Previsão de Resultados de Jogos de Futebol usando Machine Learning

Fillipe Rafael Bianek Pierin *
Bacharelado em Matemática - UFPR
bianekpierin@gmail.com

Prof. Geovani Nunes Grapiglia (Orientador)
Departamento de Matemática - UFPR
geovani_mat@outlook.com

Palavras-chave: Machine Learning, Regressão Logística, regularização, Otimização, Futebol, Previsão.

Resumo:

Machine Learning ou Aprendizagem de Máquinas é um método para análise de dados que evoluiu do estudo de reconhecimento de padrões e da teoria do aprendizado computacional em inteligência artificial. O objetivo da inteligência artificial é o desenvolvimento de "máquinas" com "inteligência" similiar à inteligência humana.

Exite dois tipos de aprendizagem de máquina: o aprendizado supervisionado e o aprendizado não supervisionado. Em mais de 70% do aprendizado de máquina ocorre o uso com aprendizado supervisionado.

No aprendizado supervisionado, o algoritmo do modelo recebe um conjunto de entradas e suas respectivas saídas, onde tais informações serão utilizadas na construção de um modelo de previsão de novas entradas. Por meio de métodos como de classificação, regressão e boosting do gradiente, o modelo utiliza padrões para prever os valores do rótulo em dados adicionais não rotulados. Matematicamente, dados $(x^{(1)}, y^{(1)})$, $(x^{(2)}, y^{(2)})$, $\ldots$, $(x^{(m)}, y^{(m)})$, onde $y^{(i)}$ representa o rótulo associado a $x^{(i)}$, o objetivo é encontrar uma função m_θ tal que $m_\theta(x^{(i)}) \approxeq y^{(i)}$, $i = 1, \ldots, m$. Assim, $m_\theta(x)$ é a previsão para o rótulo de um novo dado x.

No aprendizado não supervisionado não são atribuídas classes aos dados, de maneira que o modelo deve encontrar padrões relativos a tais. Ou seja, dados $x^{(1)}$, $x^{(2)}$, $\ldots$, $x^{(m)} \in \mathbb{R}^n$, o objetivo é encontrar um padrão a partir desse conjunto de dados.

Neste trabalho será apresentado o método de aprendizagem de máquinas com o uso do aprendizado supervisionado, nomeado regressão logística, sem e com o uso de alguns tipos de regularização. Emprega-se este método para implementação e treinamento de modelos, com o intuito de prever resultados dos jogos de futebol e comparar modelos com alguns critérios de regularização e sem regularização, verificando-se

*Iniciação Científica.

qual obtém melhor desempenho na previsão dos resultados. Esse tipo de análise pode ser visto como um problema de classificação, porque se procura classificar os resultados dos próximos jogos em vitória do time mandante (1), empate (2) ou vitória do time visitante (3).

A regularização, que é amplamente utilizada na aprendizagem de máquinas, é necessária para evitar overfitting ou underfitting, principalmente quando há um pequeno número de exemplos (rótulos) ou quando há um grande número de parâmetros a serem aprendidos. Sendo underfitting quando λ é muito grande, e todos os parâmetros acabam sendo penalizandos, deixando-os próximos de zero; e overfitting quando o erro do modelo aplicado ao conjunto de treinamento é excelente, mas quando aplicado ao conjunto de teste fica alto. Para tal, acrescentamos um termo de regularização para a função objetivo

$$f(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)} - y^{(i)})^2 + \lambda\sum_{i=1}^{n}\theta_j^2\right]$$

onde λ é o parâmetro de regularização.

Dado o exemplo usando o objetivo regularizado, obtêm-se uma curva muito mais suave que se adapta aos dados e dá uma hipótese muito melhor.

A probabilidade $m_\theta^{(i)}(x)$ do jogo de futebol pertencer a um dos resultados $i \in \{1, 2, 3\}$ é dada pelo modelo logístico

$$m_\theta^{(i)}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

onde $\theta = (\theta_0, \theta_1, \ldots, \theta_n) \in \mathbb{R}^{n+1}$ é o vetor de parâmetros do modelo e $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$.

Usa-se o método One-vs-All para classificar os resultados dos jogos futuros, em vitória do time mandante ($m_\theta^{(1)}(x)$), empate ($m_\theta^{(2)}(x)$) ou vitória do time visitante ($m_\theta^{(3)}(x)$), usando dados das rodadas anteriores. Ou seja, o método One-vs-all treina o classificador da regresão logistica $m_\theta^{(i)}(x)$, para cada classe (i), para predizer a probabilidade de $y = j$. Desta forma, pelo método One-vs-All, se

$$j = argmax_{i\in\{1,2,3\}}\left\{m_\theta^{(i)}(x)\right\}$$

prevemos que o jogo x terá como resultado j.

Os parâmetros (θ) de cada modelo são calibrados a partir de dados do campeonato brasileiro de 2017, retirados do site da Confederação Brasileira de Futebol (CBF) e do site SoccerWay. Esta calibração dos parâmetros faz-se de modo a minimizar o erro entre as previsões de cada modelo e os dados reais. Para analisar os dados utiliza-se o programa Octave.

Alguns resultados de classificação obtidos do campeonato brasileiro de 2017, com o uso do modelo regressão logística sem regularização, cuja taxas de acerto dos modelos $m_\theta^{(1)}(x)$, $m_\theta^{(2)}(x)$, e $m_\theta^{(3)}(x)$, que são respectivamente, vitória do time madante, empate e vitória do time visitante, e do One-vs-All são os seguintes:

Tabela 1: Taxa de acerto dos modelos e One-vs-All usando regressão logística.

Rodada	$m_\theta^{(1)}(x)$	$m_\theta^{(2)}(x)$	$m_\theta^{(3)}(x)$	Taxa de Acerto (One-vs-All)
6º	50.00%	70.00%	70.00%	60.00%
7º	50.00%	75.00%	75.00%	58.62%
8º	50.00%	85.71%	64.29%	57.58%
9º	56.25%	25.00%	62.50%	60.81%
10º	55.56%	50.00%	38.89%	63.41%
11º	55.00%	50.00%	35.00%	62.22%
12º	59.09%	50.00%	18.18%	62.24%
13º	50.00%	54.17%	29.17%	58.49%
14º	69.23%	34.62%	46.15%	60.53%
15º	67.86%	32.14%	50.00%	58.20%
16º	70.00%	63.33%	43.33%	60.00%
17º	71.88%	65.63%	40.63%	58.70%
18º	64.71%	73.53%	55.88%	54.11%
19º	63.16%	68.42%	52.63%	52.63%
20º	68.40%	73.70%	60.50%	52.47%

Por fim, compara-se os modelos treinados com modelos usados em sites da internet, para desta forma verificar se o modelo apresenta resultados confiáveis e que predizem o mais corretamente os resultados dos jogos de futebol.

Referências:

[1] FIGUEIRA, C. V. **Modelos de regressão logística**. 2006.

[2] PAULA, G. A. **Modelos de regressão: com apoio computacional**. [S.l.]: IME-USP São Paulo, 2004.

[3] PIERIN, F. R. B. **Previsão de resultados de jogos de futebol por meio de regressão logística multinomial**. 2016.