Борисов Дмитрий | Листок по теме „Линейная регрессия".

① Формулировка задачи

- Векторы признаков $\vec{x}_1, \dots, \vec{x}_L$ для L объектов.

$\downarrow$ (известны значения)

$y_1, \dots, y_L$

- $\tilde{y}_i \approx \hat{y}_i = \langle \vec{\omega}, \vec{x}_i \rangle = \vec{\omega}^T \vec{x}_i.$
- $Q(\vec{\omega}) = \sum_{i=1}^{L} L(y_i, \hat{y}_i) \longrightarrow \min_{\vec{\omega}}.$
- Используем квадратичную функцию потерь: $L(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2.$
- $X = \begin{bmatrix} \vec{x}_1^T \\ \vdots \\ \vec{x}_L^T \end{bmatrix}$ — матрица признаков; $\vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_L \end{bmatrix}$, $\vec{\hat{y}} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_L \end{bmatrix}.$

1) $\hat{y}_i = \langle \vec{\omega}, \vec{x}_i \rangle = \langle \vec{x}_i, \vec{\omega} \rangle = \vec{x}_i^T \vec{\omega};$

$$\Rightarrow \boxed{\vec{\hat{y}} = X\vec{\omega}}$$

2) $\underline{\underline{Q(\vec{\omega})}} = \sum_{i=1}^{L} L(y_i, \hat{y}_i) = \sum_{i=1}^{L} (y_i - \hat{y}_i)^2 = \left(\|\vec{y} - \vec{\hat{y}}\|_2\right)^2 = \underline{\underline{\|\vec{y} - X\vec{\omega}\|_2^2}};$

3) $\hat{y}$ принадлежит подпространству $L(X)$ в $\mathbb{R}^n$, порождаемыми столбцами $\vec{x}^1, \dots, \vec{x}^L$ матрицы X.

② Нормальное уравнение

1) Доказать: $\frac{\partial}{\partial \vec{x}} (A\vec{x} + \vec{b})^T (A\vec{x} + \vec{b}) = 2A^T(A\vec{x} + \vec{b}).$

- Покажем, что $\frac{\partial}{\partial \vec{x}} (\vec{x}^T \vec{b}) = \vec{b}$:

$\vec{x}^T \vec{b} = (x_1 \dots x_n) \cdot \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = x_1 b_1 + \dots + x_n b_n.$

$$\frac{\partial(\vec{x}^T \vec{b})}{\partial \vec{x}} = \begin{bmatrix} \frac{\partial(\vec{x}^T \vec{b})}{\partial x_1} \\ \vdots \\ \frac{\partial(\vec{x}^T \vec{b})}{\partial x_n} \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} = \vec{b}.$$

Аналогично получим $\frac{\partial}{\partial \vec{x}} (\vec{b}^T \vec{x}) = \vec{b}.$

- Также покажем, что $\frac{\partial}{\partial \vec{x}} (\vec{x}^T A \vec{x}) = (A + A^T)\vec{x}.$

$$A\vec{x} = \begin{bmatrix} a_{11}x_1 + \ldots + a_{1n}x_n \\ \vdots \qquad \vdots \qquad \vdots \\ a_{n1}x_1 + \ldots + a_{nn}x_n \end{bmatrix}, \quad \vec{x}^T A\vec{x} = x_1(a_{11}x_1 + \ldots + a_{1n}x_n) + \ldots + x_n(a_{n1}x_1 + \ldots + a_{nn}x_n)$$

$$\frac{\partial}{\partial \vec{x}}(\ldots) = \begin{bmatrix} \frac{\partial(\ldots)}{\partial x_1} \\ \vdots \\ \frac{\partial(\ldots)}{\partial x_n} \end{bmatrix}; \quad \frac{\partial}{\partial \vec{x}}(\vec{x}^T A \vec{x}) = \begin{bmatrix} (2a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n) + a_{21}x_2 + \ldots + a_{n1}x_n \\ a_{12}x_{1} + (2a_{22}x_2 + a_{21}x_1 + \ldots + a_{2n}x_n) + \ldots + a_{n2}x_n \\ \ldots \\ a_{1n}x_1 + \ldots a_{n-1,n}x_{n-1} + (2a_{nn}x_n + a_{n1}x_1 + \ldots + a_{n,n-1}x_{n-1}) \end{bmatrix} =$$

