Two Machine Learning Approaches to Understand the NBA Data

Panagiotis Lolas - panagd@stanford.edu

Description

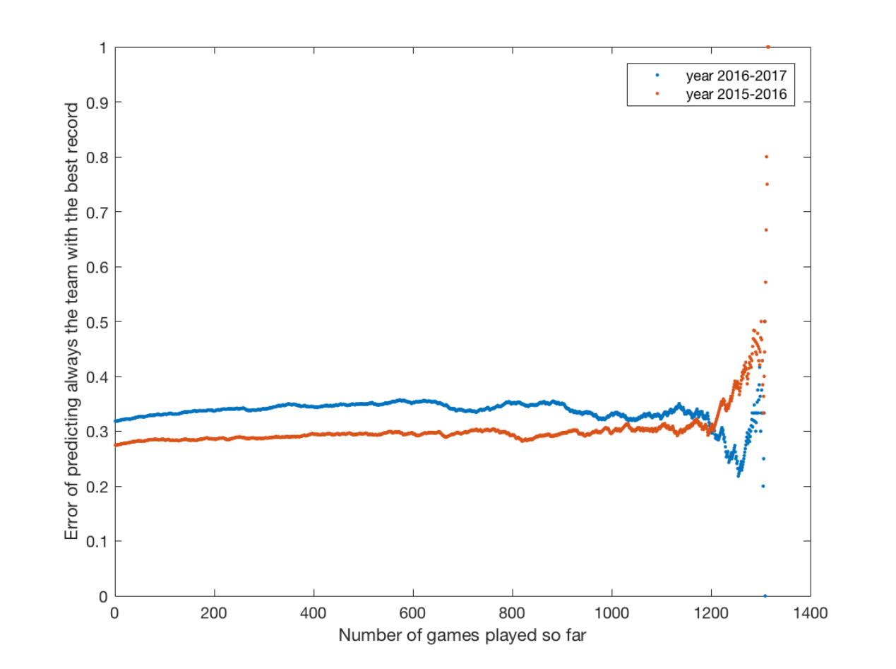
In the first part of the project we employ supervised learning in order to predict the outcomes of the NBA games. In particular, we are interested in predicting the winning team. In the second part we use unsupervised learning algorithms to discover interesting patterns behind the NBA data. Firstly, we try to detect abnormalities in the statistics of the teams and then perform a clustering of the NBA seasons based on league averages. We can interpret the clusterings to understand the evolution of basketball in the last 70 years.

Our Data

The source for our data was https://www.basketball-reference.com. For the first part we used the results of the teams in the previous games during the season and the total win percentage of home and away team to make predictions. We also considered statistics like points per game, 3P%, number of rebounds etc. For the second part we used team and league averages in categories like FG, FGA, 3P, 3PA, FT, FTA, ORB, DRB, TRB, AST. STL, BLK, TOV, PF, PTS, FG%, 3P%, FT%, Pace, eFG%, TOV%, ORB%, FT/FTA, ORtg.

Predicting the outcomes

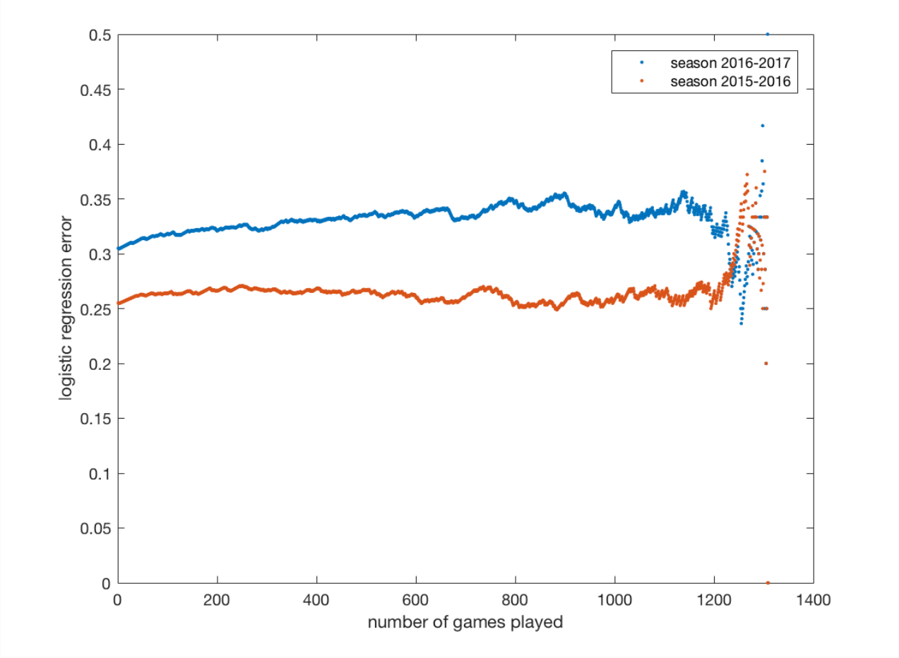
In the following plot we see the (test) error in the case that we predict always that the team with the best record wins in the past two seasons.

