Школа анализа данных
Основы статистики в машинном обучении
Домашнее задание №1

Федор Ерин

Оценки и сходимость

Задача 1 [2 балла]

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

Решение

Заметим, что $\hat{\theta} = max(X_1, \ldots, X_n) = X_{(n)}$, где $X_{(n)}$ — n-ая порядковая статистика. Найдем эмпирическую функцию распределения для $\hat{\theta}$:

$$\hat{F}_n(X_{(n)}) = \mathbb{P}(X_{(n)} < x) = \mathbb{P}(X_1 < x)^n = \begin{cases} \frac{x^n}{\theta^n}, x \in [0, \theta] \\ 1, x > \theta \\ 0, x < 0 \end{cases}$$

Зная $\hat{F}_n(X_{(n)})$, найдем по определению 1 и 2 моменты, а затем $bias, se, MSE$:

$$\mathbb{E}X_{(n)} = \int_0^\theta x d\hat{F}_n(X_{(n)}) = \int_0^\theta xn\frac{x^{n-1}}{\theta^n}dx = \frac{n}{\theta^n}\frac{1}{n+1}(x^{n+1})\Big|_0^\theta = \frac{n}{n+1}\theta$$

$$\mathbb{E}X_{(n)}^2 = \int_0^\theta x^2 d\hat{F}_n(X_{(n)}) = \frac{n}{\theta^n}\frac{1}{n+2}(x^{n+2})\Big|_0^\theta = \frac{n}{n+2}\theta^2$$

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

$$\mathbb{V}X_{(n)} = \mathbb{E}X_{(n)}^2 - (\mathbb{E}X_{(n)})^2 = \frac{n}{n+2}\theta^2 - \frac{n^2}{(n+1)^2}\theta^2 = \frac{n\theta^2(n+1)^2 - n^2\theta^2(n+2)}{(n+2)(n+1)^2} = \frac{n\theta^2}{(n+2)(n+1)^2}$$

$$se = \sqrt{\mathbb{V}X_{(n)}} = \frac{\theta}{n+1}\sqrt{\frac{n}{n+2}}$$

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

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

Задача 2 [2 балла]

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

Решение
Заметим, что $X_i \sim Exp(\lambda) = \Gamma(1, \frac{1}{\lambda})$. По свойству гамма-распределения:

$$\sum_{i=1}^{n} X_i \sim \Gamma(\sum_{i=1}^{n} 1, \frac{1}{\lambda}) = \Gamma(n, \frac{1}{\lambda})$$

Известна плотность гамма-распределения:

$$\gamma(x) = \begin{cases} \frac{\lambda^n}{\Gamma(n)} x^{n-1} e^{-\lambda x}, x > 0 \\ 0, x \le 0 \end{cases}$$

Теперь можем найти по определению 1 и 2 моменты, а затем $bias, se, MSE$:

$$\mathbb{E}\frac{1}{\overline{X^n}} = \mathbb{E}\frac{n}{\sum_{i=1}^{n} X_i} = n\int_0^\infty \frac{1}{x}\frac{\lambda^n}{(n-1)!}x^{n-1}e^{-\lambda x}dx =$$

$$= \frac{\lambda n}{(n-1)!}\int_0^\infty (\lambda x)^{n-2}e^{-\lambda x}d(\lambda x) = \frac{\lambda n}{(n-1)!}\Gamma(n-1) = \frac{\lambda n}{(n-1)!}(n-2)! = \frac{n}{n-1}\lambda$$

$$\mathbb{E}\frac{1}{(\overline{X^n})^2} = \mathbb{E}\frac{n^2}{(\sum_{i=1}^{n} X_i)^2} = n^2\int_0^\infty \frac{1}{x^2}\frac{\lambda^n}{(n-1)!}x^{n-1}e^{-\lambda x}dx =$$

$$= \frac{\lambda^2 n^2}{(n-1)!}\int_0^\infty (\lambda x)^{n-3}e^{-\lambda x}d(\lambda x) = \frac{\lambda^2 n^2}{(n-1)!}\Gamma(n-2) = \frac{\lambda^2 n^2}{(n-1)!}(n-3)! = \frac{\lambda^2 n^2}{(n-1)(n-2)}$$

