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

Федор Ерин

Задачи

Теоретический блок

Задача 1 [1 балл]

Рассмотрим задачу оптимизации, решаемую в Elastic Net:

$$\|y - Xw\|_2^2 + \lambda_1\|w\|_1^1 + \lambda_2\|w\|_2^2 \to \min_w$$

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

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

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

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

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

Решение

- Преобразуем выражение для Elastic Net, для этого заметим, что

$$\lambda_2\|w\|_2^2 \to \min_w \Longleftrightarrow \|0 - \sqrt{\lambda_2}w\|_2^2 \to \min_w$$

Будем минимизировать l_2-норму с весами в рамках нормы, где y и X. Для этого расширим вектор y и матрицу X, добавив соответствующие значения. Тогда получаем следующую запись для Elastic Net:

$$\| \begin{pmatrix} y \\ 0 \end{pmatrix} - \begin{pmatrix} X \\ \sqrt{\lambda_2} I \end{pmatrix} w^* \|_2^2 + \gamma\|w^*\|_1^1 \to \min_{w^*}$$

Допустим w и w^* связаны коэффициентом β: $w^* = \beta w$, тогда:

$$\| \begin{pmatrix} y \\ 0 \end{pmatrix} - \beta \begin{pmatrix} X \\ \sqrt{\lambda_2} I \end{pmatrix} w \|_2^2 + \beta\gamma\|w\|_1^1 \to \min_w$$

Сравним это с искусственными данными и приравняем компоненты:

$$\begin{cases} \alpha = \sqrt{\lambda_2}, \\ \frac{1}{\sqrt{1+\lambda_2}}\beta = 1, \\ \beta\gamma = \lambda_1. \end{cases} \Rightarrow \begin{cases} \alpha = \sqrt{\lambda_2}, \\ \beta = \sqrt{1+\lambda_2}, \\ \gamma = \lambda_1/\sqrt{1+\lambda_2}. \end{cases} \tag{1}$$

Ответ: $\alpha = \sqrt{\lambda_2}, \gamma = \lambda_1/\sqrt{1+\lambda_2}, \hat{w}^* = \sqrt{1+\lambda_2}\hat{w}$.

Задача 2 [2 балла]

Пусть дана выборка (x, y), $x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$, $y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$. Предположим, что справедлива следующая модель линейной регрессии:

$$y = xw + \epsilon, \quad \epsilon \sim N(0, \sigma^2).$$

Найдите OLS-оценку для w. Найдите дисперсию этой оценки. Укажите условия на распределение x, при которых оценка состоятельна при $n \to \infty$.

Решение

- Найдем OLS-оценку:

$$\|y - Xw\|^2 \to \min_w$$
$$2X^T(Xw - y) = 0$$
$$(X^T X)w = X^T y$$
$$\hat{w} = (X^T X)^{-1} X^T y$$

Найдем дисперсию этой оценки. Обозначим для удобства $X' = (X^T X)^{-1} X^T$:

$$\mathbb{V}\hat{w} = \mathbb{V}(X'\epsilon|X) = X'\mathbb{V}(\epsilon|X)X'^T = X'\mathbb{E}(\epsilon\epsilon^T|X)X'^T = X'(\sigma^2 I_n)X'^T = \sigma^2 X' X'^T =$$
$$= \sigma^2 (X^T X)^{-1} \cancel{X^T X (X^T X)^{-1}} = \sigma^2 (X^T X)^{-1}$$

Оценка $\hat{w}$ состоятельная, если $bias \to 0$ и $se \to 0$ при $n \to \infty$. Это действительно так по свойству OLS оценки, но X для этого должен быть нескоррелированным со случайными ошибками, иначе полученная оценка может быть смещенной и несостоятельной. Помимо X, ошибки ϵ должны иметь конечную одинаковую для всех X дисперсию и нескоррелированными. Также отметим, что при малом n оценка может стать несостоятельной.

