Теоретическое задание №1
«Сопряженные распределения и экспоненциальный класс распределений»

курс «Байесовские методы в машинном обучении»
кафедра ММП ВМК МГУ

Михеев Борис, 417 группа

10 октября 2022 г.

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

Запишем оценку максимального правдоподобия θ_{ML}. Т. к. $x_1, \ldots, x_N \sim U[0, \theta]$ – независимая выборка, и $p(x_i|\theta) = \frac{[0 \le x_i \le \theta]}{\theta}$ $\forall i = \overline{1, N}$, то правдоподобие выборки можно представить в виде произведения плотностей для всех x_i. Тогда:

$$\theta_{ML} = \underset{\theta}{\operatorname{argmax}}\, p(x_1, \ldots, x_N|\theta) = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} p(x_i|\theta) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \log p(x_i|\theta) =$$

$$= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \log(\frac{[0 \le x_i \le \theta]}{\theta}) = \underset{\theta}{\operatorname{argmax}}(-N \log \theta).$$

Функция $-N \log \theta$ является монотонно убывающей, следовательно, ее максимум будет достигаться при минимальном допустимом значении параметра θ. Т. к. $x_1, \ldots, x_N \sim U[0, \theta]$, т. е. $0 \le x_i \le \theta$ $\forall i = \overline{1, N}$, то $\theta \ge x_{(N)} = \max\limits_{i=\overline{1,N}} x_i$ (чтобы вся выборка попадала в интервал). Т. о., $\theta_{ML} = x_{(N)} = \max\limits_{i=\overline{1,N}} x_i$.

Перейдем к поиску сопряженного распределения $p(\theta)$. Воспользуемся подсказкой из условия задачи и рассмотрим распределение Парето:

$$p(\theta) = Pareto(\theta|a, b) = \frac{ba^b}{\theta^{b+1}}[\theta \ge a].$$

Найдем апостериорное распределение по теореме Байеса и рассмотрим, будет ли оно того же вида, что и $p(\theta)$:

$$p(\theta|x_1,\dots,x_N) = \frac{p(x_1,\dots,x_N|\theta)p(\theta)}{\int p(x_1,\dots,x_N|\theta)p(\theta)d\theta} = \frac{[0 \le x_{(1)} \le x_{(N)} \le \theta]ba^b[\theta \ge a]}{\theta^{b+1+N}\int \frac{[0\le x_{(1)}\le x_{(N)}\le\theta]ba^b[\theta\ge a]}{\theta^{b+1+N}}d\theta} =$$
$$= \frac{[\max\{x_{(N)},a\} \le \theta]}{\theta^{b+1+N}\int_{\max\{x_{(N)},a\}}^{+\infty}\theta^{-(b+1+N)}d\theta} = \frac{[\max\{x_{(N)},a\} \le \theta]}{\theta^{b+1+N}(-\frac{1}{(b+N)\theta^{b+N}}|_{\max\{x_{(N)},a\}}^{+\infty})} =$$
$$= \frac{(b+N)(\max\{x_{(N)},a\})^{b+N}}{\theta^{b+1+N}}[\max\{x_{(N)},a\} \le \theta].$$

Отметим, что при вычислениях воспользовались следующими преобразованиями индикаторов: $[0 \le x_{(1)} \le x_{(N)} \le \theta] = [0 \le x_{(1)}][x_{(1)} \le x_{(N)}][x_{(N)} \le \theta]$. Тогда в соответствующей дроби индикаторы $[0 \le x_{(1)}]$ и $[x_{(1)} \le x_{(N)}]$ сократятся (они не зависят от θ), а оставшиеся индикаторы можно преобразовать как $[x_{(N)} \le \theta][\theta \ge a] = [\max\{x_{(N)},a\} \le \theta]$.

