(x) \tag{224}$$

Ковариационная функция:

$$K(x, x') \stackrel{\text{def}}{=} \operatorname{cov}(f(x), f(x')) \tag{225}$$

Ковариационная функция обладает 2 свойствами:

- Симметричность: $K(x, x') = K(x', x)$
- Неотрицательность: $\forall h \in L_2 \Rightarrow \iint K(x, x')h(x)h(x')dxdx' \geq 0$

Таким же свойством обладают ядровые функции, и в этом плане ковариационная функция похожа на них. Теперь рассмотрим, как выглядят параметры нормального распределения из 223:

$$\mu = (m(x_1), \dots, m(x_n))^T \tag{226}$$

$$\Sigma = \begin{pmatrix} K(x_1, x_1) & \dots & K(x_1, x_n) \\ \vdots & \ddots & \vdots \\ K(x_n, x_1) & \dots & K(x_n, x_n) \end{pmatrix} \tag{227}$$

Рассмотрим понятие стационарности случайного процесса.

Определение 17. Случайный процесс является стационарным в узком смысле, если

$$p(f(x_1 + \Delta x), \dots, f(x_n + \Delta x)) = p(f(x_1), \dots, f(x_n)) \; \forall \Delta x \in \mathbb{R}^d \tag{228}$$

Этим определением очень неудобно оперировать, поэтому, введем стационарность в широком смысле.

Определение 18. Случайный процесс является стационарным в широком смысле, если функция среднего значения константна, а ковариационная функция зависит только от разницы между индексирующими элементами

$$m(x) = const \tag{229}$$

$$K(x, x') = K(x - x') \tag{230}$$

Для гауссовских процессов эти определения стационарности эквивалентны. В дальнейшем будем изучать только стационарные гауссовские процессы, будем считать, не ограничивая общности, что $m(x) = 0$. Тогда гауссовский процесс полностью определяется своей ковариационной функцией.

Ковариационная функция обладает следующим полезным свойством.

Свойство. Если ковариационная функция непрерывна в нуле, то есть $\lim\limits_{\Delta x \to 0} K(\Delta x) = K(0)$, то почти все реализации случайного процесса являются непрерывными функциями.

Рассмотрим ещё одно свойство ковариационной функции стационарного процесса.

Свойство. Для любой ковариационной функции верно:

$$K(0) \geq |K(\Delta x)| \tag{231}$$

Чтобы доказать это, заметим, что функция ковариации удовлетворяет всем свойствам скалярного произведения. Тогда для нее выполнено неравенство Коши-Буняковского

$$|\langle X, Y \rangle| \leq \|X\| \cdot \|Y\| \tag{232}$$

То есть

$$|\operatorname{cov}(X, Y)| \leq \sqrt{\mathbb{D}X\mathbb{D}Y} \tag{233}$$

Для стационарного процесса дисперсии одинаковы и равны $K(0)$, из чего вытекает (231)

Из (231) следует, что значение ковариационной функции лежит в интервале $[K(0), -K(0)]$. Как правило, на бесконечности ковариационная функция стремится к 0. Пусть известно $f(x_1)$ и что K непрерывна в нуле. То есть $\mathbb{D}f(x_1) = K(0) = \mathbb{D}f(x)$. Рассмотрим, чему равен предел $\lim\limits_{x \to x_1} f(x)$. Из непрерывности в нуле функции ковариации следует

$$\lim_{x \to x_1} \operatorname{cov}(f(x), f(x_1)) = K(0) \tag{234}$$

Тогда рассмотрим совместное распределение

$$p(f(x), f(x_1)) = \mathcal{N}\left(f(x), f(x_1) \,|\, 0, \begin{pmatrix} K(0) & K(0) \\ K(0) & K(0) \end{pmatrix}\right) \tag{235}$$

Отсюда следует, что

$$\lim_{x \to x_1} p(f(x), f(x_1)) = \delta(f(x) - f(x_1)) \tag{236}$$

Из этого следует, что $f(x)$ и $f(x_1)$ точно совпадают, то есть $\lim\limits_{x \to x_1} = f(x_1)$. Таким образом, если ковариационная функция непрерывна в 0, то почти все реализации случайного процесса будут непрерывными функциями.

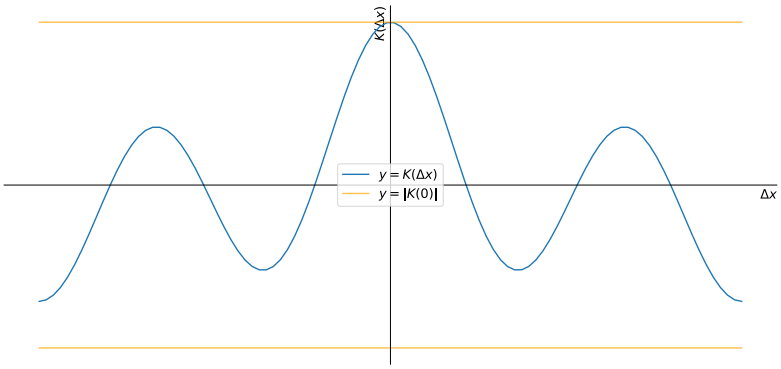


Рис. 17: Пример графика ковариационной функции

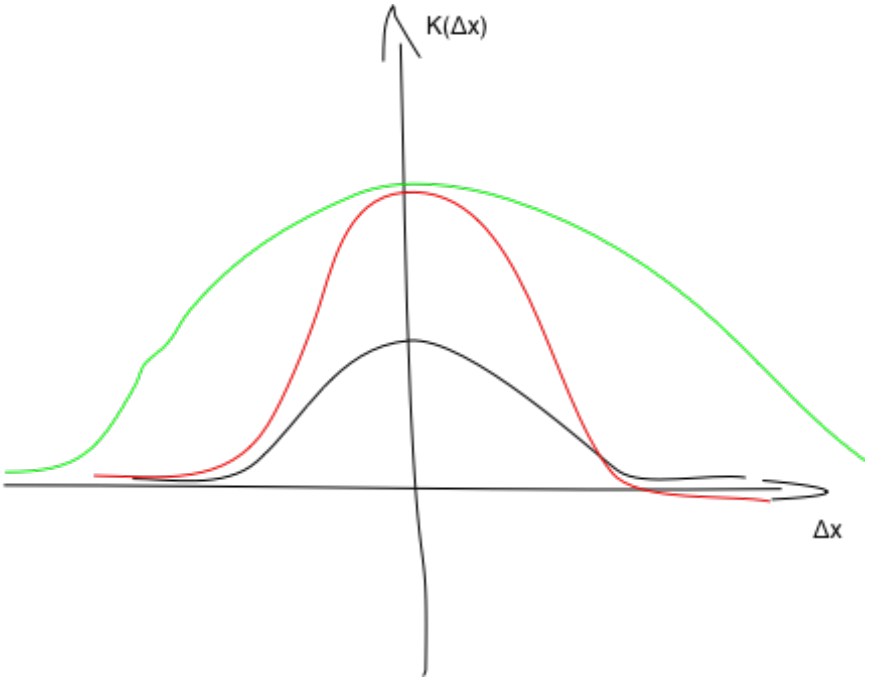


Рис. 18: Графики различных ковариационных функций

Рассмотрим несколько случайных процессов.

На рисунках (18) и (19) изображены несколько ковариационных функций и соответствующих им реализаций случайного процесса. Отметим 2 особенности: во-первых, высота «колокола», то есть значение $K(0)$ задает максимальное отклонение реализации от нуля; во-вторых, ширина «колокола» задает плавность изменение реализации, чем шире гаусси-ана ковариационной функции, тем больше ковариация соседних величин, тем более плавно изменяется реализация. Если ковариационная функция равна нулю везде, кроме точки $\Delta x = 0$, то случайные величины реализации процесса не будут скоррелированы, то есть, мы получим белый шум.

