

---

# Predicting the Outcome of College Basketball Tournament Using a Least-Squares Network Model

---

Yosuke Sugishita

CS490

yosuke.sugi@gmail.com

Henry Li

CS490

henry.s.leigh@gmail.com

## Abstract

Ranking sports teams and predicting their games' outcome is often said to be a difficult problem. In this paper, we propose a simple least squares network model to rank teams in the American college basketball league, and attempt to predict the final tournament's outcome. By training on the past 18 tournaments, we find that our model accurately predicts the outcome of a game in the final tournament with  $70(\pm 4)\%$  accuracy. We first train the model using a basic least square model, and proceed to train the same model using  $\ell_2$ -regularization, which stabilizes our predicted outcome of a new game.

## 1 Introduction

### 1.1 Background Information (Description of Data)

The biggest tournament in the American college basketball league is NCAA Basketball, which is commonly referred to as "March Madness". It is a single-elimination tournament played by 67 teams, meaning there are 66 games' outcome to predict. Roughly 3000 games are played amongst more than 300 teams in our data set every year during the regular season, which we are going to use for training our model.

### 1.2 Objective (Predictive Accuracy)

Our objective is to predict this year's (the 2014 season) tournament's outcome and we aim to maximize the simple predictive accuracy of win/lose. Therefore, we are going to use past years' tournament results as our test data set. The tested accuracy of a model is defined as follows:

$$\text{tested accuracy} = \frac{\text{number of correct predictions}}{\text{number of all predictions}} \quad (1)$$

## 2 Model

### 2.1 Model Construction

We define differences

$$\hat{d}_{ij} = \hat{x}_i - \hat{x}_j$$

where  $\hat{x}_k$  is the 'potential' for a team  $k$ . These differences represent the difference in final score between a pair of teams  $i, j$  given that they have played a game against each other. Now if  $\hat{d}_{ij} > 0$ , we know that team  $x_i$  had won, since  $\hat{x}_i > \hat{x}_j$ . Using this encoding, we can formalize an equation with which we can form predictions. With  $\hat{d}$  as an estimator of the resulting score, we can minimize

the square differences

$$\min_{x_i} \sum_{\text{all games}} (d_{ij} - \hat{d}_{ij})^2$$

### 2.1.1 Training the Model

In training the model, we take the data comprising of games played by 300 teams over the course of 18 seasons. Additionally, in each tournament, there were anywhere from 63 to 67 games. We encode this data as an incidence matrix

$$\mathbf{A} = \begin{bmatrix} a_{t_1 g_1} & a_{t_2 g_1} & \cdots & a_{t_n g_1} \\ a_{t_1 g_2} & a_{t_2 g_2} & \cdots & a_{t_n g_2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{t_1 g_n} & a_{t_2 g_n} & \cdots & a_{t_n g_n} \end{bmatrix}$$

where each column represents one team in a given season and row represents a game played; since a game can only be played between two teams, each row contains exactly 2 non-zero elements. Further, for the two teams involved in some game  $g_k$ , we apply the convention that the ‘first’ team  $t_i$ , that is the team whose column appears before the ‘second’ team  $t_j$  such that  $i < j$ , obtains the value  $a_{t_i g_k} = 1$  for the element while  $a_{t_j g_k} = -1$ . Not surprisingly this forms the machinery needed to perform many calculations of  $\hat{d}$  and thus to assign potentials to each team.

But to do this, we need to solve a system involving the actual scores for games within some tournament within the data. Let  $\mathbf{s}$  be the vector of differences between all final scores of each game in the tournament. Then to determine the potentials for each team, is a matter of solving the system

$$\mathbf{A}\mathbf{x} = \mathbf{s}$$

As a simple example with 4 teams and 3 games (we assume that teams 1 and 3 are carried forth onto the second round), we have:

$$\begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 1 & 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 32 \\ 20 \\ -2 \end{bmatrix}$$

In this case, solving for  $\mathbf{x}$  would return the potentials for each of the 4 teams.

### 2.1.2 Testing the Prediction Accuracy

As mentioned in Section 1.2. The accuracy of a model is tested as follows:

$$\text{tested accuracy} = \frac{\text{number of correct predictions}}{\text{number of all predictions}} \quad (2)$$

The accuracy is tested on past years’ tournament results, 18 years worth, after training the model on the same year’s regular season results. As there are 66 games in each tournament, resulting our test data to contain more than 1000 games to predict.

