John Conley
Adam Tong
Matt Girouard

Predicting Basketball Player Statistics

Problem Statement

There's huge potential in being able to predict a basketball player's performance for the upcoming year. With the average player's yearly contract worth several million dollars, teams must take great care in choosing their next trade. For the casual fan, too, there is value in knowing how a player will perform, as when a fan is picking the roster for his fantasy team. We would like to take advantage of 70-odd years of detailed basketball data to predict an individual player's performance over the next year, as measured by points scored, rebounds, assists, steals, and other relevant factors. We will then combine all these data points into a player-performance rating that allows us to succinctly measure the accuracy of our predictions.

Because the game of basketball changes over time and players' productivities change over their careers, we normalize each player's yearly statistics by: pace, minutes played, position, age, and year of career. The idea is that the career progressions of historical players can help us predict the future performance of current players, normalized to the same baseline.

Method

We are planning on using a regression model where our input vectors include, but are not limited to: age, year of career, position, game pace, points scored, rebounds, assists, steals, blocks, field goal percentage, etc. We are considering implementing a random forest algorithm. The main advantage with using random forests is that we will not have to worry about overfitting, which is helpful for a problem like this where it is very easy to overfit the model to the training set [3]. It also is useful for weighting variable importance, which is critical in analyzing player performance [3]. Finally, it runs efficiently on large data sets with great accuracy [2]. Our second option of algorithm is Multivariate Adaptive Regression Splines (MARS), whose implementation is outlined in Jerome Friedman's 1991 paper [1]. MARS improves on normal linear regression by introducing hinge functions, which act like piecewise functions and allow for more greater flexibility in the solution function.

Data

The data will be scraped from www.basketball-reference.com, which has all available game and player data dating back to the founding of the NBA and ABA. The NBA and ABA were eventually merged into one league, so we use data from both leagues. We will use all of the player data.

Milestone goal

At the milestone, our goal is to have collected and formatted all of the data. We will also have a rudimentary implementation of the algorithm we end up choosing to best model our regression.

References

[1] Friedman, J.H. (1991). Multivariate Adaptive Regression Splines. *Annals of Statistics* **19**, 1-67.

[2] http://www.stat.berkeley.edu/~breiman/RandomForests/

[3] http://www.whrc.org/education/indonesia/pdf/DecisionTrees_RandomForest_v2.pdf