Group Proposal - Data Mining Final Project

Aaron Gauthier, Pedro Uria, Zachary Buckley

The problem we have selected is to predict Netflix user ratings for movies based on their previous movie ratings and those of other users. We have selected this because we are aware that Netflix has provided anonymous customer ratings previously, and believe we can expand on a competition dataset that was provided between 2006 and 2009 to potentially get good prediction results using a combination of the data-mining techniques we've been learning in class.

Kaggle.com has the dataset Netflix provided for a competition aiming to improve their recommendation engine by at least 10% (https://www.kaggle.com/netflix-inc/netflix-prize-data). The data provided by Netflix does not seem to have a feature number problem, as once the data is combined into a single table, the vast majority of the data revolves around only 4 features (movie_id, user_id, rating, date). We will be doing data integration, pulling in metadata about specific movies from additional sources (currently looking at imdb and similar sites). It is likely, based on prior experience with genre information, that getting genre features into a usable form will require some cardinality reduction, particularly by 'combining two of more categories into one', as mentioned in the text.

For Instance Selection on the existing Netflix data, we plan to use Cluster Sampling (p. 157) based on movie genre to start with, and may expand to other techniques like Data Clustering (p. 159) as our analysis progresses (potentially K-means on the weighted features). After clustering the users, we will use these subsets of the data to train supervised learning models in order to predict user rating. A different model would be trained for each cluster.

We'll be using the python libraries scikit-learn, NumPy, and pandas to implement our analysis, and preprocessing code. We will base the performance of our results on comparing our predictions with the provided test dataset from the kaggle site, using RMSE (which was used for the netflix prize competition).

Rough Project Schedule follows on the next page:

Date	Milestone	Description
3/31/19	Proposal Draft	Complete Draft Proposal and Topic Selection
4/2/2019	Finalize Topic Choice	Discuss topic with Amir, and Priyanka
4/6/2019	Group Proposal Due	Due Date for Group Proposal Submission
4/7/2019	Integration Complete	Finish code for loading in netflix and imdb data
4/14/2019	Analysis Complete	Complete Data preprocessing and analysis
4/20/2019	Group Paper/Presentation	Complete Group Paper and Presentation
4/21/2019	Individual Papers/Cleanup Project Complete (Submitted)	Complete individual papers/general cleanup
4/23/2019	Final Due Date	Due Date for Group Final Report and Presentation Submission

1) Problem selected and why you selected it
2) database/dataset will you use
3) Data mining algorithm will you use, standard forms, or customized
4) Software used to implement the network (???) why?
5) What Reference material to obtain sufficient background on applying the chosen network to the problem selected?
6) How to judge the performance of your results? What metrics will be used?
7) Rough schedule for completing the project.

TODOs?
Bibliography… rather than just inline? - I think inline is fine.
Feature Selection (movie_id, movie_name, release_date, move_genre, movie_gross, avg_rotten_tomatoe_rating, ….., user_id, user_rating, rating_date)... not sure of specific technique.
Improve by 10% what… hit rate? - RMSE
Possible Feature Extractions:

- delta between avg. imdb/rottentomatoes rating and customer rating (likely applied after normalization/scaling)

Original:
From 2006 to 2009, Netflix held a competition with a reward of 1 million dollars for helping to improve on their movie recommendation engine. Our plan is to start with data provided for that competition on kaggle (https://www.kaggle.com/netflix-inc/netflix-prize-data), and add additional information regarding the movies into the equation (which does break the original competition rules, but doesn't impact the real-world feasibility of applying the predictions). Utilizing the additional information we'll attempt to predict customers ratings of movies they haven't seen yet, based on their previous ratings of other movies and the ratings of other users who have liked similar titles, and the additional information about those movies. We expect finding an answer to these problems will require the use of feature selection techniques for down-selecting the amount of information available about movies to the information relevant to our predictions. We'll likely use instance selection techniques to reduce the number of samples in the data set, as the netflix dataset contains data from 480,189 users, and their rating of up to 17,770 movies. After what will likely prove to be a generous amount of preprocessing, we expect to utilize clustering models to group customers with similar interests, and classification models to project those groups rating of various movies.

Not sure we should include:
Experiment with 'genre-bias'? Some people really love sci-fi, others really love drama… some would love the intersection of the two… some would not. Could identify these biases and address them somehow? Apply mutiple clustering filtered by other variables on the movies (like genre)... giving us customers preference grouping… within a specific genre for example.
IE: within sci-fi likes star wars, hates star trek… or something similar… but what a wierdo right?

normalization/scaling of gross movie moneys? (seems to be getting into the weeds for a proposal)

Papers by past winners:
https://www.netflixprize.com/assets/ProgressPrize2007_KorBell.pdf
https://www.netflixprize.com/assets/ProgressPrize2008_BigChaos.pdf
https://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf

Paper about data being combined with imdb leading to loss of anonymity:
https://arxiv.org/abs/cs/0610105 (haven't attempted to download it yet)

Skimming these… we probably aren't going to improve on them… it appears that they're using very complicated forms of SVD in the final GrandPrize version.