An Le

Problem Understanding

The problem is to help the Rainbow Store understand their customers' behavior through transaction history. Solving this problem will let the Rainbow Store be able to predict their customers' behavior in the future. From the type of data, this could be considered as a Market Basket analysis problem. It was first formulated by Agrawal et al. in *IEEE Transactions on Knowledge and Data Engineering* in 1995. Since then, there are many papers that offered improvements or alternatives to the original analysis method. This project aims to predict if any specific customer purchases from the store in a month window given his transaction history in the last two months.

Methodology/Approach

Since the features for our binary classification problem are not directly given, feature engineering is a very important part. I need to extract a set of features that minimizes the loss of information from the original data. Since I have no characteristic information about any of the products I have to rely on market basket analysis.

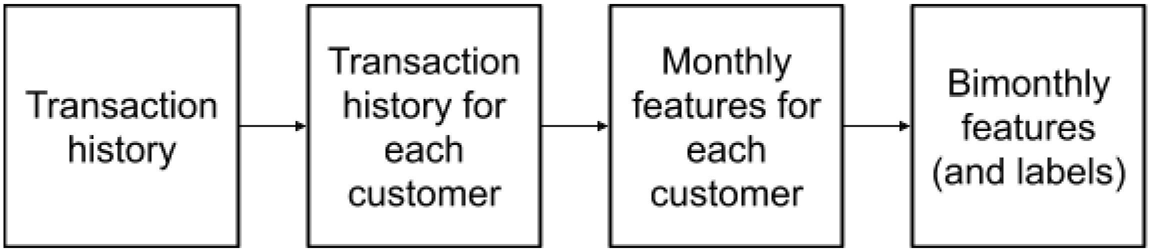
All market basket analysis methods aim to extract the relation between products in the market. In *Personalized Purchase Prediction of Market Baskets with Wasserstein-Based Sequence Matching*, 2019, Kraus and Feuerriegel proposed to represent products as vectors and the difference between two products is their vectors' cosine similarity. Because this method has been proven effective in Natural Language Processing, I decided to apply it to this project.

Combining these with other features (time differences between purchases, total price, total quantity), I build a binary classification model to predict the labels.

Workflow

Item2Vec.ipynb: The first step is to create a product embedding using Skip-gram Word2Vec. Ideally, the product vectors could be used as features, but due to limited computational power, I used OPTICS to cluster the products into groups. So occurrences of products in each group could be used as the first features.

DataPreprocessing.ipynb: Other features are extracted directly from the transaction history. Combined with group features, given data is transformed as followed:


Transaction history
Transaction history for each customer
Monthly features for each customer
Bimonthly features (and labels)

TrainingAndPrediction.ipynb: Build a LSTM model. Transformed data is splitted into train set, valid set and test set. Train and valid sets are used for training and tuning the model. Test set is then used to evaluate the model's performance. Also, the model is then used to produce the prediction for next month.

Answer on Discussion Problem

The list of customers recommended to send emails is in **Customers.txt**. This is the list of customers who have the predicted probabilities of buying from the store next month higher than 0.5. However, the cost of sending an email is very low or even none. So a lower threshold than 0.5 could be used to increase recall. In practice, many companies keep sending out email to everyone in their registered customers list until the customers unregister.

References

R. Agrawal, T. Imilienski, and A. Swami. Database Mining: A Performance Perspective. IEEE Transactions on Knowledge and Data Engineering 9, 5(6):914-925, December 1993.

Kraus, Mathias, and Stefan Feuerriegel. “Personalized Purchase Prediction of Market Baskets with Wasserstein-Based Sequence Matching.” Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, July 2019, pp. 2643–52. arXiv.org, doi:10.1145/3292500.3330791.