Теория 1. Сопряжённые распределения и экспоненциальный класс распределений

Курс: Байесовские методы в машинном обучении, 2021

1. Пусть $x_1, x_2, \ldots, x_N$ – 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 $p(\theta|x_1, \ldots, x_N)$ и вычислить его статистики: мат.ожидание, медиану и моду. Формулы для статистик нужно вывести, а не взять готовые. *Подсказка: задействовать распределение Парето.*

   Решение:

   (a) θ_{ML}:
   Здесь и далее X - выборка.
   Правдоподобие: $p(X|\theta) = \prod\limits_{i=1}^{N} p(x_i|\theta) = \prod\limits_{i=1}^{N} \theta^{-1}\mathbb{I}_{\{0 \leqslant x_i \leqslant \theta\}} = \theta^{-N}\mathbb{I}_{\{0 \leqslant X_{(1)}, X_{(N)} \leqslant \theta\}}$ - убывающая с ростом θ функция. Принимает максимальное значение при минимально возможном значении $\theta = X_{(N)}$ - левая граница θ. Поэтому оценкой θ_{ML} будет $X_{(N)}$.

   (b) $p(\theta)$:
   Правдоподобие - степенная функция от θ. Значит, априорное распределение $p(\theta)$ тоже должно быть степенной функцией. Распределение Парето подходит под это требование. Возьмем $\theta \sim p(\theta) = \frac{ba^b}{\theta^{b+1}}\mathbb{I}_{\{\theta \geqslant a\}}$ - такое априорное распределение θ будет сопряженным к правдоподобию (к семейству функций правдоподобия).

   (c) $p(\theta|x_1, \ldots, x_N)$:
   $p(\theta|X) = \frac{p(X|\theta)p(\theta)}{C} = \frac{1}{C}\frac{1}{\theta^N}\frac{ba^b}{\theta^{b+1}}\mathbb{I}_{\{\theta \geqslant X_{(N)}, \theta \geqslant a\}} = \frac{ba^b}{C\theta^{b+N+1}}\mathbb{I}_{\{\theta \geqslant max(X_{(N)}, a)\}} = \frac{\hat{b}\hat{a}^{\hat{b}}}{\theta^{\hat{b}+1}}\mathbb{I}_{\{\theta \geqslant \hat{a}\}}$, где $\hat{a} = max(X_{(N)}, a)$, $\hat{b} = b + N$, а C - нормировочная константа - отвечает за оставшуюся часть равенства.

   (d) $\mathbb{E}[\theta|X]$:
   Мат.ожидание для распределения Парето:
   $\int\limits_a^{\infty} \frac{ba^b}{\theta^{b+1}}\theta\, d\theta = ba^b \int\limits_a^{\infty} \theta^{-b}\, d\theta = ba^b \frac{\theta^{-b+1}}{-b+1}\Big|_a^{\infty} = ba^b\frac{a^{-b+1}}{b+1} = \frac{ba}{b-1}$ при $b > 1$
   Значит в нашем случае, $\mathbb{E}[\theta|X] = \frac{(b+N)max(X_{(N)}, a)}{b+N-1}$ при $N > 1$

   (e) мода:
   Так как распределение Парето - степенная функция (с отрицательной степенью), ее максимальное значение достигается при минимальном значении аргумента - то есть мода данного распределения это $\hat{a} = max(X_{(N)}, a)$.

   (f) медиана:
   По определению медиана m: $F(m) = 0.5$.
   Для распределения Парето $F(x) = \int\limits_a^x \frac{ba^b}{\theta^{b+1}}\, d\theta = ba^b\frac{\theta^{-b}}{-b}\Big|_a^x = a^b\theta^{-b}\Big|_x^a = 1 - (\frac{a}{x})^b$
   Значит, медиана $m : F(m) = 1 - (\frac{a}{m})^b = 0.5 \Rightarrow \frac{a}{m} = \sqrt[b]{0.5} \Rightarrow m = a\sqrt[b]{2}$
   Значит, в нашем случае, медиана $= max(X_{(N)}, a)\ \sqrt[b+N]{2}$

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.*

Решение:
Заметим, что распределение городов по численности населения является распределением Парето. Почему? Потому что основной принцип распределения Парето - 20% популяции владеет 80% богатства - это соотношение явно прослеживается в ситуации с городами:

"В настоящее время в России насчитывается более тысячи городов. Большую часть из них составляют малые города с населением до 50 тыс. человек. В городах-миллионниках (16 городов) проживает почти четверть населения страны". - Росстат.