$$= \begin{bmatrix} x_1(a_{11}+a_{11}) + x_2(a_{12}+a_{21}) + \ldots + x_n(a_{1n}+a_{n1}) \\ x_1(a_{12}+a_{21}) + x_2(a_{22}+a_{22}) + \ldots + x_n(a_{2n}+a_{n2}) \\ - \quad - \quad - \quad - \quad - \quad - \quad - \\ x_1(a_{1n}+a_{n1}) + \ldots \qquad + x_n(a_{nn}+a_{nn}) \end{bmatrix} = (A + A^T)\vec{x}.$$

$$\bullet \frac{\partial}{\partial x}\left[(Ax+b)^T(Ax+b)\right] = \frac{\partial}{\partial x}\left[(x^TA^T + b^T)(Ax+b)\right] = \frac{\partial}{\partial x}\left[x^TA^TAx + x^TA^Tb + b^TAx + b^Tb\right] =$$

$$= \frac{\partial}{\partial x}(x^TA^TAx) + \frac{\partial}{\partial x}(x^TA^Tb) + \frac{\partial}{\partial x}(b^TAx) + \cancelto{0}{\frac{\partial}{\partial x}(b^Tb)} =$$

$$= (A^TA + (A^TA)^T)x + A^Tb + (b^TA)^T = 2A^TAx + 2A^Tb = \underline{\underline{2A^T(Ax+b)}}.$$

---

2) $Q(\omega) = \|y - X\omega\|_2^2 = \|X\omega - y\|_2^2 = (X\omega - y)^T(X\omega - y)$

$$\frac{\partial Q}{\partial \omega} = \frac{\partial}{\partial \omega}\left[(X\omega - y)^T(X\omega - y)\right] = 2X^T(X\omega - y) = 0.$$

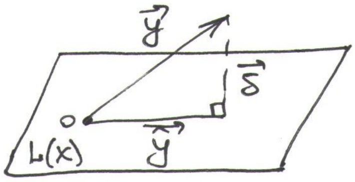
$$\cancel{2}X^TX\omega - \cancel{2}X^Ty = 0.$$

$$\boxed{\omega = (X^TX)^{-1}X^Ty}$$

---

③ Геометрическая интерпретация.

1)



$\vec{\hat{y}}$ – проекция вектора $\vec{y}$ на $L(X)$,

$\vec{\delta} = \vec{y} - \vec{\hat{y}}$ – вектор остатков.

$\vec{\delta} \perp \vec{x}^{(j)}$ $\forall j$ по свойству проекции.

Следовательно, $\langle \vec{x}^{(j)}, \vec{\hat{y}} - \vec{y} \rangle = -\langle \vec{x}^{(j)}, \vec{\delta} \rangle = \underline{0}$.

2) $X^T(\vec{\hat{y}} - \vec{y}) = \begin{bmatrix} \vec{x}^{(1)} \\ \vdots \\ \vec{x}^{(L)} \end{bmatrix} \cdot (\vec{\hat{y}} - \vec{y}) = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} = \vec{0}$.

3) $X^T(\vec{\hat{y}} - \vec{y}) = 0 \Rightarrow X^T(X\vec{\omega} - \vec{y}) = 0 \Rightarrow X^TX\vec{\omega} - X^T\vec{y} = 0 \Rightarrow$ $\boxed{\vec{\omega} = (X^TX)^{-1}X^T\vec{y}}$

(4) Вероятностная интерпретация.

4) $$f(x,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

$X \sim \mathcal{N}(\mu, \sigma^2)$

$\vec{\theta} = (\mu, \sigma^2)$ – оцениваемые параметры

$X_1, \dots, X_n$ – выборка.

$$L(X_1, \dots, X_n, \vec{\theta}) = \prod_{i=1}^{n} f_{\vec{\theta}}(X_i).$$

$$\underline{\underline{\hat{\vec{\theta}}}} = \arg\max_{\vec{\theta}} L(X_1, \dots, X_n, \vec{\theta});\quad \begin{cases} \nabla_{\vec{\theta}} \ln L = 0, \\ \nabla^2_{\vec{\theta}} \ln L < 0. \end{cases}$$

---

5) $y_i = \hat{y}_i + \varepsilon = \vec{x}_i^{\,T} \vec{w} + \varepsilon$, где $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ $\forall i = \overline{1,L}$

$$L(x_1, \dots, x_L, \vec{w}) = \prod_{i=1}^{L} f_{\vec{w}}(\varepsilon_i) = \prod_{i=1}^{L} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right),$$

$$\ln L = -L \ln(\sqrt{2\pi}\,\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{L} \varepsilon_i^2,$$

$$\ln L \propto -\sum_{i=1}^{L} \varepsilon_i^2 = -\sum_{i=1}^{L} (y_i - \vec{x}_i^{\,T} \vec{w})^2 = -\|\vec{y} - X\vec{w}\|_2^2.$$

