Premier League via Dirichilet Regression

Nascimento, Igor Ferreira do
Departamento de Administração - UnB; Laboratório de Aprendizado de Máquina em Finanças

Resumo

O artigo apresenta uma metodologia para estimar a probabilidades dos resultados dos jogos da Premier League utilizando covariáveis dos jogadores.

Keywords: Regularization,Dirichilet Regression,Premier League,Forecast

Introdução

O fascínio por eventos incertos instiga o ser humano desde os remotos tempos dos antepassados. Os historiadores XXXX mostram que desde a época o ser humano demonstrava interesse por saber ao menos algo sobre o que é incerto. Ao longo dos anos, a teoria dos jogos proporcionou o desenvolvimento de modelos probabilísticos para determinar o nível de determinados ocorrências.

O objetivo do trabalho é apresentar modelos probabilísticos para representar os resultados da Premier League por meio de covariáveis relacionado ao capital humano do time e os recursos financeiros relacionados ao time. Além disso, é apresentado uma estimação para o resultado final da tabela de classificação.

Sports Forecast

Sport is a human activity normally related to the playful environment but we can not dissemble the betting. The work like (?, ?) analyse the dissonance between the outcome of sports and the sports betting market explained by fact the set betting lines is straight related with the profit goal instead the outcome probability. On the other hand, the paper (?, ?) analysed some European soccer leagues and show that exist evidence that bookmakers learn over time about soccer prediction results. Anyway, this is interesting to either bettors or bookmakers researchers in this area.

Dito isso, pode-se destacar que os resultados esportivos possuem uma intenso interesse da sociedade como um toda, seja ela do ponto de vista financeiro ou a mera busca por entreternimento.

As redes de apostas como, XXX, YYY, GGG SÃO encontrados ao redor do mundo e fazem girar cifras ainda maiores de recursos no ato de acreditar em um determinado resultado. Esportes como o baiseball e o futebol americano possuem grande suporte dos cientistas de dados para apresentar pequenas melhoras nos resultados individuais dos atletas que aparentam ser inespressíveis, ao menos aos olhos dos descuidados, porém representam um grande impacto no coletivo e consequentemente, em toda a cadeia produtiva associada ao esporte. De encontro com tais anseios está o vertiginoso desenvolvimento de tecnologias de captação, armazenamento, tratamento e analise de dados relacionados às atividades esportivas.

O futebol é, se não o mais, um dos esportes com as movimentações de destaque na atualidade. Atualmente, existem elenco de jogadores em times que passam a casa dos 10 dígitos,

sobre tudo na Europa Ocidental. Os times Espanhóis e Britânicos lideram os valores financeiros relacionados ao capital humano disponível. No entanto, essa aparente hegemonia é, cada vez mais, desafiada com o surgimento de novos mercados da bola. O mercado Norte Americano tem investido mais de XXXXX milhões de dólares em um projeto de incentivo à prática do futebol, desafiador em uma nação com outros esportes com importância sacramentada como é o caso do Futebol Americado e Basquete.

O evento mais recente e com maior destaque no mundo do esporte foi o valor total da transferência de Neymar do Barcelona para o Paris Saint German, um total de mais 400 milhões de euros. Seria ingenuidade não encarar tal transação como um investimento. É esperado que as externalidades dessa movimentação tenha reflexos comparadas com os maiores investimentos nos mercados de financeiros do mundo. Por exemplo, a venda de camisas aumentou 123123123321% no faturamento do clube apenas com vendas de camisas.

Regressão Dirichilet

A distribuição Dirichilet permite modelar um vetor composicional de dimensão $k \leq 2$ $Y = (y_1, y_2, \cdots, y_K)$, com parâmetros de escala $\alpha_1, \alpha_2, \cdots, \alpha_K \geq 0$, sendo $\sum_{j=1}^{K} y_j = 1$ um padrão $K-1$ simplex. Dessa forma, as probabilidades presentes na distribuição multinomial tem como priori conjugada tal distribuição, com densidade probabilidade:

The Dirichilet distribution is able to model a composite vector of dimension $k \leq 2$ $Y = (y_1, y_2, \cdots, y_K)$, with scalings parameters $\alpha_1, \alpha_2, \cdots, \alpha_K \geq 0$, being $\sum_{j=1}^{K} y_j = 1$ a standard $K-1$ simplex. Thus the probabilities present in the multinomial distribution have as a priori conjugated such distribution, with probability density:

$$P(Y|\alpha) = \frac{\prod_{j=1}^{K} \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{k} \alpha_j)} \prod_{j=1}^{K} y^{\alpha_j - 1} \tag{1}$$

The work of (?, ?) that observed a good ajust upon longitudial covariates changes using Dirichilet Regression. In field of sports the work of (?, ?) show a competitive accuracy modeling player abilities in Major League Baseball.

