Ideias

Igor Patrício Michels

Fevereiro 2022

Premissas e definições

Para desenvolver algumas ideias irei considerar algumas premissas, inicialmente fortes, que tendem a ser relaxadas conforme o desenvolvimento do projeto. Uma premissa inicial seria a de que os atletas são independentes, isso é, não estamos considerando que um "racha" no elenco[1], uma boa relação entre atletas ou até conhecimento do estilo de jogo de um jogador adversário tenha alguma influência em campo. Dessa forma, a proficiência de um time se dá única e exclusivamente pela soma das proficiências de cada atleta individualmente. Por fim, vou adicionar uma premissa extra, que não deverá ser relaxada, que supõe que jogadores de proficiência similar atuam pelo mesmo clube[2].

Vamos definir um atleta i por um valor p_i que representa sua proficiência. Definimos também um clube $i \in \{1, \ldots, n\}$ no tempo t como $\mathcal{C}_{i,t}$, sendo dado por um conjunto de jogadores (ou índices de jogadores) que compõe o elenco em determinado período de tempo (em temporadas, por exemplo). Dessa forma, definimos

$$\lambda_i = \sum_{j \in \mathcal{C}_i} p_j$$

como a proficiência do clube i.

Modelagem

A ideia para o modelo será a de misturar abordagens bayesianas e frequentistas. Inicialmente vamos modelar o futebol, do último ano, a nível de clubes, por meio de vários processos de Poisson. Cada jogo entre dois clubes i e j terá um placar G_i x G_j, onde $G_i \sim Poisson\left(\frac{\lambda_i}{\lambda_j}\right)$ e $G_j \sim Poisson\left(\frac{\lambda_j}{\lambda_i}\right)$. Por meio das observações, podemos encontrar um conjunto $\lambda = \{\lambda_1, \ldots, \lambda_n\}$ que maximiza a verossimilhança dos placares dadas as proficiências.

Um fato a ser percebido agora é que esse modelo é "livre de escala", isso é, multiplicar todas as proficiências por um mesmo valor real c não afetará o modelo. Esse fato nos leva a modelagem a nível individual. Como citado nas premissas, vamos supor que os jogadores de cada clube possuem proficiência similar, assim, diremos que a proficiência de um atleta p_j é dada por $N\left(\lambda_i, \frac{\lambda_i}{2}\right)$, onde i é tal que $j \in \mathcal{C}_{i,T}$, com T sendo o tempo mais recente.

Exemplo

Modelando o futebol brasileiro, temos dados de 2013 a 2021, assim, podemos usar o modelo explicada acima para modelar o ano de 2021 e encontrar o conjunto λ que maximiza a verossimilhança definida pelas observações. Encontrado tal conjunto vamos usar, para todos os jogadores do time A de 2021, prioris da forma $N\left(\lambda_A, \frac{\lambda_A}{2}\right)$ para sua proficiência. Cada jogador terá uma distribuição à priori, que, seguindo as premissas, será independente das demais.

Feito isso, buscamos uma distribuição á posteriori e avaliamos os resultados, primeiramente com testes simples, como supor apenas 20 times com 11 jogadores em cada, durante 10 temporadas.

[1] Como o desentendimento entre Felipe Vizeu e Rhodolfo que ocorreu no jogo entre Flamengo e Corinthians pela rodada 37 do Brasileirão de 2017.

[2] Isso será usado para elicitar distribuições a priori.

Abordagem Frequentista

Durante as férias também implementei esse modelo, mas com a abordagem frequentista. Existiam 20 clubes, cada um com 11 jogadores. Usei o modelo de Poisson sugerido acima para criar dados para 10 temporadas, sendo que ao final de cada temporada os times mudavam completamente. Após isso, defini funções de log verossimilhança para otimizar visando encontrar os parâmetros. Tentei a regressão tanto usando o gradiente dessa função quanto sem o gradiente. Com o gradiente a otimização ocorre rapidamente, mas chega em um máximo local, enquanto sem o gradiente leva cerca de meia hora mas chega no ponto ótimo.

