1. Эксперимент по решению задач на регрессию и оптимизации

1.1 Подготовка к эксперименту

Время выполнения эксперимента: 120 минут

Цель

После завершения данного эксперимента вы сможете:

- Познакомьтесь с концепцией регрессионных моделей, их распространенными типами и методами оценки качества.
- Ознакомьтесь с основными методами предварительной обработки данных и анализа данных с использованием библиотек pandas и numpy.
- Ознакомьтесь с основными методами визуализации данных с использованием библиотек математического пакета matplotlib и Seaborn.
- Овладение алгоритмами регрессии из набора инструментов scilearn-kit.
- Изучение методов комбинирования моделей с использованием инструментов xgb, lgb и catboost.

Предварительная среда

Для проведения эксперимента необходимо подготовить среду работы с Python версии 3.6 и выше. Экспериментальная среда должна включать следующие элементы:

- • Anaconda

1.2 Цель эксперимента

Метод регрессии обычно используется для прогнозирования конкретного числового значения, находящегося в непрерывном диапазоне, например, цены на жильё или акции. Например, на основе данных о количестве лет с изменением уровня PM2.5 в определённом регионе можно оценить значение PM2.5 в тот день. Чем ближе прогноз к фактическому значению, тем выше надёжность алгоритма регрессионного анализа.

Цель данного эксперимента — предсказать цену подержанного автомобиля на основе таких факторов, как бренд, возраст автомобиля и пробег. В ходе эксперимента изучаются критерии оценки регрессионных моделей, методы анализа данных, инженерия признаков, а также процесс настройки параметров

моделей.

1.3 Распространённые показатели оценки задач регрессионного прогнозирования

Показатели оценки — это числовые количественные характеристики эффективности прогнозирования модели. В регрессионных задачах наиболее часто используются следующие показатели: средняя абсолютная ошибка (MAE), среднеквадратичная ошибка (MSE), $y_i \hat{y}_i \overline{y}_i$средняя абсолютная процентная ошибка (MAPE), корневая среднеквадратичная ошибка (RMSE) и коэффициент R² (R²). В формуле данного раздела реальные значения обозначаются как , прогнозные значения — как , а среднее значение выборки — как .

- *Средняя абсолютная ошибка (Mean Absolute Error, MAE) — это показатель, который лучше отражает* $MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right|$*степень несоответствия между предсказанными и фактическими значениями. Формула для расчета средней абсолютной ошибки следующая:*
- *Среднеквадратичная ошибка (MSE) рассчитывается по следующей формуле:* $MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_1\right)^2$
- Формула для расчета коэффициента R² (коэффициента квадратичности) следующая:

$$\textit{Сумма квадратов остатков:} SS_{res} = \sum\left(y_i - \hat{y}_i\right)^2$$

$$\textit{Средний общий показатель:} ss_{tot} = \sum\left(y_i - \overline{y_i}\right)^2$$

Здесь $\overline{y}R^{2-}$ обозначает среднее значение переменной y;
следовательно, выражение получается следующим образом:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum\left(y_i - \hat{y}_i\right)^2}{\sum\left(y_i - \overline{y_i}\right)^2}$$

R² показывает долю объяснённой вариации зависимой переменной, объясняемой независимыми переменными. Его значение находится в диапазоне

от 0 до 1. Чем ближе R^2 к 1, тем больше доля объяснённой части общей дисперсии, тем ближе регрессионная линия к наблюдаемым точкам, тем больше объяснённой части изменения y объясняется изменением x, и тем лучше соответствует модель реальным данным. Поэтому R^2 называют статистикой качества подгонки. Чем выше значение R^2, тем лучше подгонка.

1.4 Упражнение 1: Исследование данных

В машинном обучении методы анализа данных обычно называются EDA (Exploratory Data Analysis). Они представляют собой подходы к исследованию исходных данных при минимальном количестве предварительных допущений. С помощью графического представления данных, составления таблиц, аппроксимации данных с помощью уравнений, расчета характеристик и других методов анализаторов выявляются структура и закономерности в данных.

Основная ценность EDA заключается в следующем: 1. Понимание и знакомство с набором данных, его проверка на соответствие требованиям для последующего моделирования и анализа. 2. Анализ взаимосвязей между переменными и их связи с предсказуемыми значениями. 3. Обработка данных и выполнение этапов инженерии признаков, что позволяет сделать структуру набора данных и его признаков более надёжными для последующих задач прогнозирования.

Процесс анализа данных обычно заключается в проведении статистических анализов и подведении итогов с помощью графиков или текстовых материалов. При этом исходные данные не изменяются.

1. Загрузка различных библиотек для работы с данными и визуализации:
    - Библиотеки для работы с данными: pandas, numpy, scipy.
    - Инструменты для визуализации данных: matplotlib, Seabon.
    - Другое;
2. Загрузка данных:
    - Загрузка обучающего и тестового наборов данных;
    - Краткое описание данных (формат: head() + shape);
3. Обзор данных:
    - Изучите статистические показатели данных с помощью функции describe().

