Домашнее задание №1 по курсу «Математическая Статистика в Машинном Обучении»

Школа Анализа Данных

Оценки и сходимость

Задача 1 [2 балла]

Пусть $\boldsymbol{X}^n = \{X_1, \dots, X_n\} \sim \text{Uniform}(0, \theta)$, $\hat{\theta} = \max\{X_1, \dots, X_n\}$. Найти значения bias, se и MSE этой оценки.

Решение:

Функция распределения для равномерного распределения:

$$F(x) = \begin{cases} 0, & x < 0 \\ \frac{x}{\theta}, & x \in [0, \theta] \\ 1, & x > \theta \end{cases}$$

Функция распределения для $\hat{\theta}$: $F_{\hat{\theta}}(y) = P(\hat{\theta} \le y) = P(\max X_i \le y) = \prod_{i=1}^{n} P(X_i \le y) = (F(y))^n = \begin{cases} 0, & y < 0 \\ (\frac{y}{\theta})^n, & y \in [0.\theta] \\ 1, & y > \theta \end{cases}$

Тогда плотность распределения для $\hat{\theta}$: $p_{\hat{\theta}}(y) = \begin{cases} 0, & y \notin [0, \theta] \\ \frac{ny^{n-1}}{\theta^n}, & y \in [0.\theta] \end{cases}$

$$\mathbb{E}\hat{\theta} = \int_{-\infty}^{\infty} y p_{\hat{\theta}}(y) dy = \int_0^{\theta} \frac{ny^n}{\theta^n} dy = \frac{n\theta}{n+1}$$

$$bias(\hat{\theta}) = \mathbb{E}\hat{\theta} - \theta = -\frac{\theta}{n+1}$$

$$se(\hat{\lambda}\theta) = \sqrt{\mathbb{V}\hat{\theta}} = \sqrt{\mathbb{E}\hat{\theta}^2 - (\mathbb{E}\hat{\theta})^2} = \frac{\sqrt{n}\theta}{\sqrt{n+2}(n+1)}$$

$$MSE(\hat{\theta}) = bias^2(\hat{\theta}) + se^2(\hat{\theta}) = \frac{2\theta^2}{(n+1)(n+2)}$$

Ответ: $bias(\hat{\theta}) = -\frac{\theta}{n+1}$; $se(\hat{\lambda}\theta) = \frac{\sqrt{n}\theta}{\sqrt{n+2}(n+1)}$; $MSE(\hat{\theta}) = \frac{2\theta^2}{(n+1)(n+2)}$

Задача 2 [2 балла]

Пусть $\boldsymbol{X}^n = \{X_1, \dots, X_n\} \sim \text{Exp}(\lambda)$, $\hat{\lambda} = 1/\langle \boldsymbol{X}^n \rangle$. Найдите bias, se, MSE этой оценки. Является ли оценка смещенной? Состоятельной?

Решение:

Если $X_i \sim Exp(\lambda)$, то $\sum_{i=1}^{n} X_i \sim \Gamma(n, \frac{1}{\lambda})$ - гамма-распределение. Плотность $\Gamma(n, \frac{1}{\lambda})$: $p(x) = \begin{cases} 0, & x < 0 \\ \frac{x^{n-1} e^{-\lambda x} \lambda^n}{\Gamma(n)}, & x \ge 0 \end{cases}$

Обозначим $\sum_{i=1}^{n} X_i = y$, тогда $\hat{\lambda} = \frac{1}{\langle \boldsymbol{X}^n \rangle} = \frac{n}{y}$.

$$\mathbb{E}\hat{\lambda} = \int_{-\infty}^{+\infty} \frac{n}{y} p(y) dy = \int_0^{+\infty} \frac{ny^{n-2} e^{-\lambda y} \lambda^n}{\Gamma(n)} dy = \int_0^{+\infty} \frac{y^{n-2} e^{-\lambda y} \lambda^{n-1}}{\Gamma(n-1)} dy * \frac{\lambda n \Gamma(n-1)}{\Gamma(n)} = \frac{\lambda n}{n-1},$$

тк интеграл равен 1, а $\Gamma(n) = (n-1)!$, если n - целое неотрицательное число.