Ответ: $\hat{w} = (X^T X)^{-1} X^T y, \mathbb{V}\hat{w} = \sigma^2 (X^T X)^{-1}$.

Задача 3 [2 балла]

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

$$y = w_0 + w_1 x + \epsilon, \qquad \epsilon \sim N(0, \sigma^2).$$

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

Решение

- Рассмотрим величину $\beta = w_1 - \alpha w_0$. Тогда

$$se(\beta)^2 = se(w_1 - \alpha w_0)^2 = se(w_1)^2 + \alpha^2 se(w_0)^2 - 2cov(w_0, w_1)$$

Для модели вида $y = w_0 + w_1 x + \epsilon$ известны оценки:

$$\hat{se}(w_1) = \frac{\sigma}{S_x \sqrt{n}}, \hat{se}(w_0) = \frac{\sigma}{S_x \sqrt{n}} \sqrt{\frac{\sum_{i=1}^n X_i^2}{n}}, cov(w_0, w_1) = -\overline{X} \frac{\sigma}{S_x \sqrt{n}},$$

где $S_x^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \overline{X_n})$.

Сформулируем Тест Вальда в данном случае: если $|W| > z_{\alpha/2} \Rightarrow H_0$ отклоняем:

$$W = \frac{\hat{\beta}}{\hat{se}(\hat{\beta})} = \frac{\hat{w}_1 - \alpha \hat{w}_0}{\sqrt{\hat{se}(\hat{w}_1)^2 + \alpha^2 \hat{se}(\hat{w}_0)^2 - 2cov(w_0, w_1)}}$$

Ответ: отклоняем H_0 при $|W| > z_{\alpha/2}$, $W = \frac{\hat{\beta}}{\hat{se}(\hat{\beta})} = \frac{\hat{w}_1 - \alpha \hat{w}_0}{\sqrt{\frac{\sigma^2}{S_x^2 n} + \alpha^2 \frac{\sigma^2}{S_x^2 n} \frac{\sum_{i=1}^n X_i^2}{n} + 2\overline{X} \frac{\sigma}{S_x \sqrt{n}}}}$

Задача 4 [2 балла]

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

$$y = x^T w + \epsilon, \quad \epsilon \sim N(0, \sigma^2),$$

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

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

Решение

- Для данной модели известна оценка:

$$\hat{w} = (X^T X)^{-1} X^T y \quad \Rightarrow \quad X^* \hat{w} = y^*$$

  Тогда имеем:

$$\mathbb{E}y^* = \mathbb{E}(X^*(X^T X)^{-1} X^T y) = X^*(X^T X)^{-1} X^T y$$

$$\mathbb{V}y^* = \mathbb{V}(X^*(X^T X)^{-1} X^T y) = X^*(X^T X)^{-1} X^T \mathbb{V}y (X^*(X^T X)^{-1} X^T)^T =$$

$$= X^*(X^T X)^{-1} X^T \sigma^2 \cancel{X^T X (X^T X)^{-1}} X^{*T} = \sigma^2 X^* (X^T X)^{-1} X^{*T}$$

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

Задача 5 [2 балла]

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

$$t = x^T w + \epsilon(x),$$

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

Решение

- Известна MLE-оценка $\hat{w}$ при $\epsilon(x) \sim N(0, \sigma^2)$:

$$\hat{w} = (X^T X)^{-1} X^T y$$

  Эта оценка совпадает с оценкой МНК. В более общем случае, когда дисперсия ошибок может зависеть от X, известна оценка обобщенным МНК:

$$\tilde{w} = (X^T W X)^{-1} X^T W y,$$

  где W - ковариационная матрица ошибок. Эта матрица диагональна, если ошибки гетероскедастичны, как в данной задаче, и нет автокорреляции, причем $W_{ii} = 1/\sigma(x_i)$ для каждого X_i (взвешиваем данные на $1/\sigma(x_i)$), то есть $W = diag(1/\sigma(x_i))$.

Ответ: $\tilde{w} = (X^T W X)^{-1} X^T W y, W = diag(1/\sigma(x_i))$.