- Изучение типов данных с помощью функции info()

4. Определение отсутствия данных и аномалий
    - Проверьте наличие значений типа «nan» в каждой колонке.
    - Обнаружение аномальных значений
5. Определение распределения прогнозных значений
    - Общие характеристики распределения (распределение безграничного Джонсона и т. д.)
    - Проверьте степень смещения (skewness) и крутости (kurtosis).
    - Просмотреть конкретную частоту прогнозов
6. Характеристики делятся на категориальные и числовые, при этом для категориальных характеристик анализируется уникальное распределение.
7. Анализ цифровых характеристик
    - Анализ корреляции
    - Проверьте асимметрию и пиковые значения нескольких характеристик.
    - Визуализация распределения каждого цифрового признака
    - Визуализация взаимосвязей между цифровыми характеристиками
    - Визуализация взаимосвязей между многими переменными
8. Анализ характеристик типов
    - распределение
    - Визуализация характеристик категорий с помощью диаграмм в виде коробок
    - Визуализация характеристик категорий с использованием графического представления виолины
    - Визуализация характеристик категорий с помощью столбчатых диаграмм
    - Визуализация частоты каждого класса признаков (count_plot)
9. Создание отчета о профилировании данных с использованием библиотеки pandas_profiling

1.5 Упражнение 2: Инженерия признаков

Целью работы с признаками является дальнейший анализ этих признаков и предварительная обработка данных с целью их адаптации к процессу

моделирования. В ходе этой работы удаляются аномальные значения, восстанавливаются пропущенные данные, а также выполняются необходимые преобразования или преобразования данных в соответствии с требованиями.

К типичным инженерным характеристикам относятся:

1. Обработка исключений:
    - Удаление аномальных значений с помощью диаграммы бокс-лайн (или метода 3-Сигма).
    - Преобразование данных с использованием алгоритма BOX-COX (с учетом асимметричности распределения данных).
    - Сокращение длинного хвоста;
2. Нормализация/стандартизация характеристик:
    - Стандартизация (преобразование данных в стандартное нормальное распределение);
    - Нормализация (преобразование в диапазон [0,1]);
    - Для распределения по закону степенного распределения можно использовать формулу: $\log(1 + x1 + \text{медиана})$
3. Разделение данных на группы (бочки):
    - Разделение на равные интервалы по частоте;
    - Равномерное разделение на бочки;
    - Метод Best-KS для разделения данных на баки (аналогично использованию индекса Джини для двоичной классификации);
    - Карточные ячейки для распределения данных по критерию хи-квадрата;
4. Обработка отсутствующих значений:
    - Не обрабатывается (для моделей на основе деревьев, таких как XGBoost);
    - Удаление (слишком много данных отсутствует);
    - Интерполяция, включая среднее значение, медиану, моду, моделирование, многократную интерполяцию, компрессионное восстановление, матричное восстановление и другие методы.
    - Разделение на ящики; отсутствующие значения — в одном ящике.
5. Характеристические структуры:

- Представляет статистические характеристики: отчеты по количественным данным, суммам, процентным соотношениям, стандартным отклонениям и другим показателям.
- Характеристики времени, включая относительное и абсолютное время, праздничные дни, выходные и т. д.;
- Географическая информация, включая методы разбивки на ячейки и кодирования распределения;
- Нелинейные преобразования, включая логарифмические, квадратные и корневые преобразования;
- Комбинации признаков, перекрёстные связи между признаками
- Каждый видит по-своему.

6. фильтрация признаков
    - Фильтрация данных: сначала проводится отбор признаков, затем обучается алгоритм. Среди распространенных методов отбора признаков — методы на основе различий в значениях признаков (Relief), выбора наиболее значимых признаков по критерию дисперсии, корреляционных коэффициентов, критерия хи-квадрата и метода взаимной информации.
    - Методы оценки эффективности алгоритмов обучения, основанные на использовании специальных инструментов (например, LVM – Las Vegas Wrapper), часто применяются для определения характеристик алгоритмов, предназначенных для использования в конкретных задачах.
    - Метод встраивания (embedding) сочетает в себе подходы фильтрации и обёртывания; при обучении алгоритма происходит автоматический отбор признаков. К распространённым методам относится регрессия с использованием лассо-модели (lasso regression).
7. Снижение размерности
    - PCA/ LDA/ ICA ;
    - Выбор признаков также является способом снижения размерности данных.

