Рекомендательная система для ретейла: сравнение вероятностной и контентной моделей

Сарапулов Г. В.
Санкт-Петербургский государственный университет
Математико-механический факультет
g-eos@yandex.ru

14 февраля 2018 г.

Аннотация

В работе проведено сравнение двух подходов к построению рекомендательной системы для продуктового ретейла: на основе вероятностной модели (model-based) и на основе содержания (content-based). В рамках первого подхода построена вероятностная модель для оценки вероятности покупки товарных групп в зависимости от предыдущих покупок. Для реализации второго подхода построены векторные представления для товарных групп из ассортимента торговой сети и покупательских корзин.

I. Введение

Рекомендательные системы предназначены для предсказания того, какие объекты могут быть интересны пользователю. Их сферы применения обширны (новостные и мультимедийные сервисы, поисковые системы, e-commerce и т. д.), и на фоне последних достижений в этой области и роста вычислительных мощностей и накопленных данных в последнее десятилетие наблюдается рост интереса бизнеса к таким системам.

В работе рассмотрены подходы к построению рекомендательной системы для продуктового ретейла и проведена их сравнительная оценка на чековых данных одной из торговых сетей. В секции II рассмотрен подход составления рекомендаций на основе ранжирования товаров по вероятности покупки в зависимости от наличия других товаров в корзине покупателя, для чего использовался наивный байесовский классификатор. В секции III использован альтернативный подход, заключающийся в рекомендации товаров, похожих на приобретенные ранее. Для этой цели были построены векторные представления товаров и покупательских корзин, и список рекомендаций ранжировался по мере близости вектора товара к вектору-корзине. В секции IV приведена оценка моделей по оффлайн-метрикам (точность, покрытие).

В тексте приняты следующие обозначения:

- U - множество субъектов (users, покупатели)
- I - множество объектов (items, товары/товарные группы)
- R - матрица оценок размера $|U| \times |I|$ (например, $R[u, i] = 1$, если покупатель u купил товар i)
- x_u - вектор признаков субъекта u (демографические признаки, агрегационные данные)
- x_i - вектор признаков объекта i (характеристики товара)
- $f : U \times I \to \hat{R}$ - функция, сопоставляющая каждой паре (u, i) оценку $\hat{r}_{u,i}$
- $L(R, \hat{R})$ - функция потерь (например, кросс-энтропия или RMSE)

Задача: сформировать список рекомендаций для всех объектов $u \in U$ через нахождение функции f, которая минимизирует функцию потерь

$$f^* = argmin_f L(R, \hat{R}) \tag{1}$$

В качестве рекомендаций для каждого субъекта выбирается k объектов с наибольшими значениями $\hat{r}_{u,i}$

II. Рекомендательная система на основе наивного байесовского классификатора

Пусть $x_u = \{x_{u,1}, ..., x_{u,N}\}$ - вектор признаков покупателя u, построенный по истории транзакций, где $x_i^u = 1$, если покупатель u покупал товар i, и $x_i^u = 0$ в противном случае. Для каждого товара $i \in I$ обучается классификатор $f_i : X_u \to [0, 1]$, оценивающий вероятность покупки товара i в зависимости от предыдущих покупок, представленных вектором x_u:

$$P(y_i = 1|x_u) = \frac{P(y_i = 1) \cdot P(x_u|y_i = 1)}{P(x_u)} \tag{2}$$

Наиболее релевантным товаром является тот, вероятность покупки которого максимальна:

$$i^* = \arg\max_{i \in I \setminus I_u} f_i(x_u) \tag{3}$$

Для оценки вероятностей покупки товара i используем наивный байесовский классификатор:

$$P(y_i = 1|x_u) = \frac{P(y_i = 1) \cdot \prod_{j=1}^{N} P(x_{u,j}|y_i = 1)}{\prod_{j=1}^{N} P(x_{u,j})} \tag{4}$$

III. Рекомендательная система на основе векторных представлений товарных групп

Для каждой товарной группы $i \in I$ из ассортимента торговой сети находим векторное представление $q_i = \{q_{i,1}, ..., q_{i,K}\} \in \mathbb{R}^K$. В простейшем случае каждое слово представляется вектором q_i, в котором $q_{i,j} = 1$, где j - индекс слова i в словаре I, а остальные элементы вектора равны нулю (т.н. one-hot encoding). Более продинутым подходом к получению векторных представлений является Word2Vec[1], который и был использован в работе.

Из полученных векторных представлений товаров можно получить векторные представления покупательских корзин, например, через взвешенное среднее входящих в корзину векторов-товаров. В качестве весов берутся TF-IDF [2] веса:

$$\textit{TF-IDF}(t_k, d_j) = \frac{f_{k,j}}{\max_z f_{z,j}} \cdot \log \frac{N}{n_k} \tag{5}$$

где $f_{k,j}$ - частота встречания слова t_k в документе d_j, n_k - кол-в документов, где встречается слово t_k. Адаптируя этот подход для оценки весов товаров в корзине покупателя, примем за $f_{k,j}$ долю расходов на товар i_k в суммарных расходах покупателя u_j, за $\frac{n_k}{N}$ - долю товара i_k в обороте торговой сети. Веса дополнительно нормализуются:

$$w_{k,j} = \frac{\textit{TF-IDF}(t_k, d_j)}{\sqrt{\sum_s^{|T|} \textit{TF-IDF}(t_s, d_j)^2}} \tag{6}$$

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space // In Proceedings of Workshop at ICLR, 2013

[2] Salton, G.: Automatic Text Processing. Addison-Wesley (1989)

Таблица 1: *Точность рекомендательных систем (precision at k)*

	Precision at k	
Model	Precision at 1	Precision at 3
Naive Bayes	0.40	0.35
Item2Vec	0.37	0.32

Для поиска похожих товаров используется ранжирование по косинусной мере:

$$sim(d_i, d_j) = \frac{\sum_k w_{k,i} \cdot w_{k,j}}{\sqrt{\sum_k w_{k,i}^2 \cdot \sum_k w_{k,j}^2}} \tag{7}$$

IV. Результаты

Список литературы

[Mikolov et all, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. (2013) Efficient Estimation of Word Representations in Vector Space. *In Proceedings of Workshop at ICLR*

[Pennington et all] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for Word Representation. 2014