Московский государственный университет имени М.В. Ломоносова

Факультет вычислительной математики и кибернетики

Первое теоретическое задание: Сопряженные распределения и экспоненциальный класс распределений

Каратыщев Дмитрий Иванович, 417 группа

Дата выполнения работы:
15 октября 2024 г.

Содержание

1 Задача 1

1.1 Условие

Пусть $\mathbb{X} = (x_1, x_2, \ldots, x_n)$ - независимая выборка из непрерывного равномерного распределения $U[0, \vartheta]$. Требуется найти оценку максимального правдоподобия ϑ_{ML}, подобрать сопряжённое распределение $p(\vartheta)$, найти апостериорное распределение $p(\vartheta \mid x_1, \ldots, x_n)$ и вычислить его статистики: мат. ожидание, медиану и моду. Формулы для статистик нужно вывести, а не взять готовые. Подсказка: *задействовать распределение Парето.*

1.2 Решение

1.2.1 Нахождение оценки максимального правдоподобия

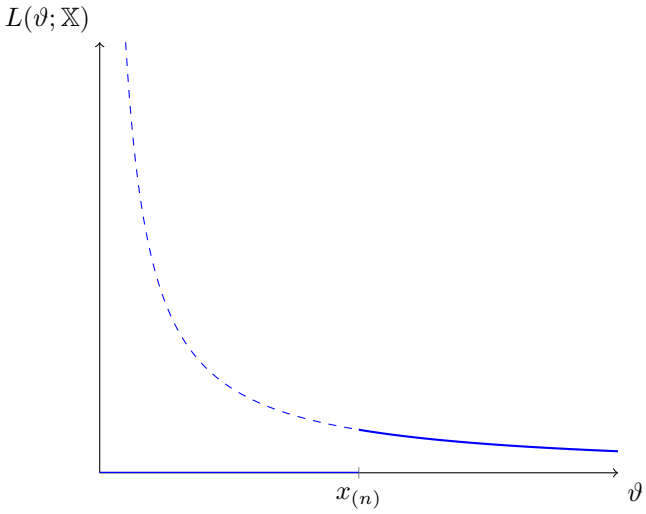
Из условия дано, что $x_i \sim U[0, \vartheta],\ \forall i \in \overline{1, n}$. Обозначим плотность данного непрерывного равномерного распределения за p_U. Известно, что плотность непрерывного равномерного распределения p_U определяется следующим образом:

$$p_U(x \mid \vartheta) = \frac{\mathbb{I}[0 \le x \le \vartheta]}{\vartheta}$$

Запишем правдоподобие выборки $\mathbb{X}$ с учётом факторизации плотности по отдельным наблюдениям в силу их независимости:

$$L(\vartheta; \mathbb{X}) = p_U(\mathbb{X} \mid \vartheta) = \prod_{i=1}^{n} p(x_i \mid \vartheta) = \frac{1}{\vartheta^n} \prod_{i=1}^{n} \mathbb{I}[0 \le x_i \le \vartheta] = \frac{1}{\vartheta^n} \mathbb{I}[0 \le \max_{1 \le i \le n} x_i \le \vartheta] = \frac{1}{\vartheta^n} \mathbb{I}[0 \le x_{(n)} \le \vartheta]$$

Помним, что оценкой максимального правдоподобия называется такое значение параметра, при котором $L(\vartheta_{ML}; \mathbb{X}) \ge L(\vartheta; \mathbb{X})\ \forall \vartheta \in \Theta$, где Θ - это пространство параметров. Покажем график зависимости функции правдоподобия от параметра при фиксированном значении выборки.



