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

Введение

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

1 Лекция 1. Байесовский подход к теории вероятностей

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

1.1 Основные понятия

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$p(x) = \int p(x \mid y)p(y)\mathrm{d}y = \mathbb{E}_y[p(x \mid y)]. \tag{8}$$

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

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

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

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

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

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

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

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

$$p(x \mid y) = \frac{p(x, y)}{p(y)} = \frac{\int p(x, y, z)\mathrm{d}z}{\iint p(x, y, z)\mathrm{d}x\mathrm{d}z}. \tag{10}$$

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

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

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

1.2 Частотный и байесовский подходы

В рамках классических курсов изучался подход, который в англоязычной литературе называют *частотным* или *фрекевентистским* (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 $p(x \mid \theta)$, т.е. рассматривать параметры θ как случайные величины, — смысл от этого не меняется. Чтобы оценить параметры θ, в классическом частотном подходе используется метод максимального правдоподобия [5] :

$$\theta_{\text{ML}} = \arg\max_\theta p(X \mid \theta) = \arg\max_\theta \prod_{i=1}^{n} p(x_i \mid \theta) = \arg\max_\theta \sum_{i=1}^{n} \log p(x_i \mid \theta). \tag{11}$$

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

Оценка максимума правдоподобия (ОМП) обладает очень хорошими свойствами:

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

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

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

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

$$p(\theta|X) = \frac{\prod_{i=1}^{n} p(x_i \mid \theta) \cdot p(\theta)}{\int \prod_{i=1}^{n} p(x_i \mid \theta) \cdot p(\theta) d\theta} \tag{12}$$

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

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

Рассмотрим примеры субъективного незнания.

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

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 $n \geqslant 0$

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

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

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

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

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

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

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

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

1.3 Приятные плюсы байесовского подхода

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

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

$$p(y \mid x, z) = \frac{p(z \mid y)p(y \mid x)}{\int p(z \mid y)p(y \mid x)\mathrm{d}y}. \tag{13}$$

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

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

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

1.4 Байесовский подход как обобщение булевой логики

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

$$p(B) = \sum_A p(B \mid A)p(A) \tag{14}$$

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

$$p(A \mid B) = \frac{p(B \mid A)p(A)}{\sum\limits_A p(B \mid A)p(A)} \tag{15}$$

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

1.5 Пример байесовских рассуждений

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

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

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

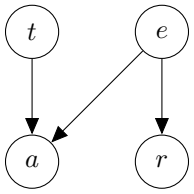

Рис. 1: Граф зависимостей в задаче про сигнализацию.

По такому графу несложно задать совместное распределение на все величины:

$$p(a, e, r, t) = p(a \mid e, t)p(r \mid e)p(t)p(e). \tag{16}$$

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

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

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

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

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

$$p(t = 1 \mid a = 1) = \frac{p(a = 1 \mid t = 1)p(t = 1)}{p(a = 1 \mid t = 0)p(t = 0) + p(a = 1 \mid t = 1)p(t = 1)}. \tag{17}$$

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

$$\begin{aligned} p(a = 1 \mid t = 0) &= p(a = 1 \mid e = 0, t = 0)p(e = 0) + p(a = 1 \mid e = 1, t = 0)p(e = 1) \\ &= 0 + 0.1 \cdot 10^{-2} = 10^{-3} \end{aligned} \tag{18}$$

Тогда

$$p(t = 1 \mid a = 1) = \frac{1 \cdot 2 \cdot 10^{-4}}{10^{-3} \cdot (1 - 2 \cdot 10^{-4}) + 1 \cdot 2 \cdot 10^{-4}} \approx \frac{1}{6} \tag{19}$$

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

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

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

$$p(t = 1 \mid a = 1, r = 1) = \frac{p(a = 1, t = 1, r = 1)}{p(a = 1, r = 1)} = \frac{\sum\limits_{e} p(a = 1, e, t = 1, r = 1)}{\sum\limits_{e,t} p(a = 1, e, t, r = 1)} = \tag{20}$$

$$= \frac{\sum\limits_{e} p(a = 1 \mid e, t = 1)p(r = 1 \mid e)p(e)p(t = 1)}{\sum\limits_{e,t} p(a = 1 \mid e, t)p(r = 1 \mid e)p(e)p(t)}. \tag{21}$$

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

$$p(t = 1 \mid a = 1, r = 1) = \frac{1 \cdot 0.5 \cdot 10^{-2} \cdot 2 \cdot 10^{-3}}{10^{-1} \cdot 0.5 \cdot 10^{-2} \cdot (1 - 2 \cdot 10^{-3}) + 1 \cdot 0.5 \cdot 10^{-2} \cdot 2 \cdot 10^{-3}}. \tag{22}$$

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 землятрясении).

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

2 Лекция 2. Сопряженные распределения, экспоненциальный класс распределений

2.1 Сопряжённые распределения

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

$$p(\theta \mid X) = \frac{p(X \mid \theta)p(\theta)}{\int p(X \mid \theta)p(\theta)\mathrm{d}\theta} \tag{23}$$

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

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

$$\theta_{MP} = \arg\max_\theta p(\theta \mid X) = \arg\max_\theta p(X \mid \theta)p(\theta) = \tag{24}$$

$$= \arg\max_\theta \left(\prod_{i=1}^n p(x_i \mid \theta)p(\theta) \right) = \arg\max_\theta \left(\sum_{i=1}^n \ln p(x_i \mid \theta) + \ln p(\theta) \right) \tag{25}$$

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 из 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.

