Two Machine Learning Approaches to Understand the NBA Data

Panagiotis Lolas

December 14, 2017

1 Introduction

In this project, I consider applications of machine learning in the analysis of nba data. To be more specific, in the first part I employ supervised learning algorithms in order to make predictions about the outcomes of nba games and in the second part I show how unsupervised learning algorithms can be used to discover interesting similarities between teams and derive important conclusions about the way basketball is evolving. The data used for this project were collected from https://www.basketball-reference.com.

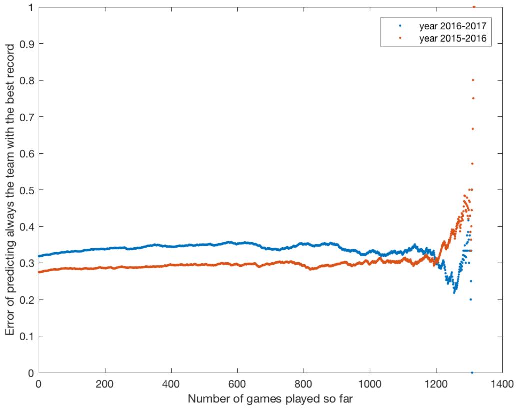
2 Related Work

In the past there have been various machine learning applications in NBA analytics. For example, in [6] they focused on the prediction of the outcomes relying on classification methods. In addition, in [3] the author used clustering algorithms to discover 8 different player types in the league. In [5] the author used machine learning in order to analyze the effect of different player positions on winning.

3 Supervised Learning Applications

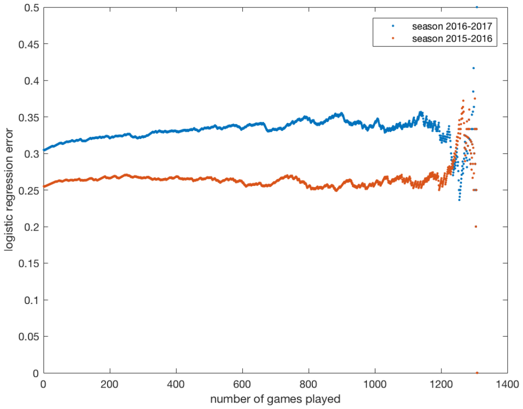
3.1 Using Records to Make Predictions

To get a sense of how simple methods perform when it comes to making predictions, I started with the simple rule according to which we always predict that the team with the best record wins. If two teams have the same record, we predict that the home team wins. Plotting the error as a function of the number of games played so far we get the following plot for the seasons 2016-2017 and 2015-2016.

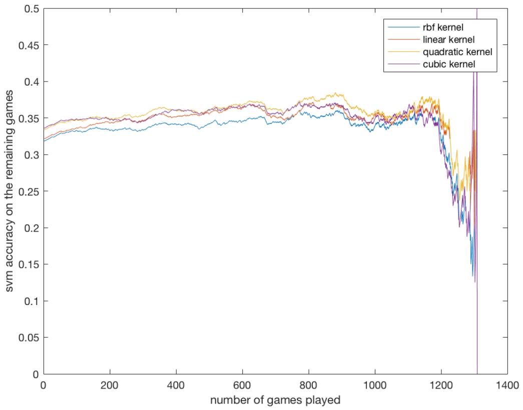


We see that it is about 28 percent for 2015-2016 and around 33 percent for 2016-2017. In the playoffs this went up for 2015-2016 (probably because the number of the test set became small and there were a lot of "unexpected results", such as the Cavaliers becoming champions against the 73-9 Warriors), while in the 2016-2017 the accuracy of this method showed smaller volatility (which again can be explained by the actual results and human sense).

