СПбГЭТУ “ЛЭТИ”, ФКТИ, МОЭВМ

Введение в анализ данных. Scikit

Для вопросов по курсу:
Иванов Дмитрий Владимирович, dmitry.ivanov@moevm.info
Префикс в теме письма [CS_23XX]

Предобработка данных. Скейлеры

Предобработка данных - стандартизация, масштабирование и нормализация данных

- обеспечить высокую точность при анализе данных
- минимизировать влияние выбросов (маленьких/больших значений)
    - например, на обучение моделей машинного обучения

Скейлеры. StandardScaler

- Что делает:
  - центрирует данные (среднее значение становится равным нулю)
  - масштабирует так, чтобы стандартное отклонение стало равным 1
- Применяется к данным с нормальным распределением
- Использование:

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaled_data = scaler.fit_transform(data)
```

$$Z_i = \frac{X_i - mean(X)}{std(X)}$$

Скейлеры. MinMaxScaler

- Что делает:
  - приводит данные к диапазону от 0 до 1
- Применяется к данным, имеющим разную амплитуду
- Использование:

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
scaled_data = scaler.fit_transform(data)
```

$$Z_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Скейлеры. MaxAbsScaler

- Что делает:
  - масштабирует данные
    - максимальное значение становится равным 1
    - все остальные значения масштабируются пропорционально.
- Применяется к данным со знаками, категориями и бинарными признаками
- Использование:

```
from sklearn.preprocessing import MaxAbsScaler
scaler = MaxAbsScaler()
scaled_data = scaler.fit_transform(data)
```

$$Z_i = \frac{x_i}{x_{max}}$$

Скейлеры. RobustScaler

- Что делает:
  - масштабирует данные с использованием медианы и интерквартильного расстояния
  - более устойчив к выбросам
- Применяется для данных с выбросами
- Использование:

```
from sklearn.preprocessing import RobustScaler
scaler = RobustScaler()
scaled_data = scaler.fit_transform(data)
```

$$Z_i = \frac{x_i - med(X)}{x_{25} - x_{75}}$$

Прогнозирование. Линейная регрессия

- Метод машинного обучения, широко используется для построения моделей и прогнозирования результатов в пространстве непрерывных значений
- Используется для анализа, исследования отношений между двумя переменными и прогнозирования новых значений
- Цель — поиск линейной функции, которая наилучшим образом соответствует заданным данным
- Общее уравнение

$$f(x_1, \dots x_n) = w_0 + w_1 x + \dots w_n x_n$$

Обработка данных. Кластеризация

- Метод машинного обучения без учителя - не использует предварительно размеченных данных для обучения модели
- Метод разделения набора данных на группы (кластеры)
  - объекты внутри одного кластера похожи (по некоторым признакам)
  - объекты из разных кластеров - сильно отличались друг от друга
- Алгоритмы кластеризации:
  - KMeans
    - метод k-средних: разбивает данные на заранее заданное количество кластеров, определяя центры каждого кластера
  - DBSCAN
    - плотностной метод кластеризации: объекты вокруг одной точки объединяются в кластер
    - работает с данными высокой размерности; когда невозможно задать заранее количество кластеров

Обработка данных. Классификация

- Метод машинного обучения с учителем - обучение модели происходит на основе заранее определенных категорий, и последующей классификации новых объектов на основе обученной модели
- Метод определения принадлежности объекта к определенному классу на основе его характеристик и заранее заданных классов
- Алгоритмы классификации:
  - Логистическая регрессия
  - Байесовские классификаторы
    - вероятностные классификаторы, предсказывающие класс с самой большой условной вероятностью для заданного вектора признаков
  - Решающие деревья
    - принятие решений/классификации в зависимости от значений характеристик

Классификация. Логистическая регрессия

- Метод машинного обучения, используется для решения задач классификации, т.е. для определения, к какому из заданных классов может быть отнесен каждый объект данных
  - используется логистическая функция
- В отличие от линейной регрессии работает с бинарными значениями или значениями классов
- Цель — классификация объектов на два класса (бинарная классификация) или на несколько классов (многоклассовая классификация).

Наивные Байесовские классификаторы

- Использование статистических методов для определения принадлежности объекта к классу, классификация объектов на основе относительных вероятностей
  - "Наивность" классификаторов — предположение и независимости признаков объекта друг от друга
  - Основаны на принципе максимального правдоподобия для оценки параметров модели — стремятся построить модель, которая наиболее правдоподобно объясняет входные данные
- Классификаторы
  - GaussianNB — "Гауссовский" классификатор, предполагаются данные с гауссовым (нормальным) распределением
  - MultinomialNB — "полиномиальный" классификатор для полиномиально распределенных данных
  - BernoulliNB — предполагаются данные с распределением Бернулли (каждый из признаков является двоичной переменной)
  - CategoricalNB — категориальный классификатор, предполагается, что каждый признак имеет свое категориальное распределение

Классификация. Решающие деревья

- Построения дерева решений — поиск наилучших условий для разделения выборки на чистые (не пересекающиеся в поддеревьях) классы
  - каждый узел (вершина дерева) — условие, проверяемое для некоторой переменной
  - каждый лист (конечная вершина дерева) — значение класса, к которому принадлежит объект
- Процесс классификации — спуск по дереву от корня до листа, классификационного ответа дерева
- Преимущество — интерпретируемость, т.е. прозрачность и понятность полученной модели
- Недостаток — переобучение и высокая чувствительность к шуму в данных

Вопросы по курсу можно задавать:

Иванов Дмитрий Владимирович
dmitry.ivanov@moevm.info