# Линейная регрессия

$X$ – выборка $\in \mathbb{R}^{n\times d}$

$n$ – кол-во эл-тов
$d$ – кол-во пр-ков

$y$ – target $\in \mathbb{R}^{n}$

$$\underset{\substack{n\times 1\\ \text{случайный}}}{\begin{pmatrix} y \end{pmatrix}} = \underset{\substack{n\times d\\ \text{не случ.}}}{\begin{pmatrix} & X & \end{pmatrix}} \underset{\substack{d\times 1\\ \text{оц. пар-ра}}}{\begin{pmatrix} \theta \end{pmatrix}} + \underset{\substack{n\times 1\\ \text{случайный}}}{\begin{pmatrix} \varepsilon \end{pmatrix}}$$

$\hat{\theta}$ – ?

$$\hat{y} = X\hat{\theta}$$

МНК: $RSS(\theta) = \|y - X\theta\|^2 \to \min\limits_{\theta}$

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

Качество модели?

$$R^2 = 1 - \frac{RSS(\hat{\theta})}{\|y - \bar{y}\|^2}$$ — коэф-т детерминации

$R \approx 1$ — хорошо $\quad\quad$ $R \approx 0$ — плохо

$d$ - большое $\Rightarrow$ $R^2$ — большой $\Rightarrow$

$\Rightarrow$ иногда используют $R^2_{ad} = 1 - (1 - R^2)\dfrac{n-1}{n-d}$ — подправленный

Проблемы линейной регрессии

- Мультиколлинеарность — ситуация с (почти) ЛЗ признаками

  $\Rightarrow \det(X^T X) \approx 0$ близка к вырожденной

  $\Rightarrow$ - $\hat{\theta}$ может быть большой
  - переобучение
  - не единственное решение
  - большая дисперсия оценок

Признаки мультиколлинеарности

- высокая корреляция пр-ков

- регрессия значимая, пр-ки нет

- $CI = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$ $\lambda$ – с.з. $X^T X$ – индекс обсл.

  $CI > 30 \Rightarrow$ модель мультиколлинеарна

  маг. константа

- коэф. вздутия по пр-ку $j$

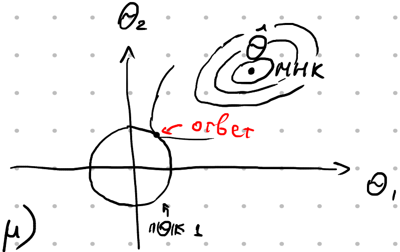
  $VIF_j = \frac{1}{1 - R_j^2}$, где $R_j^2$ – коэф. детерминации в модели $y' = X_j$ $X' = X^{-j}$ (без $X_j$)

  маг. константа

  $VIF_j > 10 \Rightarrow$ мультиколлинеарна

Борьба с мультиколлинеарностью

$$\begin{cases} \|y - X\theta\|^2 \to \min \\ \|\theta^2\| \leq \mu \end{cases}$$




$$L(\lambda) = \|y - X\theta\|^2 - \lambda\left(\|\theta\|^2 - \mu\right)$$

неприятно решать

$$\underbrace{\|y - X\theta\|^2}_{RSS(\theta)} + \underbrace{\alpha\|\theta\|^2}_{\text{регуляризатор}} \to \min$$

Ridge-regression

Предобработка данных

- Стандартизация $x_{ij} = \dfrac{x_{ij} - \overline{x_j}}{S_j}$ ← среднее по пр-ку ← стандартное отклонение

- Центрирование $y_i = y_i - \overline{y}$

Оценка

$$F(\Theta) = \| y - X\Theta \|^2 - \alpha \|\Theta\|^2 \to \min_{\Theta}$$

$$\nabla F(\Theta) = 2 X^T (y - X\Theta) + 2\alpha\Theta = 0$$

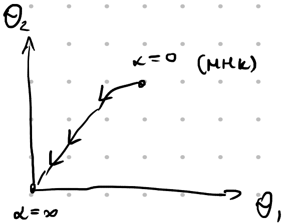
$$- X^T y + (X^T X + \alpha)\Theta = 0$$

$$(X^T X + \alpha I_d)\Theta = X^T y$$

$$\hat{\Theta} = (X^T X + \alpha I_d)^{-1} X^T y$$

ухудшили оценку
избавились от переобучения

Св-ва $\hat{\theta}$




$\alpha = 0$ — лин. регрессия

$\alpha \to \infty \Rightarrow \hat{\theta} \to 0$

$\alpha > 0 \Rightarrow \exists!\ \hat{\theta}$

$E\varepsilon = 0 \Rightarrow E\hat{\theta} = E(X^T X + \alpha I_d)X^T y = (X^T X + \alpha I_d)^{-1} X^T E y =$

$= (X^T X + \alpha I_d)^{-1} X^T X \theta \quad \alpha \neq 0 \Rightarrow$ смещ.