Отсюда видно, что $\vartheta_{ML} = x_{(n)}$, так как именно при этом значении достигается максимальное значение функции правдоподобия при заданном значении выборки. Заметим, однако, что данная оценка не является несмещённой. Это можно показать следующим образом:

$$\mathbb{P}(x_{(n)} < x) = \mathbb{P}(x_1 < x, \ldots, x_n < x) = \prod_{i=1}^{n} \mathbb{P}(x_i < x) = \frac{x^n}{\vartheta^n},\ \forall x \in [0, \vartheta]$$

$$\mathbb{E}_\vartheta[x_{(n)}] = \int_0^\vartheta x \frac{d}{dx} \left(\mathbb{P}(x_{(n)} < x)\right) dx = \int_0^\vartheta \frac{n x^n}{\vartheta^n} dx = \left. \frac{n x^{n+1}}{(n+1)\vartheta^n} \right|_0^\vartheta = \frac{n}{n+1}\vartheta \ne \vartheta$$

Поэтому возьмём в качестве несмещённой оценки максимального правдоподобия $\hat{\vartheta}_{ML} = \frac{n+1}{n} x_{(n)}$

1.2.2 Подбор сопряжённого распределения

Наша задача найти такое априорное распределение для параметра ϑ, чтобы апостериорное распределение $p(\vartheta \mid x)$ находилось в том же семействе, что и априорное. Запишем ещё раз правдоподобие нашей выборки и рассмотрим его как функцию от ϑ:

$$p_U(\mathbb{X} \mid \vartheta) = \frac{1}{\vartheta^n}\mathbb{I}[0 \le x_{(n)} \le \vartheta]$$

Можно заметить, что распределение вида $p(\vartheta \mid \alpha, \vartheta_0) = \dfrac{C}{\vartheta^\alpha}\mathbb{I}[\vartheta_0 \le \vartheta]$, где $\vartheta_0 > 0, \alpha > 1$, будет сопряжённым для равномерного распределения. Покажем, что это верно:

$$p(\vartheta \mid \mathbb{X}) \propto \frac{1}{\vartheta^n}\mathbb{I}[0 \le x_{(n)} \le \vartheta] \cdot \frac{C}{\vartheta^\alpha}\mathbb{I}[\vartheta_0 \le \vartheta] = \frac{C}{\vartheta^{n+\alpha}}\mathbb{I}[\max\{\vartheta_0, x_{(n)}\} \le \vartheta]$$

Мы получили, что новые параметры для нашего распределения - это $\alpha_1 = n+\alpha,\ \vartheta_1 = \max\{\vartheta_0, x_{(n)}\}$. Осталось лишь определить константу, которая фигурирует в предложенном сопряжённом распределении. Сделаем это через условие нормировки плотности вероятности:

$$1 = \int_{\vartheta_0}^{+\infty} \frac{C}{\vartheta^\alpha} d\vartheta = \frac{C\vartheta^{-\alpha+1}}{-\alpha+1}\bigg|_{\vartheta_0}^{+\infty} = \frac{C\vartheta_0^{-\alpha+1}}{\alpha-1}$$

$$C = (\alpha-1)\vartheta_0^{\alpha-1}$$

Если убрать знак пропорции и записать равенство, то новая константа, очевидно, будет иметь следующий вид:

$$\hat{C} = (\alpha+n-1)(\max\{\vartheta_0, x_{(n)}\})^{\alpha+n-1}$$

Полученное сопряжённое распределение называется *распределением Парето* и выглядит следующим образом:

$$p(\vartheta \mid \alpha, \vartheta_0) = \frac{(\alpha-1)\vartheta_0^{\alpha-1}}{\vartheta^\alpha}\mathbb{I}[\vartheta_0 \le \vartheta]$$

В дополнение к этому мы получили апостериорное распределение:

$$p(\vartheta \mid x_1, \dots, x_n) = C_1 \cdot p(x_1, \dots, x_n \mid \vartheta) \cdot p(\vartheta) = \frac{(\alpha+n-1)(\max\{\vartheta_0, x_{(n)}\})^{\alpha+n-1}}{\vartheta^{\alpha+n}}\mathbb{I}[\max\{\vartheta_0, x_{(n)}\} \le \vartheta]$$

Если же считать гиперпараметры сопряжённого распределения случайными величинами и определять соответствующее апостериорное распределение для них, то итоговое апостериорное распределение можно найти следующим образом:

$$p(\vartheta \mid x_1, \dots, x_n) = \int_1^{+\infty}\int_0^{+\infty} p(\vartheta \mid \alpha, \vartheta_0, x_1, \dots, x_n) p(\alpha, \vartheta_0 \mid x_1, \dots, x_n) d\vartheta_0 d\alpha$$

Для дальнейших выкладок будем считать, что α и ϑ_0 - это просто некоторые неслучайные гиперпараметры.

