№1.

Пусть $g: R^n \to R^m$, где $x \in R^n$ Д-во: 1) $a^T x = (a_1 \dots a_n)\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = a_1 x_1 + \dots + a_n x_n = \sum_{i=1}^n a_i x_i$

$$\frac{\partial g}{\partial x} = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & & & \vdots \\ \frac{\partial g_m}{\partial x_1} & \cdots & & \frac{\partial g_m}{\partial x_n} \end{pmatrix} \qquad \frac{\partial (a^T x)}{\partial x} = \begin{pmatrix} \frac{\partial \left(\sum_{i=1}^n a_i x_i\right)}{\partial x_1} \\ \vdots \\ \frac{\partial \left(\sum_{i=1}^n a_i x_i\right)}{\partial x_n} \end{pmatrix} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = a.$$

2) $A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix}$ $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ $Ax = \begin{pmatrix} \sum_{i=1}^n a_{1i} x_i \\ \sum_{i=1}^n a_{2i} x_i \\ \vdots \\ \sum_{i=1}^n a_{mi} x_i \end{pmatrix}$

$$\frac{\partial (Ax)}{\partial x} = \begin{pmatrix} \frac{\partial\left(\sum_{i=1}^n a_{1i} x_i\right)}{\partial x_1} & \cdots & \frac{\partial\left(\sum_{i=1}^n a_{1i} x_i\right)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial\left(\sum_{i=1}^n a_{mi} x_i\right)}{\partial x_1} & \cdots & \frac{\partial\left(\sum_{i=1}^n a_{mi} x_i\right)}{\partial x_n} \end{pmatrix} = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix} = A.$$

3) $x^T A x = \left(\sum_{i=1}^n a_{i1} x_i \; \sum_{i=1}^n a_{i2} x_i \dots \sum_{i=1}^n a_{in} x_i\right) \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \sum_{j=1}^n \sum_{i=1}^n a_{ij} x_i x_j.$

$$\frac{\partial (x^T A x)}{\partial x} = \frac{\partial\left(\sum_j \sum_i a_{ij} x_i x_j\right)}{\partial x} = \begin{pmatrix} 2a_{11} x_1 + \sum_{i \in \{2..n\}} a_{i1} x_i + \sum_{j \in \{2..n\}} a_{1j} x_j \\ \vdots \\ 2a_{nn} x_n + \sum_{i \in \{1..n-1\}} a_{in} x_i + \sum_{j \in \{1..n-1\}} a_{nj} x_j \end{pmatrix} =$$

$$= \begin{pmatrix} \sum_{i=1}^n x_i (a_{1i} + a_{i1}) \\ \vdots \\ \sum_i x_i (a_{ni} + a_{in}) \end{pmatrix} = (A + A^T) x$$

4) $\|x\|^2 = x_1^2 + x_2^2 + \dots + x_n^2$ $\frac{\partial \|x\|^2}{\partial x} = \begin{pmatrix} 2x_1 \\ \vdots \\ 2x_n \end{pmatrix} = 2x$

5) $g(x) = \begin{pmatrix} g(x_1) \\ \vdots \\ g(x_n) \end{pmatrix}$

$$\frac{\partial g(x)}{\partial x} = \begin{pmatrix} \frac{\partial g(x_1)}{\partial x_1} & \cdots & \frac{\partial g(x_1)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g(x_n)}{\partial x_1} & \cdots & \frac{\partial g(x_n)}{\partial x_n} \end{pmatrix} = \begin{pmatrix} g'(x_1) & 0 & \dots & 0 \\ 0 & g'(x_2) & & 0 \\ 0 & \dots & & g'(x_n) \end{pmatrix} = \mathrm{diag}(g'(x)).$$

