Machine Learning Engineer Nanodegree

Capstone Proposal

Matheus Vilachã Ferreira Pires

July 11th, 2017

Using Machine Learning to Predict NBA Games Winners

1. Domain Background

The National Basketball Association (NBA) is the major men's professional basketball league in North America and was founded in 1946. It has 30 teams (29 in the United States and 1 in Canada), and its players are the world's best paid athletes by average annual salary per player. Through the years, a lot of data has been collected based on NBA. It is useful for simulations, to analyze teams and players performances, to assist coaches and their staff, and even to predict results.

From what I've learned with the Machine Learning Engineer Nanodegree, I think it may be interesting to use both unsupervised and supervised learning to predict NBA results. The dataset in this project will be composed of the following information (for the home team and for the visitor team): stats from the last game, the average of the stats from the last two games, and the average of the stats from the last five games. I intend to use unsupervised learning to group the most relevant features on my dataset and reduce the curse of dimensionality over it, and supervised learning algorithms to predict whether the home team or the visitor team will win the game. I also want to determine which of the following is most important in this kind of analysis: the last game stats, the average of the last two games or the average of the last three games.

2. Problem Statement

Although there is a large amount of data about NBA nowadays, it is a complex job to analyze and predict results based only on it. The main objective of this project is to train a machine to achieve a reasonable prediction rate (higher than 70%) on stating whether the home team or the road team will win a NBA game using exclusively Machine Learning.

3. Datasets and Inputs

I will use the Sports Data Query Language to obtain the database from the killersports.com website. It will contain games from 2002 to 2016. I plan to use a dataset with stats (continuous variables) from the last game, the last two games (average) and the last five games (average), for both the home and the visitor team: streak wins, points made, field

goals made, field goals attempted, three points shots made, three points shots attempted, free throws made, free throws attempted, offensive rebounds, defensive rebounds, rebounds, assists, turnovers, steals, and blocks. The dataset will also contain the following features (discrete variables): number of overtimes (if any), week day, month and playoffs (dummy variable).

Since my goal is to determine if the performance in the last game, the performance in the last two games or the performance in the last five games is the most important to state if a team will win a game, I will first divide the dataset in three: stats from the last game, stats from the last two games and stats from the last five games. All of them will contain the discrete variables listed in the paragraph above. After that, I intend to use unsupervised learning techniques on the numerical features (with continuous values) to reduce the dimensionality; I believe it will improve the final results.

4. Solution Statement

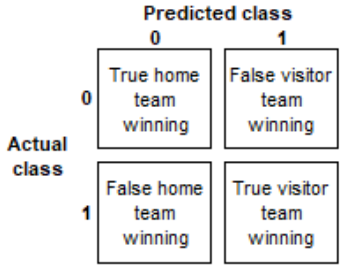
As mentioned earlier, I plan to use both unsupervised and supervised learning techniques to achieve a reasonable prediction rate. I will use unsupervised learning to find hidden information in the dataset and reduce its dimensionality. With that, supervised learning algorithms (such as logistic regression, SVM, neural network and so on) will be faster to train and the computational cost will be decreased. My goal is to determine which of the three datasets (from the last game, from the last two games or from the last five games) is the most reliable to predict NBA winning teams.

5. Benchmark Model

I will use the trained and optimized models to predict the results from the 2016-2017 regular season (which is not in the dataset). After that, I will compare the results of the three models with the results from that season. Besides that, I will use as benchmark models a logistic regression model (untuned and untouched) and a naïve assumption that the winner team will be the one with the highest number of streak wins.

6. Evaluation Metrics

To evaluate the models, I will use confusion matrix and accuracy to determine the best trained model. The confusion matrix will be used to identify true positives, true negatives, false positives and true negatives:



In other words, my goal with it is to determine if a model has difficulty to identify when the home team will win, when the home team will lose, when the visitor team will win or when the visitor team will lose.

Accuracy is calculated dividing the sum of correct predictions by the total number of predictions (it is the percentage of correct predictions).

7. Project Design

For this project, I will use Python 3.6.1 as the programming language and the scikit-learn library.

The first step in my project is to collect the data. As mentioned before, I will use Sports Query Data Language to extract it from the killersports.com website (the code used can be found in the sqdl_capstone_code.txt file).

After that, I will make it a good training dataset using data preprocessing techniques, like removing missing values, getting categorical data into shape for machine learning algorithms, and scaling numerical features using standardization approach. Then I will split the dataset in training set (70%) and testing set (30%).

The next step will be, as mentioned before, divide the dataset in three: one containing stats from the last game, one containing the average stats from the last two games, and one containing the average stats from the last five games. These three datasets will contain the categorical features mentioned earlier (number of overtimes, week day, month and if it is a post-season game). At this point, it is important to remember that one of my objectives in this project is to determine if the best dataset is from the last game, the last two games or the last five games (that's why three datasets).

After that, I will try to compress each dataset via dimensionality reduction using principal component analysis (PCA). In this part, I will use what I've learned about unsupervised learning. The next step will be use ensemble methods to combine different classification models and predict the winners (for each dataset). I plan to use AdaBoost and majority vote classifier in this step.

Finally, I will compare the three results and use the 2016-2017 regular season dataset to validate them and determine which dataset is the best to predict NBA games winners.