Generalised linear model for football forecasting

Antoine Adam

MLSA workshop

Riva del Garda

19/09/2016

KU LEUVEN

Outline

- Probabilistic model
- Player-based and team-based features
- Performance

Probabilistic model

Problem

- Challenge 1: predict winner
- Challenge 2: tournament
  => score needed for group phase
- Goal: predict the score of a match

Football match model

- Team A vs Team B
- Score: G_A G_B

Football match model

- Team A vs Team B
- Score: G_A G_B

- $G_A \rightarrow \mathcal{P}(\lambda_A)$ $G_B \rightarrow \mathcal{P}(\lambda_B)$ Poisson

Football match model

- Team A vs Team B
- Score: G_A G_B

- $G_A \rightarrow \mathcal{P}(\lambda_A)$ $G_B \rightarrow \mathcal{P}(\lambda_B)$ Poisson

- Problem: NOT INDEPENDANT

Football match model

- Team A vs Team B
- Score: G_A G_B

- $G = G_A + G_B \rightarrow \mathcal{P}(\lambda)$ Poisson

Football match model

- Team A vs Team B
- Score: $G_A \qquad G_B$

- $G = G_A + G_B \rightarrow \mathcal{P}(\lambda)$ Poisson
- $G_A \rightarrow \mathcal{B}(p, G) \qquad G_B = G - G_A$ Binomial

Football match model

- Team A vs Team B
- Score: $G_A \quad G_B$

- $G = G_A + G_B \rightarrow \mathcal{P}(\lambda)$ Poisson
- $G_A \rightarrow \mathcal{B}(p, G) \quad G_B = G - G_A$ Binomial

- λ and p depend of the features vector X

Parameters

- Generalised linear model
- Linear regression + activation function

Parameters

- Generalised linear model

- Linear regression + activation function

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

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

Parameters

- Generalised linear model

- Linear regression + activation function

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

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

Parameters

- Generalised linear model

- Linear regression + activation function

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

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

Training the model

- Dataset: M = 62 matches from june 2015 until june 2016

- Gradient descent to maximise loglikelihood with l2 regularisation

$$\sum_{k=1}^{M} \log(\mathcal{P}(g_k \mid \lambda(X)) \times \mathcal{B}(g_{A,k} \mid g_k, p(X))) - \alpha \times (\|U\|_2^2 + \|V\|_2^2)$$

Training the model

- Dataset: M = 62 matches from june 2015 until june 2016

- Gradient descent to maximise loglikelihood with l2 regularisation

$$\sum_{k=1}^{M} \log(\mathcal{P}(g_k \mid \lambda(X)) \times \mathcal{B}(g_{A,k} \mid g_k, p(X))) - \alpha \times (\|U\|_2^2 + \|V\|_2^2)$$

Training the model

- Dataset: M = 62 matches from june 2015 until june 2016

- Gradient descent to maximise loglikelihood with l2 regularisation

$$\sum_{k=1}^{M} \log(\mathcal{P}(g_k \mid \lambda(X)) \times \mathcal{B}(g_{A,k} \mid g_k, p(X))) - \alpha \times (\|U\|_2^2 + \|V\|_2^2)$$

Feature vector

List of features

- Team-based:
  - Fifa rank
  - Fifa trend
  - Eufa rank
  - Elo rank
- Player-based:
  - Eufa barometer
  - Goals scored
  - Transfermarkt value

Player-based feature

- Aggregate: average
- For training set:
  - weighted by time on the pitch
- When simulating a new match:
  - sample 11 players based on their past selections

Using the model

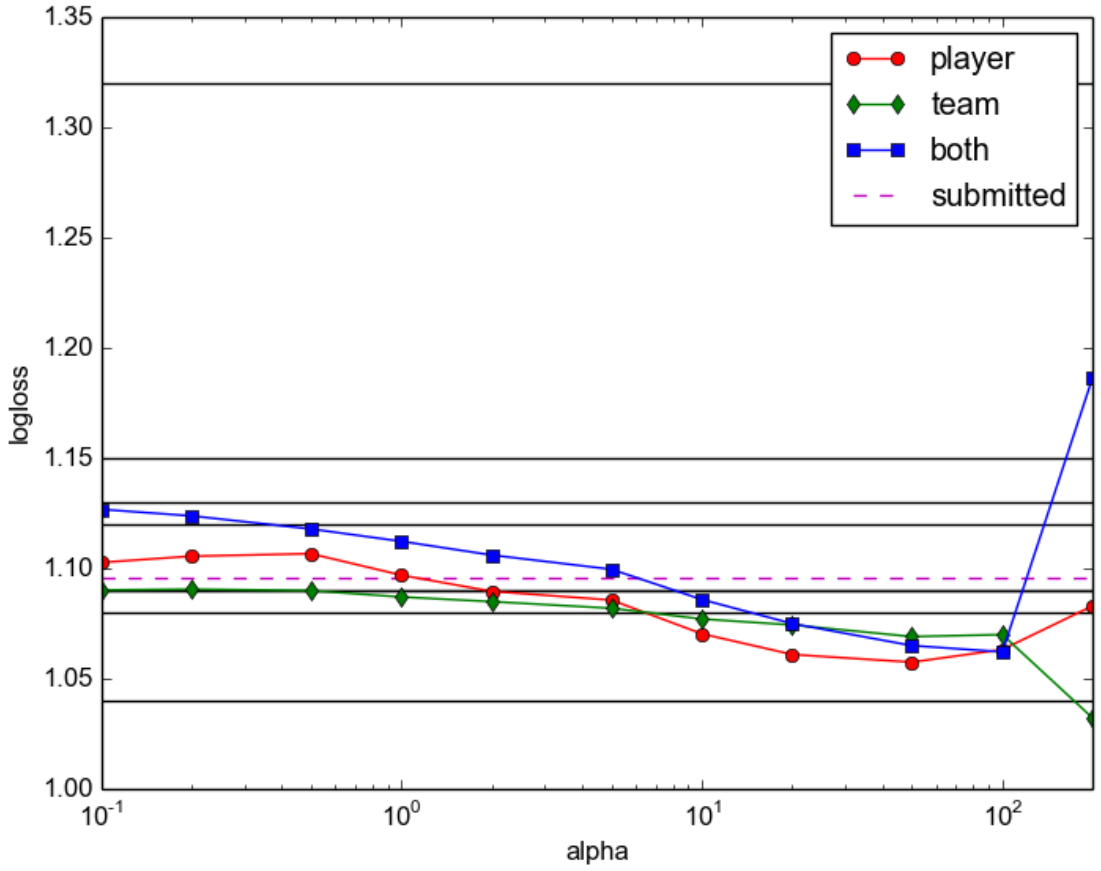
- Match simulation:
  - Compute feature vector X
  - Compute λ and p
  - Sample from Poisson and Binomial

