Основания Байесовской статистики

Зайцев А., Янович Ю.
Сколтех, ИППИ РАН
likzet@gmail.com

Содержание

1 Предисловие

На русском языке существует несколько книг, которые затрагивают использование Байесовских методов машинному обучении. В частности, можно порекомендовать учебное пособие Дмитрия Ветрова [7] на русском языке. Однако, до сих пор должного внимания не было уделено обоснованию Байесовского подхода — в частности, с точки зрения классической математической статистики.

Данное учебное пособие призвано заполнить этот пробел в литературе, доступной на русском языке. В первую очередь автор ориентировался на лекции Майкла Джордана, которые были прочитаны в Беркли в 2010 году [5]. Получившийся результат во многом посвящен основному вопросу Байесовской статистики — выбору априорного распределения, хотя уделяет внимание и другим важным вопросам.

Оказывается, что Байесовская статистика использует многие фундаментальные понятия классической математической статистики и теории информации. Например, энтропию или теорию принятия решений. Понимание принципов Байесовской статистики, таким образом, может оказаться полезным и в Байесовском машинном обучении, и для более глубокого понимания современной математической статистики.

Авторы выражают благодарность Е.В.Бурнаеву за возможность чтения курса по Байесовским методам в ИППИ РАН, ВШЭ и Сколтехе. Именно результатом опыта чтения курса и стало это пособие.

2 Байесовский подход

2.1 Основные понятия

Байесовский подход предлагает более гибкую трактовку традиционного вероятностного подхода. Введем понятия, которые используются в Байесовской математической статистике и Байесовском машинном обучении. Ограничимся здесь параметрическими моделями в конечномерных пространствах.

Обычно нас интересует значение параметра или вектора параметров $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$. Мы оцениваем значение параметра по данным $X \in \mathcal{X} \subseteq \mathbb{R}^d$. Для заданной вероятностной модели можно записать правдоподобие $p(X|\boldsymbol{\theta})$. Также задано априорное распределение $\pi(\boldsymbol{\theta})$. Помимо априорного распределения и правдоподобия определено маргинальное распределение

$$p(X) = \int_{\Theta} p(X, \boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{\Theta} p(X|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

Зная правдоподобие, апостериорное распределение и маргинальное распределение, мы можем записать апостериорную плотность распределения $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta}|X) = \frac{p(X|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(X)}.$$

Помимо апостериорной плотности распределения нас часто интересует апостериорное прогнозное распределение для новых данных X_{new}:

$$p(X_{\text{new}}|X) = \int p(X_{\text{new}}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|X) d\boldsymbol{\theta}.$$

Часто нет необходимости рассматривать маргинальное распределение, так как оно не зависит от $\boldsymbol{\theta}$, и с точностью до нормировочного коэффициента

$$p(\boldsymbol{\theta}|X) \propto p(X|\boldsymbol{\theta})\pi(\boldsymbol{\theta}).$$

Чтобы использовать введенные выше понятия для решения реальных задач нужно ответить на два вопроса: как выбрать априорное распределение и как вычислить все вероятности, определенные выше. В данном курсе мы сосредоточимся на первом вопросе — хотя по мере возможности осветим и второй.

2.2 Классические критерии статистического оценивания

Приведем критерии качества статистического оценивания, которые используются в классической математической статистике.

Состоятельность. Пусть $\widehat{\theta}_n = \widehat{\theta}(x_1, \ldots, x_n)$. Тогда оценка $\widehat{\theta}_n$ называется состоятельной, если

$$\forall \theta \in \Theta \widehat{\theta}_n \to^p \theta \text{ при } n \to \infty.$$

То есть, оценка сходится к истинному значению параметра по вероятности, если размер выборки стремится к бесконечности.

Скорость сходимости. Нам часто важен не только сам факт сходимости, но и скорость сходимости. То есть, нас интересует типичное значение величины $\|\widehat{\theta}_n - \theta\|$ в зависимости от n. Иногда сходимость порядка $\frac{1}{\sqrt{n}}$ оказывается слишком медленной.

Несмещенность. Оценку $\widehat{\theta}$ назовем несмещенной, если $\mathbb{E}\widehat{\theta} = \theta$.

Эффективность. Оценка будет эффективной, если она обеспечивает наилучшую возможную скорость сходимости. В достаточно широком наборе моделей удается получить такие оценки.

Поясним введенные выше понятия с помощью примера.

2.3 Пример использования Байесовского подхода

Пример 2.1. Пусть $x_i \sim \mathcal{N}(\theta, \sigma^2)$. Есть выборка данных $\mathrm{D} = \{x_1, \ldots, x_n\}$, причем x_i получены независимо, и они из одного и того же распределения. Задача заключается в оценке выборке значения параметра θ.

Будем использовать метод максимума правдоподобия:

$$p(D|\theta) \to \max_{\theta}.$$

Правдоподобие для такой модели имеет вид:

$$p(D|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \theta)^2\right).$$

Максимизация логарифма правдоподобия — что то же самое, что и максимизация правдоподобия — эквивалентно задаче минимизации:

$$\sum_{i=1}^{n}(x_i - \theta)^2 \to \min_{\theta}.$$

Дифференцируя и приравнивая к нуля производную, получаем оценку максимума правдоподобия (maximum likelihood estimate):

$$\widehat{\theta}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Подсчитаем теперь математическое ожидание и дисперсию такой оценки:

$$\mathbb{E}\widehat{\theta}_{MLE} = \theta,$$
$$\mathbb{V}\widehat{\theta}_{MLE} = \frac{1}{n}\mathbb{V}x = \frac{\sigma^2}{n}.$$

Таким образом, оценка максимума правдоподобия несмещенная и состоятельная. Действительно, среднее значение совпадает с истинным значением параметра и выполнено условие состоятельности:

$$\mathbb{V}\widehat{\theta}_{MLE} \to 0, n \to \infty.$$

Скорость сходимости $\mathbb{V}\|\widehat{\theta}_{MLE} - \theta\|^2 \sim \frac{1}{n}$.

Более того, оказывается, что такая оценка является эффективной. Так как нормальное распределение принадлежит экспоненциальному семейству, то выполнено неравенство Рао-Крамера:

$$\mathbb{V}\widehat{\theta}_{MLE} \geq I_\theta^{-1},$$

где $I_\theta = -\mathbb{E}\left[\frac{\partial^2 p(D|\theta)}{\partial \theta^2}\right]$ — информация Фишера, причем для заданной модели выполнено, что $I_\theta = \frac{n}{\sigma^2}$.

Таким образом, для оценки среднего значения нормального распределения выполнено, что оценка максимума правдоподобия состоятельная, несмещенная и эффективная, причем дисперсия оценки убывает как $\frac{1}{n}$.

Посмотрим теперь на Байесовскую оценку. Пусть задано априорное распределение $\pi(\theta) = \mathcal{N}(\mu, \sigma_\theta^2)$. Тогда для апостериорного распределения выполнено, что

$$p(\theta|D) \propto p(D|\theta)\pi(\theta).$$

Сделав несложные преобразования получаем, что апостериорное распределение тоже будет нормальным, а именно:

$$p(\theta|D) = \mathcal{N}\left(\frac{\frac{1}{\sigma^2}\sum_{i=1}^n x_i + \frac{\mu}{\sigma_\theta^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_\theta^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_\theta^2}}\right).$$

Легко видеть, что такая оценка состоятельна, но не является несмещенной и эффективной. С другой стороны оказывается, что такая оценка *асимптотически* несмещенная и эффективная.

3 Идеи Байесовской статистики

3.1 Проблематика Байесовского подхода

Пусть мы наблюдаем случайную величину $x \in X$ из условного распределения $p(x|\theta)$. В Байесовской статистике мы будем считать, что $\theta \in \Theta$ — тоже случайная величина. Пока будем считать, что x и θ — непрерывные случайные величины.

Классическая задача статистического оценивания заключается в оценке параметра θ по наблюдениям. Аналогичная задача решается и в Байесовской статистике — но теперь нас интересует не просто оценка $\widehat{\theta}$, а все распределение $p(\theta|x)$.

Это *апостериорное распределение* может быть получено из правдоподобия $p(x|\theta)$, априорного распределения $\pi(\theta)$ и маргинального распределения $p(x)$ с использованием формулы Байеса:

$$p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{p(x)}.$$

Так как знаменатель не зависит от θ часто удобно работать только с числителем, произведением правдоподобия на плотность априорного распределения:

$$p(\theta|x) \propto p(x|\theta)\pi(\theta).$$

Таким образом, мы ввели следующие объекты:

- априорное распределение $\pi(\theta)$,
- правдоподобие $p(x|\theta)$,
- апостериорное распределение $p(\theta|x)$,
- маргинальное распределение $p(x) = \int p(x|\theta)\pi(\theta)d\theta$.

В классической статистике мы предполагаем, что задана модель данных, которая определяет правдоподобие $p(x|\theta)$. На основании этой информации мы делаем оценку $\widehat{\theta}$. В Байесовской статистике для того, чтобы полностью задать вероятностную модель, этого оказывается недостаточно: нам нужно еще определить априорное распределение данных.

Грубо все подходы к выбору априорного распределения можно разбить на три пересекающихся группы: субъективный подход, объективный подход и прагматичный подход. Мы уделим наибольшее внимание объективному подходу, но рассмотрим и другие.

В *субъективном* подходе мы считаем, что эксперты в предметной области дали дам готовое априорное распределение, и нам его выбирать не нужно. Таким образом, в этом случае апиорное распределение уже задано и нам его выбирать не нужно. Однако, обычно это не так.

В *объективном* подходе к выбору априорного распределения мы хотим минимизировать влияние априорного распределения на наши выводы. Таким образом, нужно выбрать такое априорное распределение, у которого будут одинаковые предпочтения относительно всех $\theta \in \Theta$. Возникновение объективного подхода связано с тем, что основным направлением развития математической статистики были вероятностные методы, и альтернативы должны были в первую очередь давать правильные в смысле этого основного направления результаты. Примерами объективного подхода является априорное распределение Джеффриса и опорное априорное распределение, которые мы рассмотрим дальше. Так как индифферентность можно понимать по-разному, в этом подходе существует несколько подходов.

В прагматичном подходе нам в первую очередь интересно наличие определенных свойств у полученной оценки — например, ее численная устойчивость или разреженность. Выбор априорного распределения позволяет гарантировать некоторые такие свойства, поэтому часто Байесовские методы используют, например, для регуляризации.

3.2 Объективный Байесовский подход

Формулу Байеса знали — и скорее всего открыли примерно одновременно — Лаплас и Байес. Так как они смотрели на нее с классической точки зрения, то первым их порывом было взять в качестве априорного распределения то, которое обеспечит минимальное влияния на апостериорное распределение.

Естественный кандидат в таком случае — равномерное распределение на множестве параметров Θ. То есть, $\pi(\theta) \propto c$. В таком случае

$$p(\theta|x) \propto p(x|\theta),$$

и мы будем получать одинаковые результаты с использованием методов, которые работают и с правдоподобием, и с апостериорным распределением.

Однако, такое априорное распределение решает нашу проблему в очень узком смысле. Рассмотрим взаимно-однозначное преобразование случайной величины θ, равномерно распределенной на $\Theta = [0, 1]$. Например:

$$\begin{aligned} \rho &= \frac{\theta}{1-\theta}, \rho \in (0, \infty), \\ r &= \log\left(\frac{\theta}{1-\theta}\right), r \in (-\infty, \infty), \end{aligned}$$

В таком случае получим, что априорные распределения $\pi(\rho)$ и $\pi(r)$ уже не будут равномерными.

Таким образом, выбор в качестве априорного распределения равномерного не обеспечивает отсутствие предпочтений к различным значениям параметра.

