On Building an Efficient Recommendation System

Final Project Proposal for CSCI 8360

Team tooYoungTooSimple

Xiaodong Jiang, Yang Song, and Yaotong Cai

Nov. 1, 2016

Department of Statistics, UGA

1. Project Description:

This project is given by Santander Group as a challenge in Kaggle competition with total prize of $60,000. The question to be answered for this project is **to construct an efficient product recommendation system**. As a large retail banking company, Santander Bank offers various financial services to their customers through personalized product recommendations. By their recommendation system, a small number of Santander's customers receive many recommendations while many others rarely see any resulting in an uneven customer experience. In this competition, we are challenged to **predict which products the existing customers will use in the next month** based on their past behavior and that of similar customers. With a more effective recommendation system in place, Santander can better meet the individual needs of all customers and ensure their satisfaction no matter where they are in life.

2. Data Description

We are provided with 1.5 years of various customer behavior data from Santander Bank to predict what new products customers will purchase. The data starts at 2015-01-28 and contains monthly records of products a customer has, such as "credit card", "savings account", etc. We need to predict **additional single/multiple products** a customer will get in the "last month", 2016-06-28, in addition to what they already have at 2016-05-28.

Two files of data are provided: train.csv to be the training set and test.csv to be the testing set. They are split by time. In both data sets, column #1 to column #24 contain information such as account types, status, lengths of period as well as customers' demographic information; column #25 to column #48 in the training set are the products available to customers with 1 indicating "purchased" and 0 indicating "Not purchased". The training set has 13,647,309 observations and the testing set has 929,615 observations. The Descriptions of each column can be found at:
https://www.kaggle.com/c/santander-product-recommendation/data.

3. Proposed methodology for analyzing

To answer this question, a number of machine learning algorithms are worth to be considered, such as filtering-based approaches, various classification algorithms, rule learning methods, and neural networks. Limited by the time of this final project, we cannot try all of above algorithms, but three main approaches will be implemented.

The first algorithm we will try is neighborhood model, which is a first-and-must-try algorithm in recommendation systems. The standard approach is to take some similarity metric such as linear or nonlinear correlation measure to define similarities between pairs of financial products, take the K most similar ones under this metric (where K is perhaps chosen via cross-validation), and then use the same similarity metric when computing the weighted mean. Another approach, instead of using the similarity metric to define the interpolation weights in the mean calculations, we can perform a (sparse) linear regression to find the weights that minimize the squared error between a product's "rating" and a linear combination of the "ratings" of its neighbors.

A second algorithm we will try is matrix factorization. In its basic form, matrix factorization characterizes both products and customers by vectors of factors inferred from product "rating" patterns. High correspondence leads to a recommendation. A simple matrix factorization models map both customers and products to a joint latent factor space, such that customer-product interactions are modeled as inner products in that space. We may review more advanced and cutting-edge matrix factorization algorithm to get a better performance in this project.

The last approach we consider is Restricted Boltzmann Machines, which helped to win the famous Netflix Prize. We will follow and implement this algorithm based on the paper by Salakhutdinov, Mnih and Hinton, they have showed how a class of two-layer undirected graphical models, called Restricted Boltzmann Machines (RBM's), can be used to model tabular data, such as users' rating of movies.

Based on the above three approaches, it is worthwhile to use ensemble methods for us to combine the models and produce improved prediction results.