1.2.3 Подсчёт медианы, моды и математического ожидания апостериорного распределения

Подсчитаем моду апостериорного распределения следующим образом:

$$\vartheta_{MP} = \underset{\vartheta\in\Theta}{\operatorname{argmax}}\ p(\vartheta \mid x_1, \dots, x_n) = \underset{\vartheta\in\Theta}{\operatorname{argmax}}\ \left[\frac{(\alpha+n-1)(\max\{\vartheta_0, x_{(n)}\})^{\alpha+n-1}}{\vartheta^{\alpha+n}}\mathbb{I}[\max\{\vartheta_0, x_{(n)}\} \le \vartheta]\right] = \max\{\vartheta_0, x_{(n)}\}$$

Таким образом, **мода апостериорного распределения** ϑ_{MP} равна $\max\{\vartheta_0, x_{(n)}\}$.

Теперь найдём математическое ожидание апостериорного распределения. Пусть $\xi \sim p(\vartheta \mid \mathbb{X})$. Тогда можно записать следующее:

$$\mathbb{E}_\vartheta \xi = (\alpha+n-1)(\max\{\vartheta_0, x_{(n)}\})^{\alpha+n-1}\int_{\max\{\vartheta_0, x_{(n)}\}}^{+\infty} \frac{\vartheta}{\vartheta^{\alpha+n}} d\vartheta$$

$$\mathbb{E}_\vartheta \xi = (\alpha + n - 1)(\max\{\vartheta_0, x_{(n)}\})^{\alpha+n-1} \frac{\vartheta^{2-\alpha-n}}{2-\alpha-n}\Bigg|_{\max\{\vartheta_0, x_{(n)}\}}^{+\infty}$$

$$\mathbb{E}_\vartheta \xi = (\alpha + n - 1)(\max\{\vartheta_0, x_{(n)}\})^{\alpha+n-1} \frac{(\max\{\vartheta_0, x_{(n)}\})^{2-\alpha-n}}{\alpha+n-2} = \frac{\alpha+n-1}{\alpha+n-2} \max\{\vartheta_0, x_{(n)}\}$$

$$\mathbb{E}_\vartheta \xi = \left(1 + \frac{1}{\alpha+n-2}\right) \max\{\vartheta_0, x_{(n)}\}, \text{ при } \alpha + n > 2$$

Таким образом, **математическое ожидание апостериорного распределения** существует при $\alpha + n > 2$ и равно $\left(1 + \frac{1}{\alpha+n-2}\right) \max\{\vartheta_0, x_{(n)}\}$

Перейдём к нахождению медианы апостериорного распределения. Медиана - это квантиль порядка $1/2$. Её можно найти из следующего условия:

$$\frac{1}{2} = F_\vartheta(\vartheta_{med}) = \int_{-\infty}^{\vartheta_{med}} p(\vartheta \mid \mathbb{X}) d\vartheta, \text{ где } F_\vartheta - \text{функция апостериорного распределения}$$

Воспользуемся этим условием и получим медиану нашего распределения.

$$\frac{1}{2} = (\alpha + n - 1)(\max\{\vartheta_0, x_{(n)}\})^{\alpha+n-1} \int_{\max\{\vartheta_0, x_{(n)}\}}^{\vartheta_{med}} \frac{1}{\vartheta^{\alpha+n}} d\vartheta$$

$$\frac{1}{2} = (\alpha + n - 1)(\max\{\vartheta_0, x_{(n)}\})^{\alpha+n-1} \frac{\vartheta^{1-\alpha-n}}{1-\alpha-n}\Bigg|_{\max\{\vartheta_0, x_{(n)}\}}^{\vartheta_{med}}$$

$$\frac{1}{2} = -(\max\{\vartheta_0, x_{(n)}\})^{\alpha+n-1} \vartheta_{med}^{1-\alpha-n} + 1$$

$$\vartheta_{med}^{1-\alpha-n} = \frac{(\max\{\vartheta_0, x_{(n)}\})^{1-\alpha-n}}{2}$$

$$\vartheta_{med} = 2^{\frac{1}{\alpha+n-1}} \max\{\vartheta_0, x_{(n)}\}$$

Так, **медиана апостериорного распределения** ϑ_{med} существует и равна $2^{\frac{1}{\alpha+n-1}} \max\{\vartheta_0, x_{(n)}\}$