$$bias = \mathbb{E}\hat{\lambda} - \lambda = \frac{n}{n-1}\lambda - \lambda = \frac{\lambda}{n-1} \neq 0 \Rightarrow \text{оценка смещенная}$$

$$\mathbb{V}\hat{\lambda} = \mathbb{E}\frac{1}{(\overline{X^n})^2} - \big(\mathbb{E}\frac{1}{\overline{X_n}}\big)^2 = \frac{\lambda^2 n^2}{(n-1)(n-2)} - \frac{\lambda^2 n^2}{(n-1)^2} = \frac{\lambda^2 n^2(n-1) - \lambda^2 n^2(n-2)}{(n-1)^2(n-2)} = \frac{\lambda^2 n^2}{(n-1)^2(n-2)}$$

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

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

$$bias = \frac{\lambda}{n-1} \to 0, se = \frac{\lambda n}{(n-1)\sqrt{n-2}} \to 0 \text{ при } n \to \infty \Rightarrow \text{оценка состоятельна.}$$

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

Эмпирическая функция распределения

Задача 3 [2 балла]

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

Решение
Используя $\hat{F_n}(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x)$ и $\mathbb{E}\hat{F_n}(x) = F(x)$, распишем по определению ковариацию:

$$Cov(\hat{F_n}(x), \hat{F_n}(y)) =$$
$$= \mathbb{E}[(\hat{F_n}(x) - \mathbb{E}\hat{F_n}(x))(\hat{F_n}(y) - \mathbb{E}\hat{F_n}(y))] =$$
$$= \mathbb{E}[\hat{F_n}(x) \cdot \hat{F_n}(y)] - \mathbb{E}\hat{F_n}(x) \cdot \mathbb{E}\hat{F_n}(y) =$$
$$= \frac{1}{n^2}\mathbb{E}\big[\sum_{i=1}^{n} I(X_i \le x) \cdot \sum_{j=1}^{n} I(X_i \le y)\big] - F(x) \cdot F(y) =$$
$$= \frac{1}{n^2}\mathbb{E}\big[\sum_{i=j} I(X_i \le x)I(X_i \le y) + \sum_{i \ne j} I(X_i \le x)I(X_i \le y)\big] - F(x) \cdot F(y) =$$
$$= \frac{1}{n^2}\big[nF(min(x,y)) + n(n-1)F(x)F(y)\big] - F(x) \cdot F(y) =$$
$$= \frac{F(min(x,y)) - F(x)F(y)}{n}$$

Ответ: $Cov(\hat{F_n}(x), \hat{F_n}(y)) = \frac{F(min(x,y)) - F(x)F(y)}{n}$.

Задача 4 [2 балла]

Пусть $X^n = \{X_1, \dots, X_n\} \sim F(x)$, 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 $\hat{se}$ стандартного отклонения и $(1-\alpha)$-доверительный интервал.

Решение
Найдем $\hat{se}$ для $\hat{\theta}$, используя $\mathbb{V}\hat{F_n}(x) = \frac{F(x)(1-F(x))}{n}$ и полученную в задаче 3 формулу для ковариации:

$$\hat{se}^2 = \mathbb{V}\hat{\theta} = \mathbb{V}(\hat{F_n}(b) - \hat{F_n}(a)) = \mathbb{V}\hat{F_n}(b) + \mathbb{V}\hat{F_n}(a) - 2Cov(\hat{F_n}(b), \hat{F_n}(a)) =$$
$$= \frac{F(b)(1-F(b))}{n} + \frac{F(a)(1-F(a))}{n} - 2\frac{F(min(a,b)) - F(a)F(b)}{n} =$$
$$= \frac{1}{n}\big[F(b) - F^2(b) + F(a) - F^2(a) - 2F(a) + 2F(a)F(b)\big] =$$
$$= \frac{1}{n}\big[F(b) - F(a) - (F(b) - F(a))^2\big] =$$
$$= \frac{T(F)(1-T(F))}{n}.$$

Возьмем оценку $T(F) = T(\hat{F_n})$, откуда $\hat{se}^2 = \frac{T(\hat{F_n})(1-T(\hat{F_n}))}{n}$.
Так как $T(\hat{F_n}) \sim N(T(F), \hat{se}^2)$, то $(1-\alpha)$-доверительный интервал будет иметь вид:

$$(T(\hat{F_n}) - z_{1-\alpha/2} \cdot \hat{se}, T(\hat{F_n}) + z_{1-\alpha/2} \cdot \hat{se})$$

