

Байесовские методы в машинном обучении

Д.П. Ветров

Содержание

1	Лекция 1. Байесовский подход к теории вероятностей	3
1.1	Основные понятия	3
1.2	Частотный и байесовский подходы	5
1.3	Приятные плюсы байесовского подхода	7
1.4	Байесовский подход как обобщение булевой логики	7
1.5	Пример байесовских рассуждений	8
2	Лекция 2. Сопряженные распределения, экспоненциальный класс распределений	10
2.1	Сопряжённые распределения	10
2.2	Экспоненциальный класс распределений	12
2.2.1	Оценка параметров распределения из экспоненциального класса	13
2.2.2	Сопряженное семейство к экспоненциальному классу	14
3	Лекция 3. Байесовские методы выбора моделей. Принцип наибольшей обоснованности.	15
3.1	Бритва Оккама. Критерий фальсифицируемости Поппера.	15
3.2	Вероятностные модели	15
3.3	Обучение дискриминативных вероятностных моделей	16
3.4	Принцип наибольшей обоснованности	17
4	Лекция 4. Метод релевантных векторов для задачи регрессии. Автоматическое определение значимости.	22
4.1	Матричное дифференцирование	22
4.2	Решение системы линейных алгебраических уравнений	22
4.3	Вероятностная постановка задачи регрессии. Метод релевантных векторов.	23
5	Лекция 5. Метод релевантных векторов для задачи классификации	30
5.1	Байесовская интерпретация задачи классической логистической регрессии	30
5.2	Метод релевантных векторов	31
5.3	Приближенное вычисление обоснованности методом Лапласа	32
5.4	Оптимизация обоснованности на основе аппроксимации Лапласа	34
5.5	Вариационная нижняя оценка сигмоиды	35
6	Лекция 7. Вариационный Байесовский вывод	38
6.1	ЕМ-алгоритм	38
6.1.1	Пример применения ЕМ-алгоритма на практике	38
6.1.2	Вспоминаем ЕМ-алгоритм	38
6.1.3	Модификация модели ЕМ: априорное распределение на веса	39
6.1.4	От ЕМ-алгоритма к вариационному выводу	40
6.2	Вариационный Байесовский вывод: mean-field аппроксимация	40
6.2.1	Условная сопряженность (conditional conjugacy).	42
6.2.2	Связь mean-field аппроксимации и ЕМ-алгоритма	43
6.3	Концептуальная схема	46

7	Лекция 8. Методы Монте-Карло по схеме марковский цепей (МСМС)	48
7.1	Общие предпосылки метода Монте-Карло	48
7.2	Общие подходы и методы генерации выборок из одномерных распределений	49
7.2.1	Простейшие методы	49
7.2.2	Метод Rejection Sampling	49
7.2.3	Метод Importance sampling	52
7.3	Метод Метрополиса-Хастингса	53
8	Лекция 9а. Гамильтонов Монте-Карло	57
8.1	Общие предпосылки метода гамильтонова Монте-Карло	57
8.2	Описание классического гамильтонова Монте-Карло	57
8.2.1	Гамильтонова механика	57
8.2.2	Схема генерации точек на основе динамики Гамильтона	58
8.3	Обоснование гамильтонова Монте-Карло	59
8.4	Гамильтонов Монте-Карло на практике	60
8.5	Стохастический гамильтонов Монте-Карло	61
9	Лекция 9б. Динамика Ланжевена	64
9.1	Введение в динамику	64
9.2	Уравнение Фоккера-Планка	64
9.3	Сэмплирование	67
9.4	Применение к байесовскому выводу	67
9.5	Применение к схеме Метрополиса-Хастингса	70
9.6	Глобальная оптимизация	70
10	Лекция 11. Непараметрические байесовские методы: процессы Дирихле	71
10.1	Описание байесовских непараметрических моделей	71
10.2	Распределение Дирихле, его свойства	71
10.3	Процессы Дирихле и их применение	75
10.3.1	Определение процесса Дирихле, сравнительные характеристики	75
10.3.2	Представления процесса Дирихле	76
10.3.3	Смесь распределений с априорным распределением, заданным процес- сом Дирихле	78
11	Лекция 12. Тематическая модель Latent Dirichlet allocation (LDA)	82
11.1	Распределение Дирихле	82
11.2	Тематическая модель LDA	82
11.3	Вариационный вывод для модели LDA	83
11.3.1	Е-шаг	83
11.3.2	М-шаг	84

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* $p(x | y)$ (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

$$p(x | y) = \frac{p(x, y)}{p(y)}. \quad (1)$$

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 $p(x | y) = p(x)$. Что означает, что никакой информации об x в y не содержится.

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

$$p(x, y) = p(x | y)p(y). \quad (2)$$

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, \dots, x_n — случайные величины. Тогда их совместное распределение можно представить в виде произведения n одномерных условных распределений с постепенно уменьшающейся посылкой:

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

В дальнейшем мы часто будем сталкиваться с вероятностными 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

$$p(x | y)p(y) = p(x, y) = p(y | x)p(x). \quad (4)$$

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

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

$$p(y | x) = \frac{p(x | y)p(y)}{p(x)}. \quad (5)$$

¹Стоит 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

$$1 = \frac{\int p(x | y)p(y)dy}{p(x)} \Rightarrow p(x) = \int p(x | y)p(y)dy = \int p(x, y)dy. \quad (6)$$

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, \dots, x_n — случайные величины. Если известно их совместное распределение $p(x_1, \dots, x_n)$, то совместное распределение подмножества случайных величин x_1, \dots, x_k будет равно

$$p(x_1, \dots, x_k) = \int p(x_1, \dots, x_n) dx_{k+1} \dots dx_n. \quad (7)$$

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

$$p(x) = \int p(x | y)p(y)dy = \mathbb{E}_y[p(x | y)].$$

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

$$p(y | x) = \frac{p(x | y)p(y)}{\int p(x | y)p(y)dy}. \quad (8)$$

В концептуальной форме это правило звучит так: *апостериорное распределение* $p(y | x)$ (posterior distribution) с точностью до нормировочной константы равно произведению *правдоподобия* $p(x | y)$ (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 некоей 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 любое⁴ условное распределение. Например, скажем, что на три группы случайных величин x , y и z задана нефакторизуемая вероятностная модель $p(x, y, z)$. Как посчитать $p(x | y)$? Достаточно просто:

$$p(x | y) = \frac{p(x, y)}{p(y)} = \frac{\int p(x, y, z)dz}{\iint p(x, y, z)dx dz}. \quad (9)$$

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

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

⁴На 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 $X = (x_1, \dots, x_n)$ из параметрического распределения $p_\theta(x)$. Заметим, что такое распределение вполне можно писать как $p(x | \theta)$, т.е. рассматривать параметры θ как случайные величины, — смысл от этого не меняется. Чтобы оценить параметры θ , в классическом частотном подходе используется метод максимального правдоподобия⁵:

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

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 $n \rightarrow +\infty$ (где n — размер выборки)
- Асимптотическая несмещенность: $\theta_{\text{ML}} = \mathbb{E}[\theta]$ при $n \rightarrow +\infty$
- Асимптотическая нормальность: θ_{ML} распределена нормально при $n \rightarrow +\infty$
- 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 максимальной. Говоря на языке машинного обучения, мы подстраиваем параметры под обучающую выборку. Но мы знаем, что прямая подгонка под данные часто черевата 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:

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

Обратите внимание, что теперь ответом является новое распределение на параметры модели, в отличие 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.

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

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

	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 $n \geq 0$

Пример. Допустим, что мы подбрасываем монетку и смотрим, что выпало. В классической теории вероятностей мы привыкли считать, что исход данного эксперимента является объективной неопределённостью, 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 неопределённость, 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 неопределённостей — 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 $n \rightarrow +\infty$. В байесовском подходе такого ограничения нет: выводы можно делать при любом $n \geq 0$.⁶ Таким образом, при малых значениях n гарантии на ОМП

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

не выполняются, и лучше работает байесовский подход. А какой метод лучше применять при больших n ? Оказывается, что при больших размерах выборки один подход переходит в другой: можно показать, что при $n \rightarrow +\infty$ 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 вѣдливых 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 $n/d \rightarrow \infty$ 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

$$p(y | x, z) = \frac{p(z | y)p(y | x)}{\int p(z | y)p(y | x)dy}. \quad (12)$$

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 $p(B | A)$, то несложно посчитать $p(B)$ по правилу произведения и суммирования

$$p(B) = \sum_A p(B | A)p(A) \quad (13)$$

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

$$p(A | B) = \frac{p(B | A)p(A)}{\sum_A p(B | A)p(A)} \quad (14)$$

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 :

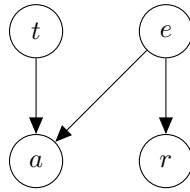


Рис. 1: Граф зависимостей в задаче про сигнализацию.

По такому графу несложно задать совместное распределение на все величины:

$$p(a, e, r, t) = p(a | e, t)p(r | e)p(t)p(e).$$

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

$p(a = 1 e, t)$	$t = 0$	$t = 1$	$p(r = 1 e)$
$e = 0$	0	1	$e = 0$ 0
$e = 1$	0.1	1	$e = 1$ 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 $p(t = 1 | a = 1)$. Для этого воспользуемся теоремой Байеса:

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

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

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

Тогда

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

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:

$$p(t = 1 \mid a = 1, r = 1) = \frac{p(a = 1, t = 1, r = 1)}{p(a = 1, r = 1)} = \frac{\sum_e p(a = 1, e, t = 1, r = 1)}{\sum_{e,t} p(a = 1, e, t, r = 1)} \quad (18)$$

$$= \frac{\sum_e p(a = 1 \mid e, t = 1) p(r = 1 \mid e) p(e) p(t = 1)}{\sum_{e,t} p(a = 1 \mid e, t) p(r = 1 \mid e) p(e) p(t)}. \quad (19)$$

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

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

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 $X = \{x_i\}_{i=1}^n$, $x_i \sim p(x | \theta)$, 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 .

$$p(\theta | X) = \frac{p(X | \theta)p(\theta)}{\int p(X | \theta)p(\theta)d\theta} \quad (21)$$

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:

$$\theta_{MP} = \arg \max_{\theta} p(\theta | X) = \arg \max_{\theta} p(X | \theta)p(\theta) = \quad (22)$$

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

Получили довольно известную регуляризацию на давно знакомую оценку максимального правдоподобия. Так, например, если в качестве априорного распределения мы возьмём нормальное распределение с нулевым матожиданием 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. Расчитав 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.

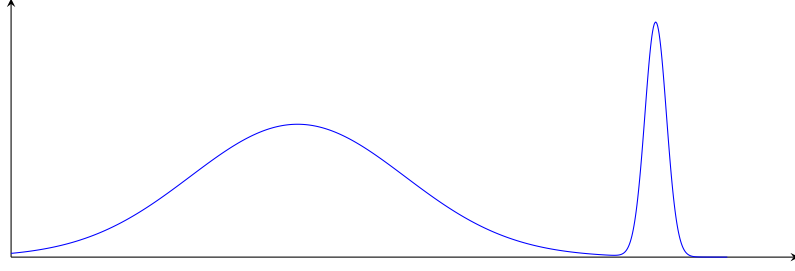


Рис. 2: Пример распределения, у которого мода нерепрезентативна.

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: $p(X | \theta) \sim \mathcal{A}(\theta)$ и $p(\theta | \beta) \sim \mathcal{B}(\beta)$. Семейства \mathcal{A} и \mathcal{B} являются *сопряжёнными* (*conjugate*) тогда и только тогда, когда $p(\theta | X) \sim \mathcal{B}(\beta')$.

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

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

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

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:

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

Теперь проверим:

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

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

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

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

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

$$p(x | \gamma) = \sqrt{\frac{\gamma}{2\pi}} \exp\left(-\frac{\gamma}{2}x^2\right)$$

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

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

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

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

$$p(x | \mu, \gamma^{-1}) = \sqrt{\frac{\gamma}{2}} \exp\left(-\frac{\gamma}{2}(x - \mu)^2\right) = \sqrt{\frac{\gamma}{2}} \exp\left(-\frac{\gamma x^2}{2} + \gamma \mu x - \frac{\gamma \mu^2}{2}\right)$$

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

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

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

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 $p(x | \theta)$ лежит в *экспоненциальном классе*, если оно может быть представлено в следующем виде

$$p(x | \theta) = \frac{f(x)}{g(\theta)} \exp(\theta^T u(x)), \quad f(\cdot) \geq 0, \quad g(\cdot) > 0, \quad (29)$$

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*:

⁷Стоит 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 $p(X | \theta) = f_1(X)f_2(\theta, a(X))$

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 $f_1(X) = f(X), f_2(\theta, u(X)) = \frac{\exp(\theta^T u(X))}{g(\theta)}$).

Рассмотрим одно замечательное свойство экспоненциального класса распределений. Заметим, что

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

Продифференцируем по θ_j

$$\frac{\partial}{\partial \theta_j} g(\theta) = \frac{\partial}{\partial \theta_j} \int f(x) \exp(\theta^T u(x)) dx = \int f(x) \exp(\theta^T u(x)) u_j(x) dx = \quad (31)$$

$$= g(\theta) \int \frac{f(x)}{g(\theta)} \exp(\theta^T u(x)) u_j(x) dx = g(\theta) \int p(x | \theta) u_j(x) dx = g(\theta) \mathbb{E}_{x \sim p(x|\theta)} u_j(x) \quad (32)$$

В итоге получаем, что

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

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

$$\frac{\partial^2}{\partial \theta_j \partial \theta_k} \log g(\theta) = \text{Cov}(u_j(x), u_k(x)) \quad (34)$$

2.2.1 Оценка параметров распределения из экспоненциального класса

Пусть нам дана выборка из распределения экспоненциального класса:

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

Оценим параметры распределения методом максимального правдоподобия

$$\theta_{ML} = \arg \max_{\theta} p(X | \theta) = \arg \max_{\theta} \sum_{i=1}^n \log p(x_i | \theta) = \arg \max_{\theta} \sum_{i=1}^n \left(\log f(x_i) - \log g(\theta) + \theta^T u(x_i) \right)$$

Продифференцировав по θ_j последнее выражение, приравняем производную к нулю и получим

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

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

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

Из выражения выше видно, что

$$\begin{aligned}\theta_1 &= \frac{\gamma}{2} & u_1(x) &= x^2 \\ \theta_2 &= \gamma\mu & u_2(x) &= x \\ g(\theta) &= \sqrt{\frac{2\pi}{\gamma}} \exp\left\{\frac{\gamma}{2}\mu^2\right\}\end{aligned}$$

2.2.2 Сопряженное семейство к экспоненциальному классу

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

$$p(\theta \mid \eta, \nu) = \exp(\theta^T \eta) \frac{1}{g^\nu(\theta)} \frac{1}{h(\eta, \nu)} \quad (35)$$

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 несуществование аналитически заданного сопряжённого семейства.

Вычислим апостериорное распределение:

$$p(\theta \mid X) = \frac{1}{Z} \prod_{i=1}^n p(x_i \mid \theta) p(\theta \mid \nu, \eta) = \quad (36)$$

$$= \frac{1}{Z} \prod_{i=1}^n [f(x_i)] \cdot \frac{1}{g^n(\theta)} \exp\left\{\theta^T \left(\sum_{i=1}^n u(x_i)\right)\right\} \exp\{\theta^T \eta\} \frac{1}{g^\nu(\theta)} \frac{1}{h(\eta, \nu)} = \quad (37)$$

$$= \frac{1}{Z'} \exp\left\{\theta^T \left(\eta + \sum_{i=1}^n u(x_i)\right)\right\} \frac{1}{g^{\nu+n}(\theta)} = \frac{1}{h(\eta', \nu')} \exp(\theta^T \eta') \frac{1}{g^{\nu'}(\theta)} \quad (38)$$

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

$$\eta' = \eta + \sum_{i=1}^n u(x_i) \quad (39)$$

$$\nu' = \nu + n \quad (40)$$

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 явления 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 распространенных постановок задач машинного обучения состоит в следующем. Дана выборка независимых одинаково распределённых объектов. Описание каждого объекта задается парой вида (x, t) :

$$(X_{tr}, T_{tr}) = \{x_i, t_i\}_{i=1}^n \quad (41)$$

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

$$p(x, t, \theta) = p(x, t \mid \theta)p(\theta) = p(t \mid x, \theta)p(x \mid \theta)p(\theta) \quad (42)$$

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

2. Дискриминативная модель

$$p(t, \theta \mid x) = p(t \mid x, \theta)p(\theta) \quad (43)$$

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

$$p(t, \theta \mid x) = p(t \mid \theta, x)p(\theta) \quad (44)$$

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} .

Таблица 2: Схема обучения и применения дискриминативной модели

Этап	Дано	Неизвестно	Хотим оценить
Обучение	$(X_{tr}, T_{tr}) = (x_i, t_i)_{i=1}^n$	θ	$p(\theta \mid X_{tr}, T_{tr})$
Тестирование	x_{test}	t_{test}	$p(t_{test} \mid x_{test}, X_{tr}, T_{tr})$

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

$$p(\theta \mid X_{tr}, T_{tr}) = \frac{p(T_{tr} \mid X_{tr}, \theta)p(\theta)}{\int p(T_{tr} \mid X_{tr}, \theta)p(\theta)d\theta} \quad (45)$$

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.

$$p(t_{test} | x_{test}, X_{tr}, T_{tr}) = \int p(t_{test} | x_{test}, \theta) p(\theta | X_{tr}, T_{tr}) d\theta \quad (46)$$

На данном этапе мы по сути делаем следующее: применяем все возможные (со всеми возможными значениями θ) алгоритмы прогнозирования $p(t_{test} | x_{test}, \theta)$ и усредняем полученные значения с весами, которые задаются нам апостериорным распределением $p(\theta | X_{tr}, T_{tr})$. Т.е. интеграл в выражении 46 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 45 невозможен, т.е. если интеграл в знаменателе формулы Байеса не берется? В этом случае есть два пути: приближенно оценить апостериорное распределение⁹ или перейти к точечной оценке параметров, воспользовавшись уже знакомым нам «Байесом для бедных»:

$$\theta_{MP} = \arg \max_{\theta} p(\theta | X_{tr}, T_{tr}) \quad (47)$$

Здесь параметры θ оцениваются только в одной точке, что соответствует замене честного апостериорного распределения 45 на дельта-функцию с центром в точке θ_{MP}

$$p(\theta | X_{tr}, T_{tr}) \approx \delta(\theta - \theta_{MP}) \quad (48)$$

Подставив данное приближение в интеграл для прогнозного распределения 46, получим

$$p(t_{test} | x_{test}, X_{tr}, T_{tr}) \approx p(t_{test} | x_{test}, \theta_{MP}) \quad (49)$$

«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 $p(t, \theta | x)$. 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:

$$p_j(t, \theta | x) = p_j(t | x, \theta) p_j(\theta), \quad j = 1, 2, 3 \quad (50)$$

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 критериев¹⁰. В нашем курсе мы рассмотрим один из них — принцип наибольшей обоснованности¹¹. 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:*

$$j^* = \arg \max_j p_j(T_{tr} | X_{tr}) = \arg \max_j \int p_j(T_{tr} | X_{tr}, \theta) p_j(\theta) d\theta \quad (51)$$

⁸Типичный пример ансамблирования или взвешенного голосования

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

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

¹¹впервые был предложен в 1992 году британским физиком Дэвидом МакКаем

Распределение $p_j(T_{tr} | X_{tr})$ называется обоснованностью (evidence). Напомним, что именно эта величина стоит в знаменателе теоремы Байеса (см. выражение 45). Заметим, что по 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 : $p_j(T, \theta | X)$. Для иллюстративности будем считать T и θ одномерными (см. Рис. 3).

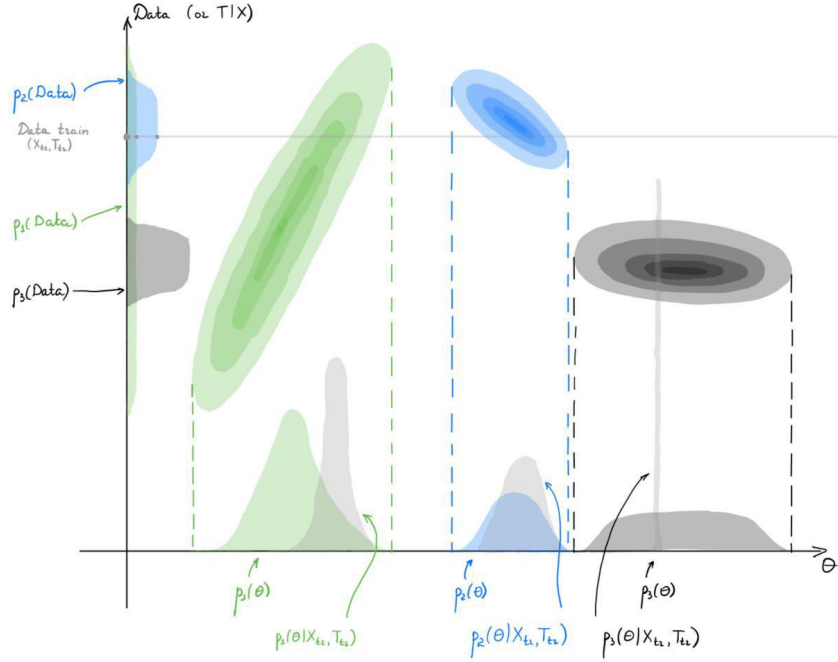


Рис. 3: Совместное распределение $p_j(T, \theta | X)$ для трех моделей. По горизонтальной оси отложен параметр θ , по вертикальной — значение целевой переменной T про условия X (для иллюстративности θ и T одномерные). Эллипсы отображают совместное распределение $p_j(T, \theta | X)$. Цветные распределения 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 $p_j(T, \theta | X)$ на ось θ . Для этого его нужно маргинализовать по T :

$$\int p_j(T, \theta | X) dT = \int p_j(T | \theta, X) p_j(\theta) dT = p_j(\theta) \quad (52)$$

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

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 $p(\theta | X_{tr}, T_{tr})$. На картинке апостериорному распределению будет соответствовать сечение совместного распределения $p(T, \theta | X)$ прямой $T_{tr} | X_{tr}$. На рисунке 3 горб $p_2(\theta | X_{tr}, T_{tr})$ ниже чем горб $p_1(\theta | X_{tr}, T_{tr})$ так как сечение второй совместной плотности шире, а площадь под горбом должна равняться единице (как интеграл от плотности распределения). Плотность распределения $p_3(\theta | X_{tr}, T_{tr})$

практически схлопывается в дельта-функцию в точке, где прямая $T_{tr} | X_{tr}$ касается линий уровня совместной плотности распределения $p_3(T, \theta | X)$ (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 $T_{tr} | X_{tr}$ 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 объясняют данные, поскольку содержат такие значения θ при которых правдоподобие данных $p(T_{tr}|X_{tr}, \theta)$ 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 распределений на вертикальную ось на рисунке 3. Эти проекции есть $p_i(T | X)$, т.е. это обоснованности моделей. Точки, в которых прямая $T_{tr} | X_{tr}$ пересекает кривые $p_i(T | X)$ равны обоснованностям наблюдаемых данных в рамках рассматриваемых моделей. Больше всего обоснованность данных у второй модели, поскольку ее плотность $p_2(T | X)$ выше всех в точке $T_{tr} | X_{tr}$. Первая модель тоже объясняет $T_{tr} | X_{tr}$, однако она может объяснить и много других значений $T | X$, поэтому ее плотность $p_1(T | X)$ «размазана» по вертикальной оси и имеет низкое значение в точке $T_{tr} | X_{tr}$. То есть чем большую совокупность фактов может объяснить модель, тем меньше у нее обоснованность для конкретных значений $T_{tr} | X_{tr}$.

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.

¹²Конкретный вид апостериорного распределения зависит от хвостов совместного распределения $p_3(T, \theta | X)$. 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.

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

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

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

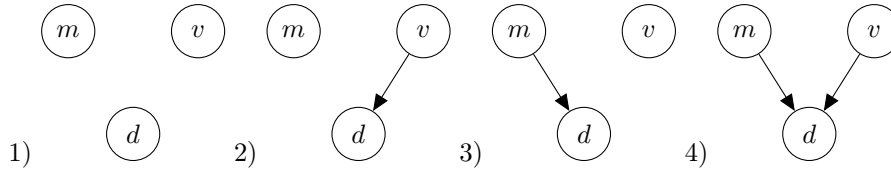


Рис. 4: Предлагаемые модели зависимости приговора d от расы убийцы m и расы жертвы v

Чтобы полностью задать байесовскую модель, необходимо ввести априорные распределения на параметры моделей (θ , α , β , γ , δ , τ , ν , χ , ξ). 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:

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

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:

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

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

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

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.