Next, I used some of the methods covered in class. I started by using logistic regression with features the records of the away team and the percentage-win of the home team before the game. Notice that this takes into account which team is home and away, as the feature vector is "ordered". An output of 0 corresponds to away win and an output of 1 to homewin. For the seasons mentioned above, my results for the error as a function of games played so far were the following. I trained the model on the first 300 games in order to have enough data, so the value around 300 corresponds to the test error. The training errors were about 20 percent for the first season and 23.67 percent for the second one (on a training set of 300). Notice that on average this is slightly better than the strategy from above, but only by a very small amount. For the regression coefficients, I got for 2016-2017 (and the same effect is observed in 2015-2016) $(-0.15, 5.41, -6.1)$, which captures idea that the team playing home has the advantage usually, in the sense that the logistic regression model predicts that the away team wins only if the win percentage of the away team is at least (6.1/5.41)×(win percentage of home team)+0.15/5.41.



Using the same features as above and training a support vector machine with linear, quadratic, cubic and Gaussian kernel I got the following plot for the errors for the season 2016-2017. In order to predict the future outcomes we take into account all the results that we have before the game day (except for the first 200 games, where I trained the model on all of the 200 because of data limitations). We see that the performance is about the same for all of them, with the Gaussian kernel being slightly better and giving about 66 percent of the time the right prediction for the test set.



If we visualize the features we see that there is significant overlap between the two classes. Hence, I thought that linear discriminant analysis might be a better method than the ones mentioned earlier. Using linear discriminant analysis, we get 69.7 percent accuracy. From the confusion matrix, we get the following evaluation metrics:

precision=72 percent

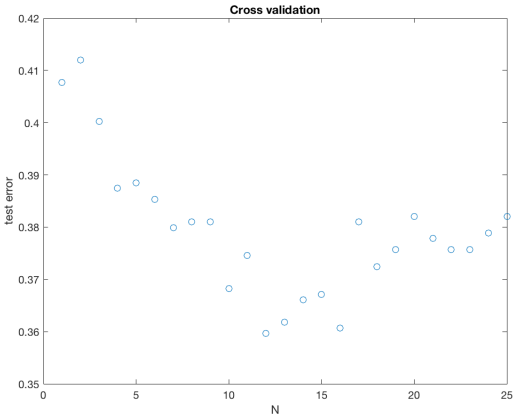
sensitivity=80 percent

specificity=59 percent

We conclude that our model has some tendency to overpredict that the home team will win.

3.2 Adding more features

In what I described above, it seems like the dimension of the features is too small to capture all the variability for the factors affecting outcomes in the NBA. The shape of teams in the last weeks for example is one thing we ignored so far. For this reason, I included the point margins of the games of the teams in the past N games in order to make prediction. In order to choose N, I used cross validation by training logistic regression using results from the first two months of the season and then testing on the rest of the season, for values of N from 3 up to 25. The test errors can be seen in the plot.



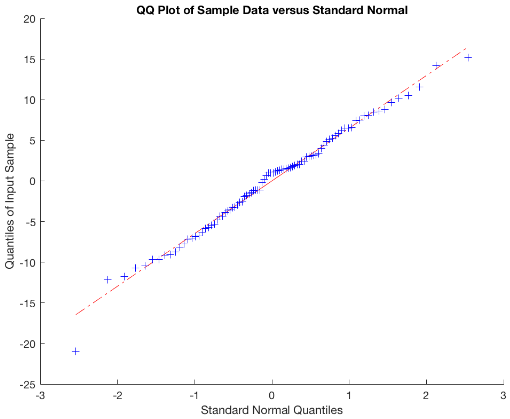
It seems like something around $N = 12$ is optimal. However, the test error seems to have gone up compared to the case when we included only the win percentages in the features. I think this is because $N = 12$ corresponds to adding 24 more parameters to the model (we need to look at 12 games for home team and 12 games for away team for each prediction). This led to increased variance of our model and the performance for making predictions for future games has decreased a little bit, even in the best case $N = 12$. The same phenomenon was observed when training an SVM using the larger feature space that we described. In addition, taking into account other statistics, such as points per game scored by a team, seems to have the same effect. In my opinion, the sports from data involve a great deal of randomness and the simplest models that we can try such as the ones explained above are expected to perform better on average than more complex models. Reading other papers on the internet it seems like 70-73 percent accuracy is close to the best people have achieved on this topic and this does not beat our methods significantly.

