9.520/6.860 Project Abstract

Subby Olubeko
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
subbyo@mit.edu

1 Overview

The NBA's Most Valuable Player and All Star awards are two of the highest honors a player can receive. These awards are given out every season to the player(s) who have been determined to have done the most to deserve recognition according to a vote from fans across the entire United States, media personalities, and their fellow players. There are an immense number of factors that could go into deciding which players will receive these honors including things such as their performance over the season, general reputation with NBA fans, individual impact they contribute to their team, improvement over recent years, and many others. The NBA keeps track of many rich statistics on players that could be collected as data points which would constitute a high dimensional data set that could be used to predict the future MVP and All Stars. However, it is doubtful that all the features in this data set would have a great deal of influence over which candidates are selected as All Stars or the MVP in any given season. For my project, I employed techniques of sparsity based regularization to select from a broad group of factors the variables that have the most weight in the choice of NBA All Stars and MVP. I then used the relevant data points determined by these methods to predict which players are most likely to be voted as MVP and All Stars. My execution made use of both Lasso and Elastic Net regularization for variable selection as well as standard and kernelized Tikhonov regularization (Ridge Regression) for regression/prediction.

2 Data

The data set I used for this project consisted of four parts. The first was the player data which were made up of ten CSV files corresponding to each NBA season from October 2007 to June 2017. I used the first seven of these seasons as training data to fit the models to and the latter three as data for validation. These files were made up of a data point for each player in the corresponding season, which included basic information about the player (e.g. height, weight, position, years of experience...), metrics of their per game and overall season performance, information and performance metrics for their team, and statistics from their Twitter profile (likes, retweets, and tweets made in a given season). The second and third parts of the data set were two sets of labels for each player indicating how many votes they received for the All Star award and the percentage share of MVP votes they received. Finally, my data set included unlabeled data points of the same form as the training and validation sets for the current NBA season up to 12/3/2018, which I used to predict the upcoming All Stars and MVP. The data for my project was obtained from https://www.basketball-reference.com and using the python-twitter API.

3 Experiment

The first step I took was normalizing the player data points to account for different patterns across seasons and also remedy the fact that the current season is still underway and thus the data points that I'd be using to predict this season's awardees would differ greatly in absolute value from previous (completed) seasons. Next I performed variable selection using both the Lasso and Elastic Net

techniques and compared the loss from the optimal parameter settings for both methods. I made use of L1 loss in selecting parameters for my models. After comparing the minimal losses for both methods, I found that Elastic Net was a better variable selector for predicting the MVP while Lasso gave me optimal results for All Star Prediction. I then condensed the player data to only include the features that were selected by the previous regularizations and used Ridge Regression on the condensed data to predict the awardees. One thing I noticed was that the model's predictions didn't seem to vary much from season to season because previous season's All Star/MVP results were being considered heavily in it's prediction criteria. So, I experimented with nonlinear transformations of the data and found that using Ridge Regression with a quadratic kernel resulted in more varied predictions due to less dependence on previous All Star/MVP votes. So my final model was a mixture of the regular and quadratic Ridge Regression models, using a mixing percentage parameter that was chosen to optimize loss on validation data.

4 Results

The selected variables for prediction were similar in both the All Star and MVP cases. The factor determined to be the most important by the regularization methods was the number of seasons in which the player had received the award in question. As a result, Lebron James was consistently predicted to have a high chance of winning the MVP award since he has won it many times in recent years. The second most important factor was determined to be an advanced statistic called VORP (Value Over Replacement Player) which essentially captures how well a player performs relative to an average free agent that could hold their position. In the case of MVP prediction, three more factors were determined to be important. These were, in respective order of importance, the number of times the player's tweets from a given season were favorited, the number of times the player had been voted as an All Star in the previous five seasons, and the total number of freethrows attempted by the player in a given season. For predicting the All Stars, there were two more relevant factors which were the number of freethrows attempted per game and the total number of freethrows attempted in a season by a player. Searching over the possible range of 0 to 1 for the optimal value for the mixing parameter for the standard linear Ridge Regression and kernelized quadratic Ridge models, I found that the optimal setting for the MVP model was a weight of 0.29 for the linear model and 0.71 for the quadratic model, while for the All Star predictor the optimal weights were 0.63 for the linear model and 0.37 for the quadratic. With these settings, my predictor was able to achieve roughly 82% accuracy at predicting MVP candidates and 72% accuracy at predicting All Star awardees for the previous three MVP seasons. My model's predictions of the upcoming All Stars and MVP seem fairly reasonable when compared to similar predictions from Basketball Reference.