В итоге получили апостериорное распределение, имеющее вид $Pareto(\theta|\max\{x_{(N)},a\}, b+N)$. Вычислим его статистики:

$$\mathbb{E}(\theta|x_1,\dots,x_N) = \int_{-\infty}^{+\infty} p(\theta|x_1,\dots,x_N)\theta d\theta = \int_{\max\{x_{(N)},a\}}^{+\infty} \frac{(b+N)(\max\{x_{(N)},a\})^{b+N}}{\theta^{b+N}}d\theta =$$
$$= (b+N)(\max\{x_{(N)},a\})^{b+N}(-\frac{1}{(b+N-1)\theta^{b+N-1}})\Big|_{\max\{x_{(N)},a\}}^{+\infty} = \frac{(b+N)\max\{x_{(N)},a\}}{(b+N-1)}.$$

Перейдем к нахождению медианы. В случае непрерывного распределения это можно сделать путем решения уравнения $1 - F_\theta(x) = \mathbb{P}(\theta \ge x) = \frac{1}{2}$. Тогда:

$$1 - F_\theta(x) = \int_x^{+\infty} p(\theta|x_1,\dots,x_N)d\theta = \int_x^{+\infty} \frac{(b+N)(\max\{x_{(N)},a\})^{b+N}}{\theta^{b+1+N}}d\theta =$$
$$= \frac{(\max\{x_{(N)},a\})^{b+N}}{x^{b+N}} = \frac{1}{2} \Rightarrow x = \max\{x_{(N)},a\}\cdot 2^{\frac{1}{b+N}} = median\ \theta.$$

Моду найдем как наиболее вероятное значение θ, т. е. точку максимума плотности апостериорного распределения:

$$mode\ \theta = \underset{\theta}{\operatorname{argmax}}\ p(\theta|x_1,\dots,x_N) = \underset{\theta}{\operatorname{argmax}}\ \frac{(b+N)(\max\{x_{(N)},a\})^{b+N}}{\theta^{b+N+1}}[\max\{x_{(N)},a\} \le \theta].$$

$p(\theta|x_1,\dots,x_N)$ убывает по θ, следовательно, ее максимум достигается при минимально допустимом значении θ. С учетом индикатора в формуле распределения получим:

$$mode\ \theta = \max\{x_{(N)},a\}.$$

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.*

Воспользуемся подсказкой из условия задачи и обратим внимание на предыдущую задачу, воспользуемся аналогиями. В целом можно принять, что номера автобусов распределены равномерно и непрерывно, и пусть при этом сами номера берутся с округлением до ближайшего целого (так можно «адаптировать» непрерывное распределение к дискретным объектам (номерам автобусов)). Такое предположение можно назвать логичным, т. к. при прочих равных автобусы с разными номерами могут прибыть с одинаковой вероятностью, а также среди номеров нет иерархии и четкого порядка, они «равноправны». Также будет удобно перевести номера в шкалу от 0 до собственно количества автобусных номеров в городе. Это всегда можно сделать, и при этом оценка наибольшего номера будет оценкой их общего числа. Т. о., будем считать, что номера $x_1, \ldots, x_N \sim U[0, \theta]$. При этом не нарушится условие нормировки вероятностей, т. к. дискретным значениям по сути будут соответствовать интервалы, дающие в объединении $[0, \theta]$, и вероятности дадут в сумме 1. Требуется с помощью байесовского подхода оценить θ – параметр данного распределения и число автобусных маршрутов в городе.

По аналогии с предыдущей задачей, в качестве априорного распределения параметров можно взять распределение Парето:

$$p(\theta) = \frac{ba^b}{\theta^{b+1}}[\theta \geq a].$$

Тогда по результатам предыдущей задачи получим апостериорное распределение вида:

$$p(\theta|x_1, \ldots, x_N) = Pareto(\theta|\max\{x_{(N)}, a\}, b + N) = \frac{(b+N)(\max\{x_{(N)}, a\})^{b+N}}{\theta^{b+1+N}}.$$

Рассмотрим его и попробуем интуитивно проинтерпретировать его параметры применительно к решаемой задаче. Оцениваемый параметр θ – число номеров автобусов в городе и верхняя граница интервала равномерного непрерывного распределения. По формуле распределения Парето $p(\theta|x_1, \ldots, x_N)$ можно предположить, что его первый параметр (в случае общей формулы вида $p(\theta) = \frac{ba^b}{\theta^{b+1}}[\theta \geq a]$ параметр a) играет роль минимального допустимого значения θ. В полученной формуле апостериорного распределения во втором параметре распределения Парето (b в случае общей формулы, $b + N$ для апостериорного распределения) фигурирует N – общее число наблюдений. Возникает предположение, что данный параметр связан с количеством наблюдений. По сути он является показателем степени, и т. о. определяет форму распределения.