Давайте подумаем, что мы вообще хотим делать с помощью случайных процессов с точки зрения машинного обучения? Мы бы хотели, имея несколько реализаций точек из случайного процесса предсказывать следующие значения. Но если же у нас белый шум, то мы это сделать не сможем, так как у нас нет никакой зависимости между проекциями.

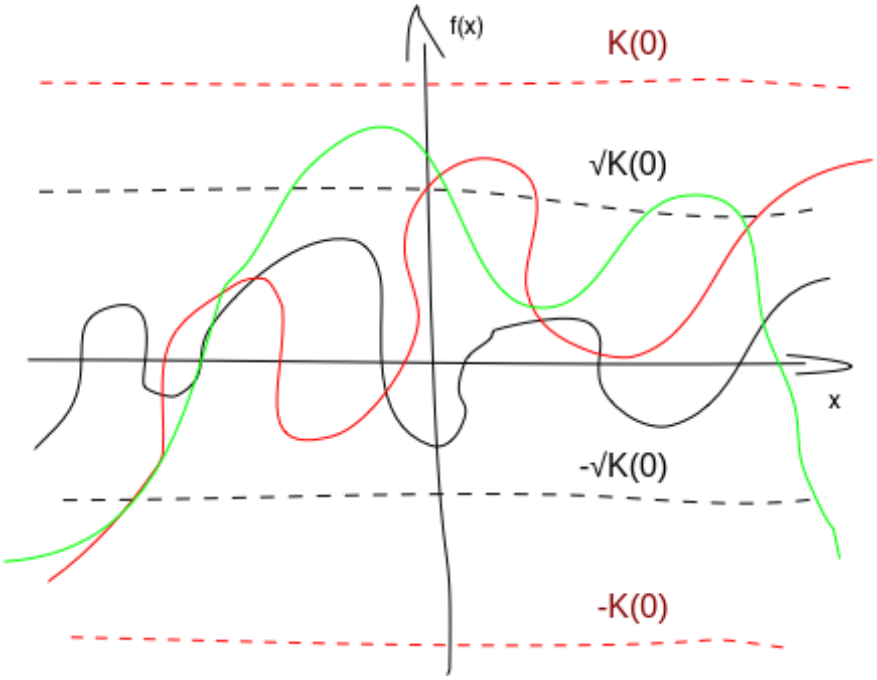


Рис. 19: Графики реализаций различных случайных процессов

11.3 Формула Андерсона

Пусть у нас есть стационарный гауссовский процесс, из которого пришли данные $F = (f(x_1), \dots f(x_n))^T$. Наша задача – построить предсказания в точках $y_1 \dots y_m$. Так как мы находимся в вероятностной постановке задачи, то более строго говоря нужно оценить вероятность для $\tilde{F} = (f(y_1), \dots f(y_m))^T$. То есть требуется найти $p(\tilde{F} \mid F)$. Для этого воспользуемся определением условной вероятности и тем фактом, что нормальное распределение самосопряженное:

$$p(\tilde{F} \mid F) = \frac{p(\tilde{F}, F)}{p(F)} = \mathcal{N}(\tilde{F} \mid \mu, \Sigma) \tag{237}$$

Как задаются параметры μ и Σ? Известно, что они определяются формулами Андерсона:

$$\mu = K^T C^{-1} F \tag{238}$$

$$\Sigma = S - K^T C^{-1} K \tag{239}$$

где матрицы $C \in R^{n \times n},\ K \in R^{n \times m},\ S \in R^{m \times m}$ определяются следующим образом:

$$C_{ij} = K(x_i, x_j),\ K_{ij} = K(x_i, y_j),\ S_{ij} = K(y_i, y_j) \tag{240}$$

Попробуем проинтерпретировать этот результат. Для этого положим $m = 1$. Тогда:

$$\mu = k^T C^{-1} F \tag{241}$$

$$\sigma^2 = K(0) - k^T C^{-1} k \tag{242}$$

Пусть ковариационная функция непрерывна в 0, тогда $\lim\limits_{y \to x_1} k(y)^T$ будет равен первой строке матрицы C, тогда:

$$\lim_{y \to x_1} k(y)^T C^{-1} = (1, 0, \dots, 0) \tag{243}$$

$$\lim_{y \to x_1} \mu(y) = f(x_1) \tag{244}$$

$$\lim_{y \to x_1} \sigma^2(y) = K(0) - k^T C^{-1} k = K(0) - K(0) = 0 \tag{245}$$

Теперь рассмотрим, что будет при $y \to \infty$. Положим для этого $K(\infty) = 0$, что верно для невырожденных случаев. В этом случае

$$\lim_{y\to\infty} \mu(y) = 0 \tag{246}$$

$$\lim_{y\to\infty} \sigma^2(x) = K(0) \tag{247}$$

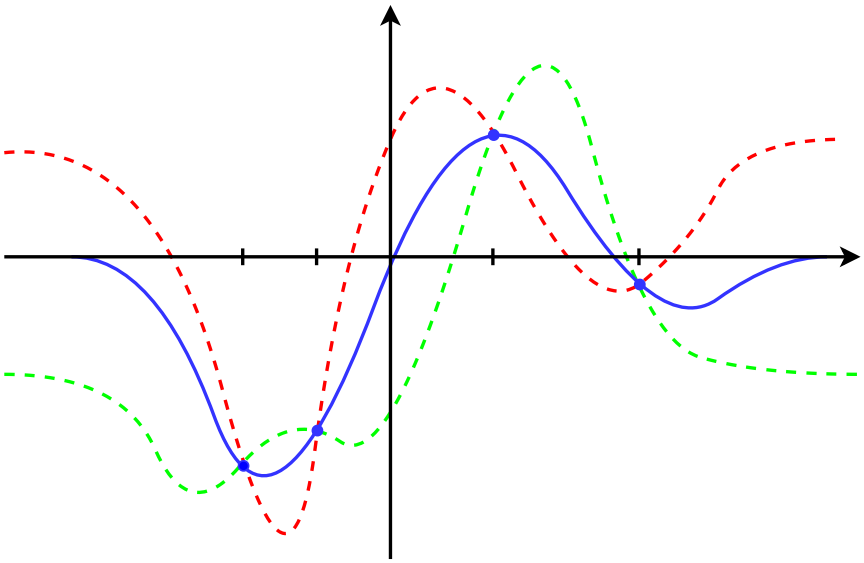


Рис. 20: График реализации процесса с изображенной дисперсией

На рисунке (20) изображена реализация процесса для решения конкретной задачи регрессии. В точках обучающей выборки $\mu(x_i) = f(x_i)$, а дисперсия в этих точках равна 0. Пунктиром отмечено значение дисперсии.

Мы научились строить предсказания. Давайте проанализируем модель. Во-первых, отметим, что сложность получения предсказаний $O(n^3)$, что делает невозможным применение метода на выборках большого объема. Этот недостаток пытаются обойти в различных исследованиях и на текущий момент. Во-вторых, встает вопрос выбора ядра. Действительно, мы любые данные можем описать с помощью белого шума, но будет ли это информативно? К счастью, с этой проблемой можно бороться с помощью принципа максимальной обоснованности.

