Министерство науки и высшего образования Российской Федерации

ФГБОУ ВО «Ульяновский государственный технический университет»

Кафедра «Вычислительная техника»

**Лабораторная работа №3**

Дисциплина: «Системы искусственного интеллекта»

«Линейная регрессия»

Вариант 12

Выполнил студент

группы ИВТАПбд-41

Меховников Е.А.

Проверил:

ассистент кафедры «ВТ»

Хайруллин И. Д.

Ульяновск, 2024

**Цель работы**

1. Написать программу, которая разделяет исходную выборку на обучающую и тестовую (training set, test set). Использовать стандартные функции.
2. С использованием библиотеки scikit-learn обучить модель линейной регрессии по обучающей выборке.
3. Проверить точность модели по тестовой выборке.
4. Построить модель с использованием полиномиальной функции. Построить графики зависимости точности на обучающей и тестовой выборке от степени полиномиальной функции.
5. Построить модель с использованием регуляризации. На основе экспериментов подобрать параметры для регуляризации. Построить графики зависимости точности модели на обучающей и тестовой выборках от коэффициента регуляризации.

**Задание по варианту**

12. Individual household electric power consumption - Индивидуальное потребление электроэнергии домохозяйствами.

Этот архив содержит 2075259 измерений, собранных в доме, расположенном в Sceaux (7 км от Парижа, Франция) в период с декабря 2006 года по ноябрь 2010 года (47 месяцев). Примечания: 1.(global\_active\_power \* 1000/60 - sub\_metering\_1 - sub\_metering\_2 - sub\_metering\_3) представляет активную энергию, потребляемую каждую минуту (в ватт-часах) в домашнем хозяйстве электрооборудованием, не измеряемым в подразделах 1, 2 и 3.

Набор данных содержит некоторые недостающие значения в измерениях (почти 1,25% строк). Все временные метки календаря присутствуют в наборе данных, но для некоторых временных меток значения измерений отсутствуют: отсутствующее значение представлено отсутствием значения между двумя последовательными разделителями атрибутов точкой с запятой. Например, в наборе данных отсутствуют значения по состоянию на 28 апреля 2007 года.

**Ход работы**

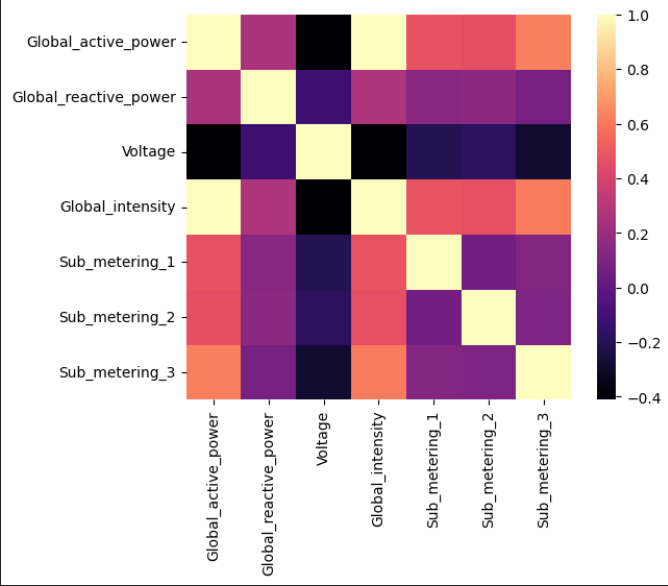
 Для начала с помощью матрицы кореляции была определена зависимость между парамтерами датасета (Рис. 1).

Рис. 1 – Матрица корреляции

В анализе данных корреляция играет важную роль, отражая взаимосвязь между различными показателями. Она позволяет понять, насколько изменения одной переменной влияют на изменения другой.

Из анализа данных следует, что Global\_active\_power демонстрирует наибольшую взаимосвязь с остальными параметрами. Поэтому именно его значение будет предсказываться.

