Теоретическое задание 1

Королев Кирилл

10 октября 2023 г.

Задание 1

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

Запишем правдоподобие выборки и найдем оценку максимального правдоподобия.

$$p(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} \frac{\mathbb{I}_{[0,\theta]}(x_i)}{\theta} \to \max_{\theta}$$

Перейдем к логарифму.

$$l(\theta) = \log p(x_1, \ldots, x_n \mid \theta) = \sum_{i=1}^{n} \log \mathbb{I}_{[0,\theta]}(x_i) - n \log \theta$$

$$\frac{\partial l}{\partial \theta} = -\frac{n}{\theta} < 0$$

Производная меньше нуля, то есть $l(\theta)$ убывающая функция, то есть максимум будет достигаться при θ, которое принимает свое наименьшее допустимое значение. Чтобы логарифм от индикатора не взорвался, таким значением будет $\theta_{ML} = \max\{x_1, \ldots, x_n\} = x_{(n)}$.

Теперь найдем сопряженное распределение $p(\theta)$. Попробуем распределение Парето, его плотность выглядит как

$$p(\theta \mid \alpha, k) = \frac{k\alpha^k}{\theta^{k+1}} \mathbb{I}_{[\alpha,+\infty)}(\theta)$$

Проверим, что после умножения на правдоподобие получается тоже распределение Парето.

$$p(\theta \mid x_1, \ldots, x_n) \propto p(x_1, \ldots, x_n \mid \theta)\, p(\theta \mid \alpha, k) = \left(\prod_{i=1}^{n} \frac{\mathbb{I}_{[0,\theta]}(x_i)}{\theta} \right) \frac{k\alpha^k}{\theta^{k+1}} \mathbb{I}_{[\alpha,+\infty)}(\theta)$$

$$0 \le x_i \le \theta,\ \theta \in [\alpha, +\infty] \Rightarrow \theta \in [\max\{x_1, \ldots, x_n, \alpha\}, +\infty)$$

Тогда можно индикатор для x_i заменить на $\mathbb{I}_{\ge 0}(x_i)$, внести верхнюю границу на x_i в индикатор для θ как $\mathbb{I}_{[\max\{x_1,\ldots,x_n,\alpha\},+\infty)}(\theta)$. Тогда выражение примет следующий функциональный вид по θ.

$$\left(\prod_{i=1}^{n} \frac{\mathbb{I}_{[0,\theta]}(x_i)}{\theta} \right) \frac{k\alpha^k}{\theta^{k+1}} \mathbb{I}_{[\alpha,+\infty)}(\theta) \propto \frac{(k+n)\max\{x_1, \ldots, x_n, \alpha\}^{k+n}}{\theta^{k+n+1}} \mathbb{I}_{[\max\{x_1,\ldots,x_n,\alpha\},+\infty)}(\theta)$$

Тут я домножил и разделил на константы, которые не зависят от θ. Получили, что

$$\theta \mid x_1, \ldots, x_n \sim Pareto(\max\{x_1, \ldots, x_n, \alpha\}, k + n)$$

Найдем статистики апостериорного распределения. Обозначим $\max\{x_1, \ldots, x_n, \alpha\} := \beta$.

$$\mathbb{E}[\theta \mid x_1, \ldots, x_n] = \int x \frac{(k+n)\beta^{k+n}}{x^{k+n+1}} \mathbb{I}_{[\beta,+\infty)}(x)\, dx =$$

$$= (k+n)\beta^{k+n} \int_{\beta}^{+\infty} \frac{1}{x^{k+n}}\, dx = (k+n)\beta^{k+n} \frac{1}{(k+n-1)\beta^{k+n-1}} =$$

$$= \frac{k+n}{k+n-1}\beta = \frac{k+n}{k+n-1} \max\{x_1, \ldots, x_n, \alpha\}$$

Найдем медиану апостериорного распределения. Функция распределения выглядит как

$$F_{\theta|x_1,\dots,x_n}(x) = 1 - \left(\frac{\beta}{x}\right)^{k+n},\ x \geq \beta$$

$$F_{\theta|x_1,\dots,x_n}(x_{med}) = \frac{1}{2}$$

$$1 - \left(\frac{\beta}{x_{med}}\right)^{k+n} = \frac{1}{2}$$

$$x_{med} = \beta 2^{\frac{1}{k+n}} = \max\{x_1,\dots,x_n,\alpha\}2^{\frac{1}{k+n}}$$

Найдем моду апостериорного распределения. Так как плотность распределения Парето убывающая функция, то максимальное значение будет в точке $\max\{x_1,\dots,x_n,\alpha\}$.

Задание 2

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.

Пусть $x_i \sim U[0,\theta]$ — непрерывные номера автобусных маршрутов. В качестве априорного распределения возьмем распределение $Pareto(\theta \mid 100, 1)$. Его первый параметр отвечает за носитель и пересчитывается как максимум того, что мы видели, что хорошо подходит для нашей задачи, а второй параметр отвечает за количество экспериментов. Из предыдущего пункта знаем, что апостериорное тоже будет распределением Парето. Если хотим сэмплировать из распределения или делать какие-то оценки на θ, то можем брать целую часть от вещественного числа.

Мода хорошо подходит в качестве оценки, так как она выражается как максимум всех номеров, которые мы пронаблюдали. При увеличении количества экспериментов, если максимум не меняется, то распределение Парето становится более вырожденным в своей моде и таким образом мы более уверенны в количестве маршрутов. Ожидание не определено после одного эксперимента. Медиана дает некоторый запас на количество маршрутов, в частности, в начале в 2 раза, и при стремлении количества экспериментов к бесконечности превращается в моду, что логично. Таким образом, адекватно брать моду и медиану с округлением.

При наблюдении автобусов с номерами 50 и 150 апостериорное распределение пересчитается как

$$\theta \mid x_1 = 50, x_2 = 150 \sim Pareto(\theta \mid \max\{100, 50, 150\}, 3)$$

Мода станет равной 150, а округленная медиана $[150 * 2^{\frac{1}{3}}] = 189$. То есть при выборе моды оценка может только не уменьшаться, при выборе медианы оценка может как увеличиться, так и уменьшиться. Кажется, стоит выбирать оценку в зависимости от потребностей в переоценивании или недооценивании.

Задание 3

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

Так как a фиксированно, хотим записать $p(x \mid a, b)$ в виде $\dfrac{f(x)}{g(b)}\exp\left(bu(x)\right)$.

$$Pareto(x \mid a, b) = \frac{ba^b}{x^{b+1}}[x \geq a] = ba^b\frac{[x \geq a]}{x}\exp\left(-\log x^b\right) = ba^b\frac{[x \geq a]}{x}\exp\left(-b\log x\right)$$

$$f(x) = \frac{[x \geq a]}{x},\ g(b) = \frac{1}{ba^b},\ u(x) = -\log x$$

На лекции доказали следующий факт.

$$\frac{\partial \log g(b)}{b} = \mathbb{E}[u(x)] = \mathbb{E}[-\log x]$$

$$\log g(b) = \log \frac{1}{ba^b} = -\log b - b \log a$$

$$\frac{\partial \log g(b)}{b} = -\frac{1}{b} - \log a$$

Таким образом, $\mathbb{E} \log x = \frac{1}{b} + \log a$.