Математическая статистика. ДЗ 14.

Прохоров Юрий, 776

Задача 1

Пусть $\mathbf{X} = (X_1, \ldots, X_n)$ — простая выборка, $X \sim \mathrm{Poiss}(\theta)$. Пусть известно априорное распределение параметра θ:

$$\theta \sim \mathrm{Gamma}(\alpha, \beta), \qquad \pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, \theta^{\alpha-1} e^{-\beta\theta}, \quad \theta > 0$$

(a) Найти оптимальную байесовскую оценку параметра θ для квадратичной функции потерь.

(b) Представить эту оценку в виде выпуклой комбинации ОМП и матожидания априорного распределения.

Решение:
Если дана *функция потерь* $\gamma(\widehat{\theta}, \theta)$ (в нашем случае квадратичная), то можно определить *функцию риска* для оценки $\widehat{\theta}(\mathbf{X})$:

$$\gamma(\widehat{\theta}, \theta) = \left(\widehat{\theta} - \theta\right)^2, \qquad R_{\widehat{\theta}}(\theta) = \mathbb{E}_\theta\left[\gamma\left(\widehat{\theta}(\mathbf{X}), \theta\right)\right]$$

Если дано априорное распределение параметра $\pi(\theta)$, $\theta \in \Theta$, то вводится *байесовский риск* $r(\widehat{\theta})$ оценки $\widehat{\theta}$:

$$r\left(\widehat{\theta}\right) = \mathbb{E}\left[R_{\widehat{\theta}}(\theta)\right] = \int\limits_\Theta R_{\widehat{\theta}}(\theta)\, \pi(\theta)\, d\theta$$

Оптимальной байесовской оценкой называется оценка, минимизирующая этот риск:

$$\widehat{\theta}^{\text{Б}}_\pi = \arg\min_{\widehat{\theta}} \left[r\left(\widehat{\theta}\right)\right] = \arg\min_{\widehat{\theta}} \left[\int\limits_\Theta R_{\widehat{\theta}}(\theta)\, \pi(\theta)\, d\theta\right]$$

Перестановкой интегралов местами и некоторыми алгебраическими преобразованиями, можно получить выражение

$$\widehat{\theta}^{\text{Б}}_\pi(\mathbf{x}) = \arg\min_{\widehat{\theta}} \left[\int\limits_\Theta \gamma(\widehat{\theta}, \theta)\, \pi(\theta \,|\, \mathbf{x})\, d\theta\right] = \arg\min_{\widehat{\theta}} \left[\int\limits_\Theta \gamma(\widehat{\theta}, \theta)\, \pi(\theta)\, L(\mathbf{x} \,|\, \theta)\, d\theta\right],$$

где $\pi(\theta \,|\, \mathbf{x})$ — плотность апостериорного распределения параметра θ:

$$\pi(\theta \,|\, \mathbf{x}) = \frac{\pi(\theta)\, L(\mathbf{x} \,|\, \theta)}{\int\limits_\Theta \pi(\theta')\, L(\mathbf{x} \,|\, \theta')\, d\theta'}, \qquad L(\mathbf{x} \,|\, \theta) \text{ — правдоподобие выборки}$$

(a) У нас квадратичная функция потерь, поэтому

$$\widehat{\theta}^{\text{Б}}_\pi(\mathbf{x}) = \arg\min_{\widehat{\theta}} \left[\int\limits_\Theta \left(\widehat{\theta} - \theta\right)^2 \pi(\theta \,|\, \mathbf{x})\, d\theta\right]$$

Дифференцируя это выражение по $\widehat{\theta}$ и приравнивая у нулю, получаем, что оптимальная байесовская оценка — матожидание апостериорного распределения:

$$\widehat{\theta}^{\text{Б}}_\pi(\mathbf{x}) = \int\limits_\Theta \theta\, \pi(\theta \,|\, \mathbf{x})\, d\theta$$

