Домашнее задание №3 по курсу «Математическая Статистика в Машинном Обучении»

Школа Анализа Данных

Задачи

Теоретический блок

Задача 1 [1 балл]

Пусть дана обучающая выборка $\{(\boldsymbol{X}, \boldsymbol{y})\colon \boldsymbol{X} \in \mathbb{R}^{n\times d}, \boldsymbol{y} \in \mathbb{R}^n\}$, $n \geq d$. Предположим, что справедлива следующая модель линейной регрессии:

$$y = \boldsymbol{x}^T\boldsymbol{w} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

где $\boldsymbol{w}$ — истинный, но неизвестный нам вектор весов. Пусть $\hat{\boldsymbol{w}}$ — MLE-оценка вектора весов $\boldsymbol{w}$.

Предположим, к нам поступили тестовые данные $\boldsymbol{X}^* \in \mathbb{R}^{m\times d}$, для которых с помощью оценки $\hat{\boldsymbol{w}}$ предсказываем вектор $\boldsymbol{y}^* \in \mathbb{R}^m$. Найдите математическое ожидание и матрицу ковариаций для вектора $\boldsymbol{y}^*$ (при условии фиксированной матрицы дизайна $\boldsymbol{X}$).

Задача 2 [1 балл]

Пусть дана выборка $(\boldsymbol{X}, \boldsymbol{t}) = \{(\boldsymbol{x}_i, t_i)\colon \boldsymbol{x}_i \in \mathbb{R}^d, t_i \in \mathbb{R}\}_{i=1}^n$, ($\boldsymbol{X} \in \mathbb{R}^{n\times d}$, $\boldsymbol{t} \in \mathbb{R}^n$, $n \geq d$). Предположим справедливость следующей модели данных

$$t = \boldsymbol{x}^T\boldsymbol{w} + \varepsilon(\boldsymbol{x}),$$

где $\varepsilon(\boldsymbol{x}) \sim \mathcal{N}(0, \sigma(\boldsymbol{x})^2)$. Найдите MLE-оценку на вектор весов $\boldsymbol{w}$ в данном случае.

Задача 3 [2 балла]

Пусть дана выборка $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_i, y_i)\colon x_i, y_i \in \mathbb{R}\}_{i=1}^n$. Пусть данные соответствуют модели

$$y_i = \beta x_i + \varepsilon_i,$$

где $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. При этом значения $\boldsymbol{x}$ наблюдаются с ошибкой, т.е. представлена не выборка $(\boldsymbol{x}, \boldsymbol{y})$, а выборка $(\boldsymbol{z}, \boldsymbol{y}) = \{(z_i, y_i)\colon z_i, y_i \in \mathbb{R}\}_{i=1}^n$, где $z_i = x_i + \delta_i$, $\delta_i \sim \mathcal{N}(0, \tau^2)$. Шумы ε_i и δ_i независимы. Оценим величину β, используя стандартный метод наименьших квадратов согласно формуле

$$\hat{\beta} = \frac{\sum_{i=1}^n z_i y_i}{\sum_{i=1}^n z_i^2}.$$

Докажите, что оценка $\hat{\beta}$ не является состоятельной. Для этого покажите, что $\hat{\beta} \xrightarrow{\mathsf{P}} a\beta$ при $n \to \infty$. Найдите явное выражение для a в предположении, что точки $\{x_i\}_{i=1}^n$ поступают из некоторого распределения $F(x)$ с конечными первыми и вторыми моментами $\mathbb{E}(X)$ и $\mathbb{E}(X^2)$.

Задача 4 [2 балла]

Пусть дана обучающая выборка $\{(\boldsymbol{x}, \boldsymbol{y})\colon \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{y} \in \mathbb{R}^n\}$. Предположим, что справедлива следующая модель линейной регрессии:

$$y = w_0 + w_1 x + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Сконструируйте асимтотический тест Вальда для проверки гипотезы $H_0\colon w_1 = \alpha w_0$.

Внимание. Замечание про асимптотичность тут не просто так.

Задача 5 [2 балла]

Рассмотрим задачу восстановления регрессии. Модель регрессии имеет вид

$$t = \boldsymbol{x}^T\boldsymbol{w} + \varepsilon,$$

где $\varepsilon \sim \mathcal{N}(0, \beta^{-1})$, и на веса $\boldsymbol{w}$ наложено априорное распределение вида $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{w}_0, \boldsymbol{S}_0)$. Пусть дана выборка $(\boldsymbol{X}, \boldsymbol{t}) = \{(\boldsymbol{x}_i, t_i)\colon \boldsymbol{x}_i \in \mathbb{R}^d, t_i \in \mathbb{R}\}_{i=1}^n$. Найдите апостериорное распределение $p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{t})$.