11.4 Восстановление регрессии

Пусть теперь дана задача регрессии:

$$(X, T) = \{x_i, t_i\}_{i=1}^n, x \in \mathbb{R}^d, t \in \mathbb{R} \tag{248}$$

И имеется следующая вероятностная модель. Пусть $f(x)$ – стационарный гауссовский процесс с непрерывной в нуле ковариационной функцией. И верно:

$$t_i = f(x_i) + \varepsilon_i,\ \varepsilon_i \sim \mathcal{N}(\varepsilon_i \,|\, 0, \sigma^2) \tag{249}$$

Связи задаются в соответствии с рис. 21. Тогда выпишем модель, соответствующую обычным стрелкам:

$$p(\tilde{T} \,|\, T) = \iint p(\tilde{T} \,|\, \tilde{F}) p(\tilde{F} \,|\, F) p(F \,|\, T) d\tilde{F} dF \tag{250}$$

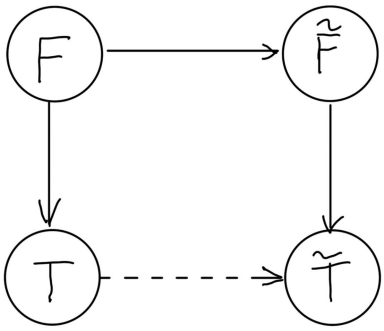


Рис. 21: Вероятностная модель

Рассмотрим каждый множитель по отдельности $p(\tilde{T} \mid \tilde{F}) = \mathcal{N}(t(x_0) \mid f(x_0), \sigma^2)$, $p(\tilde{F} \mid F)$ определяется формулой Андерсона, а для $p(F \mid T)$ воспользуемся формулой Байеса:

$$p(F \mid T) = \frac{p(T \mid F)p(F)}{\int p(T \mid F)p(F)dF} = \frac{p(T \mid F)p(F)}{p(T)} \tag{251}$$

Где $p(T) = \mathcal{N}(T \mid 0, \hat{C})$, $\hat{C}_{ij} = K(x_i - x_j) + \sigma^2[x_i = x_j]$. Заметим, что $p(T)$ – обоснованность, а значит, что, максимизируя ее, можно найти оптимальное ядро и уровень белого шума σ. Параметризуем K следующим образом (здесь θ_3 отвечает за σ):

$$K_\theta(x, x') = \theta_1 exp(-\theta_2 ||x - x'||^2) + \theta_3[x = x'] \tag{252}$$

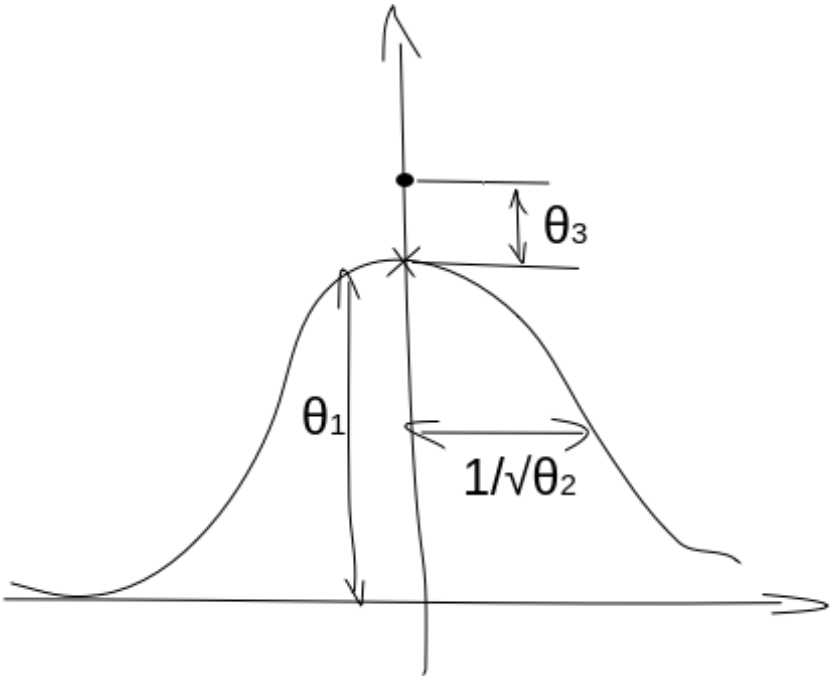


Рис. 22: Визуальное изображение параметров

Наконец, отметим, что если пойти не по сплошным стрелкам на рис. 21, а сразу по пунктирной, то получим, что получать предсказания можно по формуле Андерсона с новым ядром:

$$\hat{K}_T(x, x') = K_F(x, x') + \sigma^2[x = x'] \tag{253}$$

11.5 Гауссовские процессы для задачи классификации

Постановка задачи

$$(X, T) = \{x_i, t_i\}_{i=1}^n, x \in \mathbb{R}^d, t \in +1, -1 \tag{254}$$

Итак, в прошлом пункте мы выстроили очень удобную вероятностную модель: гауссовский процесс является скрытым, а наблюдаем мы только какие-то его косвенные величины, например, знак. Этот подход позволяет обобщить гауссовские процессы на задачу классификации. Для этого достаточно ввести:

$$p(T \mid F) = \prod_{i=1}^{n} \frac{1}{1 + exp(-t_i f(x_i))} \tag{255}$$

Аналогично с пунктом для регрессии нужно найти $p(F \mid T)$:

$$p(F \mid T) = \frac{p(T \mid F)p(F)}{\int p(T \mid F)p(F)dF} \tag{256}$$

В этот раз интеграл не берется аналитически, приблизим его ненормированной гауссианой и воспользуемся приближением Лапласса (см. Лекцию 5), $Q(F) = p(T \mid F)p(F)$:

$$\log Q(F) \approx \log Q(F_{MP}) + \frac{1}{2}(F - F_{MP})^2 \nabla^2 \log Q(F_{MP})(F - F_{MP})$$

Значит:

$$p(F \mid T) \approx \frac{p(T \mid F)p(F)}{\int Q(F_{MP}) \exp\left(\frac{1}{2}(F - F_{MP})^2 \nabla^2 \log Q(F_{MP})(F - F_{MP})\right) dF} = \mathcal{N}(F \mid \cdot, \cdot)$$

Этот интеграл берется аналитически. Итого для вычисления $p(\tilde{T} \mid T)$ нужно посчитать интеграл от произведения сигмоид и гауссианы. Как вычислять интегралы такого типа мы разбирали на семинаре по RVR.

В современных реалиях гауссовские процессы используются в задачах байесовской оптимизации, когда с помощью них подбираются гиперпараметры. Коммьюнити неоднозначно относится к этой идее, так как нет обоснований, почему гауссовские процессы лучше случайного блуждания.

12 Лекция 12. Процессы Дирихле

Процессы Дирихле являются еще одним способом строить непарметрические байесовские модели. В частности, в этой лекции мы построим модель с латентными переменными, в которой номенклатура самих латентных переменных не фиксирована и может видоизменяться при получении новых данных. В первую очередь техника процессов Дирихле необходима для автоматического определения числа кластеров в данных и при необходимости это количество увеличивать. Например на Рис. 23 при наблюдении 5 объектов можно выделить 1, 2, 3 или 5 кластеров, но при увеличении выборки четко выделяются 4 кластера, при этом увеличивая выборку число может снова поменяться. Получается, что если данные приходят динамически, то, кажется не разумно, априори фиксировать число кластеров. Тем самым мы получили необходимость в создании модели, которая адаптировала бы структуру латентных переменных (например, количество кластеров) по мере поступления дополнительных данных.

