Generalised linear model for football matches prediction

Antoine Adam

KULeuven, Departement of Computer Science, Celestijnenlaan 200A, 3001 Leuven, Belgium

Abstract. This paper presents the method we used in the prediction challenge organised by the Sports Analytics Lab of the KU Leuven for the European football(soccer) championship. We built a generalised linear model to predict the score of a match. This score was modelled as the joint probability of a Poisson distribution, representing the total number of goals, and a binomial distribution, representing the goals of one team given that total number of goals. This model was trained on the matches of the past year using gradient descent to maximise the log-likelihood with l2 regularisation. Special care was taken to construct a model that is symmetrical and does not involve any home advantage, with the exception of the host team. The features considered were both team-based and player-based, using a randomised approach to select the players based on their past selections. A simulation of the tournament was then built on this match model to predict how far each team would go in the tournament.

1 Introduction

Prediction of football (soccer) matches is becoming more and more popular. On the occasion of European football championship, the Sports Analytic Lab of the KU Leuven organised a prediction challenge which consisted of two parts. The first challenge was to predict the outcome of matches between the 24 teams of the tournament. The prediction consisted in giving the probability for each team to win, lose or draw against any other team. The second challenge was to predict how far each team would go in the tournament. The prediction consisted in giving the probability for each team to be eliminated in the group phase, round of 16, quarter final, semi final, final or to win the tournament.

In order to simulate the tournament for this second part, a model that gives the full score of a match, and not only the winner, was needed. Traditional approaches for that task predict the goals scored by the two teams. We propose to introduce an intermediate random variable representing the total number of goals and then, given that number, predict how many were scored by each team. As a team consists of a selection of players, features were built on characteristics of players and not only on those of the team, using a selection process among the players of a team.

In the next section, the probabilistic model for predicting the match outcome is explained, as well as how it was learned. The feature construction is then explained in section 3. Finally, the performance of the model in the challenge is reported in the evaluation section.

2 Probabilistic model

The goal of the first challenge was to predict match outcomes as the probability for one team of winning, loosing or drawing against any other team of the tournament. To do so, we choose to model the probability distribution of the score of a match.

2.1 Model definition

An intuitive idea would be to represent the number of goals of each team by two Poisson distributions. However, as we do not want to consider these two variables as independent, the combination of the two Poisson distribution is not trivial. What is more, we want our model to be symmetrical: the match team A vs team B should produce the same result as team B vs team A. A possible solution is to use a bivariate Poisson distribution [5]. Instead, we choose to model the total number of goals in a match as a Poisson distribution. Then, given that number of goals, the number of goals of each team can be modelled as a binomial distribution. Each team is described by a number of features that will be explained in the next section. Thus for a specific match, we have a feature vector X consisting of the features of both teams. Following the idea of Generalised Linear Models, the parameters of the distributions are formulated as the composition of a linear regression and an activation function. If we call g_i the number of goals scored by team i, the probability distribution of the score of a match between team A and team B is then defined as:

$$\begin{aligned} P(g_A, g_B|X) &= P(g|X) \times P(g_A|g, X) \\ &= \mathcal{P}oisson((g|\lambda(X)) \times \mathcal{B}inomial((g_A|n = g, p(X)) \end{aligned}$$

with

$$g = g_A + g_B$$

$$\lambda(X) = exp(U^T X + u_0)$$

$$p(X) = \frac{1}{1 + exp(-(V^T X + v_0))}$$

The symmetry of the model for the matches A vs B or B vs A can then be guaranteed by ensuring that $\lambda_{AvsB} = \lambda_{BvsA}$ and $p_{AvsB} = 1 - p_{BvsA}$. The vectors U and V are the coefficients of the linear regression for parameters λ and p, while u_0 and v_0 are the intercepts. The choice of the activation functions, exponential and sigmoid, was guided by the desired value range of λ and p.

To learn U, u_0, V and v_0, we built a training set based on the matches of the past year, competitive and friendly, between the 24 teams participating in the tournament. Unlike during this one, those matches were not played on neutral ground but at home for one of the two teams. To remove this bias, we actually produced 2 symmetrical examples from each past match. For example, the match France-Belgium on June 7th 2015 that Belgium won 4-3 produces 2 training examples X, g, g_A:

- (features of France), (features of Belgium), 7, 4
- (features of Belgium), (features of France), 7, 3

This training set being symmetrical, so was the learned model. In particular, v_0, which represents the prior of the home advantage, was always pretty close to 0.

