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

Школа Анализа Данных

Задачи

Задача 1 [12 баллов]

Рассмотрим вопросы, связанные с оценкой переобучения регрессионной модели. Пусть

- $(\boldsymbol{x}^n, \boldsymbol{y}^n)$ — заданная обучающая выборка;
- $r(\boldsymbol{x}) = \mathbb{E}_{y|\boldsymbol{x}}(y|\boldsymbol{x})$ — истинная функция регрессии;
- $a(\cdot) = a(\cdot; \boldsymbol{x}^n, \boldsymbol{y}^n) = a(\cdot; \boldsymbol{X}, \boldsymbol{y})$ — алгоритм, обученный на выборке $(\boldsymbol{x}^n, \boldsymbol{y}^n)$, матрично-векторное представление которой есть $(\boldsymbol{X}, \boldsymbol{y})$, $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, $\boldsymbol{y} \in \mathbb{R}^n$.
- $\overline{a}(\boldsymbol{x}; \boldsymbol{x}^n) = \mathbb{E}_{\boldsymbol{y}^n|\boldsymbol{x}^n} a(\boldsymbol{x}; \boldsymbol{x}^n, \boldsymbol{y}^n)$ — усредненный ответ всевозможных алгоритмов, обученных по выборке $(\boldsymbol{x}^n, \boldsymbol{y}^n)$ при фиксированной выборке $\boldsymbol{x}^n$ и всевозможных $\boldsymbol{y}^n$ из распределения $F_{y|x}(\cdot)$.

Мы хотим оценить качество модели, оценив **риск** $\mathcal{R}(a)$:

$$\mathcal{R}(a) = \mathbb{E}_{\boldsymbol{x}^n, \boldsymbol{y}^n} \mathbb{E}_{x,y} (y - a(x; \boldsymbol{x}^n, \boldsymbol{y}^n))^2.$$

Ошибки на предсказании и обучении (4 балла)

Ошибка на предсказании есть

$$\mathcal{R}(a; \boldsymbol{x}^n) = \frac{1}{n} \mathbb{E}_{\boldsymbol{y}^n|\boldsymbol{x}^n} \sum_{i=1}^{n} \mathbb{E}_{y_i^*|x_i} (y_i^* - a(x_i))^2.$$

Здесь y_i^* — случайная величина-отклик для входного признака x_i из обучающей выборки $\boldsymbol{x}^n$, т.е. y_i^* — повторное измерение отклика.

Ошибка на обучении есть

$$\mathcal{R}_{\text{tr}}(a; \boldsymbol{x}^n) = \frac{1}{n} \mathbb{E}_{\boldsymbol{y}^n|\boldsymbol{x}^n} \sum_{i=1}^{n} (y_i - a(x_i))^2.$$

Выборочная ошибка на обучении (оценка величины $\mathcal{R}_{\text{tr}}(a; \boldsymbol{x}^n)$) равна:

$$\hat{\mathcal{R}}_{\text{tr}}(a; \boldsymbol{x}^n, \boldsymbol{y}^n) = \frac{1}{n} \sum_{i=1}^{n} (y_i - a(x_i))^2,$$

$$\mathbb{E}_{\boldsymbol{y}^n|\boldsymbol{x}^n} (\hat{\mathcal{R}}_{\text{tr}}(a; \boldsymbol{x}^n, \boldsymbol{y}^n)) = \mathcal{R}_{\text{tr}}(a; \boldsymbol{x}^n).$$

Обозначим через $\Delta\mathcal{R}(a)$ разность между $\mathcal{R}(a; \boldsymbol{x}^n)$ и $\mathcal{R}_{\text{tr}}(a; \boldsymbol{x}^n)$:

$$\Delta\mathcal{R} = \mathcal{R}(a; \boldsymbol{x}^n) - \mathcal{R}_{\text{tr}}(a; \boldsymbol{x}^n).$$

Докажите, что

$$\boxed{\Delta\mathcal{R} = \frac{2}{n} \sum_{i=1}^{n} \text{Cov}(a(x_i), y_i)}.$$

Подсказка. Возможно для этого потребуется добавить и отнять y_i под скобками в выражении для $\mathcal{R}(a; \boldsymbol{x}^n)$.

Статистика Маллоу (4 балла)

Рассмотрим модель простой многомерной линейной регрессии:

$$y = \boldsymbol{x}^T \boldsymbol{\beta} + \varepsilon,$$

где $\boldsymbol{x}_i = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$ — вектор-признак, ε — случайный шум с $\mathbb{E}\varepsilon = 0, \mathbb{V}\varepsilon = \sigma^2$. В матричном виде для n объектов $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

где

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1d} \\ x_{21} & x_{22} & \ldots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nd} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}, \ \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Через $\hat{\boldsymbol{y}}$ обозначим прогноз модели на объектах из обучающей выборки: $\hat{\boldsymbol{y}} = a(\boldsymbol{X})$. Покажите, что

$$\boxed{\hat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}}, \tag{1}$$

где $\boldsymbol{S} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ в случае обычной регрессии и $\boldsymbol{S} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T$ в случае ридж-регрессии. Покажите, что если имеет место равенство (1), то

$$\boxed{\sum_{i=1}^{n} \operatorname{Cov}(\hat{y}_i, y_i) = \operatorname{trace}(\boldsymbol{S})\sigma^2}. \tag{2}$$

Подсчитайте значение $\operatorname{trace}(\boldsymbol{S})$ в случае обычной регрессии.