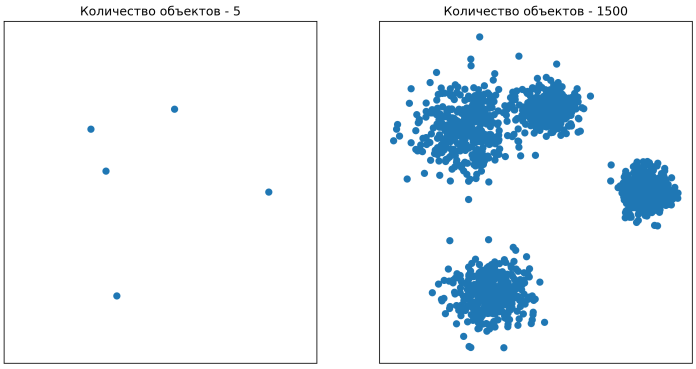


Рис. 23: Пример зависимости структуры данных от размера выборки

12.1 Предварительные сведения

Пусть $x \sim Cat(x \mid \boldsymbol{\theta})$, где $\boldsymbol{\theta} \in \mathcal{S}_k$ (напоминание $\mathcal{S}_k := \{\boldsymbol{\theta} \in \mathbb{R}^k \mid \sum\limits_{i=1}^{k} \theta_i = 1,\ \theta_i \geq 0,\ 1 \leq i \leq k\}$ — вероятностный симплекс). Понятно, что $Cat(x \mid \boldsymbol{\theta}) = \prod\limits_{i=1}^{k} \theta_i^{[x=i]}$, а распределение Дирихле является сопряженно априорным, плотность которого задается следующей формулой:

$$Dir(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \quad \boldsymbol{\theta} \in \mathcal{S}_k, \quad \boldsymbol{\alpha} = \mathbb{R}^k,\ \alpha_i > 0$$

Пусть мы наблюдаем $X = (x_1, \ldots, x_n)$ из категориального распределения, тогда получаем:

$$p(X, \boldsymbol{\theta}) = p(X \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{j=1}^{k} \theta_i^{[x_i = j]} \cdot \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \tag{257}$$

Тогда по формуле 257 получим:

$$p(X) = \int p(X \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int \prod_{i=1}^{n} p(x_i \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{258}$$

В силу того, что распределение Дирихле является сопряженным к категориальному распределению, мы можем аналитически найти интеграл в выражении 258. При этом при

маргинализации мы теряем независимость объектов, то есть $p(X) \neq \prod_{i=1}^{n} p(x_i)$. Распишем по правилу произведения:

$$p(X) = p(x_1)p(x_2 \,|\, x_1) \dots p(x_n \,|\, x_1, \dots, x_n)$$

Тогда получим (в качестве упражнения для читателя) выражение 259.

$$p(x_n = l \,|\, x_1, \dots, x_{n-1}) = \frac{\nu_l^{n-1} + \alpha_l}{n - 1 + \sum_{i=1}^{k} \alpha_i} \tag{259}$$

$$\nu_l^n := \sum_{i=1}^{n} [x_i = l]$$

Физический смысл выражения 259 неформально можно сформулировать как «богатый становится ещё богаче», вероятность увидеть конкретное значение тем больше, чем чаще оно уже наблюдалось. Часто данный принцип называют «Китайский ресторан» (Chinese restaraunt) — новый посетитель вероятнее выбирает тот столик, за которым сидят больше людей, то есть $p(x_n = l \,|\, x_1, \dots, x_{n-1})$ выше, чем больше ν_l^{n-1}.

Формула 259 также связана с одним из способов сэмплирования из распределения Дирихле, поскольку для $x_n \sim p(x_n \,|\, x_1, \dots, x_{n-1})$ вектор $(\frac{\nu_l^n}{n})_{l=1}^k$ в пределе имеет распределение $Dir(\boldsymbol{\theta} \,|\, \boldsymbol{\alpha})$.

12.2 Свойства распределения Дирихле

Для распределения Дирихле выполнено следующее свойство, которое называется свойством накопления (**aggregation property**):

$$\begin{aligned}&\text{Если } (\theta_1, \dots, \theta_k) \sim Dir(\theta_1, \dots, \theta_k \,|\, \alpha_1, \dots, \alpha_k),\\ &\text{то } (\theta_1 + \theta_2, \theta_3, \dots, \theta_k) \sim Dir(\theta_1 + \theta_2, \theta_3, \dots, \theta_k \,|\, \alpha_1 + \alpha_2, \alpha_3, \dots, \alpha_k)\end{aligned}$$

С помощью свойства накопления можно вывести:

$$\begin{aligned}&\text{Если } (\theta_1, \dots, \theta_k) \sim Dir(\theta_1, \dots, \theta_k \,|\, \alpha_1, \dots, \alpha_k),\\ &\text{то } (\theta_1, \sum_{i=2}^{k} \theta_i) = (\theta_1, 1 - \theta_1) \sim Beta(\theta_1 \,|\, \alpha_1, \sum_{i=2}^{k} \alpha_i)\end{aligned}$$

Докажем свойство независимости (**independence property**), для этого распишем вероятность $p(\theta_2, \dots, \theta_k \,|\, \theta_1)$, тогда получим по определению:

$$p(\theta_2, \dots, \theta_k \,|\, \theta_1) = \frac{p(\theta_1, \dots, \theta_k)}{p(\theta_1)} = \frac{Dir(\theta_1, \dots, \theta_k \,|\, \alpha_1, \dots, \alpha_k)}{Beta(\theta_1 \,|\, \alpha_1, \sum_{i=2}^{k} \alpha_i)} =$$

$$= \frac{1}{\mathbf{Z}} \frac{\prod_{i=1}^{k} \theta_i^{\alpha_i - 1}}{\theta_1^{\alpha_1 - 1}(1 - \theta_1)^{\sum_{i=2}^{k} \alpha_i - 1}} = \frac{1}{\mathbf{Z}'} \prod_{i=2}^{k} \left(\frac{\theta_i}{1 - \theta_1} \right)^{\alpha_i - 1} \tag{260}$$

По формуле 260, получаем:

$$\begin{aligned}&\text{Если } (\theta_1, \dots, \theta_k) \sim Dir(\theta_1, \dots, \theta_k \,|\, \alpha_1, \dots, \alpha_k), \quad \xi_i := \frac{\theta_i}{1 - \theta_1}, \quad 2 \leq i \leq k,\\ &\text{то } (\xi_2, \dots, \xi_k) \sim Dir(\xi_2, \dots, \xi_k \,|\, \alpha_2, \dots, \alpha_k)\end{aligned}$$

Свойство независимости имеет следующей смысл: наблюдение одной компоненты распределения Дирихле никакой информации о значении других компонент не дает.

Используя независимость компонент распределения Дирихле, можно получить метод сэмплирования 3 под названием «Ломка палки» (**Stick-breaking**) [35].

Algorithm 3 «Ломка палки» (Stick-breaking)

Require: $\alpha_1, \dots, \alpha_k$

$v_1 \sim Beta(v_1 \,|\, \alpha_1, \sum_{i=2}^{k} \alpha_i)$

$\theta_1 = v_1$

$N \leftarrow n$

for $l = \overline{2, k-1}$ do

$v_l \sim Beta(v_l \,|\, \alpha_l, \sum_{i=l+1}^{k} \alpha_i)$

$\theta_l = v_l \prod_{i=1}^{l-1}(1 - v_i)$

end for

$\theta_k = 1 - \sum_{i=1}^{k-1} \theta_i$

12.3 Свойства распределения Дирихле (более формально)

Пусть $\theta \in S_k$, $\theta \sim Dir(\theta \,|\, \alpha)$. Рассмотрим сужение $\tilde{\theta} \in S_m, m < k$. Скажем $\tilde{\theta}_j = \sum_{i \in I_j} \theta_i$, где $I_1, \dots, I_M$ — разбиение множества $\{1, \dots, k\}$.

Свойство агрегации:

$$\tilde{\theta} \sim Dir(\tilde{\theta} \,|\, \tilde{\alpha}_1, \dots, \tilde{\alpha}_M), \quad \tilde{\alpha}_j = \sum_{i \in I_1} \alpha_i.$$

Свойство независимости:

$$\frac{\tilde{\theta}_{-j}}{1 - \tilde{\theta}_j} \sim Dir(\tilde{\theta}_{-j} \,|\, \tilde{\alpha}_{-i}), \quad \frac{\tilde{\theta}_{-j}}{1 - \tilde{\theta}_j} \perp \tilde{\theta}_j.$$

Свойство агрегации несложно доказать. Во-первых, вспомним связь распределения Дирихле с гамма-распределением: если $z_i \sim \text{Gamma}(z_i \,|\, \alpha_i, 1)$, $i = \overline{1, K}$, то $(z_i / z_0)_{i=1}^{K} \sim \text{Dir}(\boldsymbol{\alpha})$, где $z_0 = \sum_{i=1}^{K} z_i$. Во-вторых, вспомним свойство гамма-распределения: если $x \sim \text{Gamma}(x \,|\, \alpha, 1)$, $y \sim \text{Gamma}(y \,|\, \beta, 1)$, то $x + y \sim \text{Gamma}(x + y \,|\, \alpha + \beta, 1)$.