The training set consisted of $M = 62$ matches. We use a gradient descent to maximise the loglikelihood with l2 regularisation to avoid overfitting. The function to maximise is the following:

$$\sum_{k=1}^{2M} log(P(g_{Ak}, g_{Bk}|X_k)) - \alpha \times (\|U\|_2^2 + \|V\|_2^2)$$

with α the regularisation coefficient.

The exception of France. As the tournament was held in France, the French team was an exception to this problem of symmetry of the model. To take into account the home advantage for France, a second model was trained, identical to the first one in its structure. However, the training set was only built on examples putting the home country as the first country. In the previously given example, this implies that only the first training example was used.

2.2 Simulations

The learned model can then be used to run simulations of matches. Given two teams, we first build our feature vector and compute the parameters of the distribution λ and p. We can then sample the total number of goals and given that, the goals of each team.

For the first challenge, we needed to predict for each match the probability of either one team winning, the other team winning, or having a draw between the 2 teams. To get these probabilities, we simply sampled ten thousand matches and counted the outcomes.

For the second challenge however, we needed to predict the probability for each team to reach each phase of the tournament. To evaluate these, we built a simulation of the whole tournament. As in the final phase, a match cannot end up in a draw, we adopted the following method for that phase:

1. Sample the score for regular time as before.
2. If it is a draw, sample again the total number of goal and divide it by 3, as the extra time is three times shorter than the regular time. Then, given that new number, sample the score as before using the binomial distribution.

3. If it is still a draw, randomly select a team as winner, as penalty shout-outs are random enough.

Using this method, we ran ten thousand simulations and counted the results.

These simulations approximate the probabilities, but the lack of time prevented the implementation of exact computations. This one is however not trivial because of the way the player dependant feature are built, which will be explained in the next section.

3 Feature construction

We will now explain how we built the feature vector used in the model, designated by X in the previous section. The features can be divided in two categories: team-based and player-based. While team-based features are characteristics of a team, player-based features are aggregates of characteristics of player of a team. Here is a detailed list of the features.

- Team-based:
  fifarank = FIFA rank in June 2016[2]
  fifatrend = (FIFA rank in June 2016) - (FIFA rank in January 2016)
  uefarank = UEFA rank in February 2016[1]
  elorank = ELO rank in February 2016[1]
- Player-based:
  barometer = 101 - (position[2] in UEFA barometer[3] on the 10th of June 2016)
  goals = Number of goal in national team in the whole career[1]
  value = Transfermarkt value[1]

Due to the lack of time, only the average was used as an aggregate on the player-based features.

In football as in other team sports, the players of one team vary over different matches. This is something we wanted to include in the model, for instance to take into account injured players that would not participate in the tournament. Also, some players play more often than others. For past matches, we actually know which players played and the average was weighted by their time on the pitch. For simulation matches, we draw 11 players for each team from the 23 selected for the tournament. The drawing was done with a roulette strategy with weights based on the past selections of players. To avoid selecting too much players of the same post (like 2 goalkeepers), the 11 players were actually sampled as 1 goalkeeper, 4 defenders, 4 midfielders and 2 attackers. This composition could be improved by making it more team-specific.

A main drawback of the features is that these were fixed, either in training or simulations, apart from the player selection strategy. Thus, some features, like

[1] As provided by the challenge organisers.
[2] The position is considered equal to 101 if the player does not appear on the barometer.

the FIFA ranking, were partially built on the results of the training matches, which probably overestimated their influence on the score. It would be easy to adapt the features to correct this bias. For instance, instead of the fifarank in June, use the fifarank before each match to build the training set. This was not done again by a lack of time. For the tournament prediction, these features could also be updated during the simulation. While this would be easy for team-based features, player-based ones would be more tricky: the exact algorithm of the UEFA barometer is not public, and the model does not predict which player scored the goals of a team. Also, the tournament taking place in a relatively short period of time, we consider updating these features during that period less crucial than in the training examples.

4 Evaluation

The model was evaluated in the context of the prediction challenge. For each challenge, the participants were scored using the multi-class logarithmic loss [4]. At the time the model was submitted, we used a random approach to select the home team when building the training set instead of duplicating the example. This however always resulted in an asymmetric dataset leading to an advantage for the home or away team. As a result, the submitted prediction is biased. The results we present here include the scores of the correct model as presented in this paper.

