Глава 3

Функции риска, функции потерь. Байесовские оценки. Минимаксные оценки

3.1 Базовая часть

3.1.1 Риски и потери

Определение функций риска и потерь

Поговорив об асимптотических свойствах оценок, перейдем к оцениванию на выборках фиксированного размера. Зачастую на практике мы сталкиваемся с тем, что оценивание производится с известной ценой за ошибку. В общем случае у нас задана функция потерь $L(\theta, x)$, которая отражает, сколько мы теряем, если настоящее значение параметра θ, а наша оценка x.

Будем считать, что $\Theta = \mathbb{R}^k$.

Определение 1. Функцией потерь называют такую функцию $L : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}^+$, что $L(\theta, \theta) = 0$.

Мы будем подставлять вместо x некоторые оценки $\widehat{\theta}(X_1, ..., X_n)$, где X_i — выборка. Таким образом, наши потери будут случайной величиной, которую мы, в некотором роде, хотели бы минимизировать. Со случайной потерей работать неудобно, поэтому рассматривают ее среднее.

Определение 2. Функцией риска называют отображение $R : \mathbb{R}^k \times \mathcal{L} \to \mathbb{R}^+$, заданное соотношением

$$R(\theta, \widehat{\theta}) = \mathbf{E}_\theta \left(L \left(\theta, \widehat{\theta}(X_1, ..., X_n) \right) \right).$$

Здесь $\mathcal{L}$ — множество оценок $\widehat{\theta}$, то есть измеримых отображений из множества выборок $\mathcal{X}$ в $\mathbb{R}^+$.

Функция потерь получает на вход θ и число $\widehat{\theta}$ (то есть значение статистики на данной реализации). Функция риска получает на вход θ и вид оценки $\widehat{\theta}$ (то есть форму зависимости $\widehat{\theta}$ от $X_1, ..., X_n$).

Пример 1. Пусть владелец булочной оценивает количество людей θ, которые придут к нему сегодня. Если его оценка x окажется заниженной, то он потеряет $c_1 \cdot (\theta - x)$ денег, где c_1 — выгода, получаемая владельцем с каждой булки. Если завышенной, то он потеряет $c_2 \cdot (x - \theta)$, где c_2 — цена производства одной булки. При этом он будет использовать некоторую оценку $\widehat{\theta}$, связанную со статистикой предыдущих дней, поэтому его потери будут случайными, зависящими от θ и от выборки. Для пекаря его потери — ключевая характеристика метода оценивания, превалирующая над "хорошими" свойствами оценок.

Предположим, что X_i имеют нормальное распределение со средним θ и дисперсией 1. Тогда у оценки $\widehat{\theta} = X_1$ функция риска будет равна

$$c_2 \mathbf{E}(X_1 - \theta) I_{X_1 > \theta} + c_1 \mathbf{E}(\theta - X_1) I_{X_1 < \theta} = \frac{c_1 + c_2}{\sqrt{2\pi}} \int\limits_0^\infty x e^{-x^2/2} dx = \frac{c_1 + c_2}{\sqrt{2\pi}}.$$

Три классических функции потерь

- В одномерном случае часто рассматривают квадратичную функцию потерь
$$L(\theta; x) = (x-\theta)^2.$$
В многомерном ее аналогом может быть L^2 потеря $||\theta - x||^2$ или более общая конструкция
$$||\theta - x||_C^2 = (\theta - x)C(\theta - x)^T,$$
где C – некоторая заданная матрица.
- Вторым популярным вариантом является абсолютная потеря
$$L(\theta; x) = |x-\theta|.$$
В многомерном случае можно рассматривать взамен L^1 потерю $\sum_{i=1}^k |x_i - \theta_i|$ или более общая взвешенная конструкция
$$\sum_{i=1}^k c_i |x_i - \theta_i|.$$
- Третьим популярным вариантом (преимущественно в случае, если множество значений θ не более чем счетно) является дискретная потеря
$$L(\theta; x) = I_{\theta \neq x}.$$

