Лабораторная работа №4
Методы заполнения пропущенных значений в данных

Студент группы XXX

7 мая 2025 г.

1 Введение

В данной лабораторной работе рассматриваются различные методы заполнения пропущенных значений в наборах данных. Работа включает в себя реализацию и сравнение различных методов импутации, их визуализацию и оценку эффективности.

2 Описание методов импутации

В работе реализованы следующие методы заполнения пропущенных значений:

2.1 Простые методы

- Заполнение средним значением (mean)
- Заполнение медианой (median)
- Заполнение модой (mode)
- Заполнение предыдущим значением (ffill)

2.2 Продвинутые методы

- Hot-deck импутация
- Линейная регрессия

- Стохастическая регрессия
- Сплайн-интерполяция

3 Реализация

3.1 Основная структура проекта

Проект состоит из следующих основных модулей:

- `main.py` - основной файл для запуска анализа
- `data_loading.py` - загрузка данных
- `data_preprocessing.py` - предварительная обработка данных
- `imputation_methods.py` - реализация методов импутации
- `evaluation.py` - оценка методов
- `visualization.py` - визуализация результатов

3.2 Код реализации методов импутации

```
# Пример реализации метода линейной регрессии
def fill_missing(df, method="linear_regression", **kwargs):
    if method == "linear_regression":
        target_col = kwargs.get("target_col")
        feature_cols = kwargs.get("feature_cols")

        # Заполнение пропусков в признаках
        df_filled = df.copy()
        for col in feature_cols:
            if df_filled[col].isna().any():
                df_filled[col] = df_filled[col].fillna(
                    df_filled[col].median())

        # Обучение модели
        known = df_filled[df_filled[target_col].notna()]
        unknown = df_filled[df_filled[target_col].isna()]

        model = LinearRegression()
```

```
        model.fit(known[feature_cols], known[target_col])
        predicted = model.predict(unknown[feature_cols])

        df_filled.loc[unknown.index, target_col] =
            predicted
        return df_filled
```

4 Метрики оценки

Для оценки эффективности методов импутации использовались следующие метрики:

- Средняя относительная ошибка (MeanRelativeError%) - показывает среднее отклонение предсказанных значений от истинных
- Ошибки в распределении данных - оценивают, насколько хорошо методы сохраняют статистические характеристики исходных данных:
  - Ошибка в среднем значении
  - Ошибка в стандартном отклонении
  - Ошибка в квантилях распределения

5 Методология оценки

Оценка методов проводилась следующим образом:

1. Формирование датасета из полных наблюдений
2. Внесение случайных пропусков в данные (3%, 5%, 10%, 20%, 30%)
3. Применение различных методов импутации
4. Сравнение результатов с истинными значениями
5. Оценка сохранения статистических характеристик данных

Для каждого уровня пропусков проводилось 5 запусков для получения статистически значимых результатов.

6 Результаты

6.1 Сравнение методов

Тестирование проводилось на трех наборах данных разного размера:

- Маленький набор данных (sm_dataset)
- Средний набор данных (m_dataset)
- Большой набор данных (lg_dataset)

Для каждого уровня пропусков был определен лучший метод импутации на основе средней относительной ошибки. Результаты представлены в виде графика зависимости ошибки от процента пропусков для каждого метода.

6.2 Выводы

- Простые методы (mean, median, mode) показывают хорошие результаты при малом проценте пропусков (до 5%)
- При увеличении процента пропусков более эффективными становятся продвинутые методы (линейная регрессия, сплайн-интерполяция)
- Стохастическая регрессия показывает лучшие результаты в сохранении распределения данных
- Hot-deck импутация эффективна при наличии коррелированных признаков

7 Заключение

В ходе работы были реализованы и протестированы различные методы заполнения пропущенных значений. Каждый метод имеет свои преимущества и недостатки, и выбор конкретного метода зависит от специфики данных и требований задачи.