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

4.2 Решение системы линейных алгебраических уравнений

Рассмотрим СЛАУ:

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

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 $\text{rk} A = n$).
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 $A^T A$ обратима (т.к. $\text{rk}(A^T A) = \text{rk}(A) = n$). Таким образом, получаем

$$x^* = (A^T A)^{-1} A^T b$$

Матрица $(A^T A)^{-1} A^T$ 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

$$x = (A^T A + \lambda I)^{-1} A^T b$$

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

$$x^* = \lim_{\lambda \rightarrow 0} (A^T A + \lambda I)^{-1} A^T b.$$

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 переменная¹³, $w \in \mathbb{R}^m$ - веса линейной регрессии. Пусть имеется также $(X, T) = (x_i, t_i)_{i=1}^n$ — обучающая выборка. Введём дискриминативную вероятностную модель

$$p(t, w | x) = p(t | w, x) p(w) \quad (53)$$

где $p(t | w, x)$ - функция правдоподобия, $p(w)$ - априорное распределение на веса. Правдоподобие $p(t | w, x)$ зададим нормальным распределением по t с матожиданием, равным линейной комбинации признаков $w^T x$, 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:

$$\arg \max_w \mathcal{N}(t | w^T x, \beta^{-1}) = w_{ML} = \arg \min_w \|w^T x - t\|_2^2. \quad (54)$$

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

$$p(t, w | x) = p(t | x, w) p(w) = \mathcal{N}(t | w^T x, \beta^{-1}) \mathcal{N}(w | 0, \alpha^{-1} I) \quad (55)$$

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:

$$p(w | X, T) = \mathcal{N}(w | \mu, \Sigma) \quad (56)$$

¹³В 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:

$$p(w | X, T) = \mathcal{N}(w | \mu, \Sigma) = \frac{p(T | X, w) p(w)}{\int p(T | X, w) p(w) dw} \quad (57)$$

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 (57):

$$\begin{aligned} p(T | X, w) p(w) &= \frac{\beta^{\frac{n}{2}}}{(2\pi)^{\frac{n}{2}}} \exp\left\{-\frac{\beta}{2} \|T - Xw\|^2\right\} \frac{\alpha^{\frac{m}{2}}}{(2\pi)^{\frac{m}{2}}} \exp\left\{-\frac{\alpha}{2} w^T w\right\} = \\ &= \frac{\beta^{\frac{n}{2}} \alpha^{\frac{m}{2}}}{(2\pi)^{\frac{m+n}{2}}} \exp\left\{-\frac{\beta}{2} (T^T T - 2w^T X^T T + w^T X^T X w) - \frac{\alpha}{2} w^T w\right\} \\ &= \frac{\beta^{\frac{n}{2}} \alpha^{\frac{m}{2}}}{(2\pi)^{\frac{m+n}{2}}} \exp\left\{-\frac{1}{2} w^T \underbrace{(\beta X^T X + \alpha I)}_{\Sigma^{-1}} w + \beta w^T X^T T - \frac{\beta}{2} T^T T\right\} \end{aligned} \quad (58)$$

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

$$\Sigma = (\beta X^T X + \alpha I)^{-1} \quad (59)$$

Чтобы найти μ можно выделить полный квадрат под экспонентой и провести громоздкие вычисления. Но мы поступим проще и воспользуемся тем, что матожидание нормального распределения совпадает с его модой: $\mu = w_{MP}$. Т.е. надо найти w , максимизирующий (58), то есть максимум следующего выражения:

$$-\frac{\beta}{2} (T^T T - 2w^T X^T T + w^T X^T X w) - \frac{\alpha}{2} w^T w \quad (60)$$

Найдём производную (60) по w и приравняем её к нулю:

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

Отсюда получаем формулу

$$w_{MP} = \underbrace{\beta (\beta X^T X + \alpha I)^{-1}}_{\Sigma} X^T T \quad (61)$$

Итого, мы получили значения параметров для апостериорного распределения:

$$\Sigma = (\beta X^T X + \alpha I)^{-1}, \quad (62)$$

$$\mu = w_{MP} = \beta \Sigma X^T T \quad (63)$$

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

$$p(t_* | x_*, X, T) = \int p(t_* | x_*, w) p(w | X, T) dw \quad (64)$$

Интеграл в формуле (64) 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

¹⁴Заметим 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 распределение¹⁵:

$$p(t_* | x_*, X, T) = \int p(t_* | x_*, w) p(w | X, T) dw = \mathcal{N}(t_* | x_*^T w_{MP}, \dots) \quad (65)$$

Теперь посмотрим, как наш алгоритм прогнозирования 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} 60).

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

1. $\alpha \rightarrow 0$

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

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

2. $\alpha \rightarrow \infty$

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

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

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 регуляризовались по-разному. Мы можем изменить модель так, чтобы каждому w_i соответствовал свой собственный коэффициент α_i :

$$p(t, w | x) = p(t | x, w) p(w) = \mathcal{N}(t | w^T x, \beta^{-1}) \mathcal{N}(w | 0, A^{-1}), \quad A = \begin{pmatrix} \alpha_1 & & 0 \\ & \ddots & \\ 0 & & \alpha_m \end{pmatrix} \quad (66)$$

Теперь наша вероятностная модель индексируется параметром β и параметрами $\alpha_1, \dots, \alpha_m$ 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

¹⁵Точный вид матрицы ковариации предлагаем читателю вывести самостоятельно

¹⁶Можно 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:

$$p(T | X, A, \beta) = \int p(T | X, w, \beta) p(w | A) dw \quad (67)$$

Интеграл 67 берётся (знаменатель формулы Байеса), но мы можем упростить вычисления. Обозначим $p(T | X, w, \beta) p(w, A)$ как $Q(w)$. Посмотрим что представляет собой эта функция как функция от w :

$$\begin{aligned} Q(w) &= \frac{\beta^{\frac{n}{2}}}{(2\pi)^{\frac{n}{2}}} \exp\left\{\left(-\frac{\beta}{2}\|T - Xw\|^2\right)\right\} \frac{\sqrt{\det A}}{(2\pi)^{\frac{m}{2}}} \exp\left\{\left(-\frac{1}{2}w^T A w\right)\right\} = \\ &= \frac{\beta^{\frac{n}{2}} \sqrt{\det A}}{(2\pi)^{\frac{m+n}{2}}} \exp\left\{\left(-\frac{1}{2}w^T (\beta X^T X + A) w + \beta w^T X^T T - \frac{\beta}{2}T^T T\right)\right\} \end{aligned} \quad (68)$$

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 (68) расписывается как

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

Вернемся к интегралу (67):

$$\begin{aligned} p(T | X, A, \beta) &= \int p(T | X, w, \beta) p(w, A) dw = \\ &= \int Q(w_{MP}) \exp\left\{\left(-\frac{1}{2}(w - w_{MP})^T \Sigma^{-1} (w - w_{MP})\right)\right\} dw = \\ &= Q(w_{MP}) \int \exp\left\{\left(-\frac{1}{2}(w - w_{MP})^T \Sigma^{-1} (w - w_{MP})\right)\right\} dw = \\ &= Q(w_{MP}) (2\pi)^{\frac{m}{2}} \sqrt{\det \Sigma} \rightarrow \max_{A, \beta} \end{aligned} \quad (70)$$

Теперь применим любопытный приём. рассмотрим логарифм обоснованности (70):

$$\begin{aligned} \log p(T | X, A, \beta) &= \frac{n}{2} \log \beta - \frac{n}{2} \log 2\pi + \frac{1}{2} \log \det A - \frac{\beta}{2} \|T - Xw_{MP}\|^2 - \\ &\quad - \frac{1}{2} w_{MP}^T A w_{MP} - \frac{1}{2} \log \det \Sigma^{-1} \rightarrow \max_{A, \beta} \end{aligned} \quad (71)$$

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 (62)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

¹⁷При 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:

$$\begin{cases} x_n = \arg \max_x g(x, \xi_{n-1}), \\ \xi_n = \arg \max_{\xi} g(x_n, \xi) \end{cases} \quad (72)$$

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 (71) можно рассмотреть как

$$\begin{aligned} \log p(T | X, A, \beta) &= \frac{n}{2} \log \beta - \frac{n}{2} \log 2\pi + \frac{1}{2} \log \det A - \frac{\beta}{2} \|T - X w_{MP}\|^2 - \\ &\quad - \frac{1}{2} w_{MP}^T A w_{MP} - \frac{1}{2} \log \det \Sigma^{-1} \geq \end{aligned} \quad (73)$$

$$\begin{aligned} &\geq \frac{n}{2} \log \beta - \frac{n}{2} \log 2\pi + \frac{1}{2} \log \det A - \frac{\beta}{2} \|T - X w\|^2 - \\ &\quad - \frac{1}{2} w^T A w - \frac{1}{2} \log \det \Sigma^{-1} \end{aligned} \quad (74)$$

Оценка (73) 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

$$\frac{n}{2} \log \beta - \frac{\beta}{2} \|T - X w\|^2 + \frac{1}{2} \log \det A - \frac{1}{2} w^T A w - \frac{1}{2} \log \det \Sigma^{-1} \rightarrow \max_{A, \beta, w} \quad (75)$$

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 (75) по α_j при $w = w_{MP}$ (считаем, что w_{MP} не зависит от A) и приравняем к нулю:

$$\begin{aligned} \frac{\partial}{\partial \alpha_j} \left(\frac{n}{2} \log \beta - \frac{\beta}{2} \|T - X w_{MP}\|^2 + \frac{1}{2} \log \det A - \frac{1}{2} w_{MP}^T A w_{MP} - \frac{1}{2} \log \det \Sigma^{-1} \right) &= \\ &= 0 - 0 + \frac{1}{2} \frac{\partial}{\partial \alpha_j} \log \det A - \frac{1}{2} \frac{\partial}{\partial \alpha_j} (w_{MP}^T A w_{MP}) - \frac{1}{2} \frac{\partial}{\partial \alpha_j} \log \det \Sigma^{-1} = \\ &= \left\{ \frac{\partial}{\partial \alpha_j} \log \det A = \frac{\partial}{\partial \alpha_j} \sum_{i=1}^m \log \alpha_i = \frac{1}{\alpha_j}; \frac{\partial}{\partial \alpha_j} (w_{MP}^T A w_{MP}) = (w_{MP})_j^2; \right. \\ &\quad \left. \frac{\partial}{\partial \alpha_j} \log \det \Sigma^{-1} = \text{tr} \left(\frac{\partial \log \det \Sigma^{-1}}{\partial \Sigma^{-1}} \frac{\partial \Sigma^{-1}}{\partial \alpha_j} \right) = \text{tr} (\Sigma^T I_{jj}) = \Sigma_{jj} \right\} = \\ &= \frac{1}{2\alpha_j} - \frac{1}{2} w_{MP}^2 - \frac{1}{2} \Sigma_{jj} = 0 \end{aligned} \quad (76)$$

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

$$\alpha_j = \frac{1}{w_{MP}^2 + \Sigma_{jj}} \quad (77)$$

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 (77), то сойтись процесс будет довольно медленно. Почему?

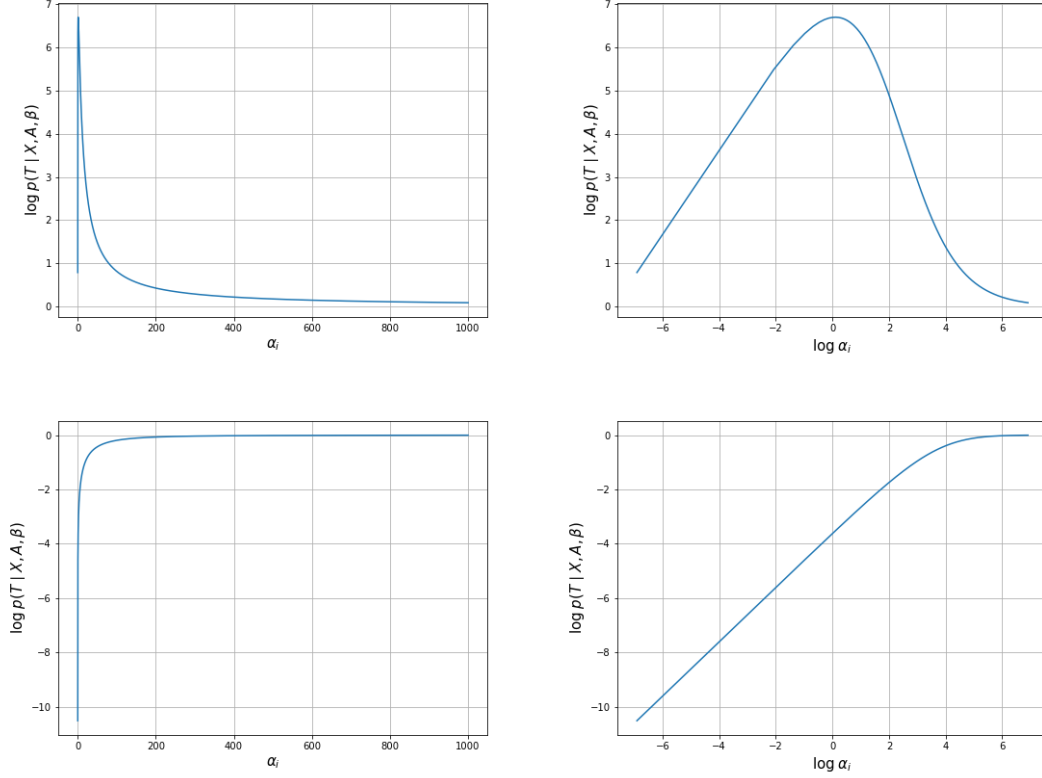


Рис. 5: Возможные виды зависимости оптимизируемой функции от α_j (левые графики) и от $\log \alpha_j$ (правые графики)

Посмотрим на график зависимости оптимизируемой функции от α_j и от $\log \alpha_j$ (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:

$$1 - \alpha_j w_{jMP}^2 - \alpha_j \Sigma_{jj} = 0 \quad (78)$$

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

В формуле (78) можем дополнительно разделить переменные:

$$1 - \alpha_j^{new} w_{jMP}^2 - \alpha_j^{old} \Sigma_{jj} = 0, \quad (79)$$

откуда получаем

$$\alpha_j^{new} = \frac{1 - \alpha_j^{old} \Sigma_{jj}}{w_{MP}^2}. \quad (80)$$

Аналогично выводится формула для β :

$$\beta^{new} = \frac{n - \sum_{j=1}^n (1 - \alpha_j^{old} \Sigma_{jj})}{\|T - Xw_{MP}\|^2} \quad (81)$$

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 метод 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 правдоподобие 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: $A = \text{diag}(\alpha_1, \dots, \alpha_d)$ приводит к разреженному 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 $A = \text{diag}(\alpha_1, \dots, \alpha_d)$.

5.1 Байесовская интерпретация задачи классической логистической регрессии

Мы наблюдаем набор независимых пар $\{(x_n, t_n)\}_{n=1}^N$: вектор признаков $x_n = (1, x_n^1, \dots, x_n^d)$ 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:

$$p(t_n | w, x_n) = \frac{1}{1 + \exp(-t_n w^T x_n)}. \quad (82)$$

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

$$p(t = -1 | x, w) + p(t = 1 | x, w) = \frac{1}{1 + e^{w^T x}} + \frac{1}{1 + e^{-w^T x}} = \frac{1 + e^{-w^T x} + e^{w^T x} + 1}{1 + e^{-w^T x} + e^{w^T x} + e^0} = 1. \quad (83)$$

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

$$p(w) = \mathcal{N}(w | 0, A^{-1}). \quad (84)$$

Итоговая вероятностная модель имеет вид:

$$p(\{(x_n, t_n)\}_{n=1}^N | w) = \left[\prod_{n=1}^N \frac{1}{1 + \exp(-t_n w^T x_n)} \right] \mathcal{N}(w | 0, A^{-1}). \quad (85)$$

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

$$w_{MP} = \arg \max_w p(w | X, T) = \arg \max_w \log p(w | X, T) = \arg \max_w \log[p(T | w, X)p(w)]. \quad (86)$$

Продолжая (86):

$$= \arg \max_w (\log p(T | w, X) + \log p(w)) = \quad (87)$$

$$= \arg \max_w \left(- \sum_{n=1}^N \log(1 + \exp(-t_n w^T x_n)) - \frac{1}{2} w^T A w \right) = \quad (88)$$

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

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

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

Читателю осталось проверить, что задача (90) 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:

$$w^{(k+1)} = \underbrace{\left(X^T R(w^{(k)}) X + \alpha I \right)^{-1}}_{d \times d} X^T R(w^{(k)}) z(w^{(k)}), \quad (91)$$

где

$$X = \begin{pmatrix} 1 & x_1^2 & \dots & x_1^d \\ 1 & x_2^2 & \dots & x_2^d \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_N^2 & \dots & x_N^d \end{pmatrix} \text{ — матрица признаков} \quad (92)$$

$$R(w) = \text{diag}(s_1(1-s_1), \dots, s_N(1-s_N)), \quad s_n = \frac{1}{1 + \exp(-t_n w^T x_n)} \quad (93)$$

$$z(w) = Xw + R^{-1}(w) \begin{pmatrix} t_1 & & 0 \\ & \ddots & \\ 0 & & t_N \end{pmatrix} \begin{pmatrix} 1-s_1 \\ \vdots \\ 1-s_N \end{pmatrix}. \quad (94)$$

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

$$\nabla^2 [\log p(T | X, w) + \log p(w)] = -(X^T R(w) X + \alpha I). \quad (95)$$

5.2 Метод релевантных векторов

Мы описали задачу логистической регрессии на «байесовском языке», введя априорное распределение на параметры модели $\mathcal{N}(w|0, A^{-1})$. Затем мы продемонстрировали связь такого выбора априорного распределения с использованием l_2 -регуляризации в задаче обучения логистической регрессии. Действуя по аналогии с предыдущей лекцией, мы можем выбрать для каждого параметра w_i свой «коэффициент регуляризации»:

$$p(w | A) = \mathcal{N}(w | 0, A^{-1}) = \prod_{i=1}^d \mathcal{N}(w_i | 0, \alpha_i^{-1}), \quad A = \text{diag}(\alpha_1, \dots, \alpha_d).$$

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

$$w_i \xrightarrow{d} 0. \quad (96)$$

Таким образом, если мы будем оптимизировать обоснованность модели по параметрам априорного распределения $\text{diag}(\alpha_1, \dots, \alpha_d)$, 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 $p(t | w, x)$ и $p(w | A)$ 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. Эту можно сделать с помощью того же самого IRLS, который в данном случае он будет выглядеть так:

$$w_{k+1} = (X^T R(w_k) X + A)^{-1} X^T R(w_k) z(w_k), \quad (97)$$

где X , $R(w)$ и $z(w)$ определены, соответственно, в (92), (93) и (94). Метод IRLS гарантирует, что $w_k \rightarrow w_{MP}$.

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 2. 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 функции — функции вида

$$\varphi_j(x) = \exp(-\gamma \|x - x_j\|^2) \quad (98)$$

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:

$$p(T | X, A) = \int p(T | X, w) p(w | A) dw \rightarrow \max_A. \quad (99)$$

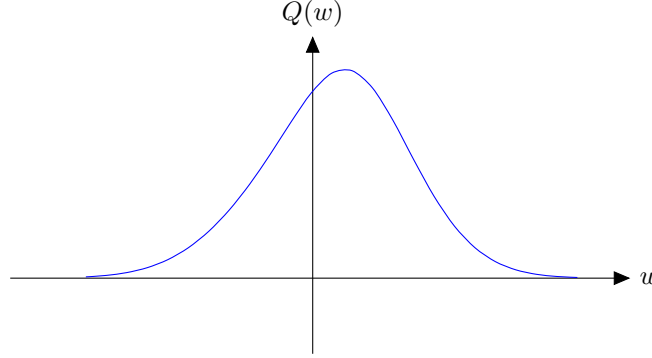
Для решения такой задачи оптимизации нужно уметь вычислять интеграл в (99), который, увы, не берется аналитически. Однако мы можем его оценить для каждого фиксированного значения параметра A ! Один из способов приблизить значение интеграла — это заменить его подынтегральную функцию на удобную оценку. По этой причине введём обозначение:

$$Q(w) := p(T | X, w) p(w | A). \quad (100)$$

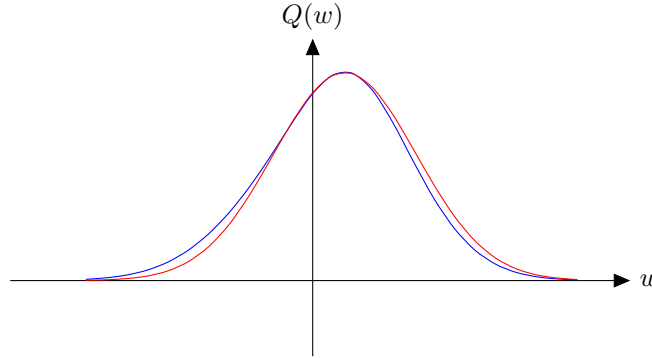
Для того чтобы предложить оценку для $Q(w)$, подумаем, что мы вообще можем сказать об этой функции. Давайте возьмём от неё логарифм:

$$\log Q(w) = - \sum_{n=1}^N \log(1 + \exp(-t_n w^T x_n)) - \sum_{j=1}^d \alpha_j w_j^2. \quad (101)$$

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 колокольчик — 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 $\log Q(w)$:

$$\begin{aligned} \log Q(w) \approx \log Q(w_{MP}) + (w - w_{MP})^T \nabla \log Q(w_{MP}) + \\ + \frac{1}{2} (w - w_{MP})^T \underbrace{\nabla^2 \log Q(w_{MP})}_{\text{гессиан}} (w - w_{MP}). \end{aligned} \quad (102)$$

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:

$$\nabla^2 \log Q(w_{MP}) = -(X^T R(w_{MP}) X + A), \quad (103)$$

где X , $R(w)$ определены выше (92), (93). Вывод формулы (103) 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:

$$\int Q(w)dw \approx \int Q(w_{MP}) \exp\left(-\frac{1}{2}(w - w_{MP})^T \Sigma^{-1}(w - w_{MP})\right) dw = Q(w_{MP})(2\pi)^{d/2} \sqrt{\det \Sigma}. \quad (104)$$

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 колокольчик, тем устойчивее будет модель, ведь $Q(w)$ будет в таком случае слабо изменяться в окрестности значений параметра w_{MP} .

Распишем чуть подробнее (104) как функцию от A :

$$\begin{aligned} \log p(T | X, A) &\approx \\ &\approx \frac{d}{2} \log(2\pi) + \log p(T | X, w_{MP}) + \log \mathcal{N}(w_{MP} | 0, A^{-1}) - \frac{1}{2} \log \det (X^T R(w_{MP})X + A). \end{aligned} \quad (105)$$

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

5.4 Оптимизация обоснованности на основе аппроксимации Лапласа

Замечание 3. Вплоть до этого момента мы обозначали w_{MP} точку максимума $p(T | X, w)p(w | A)$ при некоторой фиксированной матрице A . В данном разделе нам придётся переобозначить w_{MP} как w_{MP}^A :

$$w_{MP}^A = \arg \max_w p(T | X, w)p(w | A), \quad (106)$$

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

Итак, мы хотим решить задачу

$$\log p(T | X, A) \rightarrow \max_A. \quad (107)$$

Воспользовавшись приближением (105), оптимальную A можно найти, оптимизируя по A функцию:

$$F(A, w_{MP}^A) := \log p(T | X, w_{MP}^A) + \log \mathcal{N}(w_{MP}^A | 0, A^{-1}) - \frac{1}{2} \log \det (X^T R(w_{MP}^A)X + A). \quad (108)$$

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

$$\frac{\partial F(A, w_{MP}^A)}{\partial \log \alpha_j} = \alpha_j \frac{\partial F(A, w_{MP}^A)}{\partial \alpha_j} = 0, \quad j = 1, \dots, d. \quad (109)$$

Основная проблема заключается в том, что зависимость величины w_{MP}^A от A очень сложна, а при взятии производной (109) без нахождения $\frac{\partial w_{MP}^A}{\partial \alpha_j}$ не обойтись. Однако, можно заметить, что $F(A, w_{MP}^A) \geq F(A, w)$, для любого w , при фиксированной матрице A . Дифференцирование такой оценки аналогично взятию производной, считая $w_{MP}^A = \text{const}$ относительно A .

