MIT 15.077: Statistical Learning and Data Mining **Spring 2015**

Project Proposal

Student: Geoffrey Gunow

Description of Problem In this project, I will attempt to form an optimal strategy for ESPN's College Football Bowl Pick 'em game. The objective of this game is to correctly predict outcomes of college football games. Each player attempts to predict the winner of all N college football bowl games. In addition, the player assigns confidence points to each game, which range from 1 to N and each can only be used once. If the player correctly predicts the outcome, they receive the confidence points assigned to the game. If they incorrectly predict the outcome, they receive no points. The player with the highest point total at the end of the games wins. Therefore to maximize the point total, the player should assign the highest confidence points to games in which the player is very certain of the outcome, and the lowest points to games where the player is very uncertain of the outcome.

Analysis and Plan While this game is very straightforward to play, there are many statistical and machine learning problems that arise when trying to form an optimal strategy. First of all, data needs to be gathered to help predict outcomes. For this, I will gather data from results during the regular season. This data is provided by *Sunshine Forecast Downloadable Data Files*. This data will need to be parsed in order to select features to be used in the classification problem as well as results of bowl games from previous years. The step of selecting features should require great attention as the possible feature space is incredibly large (yards per game, yards against per game, sacks per game, yards per game in last three games, etc). Finding features that explain the outcome can be difficult. Therefore, a large amount of time will be spent determining the best features to use.

Once features are determined (or perhaps concurrently), multiple strategies should be chosen to correctly predict outcomes. These include Support Vector Machines (SVMs), logistic regression, decision trees, and perhaps others. Results from other fields suggest that SVMs produce reliable results. However, logistic regression provides the most straightforward approach to determining the confidence of a prediction as it yields probabilities for each class.

Data will need to be split into train, test, and validate sets, perhaps using a strategy such as k-fold cross-validation. There are many other aspects of this problem that can be investigated. For instance, the loss in this problem is obviously asymmetric as confidence points change for every game. However, intuitively it would seem that the largest confidence points should be assigned to the games with the lowest variance or highest probability.