Прогнозирование пола и возраста по логу посещения сайтов

Евгений Сувитов, 2019

Ключевые моменты

1. Использовались только адреса страниц.
2. Модель выбиралась только по кросс-валидации.
3. Пол и возраст были объединены в один признак.

Получение признаков

1. Доменное имя извлекалось из адреса страницы с помощью регулярного выражения - `(?:http[s]?:\/\/)+(?:www\.)?([^\/]*)`
2. Выбраны **30000** самых популярных доменов по уникальным пользователям.
3. Посещение каждого сайта из предыдущего пункта использовалось как бинарный признак(1 - пользователь посетил сайт, 0 - не посетил)

Отбор признаков

SelectPercentile + **Chi2**. **Chi2** считает хи-квадрат статистику между каждым признаком и целевым классом, **Select Percentile** оставляет заданный процент признаков с максимальной статистикой.

Оптимальные значения найденные перебором - **73%** и **74%**

Моделирование

Для моделирования использовался **наивный байесовский классификатор (MultinomialNB).** Единственный гиперпараметр **alfa** - множитель перед слагаемым с регуляризацией. Оптимальное значение - **1.78**

Выбор лучших предсказаний

Для каждого классифицируемого объекта была подсчитана оценка уверенности - разница между первой и второй по величине вероятностями принадлежности. Выбирались **50%** предсказаний с наибольшей оценкой уверенности.

Результаты

Accuracy на валидационной выборке **0.3770**

Accuracy на финальной выборке **~0.3626**

Accuracy на кроссвалидации по 10 фолдам **~0.3157** - но без выбора 50% наилучших предсказаний.

Код

https://github.com/yanezh/bigdata-10-courses/blob/master/project1/