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

$$F(A, w_{MP}^A) = - \sum_{n=1}^N \log(1 + \exp(t_n (w_{MP}^A)^T x_n)) - \frac{1}{2} (w_{MP}^A)^T A w_{MP}^A + \quad (110)$$

$$\frac{1}{2} \log \det A - \frac{1}{2} \log \det (X^T R(w_{MP}^A)X + A) + \text{const}. \quad (111)$$

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

$$\frac{\partial}{\partial \log \alpha_j} \log \det (X^T R(w_{MP}^A) X + A) = \alpha_j \frac{\partial}{\partial \alpha_j} \log \det (X^T R(w_{MP}^A) X + A) = \quad (112)$$

$$= \alpha_j \text{tr} \left((X^T R(w_{MP}^A) X + A)^{-1} E_{jj} \right) = \quad (113)$$

$$= \alpha_j \left[(X^T R(w_{MP}^A) X + A)^{-1} \right]_{jj}. \quad (114)$$

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

$$0 = \frac{\partial F(A, w_{MP}^A)}{\partial \log \alpha_j} = -\frac{\alpha_j}{2} \left[(w_{MP}^A)_j \right]^2 + \frac{1}{2} - \frac{\alpha_j}{2} \left[(X^T R(w_{MP}^A) X + A)^{-1} \right]_{jj}. \quad (115)$$

Шаг метода оптимизации для такой задачи можно записать так:

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

По факту, мы должны делать итеративно следующие два шага¹⁸:

1. Найти w_{MP}^{old} для текущей матрицы A^{old}
2. Найти A^{new} по формуле (116)

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, индексруемый 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

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

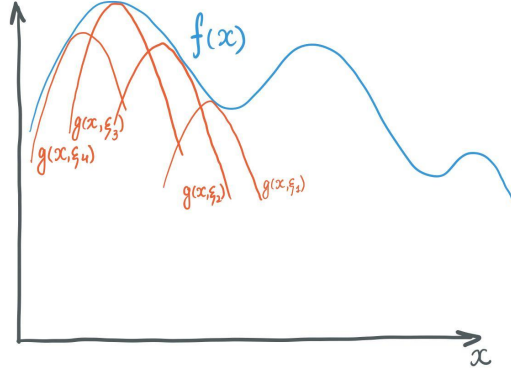


Рис. 6: Возможный вид вариационной нижней оценки при различных значениях вариационного параметра

исходным параметрам x , то такая процедура в итоге сойдется к локальному максимуму исходной функции $f(x)$ (см. выражение 72).¹⁹

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

$$p(T | X, A) = \int p(T | X, w) p(w | A) dw = \int \prod_{n=1}^N \frac{1}{1 + \exp(-t_n w^T x_n)} \mathcal{N}(w | 0, A^{-1}) dw \quad (117)$$

Попробуем оценить произведение сигмoids 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).

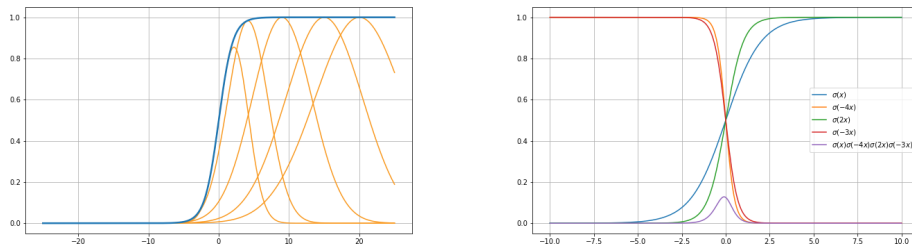


Рис. 7: Вариационная нижняя оценка сигмoids ненормированными гауссианами (слева). Сигмoids и их произведение (справа)

Итак, будем искать вариационную нижнюю оценку для сигмoids. Заметим, что если функция выпуклая, то ее вариационная нижняя оценка есть все ее касательные²⁰. Логисти-

¹⁹Заметим, 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:

$$\begin{aligned} \log \sigma(x) &= -\log(1 + \exp(-x)) = \log\left(\exp\left(-\frac{x}{2}\right)\left(\exp\left(\frac{x}{2}\right) + \exp\left(-\frac{x}{2}\right)\right)\right) = \\ &= \frac{x}{2} - \log\left(\exp\left(-\frac{x}{2}\right) + \exp\left(\frac{x}{2}\right)\right) \end{aligned} \quad (118)$$

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

$$-\log\left(e^{-\frac{x}{2}} + e^{\frac{x}{2}}\right) = -\log\left(e^{-\frac{\sqrt{y}}{2}} + e^{\frac{\sqrt{y}}{2}}\right) \quad (119)$$

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

$$\frac{d\left(-\log\left(e^{-\frac{\sqrt{y}}{2}} + e^{\frac{\sqrt{y}}{2}}\right)\right)}{dy} = -\tanh\left(\frac{\sqrt{y}}{2}\right) \frac{1}{4\sqrt{y}}. \quad (120)$$

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

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

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

$$\sigma(x) \geq \exp\left(\frac{x}{2} - \frac{1}{4|\eta|} \tanh\left(\frac{|\eta|}{2}\right)(x^2 - \eta^2) - \log\left(e^{-\frac{|\eta|}{2}} + e^{\frac{|\eta|}{2}}\right)\right) = \quad (122)$$

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

$$= \sigma(\eta) \exp\left(\frac{x - \eta}{2}\right) \exp\left(-\frac{1}{4\eta} \tanh\left(\frac{\eta}{2}\right)(x^2 - \eta^2)\right), \quad (124)$$

где мы убрали модули у второго и третьего слагаемого под экспонентой, т.к. эти функции четные, и воспользовались выражением 118.

Как мы говорили ранее, полученная оценка²¹, 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 ковариации 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.

²¹Эта 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 7. Вариационный Байесовский вывод

6.1 ЕМ-алгоритм

ЕМ-алгоритм и его возможности — одна из самых важных тем этого курса. Далеко не все аналитики данных в совершенстве владеют ЕМ-подобными процедурами, в то время как с их помощью порой можно извлечь из данных даже «больше» информации, чем, казалось бы, в них содержится исходно...

6.1.1 Пример применения ЕМ-алгоритма на практике

Рассмотрим пример совмещения ЕМ-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 ЕМ-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: «У банка поверните направо».
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.

6.1.2 Вспоминаем ЕМ-алгоритм

Дана модель с наблюдаемыми переменными X и латентными переменными Z , параметризованная вектором θ :

$$p(X, Z | \theta). \quad (125)$$

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:

$$\theta_{ML} = \arg \max_{\theta} p(X | \theta) = \arg \max_{\theta} \int p(X, Z | \theta) dZ. \quad (126)$$

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 зачастую мы умеем считать только значение совместной плотности (125) для данных X и Z и не можем посчитать неполное правдоподобие в данной точке X (не можем посчитать интеграл в правой части (126)).

Пример. Латентные переменные естественно возникают в случае, когда плотность наблюдаемых переменных $p(X | \theta)$ имеет очень сложный характер. Тогда один из способов упрощения задачи — добавление латентных переменных до тех пор, пока совместное распределение (125) 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 (126) к выпуклой путём добавления латентных переменных. Перейдём к логарифму:

$$\theta_{ML} = \arg \max_{\theta} p(X | \theta) = \arg \max_{\theta} \log p(X | \theta). \quad (127)$$

Логарифм неполного правдоподобия можно разложить на вариационную нижнюю оценку и KL-дивергенцию между *вариационным распределением* $q(Z)$ на латентные переменные и апостериорным распределением $p(Z | X, \theta)$ при заданных θ :

$$\log p(X | \theta) = \mathcal{L}(q, \theta) + KL(q(Z) \| p(Z | X, \theta)), \quad \forall q(Z). \quad (128)$$

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*:

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

Отсюда возникает итерационный ЕМ-алгоритм:

E-step

$$q_n(Z) = \arg \max_q \mathcal{L}(q, \theta_n) = p(Z | X, \theta_n) \quad (129)$$

M-step

$$\theta_{n+1} = \arg \max_{\theta} \mathcal{L}(q_n, \theta) = \arg \max_{\theta} \mathbb{E}_{q_n(Z)} \log p(X, Z | \theta) \quad (130)$$

В последнем равенстве мы воспользовались определением вариационной нижней оценки:

$$\mathcal{L}(q, \theta) = \int q(Z) \log \frac{p(X, Z | \theta)}{q(Z)} dZ = \int q(Z) \log p(X, Z | \theta) dZ - \int q(Z) \log q(Z) dZ. \quad (131)$$

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

Преимущества такой процедуры:

1. На Е-шаге можем выполнить пересчёт в явном виде (если умеем считать апостериорное распределение на Z).
2. На М-шаге возникает задача оптимизации $\mathbb{E}_{q_n(Z)} \log p(X, Z | \theta)$ — вогнутой функции по θ , так как $\log p(X, Z | \theta)$ 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.

6.1.3 Модификация модели ЕМ: априорное распределение на веса

Модифицируем модель (125). Предположим, что вероятностная модель полностью Байесовская, то есть задано совместное распределение на X, Z, θ :

$$p(X, Z, \theta) = p(X, Z | \theta) p(\theta). \quad (132)$$

Предположим, что мы хотим решать задачу не поиска оценки максимального неполного правдоподобия (126), а максимума апостериорного распределения:

$$\theta_{MP} = \arg \max_{\theta} p(\theta | X) = \arg \max_{\theta} \log p(\theta | X) = \arg \max_{\theta} [\log p(X | \theta) + \log p(\theta)]. \quad (133)$$

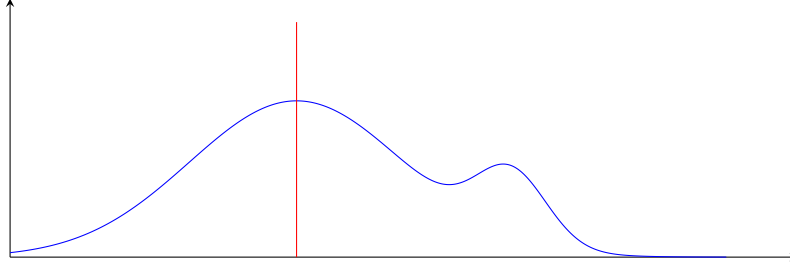


Рис. 8: Пример аппроксимации распределения в семействе дельта-функций.

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 ЕМ-алгоритм?

Выражение (128) примет вид:

$$\log p(X | \theta) + \log p(\theta) = \mathcal{L}(q, \theta) + KL(q(Z) \parallel p(Z | X, \theta)) + \log p(\theta). \quad (134)$$

На **М-шаге** (130) возникнет ещё одно аддитивное слагаемое:

$$\theta_{n+1} = \arg \max_{\theta} \mathcal{L}(q_n, \theta) + \log p(\theta) = \arg \max_{\theta} [\mathbb{E}_{q_n(Z)} \log p(X, Z | \theta) + \log p(\theta)]. \quad (135)$$

На **Е-шаге** мы максимизируем по q при фиксированном θ . Добавленное слагаемое не зависит от q , поэтому Е-шаг (129) не изменится.

Таким образом, ЕМ-алгоритм практически не меняется.

6.1.4 От ЕМ-алгоритма к вариационному выводу

Что будет, если на Е-шаге распределения не сопрягаются и мы не можем точно выполнить Байесовский вывод? Придётся выполнять его приближённо. Заметим, что на Е-шаге (129) можно эквивалентно минимизировать KL-дивергенцию:

$$q(Z) = \arg \max_q \mathcal{L}(q, \theta_n) = \arg \min_q KL(q(Z) \parallel p(Z | X, \theta_n)). \quad (136)$$

Проблема: для минимизации KL-дивергенции (136) мы должны уметь её считать, но мы не знаем $p(Z | X, \theta_n)$. 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.

Пример. Что будет, если мы ограничим семейство распределений q , к примеру, множеством дельта-функций? То есть захотим аппроксимировать $p(Z | X, \theta_n)$ в классе дельта-функций, минимизируя KL-дивергенцию между аппроксимацией и исходным распределением. Ранее мы уже выяснили, что для этого нужно взять точку в моде этого распределения (рис. 8). С точки зрения KL-дивергенции это самая репрезентативная точка. Если рассматривать другие дивергенции, ответ может меняться.

6.2 Вариационный Байесовский вывод: mean-field аппроксимация

Mean-field аппроксимация (теория среднего поля) была разработана физиками для решения полевых задач. Является частным случаем более общего подхода, который носит название *вариационный Байесовский вывод* (Байес для среднего класса, если угодно).

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:

$$p(X, Z). \quad (137)$$

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

$$q(Z) = \arg \min_{q \in Q} KL(q(Z) \parallel p(Z | X)). \quad (138)$$

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 (факторизация) и будем рассматривать лишь *факторизованные* распределения q :

$$Z = \sqcup_{i=1}^l z_i; \quad z_i \cap z_j = \emptyset; \quad q(Z) = \prod_{i=1}^l q_i(z_i). \quad (139)$$

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 (138) на эквивалентную задачу максимизации вариационной нижней оценки:

$$q(Z) = \arg \min_{q \in Q} KL(q(Z) \parallel p(Z | X)) = \arg \max_{q \in Q} \int q(Z) \log \frac{p(X, Z)}{q(Z)} dZ. \quad (140)$$

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 (140) факторизацию (139):

$$\int \prod_{i=1}^l q_i(z_i) \log \frac{p(X, Z)}{\prod_{i=1}^l q_i(z_i)} \prod_{i=1}^l dz_i = \int \prod_{i=1}^l q_i(z_i) \log p(X, Z) dZ - \int \prod_{i=1}^l q_i(z_i) \left[\sum_{k=1}^l \log q_k(z_k) \right] dZ = \quad (141)$$

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

$$= \int \prod_{i=1}^l q_i(z_i) \log p(X, Z) dZ - \sum_{k=1}^l \int q_k(z_k) \log q_k(z_k) dz_k = \quad (142)$$

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:

$$= \int q_j(z_j) \left(\int \prod_{i \neq j} q_i(z_i) \log p(X, Z) dZ_{\neq j} \right) dz_j - \int q_j(z_j) \log q_j(z_j) dz_j + Const. \quad (143)$$

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

$$\hat{p}(z_j) \equiv \exp \left(\int \prod_{i \neq j} q_i(z_i) \log p(X, Z) dZ_{\neq j} \right). \quad (144)$$

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

$$\int \prod_{i \neq j} q_i(z_i) \log p(X, Z) dZ_{\neq j} = \log \hat{p}(z_j) \quad (145)$$

$$p(z_j) = \frac{\hat{p}(z_j)}{\int \hat{p}(z_j) dz_j} \equiv \frac{\hat{p}(z_j)}{A}; \quad \hat{p}(z_j) = A \cdot p(z_j) \quad (146)$$

После перенормировки (A — нормировочная константа) $p(z_j)$ можно рассматривать как плотность вероятности. Подставим её в (143) и объединим интегралы, при этом составляющая интеграла с константой A будет вынесена в новую константу:

$$\begin{aligned} \int q(Z) \log \frac{p(X, Z)}{q(Z)} dZ &= \dots = \int q_j(z_j) \log(Ap(z_j)) dz_j - \int q_j(z_j) \log q_j(z_j) dz_j + Const \\ &= \int q_j(z_j) \log \frac{p(z_j)}{q_j(z_j)} dz_j + Const'. \end{aligned} \quad (147)$$

Напомним, что в соответствии с (140) мы хотим максимизировать это выражение по q_j . Заметим, что если поменять числитель и знаменатель под логарифмом местами, то получим KL-дивергенцию:

$$\int q(Z) \log \frac{p(X, Z)}{q(Z)} dZ = \dots = -KL(q_j(z_j) \parallel p(z_j)) + Const'. \quad (148)$$

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 $p(z_j)$ (144) с учётом нормировки и получим финальное выражение для обновления $q_j(z_j)$:

$$q_j(z_j) = \frac{\exp(\mathbb{E}_{q(Z_{\neq j})} \log p(X, Z))}{\int \exp(\mathbb{E}_{q(Z_{\neq j})} \log p(X, Z)) dz_j} \quad (149)$$

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

$$\log q_j(z_j) = \mathbb{E}_{q(Z_{\neq j})} \log p(X, Z) + Const. \quad (150)$$

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 (148)). 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.

6.2.1 Условная сопряженность (conditional conjugacy).

Определение 6. 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 (149, 150).

Примем без доказательства, но далее рассмотрим на конкретных примерах.

6.2.2 Связь mean-field аппроксимации и ЕМ-алгоритма

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

$$Z = z_1 \sqcup z_2. \quad (151)$$

Аппроксимируем апостериорное распределение факторизованным: на q_2 введем ограничение — q_2 из семейства дельта-функций:

$$p(Z \mid X) \approx q_1(z_1)\delta(z_2 - z_2^*). \quad (152)$$

Тогда формула для пересчета q_1 (на основе (150)):

$$\log q_1(z_1) = \mathbb{E}_{z_2} \log p(X, Z) + Const = \log p(X, z_1, z_2^*) + Const \quad (153)$$

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 :

$$q_1(z_1) = \frac{p(X, z_1, z_2^*)}{A} \quad (154)$$

Мы получили Е-шаг ЕМ-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:

$$\mathbb{E}_{z_1} \log p(X, Z) \rightarrow \max_{z_2}. \quad (155)$$

Тогда получим:

$$z_2^* = \arg \max_{z_2} \mathbb{E}_{z_1} \log p(X, z_1 \mid z_2) + \log p(z_2) \quad (156)$$

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

Пример. Переходим к практике. Классический пример применения ЕМ-алгоритма — разделение смеси гауссиан. В этом случае и Е-шаг и М-шаг можно выполнить аналитически.

Введем вероятностную модель:

- X — наблюдаемые переменные, Z — индексы компонент смеси
- π — априорная вероятность компоненты смеси
- μ — матожидания 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:

$$Dir(\pi | \alpha) = \frac{1}{C(\alpha)} \prod_{k=1}^K \pi_k^{\alpha_k - 1}, \quad \alpha_k > 0. \quad (157)$$

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. Поэтому таким образом удобно вводить априорные распределения, которые поощряют зануление большинства компонент.

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

$$p(X, Z, \pi, \mu, \Lambda) = p(X | Z, \mu, \Lambda) p(Z | \pi) p(\mu, \Lambda) p(\pi) = \quad (158)$$

$$= \prod_{i=1}^n p(x_i | z_i, \mu, \Lambda) p(z_i | \pi) \prod_{k=1}^K p(\mu_k, \Lambda_k) p(\pi) = \quad (159)$$

$$= \prod_{i=1}^n \left[\mathcal{N}(x_i | \mu_{z_i}, \Lambda_{z_i}^{-1}) \prod_{k=1}^K \pi_k^{[z_i=k]} \right] \prod_{k=1}^K p(\mu_k, \Lambda_k) Dir(\pi | \alpha_0). \quad (160)$$

Проблемы с базовым ЕМ-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 покрывает 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:

$$p(\mu_k, \Lambda_k) = p(\mu_k | \Lambda_k) p(\Lambda_k) = \mathcal{N}(\mu_k | m_0, (\beta \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k | \nu, W), \quad (161)$$

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

$$\mathcal{W}(\Lambda_k | \nu, W) = \frac{1}{C(\nu, W)} (det \Lambda_k)^{\frac{\nu-d-1}{2}} \exp \left(-\frac{1}{2} tr(W^{-1} \Lambda_k) \right), \quad \nu > d-1, \quad W = W^T \succ 0. \quad (162)$$

Матожидание распределение Уишарта:

$$\mathbb{E} \Lambda_k = \nu W \quad (163)$$

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 — камера. Качество съемки — не очень, каждая мышь — серый комочек. По отдельности трекал мышек еще получается, но когда мышки сбиваются в кучки — уже нет.

Первоначальная модель: отдельных мышек параметризовывали эллипсом. Упрощая, можно сказать, что вписывали гауссиану. Когда мышки собирались вместе получалась смесь гауссиан. 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 выбрали за несколько итераций), удалось качественно разделять мышей даже в «слипшихся» группах.

Подставим $p(\mu_k, \Lambda_k)$ в вероятностную модель (158):

$$\prod_{i=1}^n \left[\mathcal{N}(x_i \mid \mu_{z_i}, \Lambda_{z_i}^{-1}) \prod_{k=1}^K \pi_k^{[z_i=k]} \right] \prod_{k=1}^K \mathcal{N}(\mu_k \mid m_0, (\beta \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k \mid \nu, W) \text{Dir}(\pi \mid \alpha_0). \quad (164)$$

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

$$p(Z, \pi, \mu, \Lambda \mid X). \quad (165)$$

Есть ли сопряженность на все переменные: сопряжены ли $p(X \mid Z, \pi, \mu, \Lambda)$ и $p(Z, \pi, \mu, \Lambda)$? Если есть, то сразу выпишем, как будет выглядеть апостериорное. Без доказательства примем, что полного сопряжения не будет. Но есть условное сопряжение!

Можно показать, что условное сопряжение имеется на Z и (μ, Λ, π) , а именно, если зафиксируем (μ, Λ, π) , то будет сопряженность на Z , и наоборот.

Слегка перепишем (164):

$$\prod_{i=1}^n \prod_{k=1}^K [\mathcal{N}(x_i \mid \mu_k, \Lambda_k^{-1}) \pi_k]^{[z_i=k]} \frac{1}{C(\alpha_0)} \left(\prod_{k=1}^K \pi_k^{\alpha_0-1} \right) \left(\prod_{k=1}^K \mathcal{N}(\mu_k \mid m_0, (\beta \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k \mid \nu, W) \right). \quad (166)$$

Если мы зафиксировали все Z , то что можно сказать про π ?

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

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: $\mathcal{N}(\mu_k \mid m_0, (\beta \Lambda_k)^{-1}) \mathcal{W}(\Lambda_k \mid \nu, W)$
- Правдоподобие: $\mathcal{N}(x_i \mid \mu_k, \Lambda_k^{-1})$

Для многомерной гауссианы Уишерт-нормальное распределение является сопряженным. При фиксированном Z мы знаем, из какой гауссианы какой объект пришел.

Совместное сопряжение тоже есть, так как π и (μ, Λ) входят в разные множители. Значит, на (μ, Λ, π) есть сопряжение при фиксированном Z .

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

- Априорное распределение: $\pi_k^{[z_i=k]}$

Таблица 3: Методы Байесовского вывода.

Свойства	Метод вывода	Вид аппроксимации
$(Z, \theta) — \text{conjugacy}$	Bayes Theorem	$p(Z, \theta X)$
Conditional conjugacy Z & θ	Mean-Field	$q(Z)q(\theta)$
$Z — \text{conjugacy}$	ЕМ для оптимизации $p(\theta X)$	$p(Z X, \theta)\delta(\theta - \theta_{MP})$
$\theta — \text{conjugacy}$	МЕ для оптимизации $p(Z X)$	$p(\theta X, Z)\delta(Z - Z_{MP})$
Conditional z_j -conjugacy	Variational (mean-field) EM	$\prod_{i=1}^l q_i(z_i)\delta(\theta - \theta_{MP})$
No conjugacy	Poor man's Bayes	$\delta(Z - Z_{MP})\delta(\theta - \theta_{MP})$

- Правдоподобие: $\mathcal{N}(x_i | \mu_k, \Lambda_k^{-1})^{[z_i=k]}$

Оба распределения имеют вид произведения каких-то элементов в степени индикатора от z , поэтому при перемножении снова получим произведение каких-то элементов в степени индикатора от z , то есть на z тоже есть сопряженность (при фиксировании (μ, Λ, π)).

Вывод: можем апостериорное распределение (165) приблизить mean-field аппроксимацией:

$$p(Z, \pi, \mu, \Lambda | X) \approx q_1(Z)q_2(\mu, \Lambda, \pi). \quad (167)$$

Далее необходимо просто применить формулу (150): расписать логарифм правдоподобия и проматожидать по Z , чтобы найти q_2 , и по (μ, Λ, π) , чтобы найти q_1 .

6.3 Концептуальная схема

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

$$p(X, Z, \theta) = p(X, Z | \theta)p(\theta). \quad (168)$$

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:

1. Есть полная сопряженность на (Z, θ) . Если расписать $p(X, Z, \theta) = p(X | Z, \theta)p(Z, \theta)$, то у нас имеется полная сопряженность между функцией правдоподобия того, что мы наблюдаем, $p(X | Z, \theta)$ и априорным распределением $p(Z, \theta)$ на то, что мы не наблюдаем. Есть полная сопряженность, значит, можем получить апостериорное распределение $p(Z, \theta | X)$ 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 $q(Z)q(\theta) = \arg \min KL(q(Z)q(\theta) \parallel p(Z, \theta | X))$.
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 по сути ЕМ-алгоритм.

4. Сопряжение только на θ . Исходя из симметрии Z и θ , это будет МЕ-алгоритм: Е-шаг на θ , М-шаг на Z .
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 — аппроксимируем δ -функцией (ищем максимум апостериорной плотности). На Z — mean-field аппроксимация. Таким образом, мы применяем ЕМ-алгоритм, где на М-шаге пересчитываем точку максимума по θ , а на Е-шаге вместо честного Байесовского вывода применяем mean-field аппроксимацию.
Аналогично можем расписать условную θ -сопряженность и получить вариационный МЕ-алгоритм.
6. Нет никакой сопряженности. δ -функция для всего — Байес для бедных.

7 Лекция 8. Методы Монте-Карло по схеме марковский цепей (МСМС)

7.1 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 такой статистики является определение следующих распределений:

1. $\int p(T | X, \omega)p(\omega)d\omega = p(T | X)$, где T – имеющаяся на этапе обучения информация о предсказываемых данных, X – обучающая выборка, ω – параметры распределения.
2. $\int p(t | X, \omega)p(\omega | X, T)d\omega = p(t | x, X, T)$, где x – объект, для которого хотим получить прогноз t .

Соответственно, если мы не можем посчитать аналитически этот интеграл, то применяют численные методы. Однако классические методы возрастают экспоненциально по сложности при увеличении размерности пространства.