Этот пример и подобные ему привели к тому, что Байесовский подход в статистике практически не использовался. Фишер, Нейман, Вальд и Колмогоров строили математическую статистику, в которой не было места априорным распределениям.

Однако, со времени стало понятно, что на самом деле Байесовская и классическая статистики изучают одно и то же, но с разных сторон — подобно тому, как часть физиков в девятнадцатом веке считали, что свет это частица, а другая часть — что это волна.

Байесовский подход можно формализовать в рамках теории принятия решений. Кроме того, реабилитации Байесовского подхода в математической статистике поспособствовали три фундаментальных результата: Де Финетти удалось показать, что удачный выбор априорного распределения позволяет представить в новом виде задачу оценки свойств параметра, Джеффрису удалось развить идеи Лапласа и определить априорные распределения, которые минимальное бы влияли на апостериорное распределение, а теорема Бернштейна фон-Мизеса показала, что Байесовские оценки в достаточно общих условиях не сильно уступают по качеству классическим вероятностным оценкам. Де Финетти оправдал использование субъективного подхода в Байесовской статистике, а Джеффрис по новому взглянул на объективный подход.

Еще более важным фактором, повлиявшим на развитие Байесовских идей, стало их использование для решения прикладных задач, в том числе в машинном обучении.

3.3 Теорема де Финетти

Пускай случайные величины $\mathbf{x}_i$ таковы, что их совместная функция распределения не меняется в случае произвольной перестановки элементов выборки $\mathrm{D} = \{x_i\}_{i=1}^n$:

$$P(x_1 \leq y_1, \dots, x_n \leq y_n) = P(x_1 \leq y_{i_1}, \dots, x_n \leq y_{i_n}).$$

Такой набор случайных величин будем называть перестановочным. Будем говорить, что последовательность $x_i, i = 1, 2, \dots, n, n+1, \dots$ бесконечно перестановочна, если для любого $n > 1$ выполнено, что $x_1, \dots, x_n$ перестановочна.

Теорема 1. *Пусть x_i составляют бесконечную перестановочную последовательность, и каждое x_i принимает значения 0 или 1. Тогда для некоторого распределения $\pi(\theta)$ выполнено, что*

$$P(x_1 = v_1, \dots, x_n = v_n) = \int_0^1 \theta^{\sum_{i=1}^n v_i} (1-\theta)^{n - \sum_{i=1}^n v_i} d\pi(\theta)$$

для произвольного n и набора $v_i \in \{0, 1\}$. То есть, для заданного θ выполнено, что $x_1, \dots, x_n$ — условно независимые одинаково распределенные Бернуллиевские случайные величины с параметром θ, и априорное распределение θ — $\pi(\theta)$.

Приведенная теорема может быть обобщена на случай, если множество значений, которые принимают x_i не ограничиваются 0 и 1. Связь между приведенной выше теоремой Де Финетти и теоремой Де Финетти общего вида примерно такая же, как между теоремой Муавра-Лапласа и Центральной предельной теоремой.

Доказательство. Приведем доказательство теоремы. Обозначим

$$p(v_1, \dots, v_n) = p(x_1 = v_1, \dots, x_n = v_n).$$

Пусть $x_1 + \ldots + x_n = y_n$ для y_n из $\{1, \ldots, n\}$. Тогда для произвольной перестановки $(\tau(1), \ldots, \tau(n))$ индексов $(1, \ldots, n)$ выполнено, что

$$p(x_1 + \ldots + x_n = y_n) = C_n^{y_n} p(x_{\tau(1)}, \ldots, x_{\tau(n)}).$$

Или

$$p(x_{\tau(1)}, \ldots, x_{\tau(n)}) = \frac{1}{C_n^{y_n}} p(x_1 + \ldots + x_n = y_n).$$

Для произвольного N, такого что $N \geq n \geq y_n \geq 0$ выполнено, что

$$\begin{aligned}
&p(x_1 + \ldots + x_n = y_n) = \\
&= \sum_{y_N = y_n}^{N-(n-y_n)} p(x_1 + \ldots + x_n = y_n | x_1 + \ldots + x_N = y_N) p(x_1 + \ldots + x_N = y_N) = \\
&= \sum_{y_N = y_n}^{N-(n-y_n)} \frac{C_{y_N}^{y_n} C_{N-y_n}^{n-y_n}}{C_N^n} p(x_1 + \ldots + x_N = y_N).
\end{aligned}$$

Для фиксированного значения $x_1 + \ldots + x_N$ мы можем записать условную вероятность, используя биномиальные коэффициенты, в силу перестановочности случайных величин. То есть, мы можем записать ее как вероятность достать из урны с N шарами, y_N из которых белые, а $N - y_n$ черные, n шаров так, что из них y_n белых.

Перепишем теперь

$$\frac{C_{y_N}^{y_n} C_{N-y_n}^{n-y_n}}{C_N^n} = C_n^{y_n} \frac{(y_N)_{y_n} (N - y_N)_{n-y_n}}{(N)_n},$$

где $(N)_n = \frac{N!}{(N-n)!}$.

Таким образом,

$$\begin{aligned}
p(x_{\tau(1)}, \ldots, x_{\tau(n)}) &= \frac{1}{C_n^{y_n}} p(x_1 + \ldots + x_n = y_n) = \\
&= \frac{1}{C_n^{y_n}} \sum_{y_N = y_n}^{N-(n-y_n)} C_n^{y_n} \frac{(y_N)_{y_n} (N - y_N)_{n-y_n}}{(N)_n} p(x_1 + \ldots + x_N = y_N) = \\
&= \sum_{y_N = y_n}^{N-(n-y_n)} \frac{(y_N)_{y_n} (N - y_N)_{n-y_n}}{(N)_n} p(x_1 + \ldots + x_N = y_N)
\end{aligned}$$

Пусть $\Pi_N(\theta)$ совпадает с функцией распределения $x_1 + \ldots + x_N$, деленной на N. То есть, для $\theta < 0$ функция $\Pi_N(\theta) = 0$, в точках $\theta = \frac{y_N}{N}$ она испытывает скачок, равный $p(x_1 + \ldots + x_N = y_N)$, и не меняется в других точках.

Тогда

$$p(x_{\tau(1)}, \ldots, x_{\tau(n)}) = \int_0^1 \frac{(\theta N)_{y_n} ((1-\theta) N)_{n-y_n}}{(N)_n} d\Pi_N(\theta).$$

Для $N \to \infty$ выполнено, что

$$\frac{(\theta N)_{y_n}((1-\theta)N)_{n-y_n}}{(N)_n} \to \theta^{y_n}(1-\theta)^{n-y_n}$$

Действительно, для малых $\frac{n}{N}$ получаем:

$$\begin{aligned}\frac{C_{N\theta}^{k} C_{N(1-\theta)}^{n-k}}{C_N^n} &= \frac{N\theta!}{k!(N\theta-k)!}\frac{N(1-\theta)!}{(n-k)!(N(1-\theta)-(n-k)!}\frac{n!(N-n)!}{N!} = \\ &= \frac{n!}{k!(n-k)!}\frac{(N\theta)!(N(1-\theta))!(N-n)!}{(N\theta-k)!(N(1-\theta)-(n-k))!N!} \approx \\ &\approx \frac{n!}{k!(n-k)!}\frac{(N\theta)^k(N(1-\theta))^{n-k}}{N^n} = C_n^k\theta^k(1-\theta)^{n-k}.\end{aligned}$$

В соответствии с теоремой Хейли из последовательности $\{\Pi_N(\theta)\}$ можно выбрать сходящуюся подпоследовательность.

Таким образом, переходя к пределу по этой сходящейся подпоследовательности, получаем:

$$p(x_1,\ldots,x_n) = \int_0^1 \theta^{y_n}(1-\theta)^{n-y_n}d\Pi(\theta),$$

причем $\Pi(\theta) = \lim_{n\to\infty} p\left(\frac{\sum_{i=1}^n x_i}{n} \le \theta\right)$.

□

Приведем теперь формулировку теоремы Де Финетти в более общем виде.

Теорема 2. *Пусть x_i составляют бесконечную перестановочную последовательность с вероятностной мерой P. Тогда совместное распределение $p(x_1 = y_1,\ldots,x_n = y_n)$ можно представить в виде:*

$$p(x_1 = y_1,\ldots,x_n = y_n) = \int_{\mathcal{F}} \prod_{i=1}^n F(y_i)d\Pi(\theta),$$

где F — неизвестная или ненаблюдаемая функция распределения, и

$$\Pi(\theta) = \lim_{n\to\infty} P_n(\widehat{F}_n)$$

— вероятностная мера на пространстве функций $\mathcal{F}$, определенная как предел при $n \to \infty$ на эмпирической функции распределения $\widehat{F}_n$.

3.4 Выводы

Таким образом, в Байесовской статистике действительно есть что изучать: с одной стороны во многих случаях Байесовский подход кажется осмысленным, с другой — интуитивных идей недостаточно для построения стройной теории.

4 Асимптотическая нормальность апостериорного распределения

В классической статистике важным является установить для оценки ее асимптотическое поведение. Большинство используемых оценок — регулярные, для них можно установить асимптотическую нормальность.

Для Байесовских оценок условие сходимости апостериорного распределения к нормальному определяют теорема Бернштейна-фон Мизеса и в более общем смысле условия Ибрагимов и Хасьминского. Оба результата представлены в этом разделе.

4.1 Теорема Бернштейна-фон Мизеса

Важной проблемой Байесовской статистики с точки зрения обычной математической статистики является несоответствие между Байесовскими оценками и эффективными классическими оценкаvb.

Оказывается, что в асимптотике Байесовские оценки часто совпадают с классическими. Формальное утверждение про близость Байесовских и классических оценок составляет теорема суть теоремы Бернштейна-фон Мизеса.

Рассмотрим выборку независимых одинаково распределенных величин $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ из распределения с плотностью $f(\mathbf{x}|\theta)$. Параметр $\theta \in \Theta$, Θ — открытое подмножество $\mathbb{R}$. Будем предполагать следующие условия регулярности:

- A1 Носитель $f(\mathbf{x}|\theta)$ одинаков для всех $\theta \in \Theta$
- A2 Логарифм правдоподобия $l(\mathbf{x}|\theta) = \log p(\mathbf{x}|\theta)$ трижды непрерывно дифференцируем по θ в окрестности истинного значения $(\theta_0 - \delta, \theta_0 + \delta)$. Обозначим $\dot{l}(\mathbf{x}|\theta)$, $\ddot{l}(\mathbf{x}|\theta)$, $\dddot{l}(\mathbf{x}|\theta)$ первую, вторую и третью частные производные по параметру правдоподобия. Пусть математические ожидания $\mathbb{E}_{\theta_0}\dot{l}(\mathbf{x}|\theta)$, $\mathbb{E}_{\theta_0}\ddot{l}(\mathbf{x}|\theta)$ конечны, и
$$\sup_{\theta \in (\theta_0 - \delta, \theta_0 + \delta)} |\dddot{l}(\mathbf{x}|\theta)| < M(\mathbf{x}),$$
причем $\mathbb{E}_{\theta_0} M(\mathbf{x}) < \infty$.
- A3 Можно менять местами математическое ожидание по θ_0и дифференцирование по θ_0, так что
$$\mathbb{E}_{\theta_0}\dot{l}(\mathbf{x}|\theta_0) = 0$$
$$\mathbb{E}_{\theta_0}\ddot{l}(\mathbf{x}|\theta_0) = -\mathbb{E}_{\theta_0}(\dot{l}(\mathbf{x}|\theta))^2.$$
- A4 Информация Фишера $I(\theta_0)^2 = \mathbb{E}_{\theta_0}(\dot{l}(\mathbf{x}|\theta_0))^2 > 0$.