$bias(\hat{\lambda}) = \mathbb{E}\hat{\lambda} - \lambda = \frac{\lambda}{n-1}$, оценка смещенная.

$\mathbb{E}\hat{\lambda}^2 = \int_{-\infty}^{+\infty} (\frac{n}{y})^2 p(y) dy = \int_0^{+\infty} \frac{n^2 y^{n-3} e^{-\lambda y} \lambda^n}{\Gamma(n)} = \frac{n^2 \lambda^2}{(n-1)(n-2)}$

$se(\hat{\lambda}) = \sqrt{\mathbb{V}\hat{\lambda}} = \sqrt{\mathbb{E}\hat{\lambda}^2 - (\mathbb{E}\hat{\lambda})^2} = \frac{n\lambda}{(n-1)\sqrt{n-2}}$, оценка состоятельна, тк $se(\hat{\lambda}) \to 0$ при $n \to \infty$.

$MSE(\hat{\lambda}) = bias^2(\hat{\lambda}) + se^2(\hat{\lambda}) = \frac{\lambda^2(n+2)}{(n-1)(n-2)}$

Ответ: $bias(\hat{\lambda}) = \frac{\lambda}{n-1}$; $se(\hat{\lambda}) = \frac{n\lambda}{(n-1)\sqrt{n-2}}$; $MSE(\hat{\lambda}) = \frac{\lambda^2(n+2)}{(n-1)(n-2)}$; оценка смещенная, но состоятельная.

Эмпирическая функция распределения

Задача 3 [2 балла]

Пусть $\hat{F}_n(x)$ — эмпирическая функция распределения. Пусть x, $y \in \mathbb{R}$. Найдите ковариацию $\mathrm{Cov}(\hat{F}_n(x), \hat{F}_n(y))$.

Решение:

$$\mathrm{Cov}(\hat{F}_n(x), \hat{F}_n(y)) = \mathbb{E}(\hat{F}_n(x)\hat{F}_n(y)) - \mathbb{E}\hat{F}_n(x)\mathbb{E}\hat{F}_n(y)$$

$$\mathbb{E}\hat{F}_n(x) = \mathbb{E}\frac{\sum_{i=1}^{n} I(X_i \le x)}{n} = \mathbb{E}I(X_i \le x) = P(X_i \le x) = F(x)$$

$$\mathbb{E}(\hat{F}_n(x)\hat{F}_n(y)) = \frac{1}{n^2}\mathbb{E}(\sum_{i=1}^{n} I(X_i \le x)\sum_{j=1}^{n} I(X_j \le y)) = \frac{1}{n^2}\mathbb{E}(\sum_{i=1}^{n} I(X_i \le x, X_i \le y) + \sum_{i \ne j} I(X_i \le x)I(X_j \le y)) =$$

$$= \frac{n}{n^2}\mathbb{E}I(X_i < min(x,y)) + \frac{n(n-1)}{n^2}\mathbb{E}(I(X_i \le x)I(X_j \le y)) = \frac{1}{n}F(min(x,y)) + \frac{n-1}{n}\mathbb{E}(I(X_i \le x))\mathbb{E}(I(X_j \le y)) =$$

$$= \frac{1}{n}F(min(x,y)) + \frac{n-1}{n}F(x)F(y)$$

$$\mathrm{Cov}(\hat{F}_n(x), \hat{F}_n(y)) = \frac{1}{n}(F(min(x,y)) - F(x)F(y))$$

Ответ: $\frac{1}{n}(F(min(x,y)) - F(x)F(y))$

Задача 4 [2 балла]

Пусть $\boldsymbol{X}^n = \{X_1, \ldots, X_n\} \sim F(x)$, и пусть $\hat{F}_n(x)$ — эмпирическая функция распределения. Для фиксированных числе $a, b \in \mathbb{R}$, таких что $a < b$ определим статистический функционал $T(F) = F(b) - F(a)$. Пусть $\hat{\theta} = \hat{F}_n(b) - \hat{F}_n(a)$. Найдите оценку $\hat{\mathrm{se}}$ стандартного отклонения и $(1-\alpha)$-доверительный интервал.