Исходя из рассмотренных предположений, для априорного распределения кажется приемлемым и логичным взять параметры $a = 100$ как единственный и пока что наибольший наблюденный номер автобуса, и $b = 1$, т. к. видели всего один автобус. Можно и сказать, что выбранные значения параметров и их интерпретация согласуются с рассматриваемой вероятностной моделью и полученными формулами различных распределений, т. к. с поступлением дальнейших наблюдений и пересчетом $p(\theta|x_1, \ldots, x_N) = Pareto(\theta|\max\{x_{(N)}, a\}, b + N)$ число наблюдений будет линейно расти, а верхняя граница номеров θ как раз будет сдвигаться в

большую сторону, если среди новых пришедших наблюдений есть номер, превосходящий старую границу.

Далее рассмотрим статистики апостериорного распределения. По результатам предыдущей задачи имеем:

$$\mathbb{E}(\theta|x_1,\dots,x_N) = \frac{(b+N)\max\{x_{(N)},a\}}{(b+N-1)}.$$

$$median\ \theta = \max\{x_{(N)},a\}\cdot 2^{\frac{1}{b+N}}.$$

$$mode\ \theta = \max\{x_{(N)},a\}.$$

Данные статистики в принципе также можно округлить при желании и необходимости иметь дискретные оценки дискретных номеров. Заметим, что матожидание и медиана могут быть представлены как функции от моды в виде:

$$\mathbb{E}(\theta|x_1,\dots,x_N) = f_1(c)\cdot mode\ \theta = \frac{c\cdot mode\ \theta}{c-1},$$

$$median\ \theta = f_2(c)\cdot mode\ \theta = mode\ \theta\cdot 2^{\frac{1}{c}},$$

где $c = b+N$, $c\geq 2$, т. к. приняли $b=1$, и $N\geq 1$, $f_1(c) = \frac{c}{c-1}$, $f_2(c) = 2^{\frac{1}{c}}$.

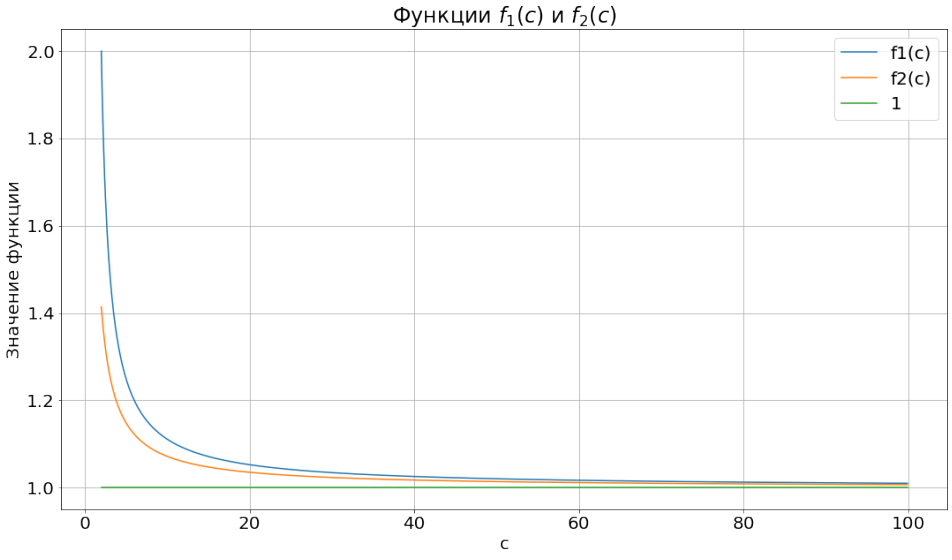

Рис. 1: Функции $f_1(c)$ и $f_2(c)$.

