НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

ВЫСШАЯ ШКОЛА ЭКОНОМИКИ

Факультет математики

Ковалев Евгений, Новиков Лев, Сухарев Иван

Техническое задание по проекту

Quora Question Pairs

3 курс, майнор «Интеллектуальный анализ данных»,
группа ИАД-4

Москва, 2017 г.

1. Описание задания

В рамках финального проекта по майнору «Интеллектуальный анализ данных» мы решили поучаствовать в соревновании «Quora Question Pairs», которое проводится на Kaggle.

`https://www.kaggle.com/c/quora-question-pairs`

Quora — ресурс, устроенный по типу вопрос/ответ, на котором можно узнать интересующую информацию самой разной тематики. Задача заключается в том, чтобы в предоставленных парах вопросов выявлять содержащие в себе одинаковые по смыслу.

Обучающая выборка состоит из 404290 объектов (пар вопросов), 5 признаков (id — идентификатор пары, qid1 и qid2 — идентификаторы вопросов, question1 и question 2 — тексты вопросов) и целевой переменной is_duplicate, принимающей значение 1, если вопросы в паре одинаковы по смыслу, и 0 — иначе. Тестовая выборка состоит из 2345796 объектов и 3 признаков (test_id — идентификатор пары, question1 и question2 — тексты вопросов). Метрика качества — Log Loss.

Понятно, что признаки-идентификаторы почти не несут в себе никакой полезной в решении задачи информации, однако на основе текстов вопросов можно извлечь относительно много признаков, которые будут определять их близость по смыслу.

2. План решения

Для решения будет использоваться язык программирования Python.

1) ***Предобработка данных.***

   - Проверка на наличие пропусков: где-то может не быть идентификатора, или текст вопроса может быть пустым. Их заполнение в случае обнаружения.
   - Извлечение признаков на основе текстов вопросов: например, длины вопросов (с точки зрения слов и символов),

наличие общих слов, лемм, расстояние между вопросами как векторами (косинусная близость, расстояние Джаккарда, расстояния в метриках Евклида и Минковского). При извлечении некоторых признаков предполагается использовать токенизацию, стэмминг и лемматизацию, возможно, морфологический парсинг и методы исправления ошибок. Также планируется попробовать разные методы для представления вопросов в векторном виде (TF-IDF, word2vec, N-граммы).

- Масштабирование признаков.
- Используемые библиотеки: fuzzywuzzy, gensim, nltk, numpy, pandas, scipy, sklearn.

2) ***Визуализация.***

- Визуализация данных: распределения признаков (гистограммы, диаграммы размаха), корреляции признаков.
- Используемые библиотеки: matplotlib, seaborn.

3) ***Построение моделей.***

- На начальном этапе предполагается попробовать достаточно незамысловатые модели: логистическую регрессию, наивный байесовский классиификатор, случайный лес, метод опорных векторов, метод k ближайших соседей. Везде произвести подбор гиперпараметров с помощью кроссвалидации.
- Затем планируется перейти к (вероятно) более мощным моделям: к нейронным сетям, бустингу.
- Применить стэкинг и построить ансамбли алгоритмов.
- Используемые библиотеки: keras, pyLightGBM, sklearn, xgboost.

4) ***Анализ результатов и выводы.***

- Сравнение полученных результатов, интерпретация, выводы. Выявление наиболее важных признаков и лучших моделей.