---

6) $$\ln L(x_1, \dots, x_L, \vec{w}) \longrightarrow \max_{\vec{w}} \iff \|\vec{y} - X\vec{w}\|_2^2 \longrightarrow \min_{\vec{w}}.$$

---

7) Воспользуемся распределение Лапласа: $f(\varepsilon_i) = \frac{\alpha}{2} \exp(-\alpha|\varepsilon_i|)$.

$$\ln L(\vec{X}, \vec{w}) \propto -\sum_{i=1}^{L} |\varepsilon_i| = -\sum_{i=1}^{L} |y_i - \vec{x}_i^{\,T} \vec{w}| = -\|\vec{y} - X\vec{w}\|_1.$$

(5) L_2-регуляризация: гребневая регрессия (ridge regression).

8) $$Q(\vec{w}) = \|X\vec{w} - \vec{y}\|_2^2 + \tau\|\vec{w}\|_2^2, \qquad \|\vec{w}\|_2^2 = \vec{w}^T\vec{w}, \quad \frac{\partial}{\partial \vec{w}}\|\vec{w}\|_2^2 = \vec{w} + \vec{w}.$$

$$\frac{\partial Q(\vec{w})}{\partial \vec{w}} = \frac{\partial}{\partial \vec{w}}\|X\vec{w} - \vec{y}\|_2^2 + \tau \frac{\partial}{\partial \vec{w}}\|\vec{w}\|_2^2 = 2X^T(X\vec{w} - \vec{y}) + 2\tau\vec{w} = 0.$$

$$\cancel{2}X^TX\vec{w} - \cancel{2}X^T\vec{y} + \cancel{2}\tau\vec{w} = 0.$$

9) $$(X^TX + \tau I)\vec{w} = X^T\vec{y}$$

$$\boxed{\vec{w} = (X^TX + \tau I)^{-1} X^T\vec{y}}$$

10) Обозначим: $A = X^T X$, $B = X^T X + \tau I = A + \tau I$.

- рассмотрим собств. значения ~~(и собств. вектора)~~ матрицы A:

$$(A - \lambda I)\vec{x} = 0 \Rightarrow \lambda_1 \le \lambda_2 \le \ldots \le \lambda_n \text{ - собств. знач. матрицы } A.$$

- собств. значение матрицы B:

$$(B - \lambda' I)\vec{x} = 0 \rightarrow (A + \tau I - \lambda' I)\vec{x} = 0 \rightarrow (A - (\lambda' - \tau) I)\vec{x} = 0.$$

обозначим $\lambda^* = \lambda' - \tau$, тогда: $(A - \lambda^* I)\vec{x} = 0$.

Снова получили собств. знач. матрицы A: $\lambda_1 \le \lambda_2 \le \ldots \le \lambda_n$, но они связаны с собств. значениями матрицы B: $\underline{\underline{\lambda'_i = \lambda_i + \tau}}$, $i = \overline{1,n}$.

- $\lambda_{min}(A) = \lambda_1$; $\lambda_{max}(A) = \lambda_n$;
  $\lambda_{min}(B) = \lambda_1 + \tau$; $\lambda_{max}(B) = \lambda_n + \tau$.

- рассмотрим числа обусловленности матриц A и B.

$$\nu(A) = \frac{\lambda_{max}(A)}{\lambda_{min}(A)} = \frac{\lambda_n}{\lambda_1},$$

$$\nu(B) = \frac{\lambda_{max}(B)}{\lambda_{min}(B)} = \frac{\lambda_n + \tau}{\lambda_1 + \tau}.$$

покажем, что $\nu(A) \ge \nu(B)$:

$$\frac{\lambda_n}{\lambda_1} \ge \frac{\lambda_n + \tau}{\lambda_1 + \tau}$$

$$\frac{\cancel{\lambda_1 \lambda_n} + \tau\lambda_n}{\lambda_1(\lambda_1 + \tau)} \ge \frac{\cancel{\lambda_1 \lambda_n} + \lambda_1 \tau}{\lambda_1(\lambda_1 + \tau)} \text{ (верно, т.к. } \lambda_n \ge \lambda_1)$$