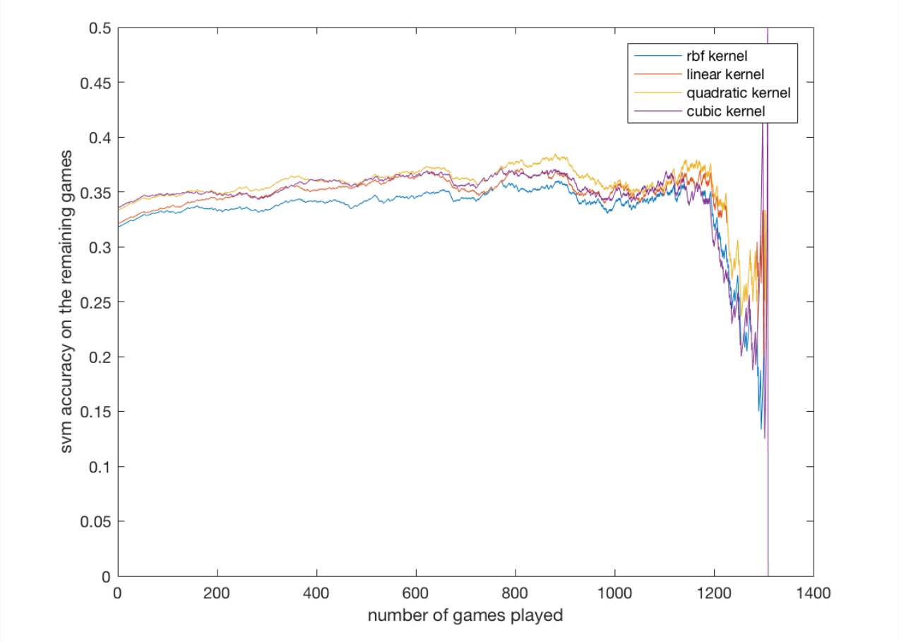


Next, we use logistic regression and support vector machines with different kernels in order to make predictions. The features initially consisted of the vectors of the win percentage of home and away teams.

...

We trained the models on the first 200 games and then tested on the rest of the season.





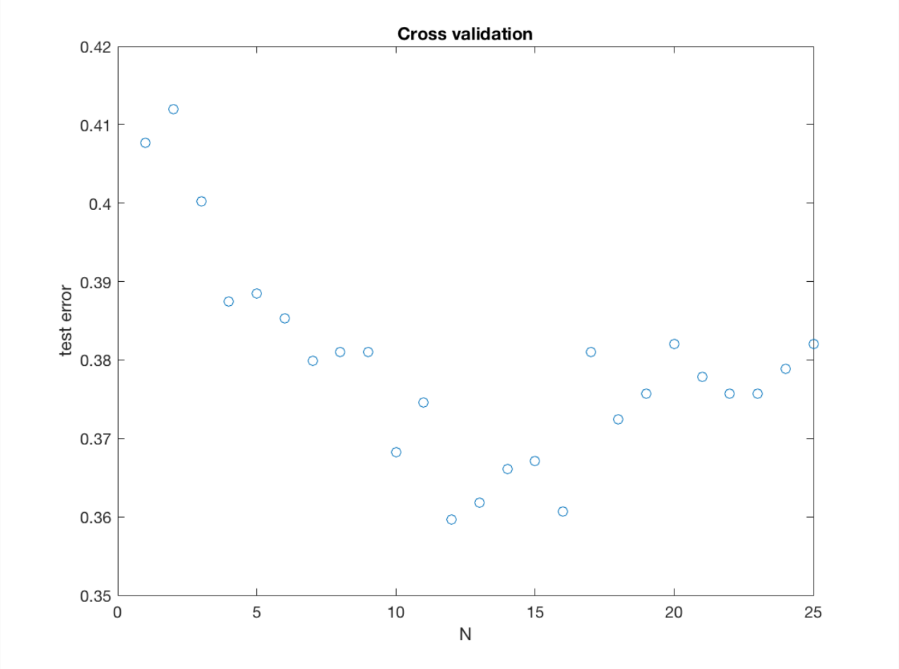
Plotting the data, linear discriminant analysis seems a better model. Using that model we get the following using the confusion matrix.

accuracy=69.7% precision=72% sensitivity=80% specificity=59%

We see that our model has a tendency to overpredict that the home team wins.

