2014 World Cup Recap

Weichen Ning, Wen Bo
Department of Electrical and Computer Engineering
Duke University

Abstract

This is the course project of Duke STA 561 Probabilistic Machine Learning (Fall 2015). We developed a Bayesian approach to predict the 2014 FIFA World Cup (especially focus on elimination games), by applying probit regression and Bayesian linear regression model with data augmentation. We selected five relevant variables like shoots on target, pass accuracy, procession rate, FIFA rank and having superstar or not of both team to determine the match result. For our model, we develop the Gibbs sampling to perform the posterior inference. We apply our algorithms on the different datasets and the results show that we could also obtain high accurate predictions.

1 Introduction

In 2010, we had Paul the Octopus to predict the results of the 2010 World Cup games. But in real life, we planned to build a model which was more convincible and precise. For the past a few weeks, we have been building Bayesian regression model and profit model to make predictions for the 2014 World Cup. By analyzing data and building a statistical models, as well as using machine learning techniques to predict outcomes of each match of elimination games, we have gotten 11 out of 16 games correct. We are giving you the keys to our prediction model so you can even build your own model and run your own predictions.

We develop a Bayesian approach to predict the 2014 FIFA World Cup (especially focus on elimination games), by applying probit regression and Bayesian linear regression model with data augmentation. For our model, we develop the Gibbs sampler to perform the posterior inference. We apply our algorithms on the different datasets and the results show that we could obtain high accurate predictions.

In our project, we aim to figure out how the use of a probabilistic model might be able to predict the results for the 2014 World Cup and compare our predictions with the real results. By treating each game played as a pairwise comparison experiment, we use Bayesian linear regression model to fit the data. Various variables have been developed to handle bias, shoots on target, FIFA rank, possession rate, passing accuracy, having superstar or not. After that, we develop the profit model to link the result to binary outcomes.

Also, we need clean data to get our model trained. We select five relevant variables of each team such as shoots on target, possession rate, passing accuracy, FIFA rank, the team owns superstar or not as well as bias. We collected the game statistics of 100 elimination games of the 1998, 2002, 2006, 2010 World Cup, 2000, 2004, 2008, 2012 UEFA European Championship and 2001, 2004, 2007, 2011, 2015 Copa America into the training set. It's hard to get the detailed game reports of the previous ones because the data at that time was not so detailed. On the other side, the data in the past a few years are more relevant to our prediction in 2014. The graphical model of our inference is shown in Figure 1.

After collecting the data we are interested in, we calculate the input as the average of all the historical game statistics of the two teams into the training weights we have calculated.