При использовании ЕМ-алгоритма и/или Mean Field аппроксимации мы получали смещённое приближение. Можно ли как-то решить задачу интегрирования в высокоразмерном пространстве с использованием несмещённой оценки? Да, на помощь приходит метод Монте-Карло:

Приближим исходный интеграл, произведя сэмплирование из распределения $p(x)$:

$$\int f(x)p(x)dx \approx \frac{1}{N} \sum_{i=1}^N f(x_i) = J, \quad x_i \sim p(x) \quad (169)$$

Видно, что J – случайная величина. Посчитаем её статистики:

$$\mathbb{E} J = \mathbb{E} \frac{1}{N} \sum_i f(x_i) = \frac{1}{N} \sum_i \mathbb{E} f(x_i) = \frac{1}{N} \mathbb{E} f(x) = \mathbb{E} f(x), \quad \text{т.е. метод несмещённый.} \quad (170)$$

$$\mathbb{D} J = \mathbb{D} \frac{1}{N} \sum_i f(x_i) = \frac{1}{N^2} \sum_i \mathbb{D} f(x_i) = \frac{1}{N} \mathbb{D} f(x) = \frac{1}{N} \int p(x)(f(x) - \mathbb{E} f(x))^2 dx \quad (171)$$

И, как следствие из ЦПТ, $J \sim \mathcal{N}(J | \mathbb{E} f(x), \frac{1}{N} \mathbb{D} f(x))$. Следовательно, чем меньше дисперсия, 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 приближения²². Если же $f(x)$ флуктуирует, то для точной оценки понадобится больше сэмплов.

Помимо отмеченных выше примеров, метод Монте-Карло может понадобиться в:

1. Теореме Байеса (для вычисления знаменателя)
2. Моделях байесовского ансамблирования: Если мы получаем какое-то апостериорное распределение, то далее можно взять усреднение алгоритмов ансамблирования по этому апостериорному распределению. Т.е. на этапе обучения делаем байесовский вывод на параметры, а на этапе валидации по апостериорному распределению проводим голосование через взвешенную сумму.

Следовательно, задача подсчёта такого интеграла сводится к генерации выборок из распределения.

²²Предельный случай - $f(x) \equiv \text{const} \Rightarrow \mathbb{D} f(x) = 0$, для приближения можно взять одну точку

7.2 Общие подходы и методы генерации выборок из одномерных распределений

7.2.1 Простейшие методы

В любой язык программирования заложена возможность генерации случайной величины, равномерно распределённой на $[0, 1]$. Пусть $\xi \sim \mathcal{U}[0, 1]$; $\mathbb{E} \xi = \frac{1}{2}$; $\mathbb{D} \xi = \frac{1}{12}$

Если мы хотим далее получить случайную величину $\eta \sim \mathcal{U}[a, b]$, то необходимо произвести следующее преобразование: $\eta = \xi(b - a) + a$

Научившись генерации равномерного распределения, возникает желание генерировать случайные величины из стандартно-нормального распределения²³. В основном, такая генерация основывается на гарантиях ЦПТ. 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 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:

$$p(x | \lambda) = \lambda e^{-\lambda x}, x \geq 0, \lambda > 0 \quad (172)$$

$$F(x) = 1 - e^{-\lambda x} = \xi \Rightarrow x = -\frac{1}{\lambda} \ln(1 - \xi) \quad (173)$$

2. Распределение Коши – имеет тяжёлые хвосты, поэтому не существует моментов²⁴. Возможно только определение моментов в смысле главного значения.

$$p(x) = \frac{1}{\pi} \frac{1}{1 + x^2} \quad (174)$$

$$F(x) = \frac{1}{\pi} \arctg(x) + \frac{1}{2} = \xi \Rightarrow x = \tg(\pi(\xi - \frac{1}{2})) \quad (175)$$

7.2.2 Метод Rejection Sampling

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

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. 9). Тогда будем принимать точку, сгенерированную из $q(x)$, только с определённой вероятностью, отражающей приближение.

²³В 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.

²⁴Этот 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

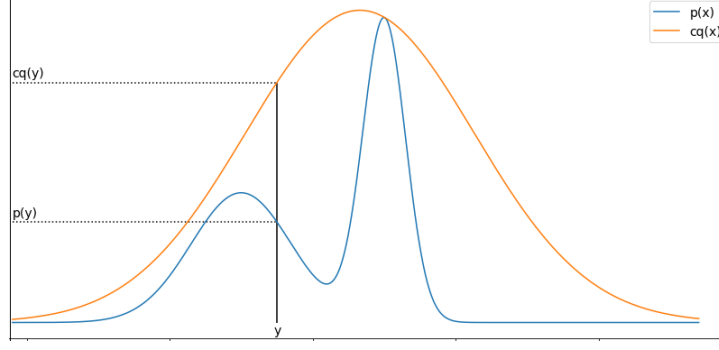


Рис. 9: Пример мажорирования плотностью в Rejection Sampling

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

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:

$$\mathbb{P}\{\text{accept} \mid y \in [a, a + \epsilon]\} = \int \mathbb{P}\{\text{accept} \mid y\} \mathbb{P}\{y \mid y \in [a, a + \epsilon]\} dy = \quad (176)$$

$$= \int_a^{a+\epsilon} \frac{p(y)}{cq(y)} \frac{q(y)}{\underbrace{\int_a^{a+\epsilon} q(y) dy}_{\text{Перенормировка на интервале}}} dy = \{\text{В предположении малости } \epsilon\} = \quad (177)$$

$$= \frac{p(a)}{cq(a)} \frac{\int_a^{a+\epsilon} q(y) dy}{\int_a^{a+\epsilon} q(y) dy} = \frac{p(a)}{cq(a)} \quad (178)$$

Тогда наша схема будет работать, если $\mathbb{P}\{x \in [a, a + \epsilon]\} = p(a)\epsilon$. Действительно,

$$\mathbb{P}\{x \in [a, a + \epsilon]\} = \mathbb{P}\{y \in [a, a + \epsilon] \mid \text{accept}\} = \frac{\mathbb{P}\{\text{accept} \mid y \in [a, a + \epsilon]\} \mathbb{P}\{y \in [a, a + \epsilon]\}}{\mathbb{P}\{\text{accept}\}} \quad (179)$$

Где

$$\mathbb{P}\{\text{accept}\} = \int \mathbb{P}\{\text{accept} \mid y\} \mathbb{P}\{y\} dy = \int \frac{p(y)}{cq(y)} q(y) dy = \frac{1}{c} \quad (180)$$

Следовательно,

$$\mathbb{P}\{x \in [a, a + \epsilon]\} = \frac{\frac{p(a)}{cq(a)} \mathbb{P}\{y \in [a, a + \epsilon]\}}{\frac{1}{c}} = \frac{p(a)}{q(a)} \int_a^{a+\epsilon} q(y) dy = \quad (181)$$

$$= \{\text{Из-за малости } \epsilon \text{ приближаем теоремой о среднем}\} = \frac{p(a)}{q(a)} q(a) \epsilon = p(a) \epsilon \quad (182)$$

■

Как видно из формул, качество приближения определяется отношением площадей под плотностями, которое равно $\frac{1}{c}$ (см. формулу 180). 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. 10)

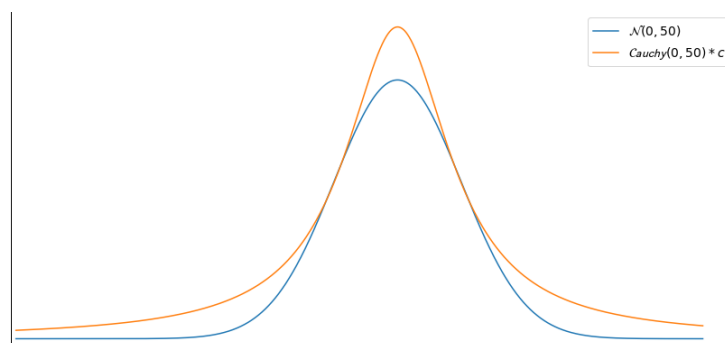


Рис. 10: Пример плохого приближения методом Rejection Sampling

В данном случае, мы сколь угодно точно можем приблизить центральную часть распределения, однако, из-за того, что распределение Коши имеет тяжёлые хвосты, нам понадобится $c \rightarrow \infty$. 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.11), то в приближении мы будем генерировать много точек, хотя распределение задано, фактически, только в одной точке.

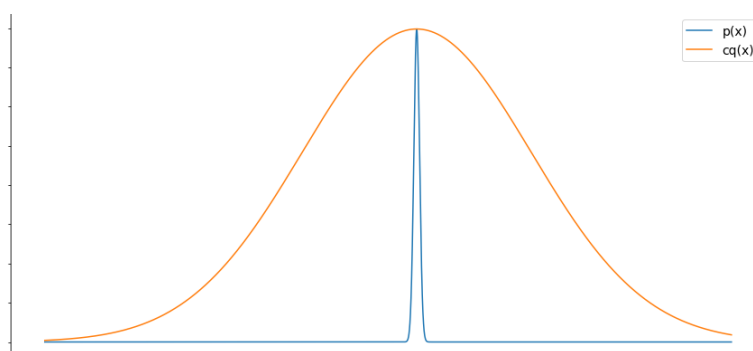


Рис. 11: Пример приближения в одной точке Rejection Sampling

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 (однако это не очень хорошая идея, т.к. не умеем считать площадь под ним). 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).

7.2.3 Метод Importance sampling

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

$$\int p(x)f(x)dx = \int q(x)\frac{p(x)}{q(x)}f(x)dx \approx \frac{1}{N} \sum_{i=1}^N \underbrace{\frac{p(x_i)}{q(x_i)}}_{v_i} f(x_i) = \sum_{i=1}^N v_i f(x_i) \quad (183)$$

Где v_i и задаёт *важность* объекта $x_i \sim q(x)$. Снова получаем несмещённую оценку, но теперь берём все точки. Однако если v_i задаются сильно равномерно (или близко к равномерному (см. рис. 12)), то большие значения распределения будут вносить бóльший вклад.

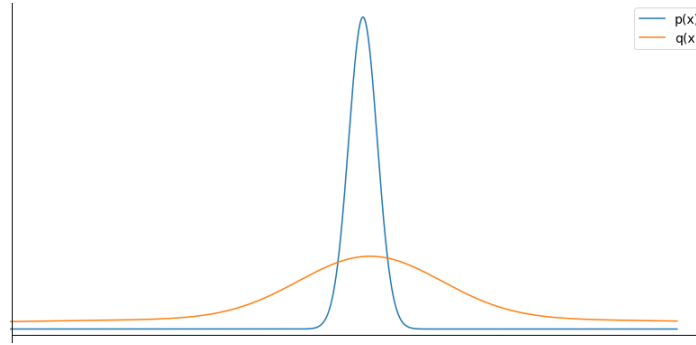


Рис. 12: Пример плохого выбора $q(x)$ в Importance sampling

Когда же важности будут распределены равномерно? 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, что важно для любого метода Монте-Карло, т.к. это медленно сходящиеся методы, в отличие от вариационной аппроксимации²⁵.

Плотность распределения $p(x)$ обычно представима в виде: $\frac{\hat{p}(x)}{Z}$. 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 нормировочную константу, то имеем дело только с $\hat{p}(x)$. Проверим, усложняет ли это схему:

$$\frac{1}{Z} \int \hat{p}(x)f(x)dx = \frac{\int q(x)\frac{\hat{p}(x)}{q(x)}f(x)dx}{\int \hat{p}(x)\frac{q(x)}{q(x)}f(x)dx} \approx \frac{\frac{1}{N} \sum_{i=1}^N \frac{\hat{p}(x_i)}{q(x_i)} f(x_i)}{\frac{1}{N} \sum_{i=1}^N \frac{\hat{p}(x_i)}{q(x_i)}} = \sum_{i=1}^N w_i f(x_i) \quad (184)$$

, где $w_i = \frac{\frac{\hat{p}(x_i)}{q(x_i)}}{\sum_{j=1}^N \frac{\hat{p}(x_j)}{q(x_j)}}$, $w_i > 0$, $\sum_i w_i = 1$. При распределении $q(x)$ далеко от $p(x)$ (см.

рис. 12) и многие w_i будут близки к 0, но будет много сэмплов, не попавших точно 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.

Однако, при работе с ненормированной плотностью, по сложности метод изменился несильно, как это можно заметить.

²⁵Имеем 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*

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.

7.3 Метод Метрополиса-Хастингса

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

Определение 7. *Марковской цепью* называется процесс порождения случайных величин (упорядоченных в виде некоторой цепочки), совместное распределение которых имеет вид:

$$p(x_1, \dots, x_n, \dots) = p_1(x_1)q_2(x_2 | x_1)q_3(x_3 | x_2)\dots q_n(x_n | x_{n-1})$$

, где индексы функций $q_i(\cdot)$ означают то, что для каждого перехода эта функция различна.

В свою очередь, важными видами марковских процессов являются те, у которых функция перехода не зависит от номера перехода.

Определение 8. Марковская цепь называется *гомогенной* (или *однородной*), если

$$\forall n q_n(x_n | x_{n-1}) = q(x_n | x_{n-1})$$

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:

$$p_2(x_2) = \int p_{12}(x_1, x_2) dx_1 = \int \underbrace{q_2(x_2 | x_1)}_{\text{Везде далее цепочки однородные}} p_1(x_1) dx_1$$

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

Эта закономерность повторяется и для дальнейших плотностей, например,

$$p_3(x_3) = \int q_3 x_3(x_3 | x_2) p_2(x_2) dx_2 = \int \int q(x_3 | x_2) q(x_2 | x_1) p_1(x_1) dx_1 dx_2$$

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 9. $p_0(x)$ называется *инвариантным распределением* (или *стационарным*) для гомогенной марковской цепи, если $p_0(x') = \int q(x' | x)p_0(x)dx \Rightarrow x'$ – стационарная точка.

Логично, что если мы оказались в стационарной точке, то уже не выйдем из неё.

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

Пример. Пусть $q(x' | x) = \delta(x' - x)$. Тогда любое распределение является инвариантным, т.к. взяв одну точку дальше повторяем её постоянно.

Пример. $q(x' | x) = \delta(x' - x - a)$. Тогда не имеем вообще инвариантных распределений.

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

Определение 10. Цепи, для которых существует и единственное инвариантное распределение будем называть *эргодическими*²⁶

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 11. Если $\forall x', x'' \in \mathcal{X}$, где \mathcal{X} – носитель марковской цепи²⁷, справедливо, что $q(x'' | x') > 0$, то для $\forall p_1(\cdot) \exists! p_0(\cdot)$ (стационарное распределение), такое, что если в момент n случайная величина x_n имела распределение p_0 , то и в $n+1$ x_{n+1} будет иметь такое же распределение.

Или, формальнее,

$$\forall n \geq N \int p(x_{n+1} | x_n) p_0(x_n) dx_n = p_0(x_{n+1})$$

Что же это значит для нас? Пусть мы захотим генерировать выборку из $p(x)$. Если мы сможем построить однородную марковскую цепь с таким 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 $p(x)$.

Но есть одна деталь – как научиться определять стационарное распределение для данной марковской цепи? Если доступ только к функции переходов $q(x' | x)$ (которая, как мы уже выяснили, является ядром линейного оператора в непрерывном случае, а в дискретном принимает вид матрицы). В дискретном случае могли бы найти собственный вектор, отвечающий единичному собственному значению. А в непрерывном случае – его аналог в Гильбертовом пространстве. На практике это довольно сложно осуществить, т.к., например, x может принимать континуум значений.

Но для некоторых частных случаев всё становится явно проще:

Теорема 6. (*Уравнение детального баланса Detailed Balance*) Если

$$\forall x, x' \in \mathcal{X} p_0(x) q(x' | x) = p_0(x') q(x | x')$$

, то p_0 – инвариантное распределение.

Доказательство. Т.е. если $x \sim p_0(x)$ и есть вероятность сгенерировать следующую случайную величину x' , то такое распределение будет у x' . Если $p(x') = p(x)$, то всё, p_0 – инвариантное.

Распишем $p(x')$:

$$p(x') = \int q(x' | x) p_0(x) dx = \{\text{Уравнение DB}\} = \int p_0(x) q(x | x') dx = p_0(x') \blacksquare$$

Почему это уравнение называется уравнением "баланса"? Пусть x, x' – страны, p_0 – ВВП страны, а $q(x' | x)$ – то, какую долю ВВП страна x инвестирует в x' . Т.е. $q(x' | x) p_0(x)$ – сумма (напр. в \$), которую x тратит на x' , а $q(x | x') p_0(x')$ – 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.

Почему уравнение называется "детальным"? Дело в том, что при детальном балансе (см. рис. 13, где А, В и С – страны) каждая страна тратит одинаковую сумму денег и всё сбалансировано. А есть модель глобального баланса (см. рис. 14), где определённые страны тратят больше денег на определённые страны, но, в целом, система сбалансирована.

²⁶В целом, в теории марковских цепей определение эргодичности цепи даётся в более сложном виде. Однако, для наших целей будет достаточно такого определения.

²⁷Т.к. можем рассмотреть формально марковскую цепь только на множестве неотрицательных случайных величин

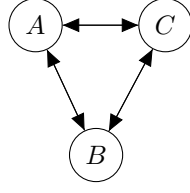


Рис. 13: Модель детального баланса

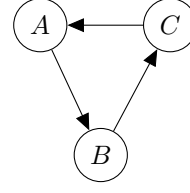


Рис. 14: Модель глобального баланса

Итак, опишем процедуру метода Метрополиса-Хастингса ²⁸:

Теорема 7. (Метод Метрополиса-Хастингса): Пусть имеется распределение $\hat{p}(x)$ с точностью до нормировочной константы и марковская цепь с $q(y | x) > 0 \forall x, y \in \mathcal{X}$. Тогда, начиная с $y \sim q(y | x)$ ²⁹ и продолжая далее брать $y \sim q(y | x_n)$, со следующей процедурой выбора следующего сэмпла:

$$x_{n+1} = \begin{cases} y, & \text{with probability } A(x_n, y) = \min \left(1, \frac{\hat{p}(y)q(x_n|y)}{\hat{p}(x_n)q(y|x_n)} \right) \\ x_n, & \text{else} \end{cases} \quad (185)$$

гарантируется, что $\forall n \geq N \ x_n \sim \hat{p}(x)$ ³⁰

Доказательство. Посмотрим, для начала, на формулу под \min : Если рассмотреть симметричное ³¹ $q(y | x) = q(x | y)$ ³². Тогда берём y с вероятностью $A(x, y) = \min \left(1, \frac{\hat{p}(y)}{\hat{p}(x_n)} \right)$. Т.о. блуждаем по области распределения, выбирая точку с вероятностью улучшения, по сравнению с предыдущей точкой. Если бы мы принимали только шаги с улучшениями, то рано или поздно стабилизировались бы в моде распределения $\hat{p}(x)$, а мы хотим сэмплы, а не оптимальные значения.

Итак, такой итеративный процесс определяет марковскую цепь. Для доказательства теоремы достаточно показать, что для неё верно уравнение детального баланса по распределению $\hat{p}(x)$. Тогда оно будет инвариантным.

Для этого рассмотрим вероятность перехода из x в y :

$$r(y | x) = \underbrace{q(y | x)}_{\text{Генерация } y \text{ из } q(y | x)} \cdot \underbrace{A(x, y)}_{\text{Вероятность его принятия}} + \quad (186)$$

$$+ \underbrace{\delta(y - x)}_{\text{Вероятность перехода в точку}} \cdot \int \underbrace{(1 - A(x, z))}_{\text{Вероятность остаться в старой точке, если сгенерировали } z} \times \quad (187)$$

$$\times \underbrace{q(z | x)dz}_{\text{Усредняем } \Rightarrow \text{ маргинальная вероятность}} \quad (188)$$

Перепишем формулу без лишних комментариев:

$$r(y | x) = q(y | x) \cdot A(x, y) + \delta(y - x) \int (1 - A(x, z))q(z | x)dz \quad (189)$$

Верно ли для неё уравнение детального баланса для функции $p(x)$?

$$p(x)r(y | x) \stackrel{?}{=} p(y)r(x | y) \quad (190)$$

²⁸Впервые метод разработал Метрополис, который работал над Манхэттенским проектом по созданию ядерного оружия, где возникла необходимость брать большое число математических ожиданий на первых компьютерах. После метод был модифицирован Хастингсом

²⁹ $q(y | x)$ называется предложенным распределением proposal distribution

³⁰При сохранении предыдущего значения, мы добавляем элемент повторно

³¹Именно таковыми были предпосылки Метрополиса. Хастингс вывел для несимметричного случая

³²Например, нормальное распределение с $\mathbb{E}y = x$ и дисперсией – любой изотропной ковариационной матрицей

Рассмотрим левую часть уравнения 190:

$$p(x)q(y | x) \cdot A(x, y) + p(x)\delta(y - x) \int (1 - A(x, z))q(z | x)dz = \quad (191)$$

$$= p(x)q(y | x) \min \left(1, \underbrace{\frac{\hat{p}(y)q(x | y)}{\hat{p}(x)q(y | x)}}_{\text{Можем перейти к нормированным плотностям, т.к. константы сокращаются}} \right) + \quad (192)$$

$$+ \underbrace{p(x)\delta(y - x) \int (1 - A(x, z))q(z | x)dz}_{\text{Выражение определено только когда } x = y \Rightarrow \text{меняем везде } x \text{ на } y} = \quad (193)$$

$$= \{\text{Занесём } p(x)q(y | x) \text{ в минимум}\} = \min(p(x)q(y | x), p(y)q(x | y)) + \quad (194)$$

$$+ p(y)\delta(x - y) \int (1 - A(y, z))q(z | y)dz = p(y)q(x | y) \min \left(\frac{p(x)q(y | x)}{p(y)q(x | y)}, 1 \right) + \quad (195)$$

$$+ p(y)\delta(x - y) \int (1 - A(y, z))q(z | y)dz = p(y)r(x | y) \quad (196)$$

Следовательно, уравнение детального баланса выполнено, значит, моделируемый итеративный процесс моделирует гомогенную марковскую цепь. И тогда процесс гарантированно сойдётся к своему инвариантному распределению из любого приближения ■

Оценка плюсов и минусов схемы:

1. Схема эффективна \Leftrightarrow точки y , которые генерируем как кандидата с высокой вероятностью принять. Иначе будет довольно неэффективная выборка с большим числом коррелирующих случайных величин, пока не сойдёмся. В принципе, это недостаток МСМС методов – высокая скоррелированность соседних случайных величин, а хотели бы выборку (Н.О.Р.С.В.).

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;

2. Схема является относительно долгой в плане выхода в стационарную точку;
3. Метод является несмещённым;
4. Нет общей идеи понимания, что вышли на стационарную точку. В основном, берут статистики для случайной подвыборки.

8 Лекция 9а. Гамильтонов Монте-Карло

8.1 Общие предпосылки метода гамильтонова Монте-Карло

В этой лекции мы продолжим рассматривать задачу генерации случайной выборки из распределения $p(x)$, заданного с точностью до нормировочной константы

$$p(x) = \frac{\hat{p}(x)}{Z}. \quad (197)$$

В прошлой лекции была предложена общая схема Метрополиса-Хастингса, в которой следующее состояние марковской цепи определялось с помощью предложного распределения. В случае непрерывного пространства состояний, простейшим вариантом могло бы служить нормальное распределение вида

$$q(x' | x) = N(y | x, \sigma I). \quad (198)$$

Такое предложное распределение не использует информации о заданной плотности $\hat{p}(x)$ и может приводить к неэффективной процедуре генерации. Во-первых, если плотность распределения $p(x)$ сконцентрирована вдоль подпространства малой размерности, большая часть сгенерированных из предложного распределения точек будет иметь низкую вероятность принятия.³³ Во-вторых, если распределение $p(x)$ имеет несколько мод на расстоянии, превосходящем характерную длину шага предложного распределения $q(x' | x)$, схеме может потребоваться значительное время чтобы сгенерировать точки из обеих мод.

Для распределений с непрерывным пространством состояний и дифференцируемой плотностью $p(x)$ гамильтонов Монте-Карло (Hamiltonian Monte-Carlo)³⁴ строит предложное распределение, способное нащупать области высокой плотности в пространстве состояний \mathcal{X} и обойти обозначенные выше трудности.

8.2 Описание классического гамильтонова Монте-Карло

Для определения предложного распределения гамильтонов Монте-Карло вводит вспомогательную механическую систему, в которой пространство состояний \mathcal{X} играет роль координат. Ниже мы приведем необходимые для обоснования метода факты из гамильтоновой механики и дадим подробное описание метода.

8.2.1 Гамильтонова механика

Гамильтонова механика - формулировка классической механики, описывающая систему с помощью функции Гамильтона (гамильтониана) $H(x, r, t)$. Аргументами гамильтониана являются координаты точек системы $x \in \mathbb{R}^d$, их импульсы $p \in \mathbb{R}^d$ и время $t \in \mathbb{R}$. Динамика системы задается как решение системы уравнений Гамильтона

$$\dot{x} = \frac{\partial H}{\partial r} \quad (199)$$

$$\dot{r} = -\frac{\partial H}{\partial x} \quad (200)$$

В дальнейшем мы опустим зависимость от времени и рассмотрим гамильтониан вида

$$H(x, r) = \Pi(x) + K(r), \quad (201)$$

где в качестве потенциальной энергии возьмем

$$\Pi(x) = -\log \hat{p}(x), \quad (202)$$

³³Более подробное интуитивное объяснение этого аргумента можно услышать в выпуске подкаста The Talking Machines <https://www.thetalkingmachines.com/episodes/generative-art-and-hamiltonian-monte-carlo>

³⁴В литературе также встречается название Hybrid Monte-Carlo, гибридный метод Монте-Карло

а качестве кинетической энергии возьмем

$$K(x) = \frac{|r|^2}{2}. \quad (203)$$

При этих ограничениях в уравнении 200 можно узнать второй закон Ньютона, в котором сила представлена потенциальным полем $-\frac{\partial \Pi}{\partial x}$, в то время как уравнение 199 связывает обобщенный импульс r и изменение координат \dot{x} .

