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

Школа Анализа Данных

Задачи

Теоретический блок

Задача 1 [1 балл]

Рассмотрим задачу оптимизации, решаемую в Elastic Net:

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda_1\|\boldsymbol{w}\|_1^1 + \lambda_2\|\boldsymbol{w}\|_2^2 \to \min_{\boldsymbol{w}}$$

Покажем, что данную задачу оптимизации можно преобразовать к виду, содержащему составляющую только для ℓ_1-регуляризации. Для этого определим искусственные данные

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

Покажите, что при таком преобразовании задачу выше можно свести к задаче

$$\|\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{w}^*\|_2^2 + \gamma\|\boldsymbol{w}^*\|_1^1 \to \min_{\boldsymbol{w}^*}.$$

Найдите требуемые для этого значения α и γ, выразив последние через λ_1 и λ_2. Найдите связь между между решениями оптимизационных задач $\hat{\boldsymbol{w}}$ и $\hat{\boldsymbol{w}}^*$.

Решение:

$$(\boldsymbol{y}^*)^T\boldsymbol{y}^* = \boldsymbol{y}^T\boldsymbol{y}$$

$$(\boldsymbol{X}^*)^T\boldsymbol{X}^* = \frac{1}{1+\lambda_2}(\boldsymbol{X}^T\boldsymbol{X} + \alpha^2\boldsymbol{I}) => \boldsymbol{X}^T\boldsymbol{X} = (1+\lambda_2)(\boldsymbol{X}^*)^T\boldsymbol{X}^* - \alpha^2\boldsymbol{I}$$

$$(\boldsymbol{y}^*)^T\boldsymbol{X}^* = \boldsymbol{y}^T\boldsymbol{X}\frac{1}{\sqrt{1+\lambda_2}} => \boldsymbol{y}^T\boldsymbol{X} = \sqrt{1+\lambda_2}(\boldsymbol{y}^*)^T\boldsymbol{X}^*$$

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda_1\|\boldsymbol{w}\|_1^1 + \lambda_2\|\boldsymbol{w}\|_2^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) + \lambda_1\|\boldsymbol{w}\|_1^1 + \lambda_2\boldsymbol{w}^T\boldsymbol{w} =$$

$$= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{w} + \boldsymbol{w}^T(\boldsymbol{X}^T\boldsymbol{X} + \lambda_2\boldsymbol{I})\boldsymbol{w} + \lambda_1\|\boldsymbol{w}\|_1^1 =$$

$$= (\boldsymbol{y}^*)^T\boldsymbol{y}^* - 2(\boldsymbol{y}^*)^T\boldsymbol{X}^*\sqrt{1+\lambda_2}\boldsymbol{w} + \boldsymbol{w}^T((\boldsymbol{X}^*)^T\boldsymbol{X}^*(1+\lambda_2) - \alpha^2\boldsymbol{I} + \lambda_2\boldsymbol{I})\boldsymbol{w} + \lambda_1\|\boldsymbol{w}\|_1^1$$

Обозначив $\boldsymbol{w}^* = \boldsymbol{w}\sqrt{1+\lambda_2}$ и приравняв $\alpha^2 = \lambda_2$ (чтобы избавиться от l_2 регуляризации), получим нужную нам форму:

$$(\boldsymbol{y}^*)^T\boldsymbol{y}^* - 2(\boldsymbol{y}^*)^T\boldsymbol{X}^*\boldsymbol{w}^* + (\boldsymbol{X}^*\boldsymbol{w}^*)^T\boldsymbol{X}^*\boldsymbol{w}^* + \frac{\lambda_1}{\sqrt{1+\lambda_2}}\|\boldsymbol{w}^*\|_1^1 = \|\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{w}^*\|_2^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}}\|\boldsymbol{w}^*\|_1^1$$

Ответ: $\alpha = \sqrt{\lambda_2}, \gamma = \frac{\lambda_1}{\sqrt{1+\lambda_2}}$

Задача 2 [1 балл]

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

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

Покажите, что модель с наибольшим значением AIC является моделью с наименьшим значением статистики Mallow C_p.