В таких предположениях состоятельная оценка максимума правдоподобия будет асимптотически нормальной.

Теорема 3. *Пусть для семейства плотностей $\{f(\mathbf{x}|\theta), \theta \in \Theta\}$ выполнены предположения [A1]-[A4] и оценка максимума правдоподобия $\widehat{\theta}_n$ состоятельна, то*

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \to^D \mathcal{N}\left(0, \frac{1}{I(\theta_0)}\right).$$

Доказательство. Утверждение теоремы следует из центральной предельной теоремы и усиленного закона больших чисел.

Обозначим $l_n(\theta) = \sum_{i=1}^n l(\mathbf{x}_i|\theta)$, а ее первую, вторую и третью производную по θ — $\dot{l}_n(\theta), \ddot{l}_n(\theta)$ и $\dddot{l}_n(\theta)$ соответственно. Разложим производную $l_n(\theta)$ по Тейлору:

$$0 = \dot{l}_n(\widehat{\theta}_n) = \dot{l}_n(\theta_0) + (\widehat{\theta}_n - \theta_0)\ddot{l}_n(\theta_0) + \frac{1}{2}(\widehat{\theta}_n - \theta_0)^2 \dddot{l}_n(\theta'),$$

где $\theta_0 \leq \theta' \leq \widehat{\theta}_n$. Тогда

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \frac{\frac{1}{\sqrt{n}}\dot{l}_n(\theta_0)}{-\frac{1}{n}\ddot{l}_n(\theta_0) - \frac{1}{2}\frac{1}{n}\dddot{l}_n(\theta')}.$$

Так как выполнена центральная предельная теорема, то числитель сходится по распределению к $\mathcal{N}(0, I(\theta_0))$. Первое слагаемое в знаменателе сходится к $I(\theta_0)$ по усиленному закону больших чисел. Второе слагаемое мало в силу состоятельности $\widehat{\theta}_n$ и ограниченности $|\dddot{l}(\mathbf{x}|\theta)|$. Следовательно, левая часть равенства сходится по распределению к $\mathcal{N}(0, 1/I(\theta_0))$. □

Для теоремы Бернштейна-фон Мизеса понадобятся дополнительные предположения:

A5 Для произвольного $\delta > 0$ существует $\varepsilon > 0$такое, что

$$P_{\theta_0}\left\{\sup_{|\theta-\theta_0|>\delta} \frac{1}{n}\left(l_n(\theta) - l_n(\theta_0)\right) \leq -\varepsilon\right\} \to 1.$$

A6 Априорная плотность распределения $\pi(\theta)$ непрерывна и положительна θ_0.

Теорема 4 (Теорема Бернштейна-фон Мизеса)**.** *Пусть выполнены предположения [A1]-[A6], и $\widehat{\theta}_n$ — состоятельная оценка максимума правдоподобия. Обозначим совместную плотность выборки $f(X|\theta)$. Тогда для $n \to \infty$:*

$$\int_{\mathbb{R}} \left| p(s|X) - \frac{1}{\sqrt{2\pi}\sqrt{I(\theta_0)^{-1}}} \exp\left(-\frac{1}{2I(\theta_0)^{-1}} s^2\right) \right| ds \to^p 0,$$

где $s = \sqrt{n}(\theta - \widehat{\theta}_n(X))$.

Доказательство этого результата — техническое: нужно разбить область интегрирования на три и в каждой из областей оценить интеграл сверху.

Эта теорема утверждает, что для выбранного априорного распределения мы можем точно описать апостериорное распределение.

Как следствие этой теоремы мы получаем асимптотическую нормальность и сходимость для Байесовской оценки.

Теорема 5. *Пусть* $\int_\Theta |\theta|\pi(\theta)d\theta < \infty$. *Будем использовать в качестве Байесовской оценки апостериорное среднее:*

$$\theta_n^* = \int_\Theta \theta p(\theta|X)d\theta.$$

Тогда

$$\sqrt{n}(\widehat{\theta}_n - \theta_n^*) \to^{p_{\theta_0}} 0.$$

Кроме того,

$$\sqrt{n}(\theta_n^* - \widehat{\theta}_n) \to^D \mathcal{N}\left(0, \frac{1}{I(\theta_0)}\right).$$

Существуют вариации представленных результатов, полученные в других предположениях о регулярности семейства. Например, получена версия теоремы Бернштейна-фон Мизеса при нарушении параметрического предположения и для конечных выборок.

4.2 Условия Ибрагимова и Хасьминского

Ибрагимов и Хасьминский предложили ряд условий для целого семейства параметрических моделей. Эти условия были проверены для различных классов нерегулярных задач и случайных процессов. Рассмотрим теперь результаты, которые получаются с использованием этих условий.

Множество значений параметров Θ является подмножеством пространства $\mathbb{R}^p$. Для упрощения изложения рассмотрим $p = 1$. Совместное распределение выборки $\mathrm{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ обозначим P_θ^n, а плотность относительно сигма-конечной меры обозначим $p(\mathrm{D}, \theta)$. Последовательность положительных констант ϕ_n сходится к 0 при $n \to \infty$. В регулярном случае, рассмотренном в предыдущем разделе, можно взять $\phi_n = \frac{1}{\sqrt{n}}$. В нерегулярном случае, как правило, сходимость $\phi_n \to 0$ может быть быстрее. Рассмотрим отображение U, определенное как $U(\theta) = \frac{1}{\phi_n}(\theta - \theta_0)$, где θ_0 — истинное значение параметра. Пусть $\mathcal{U}_n = \{U(\theta) : \theta \in \Theta\}$. Величина u является соответствующим образом масштабированной разностью между θ и θ_0. Зададим случайный процесс

$$Z_n(u, \mathrm{D}_n) = \frac{p(\mathrm{D}_n, \theta_0 + \phi_n u)}{p(\mathrm{D}_n, \theta_0)}.$$

Условия Ибрагимова-Хасьминского имеют вид:

ИХ1 Для некоторых $M > 0$, $m_1 \geq 0, \alpha > 0, n_0 \geq 1$ выполнено, что

$$\mathbb{E}_{\theta_0} \| Z_n^{\frac{1}{2}}(u_1) - Z_n^{\frac{1}{2}}(u_2)\|^2 \leq M(1 + A^{m_1})|u_1 - u_2|^\alpha, \\ \forall u_1, u_2 \in \mathcal{U}_n \text{ with } |u_1| \leq A, |u_2| \leq A$$

для всех $n \geq n_0$.

ИХ2 Для всех $u \in \mathcal{U}_n$ и $n \geq n_0$

$$\mathbb{E}_{\theta_0} \| Z_n^{\frac{1}{2}}(u)\| \leq \exp(-g_n(|u|)),$$

где g_n — последовательность действительнозначных функций, удовлетворяющих следующим условиям:

– для любого $n \geq 1$, $g_n(y) \uparrow \infty$ для $y \to \infty$,

ИХ3 для любого $N > 0$

$$\lim_{y \to \infty, n \to \infty} y^N \exp(-g_n(y)) = 0.$$

- Конечномерные распределения $\{Z_n(u) : u \in \mathcal{U}_n\}$ сходятся к конечномерным распределениям случайного процесса $\{Z(u) : u \in \mathbb{R}\}$.

Теорема 6. *Пусть Π — априорное распределение с положительной непрерывной плотностью в θ_0. Тогда если выполнены условия Ибрагимова-Хасьминского [ИХ1–ИХ3] для квадратичной функции потерь, нормализованная Байесовская оценка $\phi_n(\widetilde{\theta}_n - \theta_0)$ сходится по распределению к $\int uZ(u)du / \int Z(u)du$.*

Предложение 1. *Предположим, что $\mathbf{x}_1, \ldots, \mathbf{x}_n$ — независимые одинаково распределенные случайные величины, и Π — априорное распределение. Пусть $\widehat{\theta}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ — симметричная функция по $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Обозначим*

$$t = \phi_n^{-1}(\theta - \widehat{\theta}(\mathrm{D}_n)),$$

и A — борелевское множество. Пусть

$$\Pi(t \in A | \mathrm{D}_n) \to^{P_{\theta_0}} Y_A.$$

Тогда Y_A — константа почти всюду на P_{θ_0}.

Определение 1. *Для некоторой симметрической функции $\widehat{\theta}(\mathrm{D}_n)$ апостериорное распределение $t = \phi_n^{-1}\left(\theta - \widehat{\theta}(\mathrm{D}_n)\right)$ сходится к Q, если*

$$\sup_A \{\Pi(t \in A | \mathrm{D}_n) - Q(A)\} \to^{P_{\theta_0}} 0.$$

Тогда $\widehat{\theta}(\mathrm{D}_n)$ называют точным центрированием.

Теорема 7. *Пусть выполнены условия Ибрагимова-Хасьминского и Π — априорное распределение с непрерывной положительной плотностью в θ_0. Если точное центрирование $\widehat{\theta}(\mathrm{D}_n)$ существует, тогда существует случайная величина W, такая, что*

a) $\phi_n^{-1}(\theta_0 - \widehat{\theta}(\mathbf{x}_1, \ldots, \mathbf{x}_n))$ *сходится по распределению к* W*.*

b) Для почти всех $\eta \in \mathbb{R}$ *величина* $\xi(\eta - W) = q(\eta)$ *является неслучайной. Здесь* $\xi(u) = Z(u)/\int_{\mathbb{R}} Z(u)du, u \in \mathbb{R}$.

Если b) выполнено для некоторой случайной величины W*, то апостериорное среднее для заданной выборки* D_n *является точным центрированием с* $Q(A) = \int_A q(t)dt$*.*

5 Байесовская теория принятия решений

5.1 Задача выбора решающего правила

Будем рассматривать задачу статистического оценивания на основе выборки данных: заданы параметр $\theta \in \Theta$, определяющий распределение данных, X — наблюдения, на основе которых нужно принять решение $\delta(X)$ и риск $l(\theta, \delta(X))$, который штрафует за решение $\delta(X)$ при заданном параметре θ. Необходимо найти такое решающее правило $\delta(X)$, которое будет минимизировать риск $l(\theta, \delta(X))$.

Пример 5.1. Рассмотрим следующий естественный пример. Пусть задача состоит в оценке параметра θ, то есть $\delta(X) = \widehat{\theta}(X)$, а риск — квадратичный,

$$l(\theta, \delta(X)) = (\theta - \delta(X))^2.$$

Отметим, что мы еще не до конца сформировали нашу задачу, так как природа модели у нас вероятностная, и $l(\theta, \delta(X))$ — случайная величина.

5.2 Выбор решающего правила с использованием среднего риска

В классическом подходе к статистическому оцениванию обычно используют вероятстный или средний риск:

$$R(\theta, \delta) = \mathbb{E}_\theta l(\theta, \delta(X)) = \int l(\theta, \delta(X)) p(X|\theta) dX,$$

то есть мы усредняем риск по всем выборкам X, сгенерированным из распределения $p(X|\theta)$.

Теперь мы получим детерминированный — при заданном θ — средний риск. Однако мы все еще не можем однозначно сравнить два решающих правила $R(\theta, \delta_1)$ и $R(\theta, \delta_2)$. Чтобы понять в чем проблема, достаточно посмотреть на рисунок 1: в большинстве случаев нельзя сказать, какое решение равномерно лучше другого решения. В примере на рисунке видно, что для всех θ $R(\theta, \delta_1) \le R(\theta, \delta_2)$, однако нельзя так же сравнить решающие правила δ_1 и δ_3: для каких-то θ лучше будет использовать δ_1, а для каких-то — наоборот.

