Ветошкин Л.Н.

1 Матричное дифференцирование

1. Пусть $g: \mathbf{R}^n \to \mathbf{R}^m$, $x \in \mathbf{R}^n$. *Матрицей Якоби* называется матрица

$$\frac{\partial g}{\partial x} = \begin{pmatrix} \dfrac{\partial g_1}{\partial x_1} & \dfrac{\partial g_1}{\partial x_2} & \cdots & \dfrac{\partial g_1}{\partial x_n} \\ \dfrac{\partial g_2}{\partial x_1} & \dfrac{\partial g_2}{\partial x_2} & \cdots & \dfrac{\partial g_2}{\partial x_n} \\ \dots & \dots & \dots & \dots \\ \dfrac{\partial g_m}{\partial x_1} & \dfrac{\partial g_m}{\partial x_2} & \cdots & \dfrac{\partial g_m}{\partial x_n} \end{pmatrix}.$$

В частности, если $m = 1$ (т.е. $g(x)$ — скалярная функция векторного аргумента x), то $\dfrac{\partial g}{\partial x}$ — градиент функции g.

Доказать, что

1) если $a \in \mathbf{R}^n$, $x \in \mathbf{R}^n$, то $\dfrac{\partial(a^\top x)}{\partial x} = a$;

2) если $A \in \mathbf{R}^{m\times n}$, $x \in \mathbf{R}^n$, то $\dfrac{\partial(Ax)}{\partial x} = A$;

3) если $A \in \mathbf{R}^{n\times n}$, $x \in \mathbf{R}^n$, то $\dfrac{\partial(x^\top Ax)}{\partial x} = (A + A^\top)x$; в частности, если $A^\top = A$, то $\dfrac{\partial(x^\top Ax)}{\partial x} = 2Ax$;

4) если $x \in \mathbf{R}^n$, то $\dfrac{\partial\|x\|^2}{\partial x} = 2x$;

5) если g — скалярная функция и под $g(x)$ понимается применение функции g к каждой компоненте вектора $x \in \mathbf{R}^n$, то

$$\frac{\partial g(x)}{\partial x} = \mathrm{diag}(g'(x)),$$

где $\mathrm{diag}(a)$ — диагональная матрица с диагональю a;

6) если $h: \mathbf{R}^n \to \mathbf{R}^m$, $g: \mathbf{R}^m \to \mathbf{R}^p$, $x \in \mathbf{R}^n$, то

$$\frac{\partial g(h(x))}{\partial x} = \frac{\partial g(h(x))}{\partial h}\frac{\partial h(x)}{\partial x}.$$

№1.

$$a^T x = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} (x_1 \ldots x_n) = a_1 x_1 + a_2 x_2 + \ldots + a_n x_n.$$

$$\frac{\partial a^T x}{\partial x} = \left(\frac{\partial a^T x}{\partial x_1}, \ldots, \frac{\partial a^T x}{\partial x_n} \right) = (a_1, \ldots, a_n) = a$$

№2.

$$Ax = \begin{pmatrix} a_{11} \ldots a_{1n} \\ a_{21} \ldots a_{2n} \\ \ldots \\ a_{m1} \ldots a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \cancel{x_2, \ldots x_n} = \begin{pmatrix} a_{11} x_1 + a_{12} x_2 + \ldots + a_{1n} x_n \\ a_{21} x_1 + a_{22} x_2 + \ldots + a_{2n} x_n \\ \ldots \\ a_{m1} x_1 + a_{m2} x_2 + \ldots + a_{mn} x_n \end{pmatrix}$$

$(Ax)_i$ — i-ая строка Ax.

$$\frac{\partial Ax}{\partial x} = \begin{pmatrix} \frac{\partial (Ax)_1}{\partial x_1} & \ldots & \frac{\partial (Ax)_1}{\partial x_n} \\ \ldots & & \ldots \\ \frac{\partial (Ax)_m}{\partial x_1} & \ldots & \frac{\partial (Ax)_m}{\partial x_n} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix}$$

№3.

$$x^T A = (x_1 \ldots x_n) \begin{pmatrix} a_{11} \ldots a_{1n} \\ \ldots \\ a_{n1} \ldots a_{nn} \end{pmatrix} = \left(\sum_{i=1}^n x_i a_{i1} \ldots \sum_{i=1}^n x_i a_{in} \right)$$

$$\Rightarrow x^T A x = \left(\sum_{i=1}^n x_i a_{i1} \ldots \sum_{i=1}^n x_i a_{in} \right) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} =$$