Решение:

- Рассмотрим AIC: $l_S - |S| \to \max$

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - x_i^T w)} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\|y - X^T w\|_2^2}$$

$$l_S = \log L = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|y - Xw\|_2^2$$

Максимизация l_S эквивалентна максимизации $-\frac{1}{2\sigma^2}\|y - Xw\|_2^2$, т.к. первое слагаемое не зависит от w

AIC: $\boxed{-\frac{1}{2\sigma^2}\|y - Xw\|_2^2 - d}$ (*)

- Рассмотрим C_p-Mallow:

$$\frac{\hat{R}_{tr}(S)}{2\sigma^2} + |S| = \boxed{\frac{\|y - Xw\|_2^2}{2\sigma^2} + d} \quad (**)$$

Получаем, что модель с наибольшим значением AIC (*) является моделью с наименьшим значением статистики Mallow C_p (**)

ч.т.д.

Задача 3 [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 модель линейной регрессии:

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

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

Решение:

$$W = \frac{w_1 - \alpha w_0}{\hat{se}(w_1 - \alpha w_0)}$$

Найдем оценку для стндартного отклонения. Известно (с семинара), что

$$\mathbb{V}(\hat{w}) = \sigma^2 (X^T X)^{-1} = \frac{\sigma^2}{n(\langle X^2\rangle - \langle X\rangle^2)} \begin{pmatrix} \langle X^2\rangle & -\langle X\rangle \\ -\langle X\rangle & 1 \end{pmatrix}, \text{ где } \langle X^k\rangle = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

Значение дисперсии известно, поэтому найдем стандартное отклонение:

$$\hat{se}(w_1 - \alpha w_0) = \sqrt{\mathbb{V}(w_1 - \alpha w_0)} = \sqrt{\mathbb{V}(w_1) + \mathbb{V}(-\alpha w_0) + 2cov(w_1, -\alpha w_0)} = \sqrt{\mathbb{V}(w_1) + \alpha^2 \mathbb{V}(w_0) - 2\alpha cov(w_1, w_0)} =$$

$$= \frac{\sigma}{\sqrt{n(\langle X^2 \rangle - \langle X \rangle^2)}} \sqrt{1 + \alpha^2 \langle X^2 \rangle + 2\langle X \rangle}$$

Критерия Вальда размера β отклоняет нулевую гипотезу в пользу альтернативной тогда и только тогда, когда $|W| > z_{\beta/2}$, т.е. когда $|w_1 - \alpha w_0| > z_{\beta/2} \frac{\sigma}{\sqrt{n(\langle X^2 \rangle - \langle X \rangle^2)}} \sqrt{1 + \alpha^2 \langle X^2 \rangle + 2\langle X \rangle}$

Задача 4 [2 балла]

Пусть дана обучающая выборка $\{(\boldsymbol{X}, \boldsymbol{y})\colon \boldsymbol{X} \in \mathbb{R}^{n \times d}, \boldsymbol{y} \in \mathbb{R}^n\}$, причем данные соответствуют модели 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 $\boldsymbol{X}$).

Решение:

Из семинара известно, что $\hat{w} = (X^T X)^{-1} X^T y$.

$$\mathbb{E}y^* = \mathbb{E}(X^* \hat{w}) = X^* \mathbb{E}\hat{w} = X^* \mathbb{E}((X^T X)^{-1} X^T y) = X^* (X^T X)^{-1} X^T \mathbb{E}(Xw + \varepsilon) = X^* (X^T X)^{-1} X^T X w = X^* w$$

$$\begin{aligned} cov(y^*) = cov(y^*, y^*) = \mathbb{E}[(y^* - \mathbb{E}y^*)(y^* - \mathbb{E}y^*)^T] = \mathbb{E}[y^* (y^*)^T] - \mathbb{E}y^* \mathbb{E}(y^*)^T = \\ = \mathbb{E}[X^* (X^T X)^{-1} X^T y y^T X (X^T X)^{-1} (X^*)^T] - X^* w w^T (X^*)^T = \\ = X^* w w^T (X^*)^T + \sigma^2 X^* (X^T X)^{-1} (X^*)^T - X^* w w^T (X^*)^T = \sigma^2 X^* (X^T X)^{-1} (X^*)^T \end{aligned}$$