We object model the scale parameters by covariates X_r, and use this covariates to forecast the final classification result. Thus the model is:

$$g[\alpha_j(X)] = \beta_{j0} + \beta_{j1}X_1 + \beta_{j2}X_2 + ... + \beta_{jR}X_R + \varepsilon_j \tag{2}$$

Being $g(.) = log(.)$ has a transformation possible. The sample Y with size N the likelihood is:

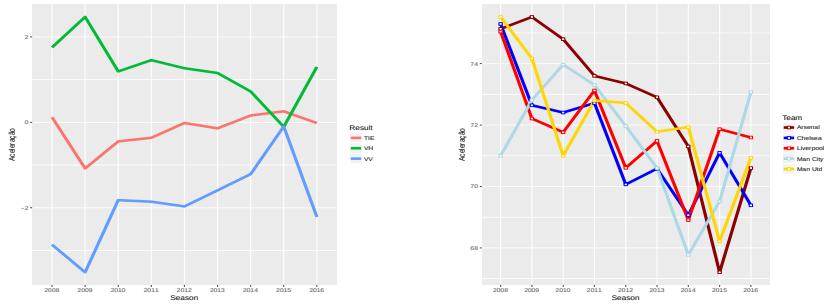
$$L(\beta|X,Y) = \prod_{i=1}^{N} \left[\frac{\prod_{j=1}^{K} \Gamma\left(\alpha_j(X)\right)}{\Gamma\left(\sum_{j=1}^{k} \alpha_c(X)\right)} \prod_{j=1}^{K} y_{ij}^{\alpha_j(X)-1} \right]$$

The coefficients β_{jr} in each α_j of Equation 2 could be mensured solvind the minimum problem as.

$$\underset{\beta}{\text{maximize}} \quad log(L(\beta|X,Y)) - \lambda \left[\phi(||\beta||) + (1-\phi)(||\beta^2||) \right] \tag{3}$$

Being

$$log(L(\beta|X,Y)) = \sum_{i=1}^{N} log \left[\Gamma \left(\sum_{j=1}^{K} \alpha_j(X_{ri}) \right) \right] - \sum_{j=1} log \left[\Gamma\left(\alpha_j(X_{ri})\right) \right] + \sum_{j=1}^{K} [1 - \alpha_j(X_{ri})] log\left(Y_{ji}\right)$$



(a) Summary acceleration mensures by match result.

(b) Acceleration mensure for period of analysis.

Figura 1. **(a)** Summary acceleration mensures by match result. Tie (TIE), Victory of Home team (VH) and Victory of visitor team (VV). **(b)** Skill acceleration mensure for period of analysis.

Thus, we using a maximization procedure to find the β coefficients in equation 2.

Nosso modelo, $K = 3$, sendo α_{Home} o parâmetro associado a probabilidade do time mandante vencer, $\alpha_{Visitor}$ a probabilidade do time visitante vencer e α_{Tie} a probabilidade para o empate dos times.

As variáveis regressoras são apresentadas a seguir.

Base de dados

Foram consideradas como regressoras as habilidades encontradas no site do jogo virtual FIFA. São elas **Aceleração**, **Altura**, **Cabeceio**, **Carrinho**, **Ch.**, **de**, **longe**, **Cobr.**, **falta**, **Combativ.**, **Contr.**, **bola**, **Cruzamento**, **Div.**, **em**, **pé**, **Dribles**, **Duração**, **Do**, **Contrato**, **Elast.**, **GL**, **Finalização**, **Fôlego**, **Força**, **Força**, **chute**, **Idade**, **Lançamento**, **Manejo**, **Marcação**, **Passe**, **curto**, **Perna**, **boa**, **Perna**, **ruim**, **Peso**, **Pique**, **pos**, **Posicion.**, **GL**, **Reação** e **Reflexos**.

Foram coletadas informações durante o período de 2008 e 2016.

