ISyE 6416 – Basic Statistical Methods - Fall 2015
Bonus Project: "Big" Data Analytics
Proposal

Team Member Names: Bella Smith & Betsie Last

Project Title: Baseball Predictions

For proposal only, please include (at least) the following sections.

Problem Statement

Which is better for predicting what baseball team will make the playoffs, k-means clustering or LDA classification?

Data Source

The data comes from the internet- baseballreference.com- sports data.

Methodology

We will use data from 1975-2015. Instead of using data from the whole season, we will get runs scored and runs against from August 16th of each year. This allows us to predict whether or not a team will make playoffs at a time most relevant to baseball fans and managers. Since July 31st is the trade deadline and September 1st is roster expansion deadline, using data through August 16th will allow us to account for finalized rosters while giving ample time before the expansion occurs for managers to make decisions on who to include in the expanded roster.

We will look at the three variables: runs scored, runs against, and previous history making playoffs three years prior. The statistic for past history will be computed as the percentage of times in the past 3 years a team went to playoffs x 100. We chose to include runs scored and runs against because they are known to have a mostly linear correlation with winning percentage, which is the main predictor for whether a team will go to playoffs. We chose to include history of the past three years because while teams who are performing well tend to stay that way, there are changing rosters that may affect whether or not they make playoffs from year to year and we thought three years would be include these changes.

Our k-means clustering will be in three dimensions: runs scored, runs against, and previous history. We will cluster in two groups: likely to make it to playoffs, and unlikely to make playoffs. Since the final clusters can be impacted by our choice of initial centers, we will run the k-means algorithm ten times. We will choose the centers for each of these ten runs by randomly selecting three data points each time. We will use the standard Euclidean distance to compute each data point's distance from the center with which it is

clustered. Each data point will be assigned to the closest initial center using the standard Euclidean distance for three dimensions, which will optimize the minimization of the total mean squared error between the data points and their clusters. Then, the new center of each cluster will be computed by averaging all of the data points assigned to the cluster. We will alternate between these two steps for 100 iterations. We choose 100 iterations because the k-means algorithm is known to usually converge quickly.

After we run the k-means algorithm ten times, we will choose the clustering that provides us with the smallest within-cluster variation. We will measure the error of this clustering by computing the percentage of teams that it incorrectly predicted.

For classification/linear discriminant analysis we will once again predict whether a team will make it to playoffs using runs scored, runs against, and previous history. In looking at the retrospective data, we divide the teams into two classifications: those that made playoffs (y=1) and those that did not make playoffs (y=0). Contrary to our k-means algorithm we will not separate the data by year and will build our boundary using the about 120 data points we get from the 40 seasons with approximately 30 data points per season. In order to run this classification, we will save 20% of our data for testing our boundary and analyzing its effectiveness and train on the other 80%.

Using our randomly-selected 96 data points (80% of our 120 data points) we will build our decision boundary which will be a plane since we are working in three 3 dimensions. We will then use this plane to classify the data points as follows: identifying points that fall on one side of the decision boundary as those teams that will (or are likely) to make playoffs and those that fall on the other side of the boundary as those teams that will not (or are not likely) to make playoffs. To build this decision boundary we will extend the formula for finding the decision boundary in two dimensions.

Expected Results

Since the variables we chose are known to be good indicators of playoff predictions, it should make sense what we get for our results. Our results should match what actually happened in regards to playoffs. Our sample size is large enough, as we have 40 years of data that we are analyzing. Therefore, our sample size is representative and there is no source of bias for concern.

In order to improve our model, we could include more variables or combine them in a different way, such as creating a new statistic from runs scored and runs against or accounting for past playoff history differently. Additionally, we could also take into consideration the changing structures of the playoffs and leagues. Over the past 40 years there have been various changes made to which teams are in which division as well as the amount of teams to make playoffs.

Our results will help coaches and managers of baseball teams, as well as fans, predict which teams will make it to the playoffs. This is useful for those that are involved in baseball, and our algorithm can be expanded to making predictions in other sports as well.