Линейная регрессия

Цели работы:

1) реализовать два способа решения задачи линейной регрессии;
2) настроить гиперпараметры у каждого алгоритма, в частности параметры одного из методов регуляризации;
3) анализ результатов.

Данные

Используйте один из этих наборов данных для тестирования алгоритмов. Каждый тест в архиве организован следующим образом:

%число признаков%
%число объектов в тренировочном наборе%
%объект тренировочного набора 1%
%объект тренировочного набора 2%
…..
%объект тренировочного набора N%
%число объектов в тестовом наборе%
%объект тестового набора 1%
%объект тестового набора 2%
…..
%объект тестового набора K%

Формат объектов совпадает с форматом из соответствующей задачи на Codeforces.

Задание

Алгоритмы

Реализуйте алгоритмы нахождения уравнения прямой для задачи линейной регрессии:

- МНК — метод наименьших квадратов (псевдообратная матрица / SVD);
- градиентный спуск минимизирующий MSE.
- градиентный спуск минимизирующий SMAPE.

Регуляризация

В реализации каждого из вышеупомянутых алгоритмов необходимо использовать регуляризацию. Для МНК гребневую регуляризацию, для градиентного один из методов на выбор:

- гребневая;
- LASSO;

- Elastic Net.

Настройка

Для каждого алгоритма найдите наилучшие гиперпараметры, а именно, параметры регуляризации, а также шаг градиентного спуска и скорость затухания.

Тестирование

- Постройте график зависимости ошибки SMAPE и MSE на тестовом множестве от параметра регуляризации для метода наименьших квадратов.
- Постройте график зависимости экспоненциального скользящего среднего эмпирического риска на тренировочном множестве для градиентных спусков.
- Оцените каждый из трёх методов на тестовом множестве данных при помощи NRMSE и SMAPE.

Реализации

- Допустимо использовать библиотеки для вычисления псевдообратной матрицы и SVD.
- Требуется реализовать стохастический или пакетный градиентный спуск.
- Для алгоритма градиентного спуска рекомендуется использовать начальную инициализацию параметров $w_i \in [-\frac{1}{2n}; \frac{1}{2n}]$, где n — число признаков (см. лекцию). Другие способы инициализации параметров использовать также *можно*.
- Алгоритм градиентного спуска необходимо запустить с ограничением по числу итераций (не более 2000 итераций).
- Для настройки гиперпараметров можно использовать любой алгоритм. Рекомендуется использовать случайный поиск. Можно использовать готовые реализации, например optuna.
- Гиперпараметры лучше искать в логарифмированном пространстве. Например для случайного поиска можно генерировать случайное вещественное число *p* от -9 до 0, а затем использовать в качестве искомого параметра exp(*p*).
- Если вы используете метод настройки отличный от случайного, то целевую функцию ошибки следует модифицировать на случай, если алгоритм построения регрессии будет ломаться на плохих гиперпараметрах и параметры будут расходиться (превращаться в NaN или Inf). $L(p) = L_{old}(p)$, если алгоритм не сломался, иначе $L(p) = 2 + 1 /$ (число итераций до поломки).