Решение:

$I(X_i \le x)$-индикаторная величина, имеющая распределение Бернулли

$$\mathbb{E}\hat{\theta} = \frac{n}{n}\mathbb{E}(I(X_i \le b) - I(X_i \le a)) = \mathbb{E}(I(a < X_i \le b)) = P(a < X_i \le b) = F(b) - F(a) = T(F)$$

$$\mathbb{V}\hat{\theta} = \mathbb{V}(\hat{F}_n(b) - \hat{F}_n(a)) = \frac{1}{n}\mathbb{V}(I(X_i \le b) - I(X_i \le a)) = \frac{1}{n}\mathbb{V}(I(a < X_i \le b)) = \frac{1}{n}P(a < X_i \le b)(1 - P(a < X_i \le b)) =$$

$$= \frac{(F(b) - F(a))(1 - F(b) + F(a))}{n} = \frac{T(F)(1 - T(F))}{n}$$

$$\hat{\mathrm{se}} = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

Доверительный интервал:

$$\frac{\hat{\theta} - T(F)}{\hat{\mathrm{se}}} \sim N(0,1) \text{ по ЦПТ}$$

$$\left|\frac{\hat{\theta} - T(F)}{\hat{\mathrm{se}}}\right| \le z_{\alpha/2}$$

$$\hat{\theta} - z_{\alpha/2}\hat{\mathrm{se}} \le T(F) \le \hat{\theta} + z_{\alpha/2}\hat{\mathrm{se}}$$

$$\hat{\theta} - z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \le T(F) \le \hat{\theta} + z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

Получили доверительный интервал

Задача 5 [2 балла]

Скачайте данные о качестве красных вин. Постройте график для $\hat{F}(x; \boldsymbol{x}^n)$ для уровня кислотности (pH). Для каждой точки x постройте:

- 95%-ый доверительный интервал на основе неравенства Дворецкого-Кифера-Вольфовица.
- Асимптотический нормальный 95%-ый доверительный интервал для значения $F(x)$.

По значениям уровня кислотности $\boldsymbol{x}^n$ подсчитайте оценку $T(\boldsymbol{x}^n)$ для функционала $T(F) = F(3.5) - F(3.4)$ и найдите оцените аналитически стандартное отклонение $\hat{se}$ оценки $T(\boldsymbol{x}^n)$. Постройте асимптотический нормальный 95%-ый доверительный интервал для $T(F)$.

Задача 6 [2 балла]

В процессе очистки питьевой воды выпадает значительный осадок. Для его уменьшения можно воздействовать на разные факторы, в т.ч. на количество микроорганизмов в жидкости, способствующих окислению органики. В группу из 261 очистительных установок был добавлен реагент, подавляющих активность микроорганизмов, а состав остальных 119 остался без изменений. Пусть θ — разность в средних значениях количества твердых частиц в этих двух группах установок. Оценить по данным `WaterTreatment` величину θ, оценить стандартную ошибку оценки, построить 95% и 99% доверительные интервалы. Какие выводы можно сделать на основе полученных результатов?

Бутстреп

Задача 7 [3 балла]

Провести моделирование, чтобы сравнить различные типы доверительных интервалов, построенных с помощью бутстрепа. Пусть $n = 50$, $T(F) = \int (x-\mu)^3 \, dF(x)/\sigma^3$ — коэффициент асимметрии, где F — логнормальное распределение. Постройте 95% доверительные интервалы для $T(F)$ (под F понимается распределение элементов выборки $X_1, \ldots, X_n$) по данным $\boldsymbol{X}^n = \{X_1, \ldots, X_n\}$, используя три подхода на основе бутстрепа.

Замечание. *Выборку из логнормального распределения можно сгенерировать из нормального, сначала сгенерировав выборку н.о.р. величин $\boldsymbol{Y}^n = \{Y_1, \ldots, Y_n\} \sim \mathcal{N}(0,1)$, после чего положив $X_i = e^{Y_i}$, $i = 1, 2, \ldots, n$.*

Задача 8 [3 балла]

Пусть $\boldsymbol{X}^n = \{X_1, \ldots, X_n\} \sim \mathcal{N}(\mu, \sigma^2)$, $\theta = e^{\mu}$ и $\hat{\theta} = e^{\langle \boldsymbol{X}_n \rangle}$. Найдите аналитически плотность распределения $p_{\hat{\theta}}(x)$ оценки $\hat{\theta} = e^{\langle \boldsymbol{X}^n \rangle}$, математическое ожидание $\mathbb{E}(\hat{\theta})$, и дисперсию $\mathbb{V}(\hat{\theta})$, а также bias, se, MSE оценки $\hat{\theta}$. Является ли оценка $\hat{\theta}$ смещенной? Состоятельной?

