Домашняя работа

Котов Артем, МОиАД ВШЭ СПб

27 октября 2021 г.

Содержание

Task 1

Условие: 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 $p(\theta|x_1, \ldots, x_n)$ и вычислить его статистики: мат.ожидание, медиану и моду. Формулы для статистик нужно вывести, а не взять готовые. *Подсказка: задействовать распределение Парето.*

Max. likelihood

Решение.

$$L(\theta) = p(X, \theta) = [\text{независимость}] = \prod_{i=1}^{n} p(x_i|\theta)$$

$$\theta_{ML} = \operatorname*{argmax}_{\theta} L(\theta) = \cdots = \operatorname*{argmax}_{\theta} \sum_{i=1}^{n} \log p(x_i|\theta)$$

Из условия мы знаем, что $p(x_i|\theta) = \frac{1}{\theta}$, тогда

$$\theta_{ML} = \operatorname*{argmax}_{\theta}(-n \log \theta)$$

Видно, что в аргументе стоит убывающая функция, следовательно, θ для максимизирования правдоподобия должна быть наименьшей из возможных. Так как мы пронаблюдали какие-то значения X, то наименьшей из возможных будет $\max X = x_{(n)}$. ■

Сопряженное к равномерному

Решение.

Рассмотрим $p(\theta|\alpha, \beta) = \frac{\alpha\beta^\alpha}{\theta^{\alpha+1}}[\beta \leqslant \theta]$ — распределение Парето. Покажем, что апостериорное так же будет иметь такой же вид, но с другими $\tilde{\alpha}, \tilde{\beta}$.

Рассмотим $p(\theta|X) = \frac{p(X|\theta)p(\theta|\alpha,\beta)}{\int_0^{+\infty} p(X|\theta)p(\theta|\alpha,\beta)d\theta}$.

Сначала разберемся с нормировочным интегралом:

$$\int_0^{+\infty} p(X|\theta)p(\theta|\alpha, \beta)d\theta = (*)$$

Здесь возникнет произведение двух индикаторных функций: $[\beta \leqslant \theta][x_{(n)} \leqslant \theta]$, что можно переписать, как $[m \leqslant \theta]$, где $m = \max(\beta, x_{(n)})$, тогда

$$(*) = \alpha\beta^\alpha \int_m^{\infty} \theta^{-n-\alpha-1} d\theta = (**)$$

Если $-n-\alpha-1 \geqslant -1$, то интеграл расходится, т.е. необходимо, чтобы $n+\alpha<0$

$$(**) = \frac{\alpha\beta^\alpha}{-n-\alpha}\theta^{-n-\alpha}\Big|_m^{+\infty} = \frac{\alpha\beta^\alpha}{(n+\alpha)m^{n+\alpha}}$$

Таким образом,

$$p(\theta|X) = (n+\alpha)m^{n+\alpha}\theta^{-n-\alpha-1}[m \leqslant \theta] = (***),$$

где $m = \max(\beta, x_{(n)})$.

Если $\tilde{\alpha} = \alpha + n$, а $\tilde{\beta} = m$, то

$$(***) = \frac{\tilde{\alpha}\tilde{\beta}^{\tilde{\alpha}}}{\theta^{\tilde{\alpha}+1}}[\tilde{\beta} \leqslant \theta],$$

что есть распределение Парето, т.е. оно действительно является сопряженным к равномерному. ■

Статистики

Мат. ожидание

Решение.

$$\mu = \mathbb{E}\theta = \int_0^{+\infty} \theta p(\theta|X)d\theta = (n+\alpha)m^{n+\alpha}\int_m^{+\infty}\theta^{-n-\alpha}d\theta$$
$$= [\tilde{\alpha} = n+\alpha] = \tilde{\alpha}m^{\tilde{\alpha}}\int_m^{+\infty}\theta^{-\tilde{\alpha}}d\theta = \tilde{\alpha}m^{\tilde{\alpha}}\frac{1}{\tilde{\alpha}-1}m^{-\tilde{\alpha}+1} = \frac{\tilde{\alpha}m}{\tilde{\alpha}-1}$$

■

Медиана

Решение.

c — медиана, причем $c \geqslant m$

$$P(c \leqslant \theta < +\infty) = (n+\alpha)m^{n+\alpha}\int_c^{+\infty}\theta^{-n-\alpha-1}d\theta = \left(\frac{m}{c}\right)^{n+\alpha} = \frac{1}{2}$$
$$\Downarrow$$
$$(n+\alpha)(\ln m - \ln c) = -\ln 2 \Longrightarrow c = \exp(\ln m + \frac{\ln 2}{n+\alpha}) = m2^{\frac{1}{n+\alpha}}$$

■

Мода

Решение.

$$\operatorname*{argmax}_\theta p(\theta|X) = \operatorname*{argmax}_\theta \theta^{-n-\alpha-1}[m \leqslant \theta]$$

Тут также возникает убывающая функция от θ, следовательно, берем наименьшее доступное, т.е. $\theta = m$, где $m = \max(\beta, x_{(n)})$ ■

Task 2: автобусы

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.*

Решение.

Номера автобусов — дискретная величина. Будем моделировать их непрерывным распределением $U[0, \theta]$ так, что под номером будет понимать $\lceil x \rceil$, где $x \sim U[0, \theta]$, т.е. целая часть сверху случайной величины. Этот выбор кажется разумным, так как у нас нет знания о частоте хождения автобусов или отношения кол-во автобусов на маршруте к общему числу автобусов, т.е. в такой модели мы предполагаем, что можем наблюдать любой номер автобуса равновероятно.