Figure 1 shows the results of the challenge of all participants (black lines) and of the proposed model for different values of the regularisation parameter *alpha*. The performance of 3 models is shown: one using only the team attributes, a second one only the team attributes and a final one using all of them. As just explained, the submitted prediction (in purple) has a error bias. For that submission, *alpha* was set to 10. We can see the regularisation improves the result by reducing the overfitting that probably occurs due to the small size of the training set. It should also reduce the bias introduced by not updating the features in the training set. The combination of both set of attributes is the best for challenge two while using only the players attributes seems to perform better for the first challenge.

5 Conclusion

We presented a new model for predicting football matches. The new idea is to first predict the total number of goals and then given that number, which goals were scored by which team. The model is built on both player-based and team-based features. This model performed decently in the prediction challenge organised by the Sports Analytics Lab of the KU Leuven.

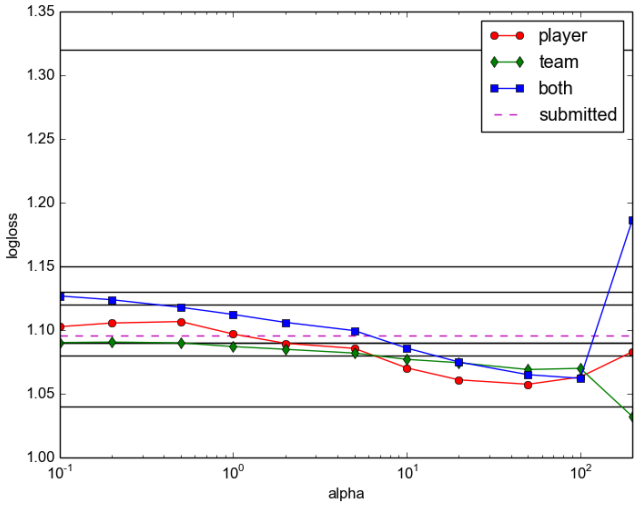
Several improvements could be brought to the model. Preserving the structure of first predicting the total number of goals and then how many are of each team, other distribution might be more fit than Poisson and Binomial. A Poisson distribution represents the number of occurrences of independent events and

goals can hardly be called independent. Although detailed experiments were not run, the variance of the Binomial distribution seemed slightly too high.

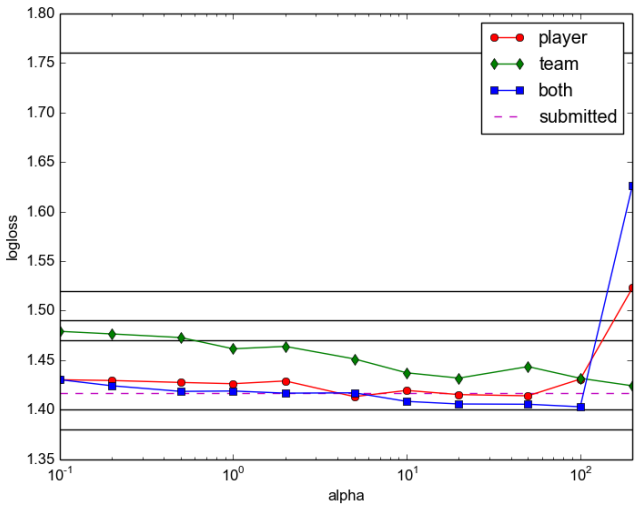
Additional features, like assists or passes, would also help to improve the model. Applying other aggregates than the average for player-based features could also be interesting. The multiplicity of features would however require some feature selection process.

References

1. McCullagh, P., & Nelder, J. A. (1989). Generalized linear models (Vol. 37). CRC press.
2. FIFA ranking. http://www.fifa.com/fifa-world-ranking/index.html.
3. UEFA Barometer (2016). http://www.uefa.com/uefaeuro/season=2016/players/faq/index.html.
4. Multi-class logarithmic loss. https://www.kaggle.com/wiki/MultiClassLogLoss.
5. Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. Journal of the Royal Statistical Society: Series D (The Statistician), 52(3), 381-393.



(a) Challenge 1: match outcome



(b) Challenge 2: tournament

Fig. 1: Results of prediction challenges. The lower the logarithmic loss, the better the prediction.