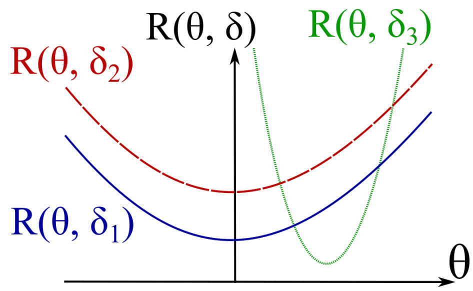

Рис. 1: Сравнение средних рисков для решающих правил δ_1, δ_2, δ_3

Перечислим подходы, которые используются для сравнения решающих правил с использованием среднего риска в классической математической статистике:

- Решающее правило δ будет *эффективным*, если для любого другого решающего правила δ' выполнено, что
$$\forall \theta \in \Theta : R(\theta, \delta) \le R(\theta, \delta').$$
Как мы увидели выше, таких решающие правила если и встречаются, то в очень ограниченном классе задач. Однако, можно говорить не про эффективность в целом, а про эффективность в более узком смысле.

- Естественно *ограничить класс решающих правил*, в котором мы будем искать целевое решающее правило. Например, в задаче оценки параметра мы можем искать только решающие правила, которые дают несмещенную оценку: $\mathbb{E}_\theta \widehat{\theta}(X) = \theta$. В классе несмещенных оценок для некоторых классов статистических моделей эффективные оценки существуют. В частности, существует классическая вероятностная теория про эффективность оценок достаточных статистик в экспоненциальном классе распределений.

- Решающее правило δ будет *минимаксно эффективным*, если
$$\forall \delta' \neq \delta : \sup_\theta R(\theta, \delta) \le \sup_\theta R(\theta, \delta').$$
Такой подход сводит сравнение средних рисков к сравнению чисел, которые агрегируют информацию про средние риски. Однако, минимаксный подход кажется излишне консервативным в большинстве случаев. Нас редко интересует то, насколько хорошо все работает в худшем случае, обычно мы хотим оценить качество работы решающего правила в среднем.

5.3 Байесовская теория принятия решений

Другая естественная идея, возникающая в теории принятия решений, — взвесить средний риск с помощью некоторой функции $\pi(\theta)$. Тогда мы опять сведем задачу сравнения двух решающих правил к задаче сравнения двух чисел $\int R(\theta,\delta)\pi(\theta)d\theta$.

Естественный кандидат на роль такой функции — априорное распределение на θ. Таким образом Байесовский подход может быть естественным образом использован в теории принятия решений.

Однако, посмотрим на нее с еще одной стороны. Введем апостериорный риск:

$$\rho(\pi,\delta(X)) = \int_\Theta l(\theta,\delta(X))p(\theta|X)d\theta.$$

Решением, которое минимизирует апостериорный риск будем называть Байесовским решением $\delta^*(X)$. Как и раньше, для выбора $\delta^*(X)$ нет необходимости уметь считать $p(X)$, достаточно уметь считать $p(x|\theta)\pi(\theta) \propto p(\theta|X)$.

Пример 5.2. Получим Байесовское решение для квадратичной функции потерь $l(\theta,\delta(X)) = (\theta-\delta(X))^2$. Апостериорный риск имеет вид:

$$\rho(\pi,\delta(X)) = \int_\Theta (\theta-\delta(X))^2 p(\theta|X)d\theta = \delta(X)^2 - 2\delta(X)\int \theta p(\theta|X)d\theta + \int \theta^2 p(\theta|X)d\theta.$$

Последний член от $\delta(X)$ не зависит. Дифференцируя разность первых двух по $\delta(X)$, получаем необходимое условие локального экстремума:

$$\frac{\partial\rho(\pi,\delta(X))}{\partial\delta(X)} = 2\delta(X) - 2\int \theta p(\theta|X)d\theta = 0.$$

Следовательно, Байесовское решение имеет вид:

$$\delta^*(X) = \int \theta p(\theta|X)d\theta,$$

то есть оно совпадает с апостериорным средним. Если мы возьмем l_1 функцию потерь $l(\theta,\delta(X)) = \|\theta-\delta(X)\|$, то получим, что Байесовское решающее правило — медиана апостериорного распределения.

Определим теперь Байесовское решающее правило, как функцию $\delta_\pi(X)$, которая минимизирует

$$r(\pi,\delta) = \int R(\theta,\delta)\pi(\theta)d\theta,$$

где $R(\theta,\delta)$ — средний риск. Таким образом мы усреднили средний риск по априорному распределению θ.

Назовем $r(\pi) = r(\pi, \delta_\pi)$. Мы можем интерпертировать его и с более Байесовской точки зрения. Действительно,

$$\begin{aligned} r(\pi, \delta) &= \int\int l(\theta, \delta(X))p(X|\theta)dx\,\pi(\theta)d\theta = \\ &= \int\int l(\theta, \delta(X))p(\theta|X)d\theta p(X)dX = \\ &= \int \rho(X, \pi)p(X)dX. \end{aligned}$$

Таким образом, $r(\pi, \delta)$ — усреднение апостериорного риска по маргинальному распределению $p(X)$. Таким образом, Байесовское решающее правило — совокупность Байесовских решений для всех X.

Отметим, что такой подход может быть использован для получения минимаксных оценок, так как во многих случаях мы можем получить Байесовское решающее правило в явном виде.

5.4 Проблемы несмещенных оценок

Приведем в завершении этого раздела несколько классических примеров задачи статистического оценивания, демонстрирующих неэффективность несмещенных оценок.

Пример 5.3 (Регрессия к среднему)**.** Рассмотрим следующий пример. Пусть x — рост матери, а y — рост дочери. x и y — случайные величины из многомерного нормального распределения:

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho \\ \rho & \sigma^2 \end{pmatrix}\right).$$

Возьмем $\mu_1 = \mu_2 = 160$см, $\mathbb{V}x = \mathbb{V}y = \sigma^2 = 1$ и $\rho = 0.5$.

Тогда

$$\mathbb{E}(y|x) = \mu_2 + \rho(x - \mu_1)$$

Следовательно,

$$\begin{aligned} \mathbb{E}_y\mathbb{E}(y|x) &= 160 + 0.5(\mathbb{E}_y x - 160) = \\ &= 160 + 0.5(160 + 0.5(y - 160) - 160) = \\ &= 160 + 0.25(y - 160) \end{aligned}$$

Ясно, что такая оценка не совпадает с y и, более того, не является несмещенной. В то же время математическое ожидание для несмещенной оценки будет иметь вид:

$$\widehat{y} = 160 + 2(x - 160).$$

Такая несмещенная оценка противоречит здравому смыслу — получается, что дочка должна в среднем сильнее отклоняться от среднего роста, чем мать. На практике наблюдается обратная ситуация, которую описал еще пионер математической статистики и автор термина регрессия Фрэнсис Гальтон: обычно дети ближе к среднему росту, если рост их родителей аномально высокий. Название феномена *регрессия к среднему* и привело к появлению термина регрессия.

Пример 5.4 (Два орла подряд)**.** Пусть мы подбросили монету n раз, причем число орлов распределено биномиально $B(n, \theta)$. Мы наблюдаем r орлов и хотим оценить θ^2, вероятность наблюдения двух орлов подряд. В таком случае мы можем получить эффективную несмещенную оценку — несмещенную оценку с минимальной дисперсией:

$$\widehat{\theta^2} = \frac{r(r-1)}{n(n-1)}.$$

Получается, что для $r = 1$ такая оценка равна нулю. То есть, вероятность получить два орла подряд равна нулю. Существует ли оценка, свободная от этого недостатка, и как ее получить?

Пример 5.5 (Парадокс Штайна)**.** Рассмотрим $\mathbf{x} = \{x_1, \dots, x_p\}$. Каждый $x_i \sim \mathcal{N}(\theta_i, \sigma^2)$. Задача состоит в оценке вектора параметров $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_p\}$. Функция потерь квадратичная, $l(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}(\mathbf{x})) = \mathbb{E}\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|^2$.

Естественная несмещенная оценка в данном случае совпадает с оценкой максимума правдподобия, $\widehat{\boldsymbol{\theta}}_{MLE} = \mathbf{x}$.

Однако, оказывается, что такая оценка не является эффективной. Рассмотрим, например, оценку Джеймса-Штайна:

$$\widehat{\boldsymbol{\theta}}_{JS} = \left(1 - \frac{(p-2)\sigma^2}{\|\mathbf{x}\|^2}\right)\mathbf{x}.$$

Для $p \geq 3$ такая оценка оказывается эффективнее, чем оценка максимума правдоподобия. Однако и она не будет самой эффективной. Оценка Джеймса-Штайна с параметром $\boldsymbol{\nu}$ оказывается еще более эффективной для $p \geq 4$:

$$\widehat{\boldsymbol{\theta}}_{JS\boldsymbol{\nu}} = \left(1 - \frac{(p-3)\sigma^2}{\|\mathbf{x} - \boldsymbol{\nu}\|^2}\right)(\mathbf{x} - \boldsymbol{\nu}) + \boldsymbol{\nu}.$$

В частности, она более эффективна, чем Байесовская оценка

$$\widehat{\boldsymbol{\theta}}_{JS\boldsymbol{\nu}+} = \left(1 - \frac{(p-3)\sigma^2}{\|\mathbf{x} - \boldsymbol{\nu}\|^2}\right)^+ (\mathbf{x} - \boldsymbol{\nu}) + \boldsymbol{\nu}.$$

6 Сопряженное априорное распределение

6.1 Определение сопряженного априорного распределения

Одним из наиболее широко используемых понятий в Байесовской статистике является понятие сопряженного априорного распределения.

Определение 2. *Пусть задана статистическая модель данных с правдоподобием данных $p(X|\boldsymbol{\theta})$. Тогда семейство априорных распределений называется* сопряженным семейством априорных распределений*, если при выборе априорного распределения из этого семейства, апостериорное распределение тоже будет ему принадлежать.*

Часто для краткости мы будем говорить не о сопряженном семействе априорных распределений, а просто о сопряженном априорном распределении.

Пример 6.1. Тривиальный пример сопряженного семейства априорных распределений — семейство всех вероятностных распределений.

Если смотреть на этот пример, не очень понятно, зачем нужно такое определение. Однако во многих случаях определение такого семейства для заданной вероятностной модели оказывается полезным.

6.2 Сопряженное распределение для мультиномиального распределения

Рассмотрим выборку размера n из мультиномиального распределения. Пускай в этом распределении k различных категорий, обозначим их $\{1, 2, \ldots, k\}$. Обозначим x_i количество наблюдений i-ой категории, а θ_i — вероятность того, что мы наблюдаем событие из i-ой категории. Тогда правдоподобие для наблюдений $\mathbf{x} = \{x_1, \ldots, x_k\}$ и вектора параметров $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_k\}$имеет вид:

$$p(\mathbf{x}|\boldsymbol{\theta}) = C_n^{x_1,\ldots,x_k} \theta_1^{x_1} \cdot \ldots \cdot \theta_k^{x_k},$$

где нормировочный коэффициент для распределения $C_n^{x_1,\ldots,x_k}$ имеет вид:

$$C_n^{x_1,\ldots,x_k} = \frac{x_1! \ldots x_k!}{n!}.$$

Пускай априорное распределение — распределение Дирихле с вектором параметров $\boldsymbol{\alpha} \in \mathbb{R}_+^k (\alpha_i \geq 0)$:

$$\pi(\boldsymbol{\theta}|\boldsymbol{\alpha}) \propto \theta_1^{\alpha_1 - 1} \cdot \ldots \cdot \theta_k^{\alpha_k - 1}.$$

Тогда апостериорное распределение тоже будет распределением Дирихле с вектором параметров $\mathbf{x} + \boldsymbol{\alpha}$:

$$p(\boldsymbol{\theta}|\mathbf{x}, \boldsymbol{\alpha}) \propto \theta_1^{x_1 + \alpha_1 - 1} \cdot \ldots \cdot \theta_k^{x_k + \alpha_k - 1}.$$

Так как распределение Дирихле для категориальных случайных величин примерно то же самое, что и нормальное распределение для непрерывных случайных величин: для него все можно посчитать аналитически, и, кроме того, у него множество других полезных свойств, то его использование в этом случае в качестве сопряженного распределения представляется крайне полезным.