Não convexidade da função de verossimilhança

Conforme a ideia de função de log-verossimilhança descrita na seção de Modelagem, podemos analisar sua convexidade. Isso se faz interessante pois existem melhores métodos para encontrar o ponto de mínimo de funções convexas, assim, se estamos trabalhando com uma função convexa, nossos cálculos computacionais ficam melhores. Para tal análise, podemos começar definindo a função de log-verossimilhança como um somatório onde cada parcela corresponde a quantidade k de gols de um time $\mathcal{C}_a$ contra um time $\mathcal{C}_d$. Como essa quantidade é vinda de uma distribuição de Poisson, temos que cada uma das parcelas será da forma

$$-\log \frac{\lambda^k \exp -\lambda}{k!} = -k \log \lambda - \lambda + \log k!,$$

onde

$$\lambda = \frac{\sum_{i \in \mathcal{C}_a} p_i}{\sum_{j \in \mathcal{C}_d} p_j} = \frac{\lambda_a}{\lambda_d}.$$

Para analisar a convexidade dessa função podemos ver que, em primeiro lugar, $\log k!$ é constante em relação aos parâmetros. Além disso, o logaritmo é uma função côncava no domínio dos parâmetros, logo, menos o logaritmo é convexa. A soma de funções convexas é convexa, bem como a composição das mesmas, assim para analisar se a função de log verossimilhança é convexa podemos ver se λ é um função convexa dos parâmetros a serem estimados.

Podemos ver se essa função é convexa por meio da sua hessiana. Se ela for positiva definida, temos que a função é convexa. Assim, podemos calcular sua hessiana, mas primeiro o gradiente. Para cada entrada i de $\nabla\lambda$, vale que:

$$\nabla\lambda_i = \begin{cases} 0, \text{ se } i \notin \mathcal{C}_a \wedge i \notin \mathcal{C}_d \\ \dfrac{1}{\lambda_d}, \text{ se } i \in \mathcal{C}_a \wedge i \notin \mathcal{C}_d \\ -\dfrac{\lambda_a}{\lambda_d^2}, \text{ se } i \notin \mathcal{C}_a \wedge i \in \mathcal{C}_d \\ \dfrac{\lambda_d - \lambda_a}{\lambda_d^2}, \text{ se } i \in \mathcal{C}_a \wedge i \in \mathcal{C}_d \end{cases}.$$

Feito isso, podemos olhar para a matriz hessiana $\nabla^2\lambda$. Agora, para cada posição j da linha i vale que:

$$\nabla\lambda_{i,j} = \begin{cases} 0, \text{ se } j \notin \mathcal{C}_a \wedge j \notin \mathcal{C}_d \\ 0, \text{ se } i \notin \mathcal{C}_a \wedge i \notin \mathcal{C}_d \\ 0, \text{ se } i \notin \mathcal{C}_d \wedge j \notin \mathcal{C}_d \\ -\lambda_d^{-2}, \text{ se } j \notin \mathcal{C}_d \wedge i \in \mathcal{C}_d \\ -\lambda_d^{-2}, \text{ se } j \in \mathcal{C}_d \wedge i \notin \mathcal{C}_d \\ 2\lambda_a\lambda_d^{-3}, \text{ se } j \notin \mathcal{C}_a \wedge i \notin \mathcal{C}_a \\ -3\lambda_d^{-4} - \lambda_d^{-2} + 2\lambda_a\lambda_d^{-3}, \text{ se } j \in \mathcal{C}_a \wedge i \in \mathcal{C}_a \\ \lambda_d^{-3}\left(2\lambda_a - \lambda_d\right), \text{ c.c.} \end{cases}.$$

Feitas tais contas, queremos ver se $\nabla^2\lambda$ é positiva definida. Para isso, devemos ter que $\forall x \neq 0 \in \mathbb{R}^n, x^\top \nabla^2\lambda x > 0$. Note que

$$x^\top \nabla^2\lambda x = \sum_j x_j \sum_i x_i \nabla^2\lambda,$$

se tomarmos x de modo que suas únicas entradas não nulas sejam dois valores iguais a 1 que estão em posições que representam dois jogadores distintos, sendo um no time atacante e um no time defensor, teremos que a entrada correspondente na hessiana será igual a $-\lambda_d^{-2} < 0$, uma vez que $\lambda_d > 0$. Assim, perceba que vale que

$$\sum_j x_j \sum_i x_i \nabla^2 \lambda = \sum_j x_j - \lambda_d^{-2} = -2\lambda_d^{-2} < 0,$$

o que mostra que a hessiana não é positiva definida, logo, λ não é uma função convexa dos parâmetros.

Na literatura, encontrei um livro de Stanford que cita que as funções desse tipo são chamadas de quasilinear.