Stat 535 Final Project Proposal

Meilan Chen
Jinchao Feng
Bing Miu
Rong Zhang

We plan to extend the Movie Recommendation Lab as our final project. The goal of the project is to use past Netflix movie ratings to predict future movie ratings. This project allows us to design and implement statistical analysis on a large data set. It also gives us an opportunity to explore Python packages such as scikit-learn, pandas, etc., and use them to solve a real-world problem.

We want to use the linear regression model with feature selection as a baseline and plan to extend the lab by refining the linear model with statistical techniques and testing additional models such as Ridge and Lasso Regression, Singular Value Decomposition (SVD), and K-Nearest Neighbors (KNN) Algorithm. And we will train these models with K-fold cross-validation on the training set to pick the best parameters of the testing models.

We plan to compare the four models and expect SVD to give the best predicting result, because it produces the best low-rank linear approximation of the original matrix. To support our claim, we will split the data set into training set and validation set, then rank the models using the Root-Mean-Square Error (RMSE) of validation set with respect to each model. We expect SVD model to have the smallest RMSE among all the above models and thus have the best prediction accuracy. We will continue to refine the models and improve the accuracy of our movie recommendation system.