$D\varepsilon = \sigma^2 I_n$, то $Dy = D(X\theta + \varepsilon) = D\varepsilon$

$D\hat{\theta} = (X^T X + \alpha I_d)^{-1} X\, Dy \cdot X^T (X^T X + \alpha I_d)^{-T} =$

$= (X^T X + \alpha I_d)^{-1} X \sigma^2 X^T (X^T X + \alpha I_d)^{-1}$

Обращать матрицы сложно и долго.

Градиентный спуск

$$\nabla F(\theta) = -2X^Ty + 2X^TX\theta + 2\alpha\theta = -2X^Ty + 2(X^TX + \alpha I_d)\theta$$

$$\theta_{k+1} = \theta_k - \gamma\nabla F(\theta_k) = \theta_k - \gamma(X^Ty + (X^TX + \alpha I_d)\theta_k) =$$

$$= \theta_k - \gamma(X^T(X\theta_k - y) + \alpha\theta_k)$$

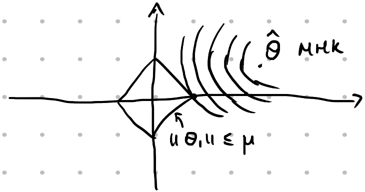
Стохаст. град. спуск.

$$\theta_{k+1} = \theta_k - \gamma\left[\frac{n}{m}\sum_{i=1}^{m}\left(x_{i_j}^T(x_{i_j}\theta_k - y_{i_j}) + \alpha\theta_n\right)\right]$$

$$I = \{i_1 \dots i_m\} \sim U\{1, \dots, n\}$$

Другая норма

$$\begin{cases} \| y - X\theta \|^2 \to \min \\ \| \theta \|_1 \leq \mu \end{cases}$$




$$L(\lambda) = \| y - X\theta \|^2 - \lambda ( \| \theta \|_1 - \mu )$$

гипотеза: часто будем в углах

$$F(\theta) = \| y - X\theta \|^2 + \beta \| \theta \|_1 \to \min_{\theta}$$ — Lasso - regression

Субградиент $f = |x|$ $\partial f = \begin{cases} -1, & x < 0 \\ [-1, 1], & x = 0 \\ 1, & x > 0 \end{cases}$

Опр. $f$ – выпуклая ф-ция, тогда ф-ция

$$\mathrm{Pr}_f(p) = \arg\min_x \left( \frac{1}{2} \| x - p \|^2 + f(x) \right)$$ наз-ся проксимальным оператором

$$g(x, p) := \frac{1}{2} \| x - p \|^2 + f(x)$$

Утв. (б/д) $g(x,p)$ – выпуклая, то $\forall p$ $\exists! \min_x$

Утв. (б/д) $x$ – экстремум выпуклой ф-ции $f$ $\Leftrightarrow$ $0 \in \partial f(x)$

Утв. $x_0 = Pr_f(p) \iff p - x_0 \in \partial f(x_0)$

▲ $x_0 = Pr_f(p) = \underset{x}{\operatorname{argmin}}\; g(x, p)$

$$0 \in \partial g(x_0, p)$$

$$\partial_x g(x_0, p) = x - p + \partial_x f(x)$$

$$\partial_x g(x_0, p) = x_0 - p + \partial_x f(x_0)$$

$$0 \in x_0 - p + \partial_x f(x_0)$$

$$p - x_0 \in \partial_x f(x_0)$$

Регуляризация $\underset{\text{ф-ция риска}}{F(\theta)} + \delta \underset{\text{регуляризатор}}{R(\theta)} \to \min\limits_{\theta}$

Lasso $RSS(\theta) + \delta \|\theta\|_1$

Утв. $\theta$ – решение задачи ↑ $\iff$ $\theta = Pr_{\partial R}(\theta - \nabla F(\theta))$

▲ $x_0 := Pr_f(p) = \theta$

$f := \delta R$

$p := \theta - \nabla F(\theta)$

$\theta = Pr_f(p) \iff -\nabla F(\theta) \in \partial f(\theta) = \delta \partial R(\theta)$

$0 \in \nabla F(\theta) + \delta \partial R(\theta) = \partial[F(\theta) + \delta R(\theta)]$ c $\iff$ $\theta$ – реш. регр.