2 Задача 2

2.1 Условие

Предположим, что вы приезжаете в новый город и видите автобус с номером 100. Требуется с помощью байесовского подхода оценить общее количество автобусных маршрутов в городе. Каким априорным распределением стоит воспользоваться (обоснуйте выбор его параметров)? Какая из статистик апостериорного распределения будет наиболее адекватной (обоснуйте свой выбор)? Как изменятся оценки на количество автобусных маршрутов при последующем наблюдении автобусов с номерами 50 и 150? *Подсказка: воспользоваться результатами предыдущей задачи. При этом обдумать, как применить непрерывное распределение к дискретным автобусам.*

2.2 Решение

Положим общее количество автобусных маршрутов в городе за ϑ. По принципу максимальной энтропии будем считать, что каждое наблюдение номера автобуса - это реализация случайной величины из дискретного равномерного распределения $Unif[1, \vartheta]$, так как именно равномерное распределение обладает максимальной энтропией в дискретном случае. Минимальным берём значение, равное 1, так как считаем, что хотя бы один автобусный маршрут существует. Тогда вероятность встретить тот или иной номер маршрута определяется так:

$$\mathbb{P}(x \mid \vartheta) = \frac{\mathbb{I}[1 \le x \le \vartheta]}{\vartheta}$$

.

В качестве априорного распределения на неизвестный параметр ϑ возьмём дискретный аналог распределения Парето, который определим следующим образом:

$$\mathbb{P}_{pareto}(\vartheta = k \mid \vartheta_0, \alpha) = \mathbb{I}[\vartheta_0 \le k] \int_k^{k+1} \frac{(\alpha-1)\vartheta_0^{\alpha-1}}{\vartheta^\alpha} d\vartheta$$

.

Во-первых, данный аналог действительно является распределением вероятностей, так как:

$$\sum_{k=-\infty}^{+\infty} \mathbb{I}[\vartheta_0 \le k] \int_k^{k+1} \frac{(\alpha-1)\vartheta_0^{\alpha-1}}{\vartheta^\alpha} d\vartheta = \sum_{k=\vartheta_0}^{+\infty} \int_k^{k+1} \frac{(\alpha-1)\vartheta_0^{\alpha-1}}{\vartheta^\alpha} d\vartheta = \lim_{n\to+\infty} \left[\sum_{k=\vartheta_0}^{n-1} \int_k^{k+1} \frac{(\alpha-1)\vartheta_0^{\alpha-1}}{\vartheta^\alpha} d\vartheta \right] =$$

$$= \lim_{n\to+\infty} \int_{\vartheta_0}^{n} \frac{(\alpha-1)\vartheta_0^{\alpha-1}}{\vartheta^\alpha} d\vartheta = \lim_{n\to+\infty} \left[-\vartheta_0^{\alpha-1} n^{1-\alpha} + 1 \right] = 1$$

Во-вторых, данное распределение является сопряжённым к дискретному равномерному распределению:

$$\mathbb{P}(\vartheta = k \mid x, \vartheta_0, \alpha) = C \cdot \mathbb{P}(x \mid \vartheta = k) \cdot \mathbb{P}(\vartheta = k \mid \vartheta_0, \alpha) = \mathbb{I}[1 \le x \le k] \cdot \mathbb{I}[\vartheta_0 \le k] \int_k^{k+1} \frac{C \cdot (\alpha-1)\vartheta_0^{\alpha-1}}{\vartheta^{\alpha+1}} d\vartheta =$$

$$= \mathbb{I}[\max\{x, \vartheta_0\} \le k] \int_k^{k+1} \frac{C \cdot (\alpha-1)\vartheta_0^{\alpha-1}}{\vartheta^{\alpha+1}} d\vartheta$$

Через условие нормировки вероятности легко показать, что $C = \dfrac{\alpha \cdot (\max\{x, \vartheta_0\})^\alpha}{(\alpha-1)\vartheta_0^{\alpha-1}}$. Таким образом, получаем:

$$\mathbb{P}(\vartheta = k \mid x, \vartheta_0, \alpha) = \mathbb{I}[\hat{\vartheta}_0 \le k] \int_k^{k+1} \frac{(\beta-1) \cdot \hat{\vartheta}_0^{\beta-1}}{\vartheta^\beta} d\vartheta$$

Это не что иное, как введённый выше дискретный аналог распределения Парето с параметрами $\beta = \alpha + 1,\ \hat{\vartheta}_0 = \max\{x, \vartheta_0\}$.

