COSC 174, Winter 2015
Machine Learning and Statistical Data Analysis
Brian Doolittle, Pratap Luitel
Project Proposal

Problem: The goal of the project is to predict how a user will rate a song given their demographic, previous ratings, music preferences, and description of the artist. This project was originally posted on kaggle.com as 24 hour hackathon taking place on July 21st, 2012 [1].

Methods: Common algorithms that competitors applied to this problem include gradient boosting, stochastic gradient descent, and random forests [2]. These algorithms can be applied to a wide array of machine learning problems and are described in detail by Murphy [3]. While these methods have proven to be successful, we will also apply relevant techniques that we learn in class.

Data Sets: There are three data sets relating to the project on Kaggle, users.csv, words.csv, and train.csv. The users.csv data set contains 48,645 samples of personalized user demographics such as gender, age, importance of music in user's life, hours spent each day listening to music, etc. The words.csv data set contains 118,301 samples that quantify how much a user likes an artist, if they own the artist's music, and words they use to describe the artist. The train.csv data set contains 118,690 samples of user track ratings and the time of year that the track was rated. We will use a subset of train.csv data as the test data [4]. Our goal is to predict a user's rating using the three data sets provided.

Goals for Milestone:

- Preprocess Data- Some of the .csv files are messy and have missing elements. We will go through the files and clean them up so they are easy to work with.

- Research Methods- We will complete thorough research into the three algorithms mentioned in the methods section as well as any other promising techniques.

- Implement an Algorithm- We will test one of our researched algorithms.

References

1."EMI Music Data Science Hackathon - July 21st - 24 Hours." *Description* -. N.p., n.d. Web. 20 Jan. 2015. <http://www.kaggle.com/c/MusicHackathon>

2."EMI Music Data Science Hackathon - July 21st - 24 Hours." *Kaggle*. EMI, 21 July 2012. Web. 20 Jan. 2015. <http://www.kaggle.com/c/MusicHackathon/forums/t/2242/code-approach-sharing>

3. Murphy, Kevin P. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT, 2012. Print.

4. "EMI Music Data Science Hackathon - July 21st - 24 Hours." *Data* -. N.p., n.d. Web. 20 Jan. 2015. <https://www.kaggle.com/c/MusicHackathon/data>.