Идея доказательства связи распределения Дирихле с гамма-распределением следующая: достаточно рассмотреть переход $(z_1, \dots, z_K) \mapsto (z_0, \theta_1, \dots, \theta_{K-1})$ и известное свойство $p(\theta) = p(z) \det J^{-1}$, где J — матрица Якоби соответствующего функционального преобразования.

12.4 Процессы Дирихле

Введем процессы Дирихле по аналогии с процессом Гаусса и Пуассона в таблице 6 сравниваются характеристики процесса Дирихле с уже знакомыми процессами. Подробнее о значении всех величин поговорим ниже.

Для определения процесса Дирихле нам понадобится некоторое универсальное множество U (например, $U = \mathbb{R}^d$), на котором задана вероятностная мера G_0, которое назовём базовым распределением, а также положительное число $\alpha > 0$ — коэффициент концентрации.

Из таблицы 6 и свойств Бета-распределения мы можем понять, какой смысл имеют параметры процесса Дирихле.

[35] Такой подход носит названия «ломки палки» из-за аналогии с отламыванием кусочков длины θ_i от палки длины 1.

Процесс	Гауссовский	Пуассоновский	Дирихле
Индексирующий элемент	$x \in \mathbb{R}^d$	$t \in \mathbb{R}_+$	Измеримое подмножество $A \subseteq U$
Реализация сл. п. Рис. 24	Функция $f(\boldsymbol{x})$ на $\mathbb{R}^d$	Последовательность поступающих событий $u(t)$	Вероятностная мера на U
Параметры	$m(x),\ K(x, x')$	$\lambda(t)$ — интенсивность	G_0 — вероятностная мера на U; $\alpha > 0$ — коэффициент концентрации
Одномерная проекция	$f(\boldsymbol{x_0}) \sim \mathcal{N}\big(m(\boldsymbol{x_0}), K(\boldsymbol{x_0}, \boldsymbol{x_0})\big)$	$u(t_0) \sim Pois(\lambda(t_0))$	$\xi(A_0) \sim Beta(\alpha G_0(A_0), \alpha(1 - G_0(A_0)))$
Многомерная проекция	$(f(\boldsymbol{x_1}), ..., f(\boldsymbol{x_n})) \sim \mathcal{N}(\mu, \Sigma)$	$u(t_1), ..., u(t_n) \sim Pois(\lambda t_1) \cdot Pois(\lambda(t_2 - t_1)) \cdot ... \cdot Pois(\lambda(t_n - t_{n-1}))$	Для разбиения $A_1, ..., A_n$: $\xi(A_1), ..., \xi(A_n) \sim Dir(\alpha G_0(A_1), ..., \alpha G_0(A_n))$

Таблица 6: Сравнение случайных процессов

Параметр G_0 имеет смысл математического ожидания для одномерных проекций, то есть $\mathbb{E}\xi(A_0) = G_0(A_0)$. Данный факт следует из формулы 261 для математического ожидания для Бета-распределения.

$$\begin{aligned} &\xi \sim Beta(\xi \,|\, \mu\nu, \nu(1-\mu)) \\ &\mathbb{E}\xi = \mu \end{aligned} \tag{261}$$

Параметр α влияет на U-образность распределения, что продемонстрировано на Рис. 25. Можем считать, что при маленьких α вероятностная масса сконцентрирована в одной точке, а при увлечении α вероятность более равномерно распределена по U. Примеры реализаций процесса Дирихле в зависимости от концентрации представлены на Рис. 24, видно что при увлечении α число атомов возрастает.

Понятно, что мы не дали формального определения процессов Дирихле, а лишь ввели их по аналогии с уже изученными процессами, что и отражено в таблице 6. Вообще говоря из наших рассуждений не следует, что такие процессы Дирихле существуют и для этого необходимо доказывать теорему существования, однако мы примем этот факт на веру.

Напомним, что разбиением множества U называется набор $A_1, \dots A_n$, такой что $\bigcup\limits_{i=1}^{n} A_i = U,\ A_i \cap A_j = \emptyset,$ при $i \neq j$. Из определения многомерной проекции, а именно для разбиения любого $A_1, ..., A_n$ множества U выполнено $\xi(A_1), ..., \xi(A_n) \sim Dir(\alpha G_0(A_1), ..., \alpha G_0(A_n))$, можно вывести реализация процесса Дирихле является атомарной вероятностной мерой. Под атомарностью мы понимаем, что вероятностная масса сконцентрированна в некоторых точках и не является непрерывной. Этот факт сложно доказать строго, но интуитивно это следствие свойства независимости распределения Дирихле, а именно, что наблюдение одной компоненты не дает знания о соотношениях других компонент. Действительно, если бы получаемая вероятностная мера была бы непрерывной, то мы бы смогли подобрать множества в разбиении так, что знание вероятности на одном из них давало бы знание о

