Various Machine Learning Approaches to Predicting NBA Score Margins

CS 229 Machine Learning (Fall 2016): Grant Avalon(gavalon), Batuhan Balci(bbalci), Jesus Guzman(guzmanj)

Motivation

We wish to better predict NBA scores for several reasons:
① To understand what features of an NBA team make them successful
② Given a matchup between two teams, we want to figure out which attributes/stats of a team will be critical against the other team.
③ To potentially make wagers on the outcomes. For this, we both wish to pick winners, as well as to pick the margin of each game.

Features & Data

As our features, we use 109 different statistics, such as turnovers, rebounds etc., for each NBA team to predict the outcome of a game between two teams. We also have data on the results of the games between two teams. Both the outcome data and the features data have been retrieved from basketballreference.com and they belong to the 2013-2014 NBA Season.

Summary of Results

We have utilized 3 machine learning algorithms, linear regression, Random Forests and PCA/SVM, to predict the margin of an NBA game between two teams(and Gaussian Discriminant Analysis for classification). PCA/SVM and Random Forest both vastly outperformed Linear Regression in accurately forecasting scoring margins.

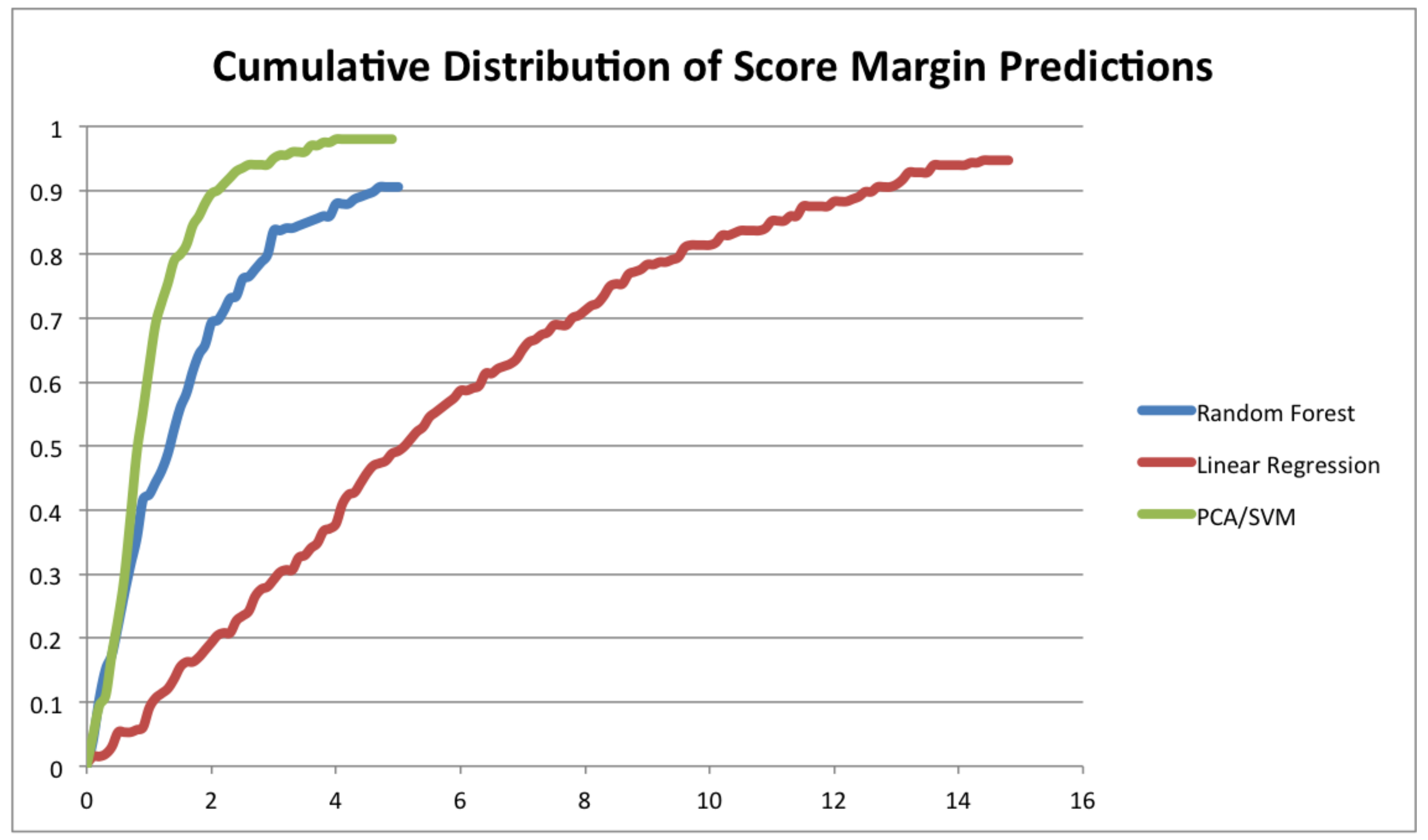