$\Rightarrow \underline{\nu(A) \ge \nu(B)}$: число обусловленности новой матрицы уменьшилось.
(это хорошо)

- $$\begin{cases} A: (A - \lambda I)\vec{x} = 0, \\ B: (A - \lambda I)\vec{x} = 0. \end{cases} \rightarrow \text{собственные векторы матриц } A \text{ и } B \text{ совпадают.}$$
  $$\underset{(B - \lambda' I)\vec{x} = 0}{\Uparrow}$$

6) L_1-регуляризация: лассо Тибширани (LASSO)

11) Рассмотрим задачу минимизации:

$$\begin{cases} Q(\omega) \rightarrow \min_{\omega}, \\ \|\omega\|_{L_1} < r. \end{cases}$$

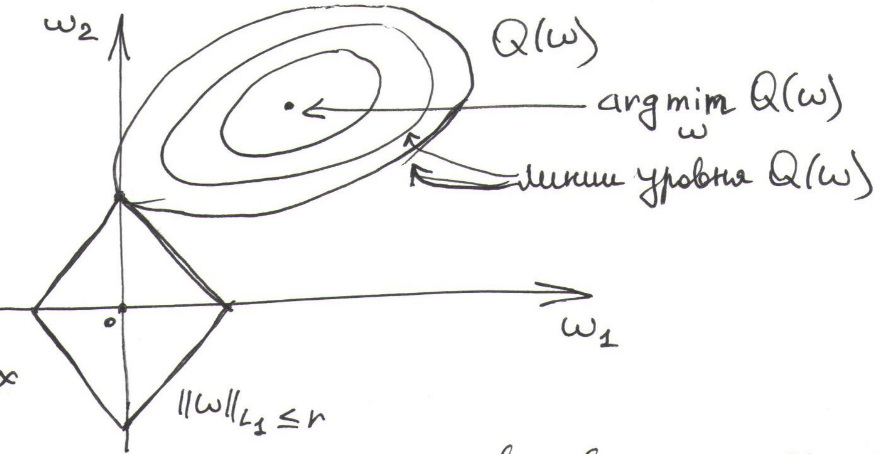
Запишем лагранжиан данной задачи:

$$L(\omega, \tau) = Q(\omega) + \tau(\|\omega\|_{L_1} - r), \quad \tau \ge 0.$$

$$\nabla_w L(\omega, \tau) = \nabla_w (Q(\omega) + \tau \|\omega\|_{L_1}) = 0.$$

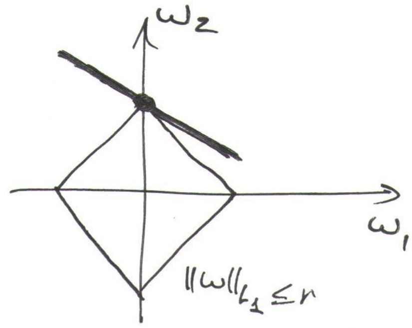
$\Rightarrow$ добавление ограничение $\|\omega\|_{L_1} < r$ равносильно добавлению штрафного слагаемого $\tau\|\omega\|_{L_1}$ в функционал для некоторого τ.

12) $\begin{cases} Q(\omega) \to \min\limits_{\omega}, \\ \|\omega\|_{L_1} \leq r. \end{cases}$



Чаще всего линии уровня пересекаются с шаром L_1-нормы в угловых точках (иначе ~~в~~ в касательных точках).

Если рассмотреть линейную аппроксимацию линий уровня в точках пересечения:



(7) Дополнительные вопросы

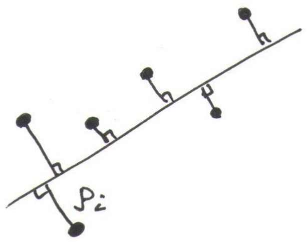
13) $y_i = \hat{y}_i + \varepsilon_i = \sum\limits_{j=1}^{n} x_i^{(j)} \omega_j + \omega_0 + \varepsilon_i, \ \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \ \forall i = \overline{1, L}$.

$$\mathbb{E} y_i = \sum_{j=1}^{n} \omega_j \mathbb{E} x_i^{(j)} + \omega_0.$$

Если отцентрировать выборку, то $\mathbb{E} y_i = 0$, $\mathbb{E} x_i^{(j)} = 0$, $j = \overline{1, n}$, $i = \overline{1, L}$

Следовательно, $\underline{\underline{\omega_0 = 0.}}$

14)



прямая: $\omega_x x + \omega_y y + \omega_0 = 0$,

квадрат расстояния $\rho_i^2 = \dfrac{(\omega_x x_i + \omega_y y_i + \omega_0)^2}{\omega_x^2 + \omega_y^2}$,

$$Q = \sum_{i=1}^{n} \rho_i^2 \to \min_{\omega_0, \omega_x, \omega_y} \Rightarrow \underline{\underline{\nabla Q = 0.}}$$

Отнормируем прямую: $\omega_x^2 + \omega_y^2 = 1$.