Правдоподобие выборки:

$$L(\mathbf{X} \,|\, \theta) = \prod_{i=1}^n \frac{\theta^{X_i}}{X_i!} e^{-\theta} = \frac{1}{X_1! \ldots X_n!}\, \theta^{X_1 + \ldots + X_n}\, e^{-n\theta}$$

Безусловное правдоподобие выборки:

$$L(\mathbf{X}) = \int\limits_{\Theta} \pi(\theta)\, L(\mathbf{X}\,|\,\theta)\, d\theta = \frac{\beta^\alpha}{\Gamma(\alpha) X_1! \dots X_n!} \int\limits_0^{+\infty} \underbrace{\theta^{n\overline{X}+\alpha-1} e^{-(\beta+n)\theta}}_{\propto \mathrm{Gamma}\left(n\overline{X}+\alpha,\, \beta+n\right)}\, d\theta = \frac{\beta^\alpha}{\Gamma(\alpha) X_1! \dots X_n!} \cdot \frac{\Gamma(n\overline{X}+\alpha)}{(\beta+n)^\alpha}$$

Апостериорное распределение:

$$\pi(\theta\,|\,\mathbf{X}) = \frac{\pi(\theta)\, L(\mathbf{X}\,|\,\theta)}{L(\mathbf{X})} = \frac{(\beta+n)^\alpha}{\Gamma(n\overline{X}+\alpha)}\, \theta^{n\overline{X}+\alpha-1} e^{-(\beta+n)\theta} \sim \mathrm{Gamma}\big(n\overline{X}+\alpha,\, \beta+n\big)$$

Тогда оптимальная байесовская оценка:

$$\widehat{\theta}^{\text{Б}}_\pi(\mathbf{X}) = \mathbb{E}\Big[\mathrm{Gamma}\big(n\overline{X}+\alpha,\, \beta+n\big)\Big] = \frac{n\overline{X}+\alpha}{\beta+n}$$

(b) Оценка максимального правдоподобия для пуассоновской выборки:

$$\widehat{\theta}^{\text{ОПМ}}(\mathbf{X}) = \overline{X}$$

Матожидание априорного распределения:

$$\mathbb{E}\big[\mathrm{Gamma}(\alpha,\beta)\big] = \frac{\alpha}{\beta}$$

Тогда оптимальную байесовскую оценку можно записать в виде:

$$\widehat{\theta}^{\text{Б}}_\pi(\mathbf{X}) = \frac{n}{\beta+n}\overline{X} + \frac{\beta}{\beta+n}\frac{\alpha}{\beta} = \frac{n}{\beta+n} \cdot \widehat{\theta}^{\text{ОПМ}}(\mathbf{X}) + \frac{\beta}{\beta+n} \cdot \mathbb{E}\big[\mathrm{Gamma}(\alpha,\beta)\big]$$

Теорема 1

> Пусть априорное распределение $\pi(\theta)$ таково, что оптимальный байесовский риск:
>
> $$r\big(\,\widehat{\theta}^{\text{Б}}_\pi\,\big) = \int\limits_{\Theta} R_{\widehat{\theta}^{\text{Б}}_\pi}(\theta)\, \pi(\theta)\, d\theta = \sup_\theta \Big[R_{\widehat{\theta}^{\text{Б}}_\pi}(\theta)\Big]$$
>
> Тогда
>
> - оптимальная байесовская оценка $\widehat{\theta}^{\text{Б}}_\pi$ является также минимаксной оценкой;
> - если такое распределение $\pi(\theta)$ единственно, то и минимаксная оценка единственна.
>
> При этом априорное распределение $\pi(\theta)$ называется *наименее благоприятным распределением.*

Теорема 2

> Пусть
>
> - $\{\pi_k(\theta)\}_{k=1}^\infty$ — последовательность априорных распределений;
> - $\big\{\widehat{\theta}^{\text{Б}}_{\pi_k}\big\}_{k=1}^\infty$ — последовательность соответствующих оптимальных байесовских оценок;
> - $\widehat{\theta}$ — такая оценка, что
>
> $$\forall \theta \in \Theta \qquad \longmapsto \qquad R_{\widehat{\theta}}(\theta) \le \lim_{k\to\infty} r\big(\,\widehat{\theta}^{\text{Б}}_{\pi_k}\,\big)$$
>
> Тогда $\widehat{\theta}$ — минимаксная оценка.