Figure 1: Cumulative distribution graph for various approaches utilized in the project

Linear Regression

- Simple To Implement
- Fast Runtime
- Naïve Benchmark to compare

- Sensitive to Outliers(happens frequently in NBA)
- Doesn't work well with linearly inseparable data
- Assumes data is independent

ⓐOffensive Rating of the Home Team
ⓑ Personal Fouls of the Away Team
ⓒ Defensive Rebounds of the Home Team

$$\hat{Y} = bX + a$$

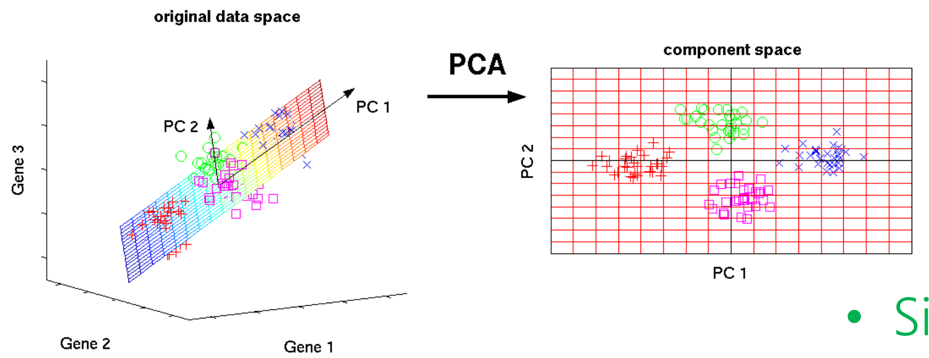

Figure 2: PCA reduces the dimensionality of the data such that there is maximum variance among the samples Source: Johnson Hsieh

Principle Component Analysis (PCA) & Support Vector Machines (SVM)

- Simple To Implement
- Fast Runtime
- Identifies Linear Boundary
- Add one more thing here

- Doesn't scale well
- Sacrifices Resolution
- Low Runtime

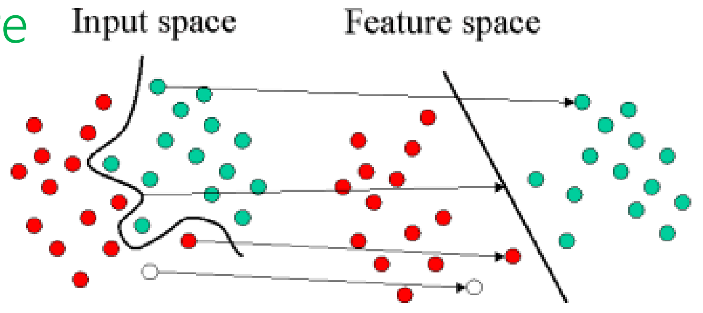

Figure 3: SVM projects linearly inseparable data onto higher dimensions to linearly separate it. Source: Statistica

ⓐ Offensive Rating of the Home Team
ⓑ Number of Turnovers of the Home Team
ⓒ Effective Field Goal Percentage

$$\begin{aligned} \min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, m \end{aligned}$$

Random Forests (RF)

- Efficient on Large Databases
- For Classification and Regression
- De-correlates Features

- Difficult to for Humans to Interpret
- Can Over Fit Data with Noisy Classification or Regression Tasks

