Machine Learning Engineer Nanodegree

Capstone Proposal

Alexandra Brinegar
April 10, 2018

Proposal

Domain Background

The National Basketball Association (NBA) is the premier men's professional basketball league of North America. It is very popular not only in North America but across the world. Many people bet on the outcome of NBA games, and the sports betting industry is worth many billions of dollars. Extensive research has been done in an attempt to determine how certain factors influence the outcome of games[1,2]. There are many factors that can impact the likelihood of a team winning any given game such as the team's 'win-loss' record, average points per game, home court advantage, injured players, etc.

When we predict which team will win a game, we are usually biased towards one or more of the influencing factors. There is a huge amount of statistics that has been collected from the NBA, and much of the data could be used to create a model that would make a prediction more accurately predict the winner of a matchup.

Problem Statement

For my capstone project, I will create a machine learning model that will predict the winning team in any NBA basketball game. It will analyze data from the two teams in any given match where the input is an away team and a home team, and it will output the team that is most likely to win the game according to the model. The success of the model can be determined by comparing the actual winner of the game (after the game is played) to the predicted winner from the model. I will run the model on many games that have been played so that I can get the success as a percentage of which winners were accurately predicted, rather than only running it on one game. This model could be used to predict winning teams in many future games.

Datasets and Inputs

I will be using data collected from regular season NBA games (not including pre-season or playoff games). To create data sets, I have consulted basketball-reference.com[3] and found many different statistics and extracted the data into a spreadsheet. I extracted data beginning with the 2012-2013 season to now (about 6 years of data). I will the team names with numbers during the data preparation phase of my project. The data includes basic statistics such as points, rebounds,

and assists as well as the outcome of every game for each season. I will test several algorithms that use the data in different ways to figure out which data is the most influential in predicting the outcome of a game.

Solution Statement

The solution to this problem is to experiment with Machine Learning methods such as linear regression or support vector machines, and to compare the results of the methods to determine which has a better prediction rate. I will analyze the data and prepare it so that I can prepare the best inputs for my method. By comparing multiple methods, we will be able to find out which model will give us higher likelihood of selecting the winning team.

Benchmark Model

For my benchmark model, I would like to compare my results to the prediction rate of the 'majority vote', which would be to simply predict that the winning team is always the team with the fewest losses. We will calculate this rate by selecting the team with the better win-loss percentage as the winner of the match. I computed this for the 2012-13 season and it had about 64% accuracy, so my goal is to get higher accuracy than that.

Evaluation Metrics

The main evaluation metric for this Machine Learning model will be the prediction accuracy. This will be calculated as a percentage of how many games the model is able to accurately predict the winning team.

Project Design

- **Programming Dependencies:** python 2.7, scikit-learn, pandas

1. **Data Preparation**

   The first step of my project will be to collect the necessary data and clean the dataset. I will extract data from basketball-reference.com into a CSV file that contains the data in numeric values.

2. **Split the Data**

   I will split the data into training and testing data. The data from the 2012-13 to 2016-17 seasons will be used as training data, and the 2017-18 (current season) data will be used as testing data.

3. **Apply Machine Learning Models**

I will then build construct a machine learning model by training different classifiers on the data. I will start by applying logistic regression and then try support vector machines. I may also try Naïve Bayes on this data.

4. **Model Calibration**

   I will select the model with the best prediction rates and experiment with it to increase its performance.

5. **Testing and Results Analysis**

   To test the model, I will use the best classifier to predict which team will win any match given a home team and an away team. I will then analyze the results in comparison to my benchmark model and see how well we were able to achieve our goal.

REFERENCES

[1] Hoffman, Lori. Joseph, Maria. 'A Multivariate Statistical Analysis of the NBA' http://www.units.miamioh.edu/sumsri/sumj/2003/NBAstats.pdf

[2] Beckler, Matthew. Wang, Hongfei. Papamichael, Michael. 'NBA Oracle' https://www.mbeckler.org/coursework/2008-2009/10701_report.pdf

[3] [NBA Statistics] https://www.basketball-reference.com/leagues/