Для нашей задачи будет логичным выбрать $\vartheta_0 = 1$ как минимальное значение количества маршрутов, которые существуют, а также $\alpha = 2$, так как данное значение параметра определяет достаточно сбалансированную модель, при которой большие значения параметра маловероятны, но всё ещё возможны. То есть мы считаем, что в городе может быть большое количество маршрутов, но вероятность этого достаточно мала.

В качестве основной статистики для апостериорного распределения выберем медиану, так как она является наиболее робастной к выбросам, а также реалистично оценивает наш параметр, считая, что маршрутов может быть несколько больше, чем максимальный номер из тех, что был в наблюдениях. В дискретном случае медиана будет иметь аналогичный вид, но с округлением вверх до ближайшего целого числа. Запишем её вид, учитывая выбранные параметры модели.

$$\vartheta_{med} = \lceil 2^{\frac{1}{n}} \max\{1, x_{(n)}\} \rceil$$

Оценки параметра будут меняться следующим образом:

1. Первое наблюдение автобуса с номером 100 обновит оценку на искомый параметр по формуле $\max\{1, 100\} = 100$. То есть минимальное значение параметра теперь равно 100.

2. Наблюдение ещё двух автобусов с номерами 50 и 150 обновит оценку на искомый параметр по аналогичной формуле: $\max\{1, 100, 50, 150\} = 150$. Поэтому теперь максимальное количество маршрутов в городе будет не меньше 150.

3. Подсчёт медианы итогового апостериорного распределения на наш параметр даст следующее значение: $\vartheta_{med} = \lceil 2^{\frac{1}{3}} \cdot 150 \rceil = 189$

3 Задача 3

3.1 Условие

Записать распределение Парето с плотностью $Pareto(x \mid a, b) = \dfrac{ba^b}{x^{b+1}}\mathbb{I}[x \geq a]$ при фиксированном a в форме экспоненциального класса распределений. Найти $\mathbb{E}\left[\log(x)\right]$ путём дифференцирования нормировочной константы.

3.2 Решение

Известно, что плотность распределения, которое принадлежит экспоненциальному классу, имеет следующий вид: $p(x \mid \vartheta) = \dfrac{f(x)}{g(\vartheta)} \exp\left(\vartheta^T u(x)\right)$. Мы рассматриваем семейство распределений Парето с фиксированным параметром a. Запишем его в экспоненциальной форме:

$$\frac{ba^b\mathbb{I}[a \leq x]}{x^{b+1}} = ba^b\mathbb{I}[a \leq x]e^{-(b+1)\log(x)} = -\frac{\vartheta+1}{a^{\vartheta+1}}\mathbb{I}[a \leq x]e^{\vartheta\log(x)}$$

Здесь $\vartheta = -b-1$, $g(\vartheta) = -\dfrac{a^{\vartheta+1}}{\vartheta+1}$, $f(x) = \mathbb{I}[a \leq x]$, $u(x) = \log(x)$. Теперь найдём $\mathbb{E}\left[\log(x)\right]$ по следующей формуле: $\mathbb{E}\left[\log(x)\right] = \dfrac{\partial \log(g(\vartheta))}{\partial\vartheta}$

$$\frac{\partial \log(g(\vartheta))}{\partial\vartheta} = \frac{\partial}{\partial\vartheta}\Big[(\vartheta+1)\log(a) - \log(-\theta-1)\Big] = \log(a) + \frac{1}{-\vartheta-1} = \log(a) + \frac{1}{b}$$

Таким образом, мы получили, что $\mathbb{E}\left[\log(x)\right] = \log(a) + \dfrac{1}{b}$