MACHINE LEARNING NO MUNDO CINEMATOGRÁFICO

Isabela da Silva Dias Coelho[1], Marcella Meirelles Ferreira[2], Marcus Vinícius Faustino[3]

[1]Universidade Federal de Minas Gerais/Departamento de Engenharia Química/Escola de Engenharia, bela.silvadc@gmail.com

[2]Universidade Federal de Minas Gerais/Departamento de Engenharia Química/Escola de Engenharia, marcellamferreira@outlook.com

[3]Universidade Federal de Minas Gerais/Departamento de Engenharia Química/Escola de Engenharia, marcus.faustino1@gmail.com

Resumo: A IMDb é uma base de dados pertencente ao grupo Amazon que reúne informações sobre músicas, filmes, programas de televisão e jogos, sendo uma ferramenta extremamente útil em uma era com o crescente consumo desses tipos de conteúdo. O presente trabalho busca, a partir de uma análise de parâmetros de filmes em um conjunto de dados da IMDb, observar tendências na indústria cinematográfica e construir um modelo computacional capaz de prever a nota média de avaliação de um filme.

Palavras-chave: IMDb, linguagem de programação, machine learning, modelo.

1. Introdução:

Machine learning, que pode ser traduzido como aprendizado de máquina, é um ramo em evolução de algoritmos computacionais projetados para emular a inteligência humana aprendendo com o ambiente circundante, podendo modificar seu comportamento autonomamente. A tal modificação comportamental consiste, basicamente, no estabelecimento de regras lógicas que visam melhorar o desempenho de uma tarefa ou, dependendo da aplicação, tomar a decisão mais apropriada para o contexto. Essas regras são geradas com base no reconhecimento de padrões dentro dos dados analisados (ALECRIM, 2018).

Técnicas baseadas em machine learning têm sido aplicadas com sucesso em diversos campos, desde reconhecimento de padrões, visão computacional, engenharia de naves espaciais, finanças, entretenimento e biologia computacional até aplicações biomédicas e médicas (EL NAQA et al., 2015). Alguns exemplos práticos em que o aprendizado de máquina vem sendo usado na área de engenharia são na predição de erros em processos químicos industriais, na análise de equipamentos e na predição de erros de cálculo em construções.

No campo do entretenimento, a indústria cinematográfica tem ganhado maior visibilidade nos últimos anos. Cada vez mais, ela se insere no mundo online e se torna mais democrática. O aumento da procura por este tipo de entretenimento gera a curiosidade dos espectadores a respeito do conteúdo dos filmes, das séries e dos programas de TV, para que possam escolher com mais precisão o que os interessa e não perder tempo assistindo algo que não condiz com suas expectativas. Uma forma de filtrar os produtos é por meio de notas dadas pelos próprios espectadores. Além da praticidade para o espectador, as notas dadas refletem diretamente os produtores, atores e hospedeiros dos títulos.

O IMDb é uma das maiores base de dados disponíveis atualmente sobre cinema e tudo o que envolve a indústria do entretenimento. Além de reunir informações sobre artistas e produções, o site também permite que usuários criem listas e avaliem seus filmes favoritos. Qualquer usuário cadastrado pode dar notas aos títulos. Os votos individuais são agregados e resumidos em uma nota única, exibida com destaque na página principal do título (MELO, 2018).

Uma análise deste banco de dados, utilizando o contexto de machine learning, é capaz de predizer notas de filmes. Conseguir prever se um filme será um fiasco ou um sucesso de bilheteria pode poupar milhões as produtoras de filmes e investidores, além de ajudar o espectador a fazer uma boa escolha e não se decepcionar (BARBOSA, 2017).

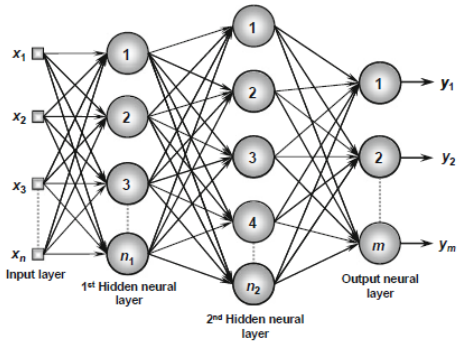


Figura 1. Exemplo de estrutura de uma rede neural de multicamadas (SILVA et al, 2017)

Entre as diversas técnicas de machine learning tem-se as Redes Neurais Artificiais, que busca reconhecer padrões através da construção de um modelo baseado na arquitetura de neurônios humanos, o que permite um processamento complexo em torno dos dados disponíveis e do problema que se busca a solução. Um tipo de Rede

Neural é a Multilayer Perceptron (MLP), que utiliza uma estrutura de múltiplas camadas com uma função de ativação para o processamento, como na Figura 1 (SILVA et al, 2017).