1. Необходимо написать программу, которая будет разделять исходную выборку на тестовую и обучающую (training set, test set). Использовать стандартные функции (Листинг 1).

Листинг 1.

|  |
| --- |
| x\_data = np.array(x\_dataframe)              # Преобразование DataFrame x\_dataframe в массив numpy  y\_data = np.array(df[column\_to\_predict])    # Преобразование данных выбранной зависимой переменной в массив numpy  # Создание объекта SimpleImputer с стратегией заполнения пропущенных значений средними  imputer = SimpleImputer(strategy='mean')  # Применение SimpleImputer к данным  x\_data\_imputer= imputer.fit\_transform(x\_data)  # Преобразование y\_data в двумерный массив  y\_data\_2d = y\_data.reshape(-1, 1)  # Применение SimpleImputer к целевым данным  y\_data\_imputed = imputer.fit\_transform(y\_data\_2d)  val\_size = 0.2                              # Установка размера валидационной выборки в 20% от общего размера данных  x\_train, x\_val, y\_train, y\_val = train\_test\_split(x\_data\_imputer, y\_data\_imputed, test\_size=val\_size)   # Разделение данных на обучающую и валидационную выборки |

Часть данных в датасете отсутствует, её можно заполнить с помощью SimpleImputer.

1. С использованием библиотеки scikit-learn обучить модель линейной регрессии по обучающей выборке (Листинг 2).

Листинг 2

|  |
| --- |
| regr = linear\_model.LinearRegression()  # Создание объекта модели линейной регрессии  regr.fit(x\_data\_imputer, y\_data\_imputed)                # Обучение модели на обучающих данных |

1. Проверить точность модели по тестовой выборке (Листинг 3).

Листинг 3.

|  |
| --- |
| y\_val\_pred = regr.predict(x\_val)            # Предсказание значений целевой переменной на валидационной выборке  r2\_val = r2\_score(y\_val, y\_val\_pred)        # Вычисление коэффициента детерминации на валидационной выборке  print(f'Коэффициент детерминации: {r2\_val}')  Коэффициент детерминации: 0.9980340233250635 |

r2\_score - это функция, которая нужна для вычисления коэффициента детерминации (R²), функция является гармоничным средним между точностью и полнотой.

Коэффициент детерминации (R²) описывает, насколько хорошо модель предсказывает изменения в целевой переменной. Его значение колеблется от 0 до 1, где 1 означает, что модель объясняет 100% вариации зависимой переменной.. Данные были отображены на графике (Рис. 2).

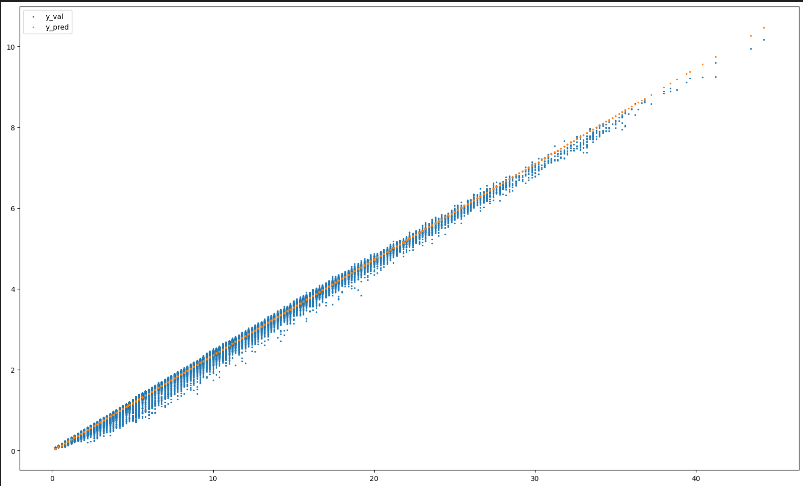


Рис. 2 – Датасет на графике.