Ответ: $\mathbb{E}y^* = X^* w, \qquad cov(y^*) = \sigma^2 X^* (X^T X)^{-1} (X^*)^T$

Задача 5 [2 балла]

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

Решение:

$$L = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2(x_i)}} e^{-\frac{1}{2} \cdot \frac{(t_i - x_i^T w)^2}{\sigma^2(x_i)}} = \underbrace{\prod_{i=1}^n (2\pi\sigma^2(x_i))^{-1/2}}_{\text{не зависит от } w} \cdot e^{-\frac{1}{2} \sum_{i=1}^n \left(\frac{t_i - x_i^T w}{\sigma(x_i)}\right)^2} \to max$$

$$L \to max \iff +\frac{1}{2} \sum_{i=1}^n \left(\frac{t_i - x_i^T w}{\sigma(x_i)}\right)^2 \to min$$

$$\sum_{i=1}^n \left(\frac{t_i}{\sigma(x_i)} - \frac{x_i^T w}{\sigma(x_i)}\right)^2 = \sum_{i=1}^n \left(t_i^* - (x_i^*)^T w\right)^2 = \|T^* - X^* w\|_2^2, \text{ где } t_i^* = \frac{t_i}{\sigma(x_i)} \text{ и } x_i^* = \frac{x_i}{\sigma(x_i)}$$

$$T^* = \begin{pmatrix} t_1^* \\ t_2^* \\ \vdots \\ t_n^* \end{pmatrix} \quad X^* = \begin{pmatrix} x_1^* \\ x_2^* \\ \vdots \\ x_n^* \end{pmatrix}$$

$$\text{Тогда } T^* = diag\left(\frac{1}{\sigma(x_1)}, \ldots, \frac{1}{\sigma(x_n)}\right) \cdot T, \quad X^* = diag\left(\frac{1}{\sigma(x_1)}, \ldots, \frac{1}{\sigma(x_n)}\right) X$$

$$\|T^* - X^* w\|_2^2 \to \min_w \implies \hat{w} = ((X^*)^T X^*)^{-1} (X^*)^T T^*$$

Ответ: $\hat{w} = (DX)^T DX)^{-1}(DX)^T DT$, где $D = \text{diag}(\frac{1}{\sigma(x_1)}, ..., \frac{1}{\sigma(x_n)})$

Задача 6 [4 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 $\mathbb{E}(X^2)$.

Решение:

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i + \delta_i)(\beta x_i + \varepsilon_i)}{\sum_{i=1}^n (x_i + \delta_i)^2} = \frac{\frac{\beta}{n}\sum_{i=1}^n x_i^2 + \frac{1}{n}\sum_{i=1}^n x_i\varepsilon_i + \frac{\beta}{n}\sum_{i=1}^n x_i\delta_i + \frac{1}{n}\sum_{i=1}^n \delta_i\varepsilon_i}{\frac{1}{n}\sum_{i=1}^n x_i^2 + \frac{2}{n}\sum_{i=1}^n x_i\delta_i + \frac{1}{n}\sum_{i=1}^n \delta_i^2} \xrightarrow{P} \frac{\beta \mathbb{E}x_i^2}{\mathbb{E}x_i^2 + \tau^2}$$

Покажем сходимость по вероятности слагаемых:

1)$\frac{1}{n}\sum_{i=1}^n x_i^2 \xrightarrow{P} \mathbb{E}x_i^2$ по ЗБЧ

2)$\mathbb{E}\delta_i^2 = \mathbb{V}\delta_i + (\mathbb{E}\delta_i)^2 = \tau^2$, поэтому по ЗБЧ$\frac{1}{n}\sum_{i=1}^n \delta_i^2 \xrightarrow{P} \tau^2$

3)$x_i, \delta_i, \varepsilon_i$ попарно независимы, поэтому $\mathbb{E}(x_i\varepsilon_i) = \mathbb{E}x_i\mathbb{E}\varepsilon_i = 0, \mathbb{E}(x_i\delta_i) = \mathbb{E}x_i\mathbb{E}\delta_i = 0, \mathbb{E}(\delta_i\varepsilon_i) = \mathbb{E}\delta_i\mathbb{E}\varepsilon_i = 0$

По ЗБЧ соответсвующие средние (те все оставшиеся) сходятся к 0 по вероятности.

