Задача 3, Круглов А.И. Б05-207

1) $\hat{\theta}$ в ridge-регрессии

Вспомним сведения о матр. производной с лекции:

$f(x) = a^T x \Rightarrow \nabla f = a$, $f(x) = x^T A x \Rightarrow \nabla f = (A + A^T) x$

Оптимизируем функционал $F(\theta) = \|Y - X\theta\|^2 + \lambda \|\theta\|^2 =$

$= (Y - X\theta)^T (Y - X\theta) + \lambda \theta^T \theta = Y^T Y - 2(Y^T X)\theta + \theta^T \cdot$

$\cdot (X^T X + \lambda E)\theta$; отметим, что $(X^T X + \lambda E)^T = (X^T X + \lambda E)$.

Тогда $\nabla F(\theta) = -2X^T Y + 2(X^T X + \lambda E)\theta = 0$, откуда

$X^T Y = (X^T X + \lambda E)\theta \Rightarrow \hat{\theta} = (X^T X + \lambda E)^{-1} X^T Y$.

В МНК было $\hat{\theta} = (X^T X)^{-1} X^T Y$, а слагаемое λE может

сделать матрицу невырожденной (обратимой), тривиальный

пример: $X^T X = 0$, $\det(X^T X) = 0$, $\det(X^T X + \lambda E) = \det(\lambda E) \neq 0$

2) Шаг град. спуска

GD: $\theta_{t+1} = \theta_t - \eta \nabla F(\theta) \Rightarrow \theta_{t+1} = \theta_t + \eta (X^T Y - (X^T X + \lambda E)\theta_t)$

SGD: тоже, но берем только строки с номерами из I ($I =$

$= \{i_1, \dots, i_k\}$, $i_1, \dots, i_k \sim U\{1, \dots, n\}$ - б/вз), т.е. $\theta_{t+1} = \theta_t + \eta \frac{n}{k} \cdot$

$\cdot (X_I^T Y_I - (X_I^T Y_I + \lambda E)\theta_t)$

3) если признаки не приведены к одинаковому масштабу, то

штраф за величину весов применяется к весам "несправедливо",

неодинаково. Стандартизация позволяет привести столбцы к равному масштабу (по сути, избавиться от размерности и оставить только информацию о распределении признака), из-за чего ridge-регрессия работает заметно лучше.