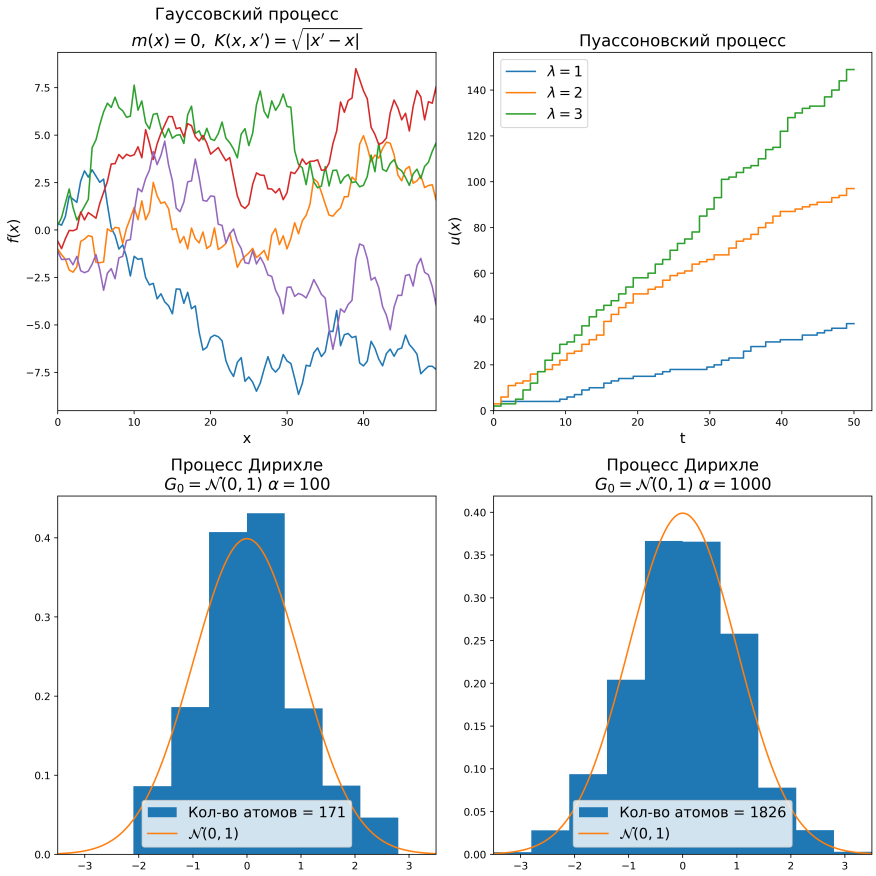


Рис. 24: Зависимость реализаций случайных процессов от параметров

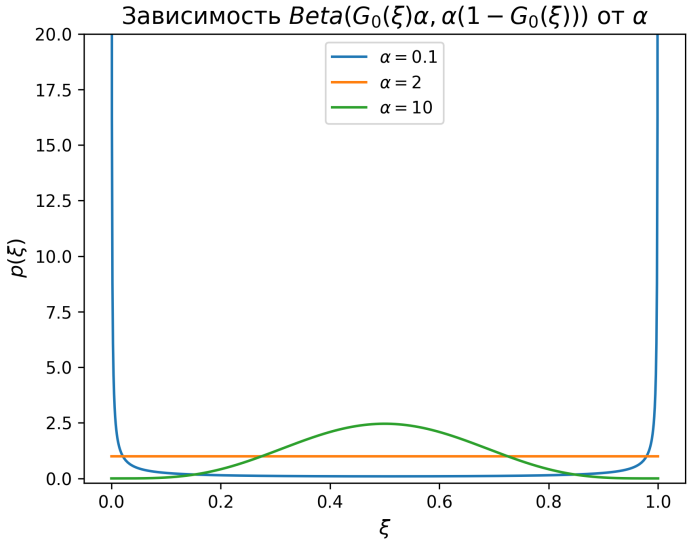


Рис. 25: Зависимость распределения одномерной проекции от параметра α, в данном примере $U = \mathbb{R}$

другом множестве в силу непрерывности функции распределения. Кроме того, количество носителей вероятностной меры бесконечно, но счетно.

12.5 Генерация реализации процесса Дирихле

Методы получения реализации процесса Дирихле очень похожи на методы сэмплирования из распределения Дирихле.

12.5.1 Процесс «Китайский ресторан» (**Chinese restaraunt process, CRP**)

Аналогично с методом сэмплирования основанном на выражении 259 мы можем получить способ генерации реализации процесса с заданными α, G_0.

Схема описанная в алгоритме 4 называется урновой схемой Блэквела-МакКвина. Как можно заметить, в данной схеме происходит разделение точек на группы (точки x_i попадают в одну группу θ_j).

Пусть k_* — количество кластеров (столов в ресторане). Новый посетитель (x_{n+1}) либо подсаживается за стол k с вероятностью $\frac{\nu_k^n}{\alpha+n}$ (посетители более предпочитают столы с уже большой компанией), либо садится за новый с вероятностью $\frac{\alpha}{\alpha+n}$, тем самым увеличивая количество кластеров (столов) на 1, таким образом, параметр α отвечает за интровертность людей. Иногда число кластеров (столов) растет слишком медленно, для этого существуют расширения процесса Дирихле (процесс Питмана-Йора) с целью изменения этого свойства.

Данный алгоритм должен сделать бесконечное число итераций и $\pi_k = \lim_{n\to\infty} \frac{\nu_k^n}{n}$. На практике критерием останова является маленькие значение $\frac{\alpha}{n+\alpha}$, а предельный переход заменяют равенством.

В результате работы алгоритм получает атомы реализуемого вероятностного распределения — θ_i, вероятности соответствующих атомов — π_i.

Algorithm 4 Процесс «Китайский ресторан»

$k_* = 0$
while $\frac{\alpha}{n+\alpha} \geq \varepsilon$ do
$$x_{n+1} = \begin{cases} k, \text{ с вероятностью } \frac{\nu_k^n}{n+\alpha} \\ k_* + 1, \text{ с вероятностью } \frac{\alpha}{n+\alpha}, \quad k_* = k_* + 1, \quad \theta_{k_*} \sim G_0 \end{cases}$$
end while
$\pi_k = \frac{\nu_k^n}{n}$

12.5.2 Процесс «Ломки палки» (**Stick-breaking process**)

Аналогично алгоритму сэмплирования 3 мы получим метод генерации реализации процесса Дирихле, который описан в алгоритме 5.

В результате работы алгоритм получает атомы реализуемого вероятностного распределения — θ_i, вероятности соответствующих атомов — π_i.

Algorithm 5 Процесс «Ломка палки»

for $k = \overline{1, \infty}$ do
$\theta_k \sim G_0$
$v_k \sim Beta(v_k \,|\, 1, \alpha)$
$\pi_k = v_k \prod_{i=1}^{k-1} (1 - v_i)$
if $\sum_{i=1}^{k} \pi_k > 1 - \varepsilon$ then
exit
end if
end for

12.5.3 Переход от процесса к распределению.

Как было сказано, процессы генерации сходятся. Это означает, что число k_* стабилизируется при фиксированном объеме наблюдаемых данных. После сходимости можно перейти от процесса Дирихле к распределению Дирихле. Тогда можно использовать полученные параметры Z для инференса на новых объектах.

Пусть $\pi \sim \mathrm{Dir}(\pi \,|\, \alpha)$, $z \sim \mathrm{Cat}(\pi)$. Распределение компонент:

$$\begin{aligned} p(Z) &= \int p(Z \,|\, \pi) p(\pi) d\pi \\ &= \int \prod_{i=1}^{N} p(z_i \,|\, \pi) p(\pi) d\pi \\ &= \int \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{[z_n = k]} \frac{\Gamma(\alpha K)}{\Gamma(\alpha)^K} \prod_{k=1}^{K} \pi_k^{\alpha - 1} d\pi \\ &= \frac{\Gamma(\alpha K)}{\Gamma(\alpha)^K} \int \prod_{k=1}^{K} \pi_i^{\sum_{n=1}^{N} [z_n = k] + \alpha - 1} d\pi. \end{aligned}$$

Можно догадаться, чему равен интеграл в полученном выражении: за счет того, что $p(z \,|\, \pi)$ и $p(\pi)$ сопряжены, под интегралом плотность распределения Дирихле со следующими параметрами: $\widehat{\alpha}_k = \alpha + \sum_{n=1}^{N} [z_n = k]$. Тогда имеем:

$$p(Z) = \frac{\Gamma(\alpha K)}{\Gamma(\alpha)^K} \frac{\prod_{k=1}^{K} \Gamma(\alpha + \sum_{n=1}^{N} [z_n = k])}{\Gamma(\alpha K + N)} \propto \prod_{k=1}^{K} \Gamma(\alpha + \nu_k^N),$$

где $\nu_k^N = \sum_{n=1}^{N} [z_n = k]$. Видим, что с помощью полученной формулы нельзя выразить $p(z_i)$ для i-го объекта независимо от остальных, поскольку нельзя получить факторизацию $p(Z) = \prod_i p(z_i)$. Поэтому можем выразить только условную вероятность:

$$\begin{aligned} p(z_N = k \,|\, z_1, \ldots, z_{N-1}) &= \frac{p(z_N = k, z_1, \ldots, z_{N-1})}{\sum_{l=1}^{K} p(z_N = l, z_1, \ldots, z_{N-1})} \\ &= \frac{\left(\prod_{i:\ i \neq k} \Gamma(\alpha + \nu_i^{N-1}) \right) \Gamma(\alpha + \nu_k^{N-1} + 1)}{\sum_{l=1}^{K} \left(\prod_{i:\ i \neq l} \Gamma(\alpha + \nu_i^{N-1}) \right) \Gamma(\alpha + \nu_l^{N-1} + 1)} \end{aligned}$$