О важности квадратичной функции потерь

Зачастую рассматриваются функции потерь, зависящие только от разности θ и $\widehat{\theta}$, как это было в примере 1 или в трех случаях выше. В таком случае особенную роль играет квадратичная функция потерь $L(\theta, x) = (\theta - x)^2$. Важность ее вытекает из следующего соображения — любая гладкая дважды дифференцируемая функция потерь L, зависящая только от $\theta - x$, разлагается в ряд Тейлора
$$L(\theta - x) = a + b(\theta - x) + c(\theta - x)^2 + \varepsilon,$$
где $a = 0$, $b = 0$ из условий неотрицательности и равенства 0 в точке 0. В свою очередь ε по порядку есть $o((\theta - x)^2)$, то есть при близко оценивающих θ оценках (например, при состоятельных $\widehat{\theta}$ и больших n) функция риска будет вести себя как $c(\theta - x)^2$.

В случае, если θ — векторный параметр, аналогичным образом можно использовать $||\theta - x||_C^2$, которая также дает основной член асимптотики.

Квадратичная потеря и дисперсия

Для квадратичной функции потерь функция риска будет просто равна $\mathbf{E}_\theta(\widehat{\theta} - \theta)^2$. В частности, для несмещенной оценки это будет $\mathbf{D}_\theta\widehat{\theta}$. Для смещенных оценок в силу соотношения
$$\mathbf{E}(X-a)^2 = \mathbf{E}(X - \mathbf{E}X + \mathbf{E}X - a)^2 = \mathbf{E}(X - \mathbf{E}X)^2 + (\mathbf{E}X - a)^2 + 2(\mathbf{E}X - a)\mathbf{E}(X - \mathbf{E}X) = \mathbf{D}X + (\mathbf{E}X - a)^2,$$
риск для квадратичной функции потерь есть $\mathbf{D}_\theta\widehat{\theta} + (\mathbf{E}_\theta\widehat{\theta} - \theta)^2$.

3.1.2 Три подхода к минимизации риска

Основной подход

Мы хотели бы выбрать оценку так, чтобы минимизировать функцию риска. Однако, функция риска при каждой оценке будет зависеть от θ и потому необходимо конкретизировать, как мы будем выбирать "меньшую" среди функции. Здесь имеются три основных подхода:

1. Ограничение рассматриваемых оценок. Если мы ограничим множество рассматриваемых оценок до какого-то класса, то может оказаться, что в этом классе одна из функций риска лежит ниже всех остальных.

   Определение 3. Оценка $\widehat{\theta}$ называется равномерно наиболее мощной в классе оценок $K \subset \mathcal{L}$, если $\widehat{\theta} \in K$ и $R(\theta, \widehat{\theta}) \leq R(\theta, \widehat{\theta}_1)$ при всех $\theta \in \Theta$ и всех $\widehat{\theta}_1 \in K$.

   Мы уже искали такие оценки для квадратичного риска на первом семинаре. Более общие подходы к таким задачам мы рассмотрим позднее.

2. Минимаксный подход. Будем использовать "подход пессимиста". Давайте предположим, что по закону подлости наверняка выпадет тот самый параметр при котором риск наибольший. Следовательно, наша задача — минимизировать максимум функции риска.

   Определение 4. Оценка $\widehat{\theta}$ такая, что $\max_\theta R(\theta, \widehat{\theta}) \leq \max_\theta R(\theta, \widehat{\theta}_1)$ для любой оценки $\widehat{\theta}_1$ называется *минимаксной.*

   Методы построения таких оценок вынесены в факультатив.

3. Байесовский подход. Будем считать, что мы знаем функцию $\pi(\theta)$ на Θ, описывающую частоту встречаемости параметров θ в окружающей реальности. Тогда осмысленно минимизировать средний риск (его называют байесовским)

$$R(\widehat{\theta}) = \int_\Theta R(\theta; \widehat{\theta}) \pi(\theta) d\theta.$$

   Определение 5. Оценка $\widehat{\theta}$ такая, что $R(\widehat{\theta}) \leq R(\widehat{\theta}_1)$ для любой оценки $\widehat{\theta}_1$ называется *байесовской.*

