Nanodegree Engenheiro de Machine Learning

Proposta de projeto final

Izabela Fonseca
24 de abril de 2018

Proposta

Histórico do assunto

No início da existência dos serviços de *streaming* durante a década de 90, a internet ainda era extremamente custosa e possuía uma performance muito menor do que a que temos acesso hoje em nossos próprios celulares. Com o surgimento dos serviços de banda larga e internet móvel, o mercado de *streaming* tomou ainda mais força e vem crescendo cada dia mais. Pesquisas e dados divulgados pela Nielsen Music no ano de 2017 explicitam que, em média, os americanos escutam mais de 32 horas de música por semana, onde a reprodução da maior parcela dessas músicas é através dos serviços de *streaming*,o que afirma o crescimento acelerado que esse mercado teve nos últimos anos.

Todo o sucesso e aumento de market share desta indústria trouxe consigo um crescimento das plataformas de *streaming*, o que tem como consequência uma competição cada vez mais acirrada para a captação e fidelização de seus clientes. Dessa forma, muitas empresas têm feito uso de *Machine Learning* principalmente para definir o perfil musical de seus clientes e fazer recomendações. A empresa mais bem sucedida nessa empreitada hoje é sem dúvida o Spotfy que, cruzando os estilos musicais, *playlists* e opinião dos usuários sobre as músicas recomendadas, têm aprendido e recomendado músicas de forma cada vez mais assertiva e eficiente, tornando este um caso muito interessante a ser estudado.

Descrição do problema

Dado o contexto introduzido acima, o problema proposto consiste em criar um modelo capaz de identificar se os os usuários em questão gostam ou não de determinada música. Dessa forma, deverá ser implementado um modelo de classificação, onde uma classe é definida para os usuários que gostam de certa música e a outra para os que não gostam. Este problema pode ser avaliado utilizando, por exemplo, a acurácia como métrica de performance.

Conjuntos de dados e entradas

A base de dados que será utilizada para implementar o projeto foi encontrada no Kaggle, intitulada de "Spotfy Song Attributes", e corresponde às características de músicas escutadas por um usuário. Tal base possui 16 colunas, onde 15 dessas representam as entradas que descrevem diversos atributos musicais técnicos das músicas e a coluna *target* que representa a saída, dizendo se o usuário gosta (1) ou não (0) de determinada música.

Dentre os atributos de entrada temos o *time.signature*, que define o andamento rítmico das músicas, o *acousticness* que revela o índice de confiança se a música é acústica ou não, *danceability* que diz se a música tem um rítmo forte e estável, defininto o quão "dançante" é, a *duration.ms* que traz a duração da música em milissegundos, *energy* que indica o nível de energia da música com base na intensidade de barulho que fazem e muitos outros, além do nome da música e do artista.

Com o mapeamento de todas essas características, torna-se possível a análise e identificação dos padrões musicais que são de preferência de determinado usuário e assim, com a criação de um modelo de classificação, a previsão de classificação desse usuário de outras músicas nunca ouvidas antes por ele. Consequentemente, as recomendações se tornam mais embasadas, personalizadas e assertivas.

Descrição da solução

Uma possível solução para este problema é implementar um modelo de classificação de aprendizado supervisionado, onde teríamos como as features as características técnicas de cada música e como reposta preenchida, ou seja, o "y", a feature *target*. Dessa forma, usaríamos parte desse base de dados para treinar o modelo com o método escolhido, e depois aplicaríamos essa modelo na parcela da base destinada para teste, e a partir do resultado obtida na base de teste, avaliaríamos o desempenho do model adotando a métrica mais adequada. Essa solução certamente poderia ser implementada para qualquer usuário, já que teríamos essas informações técnicas das músicas classificadas por outros usuários, o que torna a solução de fácil reprodução.

Modelo de referência (benchmark)

Após algumas pesquisas, foi possível ver a aplicação de diversos métodos de classificação de aprendizado supervisionado nesta mesma base, uma vez que este é um desafio do Kaggle. Dentre eles estão: árvore de decisão, *random forest*, GaussianNB e XGBoost, onde os que atingiram melhores resultados foram *random forest* e SVM *classifier*, sendo este último encontrado em um tutorial. Os valores de acurácia atingido por cada um foi em torno de 72%.

Métricas de avaliação

Como métricas de avaliação, será considerada a mais trivial de todas para rápidas percepções dos modelos gerados, e para uma análise mais profunda, o F1 Score, que nos trás as informações de falsos positivos, e valores corretamente classificados. Ambas as métricas são obtidas através da avaliação da taxa de acerto da base de treinamento, onde a acurácia contabiliza os acertos em geral do modelo nessa base e o F1 Score nos fornece a tabela de confusão, apresentando os falsos positivos, falsos negativos e acertos positivos e negativos.

Design do projeto

A estruturação do projeto se dá como segue:

1- Análise das features, entendendo seus significados e observando seus valores, média, máxima,

mínimas, dispersão de dados e outas características relevantes;

2- Pré-processamento dos dados, como por exemplo a transformação dos dados categóricos em valosres numéricos;

3- Treinamento do modelo utilizando classificadores SVM's;

4- Avaliação do modelo e implementação de técnicas para melhorar o resultado.