4 Unsupervised Learning Applications

4.1 Anomaly Detection

In this subsection we employ unsupervised learning methods to detect abnormalities behind the statistics of teams in the previous three seasons. For this part, the statistics that we use are minutes played (MP), number of field goals (FG), field goal attempts (FGA), field goal percentage (FG%), number of three-point shots made (3P), number of three-point shots attempted (3PA), three-point shot percentage (3P%), two-point shots made (2P), two-point attempts (2PA), two-point shots percentage (2P%), number of free throws made (FT), number of free throws attempted (FTA), free throw percentage (FT%), number of offensive, defensive and total rebounds (ORB, DRB and TRB respectively), number of assists (AST), number of steals (STL), number of blocks (BLK), number of turnovers (TOV), number of fouls (PF) and number of points scored (PTS).

Firstly, doing principal component analysis (as some features are clearly linearly dependent, such as ORB, DRB, TRB) we find that the first six principal components account for 95.33% of the variance in the data. For this subsection we will be using the projections of the data on this six dimensional space. By doing a qqplot and a Kolmogorov-Smirnov test, we can see that the projection of the features on this space is a multivariate Gaussian distribution. Because of the orthogonality of the principal components, here we just check this for the projection on the space spanned by the first principal component. The qq-plot that we get for this distribution is the following.



For the Kolmogorov-Smirnov test the p-value that we find is 57.97%.

Now that we know that the (projected) set of features is Gaussian, we can use a χ^2-test in order to detect abnormalities in the data set. In particular, after rescaling the data we use a χ^2_6 in order to detect data points that are very far from the origin. One interesting thing about this approach is that it takes into account the whole set of features and does not only try to detect strange performance of teams in one specific category. Using a confidence level of 10% and the test described above, we observe that the statistics of the following teams are peculiar.

Warriors 2016-2017, Rockets 2016-2017, Suns 2016-2017, Mavericks 2016-2017, Warriors 2015-2016, Kings 2014-2015, 76ers 2014-2015

Because of space limitations, we only analyze two of those and the rest can be explained similarly.

Rockets 2016-2017

The projected feature vector for this team is: $(25.6, -0.3, 2.6, 0.9, 0.8, -0.1)$. Notice that this is approximately a "big" multiple of the first principal component. Using the results from pca for the first principal component we find the following about the Rockets in 2016-2017.

This team attempted and made an extremely high number of three-point shots, while they did not attempt a lot of two-pointers. They also went to the free throw line very often, had a tendency to make turnovers and, of course, scored a great number of points.

One might argue that we could have detected the above abnormalities even without any knowledge of machine learning. A more surprising example, however, is the following.