Важным следствием из уравнений 199, 200 является закон сохранения энергии. Действительно, рассмотрим некоторое решение системы $(x(t), r(t))$ и вычислим полную производную гамильтониана:

$$\dot{H} = \frac{\partial H}{\partial x} \dot{x} + \frac{\partial K}{\partial r} \dot{r}, \quad (204)$$

подставив уравнения 199 и 200 получим

$$\dot{H} = \frac{\partial H}{\partial x} \frac{\partial K}{\partial r} + \frac{\partial K}{\partial r} \left(-\frac{\partial H}{\partial x} \right) = 0. \quad (205)$$

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 $(x(t), r(t))$ удовлетворяет уравнениям Гамильтона, то и кривая $x(-t), -r(-t)$ будет удовлетворять этим уравнениям, что можно проверить подстановкой.

Наконец, движение вдоль решений уравнений Гамильтона сохраняет объем в фазовом пространстве (пространстве пар (x, r) , в котором заданы уравнения Гамильтона). Опустив подробные выкладки, заметим что это следует из равенства нулю дивергенции задаваемого уравнениями векторного поля

$$\operatorname{div}(\dot{x}, \dot{r}) = \sum_i \left(\frac{\partial^2 H}{\partial x_i \partial r_i} - \frac{\partial^2 H}{\partial x_i \partial r_i} \right) = 0. \quad (206)$$

8.2.2 Схема генерации точек на основе динамики Гамильтона

В этом разделе мы построим предположное распределение для алгоритма Метрополиса-Хастингса на основе динамики Гамильтона, предположив что решения системы могут быть вычислены точно.

Перейдем от задачи генерации точек x из распределения $p(x)$ к задаче генерации пар (x, r) из распределения $p(x, r) = p(x)p(r)$, в котором случайные вектора x и r независимы. Если мы построим алгоритм для генерации пар (x, r) , для генерации x будет достаточно отбросить r . Выберем в качестве распределения $p(r)$ стандартное нормальное распределение, тогда логарифм плотности распределения $\log p(x, r)$ будет равен с точностью до константы отрицательной функции Гамильтона из предыдущего раздела.

$$H(x, r) = -\log p(x, r) + \text{const} \quad (207)$$

Чтобы сгенерировать точку из предположного распределения выполним следующие шаги:

1. генерируем точку $r' \sim N(r' | 0, I)$;
2. находим решение уравнения Гамильтона $x(t), r(t)$ с начальными условиями $x(0) = x_k, r(0) = r'$
3. сдвигаем точку на время T вдоль решения и меняем знак второй компоненты: $x_{k+1} = x(T), r_{k+1} = -r(T)$ ³⁵

Сдвиг по времени T здесь может быть выбран произвольно. Для маленьких T точки из предположного распределения будут близки к стартовой точке, для больших T точки окажутся дальше. Тем не менее, как мы покажем дальше, вне зависимости от выбора T

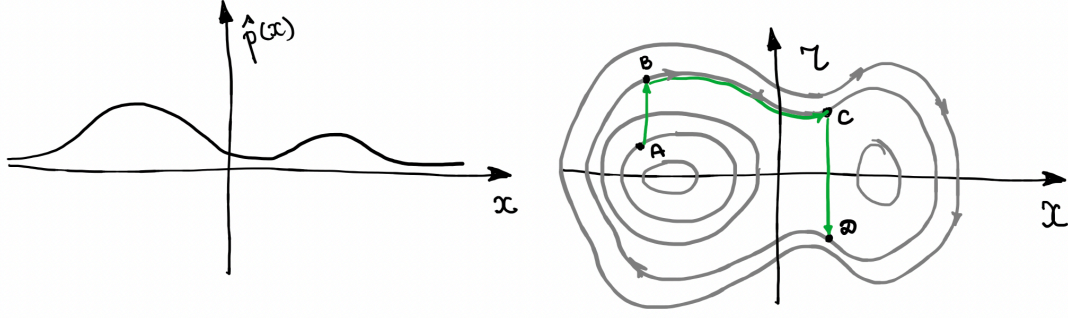


Рис. 15: Слева: Пример бимодального распределения, для которого может быть применен гибридный Монте-Карло. Справа: Решения уравнений Гамильтона на плотности (x, r) (фазовом пространстве). Для этой размерности решения совпадают с линиями уровня $H(x, r) = \text{const}$

распределение $p(x, r)$ будет инвариантным для описанной процедуры (а значит и коррекция Метрополиса-Хастингса окажется не нужна).

Иллюстрация этой схемы приведена на рисунке 15. Рассмотрим плотность бимодального распределения, заданную графиком на рисунке 15 слева. Стартуя из точки A , мы генерируем случайный импульс r и попадаем в точку B . Из точки B мы двигаемся вдоль решения до точки C , а затем переходим в точку D . Как видно из рисунка, для подходящего времени T траектория переносит систему в окрестность новой моды.

8.3 Обоснование гамильтонова Монте-Карло

В этом разделе мы покажем, что распределение $p(x, r)$ является инвариантным для предложенной в предыдущем разделе цепи. В общем случае полученная цепь может не быть эргодической, т.е. $p(x, r)$ может быть не единственным инвариантным распределением.

Два шага схемы меняют точки цепи: генерация новой точки r' и сдвиг вдоль динамики Гамильтона. Предложное распределение для первого преобразования имеет вид

$$q_1(x', r' | x, r) = \delta(x - x') \mathcal{N}(r' | 0, I), \quad (208)$$

а распределение $p(x, r)$ является инвариантным для него, поскольку для $p(r) = \mathcal{N}(r' | 0, I)$ выполнено

$$\int_{\mathcal{X} \times \mathbb{R}^d} q(x', r' | x, r) p(x, r) dx dr = \int_{\mathcal{X} \times \mathbb{R}^d} \delta(x - x') p(r') p(x, r) dx dr = p(x') p(r') = p(x', r'). \quad (209)$$

Чтобы доказать инвариантность $p(x, r)$ относительно сдвига вдоль динамики Гамильтона проверим уравнение детального баланса. Предложное распределение имеет вид

$$q_2(x, r | x', r') = \delta((x, r) - \phi_T(x', r')), \quad (210)$$

где $\phi_T(\cdot)$ означает композицию сдвига на время T вдоль решения уравнений Гамильтона и последующей смены знака r .

Поскольку $p(x, r) \propto \exp(-H(x, r))$, плотность сохраняется вдоль решения. Помимо этого, благодаря симметричности нормального распределения $p(x, r) = p(x, -r)$. Поэтому $p(x, r) = p(x', r')$. Для проверки уравнения детального баланса остается показать симметричность предложного распределения

$$q_2(x, r | x', r') = q_2(x', r' | x, r). \quad (211)$$

³⁵Смена знака у $r(T)$ не обязательна, но немного упрощает выкладки в доказательствах

Функция q_2 является обобщенной функцией, чтобы проверить это равенство необходимо сравнить действие левой и правой части уравнения на произвольную функцию $g(x, r)$, показав

$$\int_{\mathcal{X} \times \mathbb{R}^d} g(x, r) q_2(x, r \mid x', r') dr dx = \int_{\mathcal{X} \times \mathbb{R}^d} g(x, r) q_2(x', r' \mid x, r) dr dx. \quad (212)$$

По определению дельта-функции левая часть равна $g(\phi_T(x', r'))$. Для вычисления правой части сделаем замену переменных (x, r) на $y = (x', r') - \phi_T(x, r)$ для фиксированных (x', r') . После замены получим

$$\int_{\mathcal{X} \times \mathbb{R}^d} g(\phi_T^{-1}((x', r') - y)) \delta(y) \left| \frac{\partial \phi_T^{-1}}{\partial y} \right| dy = g(\phi_T^{-1}((x', r'))) \left| \frac{\partial \phi_T^{-1}}{\partial y} \right| \quad (213)$$

Заметим, что $\phi_T^{-1}(\cdot) = \phi_T(\cdot)$. Пусть $(x(t), r(t))$ решение уравнений Гамильтона с начальными условиями $x(0) = x, r(0) = r$. Тогда решением уравнений Гамильтона с начальными условиями $x(0) = x(T), r(0) = -r(T)$ будет пара $x(-t + T), -r(-t + T)$, что можно проверить подстановкой в уравнение. Сдвиг вдоль этого решения на время T приведет в точку $(x, -r)$, которая после смены знака второй компоненты станет равна изначальной точке (x, r) . Помимо этого, из свойств сохранения фазового объема следует $\left| \frac{\partial \phi_T}{\partial [x, r]} \right| = 1$, поскольку значение определителя описывает отношение дифференциально малых объемов после преобразования. Поэтому

$$g(\phi_T^{-1}((x', r'))) \left| \frac{\partial \phi_T^{-1}}{\partial y} \right| = g(\phi_T(x', r')), \quad (214)$$

что завершает проверку уравнения детального баланса:

$$p(x, r) \delta((x', r') - \phi_T(x, r)) = p(x', r') \delta((x, r) - \phi_T(x', r')). \quad (215)$$

Таким образом, распределение $p(x, r)$ инвариантно относительно шага гамильтонова Монте-Карло, поскольку шаг задан парой предположных распределений $q_1(x', r' \mid x, r), q_2(x', r' \mid x, r)$, для которых $p(x, r)$ инвариантно. Примечательно, что уравнение детального баланса не выполнено для $q_1(x', r' \mid x, r)$, а также для композиции $q(x', r' \mid x, r) = \int_{\mathcal{X} \times \mathbb{R}^d} q_2(x', r' \mid x'', r'') q_1(x'', r'' \mid x, r) dx dr$.

Аналогично, можно было бы показать, что после отбрасывания второй компоненты r мы получаем предположное распределение

$$q(x' \mid x) = \int_{\mathbb{R}^d \times \mathbb{R}^d} q(x', r' \mid x, r) p(r) dr dr', \quad (216)$$

для которого $p(x)$ является инвариантным распределением. В этом случае достаточно проверить уравнение детального баланса.

8.4 Гамильтонов Монте-Карло на практике

К сожалению, аналитические решения уравнений Гамильтона редко доступны в реальных задачах. Из-за этого приходится прибегать к численным методам решения дифференциальных уравнений, а численных решений не всегда можно гарантировать сохранение ключевых свойств решений уравнения Гамильтона. В результате распределение $p(x, r)$ может не быть инвариантным распределением для предположного распределения $q(x', r' \mid x, r)$ на основе численного решения, и для новых точек необходимо применять коррекцию Метрополиса-Хастингса. С другой стороны, если ошибка приближенной схемы решения уравнений Гамильтона достаточно мала, то при коррекции мы будем отбрасывать лишь малую долю новых точек.

В то время как для численного решения дифференциальных уравнений существует множество алгоритмов, схема *leap-frog integration* является наиболее подходящей для гамильтонова Монте-Карло. Шаг схемы задается уравнениями

$$\begin{cases} x_{t+\frac{1}{2}} = x_t + \frac{\epsilon}{2} \dot{x}_t = x_t + \frac{\epsilon}{2} r_t, \\ r_{t+1} = r_t + \epsilon \dot{r}_{t+\frac{1}{2}} = r_t + \epsilon \nabla \log \hat{p}(x_{t+\frac{1}{2}}), \\ x_{t+1} = x_{t+\frac{1}{2}} + \frac{\epsilon}{2} r_{t+1}. \end{cases} \quad (217)$$

Гамильтонов Монте-Карло после генерации новой точки $r_0 \sim \mathcal{N}(0, I)$ делает $T + 1$ шаг этой схемы, возвращая пару (x_T, r_T) . Генерация точек с помощью динамики Ланжевена, которой посвящена одна из следующих глав, может быть представлена как частный случай гамильтонова leap-frog integration с $T = 0$.

Задаваемое такой схемой предположное распределение симметрично, а доказательство этого факта практически не отличается от доказательства для аналитического решения. Схема обратима и симметрична по времени, поскольку из приведенных выше уравнений пару (x_t, r_t) можно выразить через (x_{t+1}, r_{t+1}) . Помимо этого, подобно динамике Гамильтона leap-frog integration сохраняет объем в фазовом пространстве. Действительно, каждое из уравнений задает преобразование фазового пространства. Каждое из них смещает одну переменную (x или r) вдоль оси на величину, зависящую лишь от другой переменной (r или x соответственно). Поскольку такие преобразования сохраняют объем произвольной области в фазовом пространстве, то и их композиция также сохраняет объем.³⁶

Из-за симметричности предположного распределения при коррекции не нужно учитывать $q(x', r' | x, r)$, оставляя лишь отношение ненормированных вероятностей

$$\min \left(1, \frac{\hat{p}(x')p(r')}{\hat{p}(x)p(r)} \right) = \min \left(1, \frac{\exp(-H(x', r'))}{\exp(-H(x, r))} \right). \quad (218)$$

К сожалению, вдоль численных решений уравнений Гамильтона величина $H(x_t, r_t)$ может не сохраняться, поэтому поправка для принятых точек необходима чтобы $\hat{p}(x, r)$ было инвариантным распределением схемы. На практике вероятность принятия близка к единице, из-за чего отказ от поправки не приводит к заметным ухудшениям схемы генерации.

8.5 Стохастический гамильтонов Монте-Карло

Перейдем к близкой постановке задачи, типичной для байесовского машинного обучения. При байесовском выводе мы можем вычислить ненормированную плотность $\hat{p}(\theta | X)$ на основе априорного распределения $p(\theta)$ и правдоподобия $p(x_i | \theta)$

$$\hat{p}(\theta | X) \propto p(X | \theta)p(\theta) = \left[\prod_{i=1}^n p(x_i | \theta) \right] p(\theta). \quad (219)$$

Для генерации точек $\theta \in \Theta$ из апостериорного распределения мы можем воспользоваться гамильтоновым Монте-Карло. В случае, когда выборки n велик, вычисление $\nabla_{\theta} \log \hat{p}(\theta | X)$ и поправки Метрополиса-Хастингса

$$A(\theta, \theta') = \min \left(1, \frac{q(\theta | \theta') \prod_{i=1}^n p(x_i | \theta')p(\theta')}{q(\theta' | \theta) \prod_{i=1}^n p(x_i | \theta)p(\theta)} \right) \quad (220)$$

может оказаться затруднительным из-за необходимости пройти всю выборку. С другой стороны, для обеспечения высокой вероятности принятия в алгоритме лучше использовать небольшой шаг предположного распределения ϵ , из-за соседние точки итерационной схемы будут неизбежно близки между собой.

Для ускорения гамильтонова Монте-Карло можно рассмотреть модифицированный алгоритм, в котором мы будем приближать $\log \hat{p}(\theta | X)$ и вероятность принятия $A(\theta, \theta')$ с помощью случайно выбранной под-выборки данных. Для независимых индексов под-выборки $i_1, \dots, i_m \sim U[1, n]$ оценка $\log p(\theta | X)$ будет иметь вид

$$\log p(\theta | X) = \log p(\theta) + \frac{n}{m} \sum_{j=1}^m \log p(x_{i_j} | \theta). \quad (221)$$

На практике замена $\nabla_{\theta} \log p(\theta | X)$ в динамике Гамильтона на несмещенную оценку существенно ухудшает предположное распределение, что делает коррекцию Метрополиса-Хастингса необходимой даже при малых ϵ . Для эффективного приближения коррекции

³⁶ Другое обоснование этого факта сводится подсчету Якобиана задаваемого итерационной схемой преобразования, который оказывается равен единице.

рассмотрим модификацию поправки Метрополиса-Хастингса. Пусть функция $g(\cdot)$ удовлетворяет функциональному уравнению $g(s) = \exp(-s)g(-s)$. Такие функции существуют, например

$$g(s) = \min(1, e^s), \quad (222)$$

или

$$g(s) = \frac{1}{1 + \exp(-s)}. \quad (223)$$

Ниже мы покажем, что такие функции могут быть использованы для поправки Метрополиса-Хастингса. Обозначим

$$\Delta(\theta, \theta') = \log \left[\frac{q(\theta | \theta') p(\theta') \prod_{i=1}^n p(x_i | \theta')}{q(\theta' | \theta) p(\theta) \prod_{i=1}^n p(x_i | \theta)} \right] \quad (224)$$

Теорема 8. (Лемма Баркера) Пусть марковская цепь принимает точку θ' , сгенерированную из предложного распределения $q(\theta' | \theta)$, с вероятностью $g(\Delta(\theta, \theta'))$, а в противном случае оставляет точку θ . Тогда распределение $p(\theta | X)$ будет её инвариантным распределением.

Доказательство. Сначала запишем переходную вероятность описанной выше цепи: точку из предложного распределения обозначим за θ'' , а результат коррекции обозначим за θ' . Совместное распределение будет иметь вид

$$q(\theta', \theta'' | \theta) = \delta(\theta - \theta') [1 - g(\Delta(\theta'', \theta))] q(\theta'' | \theta) + \delta(\theta'' - \theta') g(\Delta(\theta'', \theta)) q(\theta'' | \theta), \quad (225)$$

а после маргинализации по θ'' мы получим

$$q(\theta' | \theta) = \delta(\theta - \theta') \int_{\Theta} [1 - g(\Delta(\theta'', \theta))] q(\theta'' | \theta) d\theta'' + g(\Delta(\theta', \theta)) q(\theta' | \theta). \quad (226)$$

Проверим инвариантность $p(\theta | X)$ вычислив интеграл

$$\int_{\Theta} q(\theta' | \theta) p(\theta | X) d\theta \quad (227)$$

для каждого слагаемого $q(\theta' | \theta)$ по отдельности. Имеем

$$A = \int_{\Theta} \delta(\theta - \theta') p(\theta | X) \int_{\Theta} [1 - g(\Delta(\theta'', \theta))] q(\theta'' | \theta) d\theta'' d\theta = \quad (228)$$

$$p(\theta' | X) \int_{\Theta} [1 - g(\Delta(\theta'', \theta'))] q(\theta'' | \theta) d\theta'' \quad (229)$$

и, поскольку $g(\Delta(\theta', \theta)) = \exp(\Delta(\theta', \theta))g(-\Delta(\theta', \theta))$ и $g(-\Delta(\theta', \theta)) = g(\Delta(\theta, \theta'))$,

$$B = \int_{\Theta} \exp(\Delta(\theta', \theta)) g(\Delta(\theta, \theta')) p(\theta | X) q(\theta' | \theta) d\theta = p(\theta' | X) \int_{\Theta} g(\Delta(\theta, \theta')) q(\theta | \theta') d\theta, \quad (230)$$

откуда следует, что $A + B = p(\theta' | X)$. Лемма доказана.

Теперь применим лемму для построения стохастической коррекции Метрополиса-Хастингса. Эквивалентная формулировка коррекции предлагает сгенерировать $u \sim U[0, 1]$ и принять точку θ' если

$$g(\Delta(\theta', \theta)) \geq u. \quad (231)$$

Если мы в качестве $g(s)$ рассмотрим логистическую функцию $g(s) = \frac{1}{1 + \exp(-s)}$, применив g^{-1} к обеим частям неравенства мы получим критерий принятия

$$\Delta(\theta', \theta) + \xi \geq 0, \quad (232)$$

где $\xi \sim \text{Logistic}(0, 1)$. Построим теперь оценку для

$$\Delta(\theta', \theta) = \log \frac{p(\theta') q(\theta | \theta')}{p(\theta) q(\theta' | \theta)} + \sum_{i=1}^n \log \frac{p(x_i | \theta')}{p(x_i | \theta)} \quad (233)$$

на основе m точек

$$\bar{\Delta}(\theta', \theta) = \log \frac{p(\theta')q(\theta | \theta')}{p(\theta)q(\theta' | \theta)} + \frac{n}{m} \sum_{j=1}^m \log \frac{p(x_{i_j} | \theta')}{p(x_{i_j} | \theta)}. \quad (234)$$

Построенная оценка несмещенная, а её дисперсия σ^2 имеет порядок $O(\frac{1}{m})$. Её оценку можно получить посчитав выборочную дисперсию $\log p(x_{i_j} | \theta') - \log p(x_{i_j} | \theta)$ по под-выборке и умножив её на $\frac{n^2}{m}$.

Дополнительно предположим, что эта оценка имеет нормальное распределение $\mathcal{N}(\Delta(\theta', \theta), \sigma^2)$ (это предположение подкрепляется центральной предельной теоремой). В этом случае первое слагаемое в неравенстве 232 равно

$$\bar{\Delta}(\theta', \theta) = \Delta(\theta', \theta) + \sigma\varepsilon, \quad (235)$$

где $\varepsilon \sim \mathcal{N}(0, I)$.

Наконец, если мы подберем такую случайную величину η , что для независимой ε выполнено $\eta + \sigma\varepsilon \sim \text{Logistic}(0, 1)$, то неравенство 232 будет выполняться с той же частотой, что и неравенство

$$\bar{\Delta}(\theta', \theta) + \eta \geq 0. \quad (236)$$

Построение такой случайной величины возможно для достаточно малых σ , а подробности мы оставим за рамками лекций. Заменяв σ на оценку $\bar{\sigma}$ мы получим приближенную схему коррекции Метрополиса-Хастингса.

9 Лекция 9б. Динамика Ланжевена

9.1 Введение в динамику

Наша задача описать динамику изменения случайной величины X . Уравнение Ланжевена предлагает описывать её динамику как сумму случайного $G(t)$ и детерминированного $h(X, t)$ полей:

$$\dot{X} = h(X, t) + g(X, t)G(t). \quad (237)$$

Для того чтобы продвинуться вперед, мы введем несколько предположений и сузим класс рассматриваемых уравнений Ланжевена. Пусть $g(X, t) = \sigma$ и $G(t)$ это гауссовская случайная величина с нулевым средним и $\delta(t - t')$ ковариационной функцией. Предполагая детерминированное поле $G(t)$ потенциальным, перепишем уравнение 237 в форме стохастического дифференциального уравнения:

$$dX(t) = -\nabla U(X(t))dt + \sigma dB(t). \quad (238)$$

Для того чтобы 238 соответствовало 237 в описанных предпосылках, мы должны определить $B(t)$ следующим образом:

$$B(t + \tau) - B(t) = \int_t^{t+\tau} G(t')dt'. \quad (239)$$

С чисто формальной точки зрения, решение дифференциального уравнения 238 можно записать в интегральном виде:

$$X(t + \tau) = X - \int_t^{t+\tau} \nabla U(X(t'))dt' + \sigma \int_t^{t+\tau} dB(t'). \quad (240)$$

Нам осталось определить, что же такое $\int_t^{t+\tau} dB(t)$. Следуя Ито и 239 определим интегрирование относительно $dB(t)$ так:

$$\int_t^{t+\tau} f(B(t), t')dB(t') = \lim_{\Delta \rightarrow 0} \sum_{n=0}^{N-1} f(B(t_n), t_n)(G(t_{n+1}) - G(t_n)), \Delta = \max(t_{n+1} - t_n). \quad (241)$$

Особое внимание здесь стоит обратить на индексы времени t для вычисления значений интегрируемой функции f . Теперь мы можем перейти к дискретной аппроксимации, также используя явный метод:

$$X_{t+1} - X_t = -\Delta t \nabla U(X_t) + \sigma \sqrt{\Delta t} \mathcal{N}(0, 1) \quad (242)$$

Можно увидеть аналогию с градиентным спуском, где присутствует какой-то шум на каждом шаге градиента. Для такой дискретизации с шагом Δt вокруг точки x' в момент времени t мы можем записать плотность распределения:

$$(243)$$

$$p(x|x') = \mathcal{N}(x' - \nabla U(x')\Delta t, \sigma^2 \Delta t). \quad (244)$$

9.2 Уравнение Фоккера-Планка

Мы выразили распределение следующей точки при одном шаге динамики Ланжевена ???. Давайте выразим теперь плотность x в момент времени t (т.е. относительно t плотность

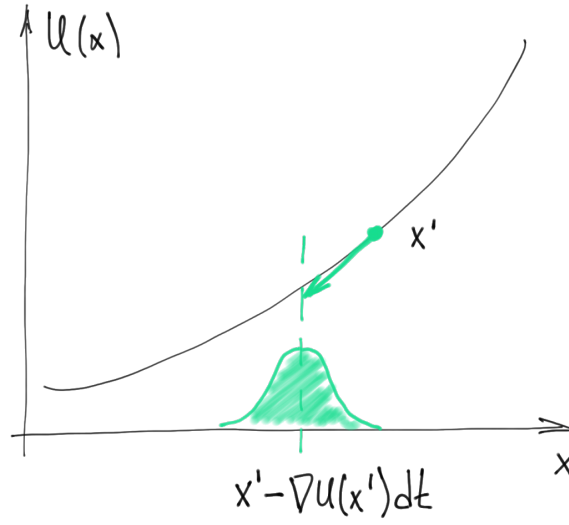


Рис. 16: Распределение следующей точки

НЕ суммируется в единицу, только по x в каждый момент времени).

$$p(x, t) = \int dx' p(x, t | x', t - dt) p(x', t - dt). \quad (245)$$

$$p(x, t | x', t - dt) = \frac{1}{(2\pi\sigma^2 dt)^{n/2}} \exp\left(-\frac{\overbrace{(x' - x - \nabla U(x') dt)^2}^y}{2\sigma^2 dt}\right), \quad (246)$$

$$p(x, t) = \int dy \left| \frac{\partial x'}{\partial y} \right| \mathcal{N}(y | 0, \sigma^2 dt I) p(x'(y) | t - dt). \quad (247)$$

Разложим $\nabla U(x')$ в ряд Тейлора:

$$y = x' - x - \left(\nabla U(x) + \frac{\partial \nabla U(x)}{\partial x} (x' - x) dt + o(x' - x) \right) dt, \quad (248)$$

$$\left(I - \frac{\partial \nabla U(x)}{\partial x} dt \right) x' = y + x + \nabla U(x) dt - \frac{\partial U(x)}{\partial x} x dt + o(dt), \quad (249)$$

$$x' = \left(I - \frac{\partial \nabla U(x)}{\partial x} dt \right)^{-1} \left(y + x + \nabla U(x) dt - \frac{\partial \nabla U(x)}{\partial x} x dt + o(dt) \right). \quad (250)$$

Воспользуемся формулой $(1 - A)^{-1} = 1 + A + A^2 + A^3 + \dots$:

$$x' = \left(I + \frac{\partial \nabla U(x)}{\partial x} dt + o(dt) \right) \left(y + x + \nabla U(x) dt - \frac{\partial \nabla U(x)}{\partial x} x dt + o(dt) \right) = \quad (251)$$

$$= y + x + \nabla U(x) dt - \frac{\partial \nabla U(x)}{\partial x} x dt + \frac{\partial \nabla U(x)}{\partial x} y dt + \frac{\partial \nabla U(x)}{\partial x} x dt + o(dt) = \quad (252)$$

$$= x + y + \nabla U(x) dt + \frac{\partial \nabla U(x)}{\partial x} y dt + o(dt). \quad (253)$$

Из 243 получаем

$$y dt = (x' - x - \nabla U(x) dt) dt = dt \sqrt{dt} \mathcal{N}(0, \sigma^2) = o(dt). \quad (254)$$

Получили формулу для x' :

$$x' = x + y + \nabla U(x) dt + o(dt). \quad (255)$$

В качестве упражнения предлагается показать, что

$$\left| \frac{\partial x'}{\partial y} \right| = 1 + \operatorname{div} \nabla U(x) dt + o(dt), \quad (256)$$

где дивергенция $\operatorname{div} \mathbf{f}(x) = \frac{\partial f_1}{\partial x_1} + \frac{\partial f_2}{\partial x_2} + \dots + \frac{\partial f_n}{\partial x_n}$.

