Christian Zinck
12/13/18

Project 3 Report

1. For my recommendation system, I decided to implement my own learning algorithm. I first constructed a similarity matrix for which each user pair combination was given a score based on how similar their known ratings are. When predicting what a user would rate a movie, I took the ratings given by the n users most similar to the user in question and performed a weighted average on them, based on the similarity, gender, and release year preference. To make recommendations, I find the 5 movies with the highest product between predicted rating and a frequency coefficient.

The hyper parameter for this algorithm is the number of similar users n which are used to calculate the predicted rating. The range of possible values is [1, 943], and ran tests with many values within this range. Low values of n do not provide enough ratings for each movie to make a decent prediction, while high values of n include ratings from users with poor similarity to the user. I have found the optimal value to be around n = 100. Using this value of n, I produce MAE = 0.811614196618 and average rating = 3.00228679294.

2. After conducting feature analysis, I have found that through 10 fold cross validation of the training set, a logistic regression model was able to predict the gender of the user with 0.170212765957% error the year the movie was released with MSE = 137.522903034, while a naive approach produced MSE = 203.192742415. These results are significant enough to conclude that the movie rating matrix contains information about both the gender of its users and the release date of its movies.

Intuitively, this makes sense, because some movies genres like action are generally more highly appreciated by men while other genres like drama are generally more highly appreciated by women. Also, people tend to prefer movies that are released in a similar time window. For example, my grandfather loved the movies Citizen Kane and It's a Wonderful Life, which were released in 1941 and 1946 respectively, and my nephew loves Finding Dory and Cars 3, which were released in 2016 and 2017 respectively. These facts are what I believe gives rise to the predictiveness of the movie rating matrix.

3. To incorporate gender into the model, I scaled the similarity score by 0.5 when the users were of different genders and did nothing when they were the same gender. To incorporate release year preference into the model, I calculated a weighted average based on known ratings to find the year preference of each user. I then calculated a coefficient between 0 and 1 to represent how close a movie is to the users year preference and multiplied that coefficient into the ratings weighted average