Круглов А.И. Б05-202 ML2

1. $L(\theta) = \sum_{i=1}^{n} w_i (y_i - x_i^T \theta)^2 \to \min_\theta$, найти реш. в матричном виде

Пусть $X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}$ - матрица признаков, $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ - вектор таргетов, $W = \begin{pmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_n \end{pmatrix}$ - диаг. матрица весов

Тогда в матричном виде: $L(\theta) = (y - X\theta)^T W (y - X\theta) \to \min_\theta$

$$\frac{\partial L}{\partial \theta} = \left(\frac{\partial}{\partial\theta}(y - X\theta)^T\right) W (y - X\theta) + (y - X\theta)^T W \left(\frac{\partial}{\partial\theta}(y - X\theta)\right) = -X^T W (y - X\theta) + (y - X\theta)^T W (-X) = -2X^T W (y - X\theta)$$

min: $\frac{\partial L}{\partial \theta} = 0$, т.е. $X^T W y = X^T W X \hat\theta$. Выразим $\hat\theta = (X^T W X)^{-1} X^T W y$

По полученной формуле также видно, что МНК - это частный случай при $w_1 = \ldots = w_n = 1$, т.е. $\hat\theta_{МНК} = (X^T X)^{-1} X^T y$, что было на лекции.

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

2. $F(\theta) = -\log L_y(\theta) + \lambda \|\theta\|_2^2 = -\ell_y(\theta) + \lambda \theta^T\theta \Rightarrow \nabla F = -\nabla \ell_y + \nabla(\lambda\theta^T\theta) = -\nabla\ell_y + 2\lambda\theta$,

$\nabla\ell_y$ выведен на лекции: $\nabla\ell_y = X^T(y - S(\theta))$, где $S(\theta) = (\sigma(x_1^T\theta), \ldots, \sigma(x_n^T\theta))^T$

GD: инициализируем веса θ_0 и на каждом шаге будем обновлять по правилу:

$$\theta_{t+1} = \theta_t - \eta \nabla F(\theta_t) = \theta_t - \eta\left(2\lambda\theta_t - X^T(y - S(\theta_t))\right)$$, где η - размер шага (learning rate), $\ominus \eta \nabla F$ - потому что $F \to \min$

SGD: похоже на GD, но каждый шаг задействует батч $I = \{i_1, \ldots, i_m\} \in U\{1, \ldots, n\}$:

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{n}{m}\left(-X_I^T (y_I - S(\theta_t))\right) - \eta \cdot 2\lambda\theta_t$$

IRLS: задаем начальные веса θ_0 и будем их обновлять. На каждом шаге вычисляем

$$W = \mathrm{diag}\left(\sigma(x^T\theta)(1 - \sigma(x^T\theta))\right) \quad и \quad \theta_{t+1} = \theta_t - (X^T W(\theta_t) X)^{-1}\left(-X^T(y - S(\theta_t))\right),$$

что подробнее выводилось на лекции

Градиентный спуск хорошо работает для малых объемов данных, т.к. требует много вычислений. Поэтому на практике чаще обращаются к SGD или IRLS