2. Explorando o Banco de Dados

Kaggle é uma renomada plataforma que incentiva o aprendizado de machine learning, onde são promovidas competições, divulgados treinamentos e disponibilizados bancos de dados para exploração dos usuários. Os dados da IMDb, utilizados no presente estudo, foram obtidos no site Kaggle, onde já foi baixado por mais de 13.400 usuários e com mais de 91.600 visualizações.

O dataset “IMDb movies” contém 85.855 filmes e 22 atributos, entre eles título, ano de publicação, diretor, avaliação, entre outros. Foram considerados no estudo o ano de lançamento, a duração, a nota média e o número de votos, bem como o idioma e gênero do filme (drama, ação, romance, entre outros). O código utilizado para executar a MLP foi construído na linguagem Python e já foi utilizado como recurso de aprendizagem desse tipo de técnica a partir da exploração de outros dados.

3. Metodologia

Foi realizado um tratamento dos dados antes de executar o código de MLP. Para isso, foram analisados quais atributos seriam mais relevantes para o modelo de predição das notas, listados na Tabela 1. Devido à forma com que os atributos foram organizados, o banco ficou formado por 85.855 linhas (filmes) e 281 colunas (atributos).

Tabela 1. Atributos utilizados na construção do modelo

Atributos utilizados na construção do modelo			
título	ano	duração	média de votos
votos	gênero	país de produção	idioma

A partir do tratamento dos dados, foram realizados ajustes no código para melhor adequação do modelo. O código separa inicialmente os dados em dois grupos: treinamento e teste. Os dados de treinamento são os que treinam o modelo, servindo de input à arquitetura da MLP, enquanto os dados de teste são aqueles que avaliam

o modelo, o que permite obter as informações se ele foi adequado ou não para a predição. Em um segundo momento os dados de treinamento subdivide em treinamento e validação, que garante a construção do modelo. Foi executado o código e feita a coleta das informações para análise dos resultados.

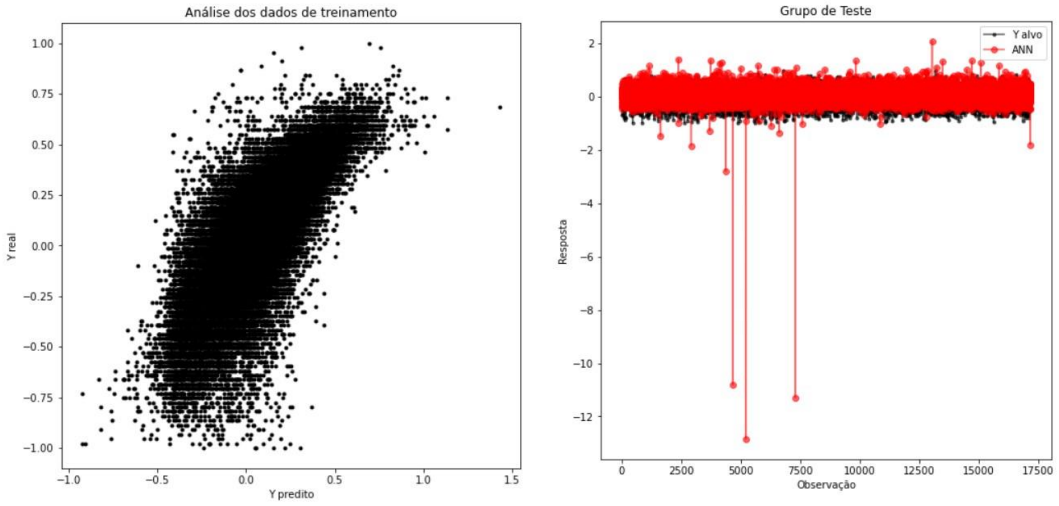
4. Análise e Interpretação dos Dados

Estão contidos na Tabela 2 os resultados obtidos de cada grupo de dados, Teste e Treinamento.

Tabela 2. Resultados dos parâmetros do modelo obtido

Grupo	Coef. Pearson	Erro Q. Médio	Erro Médio Abs	Erro Rel. Médio
Teste	0,0778	0,0729	0,1625	1,4481
Treinamento	0,5936	0,3106	0,1293	1,1964

O Coeficiente de Pearson é um teste que mede a relação estatística entre duas variáveis, que no presente trabalho são os valores das variáveis de observação (do banco de dados da IMDb) e os valores das variáveis de resposta obtidas pelo modelo. Esse valor pode variar entre -1 e 1, onde um valor diferente de 0 significa que existe uma relação entre as variáveis, podendo esta ser uma correlação negativa (menor que 0) ou positiva (maior que 0). A correlação positiva indica que, quanto mais o valor de uma variável aumenta, mais o valor da outra variável aumenta também.