б) $$\frac{\partial g(h(x))}{\partial x} = \begin{pmatrix} \frac{\partial g_1(h(x))}{\partial x_1} & \cdots & \frac{\partial g_1(h(x))}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_p(h(x))}{\partial x_1} & \cdots & \frac{\partial g_p(h(x))}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \frac{\partial g_1(h(x))}{\partial h_1} & \cdots & \frac{\partial g_1(h(x))}{\partial h_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_p(h(x))}{\partial h_1} & \cdots & \frac{\partial g_p(h(x))}{\partial h_m} \end{pmatrix} \cdot$$

$$\cdot \begin{pmatrix} \frac{\partial h_1(x)}{\partial x_1} & \cdots & \frac{\partial h_1(x)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_m(x)}{\partial x_1} & \cdots & \frac{\partial h_m(x)}{\partial x_n} \end{pmatrix} = \frac{\partial g(h(x))}{\partial h} \cdot \frac{\partial h(x)}{\partial x}$$

№2

Запишем $g(\beta) = \|X\beta - y\|^2 = (X\beta - y)^T(X\beta - y)$

Раскроем скобки: $g(\beta) = \beta^T X^T X\beta - 2y^T X\beta + y^T y$, $y^T y$ не зависит от β.

Найдём градиент по β: $\frac{\partial}{\partial \beta}(\beta^T X^T X\beta) = ((X^T X + (X^T X)^T)\beta$, но $(X^T X)^T = X^T X$

матрица симметрична $\Rightarrow \frac{\partial}{\partial \beta}(\beta^T X^T X\beta) = 2X^T X\beta$.

$\frac{\partial}{\partial \beta}(-2y^T X\beta) = -2X^T y \Rightarrow \nabla g(\beta) = 2X^T X\beta - 2X^T y = 0 \quad X^T X\beta = X^T y$

$\beta = (X^T X)^{-1} X^T y$. $\quad \frac{\partial^2 g(\beta)}{\partial\beta\partial\beta^T} = \frac{\partial}{\partial \beta^T}\left(\frac{\partial g(\beta)}{\partial \beta}\right) = \frac{\partial}{\partial \beta^T}(2X^T X\beta - 2X^T y) = 2X^T X$

№17

$$X = \begin{pmatrix} 4 & 2 & 3 \\ 0 & -3 & 2 \\ -1 & -2 & 2 \\ 3 & 1 & 1 \\ 4 & 2 & -3 \end{pmatrix} \quad \bar{x}_1 = \frac{4+0+(-1)+3+4}{5} = 2, \quad \bar{x}_2 = \frac{2+(-3)+(-2)+1+2}{5} = 0, \quad \bar{x}_3 = \frac{3+2+2+1+(-3)}{5} = 1$$

$$\Rightarrow \bar{X}^T = \begin{pmatrix} 2 & -2 & -3 & 1 & 2 \\ 2 & -3 & -2 & 1 & 2 \\ 2 & 1 & 1 & 0 & -4 \end{pmatrix}$$

Ковариационная матрица $S = \frac{1}{n-1}\bar{X}^T\bar{X}$.

$$X^T X = \begin{pmatrix} 22 & 21 & -9 \\ 21 & 22 & -9 \\ -9 & -9 & 22 \end{pmatrix} \quad S = \frac{1}{4}\begin{pmatrix} 22 & 21 & -9 \\ 21 & 22 & -9 \\ -9 & -9 & 22 \end{pmatrix} = \begin{pmatrix} 5.5 & 5.25 & -2.25 \\ 5.25 & 5.5 & -2.25 \\ -2.25 & -2.25 & 5.5 \end{pmatrix}$$

Решим $\det(S - \lambda I) = 0$ $\begin{vmatrix} 5.5-\lambda & 5.25 & -2.25 \\ 5.25 & 5.5-\lambda & -2.25 \\ -2.25 & -2.25 & 5.5-\lambda \end{vmatrix} = 0$

$$\det(C-\lambda I) = (5.5-\lambda)(5.5-\lambda)^2 + 10.5(-2.25)^2 - (5.5-\lambda)(2(-2.25)^2 + (5.25)^2) \ominus$$

$$\ominus\ 12.25 - 53.0625\lambda + 16.5\lambda^2 - \lambda^3 = 0 \Leftrightarrow -(4\lambda-1)(\lambda-4)(4\lambda-49) = 0$$

$\Rightarrow \lambda_1 = 0.25$, $\lambda_2 = 4$, $\lambda_3 = 12.25$

При $\lambda_1 = 0.25$ $(S-0.25I)V=0$ $x_1 = -x_2,\ x_3 = 0$

Если $x_1 = -1,\ x_2 = 1$: $V_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$ или $\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$.