1.6 Упражнение 3: Моделирование и настройка параметров

Ознакомьтесь с распространенными моделями машинного обучения и освоите процесс их построения и настройки. К основным моделям регрессионного прогнозирования относятся:

- Модель линейной регрессии (https://zhuanlan.zhihu.com/p/49480391)
- Модель деревьев решений (https://zhuanlan.zhihu.com/p/65304798)
- Модель GBDT (Gradient Boosting Tree) (https://zhuanlan.zhihu.com/p/45145899)
- Модель XGBoost (https://zhuanlan.zhihu.com/p/86816771)
- Модель LightGBM (https://zhuanlan.zhihu.com/p/89360721)

1. Модель линейной регрессии:
    - Требования к характеристикам при использовании метода линейной регрессии;
    - Обработка распределений с длинным хвостом;
    - Понимание модели линейной регрессии;
2. Проверка качества модели:
    - функция оценки и целевая функция;
    - Метод кросс-валидации;
    - Оставьте один метод проверки.
    - Проверка на основе временных рядов;
    - Нарисуйте кривую скорости обучения.
    - Построение кривой проверки;
3. Встраиваемый метод выбора признаков:
    - Регрессия с использованием метода лассо (Lasso regression)
    - Регрессия по методу Риджа (Ridge regression)
    - Дерево решений;
4. Сравнение моделей:
    - Часто используемые линейные модели;
    - Нерегулярные модели, часто используемые в практике;
5. Настройка параметров модели:

- Методы настройки параметров с использованием принципа жадности (greed tuning);
- Методы настройки параметров сетки;
- Метод настройки параметров по принципу Байеса;

1.7 Упражнение 4: Слияние моделей

Для получения более точных результатов прогнозирования модели, в которых используются различные параметры, применяются следующие методы их объединения:

1. Простое взвешенное слияние данных:
    - Регрессия (вероятность классификации): арифметическое среднее, геометрическое среднее.
    - Категория: Голосование
    - Комплексный метод: среднее значение по рангу (rank averaging) и логарифмическое среднее (log averaging).

2. stacking/blending:
    - Создание многоуровневой модели и использование результатов прогнозирования для дальнейшей корректировки прогнозов.

4. Методы усиления/ослабления результатов (boosting/bagging) уже используются в алгоритмах xgboost, Adaboost и GBDT.

    - Методы улучшения процесса объединения нескольких деревьев

1.8 Задание для эксперимента

Набор данных Used_car был загружен на локальный компьютер; с его помощью предсказывается цена на подержанные автомобили на основе знаний, полученных на этом занятии.

Выберите из следующих вопросов:

https://www.jianshu.com/p/20268a0fe809

Соревнования для новичков:

Kaggle: Прогнозы цен на недвижимость

Этот конкурс, являющийся одним из самых простых примеров задач на возвращение к основам, идеально подходит для тех, кто только начинает изучать машинное обучение.
Сайт:
https://www.kaggle.com/c/house-prices-advanced-regression-techniques
Классическое решение:
Решение на основе технологии XGBoost:
https://www.kaggle.com/dansbecker/xgboost
Решение Lasso:
https://www.kaggle.com/mymkyt/simple-lasso-public-score-0-12102

Дальнейшие этапы соревнований:

Kaggle – прогнозы объемов продаж
Эта задача относится к классическим задачам на основе временных рядов; её цель — предсказать общий объем продаж каждого товара и в каждом магазине в следующем месяце.
Сайт:
https://www.kaggle.com/c/competitive-data-science-predict-future-sales
Классическое решение:
LightGBM:
https://www.kaggle.com/sanket30/predicting-sales-using-lightgbm
XGBoost: https://www.kaggle.com/fabianaboldrin/eda-xgboost
Решение для первого места:
https://www.kaggle.com/c/competitive-data-science-predict-future-sales/discussion/74835#latest-503740

План проведения турнира TOP:

Kaggle: Прогноз числа посетителей ресторана
Сайт: https://www.kaggle.com/c/recruit-restaurant-visitor-forecasting
Решение:
Первый вариант решения:
https://www.kaggle.com/plantsgo/solution-public-0-471-private-0-505
Седьмой вариант решения:
https://www.kaggle.com/c/recruit-restaurant-visitor-forecasting/discussion/49259#latest-284437
Восьмой вариант решения:
https://github.com/MaxHalford/kaggle-recruit-restaurant
12-й вариант решения:
https://www.kaggle.com/c/recruit-restaurant-visitor-forecasting/discussion/49251#latest-282765
Прогнозы продаж на платформе Kaggle (CorporacionFavoritaGrocery)
Сайт: https://www.kaggle.com/c/favorita-grocery-sales-forecasting
Решение:
Первый вариант решения:
https://www.kaggle.com/c/favorita-grocery-sales-forecasting/discussion/47582#latest-360306

Второй вариант решения:
https://www.kaggle.com/c/favorita-grocery-sales-forecasting/discussion/47568#latest-278474
Третий вариант решения:
https://www.kaggle.com/c/favorita-grocery-sales-forecasting/discussion/47560#latest-302253
4-й вариант решения:
https://www.kaggle.com/c/favorita-grocery-sales-forecasting/discussion/47529#latest-271077
План 5:
https://www.kaggle.com/c/favorita-grocery-sales-forecasting/discussion/47556#latest-270515
Шестой вариант решения:
https://www.kaggle.com/c/favorita-grocery-sales-forecasting/discussion/47575#latest-269568

Машинное обучение: https://book.douban.com/subject/26708119/

Методы статистического обучения: https://book.douban.com/subject/10590856/

«Инженерия признаков для машинного обучения» https://book.douban.com/subject/26826639/

Интервью с учеными в области данных: https://book.douban.com/subject/30129410/