Интеллектуальный анализ данных
Задание 1

Задача: Прогнозирование задержки рейса
Срок сдачи: 24-го октября 2024 года
Формат представления: ссылка на репозиторий GitHub и отчет (.pdf)
Данные: Набор данных

1. Описание задачи

В этом задании вы собираетесь решать задачу оценки задержки рейса с помощью машинного обучения. Целями являются:

- Предварительная обработка, визуализация и разделение набора данных
- Выберите 2 или более подходящих моделей машинного обучения для оценки задержек рейсов (например, линейная регрессия, полиномиальная регрессия и т.д.)
- Используйте по крайней мере 1 модель машинного обучения с регуляризацией для оценки задержки рейса.
- Сравните производительность выбранных моделей машинного обучения с использованием соответствующих оценочных показателей.
- Опишите, какая модель лучше подходит, исходя из производительности тестового и обучающего наборов. Перобучена эта модель? Недообучена?
- Обнаружение и удаление выбросов

2. Набор данных

Каждая запись в файле набора данных соответствует рейсу, и данные были записаны в течение 4 лет. Эти полеты описываются в соответствии с 5 переменными. Беглый обзор набора данных можно увидеть в таблице ниже:

Departure Airport	Scheduled departure time	Destination Airport	Scheduled arrival time	Delay (in minutes)
SVO	2015-10-27 09:50:00	JFK	2015-10-27 20:35:00	2.0
OTP	2015-10-27 14:15:00	SVO	2015-10-27 16:40:00	9.0
SVO	2015-10-27 17:10:00	MRV	2015-10-27 19:25:00	14.0
MXP	2015-10-27 16:55:00	SVO	2015-10-27 20:25:00	0.0
...	...		...	

Ниже приведены 5 переменных, описывающих каждый рейс:

Имя переменной	Описание
`Departure Airport`	Название аэропорта, из которого вылетел рейс. Название задается в виде международного кода аэропорта,
`Scheduled departure time`	Время, запланированное для вылета рейса из аэропорта отправления,
`Destination Airport`	Аэропорт назначения рейса. Название указывается в виде международного кода аэропорта
`Scheduled arrival time`	Время, запланированное для приземления рейса в аэропорту назначения,
`Delay (in minutes)`	Задержка рейса в минутах

3. Предварительная обработка и визуализация данных

Самый простой способ преобразовать строковое представление в машиночитаемый формат - заменить символы уникальным целочисленным идентификатором. Этого можно легко достичь с помощью `LabelEncoder` от sklearn. Вы вольны применять другие способы обработки категориальных строковых данных и пропущенных значений. **Используйте кодировщик по вашему выбору.**

Для разработки функций дополнительные функции могут быть извлечены из временных меток (например, год, месяц, число, день недели). Эти временные характеристики можно легко дополнить с помощью pandas `pandas.Series.dt`. Для визуализации данных на плоскости 2D могут быть использованы методы уменьшения размерности, такие как **PCA** . Самый простой способ - выбрать одну значимую функцию и сопоставить ее с целевой переменной `Delay (in minutes)` . Мы предлагаем рассчитать продолжительность рейса, которая представляет собой разницу во времени между вылетом и прибытием.

Данные следует разделить на обучающие и тестовые. Данные распределяются в зависимости от запланированного времени вылета. Тренировочные данные - это все данные за период с **2015** по **2017** год. Все образцы данных, собранные в **2018** году, должны быть использованы в качестве тестового набора.

4. Обнаружение и удаление выбросов

При подготовке наборов данных для моделей машинного обучения действительно важно обнаружить все выбросы и либо избавиться от них, либо проанализировать, чтобы понять, почему они вообще появились. В моделях машинного обучения (особенно в моделях с учителем) выбросы могут нарушить процесс обучения, что приводит к увеличению времени обучения или приводит к разработке менее точных моделей.

Выбросы нелегко распознать на этапе сбора данных, однако их можно обнаружить на этапе анализа. Существует несколько способов обнаружения выбросов (т.е. Визуализации данных и определения точек данных, отличающихся от большинства данных). Чтобы проверить, существуют ли выбросы в наборе данных о задержке рейса, возьмите данные за один месяц и примените выбранный вами метод обнаружения выбросов. Для получения дополнительной информации о различных подходах к определению выбросов смотрите раздел "Ссылки".

5. Модели машинного обучения

Для оценки времени задержки рейса вам нужно будет выбрать соответствующий алгоритм машинного обучения. Из курса вы изучили несколько алгоритмов машинного обучения (например, линейную регрессию, логистическую регрессию, полиномиальную регрессию и т.д.). Если вы решите использовать алгоритм, принимающий одну переменную в качестве входных данных, то `flight duration` следует использовать как независимую (предикторную) переменную. Для оценки задержки рейса следует использовать как минимум 3 алгоритма машинного обучения, которые должны использоваться. Один из алгоритмов должен иметь регуляризацию.

6. Измерение производительности

Для измерения производительности выбранных моделей существует ряд показателей, изучаемых в ходе машинного обучения (например, MSE, precision, recall, RMSE, F1, R^2 и взвешенная оценка F1).

7. Отчет и исходный код

После выполнения сравнения моделей машинного обучения результаты должны быть представлены в виде отчета. Реализация должна быть на python. Репозиторий реализации должен быть доступен на GitHub или GitLab. Ваш репозиторий должен содержать следующее:

1. Основной скрипт python (файлы jupyter notebook могут быть включены в отдельную папку репозитория)
2. Файл Readme (т.е. Как запустить основной скрипт)
3. Документация (документация по коду и Readme)

Ваш отчет должен содержать следующее:

1. Мотивация, объяснение того, чего читателю следует ожидать от вашего отчета
2. Краткое определение задачи и описание данных
3. Если вы используете альтернативный формат ввода данных, объясните это
4. Сравнение 3 выбранных моделей. Опишите, какая модель лучше, исходя из результатов теста и результатов тренировочного набора. Подходит ли модель? Не подходит?
5. Используйте графики и таблицы для документирования результатов ваших экспериментов

Отчет должен быть представлен в формате PDF. Ссылка для сдачи.

Ресурсы

- Способы обнаружения и устранения выбросов
- Как удалить выбросы для машинного обучения
- 5 способов обнаружения выбросов / аномалий, которые должен знать каждый специалист по обработке данных)
- Sklearn: обнаружение новизны и выбросов