Задача 6 [3 балла]

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

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

где $\epsilon_i \sim N(0, \sigma^2)$. При этом значения x наблюдаются с ошибкой, т.е. представлена не выборка (x, y), а выборка $(z, y) = \{(z_i, y_i) \colon z_i, y_i \in \mathbb{R}\}_{i=1}^n$, где $z_i = x_i + \delta_i$, $\delta_i \sim N(0, \tau^2)$. Шумы $_i$ 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

- $$\hat{\beta} = \frac{\sum_{i=1}^n z_i y_i}{\sum_{i=1}^n z_i^2} = \frac{\sum_{i=1}^n (x_i + \delta_i)(\beta(x_i) + \epsilon)}{\sum_{i=1}^n (x_i + \delta_i)^2} \xrightarrow{\mathsf{P}} \frac{cov(z_i, y_i)}{\mathbb{V} z_i} = \frac{\beta \sigma_{x_i}^2}{\sigma_{x_i}^2 + \sigma_\delta^2} = \beta \frac{1}{1 + \frac{\sigma_\delta^2}{\sigma_{x_i}^2}}$$

  Причем $\sigma_\delta^2 = \tau^2, \sigma_{x_i}^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2$, тогда:

  $$a = \frac{1}{1 + \frac{\tau^2}{\mathbb{E}X^2 - (\mathbb{E}X)^2}}$$

Ответ: $a = 1/(1 + \frac{\tau^2}{\mathbb{E}X^2 - (\mathbb{E}X)^2})$.

Задача 7 [3 балла]

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

$$\vec{t} = W^T \vec{x} + \vec{\epsilon},$$

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

Подсказка. Вам могут потребоваться следующие формулы матричного дифференцирования:

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

Внимание. Во возможности ответ следует полностью записать в матричном виде, выразив всё через X и T.

Решение

- Запишем правдоподобие и найдем MLE-оценку:

  $$l(W|Y, X, \Sigma) = -\frac{1}{2} \sum_{i=1}^n (\vec{y_i} - W^T \vec{x_i})^T \Sigma^{-1} (\vec{y_i} - W^T \vec{x_i}) + const$$

$$\frac{\partial l(W|Y,X,\Sigma)}{\partial W} = -2\sum_{i=1}^{n} \vec{x_i}\vec{y_i}^T\Sigma^{-1} + 2\sum_{i=1}^{n} \vec{x_i}\vec{x_i}^T W\Sigma^{-1} = -\cancel{2}X^TY\cancel{\Sigma^{-1}} + \cancel{2}X^TXW\cancel{\Sigma^{-1}} = 0$$

$$\hat{W} = (X^TX)^{-1}X^Ty$$

Так как MLE-оценка не зависит от параметризации, подставим MLE-оценку $\hat{W}$ в выражение для ковариации:

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\vec{y_i} - \hat{\vec{y_i}})(\vec{y_i} - \hat{\vec{y_i}})^T = \frac{1}{n}Y^T[I_n - X(X^TX)^{-1}X^T]Y$$

Ответ: $\hat{W} = (X^TX)^{-1}X^Ty, \hat{\Sigma} = \frac{1}{n}Y^T[I_n - X(X^TX)^{-1}X^T]Y$.

Задача 8 [2 балла]

Рассмотрим задачу восстановления регрессии. Модель регрессии имеет вид

$$t = x^Tw + \epsilon,$$

где $\epsilon \sim N(0, \beta^{-1})$, и на веса w наложено априорное распределение вида $p(w) = N(w|w_0, S_0)$. Пусть дана выборка $(X, t) = \{(x_i, t_i) \colon x_i \in \mathbb{R}^d, t_i \in \mathbb{R}\}_{i=1}^n$. Найдите апостериорное распределение $p(w|X, t)$.