В силу свойства $\Gamma(a+1) = \Gamma(a)a$ можем записать $\Gamma(\alpha + \nu_k^{N-1} + 1) = \Gamma(\alpha + \nu_k^{N-1})(\alpha + \nu_k^{N-1})$:

$$\begin{aligned} p(z_N = k \,|\, z_1, \ldots, z_{N-1}) &= \frac{\left(\prod_{i:\ i \neq k} \Gamma(\alpha + \nu_i^{N-1}) \right) \Gamma(\alpha + \nu_k^{N-1})(\alpha + \nu_k^{N-1})}{\sum_{l=1}^{K} \left(\prod_{i:\ i \neq l}^{K} \Gamma(\alpha + \nu_i^{N-1}) \right) \Gamma(\alpha + \nu_l^{N-1})(\alpha + \nu_l^{N-1})} \\ &= \frac{\left(\prod_{i=1}^{K} \Gamma(\alpha + \nu_i^{N-1}) \right) (\alpha + \nu_k^{N-1})}{\sum_{l=1}^{K} \left(\prod_{i=1}^{K} \Gamma(\alpha + \nu_i^{N-1}) \right) (\alpha + \nu_l^{N-1})} \\ &= \frac{\alpha + \nu_k^{N-1}}{\sum_{l=1}^{K} (\alpha + \nu_l^{N-1})} \\ &= \frac{\alpha + \nu_k^{N-1}}{\alpha K + N - 1} \end{aligned}$$

Почти пришли к цели! Вспомним алгоритм 4 генерации случайного процесса Дирихле. В нем для каждого исхода $1, \ldots, k_*, k_* + 1$ были вероятности вида $\circ / (n + \alpha)$. Поэтому для «бесшовного» перехода от процесса Дирихле к распределению Дирихле в последнем достаточно взять $\overline{\alpha} = \alpha / k_*$, т.е. если z — номер компоненты нового объекта, то

$$p(z = k \,|\, Z) = \frac{\overline{\alpha} k_* + \nu_k^N}{\overline{\alpha} + N}.$$

12.6 Разделение смеси распределенний

Рассмотрим стандартную задачу разделения смеси распределений с байесовской моделью 262. Модель описывает смесь распределений для множества наблюдений $(x_1, \dots, x_N)$ со скрытыми переменными $(z_1, \dots, z_N)$, ответственными за принадлежность каждого объекту к определённому кластеру, параметрами для каждого кластера $(\theta_1, \dots, \theta_K)$ и весами кластеров $(\pi_1, \dots, \pi_K)$

$$p(X, Z, \boldsymbol{\theta}, \boldsymbol{\pi}) = p(X \mid Z, \boldsymbol{\theta}) p(Z \mid \boldsymbol{\pi}) p(\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{n} p(x_i \mid z_i, \boldsymbol{\theta}) p(z_i \mid \boldsymbol{\pi}) \cdot p(\boldsymbol{\theta}, \boldsymbol{\pi})$$

$$p(X, Z, \boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{N} \prod_{k=1}^{K} p(x_i \mid \theta_k)^{[z_i=k]} \pi_k^{[z_i=k]} \cdot p(\boldsymbol{\theta}, \boldsymbol{\pi}) \tag{262}$$

Если же мы сделаем параметр K настраиваемым, то при решении задачи $p(X, Z, \boldsymbol{\theta}, \boldsymbol{\pi}) \to \max_K$, мы получим что максимум достигается при $K = N$, а дисперсия каждой смеси нулевая. Понятно, что такое решение не имеет никакого смысла, поэтому нам необходимо добавить какие-либо ограничения, например, регуляризацию на параметр K, в качестве которой может быть процесс Дирихле. Получаем модель 263, в которой $DP(G_0, \alpha)$ - процесс Дирихле с параметрами G_0 (отвечает за априорные предпочтения параметров смеси) и α (чем больше, тем больше кластеров мы получаем). При этом при поступлении новых данных число кластеров может расти, тем самым сложность модели меняется в зависимости от данных.

$$p(X, Z, \boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{N} \prod_{k=1}^{\infty} p(x_i \mid \theta_k)^{[z_i=k]} \pi_k^{[z_i=k]} \cdot p(\boldsymbol{\theta}, \boldsymbol{\pi}), \quad p(\boldsymbol{\theta}, \boldsymbol{\pi}) := DP(G_0, \alpha) \tag{263}$$

и хотим найти $p(Z, \boldsymbol{\theta}, \boldsymbol{\pi} \mid X)$, однако аналитическую формулу вывести сложно, поэтому будем использовать приближенные способы: коллапсированная схема Гиббс, вариационный Stick-breaking.

12.6.1 Коллапсированная схема Гиббса

Пусть мы наблюдаем $X = (x_1, \dots, x_N)$. Нашей целью является получение оценки на параметры $Z, \boldsymbol{\theta}, \boldsymbol{\pi}$ с помощью апостериорного распределения $p(Z, \boldsymbol{\theta}, \boldsymbol{\pi} \mid X)$. Вывод и использование аналитической формулы достаточно трудные, поэтому вместо получения апостериорного распределения мы просэмплируем из него с помощью методов MCMC. Прогретая марковская цепь даст сэмплы из апостериорного распределения, которые мы будем использовать в качестве оценки на параметры. Известно, что методы Монте-Карло работают медленно, поэтому хотелось бы уменьшить размерность пространства сэмплирования, тем самым ускорив метод. Из определения модели 262, мы можем нам достаточно просто провести маргинализацию по π, так как $p(X, Z, \boldsymbol{\theta}) = \int p(X, Z, \boldsymbol{\theta}, \boldsymbol{\pi}) d\boldsymbol{\pi}$. Это позволит нам перейти проводить сэмплировать не $p(Z, \boldsymbol{\theta}, \boldsymbol{\pi} | X)$, а $p(Z, \boldsymbol{\theta} | X)$, тем самым уменьшив число переменных для сэмплирования и ускорив прогрев цепи.

В данной секции приведён метод сэмплирования Гиббса.

Напоминание: в методе Гиббса параметры сэмплируются последовательно по одному. То есть необходимо уметь сэмплировать из одномерных распределений $p(z_i \mid Z_{-i}, \boldsymbol{\theta}, X)$ и $p(\theta_i \mid \boldsymbol{\theta}_{-i}, Z, X)$, где A_{-i} означает $\{A_1, \dots, A_{i-1}, A_{i+1}, \dots, A_n\}$.

Шаг 0. Общее число компонент $k_* := 0$.

Шаг 1. Сэмплирование Z. Пусть число объектов, принадлежащих k-ой компоненте, без учёта i-го объекта:

$$\nu_k^{-i} = \sum_{z_j \in Z_{-i}} [z_j = k]$$

Тогда величина $z_i \sim p(z_i \mid Z_{-i}, \boldsymbol{\theta}, X)$ сэмплируется следующим образом:

$$z_i = \begin{cases} k \text{ с вероятностью, пропорциональной } p(x_i \mid \theta_k)\nu_k^{-i}, \quad k = \overline{1, k_*}, \\ k_* + 1 \text{ с вероятностью, пропорциональной } \alpha \int p(x_i \mid \theta) G_0(\theta) d\theta. \end{cases} \tag{264}$$

Последний случай означает добавление новой компоненты. В этом случае обновляем счётчик $k_* := k_* + 1$ и сэмплируем для новой компоненты её параметры $\theta_{k_*} \sim G_0$.

Шаг 2. Удаление всех k, для которых $\nu_k^N = 0$ (нет $z_i = k$), переиндексация оставшихся k с 1 до K'.

Шаг 3. Сэмплирование $\boldsymbol{\theta}$. Каждый θ_k пересчитывается, исходя из того, какие объекты x_i попали в k-ую компоненту. Для этого нужно просэмплировать из следующего распределения:

$$p(\theta_k \mid Z, \boldsymbol{\theta}_{-k}, X) \propto \prod_{i:\ z_i = k} p(x_i \mid \theta_k) G_0(\theta_k).$$

У этой вероятности есть интерпретация: значение θ_k тем вероятнее, чем больше все x_i соответствуют k-ой компоненте плюс регуляризация от априорного распределения $G_0(\theta_k)$. Распределения $p(x_i \mid \theta), G_0$ удобно выбирать сопряженными.