## 2.2 Confidence Interval Analysis

To construct the confidence interval for the tested accuracy, we assume the following for convenience sake:

$$Y_i \sim \text{Bernoulli}(\pi) \text{ (i.i.d)} \quad (3)$$

where:

$$\begin{aligned} Y_i &= 1 \text{ if the prediction for game } i \text{ is correct} \\ Y_i &= 0 \text{ if the prediction for game } i \text{ is incorrect} \end{aligned} \quad (4)$$

Under this assumption, the calculated accuracy ( $\bar{Y}$  = mean of  $Y_i$ ’s) is approximately normally distributed when the sample size is large by Central Limit Theorem.

$$\bar{Y} \sim N(E(Y_i), \text{Var}(Y_i)/n) = N(\pi, \pi(1 - \pi)/n) \quad (5)$$

Regularization parameter	Average standard deviation of potential estimates
0 (no regularization)	1.351e+15
$10^{-12}$	5.758
$10^{-9}$	5.758
0.001	5.757

Table 1: Average standard deviation of potential estimates

$\pi$  is unknown, but the lower bound for  $Var(\bar{Y}) = \pi(1 - \pi)/n$  can be found as it is maximized when  $\pi = 0.5$ .

$$Var(\bar{Y}) = \pi(1 - \pi)/n \leq 0.5(1 - 0.5)/n = 1/4n \quad (6)$$

We used this value to construct our 99% confidence interval.

$$CI(\bar{Y}) = CI(\hat{\pi}) = \hat{\pi} \pm z_{0.995} * \frac{1}{2\sqrt{n}} \quad (7)$$

It turned out, with our sample size (roughly 1200), we have:

$$\text{one-sided length of confidence interval} = z_{0.995} * \frac{1}{2\sqrt{n}} \leq 0.04 \quad (8)$$

Note that this does not depend on  $\pi$  or  $\hat{\pi}$  any more. We chose to test our model on the past tournament results partly because we were able to obtain narrow enough confidence interval this way.

### 2.3 Use of $\ell_2$ -Regularization

## 3 Results

### 3.1 Predictive Accuracy

We achieved roughly 70( $\pm 4$ )% accuracy for predicting the past 18 tournaments' outcome (for win/lose prediction). This is an out-of-sample result as we trained our model on the same season's regular game results, which preceded the tournament.

### 3.2 Regularization (Irregular Behaviour in the Results)

We noticed that differences between our potential estimates are sometimes unusually high or low (for instance, in the order of  $10^9$ ). One can see this effect easily from the average standard deviation of potential estimates for each year (Table 1).

As the potential differences are used as the predicted score difference of a game, this is clearly an irregular result. Also, as we examined the plot of predicted score differences against actual scores, we realized that the predictions can be unstable particularly when the actual score differences are small (when the games are "close").

To prevent these unusual behaviors and also to stabilize the predictions, we utilized  $\ell_2$ -regularization. The model is described in Section 2.3:  $\ell_2$ -Regularization. As expected, this model yielded more stable predictions as shown in Figure 1. This was also clear when we compared the correlation coefficients between actual score differences and predicted score differences (it improved from 0.456 to 0.468. However, it did not change the predictive accuracy significantly.

Interestingly, ranging the regularization parameter from  $10^{-9}$  to 0.001 hardly changed the results as can be seen from Figure 1.

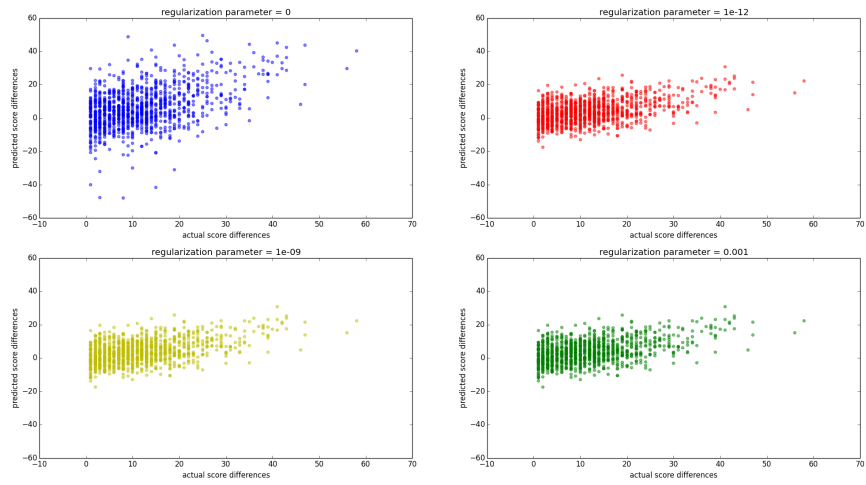


Figure 1: Predicted score differences against actual score differences

## 4 Conclusion

### 4.1 Summary

### 4.2 Further Considerations