1. Постройте модель, используя полиномиальную функцию. Постройте графики зависимости точности (r2\_score) от степени полиномиальной функции для обучающего набора и тестового набора (кривая валидации) (Рис. 3).

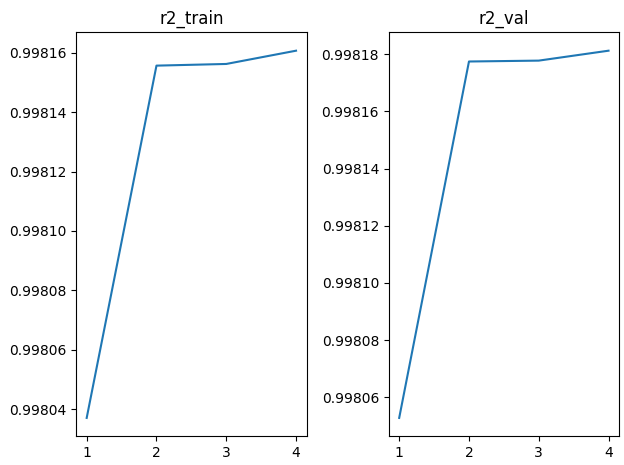


Рис. 3 – Коэффициенты детерминации при разных полиномах

На графике видно, что лучший полином – полином 4-й степени, а лучший коэффициент детерминации - 0.99818.

1. Построить модель с использованием регуляризации. На основе экспериментов подобрать параметры для регуляризации. Построить графики зависимости точности модели на обучающей и тестовой выборках от коэффициента регуляризации (Рис. 4).

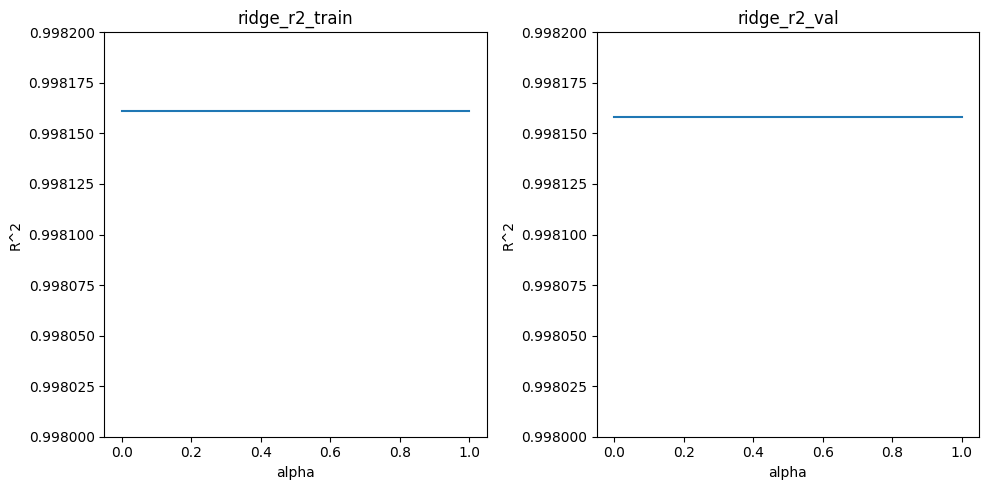
****

Рис. 4 – Коэффициенты детерминации в зависимости от alpha

**Вывод**

В ходе лабораторной работы проводились эксперименты с различными методами машинного обучения, применяемыми к обучающей выборке данных. Точность обученных моделей затем оценивалась на тестовой выборке. Вначале была построена модель линейной регрессии. Затем была применена полиномиальная функция для улучшения предсказательной способности модели. Были построены графики, показывающие зависимость точности модели от степени полиномиальной функции. Для оптимальной настройки параметров модели была использована регуляризация. В результате экспериментов были получены оптимальные значения параметров регуляризации. В конце были построены графики, демонстрирующие зависимость точности модели на обучающей и тестовой выборках от величины коэффициента регуляризации.