ПРАКТИЧЕСКАЯ РАБОТА 2

Цель:

Провести анализ данных и исследовать зависимости между признаками, чтобы лучше понять, как работает линейная регрессия.

Ход решения:

Шаг 1: Загрузка данных. Загрузите датасет, содержащий числовые признаки (например, Boston Housing, California Housing или другой подходящий из UCI или Kaggle). Импортируйте необходимые библиотеки для работы с данными, визуализацией и анализом. Загрузите данные в DataFrame и просмотрите первые строки, чтобы ознакомиться с набором данных.

Шаг 2: Исследование корреляций. Постройте корреляционную матрицу для числовых признаков. Визуализируйте корреляционную матрицу с помощью тепловой карты (heatmap). Определите пары признаков с наибольшей и наименьшей корреляцией. Для этого найдите максимальные и минимальные значения в корреляционной матрице, исключив диагональные элементы.

Шаг 3: Построение графиков зависимостей. Выберите два признака с сильной корреляцией. Замените на реальные названия признаков из вашего датасета. Постройте диаграмму рассеяния (scatter plot) и добавьте линию регрессии.

Шаг 4: Подготовка данных. Нормализуйте данные, чтобы привести все

признаки к одному масштабу, используя стандартное масштабирование. Разделите данные на обучающую и тестовую выборки в пропорции 80/20.

Шаг 5: Визуализация трендов. Используйте модель линейной регрессии для предсказания и постройте график распределения ошибок между предсказанными и реальными значениями. Постройте график ошибок (выбросов), чтобы оценить точность модели.

Шаг 6: Ответьте на вопросы. Как изменение одного признака влияет на другой? Как влияет масштабирование данных на качество модели?

Практическая работа 2 1
Цель: 1
Ход решения: 1
Теоретический материал по шагу два 5
Теория о корреляции 5
Виды корреляции 5
Корреляционная матрица 6
Визуализация с помощью тепловой карты (heatmap) 6
Определение максимальной и минимальной корреляции 6
Практическое применение 6
Теоретический материал по шагу три 8
Сильная корреляция 8
Коэффициент корреляции 8
Порог значений для сильной корреляции 8
Почему важна сильная корреляция? 9
Пример 9
Теоретический материал по шагу четыре 10
Теория о нормализации данных и разбиении выборок 10
Что такое нормализация? 10
Как применять на практике 11
Приведение к одному масштабу 12
Разделение данных на обучающую и тестовую выборки 12
Пропорции разбиения данных 12
Как использовать пропорции? 13
Теоретический материал по шагу пять 14
Линейная регрессия: Основы 14
Оценка модели с помощью визуализации 14
Построение графиков для визуализации точности 15
Точность модели 15
Виды метрик для оценки точности 15

Использование визуализации для улучшения модели 17

Теоретический материал по шагу два

Теория о корреляции

Корреляция — это статистическая мера, показывающая, насколько две переменные связаны друг с другом. Корреляция измеряется в диапазоне от -1 до 1, где:

- 1 означает полную положительную корреляцию: увеличение одной переменной приводит к увеличению другой.
- -1 означает полную отрицательную корреляцию: увеличение одной переменной приводит к уменьшению другой.
- 0 означает отсутствие корреляции: изменения одной переменной не связаны с изменениями другой.

Виды корреляции

1. Положительная корреляция: обе переменные увеличиваются или уменьшаются вместе. Пример: рост доходов обычно коррелирует с ростом расходов.
2. Отрицательная корреляция: одна переменная увеличивается, а другая уменьшается. Пример: увеличение времени работы может коррелировать с уменьшением свободного времени.
3. Отсутствие корреляции: изменения одной переменной никак не связаны с изменениями другой. Пример: количество осадков и количество проданных билетов в кино не всегда связаны.

Корреляционная матрица

Корреляционная матрица — это таблица, показывающая корреляционные коэффициенты между множеством переменных. Она помогает визуализировать, какие переменные имеют сильную или слабую связь друг с другом.

Визуализация с помощью тепловой карты (heatmap)

Тепловая карта — это графическое представление данных корреляционной матрицы, где различная интенсивность цвета указывает на уровень корреляции между переменными. Например, тёмные цвета могут показывать сильную положительную корреляцию, светлые — слабую, а противоположные цвета могут указывать на отрицательную корреляцию.

Определение максимальной и минимальной корреляции

Для анализа корреляционной матрицы важно определить пары признаков с максимальной и минимальной корреляцией, исключив диагональные элементы, которые всегда равны 1 (корреляция переменной с самой собой). Это позволяет выявить, какие признаки имеют наиболее сильное влияние друг на друга.

Практическое применение

Корреляция помогает в выявлении зависимостей в данных, что особенно полезно при построении моделей машинного обучения, когда необходимо

выбрать значимые признаки.

Также корреляция может использоваться для проверки гипотез и понимания природы данных.