Решение

- $$p(w|X,t) = \frac{p(X,t,w)}{p(X,t)} = \frac{p(t|X,w)p(w)}{p(t|X)}$$

  Знаменатель константа, найдем распределение числителя:

  $$p(t|X,w)p(w) \sim exp\big(-\frac{\beta\|t - Xw\|_2^2}{2} - \frac{\|w - w_0\|_2^2}{2S_0}\big) \Rightarrow$$

  $$\Rightarrow -\beta(t - Xw)^T(t - Xw) - \frac{1}{S_0}(w - w_0)^T(w - w_0) =$$

  $$= -\beta(t^Tt - X^Tw^Ty - tXw + w^TX^TXw) - S_0^{-1}(w^Tw - w^Tw_0 - w_0^Tw + w_0^Tw_0) =$$

  $$= -w^T(\beta X^TX + S_0^{-1})w + w(\beta tX + S_0^{-1}w_0^T) + w^T(\beta X^Tt + S_0^{-1}w_0) + \dots \qquad (1)$$

  Необходимо привести к виду, соответствующему нормальному распределению:

  $$(w - \mu)^T\Sigma^{-1}(w - \mu) = w^T\Sigma^{-1}w - \mu^T\Sigma^{-1}w - w^T\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\mu \qquad (2)$$

  Сравнивая (1) и (2), получаем:

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

  $$\mu = \Sigma(\beta X^Tt + S_0^{-1}w_0)$$

Ответ: $p(w|X,t) \sim N(\mu, \Sigma), \mu = \Sigma(\beta X^Tt + S_0^{-1}w_0), \Sigma^{-1} = \beta X^TX + S_0^{-1}$.

Задача 9 [2 балл]

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

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

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

- $$\mathbb{E}\hat{f}(x) = \mathbb{E}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\frac{x-X_i}{h}) = \frac{1}{n}\sum\int\frac{1}{h}K(\frac{x-y}{h})f(y)dy =$$

$$= \frac{1}{nh}\sum\int I[y-\frac{h}{2} < x < y+\frac{h}{2}]f(y)dy = \frac{1}{nh}\sum\int_{x-h/2}^{x+h/2}f(y)dy = \frac{1}{n}\int_{x-h/2}^{x+h/2}f(y)dy$$

$$\mathbb{V}\hat{f}(x) = \mathbb{V}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\frac{x-X_i}{h}) = \frac{1}{nh^2}\mathbb{V}K(\frac{x-X_i}{h}) =$$

$$= \frac{1}{nh^2}\big[\mathbb{E}K^2(\frac{x-X_i}{h}) - (\mathbb{E}K(\frac{x-X_i}{h}))^2\big] = \frac{1}{nh^2}\big[\int_{x-h/2}^{x+h/2}f(y)dy - \big(\int_{x-h/2}^{x+h/2}f(y)dy\big)^2\big]$$

Оценка риска:

$$R(f, \hat{f}_n) \approx \frac{1}{4}\sigma_K^4 h^4\int(f''(x))^2 + \frac{\int K^2(x)dx}{nh},$$

где $\sigma_K^2 = \int x^2K(x)dx = \int_{-1/2}^{1/2}x^2dx = 1/12, \int K^2(x)dx = 1$, откуда:

$$R(f, \hat{f}_n) \approx \frac{h^4}{4\cdot 12^2}\int(f''(x))^2 + \frac{1}{nh} \to 0 \quad \text{при} \quad h \to 0, nh \to \infty$$

С другой стороны,

$$R(f, \hat{f}_n) = \int b^2(x)dx + \int v(x)dx \to 0 \Rightarrow$$

$$\Rightarrow \int b^2(x)dx = [\mathbb{E}(\hat{f}_n(x) - f(x))]^2 \to 0, \int v(x)dx = \mathbb{V}(\hat{f}_n(x) - f(x)) \to 0 \Rightarrow$$

$$\Rightarrow \mathbb{E}(\hat{f}_n(x) - f(x))^2 = v(x) - b^2(x) \to 0 \Rightarrow \text{сходится по } L_2 \Rightarrow \text{ сходится и по P.}$$

Ответ: $\mathbb{E}\hat{f}(x) = \frac{1}{n}\int_{x-h/2}^{x+h/2}f(y)dy, \mathbb{V}\hat{f}(x) = \frac{1}{nh^2}\big[\int_{x-h/2}^{x+h/2}f(y)dy - \big(\int_{x-h/2}^{x+h/2}f(y)dy\big)^2\big]$.