$\lambda_2 = 4$: $(S-4I)=0$ $x_1 = x_2,\ x_3 = 3x_1$. Если $x_1 = 1 \Rightarrow V_2 = \begin{pmatrix} 1 \\ 1 \\ 3 \end{pmatrix}$.

$\lambda_3 = 12.25$: $(S-12.25I)V = 0$ $x_1 = x_2,\ x_3 = -1.5x_1$. Если $x_1 = -3 \Rightarrow$

$\Rightarrow x_2 = -3,\ x_3 = 2$: $V_3 = \begin{pmatrix} -3 \\ -3 \\ 2 \end{pmatrix}$. $\frac{1}{N-1}\lambda_1 = \frac{1}{N-1}\sigma_1^2 = \frac{1}{16} = 0.0625$

$\frac{1}{N-1}\lambda_2 = \frac{1}{N-1}\sigma_2^2 = 1$

$\frac{1}{N-1}\lambda_3 = \frac{1}{N-1}\sigma_3^2 = \frac{49}{16}$

$\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3} \approx 0.015$

$\frac{\lambda_2}{\lambda_1+\lambda_2+\lambda_3} \approx 0.24$

$\frac{\lambda_3}{\lambda_1+\lambda_2+\lambda_3} \approx 0.74$

Найдем сингуляр. разложение матриц:

$X_c = U\Sigma V^T$ $V = \begin{pmatrix} 1 & 1 & -3 \\ -1 & 1 & -3 \\ 0 & 3 & 2 \end{pmatrix}$ $\Sigma = \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{pmatrix} = \begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3.5 \end{pmatrix}$

$\Sigma^{-1} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 2/7 \end{pmatrix}$

$$U\Sigma = X_c V = \begin{pmatrix} 2 & 2 & 2 \\ -2 & -3 & 1 \\ -3 & -2 & 1 \\ 1 & 1 & 0 \\ 2 & 2 & -4 \end{pmatrix} \begin{pmatrix} 1 & 1 & -3 \\ -1 & 1 & -3 \\ 0 & 3 & 2 \end{pmatrix} = \begin{pmatrix} 0 & 10 & -8 \\ 1 & -2 & 17 \\ -1 & -2 & 17 \\ 0 & 2 & -6 \\ 0 & -8 & -20 \end{pmatrix}$$

$$U = X_c V \Sigma^{-1} = \begin{pmatrix} 0 & 10 & -8 \\ 1 & -2 & 17 \\ -1 & -2 & 17 \\ 0 & 2 & -6 \\ 0 & -8 & -20 \end{pmatrix} \begin{pmatrix} 2 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 2/7 \end{pmatrix} = \begin{pmatrix} 0 & 5 & -16/7 \\ 2 & -1 & 34/7 \\ -2 & -1 & 34/7 \\ 0 & 1 & -12/7 \\ 0 & -4 & -40/7 \end{pmatrix}$$

№3.

x	1	1	0	0	-1
y	4	4	0	2	6

$$X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & -1 & 1 \end{pmatrix} \quad y = \begin{pmatrix} 4 \\ 4 \\ 0 \\ 2 \\ 6 \end{pmatrix}$$

1)

2) $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$

$$X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & -1 \\ 1 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 5 & 1 & 3 \\ 1 & 3 & 1 \\ 3 & 1 & 3 \end{pmatrix}$$

$$X^T y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & -1 \\ 1 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 4 \\ 4 \\ 0 \\ 2 \\ 6 \end{pmatrix} = \begin{pmatrix} 16 \\ 2 \\ 14 \end{pmatrix}$$

$$X^T X \beta = X^T y \Rightarrow \beta = \begin{pmatrix} 1 \\ -1 \\ 4 \end{pmatrix}$$

