Тема 11. Регуляризация моделей линейной регрессии. Метрики качества моделей регрессии

Практические задания для самостоятельного выполнения

Задания выполняются по вариантам. Формулировки заданий общие для всех вариантов; набор данных, подлежащих анализу, выбирается в соответствии с номером своего варианта.

Результаты выполнения заданий необходимо представить в виде двух файлов:

1) ноутбук в формате *ipynb*, содержащий программный код, результаты его выполнения, а также все необходимые пояснения, выводы и комментарии (в текстовых ячейках);
2) файл в формате *pdf*, полученный путем экспорта (или вывода на печать) ноутбука из п. 1).

Внимание: в названии файлов должна обязательно присутствовать фамилия автора. Безымянные работы проверяться не будут.

Задание 1 (максимум 1,25 балла).

В этом задании используются наборы данных с входными признаками **f1** – **f6** и прогнозируемым признаком ***y***. Наборы данных по вариантам сохранены в csv-файлах. Имена файлов: ВариантN, где N – номер варианта.
Выполнить первичное изучение данных и построить линейную модель прогнозирования величины ***y***.

- Импортировать данные из файла и вывести несколько первых записей (для контроля корректности импорта и получения представления о наборе).
- Выполнить первичный анализ данных:
  - визуализировать парные распределения признаков; сформулировать предположения о возможных зависимостях между признаками;
  - выполнить исследование на наличие корреляции между каждым из входных и прогнозируемым признаком, а также между входными признаками;
  - сформулировать выводы по результатам проведенного анализа; все комментарии, рассуждения и выводы записать в текстовых ячейках.
- Выполнить разбиение набора данных на обучающую и тестовую выборки в соотношении 90/10, установив *random_state*, равный номеру своего варианта.
- Для обучающей выборки
  - выполнить масштабирование входных признаков;
  - создать и обучить модель линейной регрессии на основе класса LinearRegression;
  - вывести коэффициенты обученной модели, проанализировать

результаты, объяснить причины наблюдаемой ситуации, сформулировать выводы (все рассуждения и выводы записать в текстовых ячейках);
    - получить прогнозы модели на обучающих данных;
    - оценить качество прогнозов на обучающей выборке, используя метрики MSE, MAE и R^2.

- Для тестовой выборки
    - выполнить масштабирование входных признаков, используя обученный ранее обработчик (не переобучая его!);
    - получить прогнозы обученной ранее модели на тестовых данных;
    - оценить качество прогнозов на тестовой выборке, используя метрики MSE, MAE и R^2.

- Сопоставить оценки качества на обучающей и тестовой выборке, сделать вывод относительно обобщающей способности полученной модели.

- Исследовать возможности применения регуляризации модели:
    - создать модель линейной регрессии с $\boldsymbol{L_2}$-регуляризатором и коэффициентом регуляризации, равным 5; обучить ее на обучающей выборке;
    - вывести коэффициенты обученной модели, проанализировать результаты, сопоставить с коэффициентами модели LinearRegression;
    - получить прогнозы модели на обучающих и тестовых данных; выполнить оценку качества прогнозирования на обучающих и тестовых данных с помощью указанных выше метрик; сделать выводы (записать в текстовых ячейках);
    - создать модель линейной регрессии с $\boldsymbol{L_1}$-регуляризатором и коэффициентом регуляризации, равным 5; обучить ее на обучающей выборке;
    - вывести коэффициенты обученной модели, проанализировать результаты, сопоставить с коэффициентами модели LinearRegression и гребневой регрессии;
    - получить прогнозы модели на обучающих и тестовых данных; выполнить оценку качества прогнозирования на обучающих и тестовых данных с помощью указанных выше метрик; сделать выводы (записать в текстовых ячейках);
    - сформулировать общие выводы о результатах применения регуляризации.

- Обосновать выбор лучшей из всех обученных моделей.
  Для лучшей модели
    - записать в текстовой ячейке соответствующее уравнение регрессии;
    - выполнить визуализацию результатов моделирования в координатах «правильные ответы» – «прогнозы модели» отдельно для обучающей и тестовой выборки; оценить качество прогнозов на основе визуального представления.

- Дополнительное задание (+0,25 балла):
    - поэкспериментировать с различными способами разбиения исходного набора данных на обучающую и тестовую выборку путем задания разных значений параметру *random_state*; проследить, как это влияет на оценки качества модели на обучающих и тестовых данных;
    - сформулировать выводы по результатам проведенных экспериментов (записать в текстовой ячейке).

Задание 2 (максимум 1 балл).

В этом задании используется тот же набор данных, что и в задании 1, с тем же разбиением на обучающую и тестовую выборку.
Реализовать вычислительные эксперименты с регуляризацией моделей линейной регрессии, проанализировать их результаты и сделать выводы. В процессе работы выполнить следующие действия.

- Обучить на обучающих данных несколько моделей линейной регрессии на основе SGDRegressor
    - модели с L_2-регуляризатором и коэффициентом регуляризации, принимающим значения: по умолчанию, 0.01, 0.1, 0.5, 1, 5, 10, 15;
    - модели с L_1-регуляризатором и коэффициентом регуляризации, принимающим значения: по умолчанию, 0.01, 0.1, 0.5, 1, 5, 10, 15.
- Вывести коэффициенты всех обученных моделей.
- Оценить качество прогнозов всех моделей на обучающей и тестовой выборке.
- Выполнить анализ всех полученных результатов: проследить, какое влияние на коэффициенты модели и величину ошибки оказывает тип регуляризатора, значение коэффициента регуляризации. Все рассуждения и выводы записать в текстовых ячейках.
- Выполнить подбор оптимального значения коэффициента регуляризации (отдельно для каждого типа регуляризатора) по метрике MSE.

  Указание: подбор оптимального коэффициента для L_2-регуляризатора можно выполнить с помощью инструментария класса RidgeCV. Основные приемы работы те же, что в классе LassoCV.
  Для формирования сетки значений, на которой будет выполняться перебор, использовать результаты предыдущих экспериментов (создать сетку в окрестности значений коэффициента, показавших наилучшие результаты).
- Дать интерпретацию всем результатам подбора (комментарии записать в текстовых ячейках).
- Сформулировать общие выводы по результатам всех проведенных экспериментов.