Подставив в 247, получаем

$$p(x, t) = (1 + \operatorname{div} \nabla U(x) dt) \mathbb{E}_y [p(y + x + \nabla U(x) dt, t - dt)], \quad (257)$$

где $y \sim \mathcal{N}(0, \sigma^2 dt I)$.

Разложим мат ожидание в ряд Тейлора.

$$\mathbb{E}_y [p(y + x + \nabla U(x) dt, t - dt)] = \mathbb{E}_y \left[p(x, t) + \nabla_x p(x, t) (y + \nabla U(x) dt) + \right. \quad (258)$$

$$\left. + \frac{\partial}{\partial t} p(x, t) (-dt) + \frac{1}{2} (y + \nabla U(x) dt)^T \frac{\partial^2 p(x, t)}{\partial x^2} (y + \nabla U(x) dt) + o(dt) \right]. \quad (259)$$

Выпишем мат ожидание каждого слагаемого.

$$\mathbb{E}_y p(x, t) = p(x, t) \quad (260)$$

$$\mathbb{E}_y [\nabla_x p(x, t) (y + \nabla U(x) dt)] = \nabla_x p(x, t)^T \mathbb{E}_y y + dt \nabla_x p(x, t)^T \nabla U(x) = dt \nabla_x p(x, t)^T \nabla U(x) \quad (261)$$

$$\mathbb{E}_y \left[dt \frac{\partial}{\partial t} p(x, t) \right] = dt \frac{\partial}{\partial t} p(x, t) \quad (262)$$

$$\mathbb{E}_y \left[(y + \nabla U(x) dt)^T \frac{\partial^2 p(x, t)}{\partial x^2} (y + \nabla U(x) dt) \right] = \mathbb{E}_y \left[y^T \frac{\partial^2 p(x, t)}{\partial x^2} y + 2 dt \nabla U(x)^T \frac{\partial^2 p(x, t)}{\partial x^2} y + \right. \quad (263)$$

$$\left. + dt^2 \nabla U(x)^T \frac{\partial^2 p(x, t)}{\partial x^2} \nabla U(x) \right] = \mathbb{E}_y \left[\sum_{i,j} \left(\frac{\partial^2 p(x, t)}{\partial x^2} \right)_{ij} y_i y_j \right] + 2 dt \nabla U(x)^T \frac{\partial^2 p(x, t)}{\partial x^2} \mathbb{E}_y y + \quad (264)$$

$$+ o(dt) = \underbrace{\sum_{i,j} \left(\frac{\partial^2 p(x, t)}{\partial x^2} \right)_{ij} \mathbb{E}_y [y_i y_j]}_{=\Delta p(x, t)} + \sum_{i \neq j} \left(\frac{\partial^2 p(x, t)}{\partial x^2} \right)_{ij} \underbrace{\mathbb{E}_y [y_i y_j]}_{=0} + o(dt). \quad (265)$$

Подставим найденные мат ожидания в $p(x, t)$.

$$p(x, t) = (1 + \operatorname{div} \nabla U(x) dt) (p(x, t) + dt \nabla_x p(x, t)^T \nabla U(x) - dt \frac{\partial}{\partial t} p(x, t) + \quad (266)$$

$$+ \frac{1}{2} \sigma^2 dt \Delta p(x, t) + p(x, t) \operatorname{div} \nabla U(x) dt + o(dt)). \quad (267)$$

Сократив $p(x, t)$, получаем уравнение Фоккера-Планка:

$$\frac{\partial}{\partial t} p(x, t) = \nabla_x p(x, t)^T \nabla U(x) + p(x, t) \operatorname{div} \nabla U(x) + \frac{1}{2} \sigma^2 \Delta p(x, t) + \underbrace{\frac{o(dt)}{dt}}_{=0}. \quad (268)$$

Динамика Ланжевена сходится к некоторому стационарному распределению. Давайте поймем, что это за распределение. Пусть $p(x, t)$ стационарно, т.е. не зависит от времени $p(x, t) = \hat{p}(x)$. Давайте возьмем распределение Гиббса, подставим его в уравнение Фоккера-Планка и убедимся, что оно является стационарным.

$$p_G(x) = \frac{1}{Z} \exp \left(-\frac{U(x)}{T} \right), \quad Z = \int dx \exp \left(-\frac{U(x)}{T} \right) \quad (269)$$

В качестве упражнения предлагается подставить в уравнение Фоккера-Планка это распределение и найти температуру T :

$$0 = \nabla_x \hat{p}(x)^T \nabla U(x) + \hat{p}(x) \operatorname{div} \nabla U(x) + \frac{1}{2} \sigma^2 \Delta \hat{p}(x), \quad T = \frac{\sigma^2}{2} \quad (270)$$

9.3 Сэмплирование

Один из вариантов использования динамики Ланжевена – это получение семплов из стационарного распределения Ланжевена. Пусть нас интересует стационарное распределение некоторого объекта, например, частицы или веса нейросети. Предположим, что эволюция этого объекта происходит с помощью динамики Ланжевена и она имеет следующее стационарное распределение:

$$p_G(x) = \frac{1}{Z} \exp\left(-\frac{2U(x)}{\sigma^2}\right), Z = \int dx \exp\left(-\frac{2U(x)}{\sigma^2}\right). \quad (271)$$

Осталось выбрать $U(x)$, чтобы сэмплировать из некоторого распределения $p(x)$:

$$U(x) = -\log p(x), \sigma = \sqrt{2}, \quad (272)$$

$$p_G(x) = \frac{1}{Z} \exp\left(-\frac{-2\log p(x)}{2}\right) = p(x), Z = \int dx \exp(\log p(x)) = 1. \quad (273)$$

Мы показали как нужно выбрать детерминированное поле $U(x)$, чтобы стационарное распределение динамики Ланжевена совпало с $p(x)$. Часто необходимо сэмплировать из распределения $\hat{p}(x)$, которое не отнормировано, т.е. $p(x) = \frac{\hat{p}(x)}{Z}$. Это не проблема, т.к. в динамике Ланжевена нужен градиент $\nabla U(x)$, который не зависит от нормировочной константы:

$$\nabla U(x) = -\nabla \log(p(x)) = -\nabla \log \hat{p}(x) - \underbrace{\nabla \log Z'}_{=0}. \quad (274)$$

К сожалению, люди не умеют точно симулировать стохастические дифференциальные уравнения, поэтому воспользуемся дискретной аппроксимацией [242](#), подставив найденные $U(x)$ и σ :

$$X_{t+1} = X_t + dt \nabla \log p(X_t) + \mathcal{N}(0, 2dt). \quad (275)$$

Приведем более популярный вид предыдущего выражения:

$$X_{t+1} = X_t + \frac{\varepsilon}{2} \nabla \log p(X_t) + \mathcal{N}(0, \varepsilon) \quad (276)$$

$$X_t \sim p(x), \forall t > t_\infty. \quad (277)$$

В итоге, наш алгоритм двигается по градиенту и немного шумит. Остается вопрос почему дискретная аппроксимация сойдется к распределению $p(x)$, который будет разобран в следующем разделе.

9.4 Применение к байесовскому выводу

Применим динамику Ланжевена к байесовскому формализму. Есть выборка D_{train} , хотим получить новые предсказания. Обычно есть модель с латентными переменными θ и мы можем получить предсказание y , если знаем переменные θ . Используя оценку Монте-Карло и динамику Ланжевена для сэмплирования θ из $p(\theta|D_{train})$, получим несмещенную оценку для распределения $p(y|D_{train})$. Запишем более формально.

$$p(y|D_{train}) = \mathbb{E}_{p(\theta|D_{train})} p(y|\theta) \simeq \frac{1}{K} \sum_{i=1}^K p(y|\theta_i), \theta_i \sim p(\theta|D_{train}) \quad (278)$$

$$d\theta(t) = \nabla \log p(\theta(t)|D_{train}) dt + \sqrt{2} dB_t \quad (279)$$

$$\theta_{t+1} = \theta_t + \frac{\varepsilon}{2} \nabla \log p(\theta_t|D_{train}) + \mathcal{N}(0, \varepsilon) \quad (280)$$

$$\theta_{t+1} = \theta_t + \frac{\varepsilon}{2} \nabla_\theta \left(\sum_{i=1}^N \log p(\theta_t|(x_i, y_i)) + \log p(\theta_t) \right) + \mathcal{N}(0, \varepsilon) \quad (281)$$

$$\theta_{t+1} = \theta_t + \frac{\varepsilon}{2} \nabla_\theta \left(\frac{N}{B} \sum_{k=1}^B \log p(\theta_t|(x_{i_k}, y_{i_k})) + \log p(\theta_t) \right) + \mathcal{N}(0, \varepsilon) \quad (282)$$

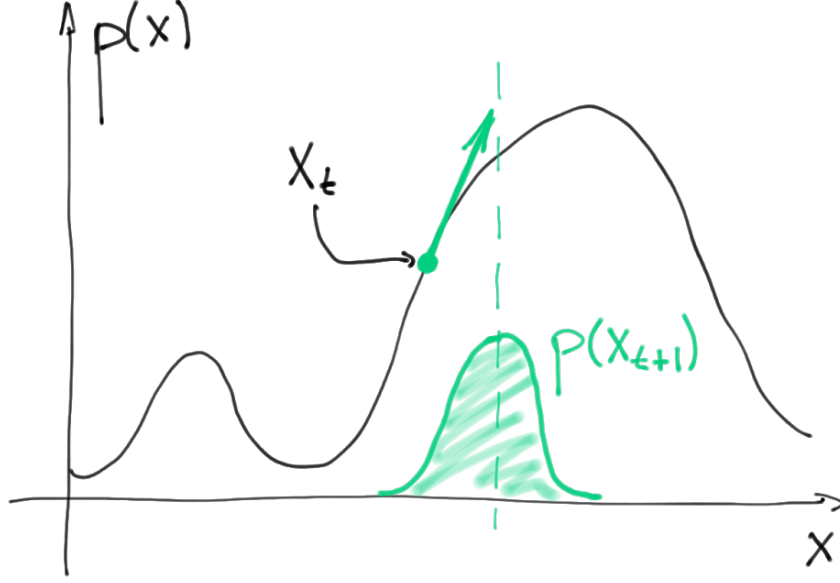


Рис. 17: Распределение следующей точки при дискретной аппроксимации

В конце, чтобы не ждать плотность по всей выборке, мы считаем по минибатчам. Осталось два вопроса: почему дискретная аппроксимация сойдется к стационарному распределению и почему можно вычислять по минибатчам. Borkar, Mitter в 1999 дали ответы на эти вопросы.

Теорема (1) (Borkar, Mitter, 1999). Рассмотрим стохастическое дифференциальное уравнение $dX(t) = h(X(t))dt + \sigma dBt$, где $X(t)$ сходится к некоторому стационарному распределению $X(t) \sim p_\sigma(x), t > t_\infty$. Возьмем дискретную аппроксимацию, которая тоже сходится к некоторому стационарному распределению:

$$X_{k+1} = X_k + \varepsilon(h(X_k) + M_k) + \mathcal{N}(0, \sigma^2 \varepsilon), X_k \sim \hat{p}_\sigma(x), k > k_\infty \quad (283)$$

$$\mathbb{E}M_k = 0, \forall k \quad (284)$$

Тогда

$$\forall \delta > 0, \exists \varepsilon : \text{KL}(p_\sigma(x) || \hat{p}_\sigma(x)) < \delta \quad (285)$$

Набросок доказательства. Возьмем линейную интерполяцию дискретной аппроксимации.

$$\tilde{X}(t) = X(0) + \int_0^t \left(h(\tilde{X}(\lfloor s \rfloor_\varepsilon) + \xi_s) \right) ds + \sigma \tilde{B}(t) \quad (286)$$

$$\lfloor s \rfloor_\varepsilon = k\varepsilon, \text{ если } s \in [k\varepsilon, (k+1)\varepsilon) \quad (287)$$

$$\xi_s = M_k, \text{ если } s \in [k\varepsilon, (k+1)\varepsilon) \quad (288)$$

$$\tilde{B}((k+1)\varepsilon) - \tilde{B}(k\varepsilon) = \mathcal{N}(0, \varepsilon) \quad (289)$$

Лемма (2).

$$\forall t \mathbb{E} \left[\|X(t) - \tilde{X}(t)\|^2 \right] \rightarrow 0 \text{ при } \varepsilon \rightarrow 0 \quad (290)$$

Прокомментируем результат леммы. Продифференцируем наше стохастическое дифференциальное уравнение и сравним с нашей аппроксимацией:

$$X(t) = X(0) + \int_0^t h(X(s))ds + \sigma B(t) \quad (291)$$

$$\tilde{X}(t) = X(0) + \int_0^t \left(h(\tilde{X}(\lfloor s \rfloor_\varepsilon) + \xi_s) \right) ds + \sigma \tilde{B}(t) \quad (292)$$

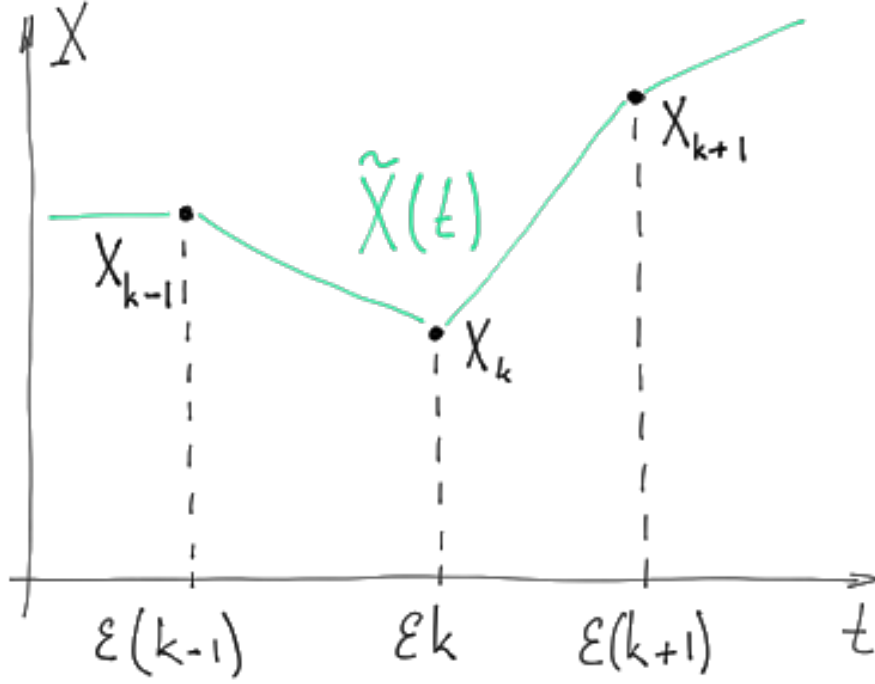


Рис. 18: Визуализация линейной интерполяции

Лемма говорит, что $X(t) = \tilde{X}(t)$, когда $\varepsilon \rightarrow 0$. Кажется, что можно получить бесплатное ускорение, считая ξ_s не по всей, а по минибатчу. Но на самом деле также надо уменьшить размер шага ε и соответственно чаще вычислять шум, чтобы сдвинуться на Δt . Графики 19 иллюстрируют это при $h = 0$, слева шум считается по половине выборки, справа по $\frac{1}{N}$.

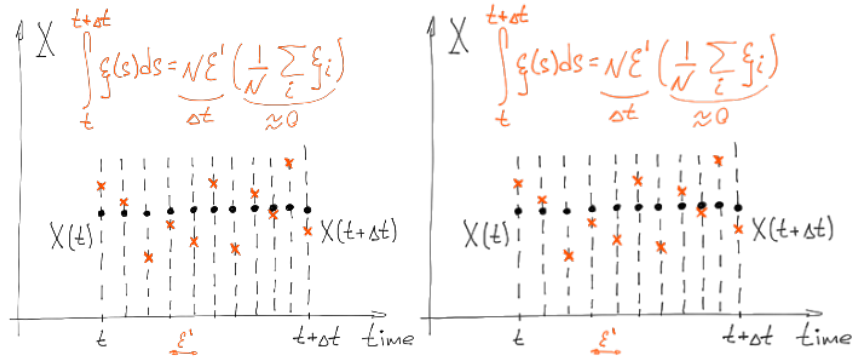


Рис. 19: Зависимость размера шага от размера минибатча

Лемма (3).

$$X(t) \sim p(x, t) \quad (293)$$

$$\text{KL}(p_\sigma(x) || p(x, t)) \text{ строго уменьшается с увеличением } t \quad (294)$$

Из лемм 2 и 3 следует доказываемая теорема. ■

9.5 Применение к схеме Метрополиса-Хастингса

Возьмем схему Метрополиса-Хастингса.

$$V_{t+\frac{1}{2}} = V_t + \frac{\varepsilon}{2} \nabla \log p(x_t) \quad (295)$$

$$X_{t+1} = X_t + \varepsilon V_{t+\frac{1}{2}} = X_t + \varepsilon V_t + \frac{\varepsilon^2}{2} \nabla \log p(x_t), V_t \sim \mathcal{N}(0, 1) \quad (296)$$

Последнее выражение соответствует динамике Ланжевена.

Следующее распределение можно взять в качестве функции $q(X_{t+1}|X_t)$:

$$X_{t+1} \sim \mathcal{N}(X_t + \frac{\varepsilon^2}{2} \nabla \log p(x_t), \varepsilon^2) \quad (297)$$

Это решает проблему выбора шага.

9.6 Глобальная оптимизация

У стационарного распределения Гиббса, которое мы выписывали ранее, в глобальном минимуме будет самая большая плотность. Опишем процедуру "temperature annealing". Возьмем распределение Гиббса и устремим температуру T к нулю.

$$p_T(x) = \frac{1}{Z} \exp\left(-\frac{U(x)}{T}\right), Z = \int dx \exp\left(-\frac{U(x)}{T}\right), T = \frac{\sigma^2}{2}, \quad (298)$$

$$dX(t) = -\nabla U(X(t))dt + \sigma dB_t = -\nabla U(X(t))dt + \sqrt{2T}dB_t, \quad (299)$$

$$p_T(x) = \exp\left(-\frac{U(x)}{T}\right) \rightarrow \pi(x) \text{ при } T \rightarrow 0, \quad (300)$$

где $\pi(x)$ - дельта-функция в глобальном минимуме. То есть симуляция одной частички рано или поздно приведет нас в глобальный минимум.

10 Лекция 11. Непараметрические байесовские методы: процессы Дирихле

10.1 Описание байесовских непараметрических моделей

Использование Гауссовских процессов в задаче восстановления регрессии и в задаче классификации, приведённых в предыдущей лекции, является примером непараметрической байесовской модели (см. таблицу 4).

Таблица 4: Примеры моделей

	Параметрические	Непараметрические
Статистические	Доверительные интервалы	Знаковые тесты, Bootstrap оценки, U-тесты, критерий Колмогорова-Смирнова, GLM
Байесовские	RVM, ЕМ, вариационные оценки	Основаны на случайных процессах (гауссовские процессы в регрессии и классификации, процессы Дирихле в LDA)

И параметрические, и непараметрические методы сводятся к оптимизации определённых параметров. Но в параметрических методах модель фиксируется заранее, т.е. изначально известна её сложность. И в рамках этой фиксированной модели параметры вписываются в наблюдаемые данные.

В непараметрических методах количество «параметров» растёт с увеличением объёма выборки, и сложность модели предварительно не фиксируется. 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.

В данной лекции будет подробно рассмотрено определение процессов Дирихле, их свойства, и описаны различные способы их использования в непараметрических методах.

10.2 Распределение Дирихле, его свойства

Определение 12. Назовём $\Delta_k = \{\theta \in \mathbb{R}^k \mid \sum_{i=1}^k \theta_i = 1, \theta_j \geq 0, j = \overline{1, k}\}$ $(k-1)$ -мерным симплексом³⁷. Распределением Дирихле называется непрерывное распределение вероятностей, определённое на $(k-1)$ -мерном симплексе, плотность вероятности которого задаётся следующей формулой:

$$Dir(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k \theta_i^{\alpha_i-1}, \quad \theta \in \Delta_k, \quad \alpha = (\alpha_i)_{i=1}^k, \quad \alpha_i > 0 \quad (301)$$

$$\text{Для удобства иногда пишут } Dir(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^k \theta_i^{\alpha_i-1} \quad (302)$$

Вектор α — параметры распределения, $\Gamma(\cdot)$ — гамма-функция, $B(\cdot)$ — многомерная бета-функция. Т.к. плотность распределения определена только на симплексе, то, фактически, плотность зависит только от $(k-1)$ -ой переменной. Это свойство будет разобрано чуть позднее.

Симметричным называется распределение Дирихле с константным вектором параметров. Примеры вида распределения Дирихле в этом случае изображены на рисунке 20.

Стоит отметить связь распределения Дирихле с другими распределениями.

³⁷ $(k-1)$ -мерный симплекс определён в k -мерном пространстве.

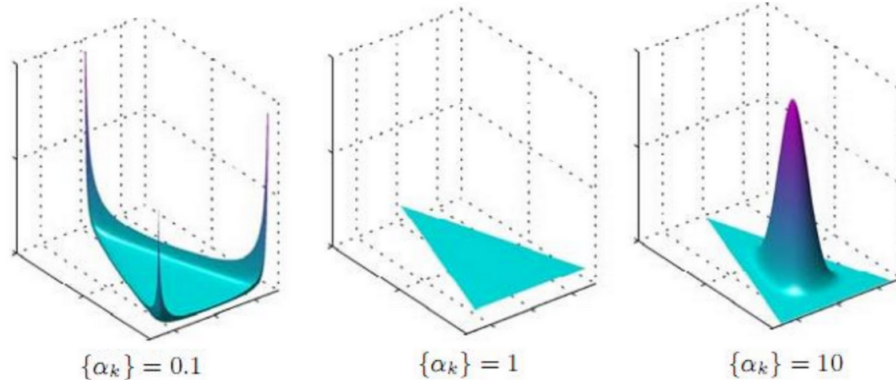


Рис. 20: Виды плотности симметричного распределения Дирихле для $k = 3$

1. В случае $k = 2$, если $x \sim \text{Beta}(x|\alpha, \beta)$, то

$$(x, 1-x) \sim \text{Dir}((x, 1-x) | (\alpha, \beta)), \quad \alpha > 0, \quad \beta > 0$$

Это следует из вида бета-распределения:

$$\text{Beta}(x|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad x \in [0, 1] \quad (303)$$

2. Распределение Дирихле является сопряжённым с мультиномиальным:³⁸

$$\text{Mult}(\mathbf{x}|k, n, \boldsymbol{\theta}) = \frac{n!}{x_1! \dots x_k!} \prod_{i=1}^k \theta_i^{x_i}, \quad x_i \in \{0, \dots, n\}, i = \overline{1, k}; \quad \sum_{i=1}^k x_i = n. \quad (304)$$

Где $k \in \mathbb{N}$ — количество возможных исходов каждого из испытаний, $n \in \mathbb{N}$ — общее число испытаний, $\boldsymbol{\theta}$ — вектор вероятностей выпадения каждого из исходов в каждом испытании.