Задача 6 [2 балла]

Пусть $\boldsymbol{x}^n \sim f(\cdot)$, и пусть $\hat{f}(\cdot) = \hat{f}(\cdot; \boldsymbol{x}^n)$ обозначает ядерную оценку плотности на основе ядра

$$K(x) = \begin{cases} 1, & x \in \left(-\frac{1}{2}, \frac{1}{2}\right); \\ 0, & \text{в противном случае.} \end{cases}$$

Найдите $\mathbb{E}[\hat{f}(x)]$ и $\mathbb{V}[\hat{f}(x)]$. Покажите, что если $h \to 0$ и $nh \to \infty$ при $n \to \infty$, то $\hat{f}(x) \xrightarrow{\mathsf{P}} f(x)$ при $n \to \infty$.

Примечание. В ответе может быть использована истинная плотность $f(x)$.

Задача 7 [4 балла]

Рассмотрим задачу непараметрической оценки плотности распределения $p(x)$ по выборке $\boldsymbol{X}^{(N)}$. Обозначим через $\hat{p}(x; \boldsymbol{X}^{(N)})$ оценку плотности, полученную некоторым образом по выборке $\boldsymbol{X}^{(N)}$. Оценка риска для $\hat{p}(x; \boldsymbol{X}^{(N)})$ имеет вид:

$$\hat{J}(h) = \int (\hat{p}(x; \boldsymbol{X}^{(N)}))^2 dx - \frac{2}{N} \sum_{i=1}^{N} \hat{p}(X_i; \boldsymbol{X}^{(N \setminus i)}),$$

где $\hat{p}(\cdot; \boldsymbol{x}^{(N\setminus i)})$ — оценка плотности распределения на основе выборки $\boldsymbol{X}^{(N\setminus i)}$, т.е. выборки без объекта X_i.

- (Гистограммная оценка) Разобьем диапазон наблюдаемых значений $\boldsymbol{X}^{(N)}$ на бины ширины h. Пусть в итоге значения $\boldsymbol{X}^{(n)}$ укладываются в M последовательных бинов $B_1, \ldots, B_M$. Пусть N_m — количество объектов выборки, попавших в B_m ($\sum_m N_m = N$). Пусть $\hat{p}_m$ — доля объектов выборки, попавших в бин B_m:

  $$N_m = \sum_{i=1}^{N} I\left[X_i \in B_m\right], \quad \hat{p}_m = \frac{N_m}{N}.$$

  Покажите, что в случае гистограммной оценки плотности оценка риска имеет вид:

  $$\hat{J}(h) = \frac{2}{h(N-1)} - \frac{N+1}{h(N-1)} \sum_{m=1}^{M} \hat{p}_m^2.$$

  Докажите или опровергните равенство

  $$\mathbb{E}[\hat{J}(h)] = \mathbb{E}[J(h)].$$

  Если равенство не верно, то чему равно $\Delta J(h) = \mathbb{E}[\hat{J}(h)] - \mathbb{E}[J(h)]$?
- (Ядерная оценка) Покажите, что в случае ядерной оценки плотности оценка риска имеет вид:

  $$\hat{J}(h) \approx \frac{1}{hN^2} \sum_{i,j} K^*\left(\frac{X_i - X_j}{h}\right) + \frac{2}{Nh} K(0),$$

  где $K^*(x) = K^{(2)}(x) - 2K(x)$ и $K^{(2)}(z) = \int K(z-y)K(y)dy$. В частности, если $K(x)$ — это плотность нормального распределения $\mathcal{N}(0,1)$, т.е. гауссово ядро, то $K^{(2)}(z)$ — плотность распределения $\mathcal{N}(0,2)$.
  Докажите или опровергните равенство

  $$\mathbb{E}[\hat{J}(h)] = \mathbb{E}[J(h)].$$

  Если равенство не верно, то чему равно $\Delta J(h) = \mathbb{E}[\hat{J}(h)] - \mathbb{E}[J(h)]$?

Задача 8 [3 балла]

Рассмотрим задачу непараметрической регрессии:

$$Y_i = f(X_i) + \varepsilon_i, \quad i \in \{1, \ldots, n\}, \quad X_i \in \mathbb{R}, \quad Y_i \in \mathbb{R}.$$

где ε_i и X_i независимы, $\mathbb{E}\varepsilon_i = 0$, $\mathbb{V}\varepsilon_i = \sigma^2$, выборка $\{X_i\}_{i=1}^n$ одномерная и сэмплируется из отрезка $[0,1]$. Необходимо по имеющимся данным оценить функцию регрессии $f(x) = \mathbb{E}(Y|X=x)$.

a) Рассмотрим следующее семейство функций

$$\mathfrak{F}_M = \left\{ f(x) = \sum_{i=1}^{M} c_i I[x \in B_i], c_i \in \mathbb{R}, i = \overline{1, M} \right\}, \text{ где } B_i = \left[\frac{i-1}{M}, \frac{i}{M}\right).$$