Adding more features

If we try to add more features, such as the results of the last N games played by each team (in an effort to capture the current shape the teams are in), using cross validation to choose the optimal value of N we observe the following for logistic regression (and similarly for SVM).



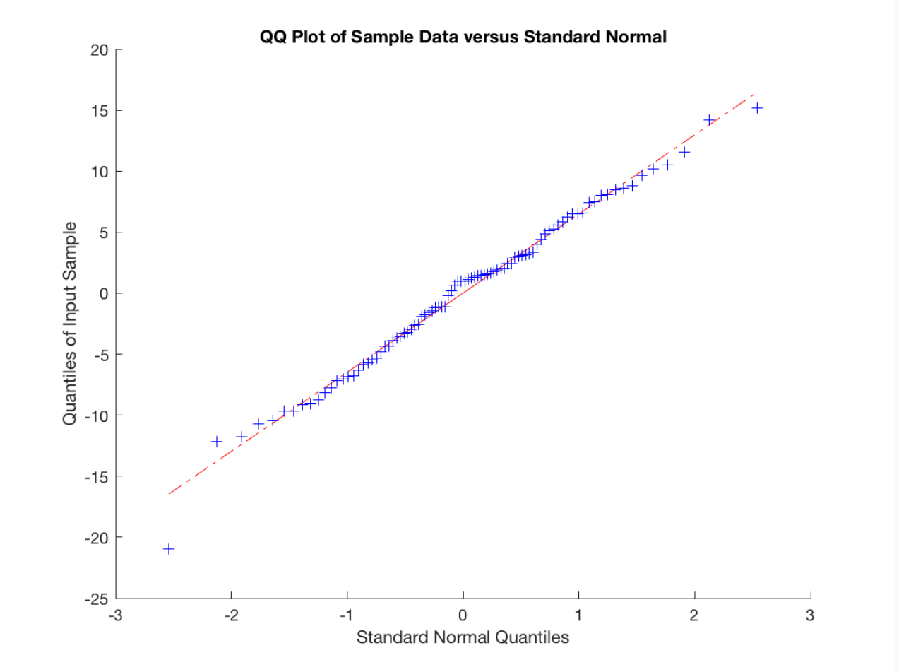
Similar phenomena observed when we add features like PPG, Field Goal Accuracy etc.

Detecting Abnormalities in the past 3 NBA seasons

We use PCA and reduce dimension to 6

...

