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

Школа Анализа Данных

Задачи

Задача 1 [1 балла]

Пусть $\boldsymbol{X}^n = \{X_1, X_2, \dots\}$ — независимые одинаково распределенные (н.о.р.) случайные величины с конечными средним $\mu = \mathbb{E}(X_1)$ и дисперсией $\sigma^2 = \mathbb{V}(X_1)$. Покажите, что величины

$$\langle \boldsymbol{X}^n \rangle = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \hat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \langle \boldsymbol{X}^n \rangle)^2.$$

являются *несмещенными* и *состоятельными* оценками среднего μ и дисперсии σ^2, т.е. что

- $\mathbb{E}(\langle \boldsymbol{X}^n \rangle) = \mu$ и $\langle \boldsymbol{X}^n \rangle \xrightarrow{\mathsf{P}} \mu$,
- $\mathbb{E}(\hat{S}_n^2) = \sigma^2$ и $\hat{S}_n^2 \xrightarrow{\mathsf{P}} \sigma^2$.

Замечание. *Конкретно в задачах статистики зачастую под $\boldsymbol{X}^n$ понимается выборка независимых значений случайной величины X. В таком случае $\langle \boldsymbol{X}^n \rangle$ и $\hat{S}_n^2$ — оценки среднего и дисперсии по выборке.*

Задача 2 [1 балла]

Пусть $\boldsymbol{X}^n = \{X_1, X_2, \dots, X_n\}$ и $\boldsymbol{Y}^m = \{Y_1, Y_2, \dots, Y_m\}$ — две выборки н.о.р. случайных величин объема n и m, полученных из одного и того же распределения. Пусть $\hat{S}_X^2$ и $\hat{S}_Y^2$ — несмещенные оценки дисперсий по выборкам $\boldsymbol{X}^n$ и $\boldsymbol{Y}^m$ соответственно. Выразите несмещенную оценку дисперсии $\hat{S}_{X,Y}$ суммарной выборки через $\hat{S}_X^2$ и $\hat{S}_Y^2$ и средние $\langle \boldsymbol{X}^n \rangle$ и $\langle \boldsymbol{Y}^m \rangle$.

Задача 3 [2 балла]

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

Задача 4 [2 балла]

Пусть $\boldsymbol{X}^n = \{X_1, \dots, X_n\} \sim \mathcal{N}(0, \sigma^2)$. Пусть для оценки параметра σ нормального распределения используется выборочное линейное отклонение $\hat{\sigma} = \langle |\boldsymbol{X}^n| \rangle = n^{-1}\sum_{i=1}^{n}|X_i|$. Найдите bias, se, MSE оценки $\hat{\sigma}$. Является ли оценка несмещенной? Если «нет», то постройте исправленную оценку. Найдите se исправленной оценки. Является ли исправленная оценка $\hat{\sigma}$ состоятельной?

Задача 5 [3 балла]

Пусть $\boldsymbol{X}^n = \{X_1, \dots, X_n\} \sim \mathcal{N}(\mu, \sigma^2)$, $\theta = e^{\mu}$ 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?

Задача 6 [2 балла]

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

Задача 7 [2 балла]

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

Задача 8 [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 $\widehat{\mathrm{se}}$ оценки $T(\boldsymbol{x}^n)$. Постройте асимптотический нормальный 95%-ый доверительный интервал для $T(F)$.

Задача 9 [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?

Задача 10 [2 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 11 [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.