DATA-643 Final Project Plan

Mohamed Elmoudni
Shazia Khan
Senthil Dhanapal

Contents

Introduction

Recommendation systems are composed of filtering algorithms that aim to predict a rating a user would assign to a given item. Recommender systems have become increasingly important across many platforms such as movies (Netflix), restaurants (Yelp), friends (Facebook and Twitter), and music (Pandora and Spotify).

Our final project will be about recommending books based on user ratings. The dataset is called Book-Crossings. The dataset is a book ratings dataset compiled by Cai-Nicolas Ziegler based on data from bookcrossing.com. It contains 1.1 million ratings of 270,000 books by 90,000 users. The ratings are on a scale from 1 to 10.

Our project plan will be based on the below high level steps:

1- Data acquisition
2- Data Exploration and Preparation
3- Model creation
4- Model Selection
5- Prediction
6- Model tuning

Purpose

The goal of the project is to build a comprehensive recommendation system for books based on user ratings. Using the User information, We plan to include user's location and age provided in the dataset to improve our recommendations.

Below is detailed highlight for the each section in our project plan:

Data acquisition

The dataset is a snapshot of www.BookCrossing.com. The motto of the site is "If you love your books, let them go!". The members of this website are located all over the world. A member registers their book and labels it and receives a book ID. The book is then shared with other members privately or publicly. The member can trace the book from one member to another and from one location to another as it travels around the world. The site mentions that at a given point they had 850,000 active users with seven million books which are traveling around 130 countries!

a. `Data source Identification`

b. `Data collection and storing`

c. `Data merging`

Data Exploration and Preparation

Looking at the data, there will be alot of cleaning and tidying to do before we can use it for creating to create models for recommender systems. The BX-Books.csv and BX-Ratings.csv files are properly delimited by semicolon and text is qualified by double quotes but the BX-Users.csv file does not seem to be in this format.The cell content is split into multiple columns as seen in the csv Excel format. We will have to make sure that all the users in ratings file are in users file just as we will check for books in ratings file are listed in book file.

a. `Variable Identification`

b. `Univariate Analysis`

c. `Bi-variate Analysis`

d. `Missing values treatment`

e. `Outlier treatment`

Model creation

These systems generally produce recommendations via one of two methods: 1) content based filtering or 2) collaborative based filtering. Content based filtering techniques use attributes of an item in order to recommend future items with similar attributes. Collaborative filtering builds a model from a user's past behavior, activities, or preferences and makes recommendations to the user based upon similarities to other users.

a. `Model identification`

b. `Model building`

c. `Model  creation`

Model Selection

The ultimate goal is for the model is to perform well on future unknown data based on the results from training data. We need to consider whether the model is overfitting or under fitting and how well the model fits the future or test data. A number of approaches are recommended for model selection such as Akaike information criteria (AIC) and Bayseian information criteria (BIC) and Adjucted R-Squared.

a. `Model comparison`

b. `Model selection`

Prediction

a. `Test data`

b. `Prediction on test/unseen data`

Conlusion