Nanodegree Engenheiro de Machine Learning

Proposta de projeto final

Matheus de Carvalho Luz
19 de agosto de 2018

Proposta

Histórico do assunto

Sistemas de recomendação possuem um bom destaque na nossa sociedade atual, onde grandes empresas como Google, Netflix e Spotify os utilizam para oferecer a melhor experiência para o usuário. Com esse tipo de sistema, é possível oferecer vídeos, músicas, playlists, filmes, e até propagandas que sejam do interesse do usuário.

Descrição do problema

Pretendo com essa proposta seguir os moldes da competição[1] na plataforma Kaggle[2], onde seu objetivo é prever quais músicas serão ouvidas novamente dentro de um período de 1 mês. Com seu dataset, é possível analisar uma série de dados que podem ajudar a determinar quais são os fatores que fazem a diferença para que o usuário queria ouvir uma música novamente.

[1] https://www.kaggle.com/c/kkbox-music-recommendation-challenge
[2] https://en.wikipedia.org/wiki/Kaggle

Conjuntos de dados e entradas

train.csv

- msno: Id do usuário
- song_id: Id da música
- source_system_tab: O nome da guia em que o evento foi acionado. As guias do sistema são usadas para categorizar as funções dos aplicativos móveis KKBOX. Por exemplo, a guia "my library" contém funções para manipular o armazenamento local e a "tab search" contém funções relacionadas à pesquisa.
- source_screen_name: Nome do layout que um usuário observa.
- source_type: O ponto de entrada onde um usuário tocou a música pela primeira vez em aplicativos móveis. Um ponto de entrada pode ser álbum, lista de reprodução online, etc.
- target: Esta é a variável alvo. target = 1 significa que há evento (s) de escuta recorrente acionado em um mês após o primeiro evento de escuta observável do usuário, caso contrário target é igual a 0.

songs.csv

- song_id: Id da música
- song_length: Duração do tempo em ms
- genre_ids: Categoria de gênero musical. Algumas músicas possuem múltiplos gêneros e são separados por "|"
- artist_name: Nome do Artista
- composer: Nome do Compositor
- lyricist: Nome do Letrista
- language: Nome da Língua

members.csv

- msno: Id do usuário
- city: Cidade
- bd: Idade
- gender: Gênero
- registered_via: método de registro
- registration_init_time: formato %Y%m%d
- expiration_date: formato %Y%m%d

song_extra_info.csv

- song_id
- song name
- isrc - Código Internacional de Gravação Padrão

Descrição da solução

O objetivo é obter uma previsão de uma categoria binária, e o algoritmo de Gradient Boosting será utilizado para realizar esta previsão.

Modelo de referência (benchmark)

O modelo de Leaderboard público com score de 0.68310 será utilizado como modelo de benchmark. Será realizada uma tentativa para que nosso modelo possa se encontrar na metade para cima top 50%) na lista de Leaderboard.

Métricas de avaliação

ROC curve

Na classificação binária, a predição de classe para cada instância é frequentemente feita com base em uma variável aleatória contínua X, que é uma "pontuação" calculada para a instância (por exemplo, o preditor linear na regressão logística). Dado um parâmetro de limite T, a instância é classificada como "positiva" se X> T e "negativa" de outra forma. X segue uma densidade de probabilidade f1 (x) se a instância realmente pertencer à classe "positivo" e f0 (x) se de outra forma. A taxa positiva verdadeira é dada por

$$\left(TPR(T) = \int_{T}^{\infty} f_1(x)\,dx \right)$$

e a taxa de falsos positivos é dada por

$$\left(FPR(T) = \int_{T}^{\infty} f_0(x)\,dx \right)$$

onde a curva ROC plota parametricamente TPR (T) versus FPR (T) com T como o parâmetro variável.

Design do projeto

Pela descrição do problema, é possível identificar que a solução seria um algoritmo de classificiação binária, e para tal utilizaremos a técnica de Gradient boosting, que é uma das mais utilizadas em competições atualmente.

Inicialmente será realizada uma data analisis, onde utilizaremos exploração visual para entender as características do nosso dataset.