Задача 10 [6 баллов]

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

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

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

- (Гистограммная оценка) Разобьем диапазон наблюдаемых значений $x^{(n)}$ на бины ширины h. Пусть в итоге значения $x^{(n)}$ 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:
$$n_m = \sum_i I[x_i \in B_m], \quad \hat{p}_m = \frac{n_m}{n}.$$
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:
$$\hat{J}(h) \approx \frac{1}{hn^2} \sum_{i,j} K^*\big(\frac{X_i - X_j}{h}\big) + \frac{2}{nh} K(0),$$
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 $N(0,1)$, т.е. гауссово ядро, то $K^{(2)}(z)$ — плотность распределения $N(0,2)$.
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

- (Гистограммная оценка) Распишем слагаемые оценки риска:
$$\int \hat{p}(x;x_n)^2 dx = \int (\sum_{m=1}^{M} \frac{\hat{p}_m}{h} I[x \in B_m])^2 dx = \frac{1}{h^2} \int \sum_{m=1}^{M} \hat{p}_m^2 I[x \in B_m] dx =$$
$$= \frac{1}{h^2} \sum_{m=1}^{M} \int_{B_m} \hat{p}_m^2 dx = \frac{1}{h^2} \sum_{m=1}^{M} h \cdot \hat{p}_m^2 dx = \frac{1}{h} \sum_{m=1}^{M} \hat{p}_m^2$$
$$\sum_{i=1}^{n} \hat{p}(x_i; x^{(n\setminus i)}) = \sum_{i=1}^{n} \sum_{m=1}^{M} \frac{\hat{p}_{m,(-i)}}{h} I[X_i \in B_m] = \frac{1}{h} \sum_{i=1}^{n} \sum_{m=1}^{M} \sum_{k=1,k\neq i}^{n} \frac{1}{n-1} I[X_k \in B_m] I[X_i \in B_m] =$$
$$= \frac{1}{h(n-1)} \sum_{i=1}^{n} \sum_{m=1}^{M} (n_m - I[X_i \in B_m]) I[X_i \in B_m] = \frac{1}{h(n-1)} \sum_{m=1}^{M} \big[n_m \sum_{i=1}^{n} I[X_i \in B_m] -$$
$$- \sum_{i=1}^{n} I^2[X_i \in B_m]\big] = \frac{1}{h(n-1)} \sum_{m=1}^{M} (n_m^2 - n_m)$$

Откуда получаем:

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

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

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

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

Получили искомое выражение. Проверим теперь равенство $\mathbb{E}[\hat{J}(h)] = \mathbb{E}[J(h)]$:

$$\Delta J(h) = \mathbb{E}\big[\cancel{\int \hat{p}^2(x)dx} - \frac{2}{n}\sum_{i=1}^{n}\hat{p}_{(-i)}(x_i)\big] - \mathbb{E}\big[\cancel{\int \hat{p}^2(x)dx} - 2\int \hat{p}(x)p(x)dx\big] =$$

$$= -\frac{2}{n}\sum_{i=1}^{n}\mathbb{E}[\hat{p}_{(-i)}(x_i)] + 2\mathbb{E}[\int \hat{p}(x)p(x)dx] \qquad (1)$$