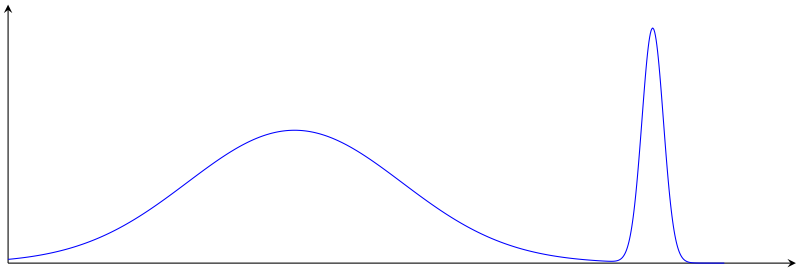

Рис. 2: Пример распределения, у которого мода нерепрезентативна.

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

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

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

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

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

$$p(\mu) = \frac{1}{\sqrt{2\pi s}} \exp\left(-\frac{\mu^2}{2s^2} + \frac{\mu m}{s^2} - \frac{m^2}{2s^2}\right) = \mathcal{N}(\mu \mid m, s^2) \tag{27}$$

   Теперь проверим:

$$p(x \mid \mu)p(\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2} + x\mu - \frac{\mu^2}{2}\right) \cdot \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{\mu^2}{2s^2} + \frac{\mu m}{s^2} - \frac{m^2}{2s^2}\right) \propto$$
$$\propto \exp\left(-\frac{\mu^2(s^2+1)}{2s^2} + \frac{\mu(m+xs^2)}{s^2} - \frac{x^2s^2+m^2}{2s^2}\right) \propto \exp\left(-\frac{s^2+1}{2s^2}\left(\mu - \frac{m+xs^2}{s^2+1}\right)^2\right) \propto$$
$$\propto \mathcal{N}\left(\mu \,\middle|\, \frac{m+xs^2}{s^2+1}, \frac{s^2}{s^2+1}\right)$$

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

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

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

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

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

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

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

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

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

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

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

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

2.2 Экспоненциальный класс распределений

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

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

$$p(x \mid \theta) = \frac{f(x)}{g(\theta)} \exp\big(\theta^T u(x)\big), \quad f(\cdot) \geqslant 0, \;\; g(\cdot) > 0, \tag{28}$$

Параметры θ называются *естественными параметрами*.

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

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

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

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

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

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

Рассмотрим одно замечательное свойство экспоненциального класса распределений. Заметим, что

$$g(\theta) = \int f(x) \exp\left(\theta^T u(x)\right) \mathrm{d}x, \quad \text{т.к.} \quad \int \frac{f(x)}{g(\theta)} \exp\left(\theta^T u(x)\right) \mathrm{d}x = 1 \tag{29}$$

Прродифференцируем по θ_j

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

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

В итоге получаем, что

$$\frac{\partial}{\partial \theta_j} \log g(\theta) = \mathbb{E}_{x \sim p(x|\theta)} u_j(x) \tag{30}$$

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

$$\frac{\partial^2}{\partial \theta_j \partial \theta_k} \log g(\theta) = Cov(u_j(x), u_k(x)) \tag{31}$$

2.2.1 Оценка параметров распределения из экспоненциального класса

Пусть нам дана выборка из распределения экспоненциального класса:

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

Оценим параметры распределения методом максимального правдоподобия

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

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 нармальное распределение

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

Из выражения выше видно, что

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

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

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

2.2.2 Сопряженное семейство к экспоненциальному классу

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

$$p(\theta \mid \eta, \nu) = \exp\bigl(\theta^T \eta\bigr) \frac{1}{g^\nu(\theta)} \frac{1}{h(\eta, \nu)} \tag{32}$$

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:

$$p(\theta \mid X) = \frac{1}{Z} \prod_{i=1}^{n} p(x_i \mid \theta) p(\theta \mid \nu, \eta) = \tag{33}$$

$$= \frac{1}{Z} \prod_{i=1}^{n} \Bigl[f(x_i)\Bigr] \cdot \frac{1}{g^n(\theta)} \exp\left\{\theta^T \left(\sum_{i=1}^{n} u(x_i)\right)\right\} \exp\bigl\{\theta^T \eta\bigr\} \frac{1}{g^\nu(\theta)} \frac{1}{h(\eta, \nu)} = \tag{34}$$

$$= \frac{1}{Z'} \exp\left\{\theta^T \left(\eta + \sum_{i=1}^{n} u(x_i)\right)\right\} \frac{1}{g^{\nu+n}(\theta)} = \frac{1}{h(\eta', \nu')} \exp\left(\theta^T \eta'\right) \frac{1}{g^{\nu'}(\theta)} \tag{35}$$

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

$$\eta' = \eta + \sum_{i=1}^{n} u(x_i) \tag{36}$$

$$\nu' = \nu + n \tag{37}$$

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