Примеры

Пример 2. Рассмотрим оценку $\overline{X}$ в схеме Бернулли. Как мы докажем чуть позже это оценка с наименьшей возможной дисперсией среди несмещенных. Значит в классе несмещенных оценок $\overline{X}$ имеет риск для квадратичной функции потерь, то есть равномерно наиболее мощна. Ее риск при этом равен $\theta(1-\theta)/n$.
При этом среди всех оценок она не является равномерно наиболее мощной, например, при $\theta = 1/2$ оценка $\widehat{\theta}(X_1, ..., X_n) = 1/2$ будет иметь нулевой риск, а $\overline{X}$ —- ненулевой.

Пример 3. Для схемы Бернулли минимаксной оценкой, как следует из одной из факультативных задач, будет вовсе не $\overline{X}$, а смещенная оценка

$$\widetilde{\theta}(X_1, ..., X_n) = \overline{X} + (1/2 - \overline{X}) \frac{1}{1 + \sqrt{n}},$$

называемая оценкой Ходжеса-Лемана. Сравним эти оценки для выборки размера 1. Для квадратичной функции потерь риск для первой есть $\theta(1-\theta)$. Для второй

$$\left(1 + \frac{-1/2}{2} - \theta\right)^2 \theta + \left(\frac{1/2}{2} - \theta\right)^2 (1-\theta) = \left(\frac{3}{4} - \theta\right)^2 \theta + \left(\frac{1}{4} - \theta\right)^2 (1-\theta) = \frac{1}{16}.$$

Очевидно, что в первом случае максимум $1/4$, что больше $1/16$.

Пример 4. Приведем пример использования байесовского подхода. Представим себе, что мы хотим провести десять испытаний Бернулли и на основе их решить, готовы ли мы заключить пари, что из следующих 1000 испытаний будет не больше 100 успехов. Проведя 10 испытаний, мы видим в них десять успехов. Это описание дает жесткое описание математическому эксперименту и, по идее, мы должны во всех реализациях такой задачи получить одну и ту же оценку параметра θ и сделать один и тот же вывод. Давайте представим себе три эксперимента. В первом из них ваш знакомый угадывает на орла или на решку упадет монета (монету подбрасываете вы, она симметричная и заведомо никаких мошенничеств нет). Во втором пожилая леди по вкусу угадывает, что раньше положили в чай — лимон или сахар. В третьем музыкальный эксперт определяет по короткому отрывку, кому принадлежит произведение — Шуману или Шуберту. Неужели в жизни вы приняли бы одно и то же решение во всех трех экспериментах?

Различные взгляды на эти три схемы Бернулли связаны с тем, что на практике у вас есть изначальное представление о том, насколько часто может встречаться тот или иной параметр. Это может быть выражено, например, в виде плотности распределения (или в дискретном случае самого распределения) параметра θ по множеству Θ. Ниже мы будем использовать термин "плотность" для обоих этих случаев. Плотность $\pi(\theta)$ может появляться как обобщение прошлых опытов (например, фармацевтическая кампания, выпуская новый препарат, может оценить его продаваемость по прошлым продажам других препаратов) или же обобщать наши представления о том, с какой частотой встречается тот или иной параметр. Такая плотность называется априорной, то есть доопытной. В нашем случае, для первого варианта можно предположить распределение с сильным пиком в $1/2$, скажем, напоминающее нормальное распределение с очень маленькой дисперсией. Для третьего, если есть мнение, что угадывающий действительно эксперт, можно взять плотность с большими значениями в области $[3/4, 1]$. Для второго в силу неизведанности этого явления, поставить равномерную на $[0, 1]$ плотность.

Итак, в байесовском случае мы сравниваем не все значения функции риска, а сравниваем функции риска с учетом вероятности параметра θ (скажем, пекарь в примере 1, не будет сильно печалиться из-за оценки, приносящей большие потери в случае, когда число клиентов за день $\theta = 6 \cdot 10^9$). Это позволяет эффективно использовать результаты прошлых опытов.

