Домашнее задание 1. Линейная регрессия.

Построить модель многомерной линейной регрессии без использования стандратных библиотек:

$$y_i = w_0 + w_1 x_{1i} + \ w_2 x_{2i} + \ldots + w_m x_{mi} + \varepsilon_i$$

где $w_0, w_1, \ldots, w_m$ - коэффициенты функции линейной регрессии, ε_i – случайная ошибка.

В качестве данных использовать следующий датасет:

- https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp
- https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236
- https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp

Замечание: каждая группа выбирает индивидуальный период времени – уникальный датасет для всех команд.

Провести анализ первичных данных данных по авиарейсам в США за определенный период времени.

Для построения модели необходимо использовать метод МНК, Стохастического градиентного спуска, AdaGrad, RMSProp, Adam.

Решить задачу при помощи метода наименьших квадратов. Напомним, что данный метод заключается в оптимизации функционала **MSE**:

$$MSE = \frac{1}{l} \sum_{i=1}^{l} \langle w, x_i - y_i \rangle^2 \to min$$

где $\{(x_i, y_i)\}_{i=1}^{l}$ - обучающая выборка, состоящая из l пар обьект-ответ.

Обучите линейную регрессию на 1000 объектах из обучающей выборки и выведите значения MSE и R^2 на этой подвыборке и контрольной выборке (итого 4 различных числа). Проинтерпретируйте полученный результат — насколько качественные прогнозы строит полученная модель? Какие проблемы наблюдаются в модели?

Далее используем $L1$- или $L2$-регуляризацию, тем самым получив Lasso и Ridge регрессии соответственно и изменив оптимизационную задачу одним из следующих образов:

$$MSE_{l1}(X, y) = \frac{1}{l} \sum_{i=1}^{l} \langle w, x_i - y_i \rangle^2 + \ \alpha \| w \|_1 \to min$$

$$MSE_{l2}(X, y) = \frac{1}{l} \sum_{i=1}^{l} \langle w, x_i - y_i \rangle^2 + \ \alpha \| w \|_2^2 \to min$$

где α — коэффициент регуляризации.

Обучите линейные регрессии с $L1$- и $L2$-регуляризатором, подобрав лучшее значение параметра регуляризации из списка alpha_grid при помощи кросс-валидации с 5 фолдами на тех же 1000 объектах. Выведите значения MSE и R^2 на обучающей и контрольной выборках.