Заметим, что $f_1(c)\geq f_2(c)\geq 1\ \forall c\geq 2 \Rightarrow \mathbb{E}(\theta|x_1,\dots,x_N)\geq median\ \theta\geq mode\ \theta$. В принципе логичным кажется брать наибольшую возможную оценку числа автобусных маршрутов, чтобы наблюдаемые номера гарантированно не превосходили верхнюю границу. Однако такая оценка может быть недостаточно точной. Оценка модой при этом является оценкой максимума апостериорной вероятности, но по сути является наибольшим из увиденных номеров

(т. к. получили $mode\ \theta = \max\{x_{(N)}, a\}$), что кажется не слишком точной и грубой оценкой. Оценка медианой выглядит нечтом средним между матожиданием и модой, судя по полученным формулам. В целом наиболее адекватной, логичной и интуитивной оценкой видится матожидание.

Теперь рассмотрим, как будут меняться данные оценки по мере поступления новых наблюдений номеров. Начнем с первого автобуса под номером 100. Параметры априорного распределения $p(\theta)$ зададим как $a = 100$, $b = 1$, как было принято ранее. Для рассчетов используем полученные формулы.

- 1. $x_1 = 100$, $N = 1$, $\max\{x_{(N)}, a\} = 100$.

  $$\mathbb{E}(\theta|x_1) = \frac{2}{1} \cdot 100 = 200.$$

  $$median\ \theta = 2^{\frac{1}{2}} \cdot 100 \approx 141.$$

  $$mode\ \theta = 100.$$

- 2. $x_2 = 50$, $N = 2$, $\max\{x_{(N)}, a\} = 100$.

  $$\mathbb{E}(\theta|x_1, x_2) = \frac{3}{2} \cdot 100 = 150.$$

  $$median\ \theta = 2^{\frac{1}{3}} \cdot 100 \approx 126.$$

  $$mode\ \theta = 100.$$

- 3. $x_3 = 150$, $N = 3$, $\max\{x_{(N)}, a\} = 150$.

  $$\mathbb{E}(\theta|x_1, x_2, x_3) = \frac{4}{3} \cdot 150 = 200.$$

  $$median\ \theta = 2^{\frac{1}{4}} \cdot 150 \approx 178.$$

  $$mode\ \theta = 150.$$

Оценки матожиданием и медианой меняются сильнее в зависимости от наблюденных данных, увеличиваются при наблюдении номера, большего текущего максимума, уменьшаются иначе. Оценка модой может лишь возрастать при появлении номеров, больших текущего максимума. При этом соотношение между статистиками сохраняется. Таким образом, оценки матожиданием и медианой кажутся более точными и гибкими, матожиданием – более логичной и естественной, хорошо подходящей к данной задаче (достаточно интуитивной и близкой к рассуждениям людей выглядит оценка матожиданием, вдвое большим единственного наблюдения 100), но она может быть несколько завышенной. Оценка модой представляется довольно грубой и плохо учитывающей новые наблюдения, оценивает θ снизу. Т. о., на основании данных рассуждений можно выбрать оценку матожиданием.

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

Распределение из экспоненциального класса распределений может быть представлено в виде:

$$p(x|\theta) = \frac{f(x)}{g(\theta)} e^{\theta^T u(x)},\ f(x) \geq 0,\ g(\theta) > 0,$$

где θ – набор параметров распределения.

Требуется определить $f(x)$, $g(\theta)$, $u(x)$ для соответствующего распределения Парето при фиксированном a. В случае рассматриваемой задачи $\theta = b$, т. к. a фиксировано. Запишем плотность данного распределения и попытаемся представить ее в нужном виде:

$$p(x|\theta) = \frac{ba^b}{x^{b+1}}[x \geq a] = \frac{ba^b e^{-b\log x}}{x}[x \geq a].$$

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

Далее требуется найти $\mathbb{E}\log x$. Воспользуемся свойством экспоненциального класса распределений: $\mathbb{E}u_i(x) = \frac{\partial \log g(\theta)}{\theta_i} \Rightarrow$

$$\Rightarrow \mathbb{E}u(x) = \mathbb{E}(-\log x) \Rightarrow \mathbb{E}\log x = -\frac{\partial \log g(\theta)}{\partial b} = -\frac{\partial \log(\frac{1}{ba^b})}{\partial b} = \frac{\partial}{\partial b}(\log b + b\log a) = \frac{1}{b} + \log a.$$