Рассмотрим $d = 1$ Lasso - regression

$$R(x) = |x| \qquad x \in \mathbb{R}$$

$$Pr_{\beta |x|} (\underbrace{\theta - \nabla F(\theta)}_{P})$$

$$g(x, p) = \frac{1}{2} \| x - p \|^2 + \beta |x|$$

$$\partial_x g(x, p) = x - p + \beta \begin{cases} -1, & x < 0 \\ [-1, 1], & x = 0 \\ 1, & x > 0 \end{cases}$$

$$0 \in \partial_x g(x, p)$$

1) $x < 0$

$x - p - \beta = 0$

$x = p + \beta$

$p + \beta < 0$

$p < -\beta$

2) $x > 0$

$x - p + \beta = 0$

$x = p - \beta$

$p - \beta > 0$

$p > \beta$

3) $x = 0$

$-p + \beta[-1, 1] \ni 0$

$p \in [-\beta, \beta]$

$$\theta_{k+1} = \begin{cases} \theta_k - \nabla F(\theta_k) + \beta, & \theta_k - \nabla F(\theta_k) < \beta \\ 0, & \theta_k - \nabla F(\theta_k) \in [-\beta, \beta] \\ \theta_k - \nabla F(\theta_k) - \beta, & \theta_k - \nabla F(\theta_k) > \beta \end{cases}$$

кря?

$$R(x) = \| x \|_1 = \sum_{i=1}^{d} |x_i|$$

$$Pr_{\beta \| x \|_1}(p) = \operatorname*{argmin}_x \left( \frac{1}{2} \sum_{i=1}^{d} (x_i - p_o)^2 + \beta \sum_{i=1}^{d} |x_o| \right) =$$

$$= \begin{pmatrix} \operatorname*{argmin}_{x_1} \left( \frac{1}{2} (x_1 - p_1)^2 + \beta |x_1| \right) \\ \vdots \\ \operatorname*{argmin}_{x_d} \left( \frac{1}{2} (x_d - p_d)^2 + \beta |x_d| \right) \end{pmatrix} = \Big( Pr_{\beta |x_j|} (p_j) \Big)_{j=1 \dots d}$$

## Градиентный спуск

$$\nabla f(\theta) = -2X^T(y - X\theta) + \beta \, \mathrm{sign}(\theta) \qquad \mathrm{sign}(\theta) = (\mathrm{sign}(\theta_j))_{j=1 \dots d}$$

$$\theta_{k+1} = \theta_k - \gamma \left[ -2X^T(y - X\theta_k) + \beta \, \mathrm{sign}(\theta_k) \right] =$$

$$= \theta_k - \gamma \left[ 2X^T(X\theta_k - y) + \beta \, \mathrm{sign}(\theta_k) \right]$$

## Стох. градиентный спуск

$$\theta_{k+1} = \theta_k - \gamma \frac{n}{m} \sum_{j=1}^{d} \left[ 2 x_{i_j}^T (x_{i_j} \theta_k - y_{i_j}) + \beta \, \mathrm{sign}(\theta_k) \right)$$

$$I = \{ i_1 \dots i_m \} \sim U\{1 \dots m\}$$

Упр. (д/з) $D\varepsilon = \sigma^2 I_n$, то $D\hat{\theta}_{Lasso} \approx \sigma^2 (X^T X + W)^{-1} X^T X (X^T X + W)^{-1}$

$$W = \frac{1}{\|\hat{\theta}\|_1 \cdot \|X^T \hat{\Sigma}\|_\infty}$$

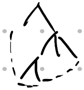
$$\hat{\Sigma} = y - X\hat{\theta}$$

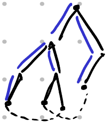
$$\|v\|_\infty = \max_j v_j$$

дерево

()
()
()

мы не влияем





5 7
3 1