6.3 Сопряженное распределение для экспоненциального семейства распределений

Экспоненциальное семейство распределений и его свойства описаны в разделе А.3.

Если $x_1, \dots, x_n$ — независимые одинаково распределенные случайные величины из одного и того же распределения из экспоненциального семейства, то

$$p(X|\boldsymbol{\theta}) = \prod_{j=1}^{n} h(\mathbf{x}_j) \exp\left(\boldsymbol{\theta}^\top \sum_{j=1}^{n} T(\mathbf{x}_j) - nA(\boldsymbol{\theta})\right).$$

Определим сопряженное априорное распределение для экспоненциального семейства как

$$\pi(\boldsymbol{\theta}|\boldsymbol{\tau}, n_0) = H(\tau, n_0) \exp(\boldsymbol{\tau}^\top \boldsymbol{\theta} - n_0 A(\boldsymbol{\theta})).$$

Это распределение тоже лежит в экспоненциальном семействе. Апостериорное распределение будет иметь такой же вид, но с параметрами

$$\boldsymbol{\tau}' = \boldsymbol{\tau} + \sum_{j=1}^{n} T(\mathbf{x}_j), n_0' = n + n_0.$$

В таблице приведены пары априорное распределение-правдоподобие.

Определим теперь $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}) = \mathbb{E}[T(\mathbf{x})|\boldsymbol{\theta}]$. Из общей теории экспоненциального семейства распределений следует, что $\boldsymbol{\mu} = \nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta})$. Найдем среднее значение $\boldsymbol{\mu}$ при фиксированном априорном распределении с параметрами $\boldsymbol{\tau}$ и n_0.

Заметим сперва, что

$$\mathbb{E}[\boldsymbol{\mu}|\boldsymbol{\tau}, n_0] = \mathbb{E}[\nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta})|\boldsymbol{\tau}, n_0].$$

С помощью прямых вычислений получим, что

$$\nabla_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta}|\boldsymbol{\tau}, n_0) = \pi(\boldsymbol{\theta}|\boldsymbol{\tau}, n_0)(\boldsymbol{\tau} - n_0 \nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta})).$$

Так как $\pi(\boldsymbol{\theta}|\boldsymbol{\tau}, n_0)$ — плотность вероятностного распределения, то она обращается в ноль в бесконечности. Тогда согласно теореме Грина:

$$\int \pi(\boldsymbol{\theta}|\boldsymbol{\tau}, n_0)(\boldsymbol{\tau} - n_0 \nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta})) d\boldsymbol{\theta} = \int_{\mathbb{R}^p} \nabla_{\boldsymbol{\theta}} \pi(\boldsymbol{\theta}|\boldsymbol{\tau}, n_0) d\boldsymbol{\theta} = 0.$$

Следовательно,

$$\mathbb{E}[\boldsymbol{\mu}|\boldsymbol{\tau}, n_0] = \mathbb{E}[\nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta})|\boldsymbol{\tau}, n_0] = \frac{\boldsymbol{\tau}}{n_0}.$$

Аналогично для апостериорного распределения:

$$\mathbb{E}[\boldsymbol{\mu}|\boldsymbol{\tau}, n_0] = \frac{\boldsymbol{\tau} + \sum_{j=1}^{n} T(\mathbf{x}_j)}{n + n_0} = \varkappa \frac{\boldsymbol{\tau}}{n_0} + (1 - \varkappa) \frac{\sum_{j=1}^{n} T(\mathbf{x}_j)}{n},$$

где $\varkappa = \frac{n_0}{n_0+n}$. Таким образом, апостериорное среднее — линейная комбинация априорного среднего и среднего достаточных статистик.

Отметим, что при естественных предположениях выполнено и обратно утверждение: если апостериорное среднее всегда выпуклая комбинация оценки максимума правдоподобия и априорного среднего, то мы работаем с распределениями из экспоненциального семейства.

7 Объективное априорное распределение

Другим популярным подходом к выбору априорного распределения — помимо прагматичного способа, описанного ранее — является объективный подход. В этом подходе наша цель — выбрать априорное распределение, которое бы больше всего соответствовало отсутствия каких-либо априорных знаний, неинформативное априорное распределение. Рассмотрим два естественных примера такого распределения.

Пример 7.1 (Априорное распределение для параметра сдвига)**.** Пускай мы хотим выбрать целевое распределение данных из семейства $f(x-\theta)$. Естественно предположить, что все параметры сдвига равновероятны — и мы не можем отдать предпочтение какому-нибудь в нашем неинформативном априорном распределении.

Тогда логично использовать равномерное априорное распределение с плотностью $\pi(\theta) \sim 1$. Если множество Θ допустимых значений θ ограничено, то мы получаем корректное априорное распределение, которое соответствует нашим требованиям.

Если множество Θ неограничено, например, $\Theta = \mathbb{R}$, то если мы возьмем плотность $\pi(\theta) \sim 1, \theta \in \Theta$, которая не будет отвечать никакому вероятностному распределению, так как интеграл $\int_\Theta \pi(\theta) d\theta$ не будет конечным. Такое априорное распределение называется *некорректным априорным распределением.*

Однако, в некоторых случаях некорректное априорное распределение оказывается полезным. Часто для некорректного априорного распределения апостериорное распределение оказывается корректным. В частности, подход к выбору θ на основе максимизации правдподобия соотвествует Байесовскому подходу с равномерным на $\mathbb{R}$ априорным распределением.

Так же можно определить такое некорректное априорное распределение как предел последовательности корректных априорных распределений. Тем самым получится математически строго работать с объектами Байесовского статистики.

Пример 7.2 (Априорное распределение для параметра масштаба)**.** Введем равномерное распределение для параметра масштаба. Параметр масштаба — такой параметр плотности, что для $\theta \in \Theta \subseteq \mathbb{R}^+$:

$$f_\theta(x) = \frac{1}{\theta} f^0\left(\frac{x}{\theta}\right),$$

отношение $\frac{1}{\theta}$ здесь нужно для нормализации распределения. Тогда если мы захотим потребовать от априорного распределения инвариантности к масштабу, то для любого $c > 0$ должно быть выполнено, что

$$\pi(\theta) = \frac{1}{c}\pi\left(\frac{\theta}{c}\right).$$

У такого функционального уравнения существует единственное с точностью до масштабирующего коэффициента решение:

$$\pi(\theta) \sim \frac{1}{\theta}.$$

Отметим, что мы — как и при выборе априорного распределения для параметра сдвига — получили некорректное неинформативное распределение для масштаба, что, на самом деле, часто не является проблемой.

Рассмотрим теперь $\rho = \log\theta$. Тогда

$$\pi(\rho) = \pi(\theta)\left|\frac{d\theta}{d\rho}\right| \sim e^{-\rho}e^{\rho} = 1.$$

То есть, неинформативное априорное распределение для такого преобразования параметра масштаба будет равномерным.

Можно получить такое априорное распределение и другим способом.

7.1 Априорное распределение Джеффриса

Мы хотели бы работать с априорным распределением, на которое не будет влиять параметризация θ. Оказывается, такое априорное распределение существует:

$$\pi_J(\theta) \sim I(\theta)^{\frac{1}{2}},$$

где $I(\theta)^{\frac{1}{2}}$ — информационная матрица Фишера или информация Фишера :

$$I(\theta) = -\mathbb{E}_\theta\left[\frac{d^2\log p(x|\theta)}{d\theta^2}\right].$$

Информационная матрица Фишера — важный объект в классической математической статистике. Легко видеть, что она локально вогнута в окрестности оценки максимума правдоподобия и глобально вогнута для экспоненциального семейства распределений.

Докажем теперь, что априорное распределение Джеффриса не зависит от параметризации.

Доказательство. Сперва докажем лемму

Лемма 1. *Если правдоподобие — регулярно (то есть, можно выносить дифференцирования по параметру за интеграл), то для математического ожидания производной логарифма правдоподобия по параметру выполнено:*

$$\mathbb{E}\left(\frac{d\log p(x|\theta)}{d\theta}\right) = 0.$$

Доказательство. Получаем результат теоремы воспользовавшись регулярностью правдоподобия и условием нормировки для вероятностного распределения $\int p(x|\theta)dx = 1$:

$$\begin{aligned}
\mathbb{E}\left(\frac{d\log p(x|\theta)}{d\theta}\right) &= \int \frac{d\log p(x|\theta)}{d\theta}p(x|\theta)dx = \\
&= \int \frac{dp(x|\theta)}{d\theta}\frac{1}{p(x|\theta)}p(x|\theta)dx = \\
&= \int \frac{dp(x|\theta)}{d\theta}dx = \frac{d}{d\theta}\int p(x|\theta)dx = \\
&= \frac{d}{d\theta}1 = 0.
\end{aligned}$$

□

Подсчитаем информацию Фишера для другой параметризации ϕ параметра θ:

$$\begin{aligned}
I(\phi) &= -\mathbb{E}\left[\frac{d^2\log p(x|\phi)}{d\phi^2}\right] = \\
&= -\mathbb{E}\left[\frac{d^2\log p(x|\theta)}{d\theta^2}\left(\frac{d\theta}{d\phi}\right)^2 + \frac{d\log p(x|\theta)}{d\theta}\frac{d^2\theta}{d\phi^2}\right] = \\
&= -\mathbb{E}\left[\frac{d^2\log p(x|\theta)}{d\theta^2}\left(\frac{d\theta}{d\phi}\right)^2\right] - \mathbb{E}\left[\frac{d\log p(x|\theta)}{d\theta}\frac{d^2\theta}{d\phi^2}\right] = \\
&= -\mathbb{E}\left[\frac{d^2\log p(x|\theta)}{d\theta^2}\right]\left(\frac{d\theta}{d\phi}\right)^2.
\end{aligned}$$

При преобразованиях мы воспользовались результатом Леммы 1 для того, чтобы избавиться от одного из слагаемых.

Следовательно,

$$I(\phi) = I(\theta)\left(\frac{d\theta}{d\phi}\right)^2.$$

Таким образом,

$$\sqrt{I(\phi)} = \sqrt{I(\theta)}\left|\frac{d\theta}{d\phi}\right|.$$

Плотность распределения при преобразовании случайной величины имеет вид:

$$\pi(\phi(\theta)) = \pi(\theta)\left|\frac{d\theta}{d\phi(\theta)}\right|$$

Получаем:

$$\pi_J(\theta) = \sqrt{I(\theta)},$$

что и требовалось доказать.