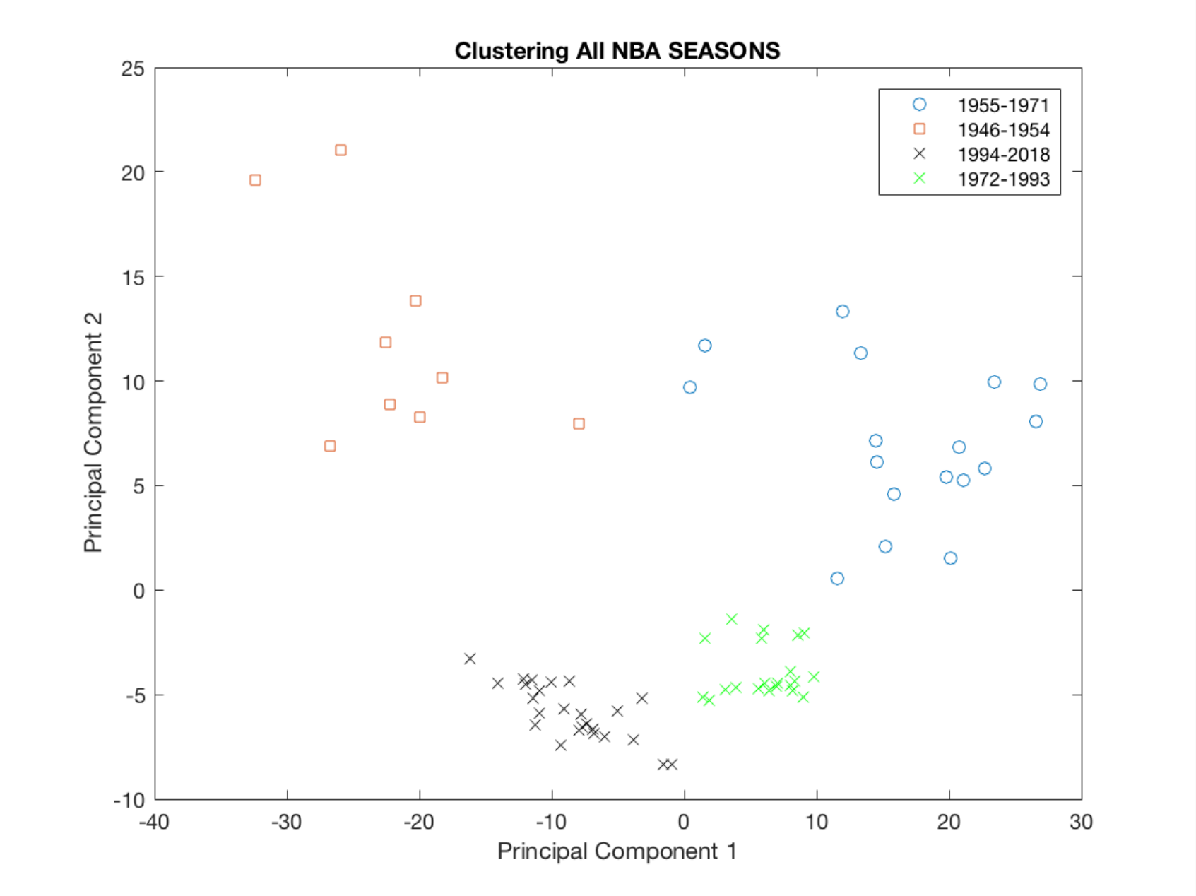
The projected data are jointly Gaussian, for example the qq-plot of the first coordinates is



. I also checked that using a KS test. Using 90% confidence level and performing a χ^2-test we detected the following possible abnormalities (explained in greater detail in the report). Warriors 2016-2017, Rockets 2016-2017, Suns 2016-2017, Mavericks 2016-2017, Warriors 2015-2016, Kings 2014-2015, 76ers 2014-2015.

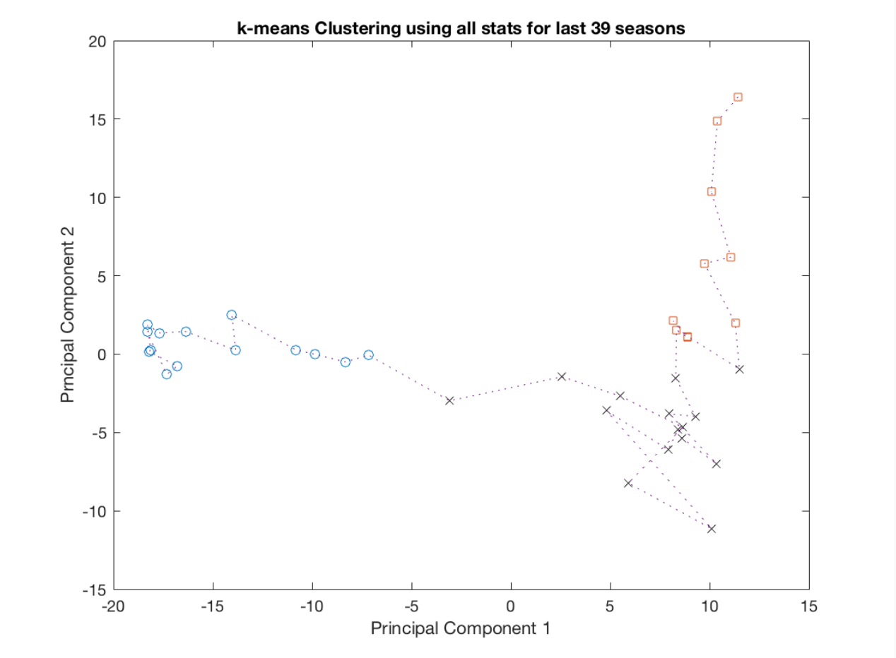
4 Eras of Basketball

We use PCA to reduce the dimension of the league averages statistics to 2 (keeping 92% of the variance). Using a mixture of Gaussians we get the following clusterings. Compating the means of the clusters we can understand the evolution of NBA in the last 70 years.



Last 39 seasons Revisited

The dotted line accounts for time and recent seasons marked as red. We can look at the dynamics in the most recent cluster to understand the way NBA



is currently changing.