Однако, метод уязвим за счет неоднозначности определения априорной плотности. Скажем, кто-то может заявить, что эксперты, отличающие Шумана от Шуберта — это миф и поставит им в соответствие плотность, которую мы предложили для случая 1, зато экстрасенсы встречаются достаточно часто и у человека, угадывающего результат выпадения монеты, априорная плотность имеет форму, предложенную нами для случая 3, а понять, что первым положили в чай — сахар или лимон, может вообще каждый с вероятностью 1.

Не стоит смущаться тем, что параметр θ, как мы говорили прежде, некоторая неизвестная, но фиксированная величина. Это не мешает считать, что он явился результатом некоторого случайного опыта. Оценивая расстояние между Нью-Йорком и Москвой, мы как будто предполагаем, что Нью-Йорк был построен на случайном расстоянии от Москвы и, по стечению обстоятельств, попал именно туда, куда попал.

3.1.3 Построение байесовской оценки

Апостериорная плотность

Итак, имея некоторые доопытные представления о распределении параметра, мы должны скорректировать их с учетом выборки. Иначе говоря, мы должны посчитать условную плотность (дискретное распределение) параметра θ при условии наблюдения выборки x.

В случае дискретных Θ, X из формулы Байеса вытекает соотношение

$$f_{\theta|X_1,...,X_n}(u|x_1,...,x_n) = P(\theta = u|X_1 = x_1,...,X_n = x_n) = \frac{P_u(X_1 = x_1)...P_u(X_n = x_n)\pi(u)}{\sum_{\theta\in\Theta} P_\theta(X_1 = x_1)...P_\theta(X_n = x_n)\pi(\theta)}.$$

Распределение $f_{\theta|X_1,...,X_n}$ называют апостериорным распределением параметра θ.

В случае абсолютно-непрерывных распределений плотностью $Y|X$ называют

$$f_{Y|X}(y|x) = \frac{f_{Y,X}(y,x)}{f_X(x)} = \frac{f_{X|Y}(x|y)f_Y(y)}{\int_{\mathbb{R}} f_{X,Y}(x,v)dv} = \frac{f_{X|Y}(x|y)f_Y(y)}{\int_{\mathbb{R}} f_{X|Y}(x|v)f_Y(v)dv}$$

Таким образом, апостериорную плотность параметра θ в абсолютно-непрерывном случае естественно задавать соотношением

$$f_{\theta|X_1,...,X_n}(u|x_1,...,x_n) = \frac{f_u(x_1)...f_u(x_n)\pi(u)}{\int_\Theta f_v(x_1)...f_v(x_n)\pi(v)dv}.$$

Пример 5. Для второго эксперимента примера 1 мы выбрали равномерную на [0,1] априорную плотность и наблюдаем успехов из 10 опытов. Для схемы Бернулли с равномерно распределенным параметром апостериорная плотность имеет вид

$$f_{\theta|X_1,...,X_n}(u|x_1,...,x_n) = \frac{u^{\sum x_i}(1-u)^{n-\sum x_i} I_{u\in[0,1]}}{\int\limits_0^1 u^{\sum x_i}(1-u)^{n-\sum x_i} du} = \frac{u^{\sum x_i}(1-u)^{n-\sum x_i} I_{u\in[0,1]}}{B(1+\sum x_i, n+1-\sum x_i)}.$$

В нашем случае имеем $B(11,1) = \frac{\Gamma(1)\Gamma(11)}{\Gamma(12)} = 1/11$, то есть плотность будет иметь вид $11u^{10}I_{u\in[0,1]}$. Как видим по рисунку, плотность "покосилась" направо.
В случае произвольных наблюдений $x_1,...,x_n$ апостериорная плотность имела бы вид

$$f_{\theta|X_1,....,X_n}(u|x_1,...,x_n) = \frac{u^{x_1+...+x_n}(1-u)^{n-x_1-...-x_n}}{\int_0^1 u^{x_1+...+x_n}(1-u)^{n-x_1-...-x_n}du}.$$

Отметим полезное соображение — можно не высчитывать знаменатель, а лишь заметить, что он не зависит от u, то есть плотность имеет вид

$$Cu^a(1-u)^b I_{u\in[0,1]},$$

где $a = x_1 + ... + x_n$, $b = n - x_1 - ... - x_n$, C — некоторая константа (то есть не зависящая от u величина, но при этом зависящая от $x_1,...,x_n$). Мы знаем распределение такого вида — это бета-распределение, в котором $C = 1/B(a+1, b+1)$. Но плотности не могут отличаться в константу раз, поэтому апостериорная плотность есть плотность бета-распределения с параметрами $x_1 + ... + x_n + 1, n - x_1 - ... - x_n + 1$.