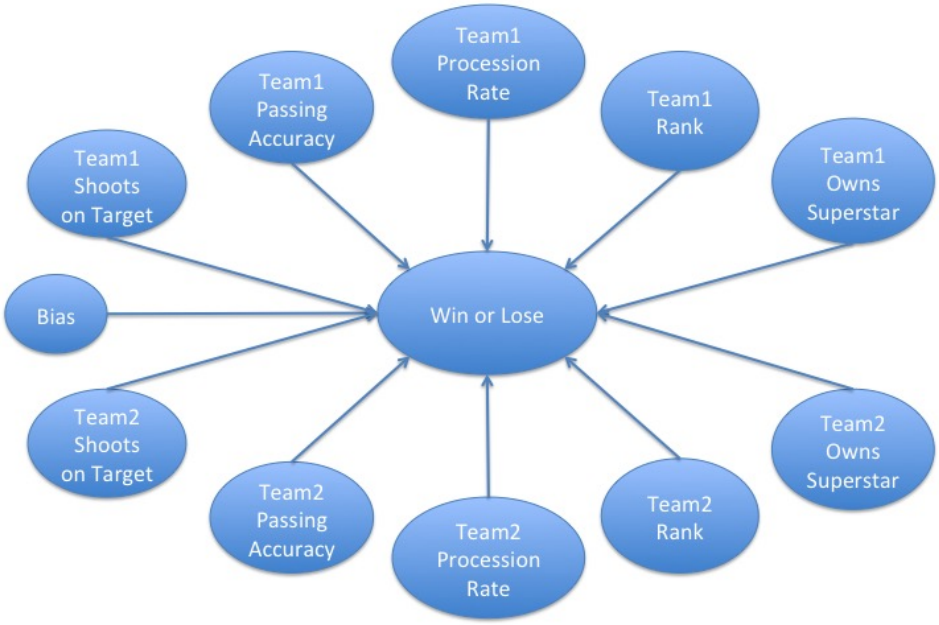

Figure 1: Graphical Model

2 Model

We aim to establish the relationship between match results $y = [y_1, y_2, \ldots, y_n]^T \in \mathbb{Z}^n$ and team statistics $X = [x_1, x_2, \ldots, x_p] \in \mathbb{R}^{n \times p}$ whose columns correspond to specific statistics. Since y is the binary data, we propose to use the following data augmentation approach to model the data: For

$i = 1, 2, \ldots, n,$

$$z_i \quad \sim \quad \mathcal{N}(z_i;\, X_i^T\beta, 1) \tag{1}$$

where z_i is the latent variable; $y_i = 1 \; if \; z_i > 0 \; and \; y_i = 0 \; otherwise$. Moreover, we assign the following prior on β:

$$\pi(\beta) \quad \sim \quad \mathcal{N}(\beta; \beta_0,\, \Sigma_0) \tag{2}$$

3 Posterior Inference

We could derive the Gibbs sampler for Bayesian linear regression as follows: The fitted parameters $\hat{\beta}$ should reflect the corresponding weights for each match statistic. We could employ the fitted $\hat{\beta}$ to accomplish prediction using new dataset $\tilde{X}$ as

$$\tilde{z} = \tilde{X}\hat{\beta} \tag{3}$$

A Gibbs sampler for Bayesian Linear Regression:

Sampling β from

$$(\beta|-) \quad \sim \quad \mathcal{N}(\beta; \hat{\beta},\, \hat{\Sigma}) \tag{4}$$

where

$$\begin{aligned} \hat{\beta} \quad &= \quad \hat{\Sigma_0}\,(\Sigma^{-1}\beta_0 + X^T z) \\ \hat{\Sigma} \quad &= \quad (X^T X + \Sigma_0^{-1})^{-1} \end{aligned}$$

Sampling z_i for $i = 1, 2, \ldots, n$ from

$$(z_i|y_i = 0) \quad \sim \quad \mathcal{N}(z_i; X_i^T\beta, 1)\{z_i \leq 0\} \tag{5}$$

$$(z_i|y_i = 1) \quad \sim \quad \mathcal{N}(z_i; X_i^T\beta, 1)\{z_i > 0\} \tag{6}$$

4 Experimental Results

We select five relevant variables of each team such as shoots on target, possession rate, passing accuracy, FIFA rank, the team owns superstar or not in addition to bias. We collected the game statistics of 100 elimination games of the 1998, 2002, 2006, 2010 World Cup, 2000, 2004, 2008, 2012 UEFA European Championship and 2001, 2004, 2007, 2011, 2015 Copa America into the training set. We calculate the input as the average of the historical game statistics of the two teams into the training weights we have calculated. Among the variables, we select the FIFA rank as the most influential variable that decide our prediction. The autocorrelation function and the distribution of FIFA rank are as follows.

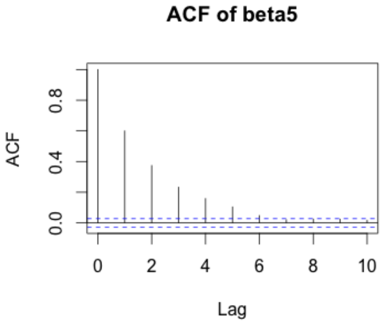

Figure 2: Autocorrelation Function

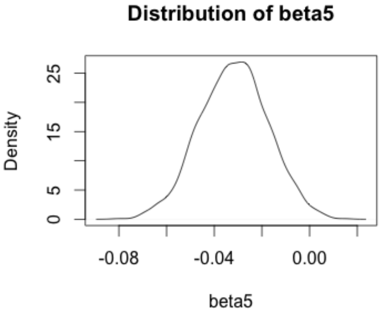

Figure 3: Density Distribution

The final results are as follows. In the game of Columbia vs Uruguay, Belgium vs USA, France vs Germany, Brazil vs Germany and Brazil vs Netherlands, we did wrong predictions according to the real data. Interestingly, some teams even win the game with the disadvantage in all fields.