- Challenge 1: sample 10 000 matches

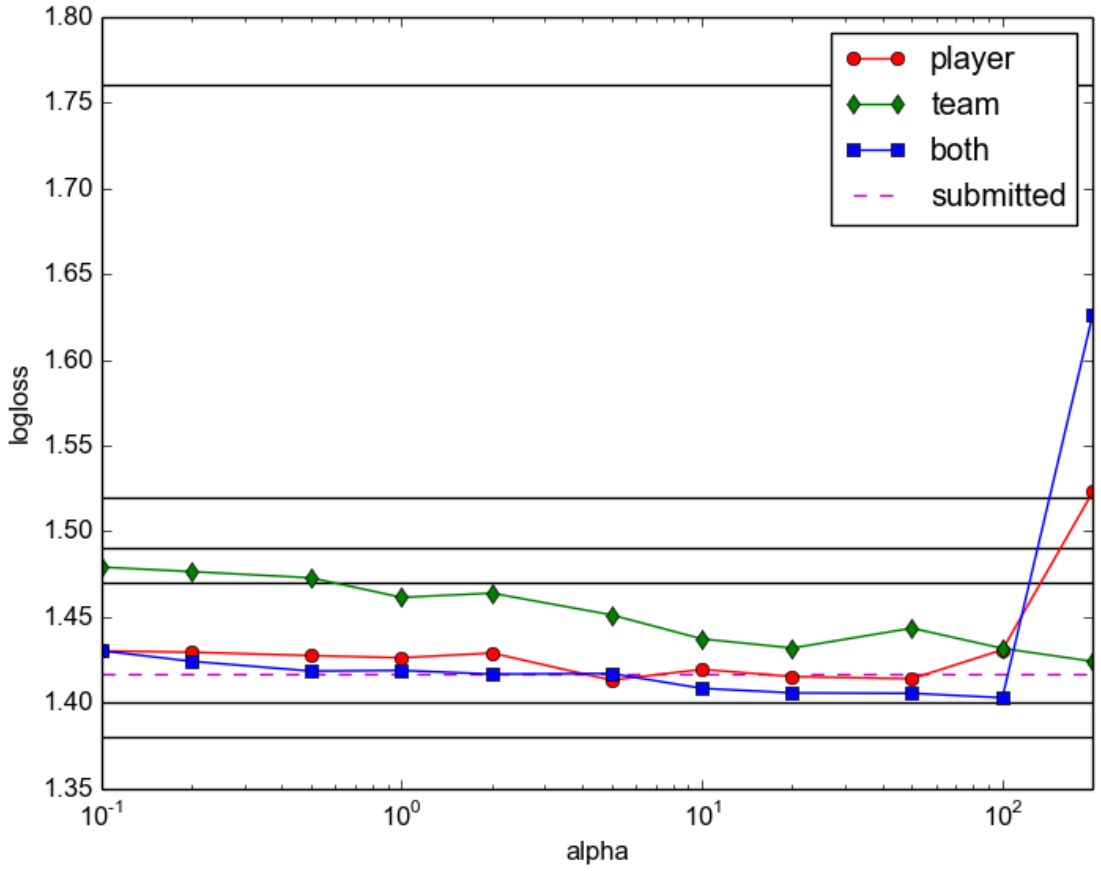
- Challenge 2: simulate 10 000 tournaments

Evaluation

Challenge1


1.35
1.30
1.25
1.20
1.15
1.10
1.05
1.00
logloss
10^{-1}
10^{0}
10^{1}
10^{2}
alpha
player
team
both
submitted

Challenge2


1.80
1.75
1.70
1.65
1.60
1.55
1.50
1.45
1.40
1.35
logloss
10^{-1}
10^{0}
10^{1}
10^{2}
alpha
player
team
both
submitted

Conclusion

- Generalised linear model
  - Poisson and binomial distribution
- Combination of player-based and team-based features