Задача 3

① Найдем $\hat{\theta}$ в модели Ridge-регрессии в матр. виде. Сравним её с оценкой коэф-тов в МНК: Запишем функцию потерь:

$F(\theta)=\|Y-X\theta\|^2+\lambda\|\theta\|^2=(Y-X\theta)^T(Y-X\theta)+\lambda\theta^T\cdot\theta=Y^TY-2Y^TX\theta+$
$+\theta^TX^T\cdot X\cdot\theta+\lambda\theta^T\theta$. Найдем оценку $\hat{\theta}$, при которой она минимальна:

$\nabla F(\theta)=\cancel{-2Y^TX+2X^TX\theta+2\lambda}=-2X^TY+2X^TX\theta+2\lambda\theta=0\Rightarrow 2X^TY=$
$=2\cdot(X^TX+\lambda I)\theta\Rightarrow\hat{\theta}=(X^TX+\lambda I)^{-1}\cdot X^TY.$

Как мы видим, в отличие от МНК ($\hat{\theta}=(X^TX)^{-1}\cdot X^TY$), вырожденность матрицы X^TX нам не мешает, ведь, подобрав нужный гипер-параметр λ, мы можем добиться того, что $(X^TX+\lambda I)$ не будет вырожденной, а значит, всё хорошо.

② Выпишем ф-лу поиска оценки коэф-тов методом GD и SGD для модели Ridge-регрессии в матр. виде:

GD: $\theta_{t+1}=\theta_t-\eta\nabla F(\theta)=\theta_t-\eta\cdot(X^T\cdot(X\theta_t-Y)+\lambda\theta_t)=\theta_t-\eta\cdot\sum_{k=1}^{n}(x_k(x_k\theta_t-y_k)+\lambda\theta_t)$

SGD: $\theta_{t+1}=\theta_t-\eta\cdot\sum_{k\in I}(x_k(x_k\theta_t-y_k)+\lambda\theta_t)$, где $I=\{k_1,\dots,k_j\}$, где $k_1,\dots,k_j\sim U\{1,\dots,n\}$.

Иными словами, I - подвыборка из всех индексов, чтобы быстрее считать,

③ Для МНК стандартизацию можно не производить; однако для модели Ridge-регрессии это необходимо. Поясним это: Заметим, что ф-я потерь в модели Ridge-регрессии зависит от $\|\theta\|^2$, а значит, непосредственно от самих признаков. Но различные признаки могут иметь различные единицы измерения. Неразумно сравнивать разные ед. изм., т.к. это приводит к смещению оценок.
В МНК же стандартизация необязательна, т.к. ~~прямой~~ зависимости непосредственно от величин признаков нет.