Теория 1. Сопряжённые распределения и экспоненциальный класс распределений

Курс: Байесовские методы в машинном обучении
Студент: Никулин Александр

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

   Вначале заметим, что θ ограничена снизу $x_{(N)}$. Тогда, для $x_1 \geq 0$, $x_n \leq \theta$:

   $$\theta_{ML} = \underset{\theta}{\operatorname{argmax}}\ L(\theta) = \underset{\theta}{\operatorname{argmax}}\ \prod_{i=1}^{n} p(x_i|\theta) = \underset{\theta}{\operatorname{argmax}}\ \frac{1}{\theta^n}$$

   Функция убывающая, поэтому для максимума нужно положить θ минимальным, то есть $\theta_{ML} = x_{(N)}$.

   Сопряженным распределением для равномерного является распределение Парето: $\theta \sim \mathrm{Pareto}(\theta|a, b)$. Найдем апостериорное распределение при $m = \max(x_{(N)}, a)$:

   $$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{\int_m^\infty p(X|\theta)p(\theta)d\theta} = \frac{\theta^{-n-b-1}[\theta \geq m]}{\frac{1}{b+n}m^{-b-n}} = \frac{(b+n)m^{b+n}}{\theta^{(b+n)+1}}[\theta \geq m] = \mathrm{Pareto}(\theta|m, b+n)$$

   Получившееся апостериорное можно легко проинтерпретировать: для каждого нового наблюдения, мы увеличиваем счетчик наблюдений, а также, если нужно, расширяем границу максимального возможного θ.
   Найдем необходимые статистики. Для этого достаточно найти их для Парето, после чего подставить обновленные параметры распределения:

   (a) $\mathbb{E}[\theta] = \int_a^\infty \theta p(\theta)\ d\theta = ba^b \int_a^\infty \theta\theta^{-b-1} = ba^b \int_a^\infty \theta^{-b} = ba^b[\frac{\theta^{-b+1}}{b-1}]\Big|_a^\infty = ba^b \frac{a^{-b+1}}{b-1} = \frac{ba}{b-1}, b \geq 1$

   (b) Мода - максимум (все, если их несколько) плотности. Возьмем производную от плотности: $(ba^b\theta^{-b-1})' = -(b+1)ba^b\theta^{-b-2}$. Видим, что функция убывающая, поэтому максимум достигается при минимальной θ, то есть в a.

   (c) Медиана - $p(\theta \leq m) = \frac{1}{2} = p(\theta \geq m)$. Найдем $p(\theta \leq m)$:

   $$p(\theta \leq m) = ba^b \int_a^m \theta^{-b-1} d\theta = -ba^b \frac{\theta^{-b}}{b}\Big|_a^m = 1 - a^b m^{-b}$$

   Приравняем к $\frac{1}{2}$:

   $$\begin{aligned} 1 - a^b m^{-b} &= \frac{1}{2} \Rightarrow \\ b(\log a - \log b) &= -\log 2 \Rightarrow \\ \log m &= \frac{\log 2}{b} - \log a \Rightarrow \\ m &= a2^{\frac{1}{b}} \end{aligned}$$

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.

Кажется логичным определить распределение номеров маршрутов как равномерное на $[0, \theta]$. Минимальным номером маршрута может быть только 0, а дальше при добавлении новых маршрутов максимальное значение увеличивается. Можно также допустить, что марштруты равномерно покрывают разные участка города и не сильно пересекаются, иначе при проезде на определенных местах можно будет наблюдать одни номера сильно чаще, чем другие, то есть не равновероятно. Мы уже знаем сопряженное распределение для априорного - Парето. Т.к. город новый и мы не можем даже приблизительно оценить его население, кол-во улиц и т.д., параметры можно инициализировать как $a = 1, b = 1$ (*интересно, можно ли их также подобрать по принципу наибольшей обоснованности на трех наблюдаемых точках? $p(X)$ зависит от a, b, а значит они влияют на обносноваnность*). Парето также формализует наше априорное знание, что больших городов сильно меньше маленьких, а значит вероятность маленьких номеров сильно больше, чем больших. Для существования мат. ожидания b должен быть больше 1, что будет верно при первом наблюдении.
При наблюдении маршрута с номером 100 мы переноценим параметры: $a = 100, b = 2$. Аналогично для последующих маршрутов с номерами 150 и 50. Для наглядности можно свести это в табличку (формулы статистик мы вывели ранее). Но прежде можно ответить, какая из статистик наиболее адекватная для оценки количества номеров. Для этого будет полезным вывести честное мат. ожидание максимального номера $m = \max(X_1, X_2, ..., X_n)$, при фиксированном θ и размере выборки n: $m|\theta$.

$$p(m \le x|\theta) = P(X_1 \le x, X_2 \le x, ..., X_n \le x|\theta) = \prod_{i=1}^{n} P(X_i \le x) = F_{X|\theta}(x)^n = \begin{cases} 0, & x \le 0 \\ 1, & x \ge \theta \\ (\frac{x}{\theta})^n, & x \in [0, \theta] \end{cases}$$

$$p(m|\theta) = \frac{dF_{X|\theta}(x)}{dx} = \frac{n}{\theta}(\frac{m}{\theta})^{n-1}$$

$$\mathbb{E}[m|\theta] = \int_0^\theta n(\frac{m}{\theta})^n \ dm = \frac{n\theta}{n+1}$$

Соберем в табличку оценки при поступлении новых данных (статистики округлены). Предположим, что реальная θ равна 300 (город маленький).

X	a	b	Мода	Медиана	Мат. ожидание	Мат. ожидание максимума
100	100	2	100	141	200	200
100, 150	150	3	150	189	225	225
100, 150, 50	150	4	150	178	200	240

Все статистики имеют тененцию недооценивать реальное значение, однако мат. ожидание дает наиболее точную оценку.

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

$$\text{Pareto}(x|a,b) = \frac{ba^b}{x^{b+1}}[x \ge a] = \frac{ba^b}{x}\exp(-b\log x)[x \ge a] = \frac{f(x)}{g(b)}\exp(bu(x))$$

Получили форму экспоненциального класса при $f(x) = \frac{[x \ge a]}{x}, g(b) = \frac{1}{ba^b}, u(x) = -\log x$. Теперь найдем мат. ожидание:

$$\frac{dg(b)}{db} = \mathbb{E}[u(x)] = -\mathbb{E}[\log x] \Rightarrow$$

$$\mathbb{E}[\log x] = -\frac{dg(b)}{db} = \frac{d\frac{1}{ba^b}}{db} = \frac{a^{-b}(b\log a + 1)}{b} = \frac{b\log a + 1}{a^b b^2}.$$