Решение:

$X_i \sim \mathcal{N}(\mu, \sigma^2)$.$X_i$ независимы, поэтому $\sum\limits_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$.

Пусть $\sum\limits_{i=1}^{n} X_i = x$. Тогда:

$$\mathbb{E}(\hat{\theta}) = \int\limits_{-\infty}^{+\infty} e^{\frac{x}{n}} \frac{1}{\sqrt{2\pi n\sigma^2}} e^{-\frac{(x-n\mu)^2}{2n\sigma^2}} dx = e^{\mu + \frac{\sigma^2}{2n}}$$

$$\mathbb{V}(\hat{\theta}) = \mathbb{E}(\hat{\theta}^2) - (\mathbb{E}\hat{\theta})^2 = \int\limits_{-\infty}^{+\infty} e^{\frac{2x}{n}} \frac{1}{\sqrt{2\pi n\sigma^2}} e^{-\frac{(x-n\mu)^2}{2n\sigma^2}} dx - e^{2\mu + \frac{\sigma^2}{n}} = e^{2\mu + \frac{2\sigma^2}{n}} - e^{2\mu + \frac{\sigma^2}{n}} = e^{2\mu + \frac{\sigma^2}{n}} \left(e^{\frac{\sigma^2}{n}} - 1 \right)$$

$$bias(\hat{\theta}) = \mathbb{E}\hat{\theta} - \theta = e^{\mu + \frac{\sigma^2}{2n}} - e^{\mu}$$

$$se(\hat{\theta}) = \sqrt{\mathbb{V}\hat{\theta}} = \sqrt{e^{2\mu + \frac{\sigma^2}{n}} \left(e^{\frac{\sigma^2}{n}} - 1 \right)}$$

$$MSE = bias^2(\hat{\theta}) + \mathbb{V}(\hat{\theta}) = e^{2\mu} \left(e^{\frac{2\sigma^2}{n}} - 2e^{\frac{\sigma^2}{2n}} + 1 \right)$$

Оценка является смещенной, тк $\mathbb{E}\hat{\theta} \neq \theta$, но состоятельной ($\mathbb{V}\hat{\theta} \to 0$, при $n \to \infty$)

Плотность распределения $p_{\hat{\theta}}(x)$:

$$F_{\hat{\theta}}(x) = P(\hat{\theta} < x) = P(e^{\langle \boldsymbol{X}_n \rangle} < x) = P(\sum_{i=1}^{n} X_i < n\ln(x)) = F_{\sum\limits_{i=1}^{n} X_i}(n\ln(x))$$

$$p_{\hat{\theta}}(x) = \frac{\partial F_{\sum\limits_{i=1}^{n} X_i}(n\ln(x))}{\partial x} = p_{\sum\limits_{i=1}^{n} X_i}(n\ln(x)) * \frac{n}{x} = \frac{n}{x\sqrt{2\pi n\sigma^2}} e^{-\frac{(n\ln(x) - n\mu)^2}{2n\sigma^2}}$$

Задача 9 [2 балла]

Пусть $\boldsymbol{X}^n = \{X_1, \ldots, X_n\} \sim \mathcal{N}(\mu, 1)$, $\theta = e^{\mu}$ и $\hat{\theta} = e^{\langle \boldsymbol{X}^n \rangle}$. Сгенерируйте выборку $\boldsymbol{X}^n$ из $n = 100$ наблюдений для $\mu = 10$. Нарисуйте гистограмму значений $\{\hat{\theta}_i^*\}_{i=1}^B$ бутстрепных оценок. Эта гистограмма является оценкой распределения $p_{\hat{\theta}}(x)$. Сравните ее с настоящим распределением $p_{\hat{\theta}}(x)$. Используя бутстреп, подсчитайте величину se и постройте тремя способами 95% доверительный интервал для θ.