□

7.2 Примеры априорных распределений Джеффриса

Пример 7.3. Получим априорное распределение Джеффриса для среднего нормального распределения. Пусть $x \sim \mathcal{N}(\mu, \sigma^2)$, причем σ^2 известно. Тогда плотность распределения:

$$p(x|\mu) \sim \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

Дифференцируя логарифм правдоподобия два раза получаем:

$$\frac{d^2 \log p(x|\mu)}{d\mu^2} = -\frac{1}{\sigma^2}.$$

Таким образом, информация Фишера в таком случае не зависит от μ, и мы получаем равномерное априорное распределение Джеффриса:

$$\pi_J(\theta) \sim I(\theta)^{\frac{1}{2}} = \frac{1}{\sigma}.$$

Пример 7.4. Найдем априорное распределение Джеффриса для еще одной широко используемой модели. Пусть $x \sim \mathrm{Bin}(n, \theta)$ — биномиальная случайная величина с параметрами n и $0 \le \theta \le 1$. Тогда правдоподобие для $x \in \mathbb{N} \cup \{0\}$имеет вид:

$$p(x|\theta) = C_n^x \theta^x (1-\theta)^{n-x}.$$

Получим априорное распределение Джеффриса:

$$\log p(x|\theta) \propto x \log\theta + (n-x)\log(1-\theta).$$

Тогда

$$\frac{d \log p(x|\theta)}{d\theta} \propto \frac{x}{\theta} - \frac{n-x}{1-\theta}.$$

И

$$\frac{d^2 \log p(x|\theta)}{d\theta^2} \propto -\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2}.$$

Для биномиального распределения

$$\mathbb{E}_\theta x = n\theta.$$

Следовательно,

$$I(\theta) = -\mathbb{E}\left[\frac{d^2 \log p(x|\theta)}{d\theta^2}\right] = \frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2} = \frac{n}{\theta} + \frac{n}{1-\theta} = \frac{n}{\theta(1-\theta)}.$$

Следовательно,

$$\pi_J(\theta) = \sqrt{I(\theta)} \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}.$$

Априорным распределением Джеффриса для такой модели $\pi_J(\theta)$ будет бета-распределение с параметрами $(\frac{1}{2}, \frac{1}{2})$. Сравнение априорного распределения Джеффриса с равномерным распределением на $[0, 1]$ (то же самое, что и распределение $\beta(1, 1)$) приведено на рисунке 2.

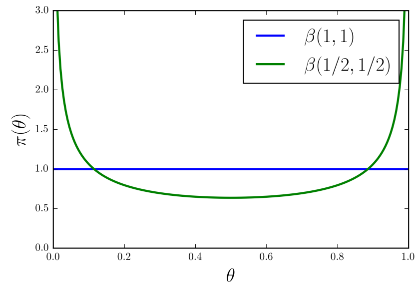

Рис. 2: Сравнение равномерного априорного распределения $\beta(1,1)$ и априорного распределения Джеффриса $\beta(\frac{1}{2},\frac{1}{2})$

7.3 Связь сопряженного априорного распределения и априорного распределения Джеффриса

Для примера с биномиальным распределением сопряженное априорное распределение и априорное распределение Джеффриса совпадают. Для нормального распределения это не так: для математического ожидания $\pi_J(\mu) \sim 1$, а $\pi_J(\sigma) \sim \frac{1}{\sigma}$, в то время сопряженное априорное распределение $\pi_C(\sigma)$ для σ — обратное гамма-распределение.

Однако, если для плотности обратного гамма-распределения с параметрами a, b:

$$\pi_{a,b}(\sigma) \propto \frac{1}{\sigma^{-(a+1)}} e^{\frac{-b}{\sigma}}$$

устремить a, b к нулю, то в пределе получим априорное распределение Джеффриса с плотностью, пропорциональной $\frac{1}{\sigma}$. Параметры a и b априорного распределения можно интерпретировать как количество наблюдений и меру концентрации параметра в области. Таким образом, устремляя эти два параметра к нулю, мы получаем априорное распределение, в котором нет «наблюдений» и параметр равномерно распределен по всему пространству.

Разумеется, модели в математической статистике не исчерпываются этими двумя примерами, и в общем случае сопряженное априорное распределение и априорное распределение Джеффриса могут быть никак не связаны.

7.4 Ограничения априорного распределения Джеффриса

Проблемы у такого подхода начинаются, когда размерность пространства параметров $p > 1$. По аналогии с одномерным случаем определим априорное распределение Джеффриса как

$$\pi_J(\boldsymbol{\theta}) = |I(\boldsymbol{\theta})|^{\frac{1}{2}}.$$

Тогда по определению

$$I(\boldsymbol{\theta})_{ij} = -\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2 \log p(X|\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right].$$

Пример 7.5. Пусть мы наблюдаем вектор $\mathbf{x}$ из многомерного нормального распределения:

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\theta}, I)$$

для $\mathbf{x} \in \mathbb{R}^p$. Задача состоит в оценке $\|\boldsymbol{\theta}\|^2$.

В таком случае априорное распределение Джеффриса будет равномерным. Апостериорное распределение в таком случае будет нецентральным χ^2 распределением с p степенями свободы. Апостериорное среднее

$$\mathbb{E}(\|\boldsymbol{\theta}\|^2|\mathbf{x}) = \|\mathbf{x}\|^2 + p.$$

Получается, что, используя априорное распределение Джеффриса, мы получаем результат, смещенный в большую сторону на p, в то время как обычно мы хотим получить в некотором роде регуляризирующую оценку. Например, математическое ожидание классической вероятностной оценки будет $\|\mathbf{x}\|^2 - p$.

Так же будет происходить и в многомерном случае. В силу того, что большая часть равномерного распределения находится на большом расстоянии от начала координат, Байесовские оценки часто будут смещены, причем в большую сторону.

Рассмотрим теперь двумерный пример.

Пример 7.6. Пусть $x \sim \mathcal{N}(\mu, \sigma^2)$, и пусть $\boldsymbol{\theta} = (\mu, \sigma^2)^\top$. Подсчитаем производные и получим информационную матрицу Фишера:

$$I(\boldsymbol{\theta}) = -\begin{pmatrix} -\frac{1}{\sigma^2} & \frac{2(x-\mu)}{\sigma^2} \\ \frac{2(x-\mu)}{\sigma^2} & \frac{3}{\sigma^4}(x-\mu)^2 - \frac{1}{\sigma^2} \end{pmatrix} =$$
$$= \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^2} \end{pmatrix},$$

так как $\mathbb{E}_{\boldsymbol{\theta}}(x-\mu) = 0$, $\mathbb{E}_{\boldsymbol{\theta}}(x-\mu)^2 = \sigma^2$. Следовательно, априорное распределение Джеффриса имеет вид:

$$\pi_J(\boldsymbol{\theta}) = |I(\boldsymbol{\theta})|^{\frac{1}{2}} \propto \frac{1}{\sigma^2}.$$

У такого априорного распределения ряд недостатков — например, низкая скорость сходимости. Сам Джеффрис предложил использовать априорное распределение $\pi_{J'}(\boldsymbol{\theta}) \propto \frac{1}{\sigma}$. Такое априорное распределение лучше с точки зрения естественных предположений и позволяет получить оценки, статистические свойства которых лучше. Оказывается, что $\pi_{J'}$ совпадает с опорным априорным распределением, про которое мы поговорим в следующей главе.

Получается, что у априорного распределение Джеффриса есть следующие недостатки:

- Равномерность априорного распределения может в некоторых случаях приводить к неразумным с точки зрения здравого смысла оценкам.
- Непонятно, как правильно обобщить его на многомерный случай.

Чтобы решить эти проблемы, было предложено использовать опорное априорное распределение. у него в меньшей степени проявляются недостатки перечисленные выше и, кроме того, оно позволяет посмотреть на задачу выбора неинформативного априорного распределения с точки зрения теории информации.

8 Опорное априорное распределение

8.1 Определение опорного априорного распределения

В конце предыдущей главы мы увидели, что априорное распределение Джеффриса не годится, если размерность пространства параметров $p > 1$. Существуют альтернативный подход к выбору неинформативного априорного распределения, который подходит и для больших размерность. В этой главе мы рассмотрим такой подход, априорные распределения, которые получаются в результате его использования, называются *опорными априорными распределениями*.

Пусть $X \sim p(x|\theta)$ и $T(X)$ — достаточная статистика для θ. Мы хотим, чтобы априорное распределение $\pi(\theta)$ и апостериорное распределение $p(\theta|t)$ для фиксированного значения достаточной статистики tбыли максимально далеки друг от друга — то есть, чтобы априорное распределение привносило в статистический вывод как можно меньше информации. Для этого будем максмизировать расстояние Кульбака-Лейблера, которое имеет вид:

$$\mathrm{KL}(p(\theta|t)|\pi(\theta)) = \int_{\theta\in\Theta} p(\theta|t)\log\frac{p(\theta|t)}{\pi(\theta)}d\theta.$$

Нам хотелось бы максимизировать такое расстояния не для одного фиксированного значения t, а по всем t — причем, взвесить разные t разумно, используя маргинальное распределение $p(t)$:

$$\mathrm{MI}_{\pi(\theta)}(\Theta, T) = \int p(t)\int p(\theta|t)\log\frac{p(\theta|t)}{\pi(\theta)}d\theta dt \to \max.$$

Такое усреднение $\mathrm{MI}(\Theta, T)$ называют взаимной информацией. В таких обозначениях искомое опорное априорное распределение будет решением следующей вариационной задачи:

$$\pi^*(\theta) = \arg\max_{\pi(\theta)} \mathrm{MI}_{\pi(\theta)}(\Theta, T).$$

Однако часто аналитическое решение такой вариационной задачи получить невозможно.

8.2 Вычисление опорного априорного распределения

Определение опорного априорного распределения, введенное выше, рассматривает статистику $T(x)$ как функцию одного наблюдения. Давайте вместо этого рассмотрим вектор $\mathbf{T}^k$, включающий значения статистик, полученные с помощью k независимых наблюдений из распределения $p(x|\theta)$.

Введем обозначения:

$$\mathrm{MI}_{\pi(\theta)}(\Theta, \mathbf{T}^k) = \int p(\mathbf{t}^k) \int p(\theta|\mathbf{t}^k) \log \frac{p(\theta|\mathbf{t}^k)}{\pi(\theta)} d\theta d\mathbf{t}^k,$$
$$\pi_k(\theta) = \arg\max_{\pi(\theta)} \mathrm{MI}_{\pi(\theta)}(\Theta, \mathbf{T}^k).$$

Затем мы получим неинформативное априорное распределение, устремив k к бесконечности:

$$\pi^*(\theta) = \lim_{k\to\infty} \pi_k(\theta).$$

Перепишем $\mathrm{MI}_{\pi(\theta)}(\Theta, \mathbf{T}^k)$ в виде:

$$\mathrm{MI}_{\pi(\theta)}(\Theta, \mathbf{T}^k) = \int \pi(\theta) \log \frac{f_k(\theta)}{\pi(\theta)} d\theta,$$

где

$$f_k(\theta) = \exp\left(\int p(\mathbf{t}^k|\theta) \log p(\theta|\mathbf{t}^k) d\mathbf{t}^k\right).$$

В решении этой вариационной задачи у нас есть дополнительное ограничение $\int \pi(\theta) d\theta = 1$. Следовательно, Лагранджиан имеет вид:

$$\pi_k(\theta) = \sup_{\pi(\theta)} \int \pi(\theta) \log \frac{f_k(\theta)}{\pi(\theta)} d\theta + \lambda\left(\int \pi(\theta) d\theta - 1\right).$$

Используя вариационное исчисление, мы получаем:

$$\pi_k^*(\theta) \propto f_k(\theta).$$

Не будем здесь приводить полного решения этой вариационной задачи. Получим только доказательство того, что

$$\pi_k^*(\theta) = f_k(\theta).$$

в дискретном случае.