Задача 2

Доказать, что при выполнении условий теоремы 2 в неравенстве

$$R_{\widehat{\theta}}(\theta) \le \lim_{k\to\infty} r\big(\widehat{\theta}^{\text{Б}}_{\pi_k}\big) \qquad (*)$$

для почти всех $\theta \in \Theta$ выполнены равенства.

Решение:

По определению оптим. байес. оценки

$$\forall\, \hat{\theta} \qquad \int R_{\hat\theta}(\theta)\,\pi_k(\theta)\,d\theta \ \geq\ \int R_{\hat\theta^{\text{Б}}_{\pi_k}}(\theta)\,\pi_k(\theta)\,d\theta = r\big(\hat\theta^{\text{Б}}_{\pi_k}\big)$$

Возьмём в кач-ве $\hat\theta$ оценку из условия

Перейдём к пределу $k\to\infty$

$$\lim_{k\to\infty} \int R_{\hat\theta}(\theta)\,\pi_k(\theta)\,d\theta \ \geq\ \lim_{k\to\infty} r\big(\hat\theta^{\text{Б}}_{\pi_k}\big)$$

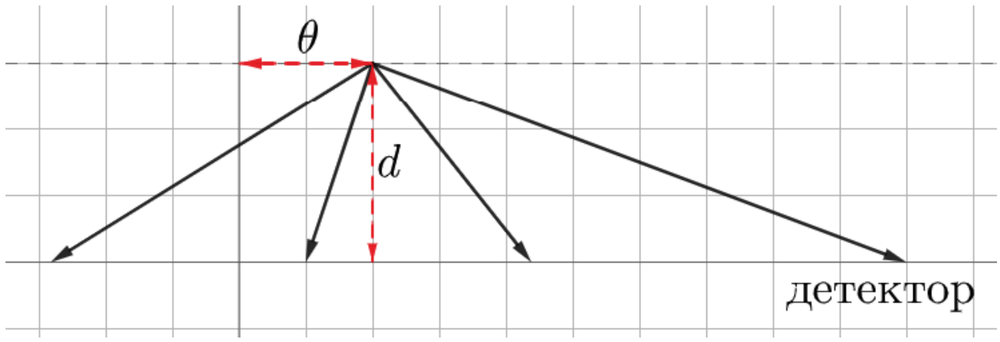
Это противоречит тому, что $\exists\, \Theta^\circ$ (не нулев. меры)

$$\forall \theta \in \Theta^\circ \qquad R_{\hat\theta}(\theta) < \lim r\big(\hat\theta^{\text{Б}}_{\pi_k}\big)$$

$$\forall \theta \notin \Theta^\circ \qquad R_{\hat\theta}(\theta) = \lim r\big(\hat\theta^{\text{Б}}_{\pi_k}\big)$$

Задача 3

Рассматривается задача обнаружения источника излучения в двумерном случае:



Точечный источник излучает в сторону плоского детектора под углом $\varphi \sim \mathcal{U}[0,1]$. На детекторе появляются независимые точки $\mathbf{X} = (X_1, \ldots, X_n)$.

(a) Найти распределение X.

(b) Показать, что $\overline{X} \stackrel{d}{=} X_1$. Показать, что оценка $\widehat{\theta}_1 = \overline{X}$ параметра θ является "плохой".

(c) Доказать, что если $\mathbf{X} = (X_1, \ldots, X_n)$ — простая выборка с плотностью f и медианой μ, причем $f(\mu) > 0$, то

$$\sqrt{n}\Big(X_{\left(\frac{n}{2}\right)} - \mu\Big) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0,\ \frac{1}{4f^2(\mu)}\right)$$

(d) Показать, что оценка $\widehat{\theta}_2 = X_{\left(\frac{n}{2}\right)}$ является "хорошей" оценкой параметра θ.