Распишем слагаемые:

$$\mathbb{E}[\hat{p}_{(-i)}(x_i)] = \frac{1}{h(n-1)}\sum_{k=1, i\neq k}^{n}\mathbb{E}_{X_i}\mathbb{E}_{X_k}K\big(\frac{X_i - X_k}{h}\big) =$$

$$= \frac{1}{n-1}\sum_{k=1, i\neq k}^{n}\frac{1}{h}\mathbb{E}_{X_i}\int K\big(\frac{X_i - y}{h}\big)p(y)dy = \frac{1}{h}\int\int K\big(\frac{X-y}{h}\big)p(x)p(y)dxdy \qquad (2)$$

$$\int \hat{p}(x)p(x)dx = \mathbb{E}_x[\mathbb{E}_X\hat{p}(x)] = \frac{1}{hn}\sum_{i=1}^{n}\mathbb{E}_x\mathbb{E}_{X_i}K\big(\frac{x - X_i}{h}\big) =$$

$$= \frac{1}{hn}\sum_{i=1}^{n}\mathbb{E}_x\int K\big(\frac{x-y}{h}\big)p(y)dy = \frac{1}{h}\int\int K\big(\frac{X-y}{h}\big)p(x)p(y)dxdy \qquad (3)$$

Из (1), (2), (3) следует равенство $\mathbb{E}[\hat{J}(h)] = \mathbb{E}[J(h)]$.

- (Ядерная оценка) Распишем слагаемые оценки риска:

$$\int \hat{p}(x;x_n)^2dx = \int\big(\frac{1}{nh}\sum_{i=1}^{n}K\big(\frac{x-X_i}{h}\big)\big)^2dx = \frac{1}{n^2h^2}\int\sum_{i,j}K\big(\frac{x-X_i}{h}\big)K\big(\frac{x-X_j}{h}\big)dx =$$

$$= \frac{1}{n^2h^2}\sum_{i,j}\int K(-y)K\big(-y + \frac{X_i - X_j}{h}\big)hdy = \frac{1}{n^2h}\sum_{i,j}K\big(\frac{X_i-X_j}{h} - y\big)K(-y)dy =$$

$$= \frac{1}{n^2h}\sum_{i,j}K^{(2)}\big(\frac{X_i - X_j}{h}\big)$$

$$-\frac{2}{n}\sum_{i=1}^{n}\hat{p}(x_i;x^{(n\setminus i)}) = -\frac{2}{n}\sum_{i=1}^{n}\frac{1}{h(n-1)}\sum_{k=1, i\neq k}^{n}K\big(\frac{X_i - X_k}{h}\big) =$$

$$= -\frac{2}{nh(n-1)}\sum_{i,k}\left[K\big(\frac{X_i - X_k}{h}\big) - nK(0)\right] = -\frac{2}{nh(n-1)}\sum_{i,k}K\big(\frac{X_i - X_k}{h}\big) + \frac{2}{h(n-1)}K(0) \approx$$

$$\approx -\frac{2}{n^2h}\sum_{i,k}K\big(\frac{X_i - X_k}{h}\big) + \frac{2}{nh}K(0)$$

Откуда получаем:

$$\hat{J}(h) = \frac{1}{n^2h}\sum_{i,k}\left[K^{(2)}\big(\frac{X_i - X_k}{h}\big) - 2K\big(\frac{X_i - X_k}{h}\big)\right] + \frac{2}{nh}K(0) = \frac{1}{n^2h}\sum_{i,k}K^*(x) + \frac{2}{nh}K(0)$$

Получили искомое выражение. Проверим теперь равенство $\mathbb{E}[\hat{J}(h)] = \mathbb{E}[J(h)]$. По теореме с лекции:

$$\forall h > 0: \quad \mathbb{E}\hat{J}(h) = \mathbb{E}J(h)$$

Доказательство аналогично выводу для гистограммной оценки.

Практический блок

Задача 12 [4 балла]

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. Выбор подмножества признаков проведите полным перебором. Позволяет ли какой-нибудь набор признаков получить значение MSE на тестовых данных меньше, чем на всех признаках?

Решение в IPYNB-ноутбуке

Задача 13 [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.

Решение в IPYNB-ноутбуке

Задача 14 [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.

Решение в IPYNB-ноутбуке