(a) (b)

Figura 2. (a) Análise dos dados de treinamento, (b) Análise dos dados de treinamento

Por meio da Tabela 2, verifica-se que o coeficiente para o grupo de Treinamento apresentou uma correlação fora do que se considera ideal, que seria próximo de 0,9000, mas ainda assim o valor de 0,5936 indica uma correlação positiva entre os dados. A Figura 2 (a) revela a plotagem dos dados de resposta real e o previsto pelo modelo obtido. Era esperado o comportamento de uma reta, uma vez que Y real fosse igual a Y predito, mas com a grande extensão dos dados verifica-se uma tendência de comportamento linear. Na Figura 2 (b) pode ser observada a relação entre Y alvo (resposta) e Y ANN (previsto pelo modelo). Esperava-se o comportamento de sobreposição dos dados, o que indicaria uma boa correlação entre os dados, mas ao contrário disso, vários dados estavam fora do previsto.

Uma das possíveis explicações para a baixa adequação do modelo na previsão das notas pode estar relacionado aos parâmetros utilizados no treinamento da rede, como o número de neurônios, o tamanho máximo das épocas, a tolerância, a função de otimização empregada e a função de ativação selecionada. A adequação desses parâmetros ao conjunto de dados analisados pode alterar a qualidade do ajuste e fazer com que o modelo explique melhor os dados. No entanto, a alteração de alguns desses parâmetros, como o número de neurônios e a tolerância, aumenta significativamente o custo computacional do algoritmo. Dessa maneira, com acesso a recursos computacionais limitados não foi possível realizar um estudo mais aprofundado da influência de cada parâmetro no ajuste.

Outra possível explicação está na seleção dos atributos utilizados para a construção do modelo. O conjunto de dados do IMDb possui diversos parâmetros não numéricos e de difícil inserção no modelo. Dessa maneira, foram selecionados parâmetros mais facilmente transformados em valores numéricos como o fato de um filme ter sido produzido em determinado país ou não. No entanto, é provável que, embora esses dados possuam alguma correlação com a qualidade de um filme, eles não consigam explicar completamente as nuances do que constitui ou não um bom filme. Além disso, alguns dados numéricos como o custo de produção e a receita não puderam ser utilizados devido ao fato de as informações desse campo se encontrarem incompletas. Nesse sentido, seria necessário um maior trabalho de limpeza e processamento dos dados, inclusive associando os dados desse banco a outros bancos disponíveis. A

título de exemplo, poderia ser construída uma referência cruzada entre o elenco de um filme e um banco de dados de ganhadores de prêmios de cinema, sendo maior a probabilidade de um filme com um elenco premiado obter uma melhor classificação. O mesmo poderia ser feito para diretores e produtoras. Além disso, poderia ser feita uma análise de palavras-chave na descrição dos filmes e nos títulos dos filmes, verificando se existe alguma correlação entre algumas palavras e as notas atribuídas a cada filme. Enfim, quanto maior o número de dados usados para descrever cada filme maior a chance de obter um modelo satisfatório.

5. Conclusão

A análise dos dados do IMDb a partir do código de MLP não foi como o esperado, ocasionando resultados inconclusivos a respeito das predições de notas de filmes. Mesmo sem atingir um dos objetivos principais do estudo, foi possível aplicar conceitos de programação e ampliar os conhecimentos no ramo. Além disso, análises gráficas e qualitativas de erros foram feitas.

Em suma, um trabalho de limpeza e processamento dos dados mais profundo deveria ser feito para que os resultados fossem mais próximos do esperado e a predição pudesse ser feita

Referências

ALECRIM, E. Machine learning: o que é e por que é tão importante. Tecnoblog, 2018. Disponível em: https://tecnoblog.net/247820/machine-learning-ia-o-que-e/. Acesso em: 3 fev. 2021.

BARBOSA, A. S. *et al*. Previsão das notas dos filmes no Internet Movie Database (IMDb). Rio de Janeiro, 20 jan. 2017. Disponível em: https://ensinandomaquinasblog.wordpress.com/2017/01/20/previsao-das-notas-dos-filmes-no-internet-movie-database-imdb/. Acesso em: 3 fev. 2021.

EL NAQA, I. *et al*. What Is Machine Learning?. *In*: EL NAQA, I. *et al*. Machine Learning in Radiation Oncology. Switzerland: Springer, Cham, 2015. p. 3-11. ISBN 978-3-319-18305-3. Disponível em: https://link.springer.com/chapter/10.1007/978-3-319-18305-3_1. Acesso em: 3 fev. 2021.

MELO, D. Como funcionam as notas do IMDb. Tecnoblog, 2018. Disponível em: https://tecnoblog.net/368275/como-funcionam-as-notas-do-imdb/. Acesso em: 3 fev. 2021.

SILVA, I. *et al*. Artificial Neural Networks, A Pratical Course. Brasil: Springer, 2017. p. 3-23. ISBN 978-3-319-43162-8