Нам надо оценить θ, которая может быть, в принципе, любым положительным целым числом. Здесь сделаем еще одно предположение: номера автобусов считаем заданными подряд с 1, т.е. если у нас всего 3 автобуса в атвопарке, то их номера будут $1, 2, 3$, а не $1, 5, 10$ или $4, 5, 6$, так сказать, предположение здравого смысла.

В качестве априорного распределения возьмем распределение Парето:

$$p(\theta|\alpha, \beta) = \frac{\alpha\beta^\alpha}{\theta^{\alpha+1}}[\theta \geqslant \beta],$$

причем в качестве α и β можно выбрать исходя их каких-то знаний об автопарке города. Так, например, если мы пронаблюдали автобус с номером 100, то $\beta = 100$, если изначально $\beta < 100$, т.к. мы точно знаем, что такой номер автобуса существует. Для дальнейшего будем считать, что $\beta = 100$.

Теперь мы наблюдаем еще один автобус, но с номером 150. Тогда β еще подправится и станет равной 150, что значит, что значения $\theta > 150$ теперь тоже стали более вероятными. Однако, если мы пронаблюдали еще и автобус с номером 50, то распределение деформируется так, что небольшие θ станут более вероятными, а большие наоборот — менее, это хорошо видно из следующей картинки, т.к. в этом случае у нас просто увеличилась α на 1:

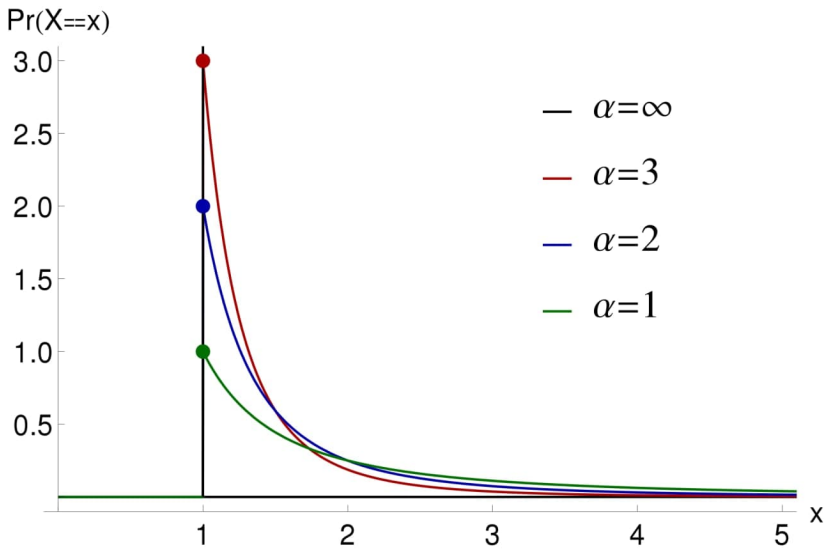

Рис. 1: Распределения Парето с различными α при $\beta = 1$

Видно, что хвост поджимается при увеличении α

Как мы видели из предыдущей задачи, мода распределения есть β. В целом, норм статистика, но понятно, что грубая, так, например, если у нас действительно всего 150 маршрутов, то все окей, но это все равно, что оценивать количество маршрутов по наибольшему наблюдаемому номеру автобуса...

Посмотрим на среднее и медиану:

$$\langle \theta \rangle = \frac{\alpha\beta}{\alpha - 1}$$
$$\theta_m = \beta \sqrt[\alpha]{2}$$

Обе эти статистики завысят оценку θ по сравнению с модой, к тому же здесь используется наше априорное знание, в отличии от моды. Какое именно выбрать из эти двух? Сложно сказать. По классике я бы выбирал среднее, но против медианы ничего, на первый взгляд, сказать не могу противного.

■

Task 3: Pareto as exp.

Условие: Записать распределение Парето с плотностью $p(x|\alpha,\beta) = \frac{\alpha\beta^\alpha}{x^{\alpha+1}}[x \geqslant \beta]$ при фиксированном β в форме экспоненциального класса распределений. Найти $\mathbb{E}\ln x$ путём дифференцирования нормировочной константы.

Решение.

Приведем распределение Парето к экспоненциальному виду:

$$p(x|\alpha,\beta) = \frac{\alpha\beta^\alpha}{x^{\alpha+1}}[x \geqslant \beta], \quad \beta - \text{fixed.}$$

$$p(x|\alpha) = \frac{f(x)}{g(\alpha)}\exp(\alpha u(x))$$

$$p(x|\alpha,\beta) = \frac{\alpha\beta^\alpha}{x}\exp(-\alpha\ln x)[x \geqslant \beta]$$

$$\Downarrow$$

$$f(x) = \frac{[x \geqslant \beta]}{x}; \quad g(\alpha) = \frac{1}{\alpha\beta^\alpha}; \quad u(x) = -\ln x$$

Теперь разберемся с матожиданием:

$$\mathbb{E}\ln x = \frac{\partial g(\alpha)}{\partial\alpha} = -\frac{1}{\alpha^2\beta^{2\alpha}}\left(\beta^\alpha + \alpha\beta^\alpha\ln\beta\right) = -\frac{1+\alpha\ln\beta}{\alpha^2\beta}$$

■