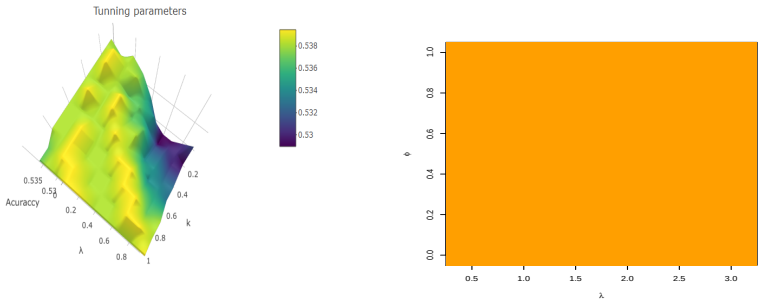
Similarmente ao *Premier League*, os dados dos disponibilizados no site da FIFA acompanham os times de cada temporada. Os dados são atualizados mais de uma vez durante semana de jogos e foram considerados os valores da versão mais atualizada de cada temporada.

Seja X_r^{home} o valor da veriável regressora r para o time mandante do confronto e, analogamente, $X_r^{visitor}$ o valor para a mesma variável do time visitante. A equação do modelo 2 é modelada pela diferença entre a habilidade na variável r entre o time mandante e o visitante. $X_r = X_r^{home} - X_r^{visitor}$. Dessa forma, valores muito positivos para a variável X_r significam superioridade do time mandante e os negativos significam que as habilidades do time visitante são superiores aos anfitriões.

O gráfico do painel (a) da imagem 1 mostra que, em média, os jogos em que a medida X_r é maior do que zero, houve vitória do time mandante. Do forma similar, nos casos em que X_r foi menor do que zero o time visitante venceu. Os casos tem empate, há equilíbrio entre os times e a medida X_r é próxima, em média, de zero. Em anexo estão as medidas para as demais variáveis.

O gráfico do painel (b) da imagem 1 mostra a evolução da variável aceleração ao longo do período de análise para os principais times da *Premier League*. É possível perceber que todos times tiveram entre 2008 e 2014 uma diminuição do valor médio da variável aceleração. O processo de renovação de elenco do Chelsea, Liverpool e Mancherster City iniciou-se em 2015. Já para Arsenal e Mancherster United a renovação ocorreu no ano seguinte em 2015.

O início de renovação do Chelse foi campeã pode ajudar a explicar o sucesso em 2015 e 2017. Uma queda dessas e outras informações das equipes inglesas podem ajudar a compreender a zebra de 2016, quando a equipe de pouca expressão Leicester foi campeã.



(a) Tunning surface parameters (b) Tunning contour parameters

Figura 2. **(a)** Surface of parameters λ and k related with equation 3.**(b)**Contour of parameters λ and k related with equation 3.

Pretendemos explicar e predizer o resultado dos jogos da Premier League, baseado nas habilidades auferidas pelo site FIFA dos jogadores de cada time.

Método

A base de dados foi dividia em:

- **base de treino:** jogos das temporadas de 2008 e 2014. A base servirá para estimar os parâmetros do modelo da equação 2.
- **base de tuning:** jogos das temporadas de 2015. A base servirá para estimar os parâmetros λ e ϕ do processo de **regularização** na equação 3.
- **base de teste:** jogos das temporadas de 2016. A base servirá para avaliar $out-of-sample$ o modelo.

Os parâmetros de ϕ e λ foram avalidados por meio do *grid-search*, sendo $\phi = (0, 0.1, 0.2, 0.3, \cdots, 0.9, 1)$ e $\lambda = (0, 0.5, 1, 1.5, 2, 2.5, 3)$. Os gráficos da figura 2 apresentam o desempenho do modelo para o *grid* de procura dos parâmetros.

O resultado para o modelo proposto não consegue identificar os resultados empate.

Seasons	Global	Home	Tie	Visitor
2008-2014	47.11	79.65	0.00	36.52
2015	38.95	73.25	0.00	28.45
2016	56.32	85.56	0.00	49.54

Tabela 1

Medidas do modelo.

A tabela 2 apresenta a simulação para o resultado da temporada de 2017/2018 baseada nos parâmetros estimados.

Isso ocorre devido o valor de X_r representar uma medida de **dissimilaridade**, isto é, quanto maior o valor X_r mais diferente são as habilidades do time mandante e visitante. Caso as habilidades dos dois times sejam iguais o valor X_r zera, inviabilizando a regressão do parâmetro α_{Tie}.

