Capstone Project 1: Statistical Analysis

Springboard 2020

Joseph Tran

The goal of this project is to become familiar with recommender systems, and the tools used to build them, focusing on model based collaborative filtering. This differs somewhat from a supervised model where a set of features, X, is used to predict a value for a target or class variable. Typically, statistical analysis is performed on the predictor variables to look for relationships, correlations, or test hypotheses one might believe about certain features. This project is different in the sense that to perform model based collaborative filtering only the user/item matrix is required. Being that the only input for the model is user/item interaction data, we can plot the distributions of the count of user events and track plays.

First, we create a column of ones to aggregate in a group-by operation, groupby by user id and track id separately and summing the count column. We now have two data frames representing each user id and how many plays they had, and track id's with play counts.

The initial distribution plot of users and listening events shows a highly right skewed distribution. This is due to a few users who had an unusually high amount of listening events. To reduce the skewness without eliminating these users with high listen counts we can perform a logarithmic transformation. After the transformation we get a clearer view of the data which appears to be exponentially distributed. We can also filter and keep only the users with at most 200 plays. This removes the users with the unusually high song plays, and accounts for 94% of the users in the dataset. What we see after plotting is a distribution that looks similar to the distribution generated from log-transform of the counts. Finally we can plot a cumulative distribution for the filtered data and see that users with between 10 and 15 plays account for roughly 25 percent of the filtered data.

We also plot the distribution of play per track. We can see in the first plot that again the data is right skewed. Performing the log transform yields a distribution similar to the user play counts showing a distribution that looks exponential. Again, we can filter the outliers by removing the track id's with play counts over 250. By filtering songs with 250 or less plays, we retain 94% of the data. Doing this yields a distribution similar to that of the log-transformed data. As we can see, there were about 16000 songs with between 15 and 20 plays. Lastly, the cumulative distribution shows that about 80% of the songs in the filtered dataset have 50 or less plays..