Курсовой проект по курсу дискретного анализа: классификация документов

Выполнил студент группы М8О-307Б-21 МАИ *Кажекин Денис*

Условие

На основе данных для обучения необходимо провести бинарную классификацию валидационных данных.

Метод решения

Классификацию можно провести из эвристики наиболее вероятного класса для конкретного документа. Для этого на основе данных обучения нужно посчитать условную вероятность от параметров (слова текста) для каждого из двух классов тестового документа и выбрать максимум (наиболее вероятную метку класса). Решение задачи сводится к реализации алгоритма наивного Байеса. В целях оптимизации также использовал логарифмизацию целевой вероятности, что позволяет перейти к сложению и вычитанию логарифмов.

Описание программы

Решение задачи сосредоточено в одном файле naivebayes.cpp

Дневник отладки

Не заметил, что ответ на каждый тестовый текст должен выводиться сразу, а не в конце, после обработки всего набора тестовых данных. Из-за этого были многочисленные ошибки долгого ожидания ответа от моей программы.

Тест производительности

Сложность данного алгоритма строго линейная, потому что вероятности считаются за константу на основе одного прохода по данным обучения

Недочёты

Недочеты не выявлены

Выводы

Очень интересная и прикладная задача. Алгоритм наивного Байеса, не смотря на свою простоту, часто используется в реальной жизни, потому что он за молниеносную скорость работы может выдавать хорошие предикты. Единственный минус – параметры, при которых считается вероятность должны быть

независимыми, то есть значение одного параметра не должно влиять на другой. Применение этого алгоритма может быть полезным, когда необходимо получать вероятности в режиме реального времени за короткие промежутки времени.