Сопряженные распределения

Наиболее удобно работать с байесовскими оценками в том случае, когда апостериорная плотность или апостериорное распределение будет того же типа, что априорное (например, и то, и то распределение будет нормальным, но с разными параметрами). Такое априорное распределение называется сопряженным (в иностранной литературе conjugate prior). Список популярных сопряженных распределений можно найти по ссылке

Байесовская оценка для квадратичной потери

Как же найти байесовскую оценку? Ответы можно получить для различных видов функции потерь. Начнем с классической ситуации квадратичной потери. В этом случае байесовская оценка есть математическое ожидание величины с апостериорной плотностью:

Теорема 1. *Для квадратичной функции потерь байесовская оценка имеет вид*

$$\theta^*(x_1,...,x_n) = \frac{\int_\Theta \theta f_\theta(x_1)...f_\theta(x_n)\pi(\theta)d\theta}{\int_\Theta f_\theta(x_1)...f_\theta x_n \pi(\theta)d\theta}.$$

При этом предполагается конечность величины в числителе и положительность и конечность величины в знаменателе.

При этом риск (не байесовский, а обычный) такой оценки будет равен математическому ожиданию от дисперсии апостериорного распределения (дисперсия апостериорного распределения есть функция выборки, а математическое ожидание берется по плотности f_θ).

Байесовская оценка для абсолютной и дискретной потери

Сформулируем еще две теоремы (их доказательства отнесены в факультативные задачи) о виде байесовской оценки для других видов функции потерь:

Теорема 2. *Пусть функция потерь абсолютная:* $L(\theta, x) = |\theta - x|$. *Назовем апостериорной функцией распределения функцию*

$$\widetilde{F}(\theta|x_1, ..., x_n) = \frac{\int_{-\infty}^{\theta} f_u(x_1)....f_u(x_n)\pi(u)du}{\int_{-\infty}^{\infty} f_u(x_1)....f_u(x_n)\pi(u)du} = \int_{-\infty}^{\theta} f_{\theta|X_1,...,X_n}(u|x_1, ..., x_n)du.$$

Пусть $\widetilde{F}$ *— непрерывная функция. Тогда байесовской оценкой* $\hat{\theta}$ *будет медиана* $\widetilde{F}^{-1}(1/2)$, *т.е.* $\widetilde{F}(\hat{\theta}(x_1, ..., x_n)|x_1, ..., x_n) = 1/2$.

Теорема 3. *Пусть* $\Theta = \{\theta_1, ..., \theta_k\}$, *функция потерь дискретная:* $L(\theta, x) = I_{\theta \neq x}$. *Иначе говоря, мы решаем задачу классификации — если мы угадываем* θ, *то ничего не теряем, если не угадываем — теряем 1. Тогда оценки будут принимать только значения* $\theta_1, ..., \theta_k$. *Байесовской оценкой при этом будет* $\widehat{\theta}(x_1, ..., x_n)$, *равная* θ_i *при тех* $x_1, ..., x_n$, *при которых*

$$\mathbf{P}(\theta = \theta_i | X_1 = x_1, ..., X_n = x_n) > \mathbf{P}(\theta = \theta_j | X_1 = x_1, ..., X_n = x_n), \ \forall j \neq i.$$

Это вполне естественно — мы выбираем то θ, *которое апостериорно наиболее вероятно.*

Пример

Пример 6. Таким образом, для квадратичного риска в схеме Бернулли байесовской оценкой при равномерной априорной плотности будет

$$\frac{B(2 + \sum x_i + 2, n + 1 - \sum x_i)}{B(1 + \sum x_i, n + 1 - \sum x_i)} = \frac{\Gamma(2 + \sum x_i)}{\Gamma(1 + \sum x_i)} \frac{\Gamma(n+2)}{\Gamma(n+3)} = \frac{1 + \sum x_i}{n+2}$$

В случае примера 5 она примет значение $11/12 \approx 0.92$. Для абсолютного риска в том же примере я должен был найти апостериорную функцию распределения (x^{11} при $x \in [0, 1]$) и найти ее медиану из соотношения $\widehat{\theta}^{11} = 1/2$, откуда $\widehat{\theta} = 2^{-1/11} \approx 0.94$.