Доказательство. Пусть T и θ — дискретны, тогда

$$\pi_k^*(\theta) = \operatorname*{argmax}_{\pi(\theta)} \sum_i \pi_i \frac{q_i}{\pi_i} + \lambda\left(\sum_i \pi_i - 1\right),$$

здесь π_i — вероятности $\pi(\theta_i)$, $q_i = f_k(\theta_i)$. Дифференцируя по π_i, получаем необходимое условие экстремума:

$$\frac{\partial}{\partial \pi_j}\left[\sum_i \pi_i \frac{q_i}{\pi_i} + \lambda\left(\sum_i \pi_i - 1\right)\right] = \log(q_j/\pi_j) + \pi_j(q_j/\pi_j)^{-1}(-q_j/\pi_j^2) + \lambda =$$
$$= -1 - \log \pi_j + \log q_j + \lambda = 0.$$

Следовательно,

$$\log \pi_j = \log q_j + \lambda - 1.$$

Таким образом,

$$\pi_j = q_j e^{\lambda - 1}.$$

Тогда

$$\pi = q,$$

что и требовалось доказать. □

Мы свели исходную задачу к задаче вычисления интеграла

$$f_k(\theta) = \exp\left(\int p(\mathbf{t}^k|\theta)\log p(\theta|\mathbf{t}^k)d\mathbf{t}^k\right).$$

К тому же нужно найти асимптотический предел для $k \to \infty$. Если мы устремим размер выборки к бесконечности, апостериорное распределение $p(\theta|\mathbf{t}^k)$ будет близко к нормальному, причем среднее этого нормального распределения будет соответствовать истинному значению оцениваемого параметра.

Формально близость апостериорного распределения к нормальному и сходство Байесовских и классических оценок параметров описывает теорема Бернштейна-фон Мизеса.

Теорема 8 (Теорема Бернштейна-фон Мизеса)**.** *Пусть для задачи статистического оценивания выполнен ряд условий регулярности: классическая эффективная оценка $\widetilde{\theta}_k$ асимптотически нормальна, и априорное распределение ведет себя достаточно регулярно, в частности в окрестности истинного значения параметра θ_0. Обозначим $\mathbf{t}^k$ вектор независимых t_j^k из распределения $p(t|\theta)$. Тогда*

$$\|p(\theta|\mathbf{t}^k) - \mathcal{N}(\widetilde{\theta}_k, I_k^{-1}(\theta_0))\| \to 0, \quad (1)$$

где $I_k(\theta_0)$ — информация Фишера для θ_0, сходимость понимается в смысле сходимости по вероятности, а $\|\cdot\|$ обозначает расстояние по вариации.

Докажем теперь, используя теорему Бернштейна-фон Мизеса, что опорное априорное распределение совпадает с априорным распределением Джеффриса в одномерном случае.

Доказательство. Любая асимптотически эффективная оценка $\widetilde{\theta}_k$ является асимптотически достаточной. Следовательно, мы можем заменить в (1) $\mathbf{t}^k$ на $\widetilde{\theta}_k$:

$$\|p(\theta|\widetilde{\theta}_k) - \mathcal{N}(\widetilde{\theta}_k, I_k^{-1}(\theta_0))\| \to 0.$$

Для $y \sim \mathcal{N}(\widetilde{\theta}_k, I_k^{-1}(\theta_0))$ плотность имеет вид:

$$p(y) = \sqrt{I_k(\theta_0)} \exp\left(-\frac{I_k(\theta_0)}{2}(y - \widetilde{\theta}_k)^2\right).$$

Для независимых наблюдений $I_k^{-1}(\theta_0) = \frac{1}{k} I^{-1}(\theta_0)$. Таким образом,

$$p(\theta|\widetilde{\theta}_k) \propto \sqrt{kI(\theta_0)} \exp\left(-\frac{kI(\theta_0)}{2}(y - \widetilde{\theta}_k)^2\right).$$

Оценка $\widetilde{\theta}_k$ — состоятельна, следовательно для больших k:

$$p(\theta|\widetilde{\theta}_k) \propto \sqrt{kI(\widetilde{\theta}_k)} \exp\left(-\frac{kI(\widetilde{\theta}_k)}{2}(y - \widetilde{\theta}_k)^2\right).$$

Далее будем действовать менее формально. Полное доказательство есть, например, в статье [1].

Рассмотрим $\theta = \theta_0$:

$$p(\theta_0|\widetilde{\theta}_k) \propto \sqrt{kI(\theta_0)} \exp\left(-\frac{kI(\widetilde{\theta}_k)}{2}(\theta_0 - \widetilde{\theta}_k)^2\right) \approx$$
$$\approx \sqrt{kI(\theta_0)} \exp\left(-\frac{kI(\widetilde{\theta}_k)}{2}(\theta_0 - \theta_0)^2\right) = \sqrt{kI(\theta_0)}.$$

Следовательно, искомый интеграл для больших k можно аппроксимировать:

$$f_k(\theta) \approx \exp\left(\int p(\mathbf{t}^k|\theta) \log \sqrt{I(\theta)} d\mathbf{t}^k\right).$$

$\sqrt{I(\theta)}$ не зависит от $\mathbf{t}^k$, а интеграл по вероятностной плотности — единица. Следовательно,

$$f_k(\theta) \approx \sqrt{I(\theta)}.$$

Получается, что опорное априорное распределение распределение совпадает с априорным распределением Джеффриса в одномерном случае. □

Пример 8.1 (Опорное априорное распределение для экспоненциального распределения)**.** Пусть $x_i \sim \mathrm{Exp}(\theta)$. Достаточная статистика для θ — среднее выборки $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$. Оценка максимума правдоподобия есть $\widehat{\theta} = \frac{1}{\overline{x}}$.

Для одномерного случая мы могли бы получить опорное априорное распределение из априорного распределения Джеффриса. Но давайте вместо этого воспользуемся подходом, описанном выше.

Пусть $\mathbf{x} = \{x_1, \ldots, x_n\}$. Тогда правдоподобие:

$$p(\mathbf{x}|\theta) = \theta^n \exp\left(-n\overline{x}\theta\right).$$

В соответствии с теоремой Бернштейна-фон Мизеса апостериорное распределение для $n \to \infty$ не зависит от априорного распределения, поэтому возьмем для удобства равномерное априорное распределение.

В силу концентрации для больших выборок для достаточной статистики $\widehat{\theta}$:

$$p(\widehat{\theta}|\theta) \approx \delta(\widehat{\theta} - \theta).$$

Следовательно,

$$f_k(\theta) \approx \exp\left[\log p(\theta|\widehat{\theta})\right] = p(\theta|\widehat{\theta}).$$

Получим апостериорное распределение по формуле Байеса:

$$p(\theta|\widehat{\theta}) = \frac{p(\widehat{\theta}|\theta)\pi(\theta)}{p(\widehat{\theta})}.$$

Априорное распределение — равномерное, правдоподобие

$$p(\widehat{\theta}|\theta) \propto \theta^n \exp\left(-n\frac{\theta}{\widehat{\theta}}\right),$$

а маргинальное распределение

$$p(\widehat{\theta}) = \int p(\widehat{\theta}|\theta)\pi(\theta)d\theta = \int \theta^n \exp\left(-n\frac{\theta}{\widehat{\theta}}\right)d\theta = \mathit{\Gamma}(n+1)\left(\frac{\widehat{\theta}}{n}\right)^{n+1}.$$

Подставляя эти выражения в исходную формулу, получаем:

$$\pi_n(\theta) = \left(\frac{n}{\widehat{\theta}}\right)^{n+1}\frac{1}{\mathit{\Gamma}(n+1)}\theta^n \exp\left(-\frac{n\theta}{\widehat{\theta}}\right)\Bigg|_{\widehat{\theta}=\theta} \propto \frac{1}{\theta}.$$

Здесь мы использовали $\widehat{\theta} = \theta$ в силу того, что выполнена теорема Бернштейна-фон Мизеса.

Проверим теперь, что опорное априорное распределение инвариантно к репараметризации — даже если не выполнены условия регулярности, при выполнении которых опорное априорное распределение совпадает с априорным распределением Джеффриса.

Доказательство. Нам нужно проверить, что взаимная информация не зависит от параметризации:

$$\begin{aligned}\mathrm{MI}_{\pi(\theta)}(\Theta, \mathbf{T}^k) &= \int p(\mathbf{t}^k)\int p(\theta|\mathbf{t}^k)\log\frac{p(\theta|\mathbf{t}^k)}{\pi(\theta)}d\theta d\mathbf{t}^k = \\ &= \int p(\mathbf{t}^k)\int p(\phi|\mathbf{t}^k)\log\frac{p(\phi|\mathbf{t}^k)}{\pi(\phi)}d\phi d\mathbf{t}^k\end{aligned}$$

При использовании другой параметризации плотность распределения нужно домножить на Якобиан:

$$p(\phi) = p(\theta(\phi)) \left| \frac{d\theta}{d\phi} \right|$$

$$p(\phi|\mathbf{t}^k) = p(\theta(\phi)|\mathbf{t}^k) \left| \frac{d\theta}{d\phi} \right|$$

Следовательно, отношение априорной и апостериорной плотностей не зависит от параметризации — Якобиан сокращается. Якобиан в $p(\phi|\mathbf{t}^k)$ сокращается в силу формулы для замены переменных под интегралом. Таким образом, внутренний интеграл не меняется, а, значит, не меняется и взаимная информация. □

8.3 Примеры опорных априорных распределений

Пример 8.2. Рассмотрим класс плотностей

$$M = \{f(x - \theta) : x \in \mathbb{R}, \theta \in \mathbb{R}\}.$$

Для такого класса плотностей зададим опорное априорное распределение $\pi(\theta)$.

Для фиксированного θ и случайной величины из распределения $f(x-\theta)$ пусть $y = x + a, \nu = \theta + a$. Определим $f'(y) = f(y - a - \theta)$. Рассмотрим семейство плотностей M', эквивалентное M:

$$M' = \{f'(y - \nu) : y \in \mathbb{R}, \nu \in \mathbb{R}\}.$$

Так как Якобиан сдвига равен единице, и плотность априорного распределения не зависит от сдвига, то $\pi'(\nu) = \pi(\theta)$. В силу инвариантности опорного априорного распределения относительно репараметризации $\pi'(\nu) = \pi(\theta+a)$. Следовательно, $\pi(\theta + a) = \pi(\theta)$ для произвольного a. Таким образом, опорное априорное распределение для одномерного параметра сдвига будет равномерным.

Пример 8.3. Рассмотрим теперь одномерный параметр масштаба. Определим семейство

$$S = \left\{ \frac{1}{\theta} f\left(\frac{x}{\theta}\right) : x > 0, \theta > 0 \right\}.$$

Взяв $y = \log x$, $\phi = \log \theta$, определим эквивалентное семейство плотностей:

$$S' = \{f(\exp(y - \phi)) : y \in \mathbb{R}, \phi \in \mathbb{R}\}.$$

Мы получили семейство распределений, для которого ϕ — параметр сдвига. В силу результата, полученного в предыдущем примере, $\pi'(\phi)$ — равномерное. Выполнено, что

$$\pi'(\phi) = \theta\pi(\theta).$$

Следовательно, опорное априорное распределение для параметра масштаба:

$$\pi(\theta) \propto \frac{1}{\theta}.$$

8.4 Использование метода Монте-Карло для получения опорного априорного распределения

Опорное априорное распределение имеет вид:

$$f_k(\theta) = \exp\left\{\int p(\mathbf{t}^k|\theta)\log\left(\frac{p(\mathbf{t}^k|\theta)h(\theta)}{\int p(\mathbf{t}^k|\theta)h(\theta)d\theta}\right)d\mathbf{t}^k\right\},$$

где $h(\theta)$ — исходное априорное распределение, от которого результат зависеть не будет.

Аналитически получить опорное априорное распределение получится только в нескольких случаях, поэтому используют приближенные подходы. Один из самых популярных — методы на основе идеи Монте-Карло. Пускай $\{x^{(i)}\}$ — выборка независимых одинаково распределенных случайных величин из распределения $p(x)$. Тогда для функции $f(x)$ можно оценить интеграл $\int f(x)p(x)dx$ как:

$$\mathbb{E}f(x) = \int f(x)p(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} f(x^{(i)}).$$

Сходимость будет, например, в силу закона больших чисел.