$= x_1 \sum_{i=1}^{n} x_i a_{i1} + x_2 \sum_{i=1}^{n} x_i a_{i2} + \ldots + x_n \sum_{i=1}^{n} x_i a_{in} = f(x)$

Найдем производную по x_1, ост. ищатся аналогич

$$\frac{\partial f(x)}{\partial x_1} = \sum_{i=1}^{n} x_i a_{i1} + a_{11} x_1 + a_{12} x_2 + a_{13} x_3 + \ldots a_{1n} x_n =$$

$$= 2a_{11} x_1 + (a_{12} + a_{21}) x_2 + (a_{13} + a_{31}) x_3 + \ldots + (a_{1n} + a_{n1}) x_n =$$

как видно $\frac{\partial f(x)}{\partial x_1}$ равен первой строке столбца $(A + A^T)x$.

№4.

$$\|x\| = \sqrt{x_1^2 + \ldots + x_n^2}, \quad \|x\|^2 = x_1^2 + \ldots + x_n^2.$$

$$\frac{\partial \|x\|^2}{\partial x} = (2x_1 \ 2x_2 \ldots 2x_n) = 2x$$

№5.

$$g(x) = \begin{pmatrix} g(x_1) \\ g(x_2) \\ \vdots \\ g(x_n) \end{pmatrix} \quad \frac{\partial g(x)}{\partial x} = \begin{pmatrix} g'(x) & 0 \ldots & 0 \\ 0 & g'(x) \ldots & 0 \\ & 0 & g'(x) \end{pmatrix}$$

№6

$$\frac{\partial g(h(x))}{\partial x} = \sum_{i=1}^{n} \frac{\partial g(h(x))}{\partial h_i(x_i)} \cdot \frac{\partial h_i(x_i)}{\partial x_i} = \frac{\partial g(h(x))}{\partial h} \cdot \frac{\partial h(x)}{\partial x}$$

2. Пользуясь №1, найдите градиент $\dfrac{\partial g(\beta)}{\partial \beta}$ и гессиан $\dfrac{\partial^2 g(\beta)}{\partial \beta^\top \partial \beta}$ функции $g(\beta) = \|\mathbf{X}\beta - \mathbf{y}\|^2$. Выведите отсюда, что решение линейной задачи наименьших квадратов $\widehat{\beta} = \operatorname{argmin} \|\mathbf{X}\beta - \mathbf{y}\|^2$ является решением нормальной системы линейных уравнений $\mathbf{X}^\top \mathbf{X}\beta = \mathbf{X}^\top \mathbf{y}$.

3. Дана обучающая выборка

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

1) изобразить точки;
2) методом наименьших квадратов построить модель вида $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$; построить график этой функции;
3) построить модель того же вида методом ридж-регрессии с параметром регуляризации $\lambda = 1$; построить график этой функции.

   Замечание: при ручных вычислениях по методу наименьших квадратов рекомендуется составить систему $\mathbf{X}^\top \mathbf{X}\beta = \mathbf{X}^\top \mathbf{y}$ и решить ее. Регуляризованная система: $(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})\beta = \mathbf{X}^\top \mathbf{y}$, где $\mathbf{I}$ — единичная матрица.

4. Рассмотрим задачу восстановления регрессии, в которой $\mathbf{y}$ распределен согласно нормальному закону $N(\mathbf{X}\beta, \sigma^2\mathbf{I})$, а β имеет априорное распределение $N(0, \tau\mathbf{I})$. Найти апостериорное распределение для β. Доказать, что β^{ridge} есть его математическое ожидание. Найти связь между параметром регуляризации λ и дисперсиями τ, σ^2.

№2.

$g(\beta) = \|X\beta - y\|^2$; $\quad \dfrac{\partial g(\beta)}{\partial \beta}$ – ? $\quad \dfrac{\partial g(\beta)}{\partial \beta^T \partial \beta}$ – ?