Ответ: $\alpha = \frac{\mathbb{E}x_i^2}{\mathbb{E}x_i^2 + \tau^2}$

Задача 7 [3 балла]

Пусть дана выборка $(\boldsymbol{X}, \boldsymbol{T}) = \{(\boldsymbol{x}_i, \boldsymbol{t}_i) \colon \boldsymbol{x}_i \in \mathbb{R}^d, \boldsymbol{t}_i \in \mathbb{R}^m\}_{i=1}^n$, $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, $\boldsymbol{T} \in \mathbb{R}^{n \times m}$. Рассмотрим модель *многомерной линейной регрессию*, т.е. регрессии, в которой независимая переменная является вектором:

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

где $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{t} \in \mathbb{R}^m$, $\boldsymbol{W} \in \mathbb{R}^{d \times m}$, $\boldsymbol{\varepsilon} \in \mathbb{R}^m$. Рассмотрим модель, в рамках которой плотность распределения вектора $\boldsymbol{t}$ при заданном векторе $\boldsymbol{x}$ имеет вид $p(\boldsymbol{t}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{t}|\boldsymbol{W}^T\boldsymbol{x}, \boldsymbol{\Sigma})$, т.е. нормальное распределение со средним $\boldsymbol{W}^T\boldsymbol{x} \in \mathbb{R}^m$ и матрицей ковариаций $\boldsymbol{\Sigma} \in \mathbb{R}^{m \times m}$. Найдите ML-оценки для матриц $\boldsymbol{W}$ и $\boldsymbol{\Sigma}$.

Подсказка. Вам могут потребоваться следующие формулы матричного дифференцирования:

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{X}^{-1} \boldsymbol{b}}{\partial \boldsymbol{X}} = -\boldsymbol{X}^{-T}\boldsymbol{a}\boldsymbol{b}^T\boldsymbol{X}^{-T}, \quad \frac{\partial \boldsymbol{a}^T \boldsymbol{X} \boldsymbol{b}}{\partial \boldsymbol{X}} = \boldsymbol{a}\boldsymbol{b}^T, \quad \frac{\partial \boldsymbol{a}^T \boldsymbol{X}^T \boldsymbol{b}}{\partial \boldsymbol{X}} = \boldsymbol{b}\boldsymbol{a}^T.$$

Внимание. Во возможности ответ следует полностью записать в матричном виде, выразив всё через $\boldsymbol{X}$ *и* $\boldsymbol{T}$.

Решение:

$$\log L = C - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^n (t_i - W^T x_i)^T \Sigma^{-1}(t_i - W^T x_i),$$

где C - некоторая константа, не зависящая от W и Σ.

Для удобства домножим выражение на (-2) и будем искать минимум (для удобства вычисления, чтобы не тащить за собой минусы). Также заметим, что Σ^{-1} - симметричная матрица, поэтому она равна транспонированной.

$$\frac{\partial(-2\log L)}{\partial \Sigma} = \frac{n}{|\Sigma|}|\Sigma|\Sigma^{-T} - \Sigma^{-T}\sum_{i=1}^n (t_i - W^T x_i)(t_i - W^T x_i)^T \Sigma^{-T} = n\Sigma^{-1} - \Sigma^{-1}(T - XW)^T(T - XW)\Sigma^{-1} = 0$$

$$=> \Sigma = \frac{1}{n}(T - XW)^T(T - XW)^T$$

$$\frac{\partial(-2\log L)}{\partial W} = \sum_{i=1}^{n}[-2x_i t_i^T \Sigma^{-1} + 2x_i x_i^T w \Sigma^{-1}] = 0$$

$$=> W = (X^T X)^{-1} X^T T$$

Ответ: $\Sigma = \frac{1}{n}(T - XW)^T (T - XW)^T, \qquad W = (X^T X)^{-1} X^T T$

Задача 8 [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 $(\boldsymbol{X}, \boldsymbol{t}) = \{(\boldsymbol{x}_i, t_i) : \boldsymbol{x}_i \in \mathbb{R}^d, t_i \in \mathbb{R}\}_{i=1}^n$. Найдите апостериорное распределение $p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{t})$.