$\hookrightarrow \rho_i^2 = (\omega_x x_i + \omega_y y_i + \omega_0)^2$.

$$\nabla Q = \begin{pmatrix} \dfrac{\partial Q}{\partial \omega_0} \\ \dfrac{\partial Q}{\partial \omega_x} \\ \dfrac{\partial Q}{\partial \omega_y} \end{pmatrix} = 0.$$

- $\dfrac{\partial Q}{\partial \omega_0} = \sum\limits_{i=1}^{n} \dfrac{\partial \rho_i^2}{\partial \omega_0} = \sum\limits_{i=1}^{n} 2(\omega_x x_i + \omega_y y_i + \omega_0) = 2\omega_x \cdot n\bar{x} + 2\omega_y \cdot n\bar{y} + 2n\omega_0 =$
$= 2n(\omega_x \bar{x} + \omega_y \bar{y} + \omega_0) = 0 \Rightarrow \boxed{\omega_x \bar{x} + \omega_y \bar{y} + \omega_0 = 0}$

$$\frac{\partial Q}{\partial \omega_x} = \sum_{i=1}^{n} \frac{\partial \rho_i^2}{\partial \omega_x} = \sum_{i=1}^{n} 2(\omega_x x_i + \omega_y y_i + \omega_0) x_i = 2\omega_x \sum_{i=1}^{n} x_i^2 + 2\omega_y \sum_{i=1}^{n} x_i y_i +$$

$$+ 2\omega_0 n \overline{x} = 2n(\omega_x \overline{x^2} + \omega_y \overline{xy} + \omega_0 \overline{x}) = 0.$$

$$\Rightarrow \boxed{(\omega_x \overline{x^2} + \omega_y \overline{xy} + \omega_0 \overline{x}) = 0}$$

$$\frac{\partial Q}{\partial \omega_y} = \sum_{i=1}^{n} \frac{\partial \rho_i^2}{\partial \omega_y} = \sum_{i=1}^{n} 2(\omega_x x_i + \omega_y y_i + \omega_0) y_i = 2n(\omega_x \overline{xy} + \omega_y \overline{y^2} + \omega_0 \overline{y}) = 0.$$

$$\Rightarrow \boxed{\omega_x \overline{xy} + \omega_y \overline{y^2} + \omega_0 \overline{y} = 0}$$

Тогда:
$$\begin{cases} (1)\ \omega_x \overline{x} + \omega_y \overline{y} + \omega_0 = 0, & \rightarrow \omega_0 = -\omega_x \overline{x} - \omega_y \overline{y} \\ (2)\ \omega_x \overline{x^2} + \omega_y \overline{xy} + \omega_0 \overline{x} = 0, \\ (3)\ \omega_x \overline{xy} + \omega_y \overline{y^2} + \omega_0 \overline{y} = 0. \end{cases}$$

$$(2): \omega_x \overline{x^2} + \omega_y \overline{xy} - \omega_x \overline{x}^2 - \omega_y \overline{x}\overline{y} = 0.$$

$$\omega_y = \omega_x \frac{(\overline{x}^2 - \overline{x^2})}{(\overline{xy} - \overline{x}\overline{y})}$$

подставим в нормировку: $\omega_x^2 + \omega_x^2 \left(\frac{\overline{x}^2 - \overline{x^2}}{\overline{xy} - \overline{x}\overline{y}}\right)^2 = 1.$

$$\rightarrow \boxed{\omega_x = \left(1 + \left(\frac{\overline{x}^2 - \overline{x^2}}{\overline{xy} - \overline{x}\overline{y}}\right)^2\right)^{-1/2}}$$

$$\rightarrow \boxed{\omega_y = \omega_x \left(\frac{\overline{x}^2 - \overline{x^2}}{\overline{xy} - \overline{x}\overline{y}}\right)}$$

$$\rightarrow \boxed{\omega_0 = -\frac{\omega_x \overline{xy} + \omega_y \overline{y^2}}{\overline{y}}}$$