$\dfrac{\partial g(\beta)}{\partial \beta} = 2X^T(X\beta - y)$ – убедимся в этом, возьмем

$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \end{pmatrix} \quad \beta = (\cancel{\beta_1 \beta_2 \dots \beta_n}) \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}$

$y = y_1$;

$$g(\beta) = \|X\beta - y\|^2 = \left(\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} - y_1 \right)^2 =$$

$$= (x_{11}\beta_1 + x_{12}\beta_2 + \dots + x_{1n}\beta_n - y_1)^2 \Rightarrow$$

$$\frac{\partial g(\beta)}{\partial \beta} = 2(x_{11}\beta_1 + x_{12}\beta_2 + \dots + x_{1n}\beta_n - y_1) \cdot \begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1n} \end{pmatrix} =$$

$$= 2(X\beta - y)X^T = 2X^T(X\beta - y) \quad (1)$$

$$\frac{\partial^2 g^2(\beta)}{\partial \beta^T \partial \beta} = \begin{pmatrix} \dfrac{\partial^2 g(\beta)}{\partial \beta_1^2} & \dots & \dfrac{\partial^2 g(\beta)}{\partial \beta_1 \partial \beta_n} \\ \vdots & & \\ \dfrac{\partial^2 g(\beta)}{\partial \beta_n \partial \beta_1} & \dots & \dfrac{\partial^2 g(\beta)}{\partial^2 \beta_n^2} \end{pmatrix} =$$

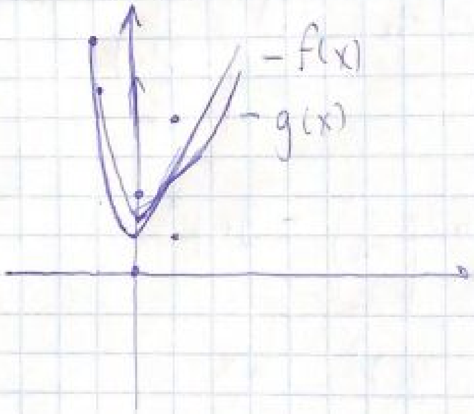
$$= 2X^TX.$$

из (1) $2X^T(X\beta - y) = 0$ => $\hat{\beta} = \arg\min \|X\beta - y\|^2 =$.

реш. ур-я => $X^T X \beta = X^T y$ т.е. $(X\beta - y) \to 0$ => $2X^T(X\beta - y)$

№3.

x	1	1	0	0	−1
y	1	4	0	2	6



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

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

$$X^T X = \begin{pmatrix} 5 & 1 & 3 \\ 1 & 3 & 1 \\ 3 & 1 & 3 \end{pmatrix}, \quad X^T y = \begin{pmatrix} 13 \\ -1 \\ 11 \end{pmatrix}$$

$$X^T X \beta = X^T y \Rightarrow \beta = \begin{pmatrix} 1 \\ -\frac{7}{4} \\ \frac{13}{4} \end{pmatrix}$$

тогда $f(x) = 1 - \frac{7}{4}x + \frac{13}{4}x^2$.

$f(-1) = 6$; $f(0) = 1$; $f(1) = 2.5$.

$\lambda = 1$

$X^T X + \lambda I = \begin{pmatrix} 6 & 1 & 3 \\ 1 & 4 & 1 \\ 3 & 1 & 4 \end{pmatrix} \Rightarrow (X^T X + \lambda I)\beta = X^T y \Rightarrow$

$\beta = \begin{pmatrix} 75/56 \\ -61/56 \\ 113/56 \end{pmatrix}$

$g(x) = \frac{75}{56} - \frac{61}{56}x + \frac{113}{56}x^2$

$g(-1) \approx 4{,}5$; $g(0) = 1.34$; $g(1) = 2{,}27$.

№4.

$y \sim N(X\beta, \sigma^2 I)$, $\beta \sim N(0, \tau I)$ $(\beta | y) \sim ?$

$y = X\beta + \lambda I \Rightarrow (y | \beta) \sim N(\lambda I, X \tau I)$.

$p(\beta | y) = \frac{p(y|\beta) \cdot p(\beta)}{p(y)} \Rightarrow (\beta | y) \sim \frac{N(\lambda I, X\tau I) \cdot N(0, \tau I)}{N(X\beta, \sigma^2 I)} =$

$= N(0, \tau, \sigma^2)$ $N\left(0, \frac{X\tau^2}{\sigma^2} I\right)$