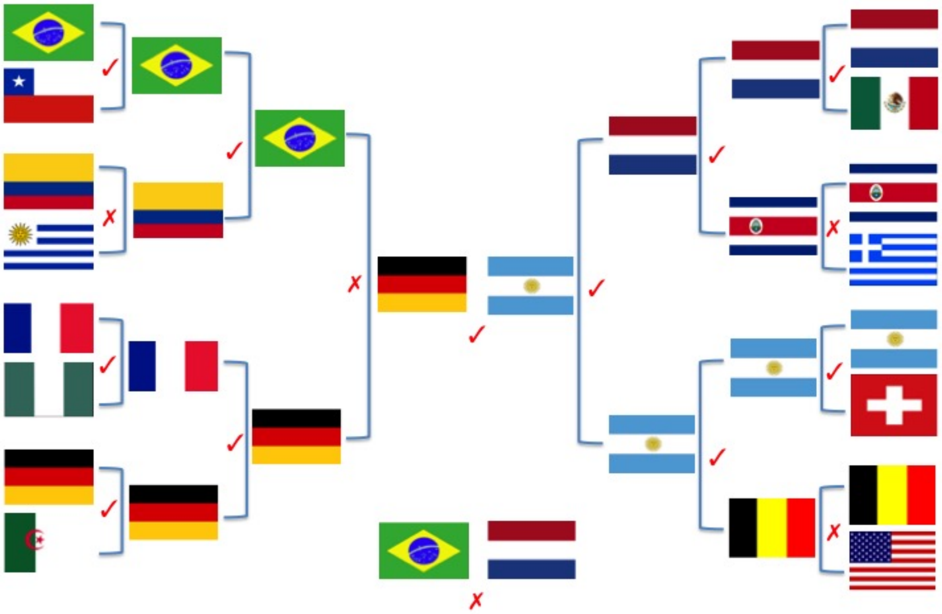

Figure 4: Comparison with 2014 FIFA World Cup Bracket

Correct	Wrong	Accuracy
11	5	68.75%

Table 1: Prediction Outcome

5 Further Thought

We plan to apply some non-linear algorithms to make the prediction, like non-linear SVM and Gaussian process regression models to see if these alternatives can do better than Bayesian linear regression by comparison.

In addition, if we analyze the five games we predicted wrong, we conclude that some team has better offensive statistics while they have worse defensive statistics than the other team. Also, the variables we selected from the game report are all about offense rather than defense. We think this is more likely to be the reason that we predicted wrong. Especially in the elimination games of World Cup, defense sometimes carries more weight than offense.

Furthermore, there are so many variables that we cannot model. For example, in the quarterfinal, Brazil 1:7 Germany and lost to Netherlands with 0:3 in the third place game. Brazil encountered a very depressing result in the game with Germany and obviously lost the motivation to play the third place game. These two games were out of expectation because Brazil have larger advantage in all aspects. There are also situations with sudden accidents such as red card, penalty kick and severe injuries. However, maybe that is also the beauty of the World Cup. You cannot explicitly model all the variables such as spiritually variables and sudden accident variables of the whole game.

6 Conclusion

In the game of Columbia vs Uruguay, Belgium vs USA, France vs Germany, Brazil vs Germany and Brazil vs Netherlands, we did wrong predictions according to the real game result. Interestingly, some teams even win the game with the disadvantage in all fields. In general, we made 11 right predictions out of 16 elimination games.

This model introduces latent parameter for making prediction. In our model, we establish the relationship between match statistics in each soccer team and the result of a match. For our method, we implement the Gibbs sampling to perform the posterior inference. The experimental results tell us that the models can fit the data well and provide the reasonable prediction.

7 Acknowledgement

We would like to express our special thanks of gratitude to our instructor (Sayan Mukherjee) who gave us the opportunity to do this interesting project of the World Cup topic as well as many support and help in the process of our project. This project combined the passion of two soccer fans and machine learning techniques. In the process of exploration, the project helped us have a deep understanding of machine learning inferences on real-world situations.

References

[1] Alexander Spermann. *The Probit Model, University of Freiburg*. Sose, 2009.

[2] Peter D Hoff. *A first course in Bayesian statistical methods*. Springer, 2009.