Ответ: $(T(\hat{F_n}) - z_{1-\alpha/2} \cdot \hat{se}, T(\hat{F_n}) + z_{1-\alpha/2} \cdot \hat{se})$, где
$T(\hat{F_n}) = \hat{F_n}(b) - \hat{F_n}(a)$,
$\hat{se} = \sqrt{\frac{T(\hat{F}_n)(1-T(\hat{F}_n))}{n}}$,
$z_{1-\alpha/2}$ – квантиль стандартного нормального распределения.

Бутстреп

Задача 8 [3 балла]

Пусть $X^n = \{X_1, \dots, X_n\} \sim N(\mu, \sigma^2)$, $\theta = e^\mu$ и $\hat{\theta} = e^{\overline{X_n}}$. Найдите аналитически плотность распределения $p_{\hat{\theta}}(x)$ оценки $\hat{\theta} = e^{\overline{X^n}}$, 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

Если $X_i \sim N(\mu, \sigma^2)$, то $e^{X_i} \sim LnN(\mu, \sigma^2)$ — логнормальное распределение. Известна плотность такого распределения:

$$p(x) = \frac{1}{x\sigma\sqrt{2\pi}} exp\big(- \frac{(\log x - \mu)^2}{2\sigma^2}\big)$$

Усредненная сумма n одинаковых нормально распределенных случайных величин $N(\mu, \sigma^2)$ распределена как $N(\mu, \sigma^2/n)$, следовательно $\hat{\theta} = e^{\overline{X_n}}$ распределена как $LnN(\mu, \sigma^2/n)$ с плотностью распределения:

$$p_{\hat{\theta}}(x) = \frac{\sqrt{n}}{x\sigma\sqrt{2\pi}} exp\big(- \frac{n(\log x - \mu)^2}{2\sigma^2}\big)$$

Используя известные матожидание и дисперсию логнормального распределения, получаем:

$$\mathbb{E}(\hat{\theta}) = \mathbb{E}[X \sim LnN(\mu, \sigma^2/n)] = e^{\mu + \sigma^2/2n}$$

$$\mathbb{V}(\hat{\theta}) = \mathbb{V}[X \sim LnN(\mu, \sigma^2/n)] = e^{2\mu + \sigma^2/n}(e^{\sigma^2/n} - 1)$$

$$bias = \mathbb{E}(\hat{\theta}) - \theta = e^{\mu + \sigma^2/2n} - e^\mu = e^\mu(e^{\sigma^2/2n} - 1) \neq 0 \Rightarrow \text{оценка смещенная}$$

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

$$MSE = bias^2 + se^2 = e^{2\mu}(e^{\sigma^2/2n} - 1)^2 + e^{2\mu + \sigma^2/n}(e^{\sigma^2/n} - 1) = e^{2\mu}(e^{\sigma^2/2n} - 1)(e^{\sigma^2/2n} - 1 + e^{\sigma^2/n})$$

$$bias = e^\mu(e^{\sigma^2/2n} - 1) \to 0, se = e^{\mu + \sigma^2/2n}\sqrt{e^{\sigma^2/n} - 1} \to 0 \text{ при } n \to \infty \Rightarrow \text{оценка состоятельна.}$$

Ответ: $p_{\hat{\theta}}(x) = \frac{\sqrt{n}}{x\sigma\sqrt{2\pi}} exp\big(- \frac{n(\log x - \mu)^2}{2\sigma^2}\big), \mathbb{E}(\hat{\theta}) = e^{\mu + \sigma^2/2n}, \mathbb{V}(\hat{\theta}) = e^{2\mu + \sigma^2/n}(e^{\sigma^2/n} - 1), bias = e^\mu(e^{\sigma^2/2n} - 1), se = e^{\mu + \sigma^2/2n}\sqrt{e^{\sigma^2/n} - 1}, MSE = e^{2\mu}(e^{\sigma^2/2n} - 1)(e^{\sigma^2/2n} - 1 + e^{\sigma^2/n})$, смещенная, состоятельна.