3.2 Факультатив

3.2.1 Доказательство теоремы 1

Доказательство. Доказательство достаточно просто:

$$R(\widehat{\theta}) = \int_{\Theta} R(\theta, \widehat{\theta})\pi(\theta)d\theta = \int_{\Theta}\int_{R^n} (\widehat{\theta} - \theta)^2 f_\theta(x_1)...f_\theta(x_n)dx_1...dx_n\pi(\theta)d\theta =$$
$$\int_{R^n}\int_{\Theta} (\widehat{\theta}(x_1, ..., x_n) - \theta)^2 \pi(\theta) f_\theta(x_1)...f_\theta(x_n)d\theta dx_1...dx_n.$$

Внутренний интеграл есть

$$\widehat{\theta}^2(x_1, ..., x_n)\int_{\Theta} f_\theta(x_1)...f_\theta(x_n)\pi(\theta)d\theta - 2\widehat{\theta}(x_1, ..., x_n)\int_{\Theta} \theta f_\theta(x_1)...f_\theta(x_n)\pi(\theta)d\theta +$$
$$\int_{\Theta} \theta^2 f_\theta(x_1)...f_\theta(x_n)\pi(\theta)d\theta = a\widehat{\theta}^2 + b\widehat{\theta} + c.$$

При каждом наборе $x_1, ..., x_n$ он будет достигать минимума при $\widehat{\theta}(x_1, ..., x_n)$, являющимся вершиной соответствующей параболы, то есть $-b/(2a)$. Значит и максимум интеграла будет достигнут при такой $\widehat{\theta}$. Аналогичным образом можно решать задачи для неквадратичного риска. □

3.2.2 Минимаксные оценки

Для построения минимаксной оценки удобно использовать следующие несложные результаты

Теорема 4. *Если оценка $\widehat{\theta}$ имеет постоянный по θ риск и найдется такая априорная плотность π, что оценка $\widehat{\theta}$ — байесовская, то $\widehat{\theta}$ — минимаксная оценка.*

Доказательство. Пусть $\widehat{\theta}$ не минимаксна. Тогда найдется такая $\widehat{\theta}_1$, что

$$\sup_{\theta\in\Theta} R(\theta;\widehat{\theta}_1) < \sup_{\theta\in\Theta} R(\theta;\widehat{\theta}) = R(\theta,\widehat{\theta}).$$

Но тогда при всех θ справедливо неравенство

$$R(\theta;\widehat{\theta}_1) < R(\theta,\widehat{\theta}).$$

Но поскольку оценка $\widehat{\theta}$ байесовская,

$$\int_\Theta R(\theta,\widehat{\theta})\pi(\theta)d\theta \leq \int_\Theta R(\theta,\widehat{\theta}_1)\pi(\theta)d\theta.$$

Эти два неравенства противоречат друг другу. Следовательно, $\widehat{\theta}$ минимаксна. □

Таким образом, удачно подобрав априорную плотность, мы можем доказывать минимаксность оценки с постоянным риском. Более того, справедлива и более общая теорема:

Теорема 5. *Если оценка $\widehat{\theta}$ имеет постоянный по θ риск и найдется такая последовательность априорных плотностей π_n, что $R(\widehat{\theta},\theta) \leq \liminf_{n\to\infty} R(\widehat{\theta}_n)$ при всех θ, где $\widehat{\theta}_n$ — байесовские с π_n, R — их байесовский риск, то $\widehat{\theta}$ — минимаксная оценка.*

Доказательство практически аналогично.