$$f(x) = 1 - x + 4x^2.$$

3) $\lambda = 1$

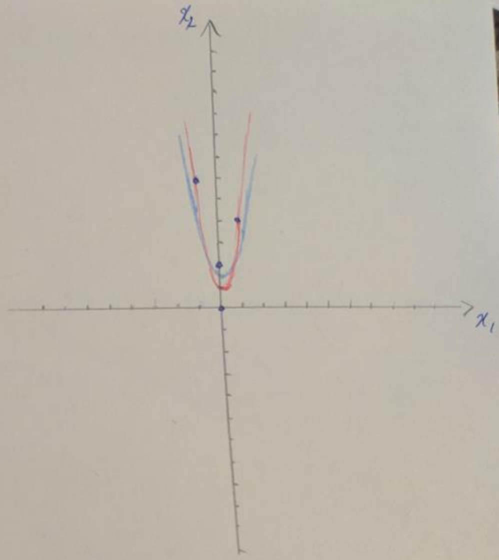
$$X^T X + \lambda I = \cancel{\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & -1 \\ 1 & 1 & 0 & 0 & 1 \end{pmatrix}}$$

$$= \begin{pmatrix} 5 & 1 & 3 \\ 1 & 3 & 1 \\ 3 & 1 & 3 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 6 & 1 & 3 \\ 1 & 4 & 1 \\ 3 & 1 & 4 \end{pmatrix}$$

$$(X^T X + \lambda I)\beta = X^T y \Rightarrow$$

$$\Rightarrow \beta = \begin{pmatrix} 3/2 \\ -1/2 \\ 5/2 \end{pmatrix}$$

$$f(x) = 3/2 - \frac{1}{2}x + \frac{5}{2}x^2$$



№9

№37

Нужно чтобы опт. прогноз был $\mathrm{mode}(Y|X=x) = \arg\max_y Pr(Y=y|X=x)$.

Возьмем нуль-единичную ф-цию потерь $L(a,y) = \begin{cases} 0, & a=y \\ 1, & a \neq y. \end{cases}$

$R(a|x) = \sum_y L(a,y) Pr(Y=y|X=x) = \sum_{y\neq a} Pr(Y=y|x) = 1 - Pr(Y=a|x)$.

Минимизируя $R(a|x)$ по a эквивалентно макс $Pr(Y=a|x) \Rightarrow$

$\Rightarrow \arg\min_a R(a|x) = \arg\max_a Pr(Y=a|x)$, т.е усл. мода опт. прогноз.

№38

Ф-ция потерь задана как $L(0,0) = L(1,1) = 0$, $L(1,0) = l_1$, $L(0,1) = l_0$.

Д-ть $f^*(x) = \arg\max_{y\in\{0,1\}} l_y Pr(y|x)$.

Для x вычислим усл. риск для каждого возможного предсказания $a \in \{0,1\}$: для $a=0$

$R(0|x) = L(0,0) Pr(Y=0|x) + L(0,1) Pr(Y=1|x) = 0 \cdot Pr(Y=0|x) + l_0 Pr(Y=1|x) =$
$= l_0 Pr(Y=1|x)$.

для $a=1$: $R(1|x) = L(1,0) Pr(Y=0|x) + L(1,1) Pr(Y=1|x) = l_1 Pr(Y=0|x) +$
$+ 0 \cdot Pr(Y=1|x) = l_1 Pr(Y=0|x)$.

$f^*(x) = \begin{cases} 0, & \text{если } l_0 Pr(Y=1|x) < l_1 Pr(Y=0|x), \\ 1, & \text{иначе}. \end{cases}$

Упростим и перставим члены, при $a=0$:

$l_0 Pr(Y=1|x) < l_1(1 - Pr(Y=1|x)) \Leftrightarrow (l_0+l_1) Pr(Y=1|x) < l_1 \Leftrightarrow$

$\Leftrightarrow Pr(Y=1|x) < \frac{l_1}{l_0+l_1}$. Аналогично $a=1$ соотв. $Pr(Y=1|x) > \frac{l_1}{l_0+l_1}$

$f^*(x) = \arg\min_{a\in\{0,1\}} R(a|x) = \arg\min_{a\in\{0,1\}} l_a Pr(Y\neq a|x) = \arg\max_{a\in\{0,1\}} l_a Pr(Y=a|x)$,

$l_a Pr(Y\neq a|x) = l_a(1 - Pr(Y=a|x))$ и минимизация этого по a эквив. максимизации $l_a Pr(Y=a|x)$. *

№39

$R(f) = E_X[E[l(f(X), Y) \mid X]]$ $\quad R(f \mid X=x) = E[l(f(x), Y) \mid X = x]$.

