Домашнее задание 4
Машинное обучение, БИБ20

Линейная регрессия

Пусть функция потерь обозначена как $S(y_i, \widehat{y_i})$, где y_i – "реальное"значение переменной, а $\widehat{y_i}$ – регрессионное предсказание. Как правило, все функции потерь рассчитывают разницу между y_i и $\widehat{y_i}$.

Например, для MSE функция потерь выглядит так:

$S_{MSE} = \frac{1}{n}\sum_{i=0}^{n}(y_i - \widehat{y_i})^2$ (n – количество объектов выборки).

Как рассчитать $\widehat{y_i}$? Линейная регрессия (на плоскости): $\widehat{y_i} = ax_i + b$, а значит $S_{MSE} = \frac{1}{n}\sum_{i=0}^{n}(y_i - (ax_i + b))^2$;

Зная предсказание x, нам необходимо вычислить коэффициенты a и b. Это можно сделать с помощью алгоритма градиентного спуска.

Сначала вычислим частные производные:

$$\frac{\partial S_{MSE}}{\partial a} = \frac{1}{n}\sum_{i=0}^{n} 2(y_i - (ax_i + b))(-x_i) = \frac{-2}{n}\sum_{i=0}^{n} x_i(y_i - (ax_i + b))$$

$$\frac{\partial S_{MSE}}{\partial b} = \frac{-2}{n}\sum_{i=0}^{n}(y_i - (ax_i + b))$$

;

Пусть дан шаг алгоритма ε, количество итераций p и исходные значения $a = a_0$, $b = b_0$. Тогда, согласно градиенту, можно итерационно для всех $j = \overline{1, p}$ вычислять

$$a_j = a_{j-1} - \varepsilon\frac{\partial S_{MSE}}{\partial a}|_{a=a_{j-1}}$$

$$b_j = b_{j-1} - \varepsilon\frac{\partial S_{MSE}}{\partial b}|_{b=b_{j-1}}$$

Тогда $a = a_p$, $b = b_p$ и уравнение аппроксимирующей прямой будет выглядеть как $y(x) = ax + b$.

1. Используйте датасет **sklearn.datasets.load_diabetes()**. Разобраться с тем, какие данные в нём содержатся, а также какая переменная является целевой, можно по ссылке;

2. Используйте любой известный алгоритм понижения размерности (например, LDA) для того, чтобы снизить количество признаков до одного

(вариант примитивнее – взять любую переменную исходного датасета, которую Вы считаете наиболее значимой).

3. Реализуйте алгоритм линейной регрессии с использованием градиентного спуска и функциями потерь S_{MSE} (см. выше) и S_{MAE} (продифференцируйте самостоятельно). Обратите внимание, что для данного пункта *запрещается* использовать готовые реализации методов (LinearRegression, mean_squared_error и т.д.);

4. Теперь постройте прогнозы, используя стандартную реализацию LinearRegression из sklearn;

5. Сравните основные метрики качества для "собственной"реализации и варианта из sklearn – MSE, MSLE, MAE, R^2, RMSE. Какой из двух алгоритмов оказался эффективнее? Какой менее подвержен переобучению?

6. Постройте на плоскости графики прямых (регрессий) для "собственной"реализации и варианта из sklearn.