Принцип Парето вообще наблюдается там, где сила притяжения к объекту (такому как массивная звезда или большой город) пропорциональна его массе. А поскольку между числом жителей города и числом автобусных маршрутов в нем есть прямая связь, то распределение городов по числу автобусных маршрутов - это тоже распределение Парето.
Поэтому примем априорное распределение $p(\theta|a,b) = \text{Pareto}(\theta|a,b) = \frac{ba^b}{\theta^{b+1}}\mathbb{I}_{\{\theta \geqslant a\}}$. Параметр a отвечает минимальному значению θ и одновременно наиболее вероятному значению θ, а b - это коэффициент масштаба. Поэтому в нашем случае положим $a = 1, b = 1$.

Предположим, что все маршруты от 1 до θ равновероятны (дискретное равномерное распределение). Тогда наблюдать автобус с номером $x_1 = 100$ можно с вероятностью $p(X_1 = x_1|\theta) = \theta^{-1}\mathbb{I}_{\{1 \leqslant x_1 \leqslant \theta\}}$.
Тогда апостериорное распределение $p(\theta|X) = \frac{p(X_1=100|\theta)p(\theta)}{C} = \frac{\hat{b}\hat{a}^{\hat{b}}}{\theta^{\hat{b}+1}}\mathbb{I}_{\{\theta \geqslant \hat{a}\}}$, где $\hat{a} = max(100, 1) = 100$, $\hat{b} = b + 1 = 2$.
Если увидеть автобусы 50 и 150 дополнительно, то параметры прогнозируемого распределения будут следующими: $\hat{a} = max(50, 100, 150, 1) = 150$, $\hat{b} = b + 3 = 4$.
Выберем среди оценок: мат.ожидание $= \frac{4 \cdot 150}{3} = 200$, мода $= 150$, медиана $= 150\sqrt[4]{2} \sim 179$. Мат.ожидание Парето - дело непредсказуемо безграничное, мода всегда выдает наименьшие параметр, поэтому, вероятно, лучшую оценкц даст медиана.

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

Решение:
$\text{Pareto}(x|a,b) = \frac{ba^b}{x^{b+1}}\mathbb{I}_{\{x \geqslant a\}} = \frac{x^{-1}\mathbb{I}_{\{x \geqslant a\}}}{(ba^b)^{-1}}e^{-b\ln x} = \frac{f(x)}{g(\theta)}e^{\theta^T u(x)}$, где
$\theta = (\theta_1, \theta_2) = (a, b), \quad f(x) = x^{-1}\mathbb{I}_{\{x \geqslant \theta_1\}}, \quad g(\theta) = (\theta_2\theta_1^{\theta_2})^{-1}, \quad u(x) = (0, -\ln x)$
Таким образом, распределение Парето принадлежит экспоненциальному классу распределений. Тогда по свойству экспоненциального класса:
$\frac{\partial \ln g(\theta)}{\partial \theta_i} = \mathbb{E}u_i(x)$
Значит, $\mathbb{E}\ln(x) = -\mathbb{E}[-\ln(x)] = -\frac{\partial \ln g(\theta)}{\partial \theta_2} = -\frac{\partial \ln((\theta_2\theta_1^{\theta_2})^{-1})}{\partial \theta_2} = \frac{\partial \ln(\theta_2\theta_1^{\theta_2})}{\partial \theta_2} = \frac{\partial(\ln\theta_2 + \theta_2\ln\theta_1)}{\partial \theta_2} = \theta_2^{-1} + \ln\theta_1$
Значит, $\mathbb{E}\ln(x) = b^{-1} + \ln a$