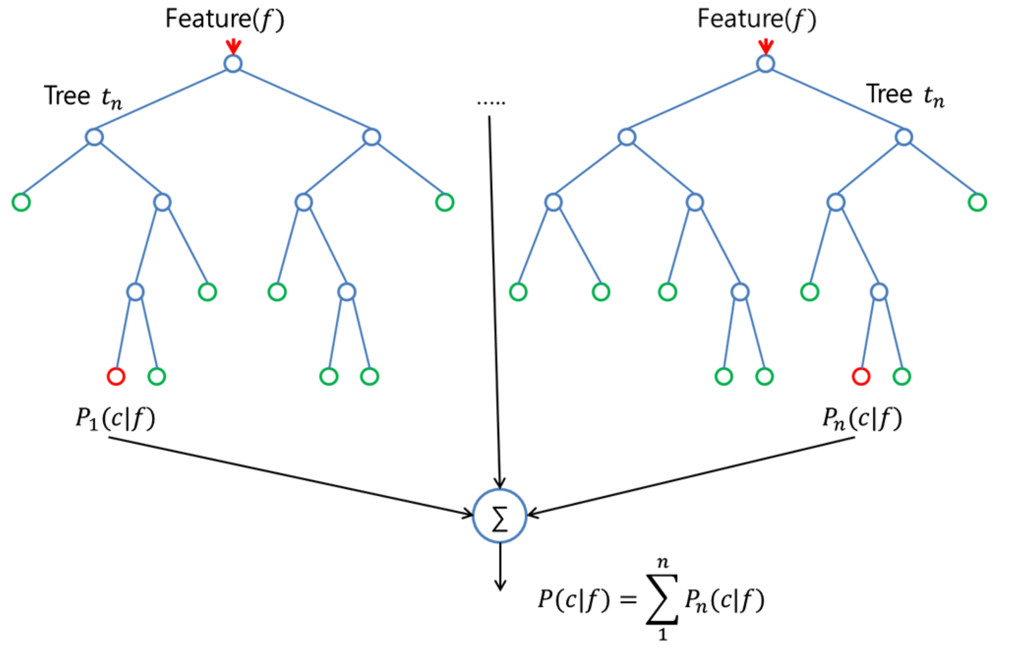

Figure 4: Two decision trees. Each can be thought of as a representation of the training data that is split into subpopulations based on a strong differentiating variable. A random forest algorithm is an ensemble learning algorithm that spawns a lot of decisions trees based on random selection of the training data. Source: stat.Berkeley.edu

Gaussian Discriminant Analysis(GDA)

- Convenient for continuous data
- Works well on normally distributed data

- Can only be used for classification
- The most important features of a team can not be retrieved

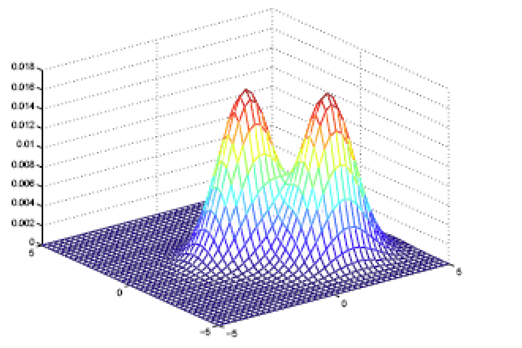

Figure 5: Gaussian Discriminant Analysis assumes that the features given the labels are distributed according to a multivariate Gaussian

$$p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}$$

In addition to the regression task, we also wanted to test the accuracy of our algorithms as classification problems. In other words, we wanted to test how well our algorithms predicted the 'win' or 'loss' outcome of a game. After having trained our models on more than 1052 games, we tested each approach on a set of 264 games, and performed 20-fold cross validation to analyze accuracy, precision and recall.

Approach	Accuracy	Precision	Recall	F1 Score
Linear Regression	64.26%	65.36%	78.52%	71.34%
Gaussian Discriminant Analysis	65.53%	67.90%	73.83%	70.74%
PCA & SVM	61.96%	64.04%	80.22%	71.22%
Random Forest	61.36%	55.71%	35.40%	43.33%

Figure 6: Various evaluation results of the approaches.

Analysis & Future Directions

From our results, we found:

- Each algorithm has niches (similarity, features, texture, etc.)
- PCA scales better on larger datasets, but LDA is more popular than PCA given the accuracy-time tradeoff
- SVM's complement PCA, but too expensive compared to LBP
- LBP has superior accuracy in cases of poor lighting/training sets
- Neural Networks and Gabor Filters are complex and effective, but LBP serves as an effective alternative

Looking ahead, we hope to:

- Implement SVM on top of LDA; Extend LBP to circular LBP
- Train our algorithms for dynamic images (video tracking)
- Combine CNN (feature analysis), Gabor Filters (edge detection), and LBP (texture) into 1 facial recognition algorithm

Stanford University