Итак, если $\mathbf{x} \sim \text{Mult}(\mathbf{x}|k, n, \boldsymbol{\theta})$, на $\boldsymbol{\theta}$ задано априорное распределение Дирихле ($\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})$), то апостериорное распределение на $\boldsymbol{\theta}$ также будет распределением Дирихле:

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{1}{Z} p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{1}{Z} \left(\frac{n!}{x_1! \dots x_k!} \prod_{i=1}^k \theta_i^{x_i} \right) \left(\frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k \theta_i^{\alpha_i-1} \right) = \quad (305)$$

$$= \frac{1}{Z} \prod_{i=1}^k \theta_i^{\alpha_i + x_i - 1} = \text{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha} + \mathbf{x}) \quad (306)$$

■

3. Свойство накопления (aggregation property)

Суммы компонент вектора, распределённого по Дирихле, также распределены по Дирихле. Пусть $\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})$ и пусть множество $\{1, \dots, k\}$ разбито на r непересекающихся подмножеств A_1, \dots, A_r . Тогда

$$\left(\sum_{i \in A_1} \theta_i, \dots, \sum_{i \in A_r} \theta_i \right) \sim \text{Dir} \left(\hat{\boldsymbol{\theta}} \mid \sum_{i \in A_1} \alpha_i, \dots, \sum_{i \in A_r} \alpha_i \right), \quad \hat{\boldsymbol{\theta}} \in \Delta_r \quad (307)$$

Для доказательства данного свойства понадобится ещё одно свойство, используемое в простейшем случае для генерации случайных величин из распределения Дирихле.

³⁸Биномиальное распределение — это случай $k = 2$ для мультиномиального распределения. Таким образом, этот факт является обобщением свойства сопряжённости бета-распределения и биномиального.

4. Генерация распределения Дирихле с использованием гамма-распределений

Напомним, что Гамма-распределение имеет следующий вид:

$$\Gamma(\lambda|a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}, \quad \lambda > 0, \quad a, b > 0 \quad (308)$$

В схеме сэмплирования используется следующее свойство Гамма-распределения. Если $z_i \sim \Gamma(z_i | a_i, b)$ (т.е. порядковый параметр³⁹ фиксирован) — независимы для $i = \overline{1, n}$. Тогда $\sum_{i=1}^n z_i \sim \Gamma(\sum_{i=1}^n z_i | \sum_{i=1}^n a_i, b)$.

Процедура генерации выборки с использованием гамма-распределений описывается следующим образом:

- (a) Генерируем $z_i \sim \Gamma(z_i | \alpha_i, 1)$, $i = \overline{1, k}$
- (b) Тогда $\theta_i = \frac{z_i}{\sum_{j=1}^k z_j}$, $i = \overline{1, k}$ и $\theta \sim Dir(\theta | \alpha)$

Докажем это свойство, используя формулу замены переменных (чтобы доказать, что плотность θ — это плотность распределения Дирихле). Итак, в результате процедуры генерации из исходного набора переменных $\{z_i\}_{i=1}^k$ получен новый набор $\theta_i = \frac{z_i}{Z}$, $i = \overline{1, k-1}$, $Z = \sum_{j=1}^k z_j$.⁴⁰ Обозначим отображение из $(Z, \theta_{\setminus k})$ в \mathbf{z} за T :

$$(z_1, \dots, z_k) = T(Z, \theta_1, \dots, \theta_{k-1}) = \left(Z\theta_1, \dots, Z\theta_{k-1}, Z \left(1 - \sum_{j=1}^{k-1} \theta_j \right) \right) \quad (309)$$

Формула замены переменных в плотности новых переменных имеет следующий вид:

$$p(Z, \theta_1, \dots, \theta_{k-1}) = p(z_1, \dots, z_k) \circ T \times \det(J(T)) \quad (310)$$

Якобиан перехода вычисляется по следующей формуле:

$$\frac{\partial \mathbf{z}}{\partial \theta} = J(T) = \begin{pmatrix} \theta_1 & Z & 0 & \dots & 0 \\ \theta_2 & 0 & Z & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{k-1} & 0 & 0 & \dots & Z \\ 1 - \sum_{j=1}^{k-1} \theta_j & -Z & -Z & \dots & -Z \end{pmatrix}; \quad \det(J(T)) = Z^{k-1} \quad (311)$$

Запишем теперь итоговую формулу совместной плотности вероятности:

$$p(Z, \theta_1, \dots, \theta_{k-1}) = \left(\prod_{i=1}^{k-1} (Z\theta_i)^{\alpha_i-1} \frac{e^{-Z\theta_i}}{\Gamma(\alpha_i)} \right) \left(Z \left(1 - \sum_{i=1}^{k-1} \theta_i \right) \right)^{\alpha_k-1} \frac{e^{-Z \left(1 - \sum_{i=1}^{k-1} \theta_i \right)}}{\Gamma(\alpha_k)} \cdot Z^{k-1} \quad (312)$$

Проинтегрируем по Z для получения маргинального распределения на $\{\theta_i\}_{i=1}^{k-1}$:

$$p(\theta_1, \dots, \theta_{k-1}) = \int_0^\infty p(Z, \theta_1, \dots, \theta_{k-1}) dZ = \frac{\left(\prod_{i=1}^{k-1} \theta_i^{\alpha_i-1} \right) \left(1 - \sum_{i=1}^{k-1} \theta_i \right)^{\alpha_k-1}}{\prod_{i=1}^k \Gamma(\alpha_i)} \times \quad (313)$$

$$\times \int_0^\infty Z^{(\sum_{i=1}^k \alpha_i)-1} e^{-Z} dZ = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \left(\prod_{i=1}^{k-1} \theta_i^{\alpha_i-1} \right) \left(1 - \sum_{i=1}^{k-1} \theta_i \right)^{\alpha_k-1} \quad (314)$$

³⁹В английской литературе данный параметр в такой форме Гамма-распределения называют the rate parameter.

⁴⁰Так как последняя компонента θ_k однозначно получается из всех прочих компонент путём вычитания их из единицы, её мы для удобства опустим.

Что является плотностью распределения Дирихле. ■

Данный способ генерации подходит для малого числа k , поскольку для генерации i -ой компоненты необходимо знать все k чисел. Для случая генерации бесконечного числа объектов (как в процессе Дирихле, об этом далее) этот способ не работает.

С использованием связи распределения Дирихле с гамма-распределениями теперь можно доказать свойство накопления. Пусть $(\theta_1, \dots, \theta_k) \sim Dir(\alpha)$. Т.к. известно, что

$$\theta_i = \frac{z_i}{\sum_{j=1}^k z_j}, \quad i = \overline{1, k}, \quad z_i \sim \Gamma(\alpha_i, 1) \quad (315)$$

То

$$\left(\sum_{i \in A_1} \theta_i, \dots, \sum_{i \in A_r} \theta_i \right) = \frac{1}{\sum_{i=1}^k z_i} \left(\sum_{i \in A_1} z_i, \dots, \sum_{i \in A_r} z_i \right) \sim \quad (316)$$

$$\sim \frac{1}{\sum_{i=1}^k \Gamma(\alpha_i, 1)} \left(\Gamma\left(\sum_{i \in A_1} \alpha_i, 1\right), \dots, \Gamma\left(\sum_{i \in A_r} \alpha_i, 1\right) \right) \stackrel{d}{=} Dir\left(\sum_{i \in A_1} \alpha_i, \dots, \sum_{i \in A_r} \alpha_i\right) \quad (317)$$

■

5. **Свойство нейтральности.** Каждая компонента вектора θ влияет на распределение остальных только через нормировку. Т.е. случайная величина θ_i и случайный вектор $(\frac{1}{1-\theta_i}\theta_{\setminus i})$ являются независимыми.⁴¹

6. **Маргинальные распределения.** Следствием свойства накопления является то, что маргинальное распределение одной компоненты θ_i вектора θ , распределённого по Дирихле, является Бета-распределением.

Если $\theta \sim Dir(\theta \mid \alpha)$, то $\theta_i \sim Beta(\theta_i \mid \alpha_i, \sum_{j \neq i} \alpha_j)$.

7. Из предыдущих свойств вытекает следующий факт:

$$p(\theta_2, \dots, \theta_k \mid \theta_1) = \frac{p(\theta_1, \dots, \theta_k)}{p(\theta_1)} = \left[p(\theta_1) = Dir(\theta_1, 1 - \theta_1 \mid \alpha_1, \sum_{j=2}^k \alpha_j) \right] = \quad (318)$$

$$= \left[Beta(\theta_1 \mid \alpha_1, \sum_{j=2}^k \alpha_j) = \frac{1}{Z} \theta_1^{\alpha_1-1} (1 - \theta_1)^{\sum_{j=2}^k \alpha_j - 1} \right] = \quad (319)$$

$$= \frac{1}{Z} \prod_{j=2}^k \left(\frac{\theta_j}{1 - \theta_1} \right)^{\alpha_j - 1} = Dir\left(\frac{\theta_2}{1 - \theta_1}, \dots, \frac{\theta_k}{1 - \theta_1} \mid \alpha_2, \dots, \alpha_k\right) \quad (320)$$

Таким образом:

$$(\theta_{\setminus i} \mid \theta_i) \sim (1 - \theta_i) Dir(\alpha_{\setminus i}) \quad (321)$$

8. **"Ломка палки" (Stick-breaking).** Прежде чем перейти к процессам Дирихле, необходимо решить проблему сэмплирования бесконечного числа элементов из распределения Дирихле. Благодаря свойствам распределения Дирихле, процесс генерации можно осуществить последовательно, генерируя сначала из маргинального распределения $p(\theta_1)$, затем из условного $p(\theta_2 \mid \theta_1)$, из $p(\theta_3 \mid \theta_1, \theta_2)$ и т.д.⁴²

Маргинальное распределение в этом случае $p(\theta_1) = Beta(\theta_1 \mid \alpha_1, \sum_{i=2}^k \alpha_i)$. Генерируем θ_1 из этого распределения с помощью схемы с гамма-распределениями, т.к. Бета-распределение — частный случай распределения Дирихле. Тогда по формуле, полученной выше, $((\theta_2, \theta_3, \dots, \theta_k) \mid \theta_1) \sim (1 - \theta_1) Dir(\alpha_2, \alpha_3, \dots, \alpha_k)$.

⁴¹Подробное доказательство приведено в В. А. Frigiyk, A. Kapila, and M. R. Gupta, "Introduction to the Dirichlet distribution and related processes", UWEE, Tech. Rep. UWEE-2010-0006, 2010.

⁴²Такой подход носит названия "ломки палки" из-за аналогии с отламыванием кусочков длины θ_i от палки длины 1.

Для $2 \leq j \leq k-2$ продолжаем рекурсивно. Пусть уже отломлено $j-1$ частей; используя маргинальное распределение, $(\theta_j \mid (\theta_1, \theta_2, \dots, \theta_{j-1})) \sim \left[\prod_{i=1}^{j-1} (1 - \theta_i) \right] \text{Beta}(\alpha_j, \sum_{i=j+1}^k \alpha_i)$.

Тогда остаток распределён как

$$((\theta_{j+1}, \theta_{j+2}, \dots, \theta_k) \mid (\theta_1, \theta_2, \dots, \theta_j)) \sim \left[\prod_{i=1}^j (1 - \theta_i) \right] \text{Dir}(\alpha_{j+1}, \alpha_{j+2}, \dots, \alpha_k) \quad (322)$$

На последних шагах $j = k-1, k$ происходит следующее: мы получаем результат с шага $k-2$, аналогично ранним шагам генерируем элемент θ_{k-1} из Бета-распределения, а остаток (то, что останется после вычитания из единицы — общей длины палки) присваиваем θ_k .

Таким образом, итоговая схема записывается в следующем виде:

$$v_i \sim \text{Beta}(\alpha_i, \sum_{j=i+1}^k \alpha_j), \quad i = \overline{1, k-1}, \quad (323)$$

$$v_k = 1. \quad (324)$$

$$\theta_i = v_i \prod_{j=1}^{i-1} (1 - v_j), \quad i = \overline{1, k} \quad (325)$$

Нетрудно проверить, что выполнено равенство $\sum_{i=1}^k \theta_i = 1$.

10.3 Процессы Дирихле и их применение

10.3.1 Определение процесса Дирихле, сравнительные характеристики

В предыдущем разделе не было необходимости введения строгого определения случайного процесса. Для описания свойств процесса Дирихле оно потребуется.

Определение 13. *Случайным процессом* называется функция двух аргументов

$$\xi(\omega, x) : \Omega \times \mathcal{X} \rightarrow \mathbb{R},$$

где ω — элементарный исход, x — индексирующий элемент. Множество $\xi(\cdot, x) = \{\xi(\omega, x)\}_{\omega \in \Omega}$ — различные *проекции* случайного процесса, является случайной величиной, а множество $\xi(\omega, \cdot) = \{\xi(\omega, x)\}_{x \in \mathcal{X}}$ — функция, отображающая \mathcal{X} в \mathbb{R} , — *реализации случайного процесса*.

Для определения процесса Дирихле нам понадобится некоторое *универсальное множество* I (например, $I = \mathbb{R}^d$), на котором задана вероятностная мера G_0 , которое назовём *базовым распределением*, а также положительное число $\alpha > 0$ — *коэффициент концентрации*.

Определение 14. **Процесс Дирихле как случайное вероятностное распределение вероятностей.** Вероятностная мера G *распределена согласно процессу Дирихле* с параметрами G_0 и $\alpha > 0$, если для любого конечного измеримого разбиения A_1, \dots, A_n случайный вектор $G(A_1), \dots, G(A_n)$ распределён согласно распределению Дирихле с параметрами $(\alpha G_0(A_1), \dots, \alpha G_0(A_n))$:

$$\forall A_1, \dots, A_n : A_i \cap A_j = \emptyset, \quad i \neq j, \quad \cup_{i=1}^n A_i = I \quad (\xi(\omega_0, I) \equiv 1) \quad (326)$$

$$(G(A_1), \dots, G(A_n)) \sim \text{Dir}(\alpha G_0(A_1), \dots, \alpha G_0(A_n)) \quad (327)$$

Обозначается $G \sim DP(G_0, \alpha)$.

Параметр базового распределения носит характер среднего для процесса Дирихле: для любого измеримого множества $A \subseteq I$ $\mathbb{E}[G(A)] = G_0(A)$. Параметр концентрации тесно связан с дисперсией: $\mathbb{D}[G(A)] = G_0(A)(1 - G_0(A))/(\alpha + 1)$. Т.е. чем больше α , тем меньше дисперсия. Логично, что $G(A) \rightarrow G_0$ при $\alpha \rightarrow \infty$.

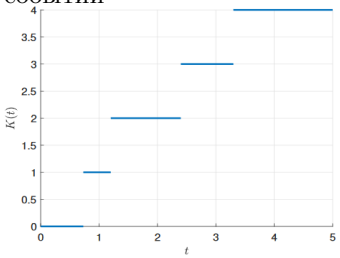
Проведём сравнительную характеристику процесса Дирихле с процессами Пуассона и Гаусса (см. таблицу 5). Ключевым отличием процесса Дирихле от остальных является то, что реализациями процесса являются вероятностные меры. Известно, что для любой вероятностной меры G_0 над \mathbb{R}^d и любого $\alpha > 0$ процесс $DP(G_0, \alpha)$ существует и единственен.

Также стоит отметить важное теоретическое свойство — реализация процесса Дирихле, с вероятностью 1, является дискретной вероятностной мерой, плотность которой можно записать в виде:

$$p(x) = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}(x), \quad p_i > 0, \quad \sum_{i=1}^{\infty} p_i = 1, \quad \theta_i \in I, \quad (328)$$

где $\delta_{\theta_i}(x)$ — дельта-функция Дирака с параметром θ_i ; p_i называются *вероятностями атомов* дискретной меры, а θ_i — *позициями атомов*.

Таблица 5: Сравнение некоторых случайных процессов

Процесс	Гаусса	Пуассона	Дирихле
Индексирующий элемент	$x \in \mathbb{R}^d$	$t \in \mathbb{R}_+$	Измеримое подмножество $A \subseteq I$
Реализация (фиксировано ω_0)	Некоторая функция $\xi(\omega_0, \cdot) : \mathbb{R}^d \rightarrow \mathbb{R}$	Посл-ть поступающих событий 	Вероятностная мера на I $\xi(\omega_0, \cdot) : 2^I \rightarrow [0, 1]$
Параметры	$m(x)$ — мат. ожидание; $K(x, x')$ — ковар. ф-я	λ (или $\lambda(t)$) — интенсивность	G_0 — вер. мера на I ; $\alpha > 0$ — коэф-т концентрации
Одномерная проекция	$f(x_0) \sim \mathcal{N}(m(x_0), K(x_0, x'_0))$	$K(t_0) \sim \text{Pois}(\lambda t_0)$ (или $\text{Pois}(\lambda(t_0))$)	$\xi(\cdot, A_0) = \xi(A_0) = \text{Beta}(\alpha G_0(A_0), \alpha(1 - G_0(A_0)))$
Многомерная проекция	$f(x_1), \dots, f(x_n) \sim \mathcal{N}(\bar{m}, C)$	$K(t_1), \dots, K(t_n) \sim \text{Pois}(\lambda t_1) \cdot \text{Pois}(\lambda(t_2 - t_1)) \cdot \dots \cdot \text{Pois}(\lambda(t_n - t_{n-1}))$	Для разбиения A_1, \dots, A_n : $\xi(A_1), \dots, \xi(A_n) \sim \text{Dir}(\alpha G_0(A_1), \dots, \alpha G_0(A_n))$

10.3.2 Представления процесса Дирихле

Процесс “Ломки палки” Исходя из вида плотности реализаций процесса Дирихле (уравнение (328)) можно получить конструктивное определение процесса Дирихле на основании процесса “Ломки палки”, описанного в свойствах распределения Дирихле.

Определение 15. Если $\alpha > 0$ и G_0 — вероятностная мера на I , тогда вероятностная мера

G с плотностью $p(x)$ (328), генерируемая следующим процессом:

$$v_1, v_2, \dots \sim_{iid} \text{Beta}(1, \alpha) \rightarrow p_k = v_k \prod_{j=1}^{k-1} (1 - v_j) \quad (329)$$

$$\theta_1, \theta_2, \dots \sim_{iid} G_0 \quad (330)$$

является реализацией процесса Дирихле с базовой мерой G_0 и концентрацией α .⁴³

Интуитивно понятно, что из-за наличия ограничения $\sum_{i=1}^{\infty} p_i = 1$ вероятности атомов являются зависимыми величинами. Поэтому генерация величин p_1, p_2, p_3, \dots происходит последовательно на сегментах $[0, 1], [0, 1 - p_1], [0, 1 - p_1 - p_2] \dots$, а позиции θ_i получаются независимыми и одинаково распределёнными по G_0 .

Апостериорное распределение Для осуществления байесовского вывода в моделях с процессами Дирихле необходимо умение получать апостериорное распределение, если априорным является процесс Дирихле:

$$G \sim DP(G_0, \alpha), \quad (331)$$

$$(\theta_1, \dots, \theta_n | G) \sim_{iid} G \quad (332)$$

Сначала, согласно априорному распределению $DP(G_0, \alpha)$ выбирается вероятностная мера G , из которой затем независимо генерируются точки. В рамках этого процесса как бы явно не наблюдается мера G — поясним, что имеется в виду. Величина θ_1 имеет распределение G_0 (согласно рассмотренной выше схеме с ломкой палки). Про вероятность θ_2 при условии θ_1 можно сказать то, что при $\theta_1 = \theta$ в плотности распределения, из которой генерируется θ_2 , должен присутствовать атом θ . Тогда θ_2 либо равно θ снова, либо станет новым атомом. Этот факт отражается в следующей теореме.

Теорема 9. Пусть G — процесс Дирихле $DP(\alpha, G_0)$, и пусть наблюдения $\theta_1, \dots, \theta_n$ сгенерированы процессом (332). Тогда апостериорное распределение на меру G является процессом Дирихле:

$$G | \theta_1, \dots, \theta_n \sim DP\left(\frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{i=1}^n \delta_{\theta_i}, \alpha + n\right), \quad (333)$$

т.е. апостериорное распределение — это взвешенное усреднение базового распределения G_0 и эмпирического $\frac{\sum_{i=1}^n \delta_{\theta_i}}{n}$.

Докажем это.

Т.к. $(G(A_1), \dots, G(A_r)) \sim \text{Dir}(\alpha G_0(A_1), \dots, \alpha G_0(A_r))$ и по свойству сопряжённости распределения Дирихле и мультиномиального распределения имеем:

$$p((G(A_1), \dots, G(A_r)) | \theta_1, \dots, \theta_n) = \frac{1}{Z} p(\theta_1, \dots, \theta_n | (G(A_1), \dots, G(A_r))) \times \quad (334)$$

$$\times p((G(A_1), \dots, G(A_r))) = \frac{1}{Z} \frac{1}{Z_\theta} \prod_{i=1}^r G(A_i)^{\sum_{j=1}^n [\theta_j \in A_i]} \cdot \frac{1}{Z_G} \prod_{i=1}^r G(A_i)^{\alpha G_0(A_i) - 1} = \quad (335)$$

$$= \frac{1}{Z} \prod_{i=1}^r G(A_i)^{(\alpha G_0(A_i) + \sum_{j=1}^n [\theta_j \in A_i]) - 1} = \text{Dir}(\alpha G_0(A_1) + n_1, \dots, \alpha G_0(A_r) + n_r) \quad (336)$$

■

В случае $\alpha \rightarrow 0$ априорное распределение становится неинформативным и предсказания опираются на эмпирическое распределение. То же происходит и в случае $n \gg \alpha$, когда апостериорное распределение доминируется правдоподобием.

⁴³Доказательство корректности схемы в P. Orbanz, Lecture Notes on Bayesian Nonparametrics.

Для генерации из процесса Дирихле на практике самым важным следствием предыдущих рассуждений является следующее:

$$\forall A \subseteq I \ p(\theta_{n+1} \in A \mid \theta_1, \dots, \theta_n) = \mathbb{E}[G(A) \mid \theta_1, \dots, \theta_n] = \frac{1}{\alpha + n} \left(\alpha G_0(A) + \sum_{i=1}^n \delta_{\theta_i}(A) \right), \quad (337)$$

откуда следует

$$\theta_{n+1} \mid \theta_1, \dots, \theta_n \sim \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{i=1}^n \delta_{\theta_i}. \quad (338)$$

Процесс “Китайский ресторан” (Chinese restaurant process, CRP) Для генерации из процесса Дирихле выражение (338) удобнее переписать в виде взвешенной суммы:

$$\theta_{n+1} \mid \theta_1, \dots, \theta_n \sim \frac{\alpha}{\alpha + n} G_0 + \frac{n}{\alpha + n} \frac{\sum_{i=1}^n \delta_{\theta_i}}{n} = \begin{cases} G_0 & , \text{ с вероятностью } \frac{\alpha}{\alpha + n} \\ \delta_c & , \text{ с вероятностью } \frac{\sum_{j=1}^n [\theta_j = c]}{\alpha + n} \end{cases} \quad (339)$$

Эта схема называется урновой схемой Блэквела-МакКвина. Идея опирается на рассуждения вида “богатый становится ещё богаче”: вероятность увидеть конкретное значение тем больше, чем чаще оно уже наблюдалось.

Как можно заметить, в данной схеме происходит разделение точек на группы (точки θ_i попадают в одну группу, если значения совпадают). Такое разбиение называется кластеризацией точек, а процесс разбиения назван процессом “Китайский ресторан” ($CRP(\alpha, n)$) из-за следующей аналогии.

Пусть m — количество кластеров (столов в ресторане). Новый посетитель (θ_{n+1}) либо подсаживается за стол k с вероятностью $\frac{\sum_{j=1}^n [\theta_j = k]}{\alpha + n}$ (посетители более предпочитают столы с уже большей 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 “мизантропность” человека).

Новый кластер при генерации новой точки i образуется с вероятностью $\frac{\alpha}{\alpha + i - 1}$, а потому мат. ожидание на количество кластеров m можно вычислить так:

$$\mathbb{E}[m] = \mathbb{E} \left[\sum_{i=1}^n [\text{+кластер в } i] \right] = \sum_{i=1}^n \frac{\alpha}{\alpha + i - 1} = \alpha (\psi(\alpha + n) - \psi(\alpha)) \simeq \alpha \log \left(1 + \frac{n}{\alpha} \right) \quad (340)$$

Отсюда можно сделать следующие выводы:

1. Параметр α , фактически, линейно влияет на среднее число кластеров.
2. Количество кластеров растёт логарифмически с ростом числа точек.⁴⁴

10.3.3 Смесь распределений с априорным распределением, заданным процессом Дирихле

Рассмотрим классическую модель смеси распределений для множества наблюдений $\{x_1, \dots, x_N\}$ со скрытыми переменными $\{z_1, \dots, z_N\}$, ответственными за принадлежность каждого объекта к определённому кластеру, параметрами для каждого кластера $\{\theta_1, \dots, \theta_K\}$ и весами кластеров $\{\pi_1, \dots, \pi_K\}$:

$$p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\pi}, \boldsymbol{\theta}) = \prod_{n=1}^N \prod_{k=1}^K [\pi_k p(x_n \mid \theta_k)]^{[z_n=k]} \quad (341)$$

$$p(\mathbf{x} \mid \boldsymbol{\pi}, \boldsymbol{\theta}) \rightarrow \max_{\boldsymbol{\pi}, \boldsymbol{\theta}} \quad (342)$$

⁴⁴Порой это слишком медленно, поэтому существуют расширения процесса Дирихле (процесс Питмана-Йорса) с целью изменения этого свойства.

