Волков Антон

Алгоритмы обработки больших данных

Семестровый проект

Постановка задачи

- Используя информацию о сообщениях пользователей социальной сети Twitter и информацию о классах («0» или «1») для части из них, обучить регрессионную модель
- С помощью обученной модели предсказать, с какой вероятностью неразмеченные пользователи принадлежат классу «1»

Признаки

- В качестве признаков использовались частоты токенов в твитах пользователей
- По возможности, из твитов были удалены символы, не являющиеся буквами английского алфавита
- Слова твитов, не являвшиеся шумовыми, были приведены к нижнему регистру, к каждому была применена лемматизация
- Преобразованные слова и стали токенами

Подбор модели

- Исследовалось применение следующих моделей из пакета scikit-learn к обучающей выборке:
  - SGDRegressor - стохастический градиентный спуск
  - Ridge - линейная регрессия, накладывающая ограничения на величину значений весов
  - PassiveAggressiveRegressor - итеративный алгоритм, правящий веса в зависимости от того, насколько предсказанное значение целевой переменной было близко к реальному ее значению

Подбор модели (2)

- Метрика определения качества - площадь под кривой
- Использовалась информация об оценках по данной метрике с кросс-валидации на обучающей выборке
- SGD имел средний score 0.51, а на некоторых этапах кросс-валидации он был ниже 0.5, потому его было решено не использовать
- Лучший средний score для Ridge (с округлением): 0.591
- Лучший средний score для PA (с округлением): 0.595

Выбранные модели

- Kaggle позволяет выбрать два варианта submisson, которые будут использованы для определения финального места
- Было решено в качестве основных моделей взять PassiveAggressiveRegressor с числом итераций 5, с функцией потери, не чувствительной к epsilon, и его же в тандеме с Ridge