LOO-проверка (4 балла)

LOO-оценка риска имеет вид

$$\hat{\mathcal{R}}_{\text{LOO}}(a; \boldsymbol{x}^n, \boldsymbol{y}^n) = \frac{1}{n}\sum_{i=1}^{n}(y_i - a(\boldsymbol{x}_i; \boldsymbol{x}^{(n\setminus i)}, \boldsymbol{y}^{(n\setminus i)}))^2,$$

где $(\boldsymbol{x}^{(n\setminus i)}, \boldsymbol{y}^{(n\setminus i)})$ — выборка без i-го объекта. Докажите, что в случае обычной многомерной линейной регрессии LOO-оценку можно вычислить гораздо проще по формуле

$$\boxed{\hat{\mathcal{R}}_{\text{LOO}}(a; \boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - a(x_i; \boldsymbol{X}, \boldsymbol{y})}{1 - \|\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\|_{ii}}\right)^2}.$$

Подсказка. Вам потребуется тождество Шермана-Моррисона-Вудбери:

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

Задача 2 [1 балл]

Для регрессионной модели $y = \beta_0 + \beta_1 x + \varepsilon$ сконструировать тест Вальда для различения двух гипотез $H_0\colon \beta_1 = 17\beta_0$ и $H_1\colon \beta_1 \neq 17\beta_0$.

Задача 3 [2 балла]

Рассмотрим следующую модель регрессии:

$$y = \sum_{j=1}^{k}\beta_j x_j + \varepsilon = \boldsymbol{x}^T\boldsymbol{\beta} + \varepsilon,$$

где шум $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ и дисперсия σ^2 известна. Показать, что модель с наибольшим значением AIC является моделью с наименьшим значением статистики Mallow C_p.

Задача 4 [6 баллов]

Скачать данные о связи между уровнем антигена, специфичного при наличии рака простаты, с рядом клинических показателей, которые были измерены у мужчин непосредственно перед проведением операции. Используя для оценки обобщающей способности регрессии критерии типа AIC, BIC, 10-кратную и 5-кратную кросс-валидации, а также leave-one-out cross-validation, выбрать наилучшее подмножества регрессоров, описывающих выходную переменную. Выбор подмножества признаков производить полным перебором. Сравнить с выбором подмножества признаков, в котором применяется жадный перебор как прямым, так и обратным вариантами пошагового выбора. Проанализировать полученные результаты.

Задача 5 [4 балла]

Пусть дана модель регрессии вида

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

где $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, $\boldsymbol{\varepsilon} \in \mathbb{R}^n$. Допустим, что шум нормально распределен, то есть $\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}\right)$, и априорное распределение коэффициентов линейной регрессии также нормальное, т.е. $\boldsymbol{\beta} \sim \mathcal{N}\left(0, \tau\boldsymbol{I}\right)$, где τ — некоторый параметр. Показать, что оценка параметров $\boldsymbol{\beta}$ с помощью гребневой регрессии совпадает по структуре со средним значением (и модой) апостериорного распределения параметров регрессии $\boldsymbol{\beta}$, найти отсюда связь между параметром регуляризации в гребневой регрессии λ и параметрами σ^2 и τ.

Задача 6 [1 балла]

Рассмотрим оптимизационную задачу

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_1\|\boldsymbol{\beta}\|_1^1 + \lambda_2\|\boldsymbol{\beta}\|_2^2\right).$$

Определим искусственные данные

$$\boldsymbol{X}^* = \frac{1}{\sqrt{1+\lambda_2}}\begin{pmatrix}\boldsymbol{X}\\ \sqrt{\lambda_2}\boldsymbol{I}\end{pmatrix} \in \mathbb{R}^{(n+d)\times d}, \qquad \boldsymbol{y}^* = \begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{0}\end{pmatrix} \in \mathbb{R}^{n+d},$$

Пусть $\gamma = \lambda_1/\sqrt{1+\lambda_2}$ и $\boldsymbol{\beta}^* = \sqrt{1+\lambda_2}\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}}^* = \arg\min_{\boldsymbol{\beta}^*} \left(\|\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*\|^2 + \gamma\|\boldsymbol{\beta}^*\|_1^1\right).$$

Докажите, что

$$\boxed{\hat{\beta} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\boldsymbol{\beta}}^*}.$$

Задача 7 [4 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

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

(a) Показать, что

$$\mathbb{E}\hat{f}(x) = \frac{1}{h}\int_{x-(h/2)}^{x+(h/2)} f(y)dy.$$

и

$$\mathbb{V}\hat{f}(x) = \frac{1}{nh^2}\left[\int_{x-(h/2)}^{x+(h/2)} f(y)dy - \left(\int_{x-(h/2)}^{x+(h/2)} f(y)dy\right)^2\right].$$

(b) Показать, что если $h \to 0$ и $nh \to \infty$ при $n \to \infty$, то $\hat{f}(x) \xrightarrow{\mathsf{P}} f(x)$ при $n \to \infty$.

Задача 8 [5 баллов]

Скачать данные о значении коэффициента преломления для разных типов стекла; первый столбец). Оцените плотность распределение этих значений, используя гистограмму и ядерную оценку. Для подбора ширины ячейки или ширины ядра используйте перекрестную проверку (кросс-проверку). Для выбранных значений ширины ячейки и ширины ядра постройте 95%-ые доверительные интервалы для полученной оценки плотности.

Задача 9 [5 баллов]

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. Постройте 95%-ые доверительные интервалы для полученной оценки функции регрессии.

Задача 10 [5 баллов]

Пусть

$$\hat{J}(h) = \int (\hat{f}(x))^2 dx - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{-i}(x),$$

где $\hat{f}_{-i}(\cdot)$ — оценка плотности распределения на основе выборки $\boldsymbol{x}^{(n\setminus i)}$, т.е. выборки без объекта i.

Доказать, что для любого $h > 0$

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

Также

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

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 $\mathcal{N}(0,2)$.