

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

Школа Анализа Данных

Задача 1 [2 балла]

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

Задача 2 [2 балла]

Пусть н.о.р. выборка¹ $\mathbf{X}_n = \{X_1, \dots, X_n\}$ получена из распределения $F(x) = 1 - e^{\theta-x}$, $x > \theta$. Постройте несмещенную оценку $\hat{\theta}$ параметра θ . Найдите se и MSE этой оценки. Является ли построенная оценка состоятельной?

Задача 3 [2 балла]

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

Задача 4 [2 балла]

Пусть $\mathbf{X}_n = \{X_1, \dots, X_n\} \sim F(x)$, и пусть $\hat{F}_n(x)$ — эмпирическая функция распределения. Для фиксированных чисел $a, b \in \mathbb{R}$, 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 $\hat{\sigma}$ стандартного отклонения и $(1 - \alpha)$ -доверительный интервал.

Задача 5 [2 балла]

Скачайте данные об амплитудах землетрясений вблизи Фиджи. Постройте график для $\hat{F}_n(x)$. Подсчитайте и постройте приближенные 95% доверительные интервалы для значений $F(x)$. Подсчитайте и постройте приближенный 95% доверительный интервал для значения $F(4.9) - F(4.3)$.

Задача 6 [2 балла]

В 1975 г. проводился эксперимент, в ходе которого пытались выяснить, действительно ли распыление специальных реагентов позволяет повысить вероятность дождя. Таким образом в 26 облаках было проведено распыление реагентов, также была рассмотрена контрольная группа из 26 других облаков. Пусть θ — разность в средних значениях выпавших осадках из этих двух групп облаков. Оценить по данным CloudSeeding величину θ , оценить стандартную ошибку оценки, построить 95% и 99% доверительные интервалы. Какие выводы можно сделать на основе полученных результатов?

Задача 7 [3 балла]

Провести моделирование, чтобы сравнить различные типы доверительных интервалов, построенных с помощью бутстрепа. $T(F) = \int (x - \mu)^3 dF(x) / \sigma^3$ — эксцесс, где F — распределение χ_k^2 с тремя степенями свободы ($k = 3$). Постройте 95% доверительные интервалы для $T(F)$ по выборке $\mathbf{X}_n = \{X_1, \dots, X_n\}$, используя три подхода на основе бутстрепа. Поэкспериментируйте с размерами выборки n , чтобы добиться приемлемой схожести (оценка эксцесса не должна отличаться от его истинного значения более чем на 10%).

Примечание. В данной может потребоваться в качестве значений n взять достаточно большие числа (порядка 1000-10000).

¹Выборка независимых одинаково распределенных случайных величин

Задача 8 [3 балла]

Пусть $\mathbf{X}_n = \{X_1, \dots, X_n\} \sim \text{Exp}(\lambda)$, $\theta = e^{\frac{1}{\lambda}}$ и $\hat{\theta} = e^{\overline{\mathbf{X}_n}}$. Найдите аналитически распределение оценки $\hat{\theta} = e^{\overline{\mathbf{X}_n}}$, математическое ожидание $\mathbb{E}(\hat{\theta})$, дисперсию $\mathbb{V}(\hat{\theta})$, а также bias, se, MSE оценки $\hat{\theta}$. Является ли оценка $\hat{\theta}$ смещенной? Состоятельной?

$e^{\overline{\mathbf{X}_n}}$ — оценка на основе выборки $\overline{\mathbf{X}_n}$. Обозначим через $\mathbb{E}(\hat{\theta}^* | \mathbf{X}_n)$ математическое ожидание бутстрепной оценки при фиксированной выборке \mathbf{X}_n . Верно ли, что $e^{\overline{\mathbf{X}_n}} \leq \mathbb{E}(\hat{\theta}^* | \mathbf{X}_n)$? Если «да», то почему?

Задача 9 [2 балла]

Пусть $\mathbf{X}_n = \{X_1, \dots, X_n\} \sim \text{Exp}(\lambda)$, $\theta = e^{\frac{1}{\lambda}}$ и $\hat{\theta} = e^{\overline{\mathbf{X}_n}}$. Сгенерируйте выборку \mathbf{X}_n из $n = 1000$ наблюдений для $\lambda = 0.25$. 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 $p_{\hat{\theta}}(x)$, вычисленным аналитически в предыдущей задаче. Используя бутстреп, подсчитайте величину se и постройте тремя способами 95% доверительный интервал для θ .