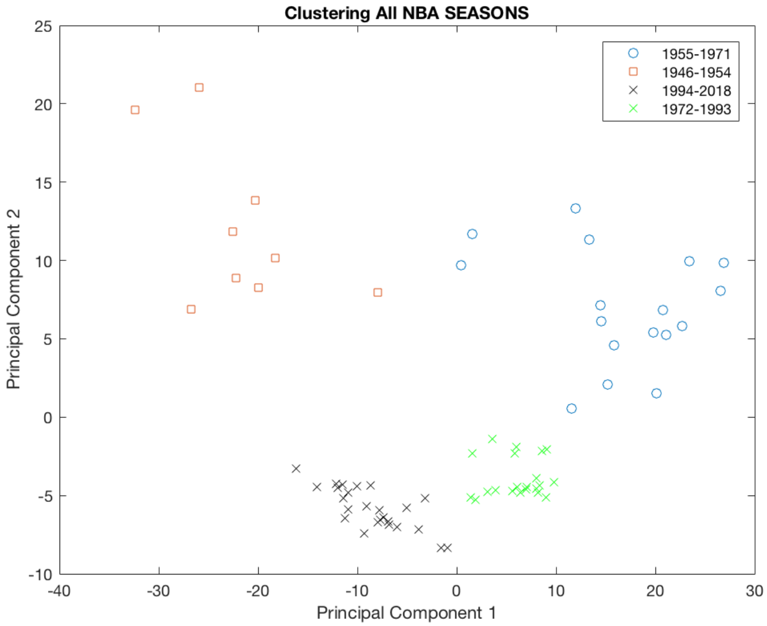
Suns 2016-2017

The feature vector in the 6-dimensional space of features is $(-1.8, 8.9, 4.9, 4.2, 3.3, -1.8)$. We see that this is somewhat "split" between the different components and this is what made the statistics of the Phoenix Suns in 2016-2017 look strange.

Using the values for the first six principal components that we found above, we can describe the Suns in 2016-2017 as follows: They attempted a lot of field goals, but the number of attempted three pointers compared to the season averages was particularly low. They attempted a lot of two-point shots (in fact, if we look at the data, they shot more 2's than any other team in that season) and went to the free throw line very often (which seems reasonable, as they tried to go closer to the basket). They also grabbed a lot of offensive rebounds and made very few assists. Finally, they made a lot of turnovers and made more fouls than any other team in the league during 2016-2017. Definitely a strange combination of statistics!

4.2 The 4 Eras of Basketball

In this section we consider the league averages of all NBA seasons in the following categories: FG, FGA, FT, FTA , FT, FTA, AST, PF, PTS, FG%, FT%, eFG% (which is field goal percentage adjusted by the fact that three point shots are worth one more point) and FT/FTA. These are used because of being measured for all seasons in the history of the NBA (in contrast, for example, to the introduction of the 3 point line). After principal component analysis, we see that for this data set the first two principal components explain more than 92% of the variability in the data. Projecting the data on this 2 dimensional space, we can visualize the data and observe that there are probably 4 different types of distributions in the figure. I used the EM algorithm to fit a mixture of 4 Gaussian distributions to this data set and I got the following.



Surprisingly, our analysis indicates that the NBA basketball can be clustered in 4 different eras, namely years 1946-1954, 1955-1971, 1972-1993, 1994-2018.

We are now going to investigate the different characteristics that define those four eras. We are going to find the vectors connecting the means of the Gaussians that we found above and use them as measures of the dissimilarity between the clusters.

For the transition between era 1 and era 2, this measure indicates that after 1954 there was (on average) a significant increase in field goals made and attempted, as well as in points scored, a slight increase in assists and fouls. This is not surprising at all, since 1955 was the first season that the 24 second clock was introduced in the NBA!

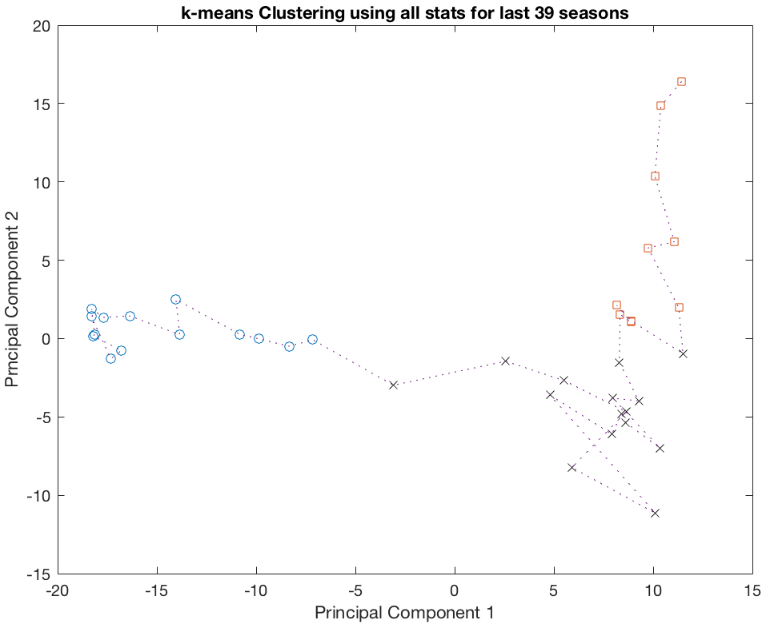
For the transition between era 2 and era 3, we observe that on average there was a decline in the field goals made and attempted, a decline in free throws made and attempted, a small increase in assists and a decrease in the number of points scored.