Шаги 1-3 составляют одну итерацию схемы Гиббса. Сходимость на практике быстрая. При этом вспомним, что среди параметров еще фигурирует $\boldsymbol{\pi}$. Вместо того чтобы сэмплировать вместе с $Z, \boldsymbol{\theta}$ ещё и $\pi_1, \ldots, \pi_{k_*}$, можно использовать оценку $\pi_k \approx \nu_k^N / N$ — число объектов, отнесённых к k-ой компоненте.

Схема сэмплирования Гиббса применяется непосредственно для маргинализованного (коллапсированного) апостериорного распределения $p(Z, \boldsymbol{\theta} \mid X) = \int p(Z, \boldsymbol{\theta}, \boldsymbol{\pi} \mid X) d\boldsymbol{\pi}$. Благодаря маргинализации мы уменьшили размерность пространства, в котором сэмплируем, тем самым ускорили алгоритм, что важно для MCMC в силу их низкой скорости работы.

12.6.2 Гастрономическая интерпретация

Как упоминалось, второе название приведённого алгоритма — схема «китайский ресторан». Чтобы лучше осознать и запомнить, вот её интерпретация:

Шаг 0. Изначально в китайском ресторане все столики пусты ($k_* := 0$), нет ни одного посетителя.

Шаг 1. Заходят N посетителей, каждый из них по очереди выбирает номер столика z_i, за который сесть. Если посетитель выбирает пустой столик (как, например, всегда делает самый первый посетитель), то число непустых столиков увеличивается ($k_* := k_* + 1$) и на столик приносят заказанную еду θ_k из G_0. Можно интерпретировать G_0 как цену блюда, то есть чем дороже еда, тем менее вероятно ресторан поставит ее на стол. При этом выбор $z_i = k$ тем вероятнее, чем больше ν_k^{-i} — число людей, сидящих за этим столиком, и чем больше $p(x_i \mid \theta_k)$ — соответствие еды на столике θ_k вкусовым предпочтениям x_i.

Шаг 2. Все столики без посетителей ($\nu_k^N = 0$) убирают, а оставшиеся столики перенумеруют от 1 до текущего k_*.

Шаг 3. На каждом столике меняют еду θ_k в соответствии со вкусовыми предпочтениями людей $\prod_{i:z_i=k} p(x_i \mid \theta_k)$, сидящих за ним, и в соответствии с ценой на блюда $G_0(\theta_k)$. После шага 3 все посетители по очереди меняют столик, т.е. переходят к шагу 1.

Данная интерпретация показывает смысл регуляризации. Каждый посетитель выбирает: садиться ли ему за новый стол или присоединиться к кому-то. Если бы все блюда стоили одинаково, то есть $G_0(\cdot) = \text{const}$, то новый посетитель бы искал свое самое любимое блюдо, а если бы не находил, то садился бы за новый стол. Новый стол стоит выбрать, так как все цены одинаковые и ресторану не важно какую еду готовить, то есть можно попытать удачу в надежде на любимую еду. Если же $G_0(\cdot) \neq \text{const}$, и посетитель любит очень дорогое блюдо, то даже не найдя его, он сядет за менее привлекательный стол к более шумной компании вместо одиночества, так как скорее всего ресторан не вынесет «черную икру» ради одного человека. Если бы владельцы ресторана знали бы предпочтения всех клиентов, то выбор G_0 (цены блюда) мог бы усадить всех за 1 стол (самое любимое блюдо сделать самым дешевым), либо мотивировать занимать новые столы (сделать все основные блюда очень дорогими).

12.6.3 Вариационный вывод

В случае вариационного подхода перепишем исходную модель 263 в терминах переменных $\boldsymbol{\theta}$ и $\boldsymbol{v}$ как в алгоритме 5:

$$p(X, Z, \boldsymbol{\theta}, \boldsymbol{v} \mid \alpha G_0) = \prod_{i=1}^{N} \prod_{k=1}^{\infty} p(x_i \mid \theta_k)^{[z_i=k]} \Bigg(\underbrace{v_k \prod_{j=1}^{k-1} (1 - v_j)}_{\pi_k} \Bigg)^{[z_i=k]} \cdot \prod_{k=1}^{\infty} G_0(\theta_k) Beta(v_k \mid 1, \alpha)$$

Как можно заметить, в сэмплировании Гиббса процесс Дирихле был представлен в виде схемы «Китайский ресторан», здесь же используется представление в виде процесса «Stick breaking», где вероятности кластеров π_i интерпретируются в виде последовательной генерации величин v_i.

Поскольку в данном представлении явно видна факторизованность, логично применить Mean-Field схему:

$$p(\boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{v} \mid \boldsymbol{x}) \approx q(\boldsymbol{z}) q(\boldsymbol{\theta}) q(\boldsymbol{v})$$

Проделаем вариационный вывод для параметров вероятностей кластеров:

$$\log q(\boldsymbol{v}) = const + \mathbb{E}_{q(\boldsymbol{z})q(\boldsymbol{\theta})} \log p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{v}) =$$

$$= const + \mathbb{E}_{q(\boldsymbol{z})q(\boldsymbol{\theta})} \left[\sum_{i=1}^{N} \sum_{k=1}^{\infty} [z_i = k] \left[\log v_k + \sum_{j=1}^{k-1} \log(1 - v_j) \right] + \sum_{k=1}^{\infty} \log Beta(v_k \mid 1, \alpha) \right] =$$

$$= \left\{ \log Beta(v_k \mid 1, \alpha) = \log \left(\frac{v_k^{1-1}(1 - v_k)^{\alpha - 1}}{B(1, \alpha)} \right) = (\alpha - 1) \log(1 - v_k) + const \right\} =$$

$$= const + \sum_{k=1}^{\infty} \left[\log v_k + \sum_{j=1}^{k-1} \log(1 - v_j) \right] \left[\sum_{n=1}^{N} \mathbb{E}_{q(\boldsymbol{z})}[z_n = k] \right] + \sum_{k=1}^{\infty} (\alpha - 1) \log(1 - v_k)$$

При достаточных статистиках Бета-распределения стоят следующие константы:

$$\log v_k : \sum_{i=1}^{N} \mathbb{E}_{q(\boldsymbol{z})}[z_i = k]$$

$$\log(1 - v_k) : \alpha - 1 + \sum_{l=k+1}^{\infty} \sum_{i=1}^{N} \mathbb{E}_{q(\boldsymbol{z})}[z_i = l]$$

Следовательно,

$$q(\boldsymbol{v}) = \prod_{k=1}^{\infty} Beta(v_k \mid 1 + \sum_{i=1}^{N} \mathbb{E}_{q(\boldsymbol{z})}[z_i = k], \alpha + \sum_{l=k+1}^{\infty} \sum_{i=1}^{N} \mathbb{E}_{q(\boldsymbol{z})}[z_i = l])$$

Аналогично из модели получаем $q(\boldsymbol{z})$ и $q(\boldsymbol{\theta})$

Для $\boldsymbol{z}$:

$$q(\boldsymbol{z}) = \prod_{i=1}^{N} q(z_i)$$

$$q(z_i = k) = \frac{1}{A} \exp \left(\mathbb{E}_{q(\boldsymbol{v})q(\boldsymbol{\theta})} \left[\log p(x_i \mid \theta_k) + \log v_k + \sum_{j=1}^{k-1} \log(1 - v_j) \right] \right)$$

Для $\boldsymbol{\theta}$:

$$q(\boldsymbol{\theta}) = \prod_{k=1}^{\infty} q(\theta_k)$$

$$q(\theta_k) = \frac{1}{A} \exp\left(\mathbb{E}_{q(\boldsymbol{v})q(\boldsymbol{z})} \left[\sum_{i=1}^{N} [z_i = k] \log p(x_i \mid \theta_k) + \log G_0(\theta_k) \right] \right)$$