Sujeet Yeramareddy
CSE 158
Kaggle username: syeramar

Assignment #1 Report

Play Prediction Task

For this task, we were asked to predict whether a random given user would play a given game. This is a supervised classification problem because we are trying to classify whether the user would (1) or wouldn't (0) play the game. I used logistic regression to make a model that can accurately predict this given various features. When first exploring the training set, I noticed the distribution of 'hours_transformed' was skewed right, and there were some really high values, and decided to remove all 'hours_transformed' that was greater than 8.5, which ended up removing ~ 5% of our data, which is reasonable considering we have well over 150,000 data observations. Most importantly, all our training data consists of positive (1) classifications, and wouldn't make for a balanced model, therefore for every positive user, I created an observation of a game they didn't play, doubling my training set size. For my features I used Jaccard Similarity, Popularity threshold (totalPlayed/1.5) with average hours the user plays, average hours spent on the game, and how often the game is offered for free. For the Jaccard similarity, I did a user-user comparison by comparing how similar the users of the given game are with the users of all games played by the given user. In my feature vectors, I added whether the maximum Jaccard Similarity was greater than 0.021, optimizing BER.

Category Prediction Task

For this task, we were asked to predict the 'genreID' based on the textual features of a given review. This is a supervised multi-classification problem because we are using textual features to classify the review to a specific genre of games. I also used logistic regression for this problem using a bag of words model to create my features. I used the 3500 most occurring words out of all reviews, making the matrix of size (nObs, 3501), including the constant. Additionally, I added the length of the review in numWords making each feature vector of length 3502. The max_iter parameter had to be increased as we have lot of data, and our model performed better when our regularizer constant (C) was 10. I tried using the TF-IDF approach, by using ~3500 of the highest TF-IDF values in the dataset, however this reported a lower accuracy so I chose to use my bag of words model.