Dessa forma, propomos uma medida de **similaridade** para regressão do parâmetro α_{Tie} relativo ao empate. Faremos:

$$X_r^{Tie} = \gamma \exp\left\{-\sigma X_r^2\right\} \tag{4}$$

Team	PLC	CCL	CEL	RET
Man Utd	16.50	49.02	14.96	1.77
Man City	16.08	48.46	14.94	1.58
Tottenham	12.90	43.01	15.79	1.98
Chelsea	10.99	37.00	14.25	3.62
Arsenal	9.32	35.04	14.77	3.83
Southampton	8.70	36.16	14.13	3.08
Newcastle	4.51	22.01	12.36	7.78
Crystal Palace	3.85	19.93	11.86	9.30
Stoke City	3.31	17.91	11.30	9.36
Everton	2.79	16.15	10.55	11.90
Leicester	1.85	10.63	8.78	16.67
Liverpool	1.77	12.71	8.93	14.09
Watford	1.69	9.36	7.82	16.87
West Brom	1.52	10.38	8.18	17.85
Bournemouth	1.52	8.61	7.53	19.35
Swansea	1.14	8.78	7.97	19.50
Burnley	0.60	4.68	5.03	29.40
West Ham	0.46	4.25	4.81	31.67
Brighton	0.40	4.31	4.06	32.00
Huddersfield	0.10	1.62	1.98	48.40

Tabela 2

Detailed table of competition where the values are probability of the simulation. Premier League Champion (PLC), Classify to UEFA Champions League (CCL), Classify to UEFA Europa League (CEL) Relegated team (RE).

Seasons	Data	**Global**	Home	Tie	Visitor
2008-2014	Trainning	56.97	80.91	8.77	56.30
2015	Tuning	44.21	71.34	3.74	44.83
2016	Test	51.05	66.84	4.76	59.63

Tabela 3

Medidas do modelo que inclue o ajuste para similaridade do empate.

Nessa estratégia, os parâmetros λ, ϕ, σ e γ são determinados utilizando a procedimento *tuning*. O modelo para α_{Tie} captará o nível de similaridade dos times.

Apesar de pequena, o tratamento para a regressão do coeficiente do emapate contribui para melhorar o poder de previsibilidade do modelo.

Resultados

Team	PLC	CCL	CEL	RET
Liverpool	39.00	86.00	7.00	0.00
Man City	31.00	78.00	12.00	0.00
Chelsea	13.00	62.00	17.00	0.00
Leicester	6.00	34.00	24.00	0.00
Tottenham	5.00	32.00	24.00	2.00
Man Utd	4.00	26.00	19.00	0.00
Stoke City	1.00	26.00	26.00	0.00
Everton	1.00	14.00	8.00	3.00
West Ham	0.00	9.00	13.00	10.00
West Brom	0.00	11.00	17.00	4.00
Watford	0.00	0.00	3.00	21.00
Swansea	0.00	1.00	5.00	17.00
Southampton	0.00	6.00	6.00	16.00
Newcastle	0.00	3.00	3.00	9.00
Huddersfield	0.00	0.00	0.00	64.00
Crystal Palace	0.00	1.00	5.00	21.00
Burnley	0.00	0.00	0.00	84.00
Brighton	0.00	2.00	1.00	18.00
Bournemouth	0.00	0.00	1.00	27.00
Arsenal	0.00	9.00	9.00	4.00

Tabela 4

Detailed table of competition where the values are probability of the simulation with similarity for tie. Premier League Champion (PLC), Classify to UEFA Champions League (CCL), Classify to UEFA Europa League (CEL) Relegated team (RE).