Теоретический материал по шагу три

Сильная корреляция

Сильная корреляция — это ситуация, когда две переменные имеют тесную взаимосвязь, что выражается высоким по абсолютному значению коэффициентом корреляции.

Коэффициент корреляции

Коэффициент корреляции измеряет степень и направление связи между двумя переменными и варьируется от -1 до 1:

- 1 означает идеальную положительную корреляцию.
- -1 означает идеальную отрицательную корреляцию.
- 0 означает отсутствие связи между переменными.

Порог значений для сильной корреляции

Хотя пороги могут варьироваться в зависимости от области применения, общепринятые границы для оценки силы корреляции следующие:

- Слабая корреляция: от 0 до ±0.3
- Средняя корреляция: от ±0.3 до ±0.7
- Сильная корреляция: от ±0.7 до ±1

Сильная положительная корреляция (ближе к 1) означает, что увеличение одной переменной ведет к увеличению другой.

Сильная отрицательная корреляция (ближе к -1) означает, что увеличение одной переменной ведет к уменьшению другой.

Почему важна сильная корреляция?

1. Прогнозирование: В случае сильной корреляции изменение одной переменной позволяет с высокой точностью прогнозировать изменение другой.
2. Анализ взаимосвязей: Сильная корреляция помогает понять, какие факторы наиболее сильно влияют друг на друга, что особенно важно при построении моделей и принятии решений.
3. Выбор признаков: В машинном обучении признаки с сильной корреляцией могут быть полезными для построения моделей, но важно следить за мультиколлинеарностью (избыточностью признаков).

Пример

Допустим, у вас есть данные о высоте и весе людей. Если корреляция между этими переменными равна 0.8, это указывает на сильную положительную связь: чем выше человек, тем, как правило, больше его вес.

Использование таких понятий и подходов помогает при построении графиков зависимости, анализе данных и создании более точных моделей прогнозирования.

Теоретический материал по шагу четыре

Теория о нормализации данных и разбиении выборок

Нормализация данных — это процесс преобразования данных в такой формат, при котором все признаки (параметры) находятся в одном масштабе. Это помогает улучшить работу алгоритмов машинного обучения, так как многие модели чувствительны к различным масштабам признаков.

Что такое нормализация?

Нормализация — это приведение данных к стандартному диапазону, чаще всего путем изменения масштаба значений признаков.

Основные методы нормализации:

1. Min-Max Scaling (масштабирование до диапазона от 0 до 1):

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Все значения приводятся к диапазону [0, 1].

Пример: Если значения признака находятся в диапазоне от 10 до 50, то после применения Min-Max Scaling они будут варьироваться от 0 до 1.

2. Standard Scaling (стандартное масштабирование):

$$X_{norm} = \frac{X - \mu}{\sigma}$$

- μ — среднее значение признака
- σ — стандартное отклонение

После такого преобразования данные имеют среднее значение 0 и стандартное отклонение 1.

Пример: Если признак имеет значения с средним 50 и стандартным

отклонением 10, после применения Standard Scaling среднее значение станет 0, а отклонение — 1.

3. Robust Scaling (Масштабирование на основе медианы и межквартильного размаха). Этот метод особенно полезен, когда данные имеют выбросы, так как он использует медиану и межквартильный размах.

$$X_{scaled} = \frac{X - Median(X)}{IQR}$$

где: IQR (Interquartile Range) — разность между 75-м и 25-м процентилями.

4. Normalization (Нормализация на основе нормы вектора). Этот метод изменяет данные так, чтобы длина вектора значений признака была равна 1, что особенно полезно в задачах классификации.

$$X_{normalized} = \frac{X}{||X||}$$

Как применять на практике

1. Min-Max Scaling и Standard Scaling чаще всего применяются через библиотеку scikit-learn:

```
from sklearn.preprocessing import MinMaxScaler, StandardScaler

# Для Min-Max Scaling
scaler = MinMaxScaler()
X_scaled = scaler.fit_transform(X)

# Для Standard Scaling
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

2. Robust Scaling полезен для данных с выбросами:

```
from sklearn.preprocessing import RobustScaler

scaler = RobustScaler()
X_scaled = scaler.fit_transform(X)
```

Приведение к одному масштабу

Приведение признаков к одному масштабу важно для многих алгоритмов, таких как линейная регрессия, метод k-ближайших соседей и нейронные сети, поскольку они чувствительны к масштабам признаков. Например, если один признак имеет диапазон значений от 0 до 1, а другой — от 0 до 1000, алгоритм может неправильно оценивать значимость признаков.

Разделение данных на обучающую и тестовую выборки

Разделение данных — это процесс деления набора данных на две части: обучающую и тестовую выборки.

- Обучающая выборка (Training set): используется для обучения модели, то есть для нахождения зависимостей между признаками и целевой переменной.
- Тестовая выборка (Test set): используется для оценки качества модели на данных, которые она ранее не видела, что позволяет проверить её обобщающую способность.

Пропорции разбиения данных

Пропорции разбиения данных обычно выбираются в зависимости от

объема данных и задачи. Чаще всего используются следующие пропорции:

- 80/20: 80% данных идут на обучение модели, а 20% — на тестирование. Это популярное соотношение, обеспечивающее достаточно данных для обучения при наличии небольшой тестовой выборки для оценки.
- 70/30: используется, когда необходимо больше данных для тестирования.
- 90/10: применяется, когда доступно большое количество данных, и нужно сохранить как можно больше для обучения.

Как использовать пропорции?

1. Выберите соотношение, например, 80/20.
2. Разделите данные случайным образом, чтобы обе выборки были репрезентативны.
3. Используйте обучающую выборку для создания модели и тестовую выборку для оценки её производительности.

Такой подход помогает проверить модель на устойчивость и избежать переобучения, позволяя убедиться, что модель не подстраивается под конкретные данные, а может обобщать результаты на новые данные.

Теоретический материал по шагу пять

Линейная регрессия: Основы

Линейная регрессия — это метод статистического моделирования, который используется для анализа и предсказания зависимой переменной на основе одной или нескольких независимых переменных. Она строит линейную зависимость между переменными, что позволяет находить тренды и делать прогнозы.

Оценка модели с помощью визуализации

1. График распределения ошибок (Residual Plot)

График распределения ошибок показывает разницу между предсказанными и реальными значениями (остатки). Это полезный инструмент для оценки качества модели, так как он помогает увидеть, насколько хорошо модель справляется с данными.

- Ось X: предсказанные значения.
- Ось Y: ошибки (остатки), то есть разница между реальными и предсказанными значениями.

Идеальная модель будет иметь остатки, распределенные случайным образом вокруг нуля, без явных паттернов или трендов. Это означает, что модель хорошо описывает данные и не систематически ошибается.

2. График ошибок (выбросов)

График ошибок (график выбросов) помогает обнаружить, где модель ошибается наиболее значительно, что может указывать на выбросы или аномалии в данных. Такие данные часто оказывают сильное влияние на модель и могут искажать результаты.

Выбросы: точки, которые значительно отличаются от других значений в данных. Они могут быть вызваны ошибками измерения, редкими событиями или необычными условиями.

Построение графиков для визуализации точности

1. График распределения ошибок

- Постройте график остатков: если остатки распределены случайно, модель работает хорошо.
- Ищите паттерны: если остатки не случайны, это сигнал, что модель не учла какую-то важную зависимость в данных.

2. График ошибок (выбросов)

- Постройте график, показывающий предсказанные значения против реальных значений.
- Определите выбросы: точки, которые отклоняются от общей тенденции, могут указывать на проблемы с данными или недочеты модели.

Точность модели

Точность модели — это метрика, которая измеряет, насколько хорошо модель машинного обучения предсказывает результаты. Она помогает оценить производительность модели и понять, насколько успешно она справляется с задачей на данных, которые ей предоставлены.

Виды метрик для оценки точности

1. Точность (Accuracy) для классификации.

Для задач классификации, таких как определение принадлежности объекта к классу (например, «да» или «нет»), точность измеряет долю правильных предсказаний.

$$\text{Точность} = \frac{\text{Количество верных предсказаний}}{\text{Общее количество предсказаний}}$$

Подходит для задач, где классы равномерно сбалансированы.

Не рекомендуется для несбалансированных классов, так как может вводить в заблуждение (например, модель, которая всегда предсказывает «нет», будет иметь высокую точность на сильно несбалансированных данных).

2. Среднеквадратичная ошибка (Mean Squared Error, MSE) для регрессии.

Для задач регрессии (прогнозирование числовых значений), как, например, предсказание цен на жилье, используется среднеквадратичная ошибка. MSE измеряет среднее квадратичное отклонение предсказанных значений от реальных.

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

Где:

- y_i — реальное значение.
- $\hat{y}_i$ — предсказанное значение.
- n — количество наблюдений.

Чем меньше значение MSE(от 0 до бесконечности), тем точнее модель предсказывает данные.

3. Средняя абсолютная ошибка (Mean Absolute Error, MAE).

MAE измеряет среднюю абсолютную разницу между предсказанными и реальными значениями, что делает её интерпретируемой и устойчивой к выбросам по сравнению с MSE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

4. Коэффициент детерминации R^2.

R^2 показывает, какую долю вариации в данных модель объясняет. Значение R^2 варьируется от 0 до 1:

$R^2 = 1$: модель идеально объясняет данные.

$R^2 = 0$: модель не лучше, чем среднее значение.

5. F1-мера для классификации.

Используется в задачах классификации для оценки модели на несбалансированных данных. Она объединяет точность (precision) и полноту (recall) в одну метрику.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Использование визуализации для улучшения модели

- Диагностика проблем: если остатки показывают паттерны, это может указывать на то, что линейная модель не подходит и нужно попробовать другие подходы, такие как полиномиальная регрессия или добавление новых признаков.
- Выявление выбросов: выбросы можно либо исключить из анализа, либо изучить их природу для улучшения модели.