Т.к внешнее ожидание берет среднее по всем x, а $f(x)$ можно выбрать независимо от каждого x, то оптимум:

$$f^*(x) = \arg\min_{a \in \{1, \dots, K\}} R(a|x),$$ где $R(a|x)$ - риск

Выражаем $R(a|x) = E[l_{aY} \mid X = x]$. Т.к Y принимает $1, \dots, K$ раскроем:

$R(a|x) = \sum_{y=1}^{K} l_{ay} Pr(Y=y \mid X=x)$. - байесовское правило для произвольной матрицы потерь.

Если $l_{ay} = \begin{cases} 0, & a = y \\ 1, & a \neq y \end{cases}$, тогда $R(a|x) = \sum_{y \neq a} Pr(Y=y \mid X=x) = 1 - Pr(Y=a \mid X=x)$.

Минимизация по a эквив.: $f^*(x) = \arg\max_a Pr(Y=a \mid X=x)$, что дает MAP классификатор.

Итого $f^*(x) = \arg\min_{a \in \{1, \dots, K\}} \sum_{y=1}^{K} l_{ay} Pr(Y=y \mid X=x)$ и оно следует из минимизации

$$f^*(x) = \arg\min_a E[l_{aY} \mid X = x].$$

№35

$L(y', y) = (y' - y)^2 \quad R(f^*) - ?$

Для фиксированного x опр. ф-ция риска как ф-ция от c. По опр. усл. мат. ожидания $g(c) = \int (y-c)^2 p_{Y|X}(y|x)\,dy$. Это квадратичная ф-ция по c, так что минимум достигается в c, где производная по c равна 0.

$(Y-c)^2 = Y^2 - 2cY + c^2$, $g(c) = E[Y^2|X=x] - 2c\,E[Y|X=x] + c^2$.

$E[Y^2|X=x]$ - константа по c.

$g'(c) = -2E[Y|X=x] + 2c = 0 \Leftrightarrow c = E[Y|X=x]$.

$g''(c) = 2 > 0$ значит минимум. $\Rightarrow f^*(x) = E[Y|X=x]$.

Теперь $R(f^*) = E[(Y - f^*(X))^2] = E[(Y - E[Y|X])^2]$.

$E[(Y - E[Y|X])^2] = E[Var(Y|X)] = E[Y^2] - E[E[Y|X]^2]$.

№36

$L(y', y) = |y' - y|$ показать $f^*(x) = median(Y|X=x)$.

Зафиксируем x и рассмотрим $g(c) = E[|Y-c|\,|X=x] = \int |y-c|\, p_{Y|X}(y|x)\,dy$.

Найдем c минимизирующее $g(c)$.

Субградиент по c: $\frac{d}{dc}|y-c| = \begin{cases} -1, & y > c \\ (-1, 1), & y = c \\ 1, & y < c. \end{cases}$ Поэтому субградиент $G(c)$ для $g(c)$:

$G(c) = E[\operatorname{sign}(c - Y)|X=x] = Pr(Y<c|x) - Pr(Y>c|x)$, c минимизирует $g(c)$, если $0 \in \partial g(c)$, т.е. если $Pr(Y<c|x) \le \frac{1}{2}$ и $Pr(Y \le c|x) \ge \frac{1}{2}$, где c - усл. медиана распр. $Y|X=x$

Для непр. распр. $Pr(Y \le c|x) = 1/2$.

Если сдвинуть c на $\epsilon > 0$: $g(c+\epsilon) - g(c) \approx \epsilon(Pr(Y<c|x) - Pr(Y>c|x))$