Considerações finais

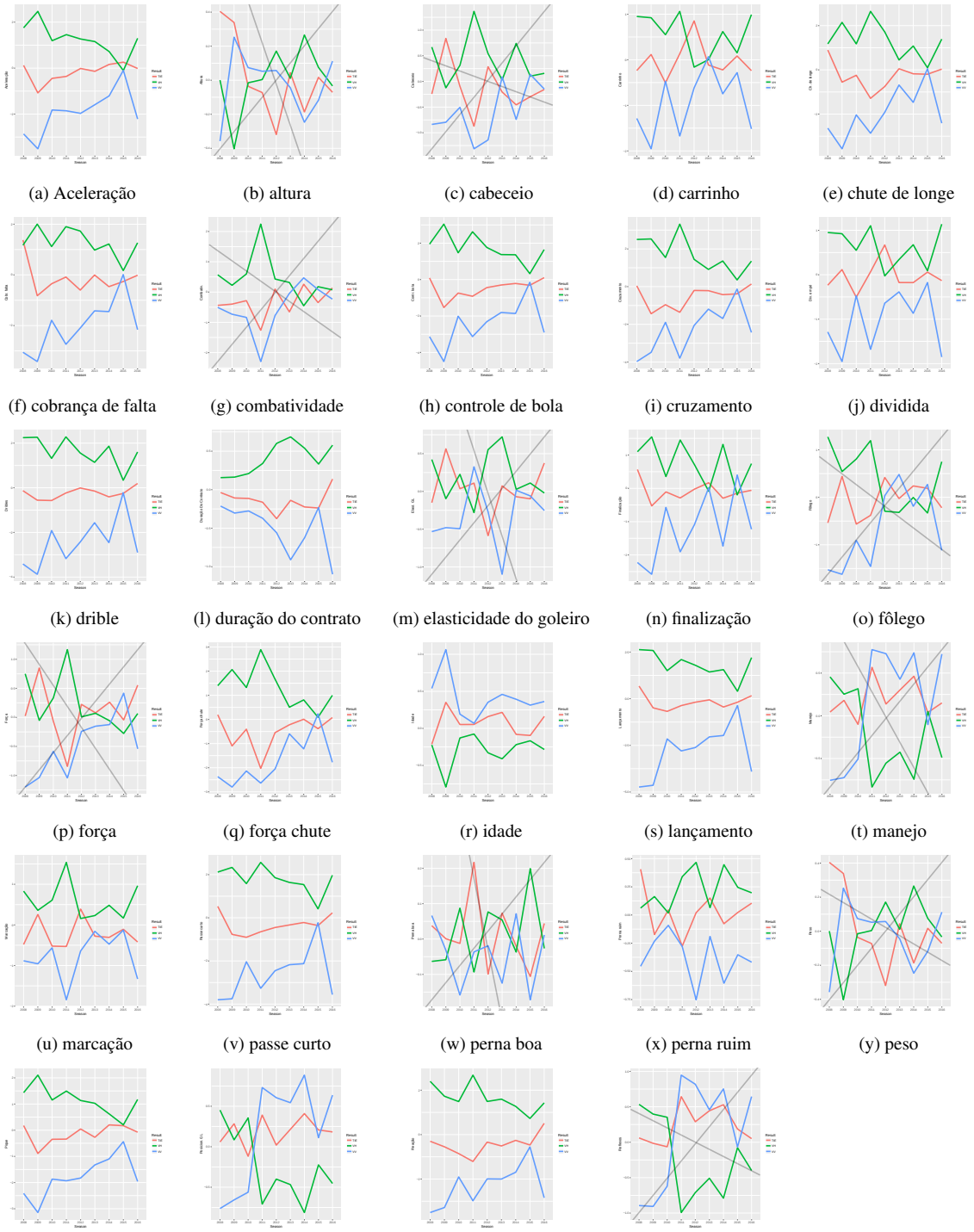
Referências

Hijazi, R. H., & Jernigan, R. W. (2009). Modeling Compositional Data Using Dirichlet Regression Models. *Journal of Applied Probability and Statistics*.

Kain, K. J., & Logan, T. D. (2014). Are Sports Betting Markets Prediction Markets? *Journal of Sports Economics*. doi: 10.1177/1527002512437744

Null, B. (2009). Modeling Baseball Player Ability with a Nested Dirichlet Distribution. *Journal of Quantitative Analysis in Sports*. doi: 10.2202/1559-0410.1175

Štrumbelj, E., & Šikonja, M. R. (2010). Online bookmakers' odds as forecasts: The case of European soccer leagues. *International Journal of Forecasting*. doi: 10.1016/j.ijforecast.2009.10.005



(a) Aceleração (b) altura (c) cabeceio (d) carrinho (e) chute de longe

(f) cobrança de falta (g) combatividade (h) controle de bola (i) cruzamento (j) dividida

(k) drible (l) duração do contrato (m) elasticidade do goleiro (n) finalização (o) fôlego

(p) força (q) força chute (r) idade (s) lançamento (t) manejo

(u) marcação (v) passe curto (w) perna boa (x) perna ruim (y) peso

(z) pique (aa) posição do goleiro (ab) reação (ac) reflexo

Figura 3. Medidas resumo alongo dos anos para cada variável disponível.