Решение:

$$p(w|x,t) = \frac{p(w,x,t)}{p(x,t)} = \frac{p(t|x,w)\cdot p(x,w)}{p(x,t)} = \frac{p(t|x,w)\,p(x)\,p(w)}{p(x)\,p(t)} =$$

$$= \frac{\prod_{i=1}^{n} p(t_i|x_i,w)\cdot p(w)}{p(t)}$$

$$p(w|x,t) \propto \frac{1}{Z}\exp\left\{-\frac{\beta\|Xw-t\|_2^2}{2} - \frac{(w-w_0)^T S_0^{-1}(w-w_0)}{2}\right\}$$

Покажем, что $\exp\{\ldots\}$ можно записать в виде $\exp\{-\frac{1}{2}(w-w_n)^T\Sigma^{-1}(w-w_n)\}$, тогда получим нормальное распределение $N(w_n, \Sigma)$.
($\frac{1}{Z}$ - некоторая нормировочная константа)

$$\beta\|Xw-t\|_2^2 + (w-w_0)^T S_0^{-1}(w-w_0) = \beta(Xw-t)^T(Xw-t) + (w-w_0)^T S_0^{-1}(w-w_0) =$$

$$= \beta(w^T X^T X w - w^T X^T t - t^T X w + t^T t) + (w-w_0)^T S_0^{-1}(w-w_0) =$$

$$= \underline{w^T(\beta X^T X + S_0^{-1})w} - \beta\underline{w}^T X^T t - \beta t^T X w + \beta t^T t - \underline{w}^T S_0^{-1} w_0 - w_0^T S_0^{-1} w + w_0^T S_0^{-1} w_0 =$$

$$= \underline{w^T \Sigma^{-1} w} - w^T \Sigma^{-1} w_n - w_n^T \Sigma^{-1} w + w_n^T \Sigma^{-1} w_n$$

$$\Sigma^{-1} = \beta X^T X + S_0^{-1}$$

$$-w^T \Sigma^{-1} w_n = -w^T(\beta X^T t + S_0^{-1} w_0) => \Sigma^{-1} w_n = \beta X^T t + S_0^{-1} w_0 =>$$

$$=> w_n = \Sigma(\beta X^T t + S_0^{-1} w_0) = (\beta X^T X + S_0^{-1})^{-1}(\beta X^T t + S_0^{-1} w_0)$$

$$p(w|x,t) = N\left((\beta X^T X + S_0^{-1})^{-1}(\beta X^T t + S_0^{-1} w_0), (\beta X^T X + S_0^{-1})^{-1}\right)$$

Ответ: $N(\boldsymbol{w}|(\beta\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{S}_0^{-1})^{-1}(\beta\boldsymbol{X}^T\boldsymbol{t} + \boldsymbol{S}_0^{-1}\boldsymbol{w}_0), (\beta\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{S}_0^{-1})^{-1})$

Задача 9 [2 балл]

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 $n \to \infty$.

Решение:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

$$\mathbb{E}[\hat{f}(x)] = \frac{1}{h}\mathbb{E}K\left(\frac{x - X_i}{h}\right) = \frac{1}{h}\int K\left(\frac{x-u}{h}\right)f(u)du = \frac{1}{h}\int\limits_{x-h/2}^{x+h/2} f(y)dy$$

$$\mathbb{V}[\hat{f}(x)] = \frac{1}{nh^2}\mathbb{V}K\left(\frac{x - X_i}{h}\right) = \frac{1}{nh^2}\bigg([\mathbb{E}K^2] - [\mathbb{E}K]^2\bigg)\bigg) = \frac{1}{nh^2}\int\limits_{x-h/2}^{x+h/2} f(y)dy\left(1 - \int\limits_{x-h/2}^{x+h/2} f(y)dy\right)$$