Предложим алгоритм сэмплирования из опорного априорного распределения на основе идеи Монте-Карло.

9 Что еще читать про Байесовскую математическую статистику

Существует множество хороших книг, в которых описано современное состояние Байесовской статистики и Байесовского машинного обучения.

В машинном обучении стоит начать с книги К.Бишопа [2]. Большую часть других вопросов покрывает более современная книга А.Гельмана [3]. Следует использовать последнее третье издание, в которое включен, например, раздел, посвященный процессам Дирихле. С другой стороны следует отметить, что эту книгу следует использовать скорее как справочник, чем как книгу, которую следует читать подряд.

В математической статистике данное пособие больше всего коррелирует с книгой [6], богатой на примеры использования Байесовской математической статистики и ее апологии.

Наиболее полно непараметрическая Байесовская статистика изложена в книге Дж.Гоша [4]. В этой книге довольного много опечаток, а изложение не всегда ясное и последовательное. Однако она наиболее полно представляет широкое многообразие результатов как в параметрической, так и в непараметрической Байесовской статистике.

A Основные вероятностные распределения

A.1 Многомерное нормальное распределение

Многомерное нормальное распределение или гауссовское распределение — такое вероятностное распределение $p(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ на $\mathbf{x} \in \mathbb{R}^d$, что его плотность имеет вид:

$$p(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right\}.$$

Два набора параметров распределения — вектор $\boldsymbol{\mu}$ и матрица Σ — определяют его среднее значение и ковариационную матрицу соответственно. Такое распределение обозначают $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$.

У нормального распределения множество замечательных свойств, о которых можно прочитать в отдельных главах этой книги или в более общих книгах, таких как книга Бишопа [2].

A.2 Распределение Дирихле

Носитель распределение Дирихле — симплекс. Для k-мерного распределения Дирихле симплекс есть множество точек, для которых:

$$S_k = \left\{\mathbf{x} : \sum_{i=1}^{k} x_i = 1, x_i \geq 0, i \in \{1, \ldots, k\}\right\}.$$

Легко видеть, что существует взаимнооднозначное соотвествие между вероятностными распределениями на конечном множестве $\{1, \ldots, k\}$ и точками такого симплекса.

Плотность распределения Дирихле с вектором параметров $\boldsymbol{\alpha} \in \mathbb{R}_+^k (\alpha_i \geq 0)$ есть:

$$p(\mathbf{x}|\boldsymbol{\alpha}) \propto x_1^{\alpha_1 - 1} \cdot \ldots \cdot x_k^{\alpha_k - 1}.$$

Получим нормировочный коэффициент для такого распределения:

$$\int_{\mathbf{x} \in S_k} x_1^{\alpha_1 - 1} \cdot \ldots \cdot x_k^{\alpha_k - 1} d\mathbf{x} = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k} \alpha_k)},$$

здесь $\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$ — гамма функция, $\Gamma(n+1) = n!$ для $n \in \mathbb{N}$ и $\Gamma(a+1) = a\Gamma(a)$.

Если мы рассмотрим $k = 2$, то получим бета-распределение. В частности, выполнено, что:

$$\int_0^1 \theta^{\alpha_1 - 1}(1 - \theta)^{\alpha_2 - 1} d\theta = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)}.$$

Пример A.1. Найдем среднее бета-распределения.

$$\begin{aligned}\mathbb{E}x &= \int_0^1 \theta \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1} d\theta = \\ &= \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^1 \theta^{\alpha_1+1-1}(1-\theta)^{\alpha_2-1} d\theta = \\ &= \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{\Gamma(\alpha_1+1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2+1)} = \frac{\alpha_1}{\alpha_1+\alpha_2}.\end{aligned}$$

Для распределения Дирихле с вектором параметров $\boldsymbol{\alpha}$ математическое ожидание равно $\mathbb{E}x_j = \frac{\alpha_j}{\sum_{i=1}^k \alpha_k}$.

A.3 Экспоненциальное семейство распределений

В этом разделе определим экспоненциальное семейство распределений и получим ряд его полезных свойств.

Определение 3. *Будем говорить, что распределение принадлежит экспоненциальному семейству, если его плотность (относительно меры Лебега) имеет вид:*

$$p(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x}) \exp\left(\boldsymbol{\theta}^\top T(\mathbf{x}) - A(\boldsymbol{\theta})\right).$$

Параметризация $\boldsymbol{\theta}$, для которой правдоподобие имеет такой вид, называется канонической, а вектор $T(\mathbf{x})$ — вектор достаточных статистик для модели, то есть такая функция данных $\mathbf{x}$, что условное распределение $P(\mathbf{x}|\boldsymbol{\theta})$ совпадает с условным распределением $P(\mathbf{x}|T, \boldsymbol{\theta})$. Эквивалентное утверждение, необходимое и достаточное условие того, что статистика является достаточной: $P(\boldsymbol{\theta}|\mathbf{x}, T) = P(\boldsymbol{\theta}|T)$.

Экспоненциальному семейству распределений принадлежат почти все используемые в математической статистике распределения: нормальное, биномиальное, Пуассоновское. Среди известных распределений, которые не принадлежат этому семейству распределений, — распределение Коши.

Приведем два примера экспоненциального семейства, канонических параметризаций и достаточных статистик для них.

Пример A.2. Рассмотрим распределение Бернулли, определенное на $x \in \{0, 1\}$.

$$\begin{aligned}p(x|\alpha) &= \alpha^x(1-\alpha)^{1-x} = \\ &= \exp\left[\log(\alpha^x(1-\alpha)^{1-x})\right] = \\ &= \exp\left[x\log\alpha + (1-x)\log(1-\alpha)\right] = \\ &= \exp\left[x\log\frac{\alpha}{1-\alpha} + \log(1-\alpha)\right] = \\ &= \exp\left[x\theta - \log(1+e^\theta)\right].\end{aligned}$$

Для распределения Бернулли

$$T(x) = x, \theta = \log \alpha 1 - \alpha, A(\theta) = \log(1 + e^{\theta}).$$

Покажем теперь, что нормальное распределение тоже принадлежит экспоненциальному семейству

Пример A.3.

$$\begin{aligned} p(x) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) = \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\log\sigma - \frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right) = \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(\boldsymbol{\theta}^\top T(x) - \log\sigma - \mu^2/(2\sigma^2)\right), \end{aligned}$$

$h(x) = \frac{1}{\sqrt{2\pi}}$, $A(\boldsymbol{\theta}) = \log\sigma + \mu^2/(2\sigma^2)$. Получаем достаточные статистики:

$$T(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$$

Каноническую параметризацию:

$$\boldsymbol{\theta} = \begin{pmatrix} \mu/\sigma^2 \\ -1/(2\sigma^2) \end{pmatrix}$$

Компонента $A(\theta)$ имеет следующий вид как функция от канонических параметров:

$$A(\theta) = \frac{\mu}{2\sigma^2} + \log\sigma = -\frac{\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2).$$

Приведем еще три важных свойства распределений из этого семейства.

Теорема 9.

$$\frac{dA(\boldsymbol{\theta})}{d\boldsymbol{\theta}} = \mathbb{E}_{p_{\boldsymbol{\theta}}} T(\mathbf{x}).$$

Доказательство. Используя то, что интеграл плотности вероятностного распределения равен 1, получим:

$$A(\boldsymbol{\theta}) = \log\left[\int_{\mathbb{R}^d} h(\mathbf{x}) \exp(\boldsymbol{\theta}^\top T(\mathbf{x})) d\mathbf{x}\right].$$

Обозначим $Q(\boldsymbol{\theta}) = \int_{\mathbb{R}^d} h(\mathbf{x}) \exp(\boldsymbol{\theta}^\top T(\mathbf{x})) d\mathbf{x}$. Подсчитаем производную:

$$
\begin{aligned}
\frac{dA(\boldsymbol{\theta})}{d\boldsymbol{\theta}} &= \frac{1}{Q(\boldsymbol{\theta})} \frac{dQ(\boldsymbol{\theta})}{d\boldsymbol{\theta}} = \frac{Q'(\boldsymbol{\theta})}{Q(\boldsymbol{\theta})} = \\
&= \frac{\int_{\mathbb{R}^d} h(\mathbf{x}) \exp(\boldsymbol{\theta}^\top T(\mathbf{x})) T(\mathbf{x}) d\mathbf{x}}{\int_{\mathbb{R}^d} h(\mathbf{x}) \exp(\boldsymbol{\theta}^\top T(\mathbf{x})) d\mathbf{x}} = \\
&= \frac{\int_{\mathbb{R}^d} h(\mathbf{x}) \exp(\boldsymbol{\theta}^\top T(\mathbf{x}) - A(\boldsymbol{\theta})) T(\mathbf{x}) d\mathbf{x}}{\int_{\mathbb{R}^d} h(\mathbf{x}) \exp(\boldsymbol{\theta}^\top T(\mathbf{x}) - A(\boldsymbol{\theta})) d\mathbf{x}} = \\
&= \int_{\mathbb{R}^d} h(\mathbf{x}) \exp(\boldsymbol{\theta}^\top T(\mathbf{x}) - A(\boldsymbol{\theta})) T(\mathbf{x}) d\mathbf{x} = \\
&= \mathbb{E}_{p_{\boldsymbol{\theta}}} T(\mathbf{x}).
\end{aligned}
$$

Получается, что мы явно можем выразить эту производную как математическое ожидание достаточной статистики:

$$
\frac{dA(\boldsymbol{\theta})}{d\boldsymbol{\theta}} = \mathbb{E}_{p_{\boldsymbol{\theta}}} T(\mathbf{x}).
$$

□

Теорема 10. *Функция $A(\boldsymbol{\theta})$ выпуклая.*

Доказательство. Если мы возьмем вторую производную, то получим:

$$
\frac{d^2 A(\boldsymbol{\theta})}{d\boldsymbol{\theta}^2} = \mathrm{Cov}_{p_{\boldsymbol{\theta}}} T(\mathbf{x}).
$$

Ковариационная матрица случайного вектора $\mathrm{Cov}_{p_{\boldsymbol{\theta}}} T(\mathbf{x})$ неотрицательно определена. Поэтому функция $A(\boldsymbol{\theta})$ выпуклая. □

Наконец обозначим $\boldsymbol{\mu} = \mathbb{E} T(\mathbf{x})$.

Теорема 11. *Пусть мы наблюдаем выборку независимых одинаково распределенных случайных величин $D = \{x_1, \dots, x_n\}$. Тогда оценка максимума правдоподобия $\widehat{\boldsymbol{\mu}}_{MLE}$:*

$$
\widehat{\boldsymbol{\mu}}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} T(x_i).
$$

Доказательство. Используем утверждение 9 и явно дифференцируем плотность. □

Отметим, что оценки максимума правдоподобия $\widehat{\boldsymbol{\mu}}_{MLE}$ будут несмещенными и эффективными (для них будет выполнено неравенство Рао-Крамера).

Предметный указатель

Список литературы

[1] J. Bernardo. Reference analysis. *Handbook of statistics*, 25:17–90, 2005.

[2] C.M. Bishop. Pattern recognition. *Machine Learning*, 128, 2006.

[3] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian data analysis*, volume 2. CRC press Boca Raton, FL, 2014.

[4] J.K. Ghosh and R.V. Ramamoorthi. *Bayesian Nonparametrics*. Springer Science & Business Media, 2003.

[5] M.I. Jordan. Lecture notes in stat260: Bayesian modeling and inference, January 2010.

[6] C. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

[7] Д.П. Ветров and Д.А. Кропотов. *Байесовские методы машинного обучения, учебное пособие по спецкурсу*. 2007.