Predicting the NBA's MVP and All Star Award Recepients Using Sparsity Based Regularization

Subby Olubeko[1]
subbyo@mit.edu

9.520/6.860 Statistical Learning Theory
Massachusetts Institute of Technology

Abstract

Background:

The NBA's Most Valuable Player and All Star 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.

Objective:

To use techniques of sparsity based regularization to select from the broad group of factors available in the data maintained on players by the NBA, which of these variables have the most weight in the choice of All Stars and MVP. And to then use the data determined to be relevant by these methods to predict which current players are most likely to be voted as MVP and All Stars.

Methods:

Variable selection with Lasso and Elastic Net regularization. Prediction with standard Tikhonov regularization and kernelized Ridge Regression.

Data

- **Player data**
  - 10 CSV files, each containing data from corresponding NBA seasons from October 2007 - June 2017
  - October 2007 – June 2014 used as training data to fit models
  - October 2014 – June 2017 used for validation
  - One data point for each player made up of features from 6 categories:
    - Basic Info (Height, Weight, Position, Years of Experience, Home Country, Draft Position)
    - Per game season stats (Minutes, Points, Steals, Rebounds, Blocks, Assists…)
    - Overall season stats (Salary, Points, Freethrow Attempts…)
    - Team season stats (Wins, Losses, Attendance…)
    - Indicator fields for teammates
    - Twitter profile season statistics (Tweets, Retweets, and Favorites)
  - Total number of features = 1371
- **Labels**
  - MVP prediction: Percentage of MVP votes received by each player. In range (0, 1)
  - All Star Prediction: Number of fan votes received by each player (typically in hundreds of thousands to millions)
- **Unlabeled Data from Current Season**
  - Same form as training and validation data.
  - Up to date as of 12/3/2018
  - Used to predict the upcoming All Stars and MVP
- Obtained from https://www.basketball-reference.com and using Python-twitter API

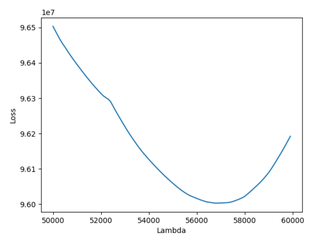
Preprocessing

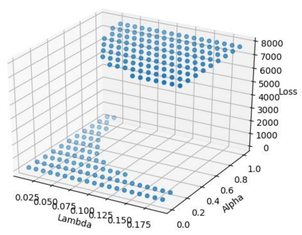
Normalized player data points

- Mean centered and scaled by standard deviation
- Account for different patterns across seasons
- Adjust for large differences in current season data due to the fact that it is still underway and thus many statistics would be much smaller in value than complete season data

Variable Selection

- **Minimized L1 Loss**
- **Lasso**
  - Performed linear search for optimal parameter λ
- **Elastic Net**
  - Performed grid search for optimal λ and α parameters
- Used non-zero coefficients in optimal **ω** vector from model with minimal loss to determine relevant variables
  - Lasso optimal for All Star Prediction
  - Elastic Net optimal for MVP Prediction





Prediction

- Condensed player data to only include selected variables
- Ridge Regression on condensed data to predict award winners
- Kernelized Ridge Regression
  - Low variance in predicted winners between seasons because previous awards were being considered heavily
  - Transformed data using quadratic kernel
- Mixed linear and quadratic models to attain better results

Results

- **Most Relevant Features:**
  - MVP
    - Number of times voted MVP in previous 5 seasons
    - VORP (Value over Replacement Player)
    - Number of times Tweets were favorited in season
    - Number of times voted All Star in previous 5 seasons
    - Number of freethrows attempted in season
  - All Stars
    - Number of times voted All Star in previous 5 seasons
    - VORP (Value over Replacement Player)
    - Number of freethrows attempted per game
    - Number of freethrows attempted in season
- **Accuracy on previous 3 seasons:**
  - MVP Candidates – 82%
  - All Stars – 72%
- **Predictions:**

MVP


LeBron James
34%


James Harden
13.2%


Russell Westbrook
8.2%


Stephen Curry
8.2%


Damian Lillard
6.2%


Kevin Durant
5.6%


Giannis Antetokounmpo
5.5%


DeMarcus Cousins
4.6%

All Stars

East: LeBron James, Kyrie Irving, John Wall, Giannis Antetokounmpo, DeMar DeRozan, Kyle Lowry, Dwight Howard, Al Horford, Andre Drummond, Dwyane Wade, Kemba Walker, Kevin Love, Joel Embiid, Ben Simmons, Victor Oladipo, Isaiah Thomas, Bradley Beal, Kristaps Porzingis, Joakim Noah, Aaron Gordon, Jayson Tatum, Kyle Korver, Otto Porter, Jeremy Lamb, Robert Covington, DeMarre Carroll

West: James Harden, Russell Westbrook, Stephen Curry, Kevin Durant, Anthony Davis, DeMarcus Cousins, LaMarcus Aldridge, Damian Lillard, Paul George, Paul Millsap, Carmelo Anthony, Chris Paul, Blake Griffin, Jimmy Butler, Marc Gasol, Klay Thompson, Draymond Green, Pau Gasol, Tony Parker, Dirk Nowitzki, Kawhi Leonard, DeAndre Jordan, Jeff Teague, Nikola Jokic, Devin Booker

Conclusions

- The MVP and All Star award recipients can be predicted very accurately using only a few (4-5) of the available features
- Dependence on this features is not completely linear and can be better captured by a mix of linear and quadratic regression.