Теперь покажем сходимость по вероятности. Для этого воспользуемся выражениями из лекции:

$$\mathbb{E}[\hat{f}(x)] = \mathbb{E}\frac{1}{h}K\left(\frac{x - X_i}{h}\right) = f(x) + \frac{1}{2}h^2 f^{''}(x)\int u^2 K(u)du + ...$$

$$\mathbb{V}[\hat{f}(x)] \approx \frac{f(x)\int K^2(x)dx}{nh}$$

Запишем неравенство Чебышева и покажем, что при $h \to 0$ и $nh \to \infty$ выполняется сходимость по вероятности $\hat{f}(x) \xrightarrow{\mathsf{P}} f(x)$ при $n \to \infty$:

$$P\left(\left|\hat{f}(x) - \mathbb{E}[\hat{f}(x)]\right| \geq \varepsilon\right) = P\left(\left|\hat{f}(x) - f(x)\right| \geq \varepsilon\right) \leq \frac{\mathbb{V}[\hat{f}(x)]}{\varepsilon^2} \approx \frac{f(x)}{nh\varepsilon^2} \to 0$$

Что и требовалось показать

Ответ: $\mathbb{E}[\hat{f}(x)] = \frac{1}{h}\int\limits_{x-h/2}^{x+h/2} f(y)dy, \quad \mathbb{V}[\hat{f}(x)] = \frac{1}{nh^2}\int\limits_{x-h/2}^{x+h/2} f(y)dy\left(1 - \int\limits_{x-h/2}^{x+h/2} f(y)dy\right)$

Задача 10 [6 баллов]

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:

$$\hat{J}(h) = \int (\hat{p}(x; \boldsymbol{x}^{(n)}))^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{p}(x_i; \boldsymbol{x}^{(n\backslash i)}),$$

где $\hat{p}(\cdot; \boldsymbol{x}^{(n\backslash i)})$ — оценка плотности распределения на основе выборки $\boldsymbol{x}^{(n\backslash i)}$, 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. Пусть $\hat{p}_m$ — доля объектов выборки, попавших в бин B_m:

  $$n_m = \sum_i I\left[x_i \in B_m\right], \quad \hat{p}_m = \frac{n_m}{n}.$$

  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 $\Delta J(h) = \mathbb{E}[\hat{J}(h)] - \mathbb{E}[J(h)]$?

  Решение:

  $$\hat{J}(h) = \sum_{m=1}^{M}\frac{\hat{p}_m^2}{h^2}h - \frac{2}{nh}\sum_{i=1}^{n}\sum_{m=1}^{M}\frac{n_m - 1}{n-1}I[x_i \in B_m] = \frac{1}{h}\sum_{m=1}^{M}\hat{p}_m^2 - \frac{2}{h}\sum_{m=1}^{M}(\hat{p}_m - \frac{1}{n})\frac{n_m}{n-1} =$$

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

Оценка смещенная, почему?

- (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 $\Delta J(h) = \mathbb{E}[\hat{J}(h)] - \mathbb{E}[J(h)]$?
Решение:

$$\hat{J}(h) = \int\left(\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x-X_i}{h}\right)\right)^2 dx - \frac{2}{n}\sum_{i=1}^{n}\frac{1}{(n-1)h}\sum_{j=1}^{n} k\left(\frac{X_i-X_j}{h}\right) =$$

$$= \frac{1}{n^2h^2}\int\left[\sum_{i=1}^{n} k^2\left(\frac{x-X_i}{h}\right) + \sum_{i\neq j} k\left(\frac{x-X_i}{h}\right)k\left(\frac{x-X_j}{h}\right)\right]dx -$$

$$- \frac{2}{hn(n-1)}\sum_{i,j} k\left(\frac{X_i-X_j}{h}\right) + \frac{2}{hn}k(0) =$$

$$= \frac{h}{n^2h^2}\sum_{i,j}\int k\left(\frac{X_i-X_j}{h} - y\right)k(y)dy - \frac{2}{n(n-1)h}\sum_{i,j} k\left(\frac{X_i-X_j}{h}\right) + \frac{2}{nh}k(0) \approx$$

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

Покажем, что $\mathbb{E}[\hat{J}(h)] = \mathbb{E}[J(h)]$

$$\hat{J}(h) = \int_{-\infty}^{+\infty} \hat{f}_n^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{(-i)}(X_i)$$

$$J(h) = \int_{-\infty}^{+\infty} \hat{f}_n^2(x)dx - 2\int_{-\infty}^{+\infty} \hat{f}_n(x) f(x)dx$$