Ограничением данной модели является конечное заранее заданное число компонент K . Более общей моделью будет добавление процесса Дирихле в качестве априорного распределения на параметры компонент смеси. В этом случае может быть задана смесь распределений из бесконечного числа компонент с параметрами $\{\pi_k, \theta_k\}_{k=1}^{\infty} \sim DP(\alpha, G_0)$:

$$p(\mathbf{x}, \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\theta} \mid \alpha, G_0) = p(\boldsymbol{\pi}, \boldsymbol{\theta} \mid \alpha, G_0) p(\mathbf{x}, \mathbf{z} \mid \boldsymbol{\pi}, \boldsymbol{\theta}) = \quad (343)$$

$$= p(\boldsymbol{\pi} \mid \alpha) \prod_{k=1}^{\infty} G_0(\theta_k) \prod_{n=1}^N \prod_{k=1}^{\infty} [\pi_k p(x_n \mid \theta_k)]^{[z_n=k]} \quad (344)$$

Здесь мы воспользовались свойством ломки палки для реализаций процесса Дирихле: параметры $\boldsymbol{\theta}$ являются независимыми одинаково распределёнными согласно мере G_0 , а веса смесей $\boldsymbol{\pi}$ генерируются последовательно.

Зачастую на практике нас интересует именно распределение объектов по кластерам. Для этого необходимо модель (343) маргинализировать по $\boldsymbol{\pi}$:

$$p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta} \mid \alpha, G_0) = \int_{\boldsymbol{\pi}} \left[p(\boldsymbol{\pi} \mid \alpha) \prod_{k=1}^{\infty} G_0(\theta_k) \prod_{n=1}^N p(x_n \mid \boldsymbol{\theta}, z_n) p(z_n \mid \boldsymbol{\pi}) \right] d\boldsymbol{\pi} = \quad (345)$$

$$= \prod_{k=1}^{\infty} G_0(\theta_k) \prod_{n=1}^N p(x_n \mid \boldsymbol{\theta}, z_n) p(\mathbf{z} \mid \alpha), \quad (346)$$

причём маргинальное распределение $p(\mathbf{z} \mid \alpha) = CRP(\alpha, N)$ — процесс “китайского ресторана”, т.е. для каждого объекта x_n мы либо с вероятностью $\frac{\alpha}{\alpha+n-1}$ заводим новый кластер, либо добавляем его в уже имеющийся с вероятностью, пропорциональной количеству точек в этом кластере.

Разберём два метода вывода в, соответственно, маргинализованной (345) и исходной (343) моделях — сэмплирование по Гиббсу и вариационный вывод. В дальнейшем для удобства мы будем опускать в формулах зависимость от параметров G_0 и α .

Схема Гиббса Схема сэмплирования Гиббса применяется непосредственно для маргинализованного апостериорного распределения $p(\mathbf{z}, \boldsymbol{\theta} \mid \mathbf{x})$ — так называемый коллапсированный Гиббс. Для этого необходимо уметь сэмплировать из одномерных распределений $p(z_i \mid \mathbf{z}_{\setminus i}, \boldsymbol{\theta}, \mathbf{x})$ и $p(\theta_i \mid \boldsymbol{\theta}_{\setminus i}, \mathbf{z}, \mathbf{x})$. Для начала, выпишем следующее условное распределение из основной модели:

$$p(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{z}) = p(\boldsymbol{\theta}) p(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{z}) = \prod_{k=1}^{\infty} G_0(\theta_k) \prod_{n=1}^N p(x_n \mid \theta_{z_n}) \quad (347)$$

Из (347) получается второе необходимое условное распределение:

$$p(\theta_k \mid \boldsymbol{\theta}_{\setminus k}, \mathbf{z}, \mathbf{x}) \propto p(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{z}) \propto G_0(\theta_k) \prod_{n: z_n=k} p(x_n \mid \theta_k) \quad (348)$$

По определению CRP:

$$p(z_n \mid \mathbf{z}_{\setminus n}) = \begin{cases} \frac{\sum_{m \neq n} [z_m=k]}{\alpha + N - 1} & , \text{ если } z_n \text{ присоединяется к кластеру } k \\ \frac{\alpha}{\alpha + N - 1} & , \text{ если } z_n \text{ образует новый} \end{cases} \quad (349)$$

Перемножением (349) и правдоподобия $p(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{z})$ получаем требуемое условное распределение, но имеющее разный вид для случаев определения нового кластера и отнесения к старому.

В случае отнесения к старому классу:

$$p(z_n = k \mid \mathbf{z}_{\setminus n}, \boldsymbol{\theta}, \mathbf{x}) \propto \frac{\sum_{m \neq n} [z_m = k]}{\alpha + N - 1} p(x_n \mid \theta_k) \quad (350)$$

Если же возникает новый кластер, то в модели необходимо учесть его появившийся параметр θ_{new} :

$$p(z_n = new, \theta_{new} | \mathbf{z}_{\setminus n}, \boldsymbol{\theta}, \mathbf{x}) \propto \frac{\alpha}{\alpha + N - 1} p(x_n | \theta_{new}) G_0(\theta_{new}) \quad (351)$$

Проинтегрировав по θ_{new} , находим

$$p(z_n = new | \mathbf{z}_{\setminus n}, \boldsymbol{\theta}, \mathbf{x}) \propto \frac{\alpha}{\alpha + N - 1} \int p(x_n | \theta_{new}) G_0(\theta_{new}) d\theta_{new} \quad (352)$$

Нормировочная константа находится путём суммирования (350) и (352).

Вариационный вывод В случае вариационного подхода перепишем исходную модель (343) в терминах переменных $\boldsymbol{\theta}$ и \mathbf{v} :

$$p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta}, \mathbf{v} | \alpha, G_0) = \prod_{n=1}^N \prod_{k=1}^{\infty} \left[p(x_n | \theta_k) v_k \prod_{i=1}^{k-1} (1 - v_i) \right]^{[z_n=k]} \left[\prod_{k=1}^{\infty} G_0(\theta_k) \text{Beta}(v_k | 1, \alpha) \right] \quad (353)$$

Как можно заметить, в сэмплировании Гиббса процесс Дирихле был представлен в виде схемы CRP, здесь же используется представление в виде процесса “Ломки палки”, 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:

$$p(\mathbf{z}, \boldsymbol{\theta}, \mathbf{v} | \mathbf{x}) \approx q(\mathbf{z}) q(\boldsymbol{\theta}) q(\mathbf{v}) \quad (354)$$

Проделаем вариационный вывод для параметров вероятностей кластеров:

$$\log q(\mathbf{v}) = const + \mathbb{E}_{q(\mathbf{z})q(\boldsymbol{\theta})} \log p(\mathbf{x}, \mathbf{z}, \boldsymbol{\theta}, \mathbf{v}) = \quad (355)$$

$$C + \mathbb{E}_{q(\mathbf{z})q(\boldsymbol{\theta})} \left[\sum_{k=1}^{\infty} \log \text{Beta}(v_k | 1, \alpha) + \sum_{n=1}^N \sum_{k=1}^{\infty} [z_n = k] \left[\log v_k + \sum_{i=1}^{k-1} \log(1 - v_i) \right] \right] = \quad (356)$$

$$= \{ \log \text{Beta}(v_k | 1, \alpha) = \log \left(\frac{v_k^{1-1} (1 - v_k)^{\alpha-1}}{B(1, \alpha)} \right) = (\alpha - 1) \log(1 - v_k) + c \} = \quad (357)$$

$$= C + \sum_{k=1}^{\infty} (\alpha - 1) \log(1 - v_k) + \sum_{k=1}^{\infty} \left[\log v_k + \sum_{i=1}^{k-1} \log(1 - v_i) \right] \left[\sum_{n=1}^N \mathbb{E}_{q(\mathbf{z})} [z_n = k] \right] \quad (358)$$

При достаточных статистиках Бета-распределения стоят следующие константы:

$$\log v_k : \sum_{n=1}^N \mathbb{E}_{q(\mathbf{z})} [z_n = k] \quad (359)$$

$$\log(1 - v_k) : \alpha - 1 + \sum_{l>k} \sum_{n=1}^N \mathbb{E}_{q(\mathbf{z})} [z_n = l] \quad (360)$$

Следовательно,

$$q(\mathbf{v}) = \prod_{k=1}^{\infty} \text{Beta}(v_k | 1 + \sum_{n=1}^N \mathbb{E}_{q(\mathbf{z})} [z_n = k], \alpha + \sum_{l>k} \sum_{n=1}^N \mathbb{E}_{q(\mathbf{z})} [z_n = l]) \quad (361)$$

Аналогично из модели получаем:

Для \mathbf{z} :

$$q(\mathbf{z}) = \prod_{n=1}^N q(z_n) \quad (362)$$

$$q(z_n = k) = \frac{1}{A} \exp \left(\mathbb{E}_{q(\mathbf{v})q(\boldsymbol{\theta})} \left[\log p(x_n | \theta_k) + \log v_k + \sum_{i=1}^{k-1} \log(1 - v_i) \right] \right) \quad (363)$$

Для $\boldsymbol{\theta}$:

$$q(\boldsymbol{\theta}) = \prod_{k=1}^{\infty} q(\theta_k) \quad (364)$$

$$q(\theta_k) = \frac{1}{A} \exp \left(\mathbb{E}_{q(\boldsymbol{v})q(\boldsymbol{z})} \left[\sum_{n=1}^N [z_n = k] \log p(x_n \mid \theta_k) + \log G_0(\theta_k) \right] \right) \quad (365)$$

11 Лекция 12. Тематическая модель Latent Dirichlet allocation (LDA)

11.1 Распределение Дирихле

Случайная величина $\boldsymbol{\theta} \in \mathbb{R}^K$, определенная на симплексе ($\theta_k \geq 0, \sum_{k=1}^K \theta_k = 1$), имеет распределение Дирихле, если её плотность определяется как:

$$p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \text{Dir}(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_k^{\alpha_k-1}, \alpha_k > 0, \quad (366)$$

где $\Gamma(\cdot)$ — гамма-функция, $\boldsymbol{\alpha}$ — набор параметров распределения. Введем обозначение для нормировочной константы: $B(\boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)}$.

Распределение Дирихле относится к экспоненциальному классу распределений с набором параметров $[\alpha_1 - 1, \dots, \alpha_K - 1]$ и достаточными статистиками $\mathbf{u}(\boldsymbol{\theta}) = [\log \theta_1, \dots, \log \theta_K]$.

Для распределения из экспоненциального класса все моменты достаточных статистик определяются соответствующими производными нормировочной константы, поэтому

$$\mathbb{E}_p \log \theta_i = \frac{\partial}{\partial (\alpha_i - 1)} \log \frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)} = \Psi(\alpha_i) - \Psi\left(\sum_k \alpha_k\right), \quad (367)$$

где $\Psi(x) = \frac{d}{dx} \log \Gamma(x)$ — дигамма-функция.

Распределение Дирихле часто используется в качестве распределения для набора дискретных вероятностей (распределение над дискретными распределениями). Например, в тематических моделях априорное распределение на тематики можно задать как распределение Дирихле.

11.2 Тематическая модель LDA

Для описания текстов в LDA используется модель мешка слов, т.е. каждый документ рассматривается как набор терминов (слов, словосочетаний), которые в нем используются. При этом порядок употребления терминов в документе игнорируется. Набор текстов подвергается предобработке, которая выделяет термины в каждом документе и отбрасывает слова, которые встречаются практически в каждом тексте, — союзы, предлоги, вводные слова и др.

Основная идея тематической модели LDA состоит в том, что каждый документ имеет несколько тем, смешанных в некоторой пропорции (тематический профиль документа). В качестве априорного распределения на тематические профили используется распределение Дирихле. Если при этом предполагается, что в документе не может присутствовать очень большое количество тем, то параметры распределения Дирихле выбираются так, чтобы поощрялась разреженность тематического профиля. Тема в свою очередь рассматривается как некоторое распределение вероятностей в пространстве слов из общего словаря. Например, генетическая тема задает высокие вероятности для слов «ген», «секвенирование» и т.д., а компьютерная тема задает высокие вероятности для слов «вычисления», «компьютер», «память», «алгоритм» и т.д.

LDA — это генеративная модель. Она предполагает, что процесс порождения текста состоит из двух этапов. На первом этапе для текста выбирается некоторое распределение на темы (тематический профиль). На втором этапе для каждого слова сначала выбирается тема из распределения вероятностей на темах, а затем само слово генерируется из распределения на слова, соответствующего выбранной теме.

Ниже вводятся обозначения для используемых понятий.

$w \in \{1, \dots, W\}$	— номер слова в словаре
$t \in \{1, \dots, T\}$	— номер темы
$d \in \{1, \dots, D\}$	— номер документа
N_d	— число слов в документе d
$\mathbf{w}_d = [w_{d,1}, \dots, w_{d,N_d}]$	— слова в документе d , $w_{d,n} \in \{1, \dots, W\}$
$\boldsymbol{\theta}_d = [\theta_{d,1}, \dots, \theta_{d,T}]$	— вероятности тем в документе d , $\theta_{d,t} \geq 0$, $\sum_{t=1}^T \theta_{d,t} = 1$
$\mathbf{z}_d = [z_{d,1}, \dots, z_{d,N_d}]$	— темы слов в документе d , $z_{d,n} \in \{1, \dots, T\}$
$\boldsymbol{\varphi}_t = [\varphi_{t,1}, \dots, \varphi_{t,W}]$	— вероятности слов в теме t , $\varphi_{t,w} \geq 0$, $\sum_{w=1}^W \varphi_{t,w} = 1$
$\Theta = [\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_D]^T \in \mathbb{R}^{D \times T}$	— вероятности тем во всех документах
$\Phi = [\boldsymbol{\varphi}_1, \dots, \boldsymbol{\varphi}_T]^T \in \mathbb{R}^{T \times W}$	— вероятности слов во всех темах
$(\mathcal{W} = \{\mathbf{w}_1, \dots, \mathbf{w}_D\})$	— полный корпус слов во всех документах
$\mathcal{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_D\}$	— темы всех слов во всех документах корпуса d

Вероятностная модель LDA задаётся следующим образом:

$$p(\mathcal{W}, \mathcal{Z}, \Theta \mid \Phi, \boldsymbol{\alpha}) = \prod_{d=1}^D \left(p(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}) \prod_{n=1}^{N_d} p(w_{d,n} \mid z_{d,n}, \Phi) p(z_{d,n} \mid \boldsymbol{\theta}_d) \right), \quad (368)$$

$$p(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}) = \text{Dir}(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}), \quad (369)$$

$$p(w_{d,n} \mid z_{d,n}, \Phi) = \Phi_{z_{d,n}, w_{d,n}}, \quad p(z_{d,n} \mid \boldsymbol{\theta}_d) = \theta_{d, z_{d,n}}. \quad (370)$$

Таким образом

$$p(\mathcal{W}, \mathcal{Z}, \Theta \mid \Phi, \boldsymbol{\alpha}) = \prod_{d=1}^D \left(\text{Dir}(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}) \prod_{n=1}^{N_d} \Phi_{z_{d,n}, w_{d,n}} \theta_{d, z_{d,n}} \right) = \quad (371)$$

$$= \prod_{d=1}^D \left(\frac{1}{B(\boldsymbol{\alpha})} \prod_{t=1}^T \theta_{d,t}^{\alpha_t - 1} \prod_{n=1}^{N_d} \prod_{t=1}^T \varphi_{t, w_{d,n}}^{[z_{d,n}=t]} \theta_{d,t}^{[z_{d,n}=t]} \right). \quad (372)$$

11.3 Вариационный вывод для модели LDA

В задаче построения тематической модели нам дан корпус документов \mathcal{W} и параметры $\boldsymbol{\alpha}$ априорного распределения на тематические профили. Требуется оценить параметры модели Φ , а также найти апостериорное распределение $p(\mathcal{Z}, \Theta \mid \mathcal{W}, \Phi)$. Поскольку параметры $\boldsymbol{\alpha}$ априорного распределения фиксированы, для упрощения будем опускать их в дальнейших формулах.

Рассмотрим решение задачи обучения модели LDA с помощью метода максимального правдоподобия:

$$p(\mathcal{W} \mid \Phi) \rightarrow \max_{\Phi} \quad (373)$$

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 \mathcal{Z} , что соответствует суммированию по $T^{\sum_d N_d}$ слагаемым.

Воспользуемся вариационным ЕМ-алгоритмом:

$$\text{Е-шаг: } q(\mathcal{Z}, \Theta) \simeq p(\mathcal{Z}, \Theta \mid \mathcal{W}, \Phi), \quad (374)$$

$$\text{М-шаг: } \mathbb{E}_q \log p(\mathcal{W}, \mathcal{Z}, \Theta \mid \Phi) \rightarrow \max_{\Phi}. \quad (375)$$

11.3.1 Е-шаг

На Е-шаге ЕМ-алгоритма необходимо найти апостериорное распределение $p(\mathcal{Z}, \Theta \mid \mathcal{W}, \Phi)$. Распределения $p(\mathcal{Z} \mid \Theta)p(\Theta)$ и $p(\mathcal{W} \mid \mathcal{Z}, \Theta, \Phi)$ не сопряжены, следовательно, нельзя аналитически вывести Е-шаг. Поэтому воспользуемся вариационным подходом и будем искать аппроксимирующее распределение q в семействе факторизованных распределений:

$$q(\mathcal{Z}, \Theta) \approx q(\mathcal{Z})q(\Theta). \quad (376)$$

Тогда относительно Θ правдоподобие $p(\mathcal{Z} \mid \Theta)$ и априорное распределение $p(\Theta)$ сопряжены. Аналогично по параметру \mathcal{Z} правдоподобие $p(\mathcal{W} \mid \mathcal{Z}, \Theta, \Phi)$ и априорное распределение $p(\mathcal{Z} \mid \Theta)$ сопряжены. Значит, имеется условная сопряжённость распределений, и можно использовать метод Mean Field Approximation для выполнения Е-шага.

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 $\text{const}(\Theta)$:

$$\log q(\Theta) = \mathbb{E}_{q(\mathcal{Z})} \log p(\mathcal{W}, \mathcal{Z}, \Theta \mid \Phi) = \quad (377)$$

$$= \mathbb{E}_{q(\mathcal{Z})} \sum_{d=1}^D \left(\sum_{t=1}^T (\alpha_t - 1) \log \theta_{d,t} + \sum_{n=1}^{N_d} \sum_{t=1}^T [z_{d,n} = t] (\log \varphi_{t,w_{d,n}} + \log \theta_{d,t}) \right) + \text{const}(\Theta) = \quad (378)$$

$$= \sum_{d=1}^D \left(\sum_{t=1}^T (\alpha_t - 1) \log \theta_{d,t} + \sum_{t=1}^T \sum_{n=1}^{N_d} \mathbb{E}_{q(\mathcal{Z})} [z_{d,n} = t] \log \theta_{d,t} \right) + \text{const}(\Theta) = \quad (379)$$

$$= \sum_{d=1}^D \left(\sum_{t=1}^T \log \theta_{d,t} \left(\alpha_t - 1 + \sum_{n=1}^{N_d} \mathbb{E}_{q(\mathcal{Z})} [z_{d,n} = t] \right) \right) + \text{const}(\Theta). \quad (380)$$

Пусть $\mu_{d,n,t} = \mathbb{E}_{q(\mathcal{Z})} [z_{d,n} = t]$. Тогда можем записать

$$q(\Theta) = \prod_{d=1}^D q(\theta_d) = \prod_{d=1}^D \text{Dir}(\theta_d \mid \alpha_1 + \sum_{n=1}^{N_d} \mu_{d,n,1}, \dots, \alpha_T + \sum_{n=1}^{N_d} \mu_{d,n,T}). \quad (381)$$

Теперь распишем $\log q(\mathcal{Z})$, группируя все члены, не зависящие от \mathcal{Z} , в $\text{const}(\mathcal{Z})$:

$$\log q(\mathcal{Z}) = \mathbb{E}_{q(\Theta)} \log p(\mathcal{W}, \mathcal{Z}, \Theta \mid \Phi) = \quad (382)$$

$$= \mathbb{E}_{q(\Theta)} \sum_{d=1}^D \left(\sum_{t=1}^T (\alpha_t - 1) \log \theta_{d,t} + \sum_{n=1}^{N_d} \sum_{t=1}^T [z_{d,n} = t] (\log \varphi_{t,w_{d,n}} + \log \theta_{d,t}) \right) + \text{const}(\mathcal{Z}) = \quad (383)$$

$$= \sum_{d=1}^D \sum_{n=1}^{N_d} \sum_{t=1}^T [z_{d,n} = t] (\mathbb{E}_{q(\Theta)} \log \theta_{d,t} + \log \varphi_{t,w_{d,n}}) + \text{const}(\mathcal{Z}). \quad (384)$$

Следовательно

$$q(\mathcal{Z}) = \prod_{d=1}^D \prod_{n=1}^{N_d} q(z_{d,n}), \quad (385)$$

где

$$q(z_{d,n} = t) = \frac{\exp (\mathbb{E}_{q(\Theta)} \log \theta_{d,t} + \log \varphi_{t,w_{d,n}})}{\sum_{s=1}^T \exp (\mathbb{E}_{q(\Theta)} \log \theta_{d,s} + \log \varphi_{s,w_{d,n}})} = \mu_{d,n,t}. \quad (386)$$

11.3.2 М-шаг

На М-шаге ЕМ-алгоритма необходимо оптимизировать вариационную нижнюю оценку $\mathbb{E}_{q(\mathcal{Z}, \Theta)} \log p(\mathcal{W}, \mathcal{Z}, \Theta \mid \Phi)$ по параметрам модели Φ . Распишем данную оптимизационную задачу:

$$\mathbb{E}_{q(\mathcal{Z})q(\Theta)} \sum_{d=1}^D \left(\sum_{t=1}^T (\alpha - 1) \log \theta_{d,t} + \sum_{n=1}^{N_d} \sum_{t=1}^T [z_{d,n} = t] (\log \varphi_{t,w_{d,n}} + \log \theta_{d,t}) \right) + \text{const}(\Theta, \mathcal{Z}) \rightarrow \max_{\Phi} \quad (387)$$

$$\Rightarrow \sum_{d=1}^D \sum_{n=1}^{N_d} \sum_{t=1}^T (\mu_{d,n,t} \log \varphi_{t,w_{d,n}}) \rightarrow \max_{\Phi} \quad (388)$$

Также знаем, что

$$\sum_{v=1}^W \varphi_{t,v} = 1 \quad \forall t = \overline{1, T}. \quad (389)$$

Составим функцию Лагранжа:

$$\mathcal{L} = \sum_{d=1}^D \sum_{n=1}^{N_d} \sum_{t=1}^T (\mu_{d,n,t} \log \varphi_{t,w_{d,n}}) + \sum_{t=1}^T \lambda_t \left(\sum_{v=1}^W \varphi_{t,v} - 1 \right). \quad (390)$$

Для нахождения $\varphi_{t,v}$ найдем экстремум функции Лагранжа по этим переменным:

$$\frac{\partial \mathcal{L}}{\partial \varphi_{t,v}} = \sum_{d=1}^D \sum_{n=1}^{N_d} \mu_{d,n,t} \frac{1}{\varphi_{t,v}} [w_{d,n} = v] + \lambda_t = 0 \quad (391)$$

$$\Rightarrow \varphi_{t,v} = - \frac{\sum_{d=1}^D \sum_{n=1}^{N_d} \mu_{d,n,t} [w_{d,n} = v]}{\lambda_t}. \quad (392)$$

Суммируя это равенство по v и зная, что $\sum_{v=1}^W \varphi_{t,v} = 1$, найдем λ_t :

$$1 = \sum_{v=1}^W \varphi_{t,v} = \sum_{v=1}^W - \frac{\sum_{d=1}^D \sum_{n=1}^{N_d} \mu_{d,n,t} [w_{d,n} = v]}{\lambda_t} \quad (393)$$

$$\Rightarrow \lambda_t = - \sum_{d=1}^D \sum_{n=1}^{N_d} \mu_{d,n,t}. \quad (394)$$

Таким образом, получим выражение для $\varphi_{t,v}$:

$$\varphi_{t,v} = \frac{\sum_{d=1}^D \sum_{n=1}^{N_d} \mu_{d,n,t} [w_{d,n} = v]}{\sum_{d=1}^D \sum_{n=1}^{N_d} \mu_{d,n,t}}. \quad (395)$$

Е- и М-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.