A Market Basket Recommender System

Angéllica C. Araujo[1]

[1]Rio de Janeiro - RJ, Rio de Janeiro Brazil

angellica.c.a@gmail.com

1. Domain Background

The task of identifying correlations in sequential user shopping data is called Market Basket Analysis. The most common application of this prediction task is the indication of the next items to be consumed by a user. Therefore, the purpose of discovering a user's frequent buying patterns is to obtain information about the user's buying behavior [Kaur and Kang 2016].

2. Problem Statement

Let $C = \{c^1, c^2, c^3, ..., c^n\}$ be a set of customers, $I = \{i^1, i^2, i^3, ..., i^n\}$ the set of itens and $E_c = \{E_c^1, E_c^2, E_c^3, ..., E_c^n\}$ the set of consumer events. A consumer event can be defined as $E_c^t, b \subset I, \forall c, t$ where for each set b of consumed items there is a transaction that contains one or more items, for every user in a time frame t. An algorithm should consider a customer's consumption event history E_c^t and predict which items are most likely to be consumed in addition to estimating the likelihood of interest in an item. Thus, we can consider our problem as a classification modeling task.

3. Datasets and Inputs

O dataset fornecido pelo Instacart disponibiliza uma séries de dados no formato csv. Para esse trabalho, vamos focar nos dados dos arquivos *order products* e *orders*.

The *order products* file specify which products were purchased in each order and the 'reordered' attribute indicates that the customer has a previous order that contains the product. We will use the *orders* file to select the set for training and testing in addition to building our baskets. [Kaggle 2017]. So, our work will receive as a input of orders the following data [Instacart 2017]:

orders (3.4m rows, 206k users):

- order_id: order identifier
- user_id: customer identifier
- eval_set: which evaluation set this order belongs in (prior, train or test)
- order_number: the order sequence number for this user (1 = first, n = nth)
- order_dow: the day of the week the order was placed on
- order_hour_of_day: the hour of the day the order was placed on
- days_since_prior: days since the last order, capped at 30 (with NAs for order_number = 1)

order products (prior, train and test) (30m+ rows):

- order_id: foreign key
- product_id: foreign key
- add_to_cart_order: rder in which each product was added to cart
- reordered: 1 if this product has been ordered by this user in the past, 0 otherwise

4. Solution Statement

A classification model will be applied for this work. Our intention is with a routine that produces as a result a vector of probabilities on the items contained in the set I. The main objective is to maximize the F-score metric, defined as a quality parameter in Kaggle's prediction competition and calculate $Pr(c$ being interested in i in a time frame $t_c + 1|E_c)$.

Once we have a formulated hypothesis we can make use of mathematical tools aimed at predicting ocurrence probabilities since we are dealing with a classification problem that works with prior knowledge to evaluate the probability of our hypothesis [Science 2018].

As input, our solution receives a consumption event E_c, defined by a sequence of chosen items. Based on this, we will test results of a Naive Bayes Classifier and a Random forest model to provide the next most likely item to be chosen by the consumer c, complementing the shopping basket like a recommender system [Science 2018].

5. Benchmark Model

Inspired on the Kaggle competition success parameters, as our first choice, this work will use the value of 0.4091449 as the ideal maximum based on Leaderboard panel in Kaggle. We will consider as acceptable a result in the interval between 0.4074450 and 0.4047891, referring to the fifty best results of the competition [Kaggle 2017].

As a second strategy, we will compare the answers obtained by the two classifiers. The objective here is to verify which of the two models will perform better, producing more accurate results. For example, we can structure our Naive Bayes classifier based on features that describe a rule of consumer's decision to buy the product i and then compare the results with the Random Forest model which works with a random selection of features based on prior occurrences [Science 2018].

6. Evaluation Metrics

We will use the F1 Score as proposed by the Kaggle challenge [Kaggle 2017]. Recall and precision fractions are considered in the calculation giving us a metric whose result is a balance between both fractions [Prentice-hall]. The general formula for F1 Score is:

$$F_\beta = (1 + \beta^2) \left(\frac{precision.recall}{(\beta^2.precision) + recall} \right)$$

Recall is a measure of proportionality analysis between positive outcomes and outcomes identified as positive. In this way, recall allows us to understand how complete the end result is. Precision allows us to analyze how accurate our result is, returning the ratio between the results identified as positive and all positive indefensions, including false positives [Prentice-hall].

7. Project Design

This work intends to reach the proposed objectives through the accomplishment of the sequence of steps described below:

1. Preprocess the set of orders to obtain the Baskets per customer
2. Train classification models considering the consumption events E_c of customer c
3. Train both models considering different levels of temporal granularity

4. For each model produce as output a vector of probability of consumption per item
5. Analyze the results obtained for each model with evaluation metrics

In preprocessing, baskets will be produced by collecting the information available in *orders* and *order products* csv files, in order to gather the items purchased by a consumer c into an array, thus characterizing the baskets of the consumer c and consumption events E_c. The temporal granularity can be selected on the values of the attributes *order dow*, *order hour of day*.

In the training phase, for the Naive Bayes Classifier, features considered relevant in a consumer c choice for a product i will be selected manually. The Random Forest model will play the role of choosing features automatically. Our intervention in the Random Forest model will be limited to the variation of parameters as the maximum number of features and estimators. Finally, the proposed evaluation metric will be applied to compare the performance of both classifiers results.

Reference

Instacart (2017). The instacart online grocery shopping dataset 2017 data descriptions.

Kaggle (2017). Instacart market basket analysis.

Kaur, M. and Kang, S. (2016). Market basket analysis: Identify the changing trends of market data using association rule mining. *Procedia Computer Science*, 85:78 – 85. International Conference on Computational Modelling and Security (CMS 2016).

Prentice-hall, M. Probability and statistics for engineers and scientists, by r.e. walpole, r.h. myers, and s.l.

Science, T. D. (2018). The random forest algorithm.