Достаточно показать, что

$$\mathbb{E}\int \hat{f}_n(x) f(x)dx = \mathbb{E}\left[+\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{(-i)}(X_i)\right]$$

$$\mathbb{E}\int \hat{f}_n(x) f(x)dx = \mathbb{E}\int \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) f(x)dx =$$

$$= \frac{1}{h}\mathbb{E}\int K\left(\frac{x-X_i}{h}\right) f(x)dx = \frac{1}{h}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} K\left(\frac{x-u}{h}\right) f(x) f(u)dxdu \quad (*)$$

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{(-i)}(X_i)\right] = \mathbb{E}\hat{f}_{(-i)}(X_i) = \mathbb{E}\frac{1}{(n-1)h}\sum_{\substack{j=1\\ i\neq j}}^{n} K\left(\frac{X_i-X_j}{h}\right) =$$

$$= \frac{1}{h}\mathbb{E}K\left(\frac{X_i-X_j}{h}\right) = \frac{1}{h}\iint_{-\infty\,-\infty}^{+\infty\,+\infty} K\left(\frac{x-u}{h}\right) f(x) f(u)dxdu \quad (**)$$

$(*) = (**) \Rightarrow$ равенство доказано

Задача 11 [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:

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

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?

Решение:

а) $\underset{f(x)\in \mathcal{F}(x)}{\arg\min} \sum_{i=1}^{n} (y_i - f(x_i))^2$

Чтобы минимиз. сумму квадратов, нужно минимиз. её на каждом бину B_m, т.е. $C_m = \underset{x_i \in B_m}{\text{mean}}(y_i)$. Тогда получаем, что

$$\boxed{f(x) = \sum_{i=1}^{M} I(x \in B_i)\, \underset{x_j \in B_i}{\text{mean}}(y_j)}$$

б)

$$\underset{y \in \mathbb{R}}{\arg\min} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)(y_i - y)^2$$

$$\frac{\partial r}{\partial y} = -\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \cdot 2(y_i - y) = +2\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)(y - y_i) = 0$$

$$\Rightarrow \boxed{y = \frac{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)}}$$

в) $\underset{a, b \in \mathbb{R}}{\arg\min} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)(y_i - a - bx_i)^2$

$$\begin{cases} \frac{\partial r}{\partial a} = -2\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)(y_i - a - bx_i) = 0 \\ \frac{\partial r}{\partial b} = -2\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)(y_i - a - bx_i)x_i = 0 \end{cases}$$

$$\Rightarrow \boxed{\begin{gathered} b = \frac{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) x_i \cdot \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) y_i - \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \cdot \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) x_i y_i}{\left(\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) x_i\right)^2 - \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \cdot \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) x_i^2} \\[2ex] a = \frac{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) x_i y_i \cdot \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) x_i - \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) y_i \cdot \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) x_i^2}{\left(\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) x_i\right)^2 - \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \cdot \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) x_i^2} \end{gathered}}$$

Практический блок

Задача 12 [7 баллов]

Скачайте по ссылке данные о связи между оценкой качества вина от различных характеристик вина `https://archive.ics.uci.edu/ml/datasets/wine+quality`. По ссылке представлено два набора данных: для белых и для красных вин. Далее предполагается использование данных для белых вин (`winequality-white.csv`). Разбейте данные на обучающую и тестовую выборку: для тестовой выборки возьмите 25% данных.

- 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 и 10-кратную кросс-проверку. Выбор подмножества признаков проведите полным перебором. Позволяет ли какой-нибудь набор признаков получить значение MSE на тестовых данных меньше, чем на всех признаках .
- Обучите гребневую регрессию и Lasso. Оптимальные параметры подберите с помощью поиска по сетке и 10-кратной кросс-проверки. Получилось ли обучить модель, имеющую лучшее качество на тестовой выборке, чем простая линейная регрессия.

Задача 13 [5 баллов]

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 14 [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. Построить 95%-ые доверительные интервалы для полученной оценки функции регрессии.