The transition between era 2 and era 3 is similar, observing (on average) a decline in all categories FG,FGA, FT, FTA, AST, PF, PTS (in different

”proportions” than in the previous transition).

4.3 Revisiting the last 39 seasons - A new era emerging?

If we look at the data of the current era using the principal components that we used above, we observe that right now there is a tendency towards bigger first principal component and smaller second principal component in a linear way. This is true roughly for data from the last 10 years. At this point we consider the last 39 seasons, for which much more data is available and since the 3 point line introduced 40 years ago is a significant change in the NBA. To the features used earlier we now include statistics related to 3 point shots, ORB and DRB, STL, BLK, TOV, Pace (possessions per 48 minutes), Offensive Rating (ORtg), ORB% (an estimate of the offensive rebounds a player grabbed while on the floor). Reducing the dimension to 2 using again PCA allows a more descriptive clustering. Here we used k-means cluster for $k = 3$ (chosen by silhouette analysis) and the dotted line shows the evolution of the features in time (the most recent season is on the upper right corner). The first black point corresponds to the season 1994-1995, so our here clustering agrees with the one in the previous subsection for years before 1994.



We conclude from the plot above that for the last decade there has been a single component which drives most of the movements in the league averages in the NBA. Looking at the principal components we see that this trend corresponds to an important increase in points scored and field goals attempted, in 3 point shots made and attempted, a dramatic change in pace and offensive rating (compared to the decade before) and a huge decrease in ORB%. In my opinion, this shows a tendency for a new era in the NBA. Of course, we do not have enough data points at this time to investigate this claim further. However, there is some reason to believe that this is indeed the case. One way to understand this is the following: We expect that the sport changes when most of the teams adopt a particular style of play. This leads to the variability of the data being explained to a great extent by a single component, something that we also observe between the first two clusters in the last plot. In other words, a trend in the NBA is exactly the same as the majority of the teams focusing on a particular style of play for several seasons in a row. This claim is going to be verified or disproved in the future seasons.

5 Conclusions and Future Work

In this project we saw how supervised learning can be used to predict outcomes of NBA games. The randomness in the outcomes made this task extremely difficult and did not allow great performance of our algorithms. After that, we explored how we can understand the evolution of basketball by using unsupervised learning. This led to the conjecture described in the previous section for the dynamics currently governing the NBA statistics.

An interesting direction to explore in the future might be to use similar techniques to the ones from the second to analyze the evolution of the basketball positions throughout time. This can be extremely useful in finding combinations that make a roster achieve a lot of wins. If such an approach was effective, teams might be able to reduce risk in building their roster and find optimal allocations of their salary cup.

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning.* New York, NY, 2006.

[2] Trevor Hastie, Robert Tibshirani, Jerome Friedman. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*

[3] Alex Cheng. *Using Machine Learning to Find the 8 Types of Players in the NBA* https://fastbreakdata.com/classifying-the-modern-nba-player-with-machine-learning-539da03bb824

[4] Lehmann, Erich L., Romano, Joseph P. *Testing Statistical Hypotheses* Springer, 2010.

[5] Dwight Lutz. *A cluster analysis of NBA players* http://www.sloansportsconference.com/wp-content/uploads/ 2012 / 02/ 44-Lutz-cluster-analysis-NBA.pdf

[6] Amorim Torres, Renato. *Prediction of NBA games based on Machine Learning Methods.* Computer-Aided Engineering, University of Wisconsin, Dec. 2013.