Отчёт об использованных методах

Пехтерев Денис Олегович, МСКМ-181

10 апреля 2019 г.

1 Описание наборов данных

Итак, имеем определённый набор данных, в данной работе использовался набор данных - отзывов на категории товаров, представленных на сайте Amazon в период с мая 1996 по июль 2014 года. Были взяты наборы данных по следующим категориям:

1. Отзывы на товары категории "Рецензии на книги"
2. Отзывы на товары категории "Здоровье и личная гигиена"
3. Отзывы на товары категории "Кино и Телевидение"
4. Отзывы на товары категории "Электроника"

Наборы данных были представлены в одинаковом формате json и состояли из следующих колонок:

- reviewerID – будет описание
- asin – будет описание
- reviewerName – будет описание
- helpful – будет описание
- reviewText – будет описание
- overall – будет описание
- summary – будет описание
- unixReviewTime – будет описание
- reviewTime – будет описание

Из этих столбцов для нашей задачи будут интересны только reviewText - сам отзыв, оставленный посетителем сайта и overall - оценка пользователя, которая ставится в соответствии с его отзывом.

Итак, обработав отзывы в формате json и переведя их в формат pandas - DataFrame, который является наиболее удобным для работы с наборами данных, мы получили 3 набора данных. Из-за того, что оперативная память на компьютере ограничена (8gb.), то были рассмотрены только часть отзывов - 1 млн. записей.

2 Обработка данных

Для начала, в данном упражнении, переведём оценки из формата 1-5 в бинарный (-1/1), для этого воспользуемся следующей логикой:

- при overall $\in$ (1,2) присвоим значение 0, что будет означать, что отзывы данной категории имеют отрицательную тональность;
- при overall $\in$ (3) - не будем присваивать никакого значения, так как для первой задачи мы будем рассматривать задачу бинарной классификации;
- при overall $\in$(4,5) - присвоим значение 1, так как оставляя отзывы с такими оценками, мы можем быть уверены, что пользователь был удовлетворён своей покупкой, а значит отзывимеет положительную тональность;

Имеем 4 набора данных, состоящих из отзыва пользователя и его оценки, в каждом по миллиону записей. Так как в задачах машинного обучения очень важно иметь сбалансированные выборки, то 1 миллион записей набирался по следующему алгоритму: выбирались 500000 положительных отзывов (оценка: 1) и 500000 отрицательных отзывов (оценка: -1).

2.1 Векторизация отзывов (метод «Мешка слов»)

Первым этапом при анализе тех или иных текстов методами машинного обучения, текстовые данные требуется перевести в числовой формат.

Для этого существует определённых набор методик, таких как:

- Метод "Мешок слов" (BoW, bag-of-words), является простым алгоритмом, который переводит набор данных в матричный формат, где каждая строка является отдельным предложением или отедельным отзывом, тогда как столбец отвечает за 1 слово в этом предложении/отзыве. Элементами же матрицы являются частоты с которыми данные слова встречаются в предложении/строки
- -будет описан: TF-IDF
- и ещё один

При построении числового представления отзывов (векторизации) я использовал самый проверенный метод !!!!! - метод "Мешок слов". Удалив из данных знаки препинания и другие лишние символы (отмечу, что они могут

быть исследованы при более глубоком анализе), я воспользовался библиотекой sklearn, а точнее CountVectorizer (внутри задаю binary = True, так как сначала хочется изучить зависимость выходного результата от вхождения определённых слов в предложение), в котором зашита логика метода Мешок слов.

2.2 Формальная постановка задачи классификации

Формальная постановка рассматриваемой задачи: X - множество объектов, в нашем случае это одномерный вектор, который представляет из себя текстовые отзывы покупателей о приобритённом товаре. $x_1, ..., x_l \subset X$ - обучающая выборка. $D_j = 0, 1$ - бинарный признак, который отвечает за тональность отзыва: положительный/отрицательный. $f_j : X \to D_j, j = 1, .., n$ - признаки объектов (от англ. features) Тогда векторное описание объекта имеет следующий вид:

$$F = ||f_j(x_i)||_{l*n} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \vdots & \ddots & \vdots \\ f_1(x_l) & \dots & f_n(x_l) \end{pmatrix}$$

2.3 Логистическая регрессия

Для решения данной задачи бинарной классификации, будем использовать логистическую регрессию, которая является одним из методов построения линейного классификатора и при этом даёт оценивать апостериорные вероятности принадлежности объектов классам. !тут вставлю картинку с тем, что из себя представляет!

В нашем случае имеем $Y = (-1, 1)$. (пересказ из источника про то, как работает метод с формулами и ост, нужен?)

—

Также будет добавлено про то, как выбирался (подбирался) показатель регуляризации для классификации.

2.4 Оценка точности классификации

Исследователями выделяется большое количество оценки точности классификации.

Самым популярным методом является метод Accuracy: суть заключается в том, что при исследовании точности модели мы смотрим на долю правильных ответов, которая может быть записана в следующем виде:

$$Accuracy\text{Rate} = \text{argmax}\ \frac{1}{l}\sum_{i=1}^{l}[y_o = y_i]$$

Данную формулу также принято записывать в следующем виде:

$$Accuracy\text{Rate} = \frac{\text{TP+TN}}{TP + FP + FN + TN}$$

В рассматриваемом примере, данный показатель будет вычисляться на основе сравнения результатов модели с заданными результатами последних 10000 отзывов в каждом наборе данных.

2.5 Результаты:

Результаты работы классификатора:

Название категории	AccuracyRate, %
"Рецензии на книги"	89.96%
"Здоровье и личная гигиена"	73.35%
"Кино и Телевидение"	88.61%
"Электроника"	87.54%

Здесь мы можем увидеть результаты работы классификатора, реализованного с использованием логистической регрессии. Интересно изучить слова, которые оказали наибольшее влияние на работу классификатора. Так, например, при анализе отзывов на категорию книги одним из важных факторов для модели при определении тональности стало наличие в предложении автора: Joseph Klausner. Скорее всего, это связано с тем, что под произведениями данного автора, представленными на сайта Amazon стояло больше всего положительных (как мы помним, оценки 4,5) отзывов, что свидетельствует о любви читателей к определённому автору в рамках рассматриваемого набора данных.

Далее выборки данных будут очищены, при этом будет произведена более тщательная обработка входных данных.