

Universidade Federal de São João del-Rei
Cálculo Numérico
Trabalho Prático

Felipe Francisco Rios de Melo
Thales Mrad Leijoto

1. INTRODUÇÃO

Regressão é o ato de prever, a partir de dados históricos, dados no futuro. Estudando a relação entre duas ou mais variáveis, em que uma depende de outra ou outras.

Em Estatística, aprende-se, um conceito chamado correlação, isto é, uma variável pode depender positivamente ou negativamente em relação a alguma outra. Por exemplo: imagine encontrar uma correlação entre o preço de um imóvel e seu tamanho. É difícil obter uma correlação perfeita, mas podemos entender que há um padrão nos dados, e que um valor depende de outro ou outros. A Regressão, já é a descoberta dessa função, que expressa o padrão nos dados. Ou seja, quando se desenvolve um trabalho de Análise de Regressão, busca-se entender o padrão de um valor dependendo de outro ou outros, e assim encontrar uma função que expressa esse padrão.

Esse conceito é usado em muitas áreas da matemática e computação, como por exemplo: estatística, *machine learning* e análise de dados em geral.

O objetivo deste trabalho, então, foi aplicar o conceito de regressão linear em uma base de dados, onde que dado um conjunto de dados de entrada, tentar prever a saída através da modelagem de um polinômio.

2. DESENVOLVIMENTO

2.1. DETALHES TÉCNICOS DE IMPLEMENTAÇÃO

O algoritmo foi implementado em Python 3, a manipulação do arquivo de dados .csv foi feita com o auxílio da biblioteca *Pandas* e a manipulação das matrizes foi realizada com a biblioteca *NumPy* e com a biblioteca implementada por nós, *Matrix*.

2.2. DATASET

Para a elaboração do trabalho, foi utilizado uma base de dados intitulada “FIFA 19 complete player dataset”, que é uma base do jogo eletrônico de futebol FIFA 19 que lista todos os jogadores, bem como todos os seus atributos. Ela contém 18206 linhas referentes a todos os jogadores e 89 colunas referentes a todos os atributos de cada jogador.

2.3. SOLUÇÃO PROPOSTA

Como mostrado anteriormente, o objetivo deste trabalho foi prever uma saída coerente aos dados de entrada. Então, dada a base de dados, pretendíamos prever o atributo *overall* dos jogadores baseado em outros atributos.

Apenas a fim de contextualização, no FIFA 19, *overall* é uma pontuação geral que se dá a um jogador.

A princípio, o intuito era prever o *overall* de todos os 18206 jogadores, porém estudando um pouco a respeito do FIFA 19, um mesmo atributo pode ser muito significativo para algumas posições e pouco para outras (por exemplo, um atacante deve ter uma boa finalização, mas para um zagueiro isso não é tão importante). Por este fato, a fim de simplificação optamos por restringir a amostra à apenas jogadores com o ofício de goleiro.

Foram considerados seis atributos relevantes para a predição do *overall* de um goleiro, são eles: habilidade com as mãos (*GKHandling*), chute (*GKKicking*), elasticidade (*GKDivining*), posicionamento (*GKPositioning*), reflexos (*GKReflexes*) e reputação internacional (*InternationalReputation*).



Figura 1 - Exemplo de goleiro junto a seus atributos relevantes no FIFA 19.

Como dito na primeira estrofe da introdução, regressão é o ato de prever, a partir de dados históricos, dados no futuro. Tomando como dados históricos 70% dos jogadores, queremos prever o *overall* dos outros 30%. Aqui, chamaremos esta relação de treino e teste. Vamos treinar o algoritmo com 70% dos dados, ela gerará uma função baseada na correlação entre os atributos, e por fim, aplicaremos essa função ao restante dos dados.

Na regressão linear, deseja-se ajustar os parâmetros x , a fim de que a função gerada fique o mais próximo possível original, isso é feito através da resolução do seguinte sistema linear

$$G^T G x = G^T y$$

Onde:

- G é uma matriz em que as linhas representam todos os goleiros, a primeira coluna é toda preenchida com 1 (intercepto) e as seguintes correspondem aos seis atributos de cada goleiro;
- G^T é a matriz transposta de G ;
- y é uma matriz unidimensional onde que as linha representa o *overall* de cada jogador;
- x é uma matriz unidimensional de parâmetros que queremos encontrar.

Para a resolução do sistema linear, optamos por implementar o algoritmo iterativo de Gauss-Seidel, por convergir mais rápido que o método de Gauss-Jacobi.

