Machine Learning Engineer Nanodegree

Capstone Proposal

Instacart Market Basket Analysis
Which products will an Instacart consumer purchase again?

Fu-Chang Sun
May 21st, 2017

Proposal

Domain Background

Have you ever go to supermarket with the thoroughly planned grocery list and eventually realize you forget to buy something on the way going back home? Based on the purchase history, there should have a pattern of what we buy and how ofen we buy. However, it is hard to remember that whether your daughter's favorite cereal is running out or you only have one last roll of toilet paper in the bathroom. What if there is a housekeeper tracking all you need and make an order for you? We could save time and extra trouble on going back store to get that missing item.

Furthermore, buyer might "need" something that they didn't think of while making grocery list. What if the housekeeper provide the expert suggestion while we fill up our shopping cart? For instance, if someone have tomoato and pasta in the shopping list, the system can recommend buyer a package of italian suasage, a thin loaf of French bread, or a good bottle of white wine. There is an old saying, "customers don't know what they want". With what machine learning techniques can offer, the shopping experience might become easy and delightful. The competition is proposed by Instacart very recently on kaggle.

Problem Statement

I would like to use the purchase history to predict the items that buyer is going to buy next time. Moreover, I would also like to provide some "wisdom" suggestion on what the buyer might want to buy. The first part is related to forecasting problem, and second part is associated with the recommendation system.

Datasets and Inputs

The dataset is provided through Instacart from kaggle competition. There are 7 csv files, which are aisles, departments, order_products_prior, order_products_train, orders, products, and sample_submission. These files are associated with the customers' orders over time based on 3 million grocery orders from more than 200,000 Instacart users. For each user, there are about 4 to 100 orders, with the sequence of products. The week and hour making order is also provided. Our goal is to predict which products will be in a user's next order. These files are self-explanatory as

follows,

aisles.csv

- 1,prepared soups salads
- 2,specialty cheeses
- 3,energy granola bars
- ...

departments.csv

- 1,frozen
- 2,other
- 3,bakery
- ...

order_products__*.csv

- 1,49302,1,1
- 1,11109,2,1
- 1,10246,3,0
- ...

prodocts.csv

- 1,Chocolate Sandwich Cookies,61,19
- 2,All-Seasons Salt,104,13
- 3,Robust Golden Unsweetened Oolong Tea,94,7
- ...

sample_submission.csv

- 17,39276
- 34,39276
- 137,39276
- ...

The example of the order is provided in the blog.

Solution Statement

Our goal is using customers' orders over time to predict which previous purchased products will be in a user's next order. From "order_products_*.csv" file, the label "reordered" indicates 0(nope) or 1(yep), which is a classification problem. Because there are various products (49688 items), the dimensionality of the problem might explode if direct one-hot encoding is implemented. Due to high dimensionality of the dataset, Gaussian Naive Bayes and Ensemble methods are good candidates for such problem. I plan to implement multiple supervised learning models and discuss its applicability/performance.

Furthermore, using machine learning techniques can provide personalized product recommendations to increae the size of the shopping cart. I will apply collaborative filtering and/or content-based filtering approach in the recommendation engine and discuss the findings.

Benchmark Model

As a naive guess, I would choose the benchmark model such that the buyer always buys the products that he has purchased in previous order. The reordered lable is always 1, namely buy it again. It is a reasonable assumption but of course won't be the case in reality.

Evaluation Metrics

In order to test our benchmark model, I would evaluate model accuracy and F-score. The performance metrics can be split into two part, training and predicting. The execution time is recorded to compare its scalability. By viewing its performance among different methods, the optimal approach could be proposed for not only this, but also similar problem. Moreover, scoring parameter could be evaluated for fine tune the estimator.

```
from IPython.core.display import HTML
HTML('<img src="wf.png", style="margin:-240px 0px -240px -80px">')
```

Project Design

For this project, we are going to follow the workflow like every machine learning problem does.

Evaluate problem:

Based on the description of instacart kaggle competition, our final goal is using previous purchase history to predict future order. The information of the product, such as aisle number, department id, product name, order time is provided from 3 million grocery orders. Since the outcome is evaluated as reorder or not, the machine learning technique for classification could be implemented. Furthermore, one can provide the recommendation for the users who have similar shopping behavior using collaborative filtering approach.

Data exploration and preparation:

Seven separate data files are provided with unique id. The products are categorized in different aisle and department. Each transaction is reported with detailed description of purchasing time and time gap of previous order. The statistical description could offer a better understanding of dataset, such as the number of category,

Model development and training:

I would compare the preliminary results between different supervised algorithms and discuss the findings. The data should be carefully shuffled and randomly split into trainning and testing subset. There are quite a few classfication methods available, but it is too early to give solid comment on the best among them. Therefore, I would "wisely" pick the stardard method such as logistic regression and support vector mechines (SVM) as a starting point. More advanced approaches, such as Gaussian Naive bayes and ensemble methods are considered. The general and in-depth discussion could provide the complete picture of the problem, as long as clarify the subtlety of our methodology.

Model testing and deployment:

The performance of each approaches is evaluated through computational cost (execution time), accuracy, and F-score. The required execution time from smaller dataset can give us the idea of whether the implemented approach can be scaled up to the larger problem. The evalution metric of accuracy and F-score is utilized in order to identify the best model. The parameters can be further tuned, using grid search approach to optimize our model.

I would conclude the study by data exploration, model evaluation, and result visualization. The critical factor of solving this problem will be revealed in the final report. Moreoever, further investigation and/or future work, for generalization of the similar problem, will be discussed as well.

Reference

- https://www.kaggle.com/c/instacart-market-basket-analysis/leaderboard
- http://machinelearningmastery.com/make-predictions-time-series-forecasting-python/
- https://cloud.google.com/ml-engine/docs/concepts/ml-solutions-overview