Thomas Klimek
NickName: TK
Project 3 Report
12/13/2018

1. Learning Algorithms and Parameter Settings

For collaborative filtering I used singular value decomposition, or SVD for short, to find vector representations for users and items from the incomplete matrix of ratings. I imported the Python Surprise package, and used their module for SVD as well as their K-fold and GridSearch modules for hyper parameter tuning and selection. My procedure for model selection includes iterating over a list of hyper parameters using grid search and optimizing for the best MAE score. I iterated over hyper parameters of `{'reg_all': [0.2, 0.4, 0.6], 'n_factors': [1, 2, 5, 10]}`. In Surprise, reg_all corresponds with lambda and n_factors corresponds with K in the given training objective for SVD. The hyperparameters selected for best performance are `{'reg_all': 0.2, 'n_factors': 10}` for scores of

```
MAE:  0.7334
RMSE: 0.9351
```

For Vector Analysis, I predict genders using scikit learn SVC module which is an implementation of support vector classification. For model selection I tune over parameters `{'C':[1,10,100,1000],'gamma':[1,0.1,0.001,0.0001],'kernel':['linear', 'rbf']}` using scikit learn GridSearchCV. I use this model with parameters `{'C': 1, 'gamma': 1, 'kernel': 'linear'}` and get results of:

```
SVM Classifier Lowest CV Error:  0.708994708994709
```

For predicting release years I use scikit learn linear regression. I compare the MSE error of the linear regression model to a naive method that takes the mean of all the years and uses this as its input vector. I generate results of

```
linear regression MSE:  135.41828240249131
naive method MSE:  203.04368486931486
```

2. Result Analysis

For collaborative filtering, the results vary a bit from what I expected. The average rating of the top K=5 recommendations generated is only 2.75. We would expect our recommendations to generate higher ratings, because as mentioned in the assignment a rating of 2 means the user didn't like the item very much. This is most likely due to the fact that a large amount of user and item pairs where not found in the test set, in which case we replaced the rating with a 2. This probably skewed the average rating of the recommended items, however because our rating is higher than 2 we can see that in the cases that we did find the user item pair in the test set it had a higher rating.

For gender prediction, we did produce a high accuracy of 70%, however looking at the distribution there are about 70% of genders labeled 1. This means if our model where to always predict the label 1 it would achieve similar accuracy to our model, however I did suspect that it would be difficult to use these features to predict gender. For predicting release years our models did behave as expected. We see the MSE of the linear regression model is significantly lower than the naive approach, signifying that our linear regression was successful in learning a trend in the data.

3. Incorporate User & Movie Information in Recommendation Model

To incorporate further user and movie information in a recommendation system, you could create a multi-dimensional system for additional information as opposed to the 2 dimensional user-item model. This additional information could include facts about the user such as different personality traits, as well as additional facts about the movie and even temporal information such as day of the week, time, and location. One such approach is proposed in the research paper *Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach* [1]. This paper suggests a reduction-based approach that segments the ratings based on user-specified criteria and then applies collaborative filtering or other two-dimensional rating estimation methods to the resulting two-dimensional segment. They demonstrate using empirical results that this further contextual information can improve recommendation systems. To develop sophisticated recommendation systems it is important to consider a multi dimensional approach in order to generate better recommendations.