Resolvido o sistema linear, temos os valores dos parâmetros x , e então podemos aplicar a função polinomial aos testes.

$$P(\text{InternationalReputation}, \dots, \text{GKReflexes}) = x_0 + \text{InternationalReputation} * x_1 + \text{GKdiving} * x_2 + \text{GKHandling} * x_3 + \text{GKKicking} * x_4 + \text{GKPositioning} * x_5 + \text{GKReflexes} * x_6$$

E por fim, obter a previsão do *overall* dos jogadores.

3. RESULTADOS E DISCUSSÕES

No *dataset* tem-se um total de 2022 goleiros, 70% deles foram usados para o treino do algoritmo, o que equivale a 1415 e os outros 607 foram usados como teste.

A cada execução do algoritmo, os dados selecionados como treino e teste alteram, pois antes da seleção as linhas da base de dados são misturadas de forma aleatória, a fim de testar o modelo para casos distintos e pode verificar sua consistência.

Abaixo estão os parâmetros x calculados de três execuções do programa:

Execuções	Vetor de parâmetros x
Execução 1	[-0.33946443 , 0.28997661, 0.22408178, 0.22697313, 0.0548692, 0.23906576, 0.26125187]
Execução 2	[-0.32581558, 0.22527236, 0.22743068, 0.2266009, 0.05476551, 0.24170368, 0.25676958]
Execução 3	[-0.33892573, 0.23493783, 0.23167062, 0.22897034, 0.05639082, 0.23845169, 0.25171962]

Tabela 1 - Parâmetro x de três execuções distintas

É possível, então, notar que a diferença entre os parâmetros gerados de cada execução é mínima. O que gera certa confiança no modelo, pois independente do conjunto de entradas usados como teste, o polinômio gerado será consistente e, logo, a ordem de grandeza do erro não irá variar de uma execução para outra.

Ainda, pegando a Tabela 1 como referência, podemos ver a correlação entre os atributos da formação do *overall*. É possível observar que só há correlações positivas (com exceção da variável livre), onde que o atributo que tem menor influência no cálculo do *overall* é o GK Kicking (chute), pois ele é multiplicado pelo coeficiente de apenas 0.05, enquanto que os outros coeficientes flutuam em 0.2.

Quanto a análise da previsão do valor de *overall* dos 607 jogadores, segue a tabela abaixo de 5 goleiros aleatórios:

Valor real	Valor previsto	Erro absoluto	Erro relativo
72	71.8660116325575	0.13398837	0.00186095
68	68.10162904731901	0.10162905	0.00149454

68	67.78986412469072	0.21013588	0.00309023
48	47.58034412749006	0.41965587	0.00874283
59	60.41667002321033	1.41667002	0.02401136

Tabela 2 - Previsão e erro de 5 goleiros aleatórios

Pegando os 607 jogadores, e calculando a média do erro absoluto e erro relativo temos os seguintes resultados:

- Erro absoluto médio: 0.55286868
- Erro relativo médio: 0.00872573

O erro absoluto é a diferença entre o valor exato de *overall* e seu valor aproximado. O erro absoluto não permite uma avaliação da precisão entre dois resultados de forma correta porque não leva em consideração a grandeza destes números, o erro relativo cumpre este papel.

Porém o atributo *overall* neste *dataset* tem um intervalo [47, 90], são valores da mesma grandeza. Portanto, para a análise do erro, vamos nos concentrar no erro absoluto.

O erro absoluto médio de 0.55286868 superou as expectativas, pois em uma base de dados com tantos atributos, apenas seis deles foi capaz de chegar tão próximo do *overall* real, tendo uma acurácia média superior a 98.8%

4. CONCLUSÃO

Os resultados foram muitos satisfatórios, foi possível aplicar os conceitos de regressão linear múltipla e vários outros conceitos da disciplina de Cálculo Numérico neste trabalho.

Quanto às limitações da solução e dificuldades de implementação, a maior delas acreditamos ser o fato de realizar a previsão apenas para os goleiros, e não para qualquer jogador. Porém se abrangêssemos para qualquer jogador a complexidade da implementação iria aumentar muito (e este não é o objetivo inicial do trabalho), devido ao fato de que cada posição específica tem seus atributos relevantes.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Implementando Regressão Linear Simples em Python. Disponível em:

<<https://medium.com/data-hackers/implementando-regress%C3%A3o-linear-simples-em-python-91df53b920a8>>

FIFA 19: Domine os atributos dos jogadores e aperfeiçoe sua equipe no Ultimate Team. Disponível em:

<https://www.espn.com.br/esports/artigo/_/id/4820667/fifa-19-domine-os-atributos-dos-jogadores-e-aperefeicoe-sua-equipe-no-ultimate-team>

Análise de Erros Numéricos. Disponível em:

<https://ufsj.edu.br/portal-repositorio/File/prof_ngoulart/notas_aula/Calculo_Numerico_Erros.pdf>