Московский авиационный институт (национальный исследовательский университет)

Факультет информационных технологий и прикладной математики

Кафедра вычислительной математики и программирования

Лабораторная работа №1 по курсу «Искусственный интеллект»

Студент: А. С. Усков
Преподаватель: А. С. Халид
Группа: М8О-306Б

Дата:
Оценка:
Подпись:

Москва, 2020

Московский авиационный институт (национальный исследовательский университет)

Факультет информационных технологий и прикладной математики

Кафедра вычислительной математики и программирования

Лабораторная работа №1 по курсу «Искусственный интеллект»

Студент: А. С. Усков
Преподаватель: А. С. Халид
Группа: М8О-306Б

Дата:
Оценка:
Подпись:

Москва, 2020

Лабораторная работа №2

Задача:

Необходимо сформировать два набора данных для приложений машинного обучения. Первый датасет должен представлять из себя табличный набор данных для задачи классификации. Второй датасет должен быть отличен от первого, и может представлять из себя набор изображений, корпус документов, другой табличный датасет или датасет из соревнования Kaggle, предназначенный для решения интересующей вас задачи машинного обучения. Необходимо провести анализ обоих наборов данных, поставить решаемую вами задачу, определить признаки необходимые для решения задачи, в случае необходимости заняться генерацией новых признаков, устранением проблем в данных, визуализировать распределение и зависимость целевого признака от выбранных признаков. В отчете описать все проблемы, с которыми вы столкнулись и выбранные подходы к их решению.

Датасеты:

1. https://www.kaggle.com/jsphyg/weather-dataset-rattle-package
2. https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

1 Описание

Первый датасет ставит задачу бинарной классификации.

Первым с датасетом, бинаризовал категориальные факторы и привел бинарные факторы из строкового к флотовому типу. Замерил корреляцию, и подсчитал количество NaN значений на каждый фактор - оказалось, что у четырёх факторов примерно половина данных не заполнена. Можно было бы заменить эти пропуски медианами, но датасет посвящён погоде, а значит не исключены сезонные явления, а также показатели собирались в разных локациях, поэтому более логичным мне показалось обучить модель для заполнения данных. Флотовые фичи заполняются при помощи линейной регрессии, бинарная фича о том, был ли дождь в текущий день заполняется при помощи случайного леса. Модели обучаются на строках, в которых нет ни одного пробела, если при применении в фичах есть пробел, что фича заменяется на медиану. В первом заходе при заполнении фичей я использовал таргет, что, как оказалось, было не правильным шагом. Среди категориальных фичей также были пропуски, но семантически фичи с пропусками означали направление ветра, так что я посчитал, что пробел может оказаться информативен - выделить явное направление ветра не удалось. После заполнения пробелов я ещё раз подсчитал корреляцию факторов и таргета - фичи полученные бинаризацией категориальных факторов часто не являются высоко скоррелированными, но я допускаю, что они могут быть полезны. В качестве результатов я записал датасет со всеми фичами и датасет с фичами с корреляцией выше 0.1.

Второй датасет, как я сперва посчитал, ставит задачу мультиклассовой классификации, но более внимательно прочитав комментарий к датасету я понял, что на нём предлагается проводить бинарную классификацию. Более внимательно прочитать условия меня сподвиг тот факт, что некоторые классы представлены буквально десятком элементов и как я посчитал данных будет недостаточно. В этом датасете нет пробелов, поэтому я просто посмотрел на графики и коэффициенты корреляции взял наиболее скоррелированные фичи.