Последний отрезок B_M включает обе граничные точки. Найдите функцию из класса $\mathfrak{F}_M$, которая минимизирует сумму квадратов ошибок:

$$r(x; \boldsymbol{X}^n) = \arg \min_{f(x) \in \mathfrak{F}_M} \sum_{i=1}^{n} \left(Y_i - f(X_i)\right)^2$$

b) Найдите функцию регрессии поточечно, решив в каждой точке x следующую оптимизационную задачу:

$$r(x; \boldsymbol{X}^n) = \arg\min_{y \in \mathbb{R}} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)(Y_i - y)^2,$$

где $K(x)$ — заданная ядерная функция, h — ширина ядра.

c) Какая оценка получится, если изменить задачу на следующую:

$$r(x; \boldsymbol{X}^n) = \arg\min_{a, b \in \mathbb{R}} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)(Y_i - a - bX_i)^2,$$

где $K(x)$ — заданная ядерная функция, h — ширина ядра?

Практический блок

Задача 9 [2 балла]

Винни-Пуху на день рождения Сова подарила 5 горшочков с медом, каждый приблизительно весом 1 кг (исходя из объема горшочка и плотности мёда). Однако из проверенных источников (от Пятачка), Винни-Пух получил информацию, что один горшочек предположительно содержит неправильный мёд, причем его вес должен отличаться от 1 кг (из-за содержания неправильных веществ). Для проведения следственных мероприятий у ослика Иа-Иа были изъяты самодельные весы. Взвесив каждый горшочек индивидуально, Винни-Пух обнаружил, что весы явно имеют некоторую неизвестную погрешность взвешивания, так что проделанные измерения не позволяют однозначно проверить информацию о неправильности мёда в одном из горшочков. Поэтому Винни-Пух почему-то решил взвешивать горшочки сразу по два, но как только он закончил эти 10 взвешиваний, как ослик Иа-Иа, пригрозив судебными разбирательствами, в принудительном порядке затребовал свои весы обратно, оставив Винни-Пуха с результатами 15-и взвешиваний.

Нам даны результаты этих взвешиваний — бинарная матрица $\boldsymbol{X} \in \{0,1\}^{n \times d}$, где $n = 15$ и $d = 5$, и вектор $\boldsymbol{y}$ с результатами взвешиваний (`honey_X.csv` и `honey_y.csv`). По этим данных для каждого горшочка найдите p-value для гипотезы о том, что данный горшочек содержит неправильный мёд. Если ли среди горшочков такой, который на уровне значимости 95% содержит неправильный мёд?

Дополнительное задание на 1 балл. Дисперсия веса горшочка зависит от дизайна взвешиваний (выбора матрицы $\boldsymbol{X}$). Возможно Винни-Пух ошибся, начав взвешивать горшочки сразу по два, и вместо этого стоило взвесить каждый горшочек отдельно от других ещё по два раза, получив в результате те же самые 15 взвешиваний до того момента, как весы были возвращены Иа-Иа. Найдите отношение дисперсии оценки веса горшочка в случае дизайна, предложенного Винни-Пухом, к дисперсии оценки веса горшочка в случае предложенного «индивидуального» дизайна. Какой дизайн лучше с точки зрения поиска горшочка с неправильным мёдом?

Задача 10 [3 балла]

Скачайте данные `data.csv`, содержащие 12 столбцов независимых переменных и 1 столбец с зависимой переменной. Первые 250 строк отведите под обучение, а оставшиеся 1250 под тест (да, под обучение отводим сильно меньше).

- Обучите простую линейную регрессию по обучающей выборке. Примените модель к тестовой выборке и найдите MSE.
- По обучающей выборке оцените наилучший набор признаков, описывающих выходную переменную. Используйте для этого статистику Cp Mallow, AIC-критерий, BIC-критерий, LOO-проверку. Выбор подмножества признаков проведите полным перебором. Позволяет ли какой-нибудь набор признаков получить значение MSE на тестовых данных меньше, чем на всех признаках?
  Внимание. В ответе должно быть понятно, какой набор признаков был выбран согласно каждому из критериев.

Задача 11 [4 балла]

Скачать данные со страницы курса (значения коэффициента преломления для разных типов стекла; первый столбец). Оценить плотность распределения этих значений, используя гистограмму и ядерную оценку. Для подбора ширины ячейки или ширины ядра использовать перекрестную проверку (кросс-проверку). Для выбранных значений ширины ячейки и ширины ядра построить 95%-ые доверительные интервалы для полученной оценки плотности.

Задача 12 [4 балла]

По данным из предыдущей задачи, используя в качестве выходной переменной y значения преломления для разных типов стекла, а в качестве входной переменной x — данные о содержании алюминия (четвертая переменная в матрице данных), восстановить зависимость между y и x с помощью ядерной непараметрической регрессии. Оценку ядра проводить с помощью перекрестной проверки. Построить 95%-ые доверительные интервалы для полученной оценки функции регрессии.