Домашнее задание 2
Машинное обучение в прикладных задачах

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.

Кластеризация: о вкусах не спорят

Наверняка у каждого из нас есть любимый жанр музыки, песня или исполнитель.

Также наверняка мы хоть раз в жизни спорили с другими людьми о том, что именно наш любимый жанр или исполнитель – самый лучший и слушабельный.

Но так ли велика разница между различными жанрами музыки, если мы будем основываться только на фрагментах текста? В этом задании вам предлагается провести небольшое исследование на подобную тему.

Выберите несколько жанров музыки и самостоятельно составьте датасет, включающий в себя следующую информацию: отрывок некоторой песни (например, припев) и жанр этой песни. Вместо жанра можно использовать также определенных исполнителей.

Возможный датасет (пример №1):

text	genre
death blood satan...	rock
money girls money...	rap
you are the best...	pop
death again death...	rock

Возможный датасет (пример №2):

text	artist
death blood satan...	SlipKnot
money girls money...	Lil Pump
you are the best...	Katy Perry
death again death...	SlipKnot

В данном примере жанр (или исполнитель) являются **кластером**, а наша задача – по исходным данным провести **кластеризацию**, то есть отнести песню к определенному жанру (исполнителю) по отрывку из её текста.

Требования к датасету: минимум 150 объектов (песен) и от 4 до 20 кластеров (жанров или исполнителей).

Как набирать данные? Попробуйте найти тематические сайты с текстами песен, напишите небольшой веб-парсер, выкачивающий тексты для тех исполнителей, которых вы укажете. Также можно делать это руками без скриптов – смотрите сами, какой вариант Вам кажется быстрее.

Что использовать? Кластеризацию необходимо проводить методом k-means с использованием любых вспомогательных средств (TF-IDF, MiniBatch, нейронные сети, ...).

На что ещё обратить внимание? Для хорошего результата все тексты должны быть приведены к нижнему регистру и очищены от всех знаков препинания ("слово1 слово2 слово3 ..."). Перед самой кластеризацией тексты необходимо прогнать через Encoder (OneHot, Label, ...) – вспоминаем Д/З №2.

После разделения на кластеры сверьте то, насколько предсказание алгоритма совпало с реальным жанром (исполнителем) песни.

Согласны ли Вы с тем, что разделять песни по жанрам только на основе текста – это плохая идея?