Решение:

(a) Из курса теории вероятности мы знаем, что такая постановка задачи приводит к распределению Коши:

$$X \sim \text{Cauchy}(\theta, d), \qquad f(x) = \frac{1}{\pi d\left[1 + \left(\frac{x-\theta}{d}\right)^2\right]}$$

(b) Данное семейство распределений замкнуто относительно линейных операций:

- $\xi \sim \text{Cauchy}(x_0,\, d) \qquad \Longrightarrow \qquad k\xi + l \sim \text{Cauchy}(kx_0 + l,\, |k|d)$
- $\xi \sim \text{Cauchy}(x_1,\, d_1), \quad \eta \sim \text{Cauchy}(x_2,\, d_2) \qquad \Longrightarrow \qquad \xi + \eta \sim \text{Cauchy}(x_1 + x_2,\, d_1 + d_2)$

Отсюда следует, что

$$X_1 + \ldots + X_n \sim \text{Cauchy}(n\theta,\, nd), \qquad \overline{X} \sim \text{Cauchy}(\theta,\, d)$$

Оценка $\widehat{\theta}_1 = \overline{X}$ не обладает никакими хорошими свойствами. Она не является несмещенной (у нее нет матожидания), она имеет бесконечную дисперсию.

(c) Пусть для простоты $n = 2k+1$. Введем случайные величины $Y_i = F(X_i) \sim \mathcal{U}[0,1]$.

Найдем плотность $Y_{\left(\frac{n}{2}\right)} = Y_{(k+1)}$:

$$\begin{aligned}
f_{Y_{(k+1)}}(x) &= \lim_{\Delta x \to 0} \frac{1}{\Delta x}\mathbb{P}\left\{Y_{(k+1)} \in [x, x+\Delta x]\right\} = \\
&= \lim_{\Delta x \to 0} \frac{1}{\Delta x}\left[\begin{array}{c}\#\text{ способов}\\ \text{разбить выборку}\\ \text{на 3 группы}\\ \text{по } k,\ 1 \text{ и } k \text{ элементов}\end{array}\right] \cdot \mathbb{P}\left\{\begin{array}{c}\text{1 группа}\\ \text{левее } x\end{array}\right\} \cdot \mathbb{P}\left\{\begin{array}{c}\text{2 группа}\\ \in [x, x+\Delta x]\end{array}\right\} \cdot \mathbb{P}\left\{\begin{array}{c}\text{3 группа}\\ \text{правее } x+\Delta x\end{array}\right\} = \\
&= \lim_{\Delta x \to 0} \frac{1}{\Delta x}\left[\frac{(2k+1)!}{k!k!}\right] \cdot (x)^k \cdot (1 \cdot \Delta x) \cdot (1-x)^k = \frac{(2k+1)!}{k!k!}\left[x(1-x)\right]^k
\end{aligned}$$

Тогда плотность $Z_k = \sqrt{2k+1}\left(Y_{(k+1)} - \frac{1}{2}\right)$:

$$\begin{aligned}
f_{Z_k}(x) &= \frac{1}{\sqrt{2k+1}} f_{Y_{(k+1)}}\left(\frac{x}{\sqrt{2k+1}} + \frac{1}{2}\right) = \frac{1}{\sqrt{2k+1}}\frac{(2k+1)!}{k!k!}\left[\left(\frac{1}{2} + \frac{x}{\sqrt{2k+1}}\right)\left(\frac{1}{2} - \frac{x}{\sqrt{2k+1}}\right)\right]^k = \\
&= \frac{1}{\sqrt{2k+1}}\frac{(2k+1)!}{k!k!}\left(\frac{1}{4} - \frac{x^2}{2k+1}\right)^k = \Bigg/ \begin{array}{c}\text{формула Стирлинга:}\\ k! \sim \sqrt{2\pi k}\left(\frac{k}{e}\right)^k\end{array} \Bigg/ \overset{k\to\infty}{\sim} \\
&\sim \frac{1}{\sqrt{2k+1}} \cdot \frac{\sqrt{2\pi}\sqrt{2k+1}(2k+1)^{2k+1}}{e^{2k+1}} \cdot \frac{e^{2k}}{2\pi k k^{2k}} \cdot \frac{1}{2^{2k}} \cdot \underbrace{\left[\left(1 - \frac{1}{\frac{2k+1}{4x^2}}\right)^{-\frac{2k+1}{4x^2}}\right]^{-\frac{4x^2}{2k+1}\cdot k}}_{\longrightarrow\, e} \overset{k\to\infty}{\sim} \\
&\sim \frac{(2k+1)^{2k+1}\cdot 2}{e(2k)^{2k+1}\sqrt{2\pi}} \cdot e^{-2x^2} \sim \sqrt{\frac{2}{\pi}} \cdot \frac{1}{e}\underbrace{\left(1 + \frac{1}{2k}\right)^{2k}}_{\longrightarrow e}\left(1 + \frac{1}{2k}\right)e^{-2x^2} \xrightarrow{k\to\infty} \sqrt{\frac{2}{\pi}}e^{-2x^2}
\end{aligned}$$

Получили плотность распределения $\mathcal{N}\left(0,\, \frac{1}{4}\right)$. Значит, мы доказали, что

$$Z_k = \sqrt{2k+1}\left(Y_{(k+1)} - \frac{1}{2}\right) \xrightarrow[k\to\infty]{d} \mathcal{N}\left(0,\, \frac{1}{4}\right)$$

Теперь вернемся обратно к X_i. Так как функция распределения F монотонно возрастает, то порядковые статистики переходят друг в друга:

$$Y_{(2k+1)} = F\big(X_{(2k+1)}\big), \qquad X_{(2k+1)} = F^{-1}\big(Y_{(2k+1)}\big)$$

Теорема (дельта-метод)

> Пусть
>
> - $\{X_n\}_{n=1}^{\infty}$ — такая последовательность случайных величин, что
>
> $$\sqrt{n}\big(X_n - \alpha\big) \xrightarrow[n\to\infty]{d} \mathcal{N}(0,\, \sigma^2), \qquad \alpha \in \mathbb{R}$$
>
> - $g(x)$ — непрерывно дифференцируемая функция
>
> Тогда выполнено
>
> $$\sqrt{n}\big(g(X_n) - g(\alpha)\big) \xrightarrow[n\to\infty]{d} \mathcal{N}\Big(0,\, \sigma^2 \cdot \big[g'(\alpha)\big]^2\Big)$$

В нашем случае $g(x) = F^{-1}(x)$, и ее производная в точке $\alpha = \frac{1}{2}$:

$$F\left(F^{-1}\left(x\right)\right) = x \qquad \overset{\frac{d}{dx}}{\Longrightarrow} \qquad F'\left(F^{-1}(x)\right) \cdot \left(F^{-1}(x)\right)' = 1 \qquad \overset{x=\frac{1}{2}}{\Longrightarrow} \qquad \left(F^{-1}(x)\right)'\Big|_{x=\frac{1}{2}} = \frac{1}{F'(\mu)} = \frac{1}{f(\mu)},$$

где $\mu = F^{-1}\left(\frac{1}{2}\right)$ — медиана выборки.

По теореме о дельта-методе получаем

$$\sqrt{2k+1}\big(X_{k+1} - \mu\big) \xrightarrow[k\to\infty]{d} \mathcal{N}\left(0,\, \frac{1}{4f^2(\mu)}\right)$$

(d) "Хорошей" оценкой неизвестного параметра θ будет медиана:

$$\widehat{\theta}_2(\mathbf{X}) = X_{\left(\frac{n}{2}\right)}$$

Эта оценка является несмещенной, асимптотически нормальной. Отсюда следует, что она является состоятельной.