lab1

Условие

Данная лабораторная работа будет представлять собой соревнование на Kaggle.

В рамках этого соревнования вам предстоит обучить модель линейной регрессии на основе датасета из «train.csv», содержащего около 10,000 строк. Предсказываемой переменной является «***RiskScore***». После обучения вы сможете проверить эффективность вашего алгоритма на тестовом наборе данных из «test.csv», а затем отправить в тестирующую систему. Более подробную информацию по данным и посылкам решений можной найти на странице соревнования.

Использование других моделей (кроме линейной регрессии), в том числе для отбора признаков, в данной работе **запрещено**.

Оценивание

Разбалловка

Работа оценивается в 15 баллов и складывается из двух компонент:

10р. *Обязательные задания по коду,*

5р. *Защита лабораторной работы.*

Задания по коду

Список обязательных заданий по коду (10 баллов) и их оценивание в баллах:

1р. *Провести разведочный анализ данных (EDA): построить графики зависимости некоторых признаков друг от друга, график целевой переменной и матрицу корреляций, сделать выводы.*

1р. *Реализовать нормализацию данных с помощью z-score и min-max.*

3р. *Реализовать класс линейной регрессии с обязательными методами fit и predict, метод fit реализовать через аналитическую формулу, через градиентный спуск и стохастический градиентный спуск. Сравнить полученные результаты с реализациями sklearn.*

3р. *Реализовать кросс-валидацию k-fold и leave-one-out*

0.5р. *Реализовать метрику MSE, протестировать и сравнить полученный результат с MSE из sklearn*

0.5р. *Реализовать MAE, протестировать и сравнить с метрикой из sklearn.*

0.5р. *Реализовать R^2, протестировать и сравнить с метрикой из sklearn.*

0.5р. *реализовать MAPE, протестировать и сравнить с метрикой из sklearn.*

Дополнительные 2 балла можно получить за реализацию и проверку линейной регрессии вместе с разными типами регуляризации: L_1, L_2, L_1+L_2. Ещё один дополнительный балл можно получить за реализацию L_p регуляризации, где значение p передаётся пользователем.

Основная метрика

Для получения баллов за код и последующего допуска к защите необходимо получить приемлемое значение основной метрики задачи.

В качестве основной метрики используется MSE (Mean Squared Error), которая рассчитывается по следующей формуле:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

где:

— n — количество наблюдений в тестовом наборе,

— y_i — истинное значение (целевое значение) для i-го наблюдения,

— $\hat{y}_i$ — предсказанное значение для i-го наблюдения.

MSE измеряет среднее значение квадратов ошибок, то есть разницу между предсказанными и истинными значениями. Чем меньше значение MSE, тем лучше модель предсказывает целевую переменную.

Максимальное пороговое значение MSE, которое можно получить при предсказаниях на тестовой выборке и при котором можно допуститься к защите работы составляет 18.00.

Важно отметить, что хотя MSE является основной метрикой для оценки качества модели и окончательный скор будет вестись только по MSE, вы должны использовать и другие метрики для самопроверки, такие как MAE (Mean Absolute Error), R^2 (коэффициент детерминации) и MAPE (Mean Absolute Percentage Error).

Загрузка ноутбука

После получения приемлемой метрики ($MSE < 18.00$) после посылки на Kaggle, необходимо загрузить ваш ноутбук («.ipynb»-файл) на свой GitHub для оценивания заданий по коду, а также для проверки на оригинальность. Затем нужно оставить комментарий в Google-таблице с оценками в столбце «lab1» в строке со своей фамилией о том, что вы выполнили работу с указанием ника на Kaggle.

Защита работы

После оценки заданий по коду, вы должны защитить лабораторную работу. За защиту можно получить до 5 баллов. На самой защите необходимо кратко рассказать про своё решение, а также ответить на вопросы по теории, связанные с линейной регрессией и обработкой данных.

Оценку за задания по коду можно повышать до дедлайна, оценку за защиту поднять нельзя.