Nanodegree Engenheiro de Machine Learning

Proposta de projeto final

Ricardo Luiz Araujo Junior
7 de abril de 2018

Modelo de Sistema de Recomendação de Filmes

O uso dos sistemas de recomendação nos últimos anos vem aumentado, tanto por questões comerciais quanto para melhorar a experiência do usuário. Principalmente no setor de e-commerce, como a Amazon, e em sites de streaming de filmes e música, podem-se destacar Netflix e Spotify. Um sistema de recomendação basicamente seleciona certo conteúdo baseado no perfil do usuário para lhe fornecer uma experiência personalizada.
Em 2006 a Netflix lançou uma competição para criação do melhor algoritmo de filtragem colaborativa, o mesmo serve para prever classificações de usuários, baseado em classificações prévias. Após esta competição muitas outras surgiram, como o da empresa de streaming de músicas Spotify.
Os sistemas de recomendação estão em constante aperfeiçoamento, conforme os usuários ficam mais exigentes e surgem mais features para avaliarmos. Um dos meus maiores motivos para começar a estudar machine learning foram os sistemas de recomendação. Por isso minha maior motivação para o problema é desenvolver este projeto baseado neste tema, criando um modelo que ofereça o melhor conteúdo de para o usuário, personalizando cada vez mais sua experiência.

Descrição do problema

Irei desenvolver um modelo de sistema de recomendação de filmes, tendo como foco principal prever as classificações de usuários e apresentar os filmes mais recomendados para o mesmo. Podemos utilizar diversos algoritmos para a filtragem colaborativa, que é a base para prever a classificação do usuário para os filmes que ele não assistiu. Os scores de diversos algoritmos serão comparados para que possamos decidir aquele que melhor resolve o problema.

Conjunto de dados

O dataset a ser utilizado será um dos fornecido pela Grouplens, um laboratório de pesquisa do Departamento de Ciência da Computação e Engenharia da Universidade de Minnesota. Este

dataset descreve classificações de 1 a 5 estrelas de usuários do MovieLens, um serviço de recomendação de filmes. Ele contém 100004 classificações de 9125 filmes. Os dados foram criados por 671 usuários entre 9 de janeiro de 1995 e 16 de outubro de 2016. Os usuários foram selecionados de forma aleatória e todos têm pelo menos classificado 20 filmes. O dataset se encontra nesta página https://goo.gl/sm6ggD.
Abaixo segue a descrição do conjunto de dados que irei utilizar e os seus atributos:

Estrutura do dataset de filmes (movies.csv)

Cada linha deste dataset representa um filme.
movieId: ID do respectivo filme.
title: Nome do filme.
genres: Gênero do filme (pode ser Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western, (no genres listed)).

Estrutura do dataset de classificações de filmes (ratings.csv)

Cada linha deste dataset representa a classificação de um filme feita por um usuário.
userId: ID do usuário.
movieId: ID do filme classificado pelo usuário.
rating: Como o usuário classificou aquele filme. Em escala de 5 estrelas, com incrementos de meia (0.5) estrela.
timestamp: Horário da classificação em formato timestamp.

Solução

Para a resolução do problema irei utilizar um algoritmo de filtragem colaborativa. O algoritmo de filtragem colaborativa pode ter como base um algoritmo SVD (Singular-value decomposition), K-NN, Slope One, etc. Irei utilizar o Coeficiente de Correlação de Pearson para medir os pares correlacionados de notas de que usuários deram a certo filme. Os principais algoritmos passarão por uma Busca de Grade exaustiva e mensurados utilizando cálculo do erro quadrático médio. Vou utilizar a biblioteca Surprise feita em Python que possui um design muito parecido com o do Sklearn. Ela foi criada especificamente para desenvolver sistemas de recomendação e se encontra na página surpriselib.com.

Benchmark

Um modelo que serve como benchmark foi um criado por um usuário do Kaggle e utilizado o dataset do Netflix Prize: https://goo.gl/e7SA2h. Ele utilizou o algoritmo SVD em um dataset de 100 mil linhas e para mensurar utilizou erro quadrático médio e erro absoluto médio, com

pontuações médias de 0.98 e 0.78. Para medir a correlação linear entre os filmes, foi utilizado o coeficiente de correlação de Pearson.

Métricas

As métricas de avaliação que podem ser utilizadas são Erro quadrático médio (RMSE), erro absoluto médio (MAE) e o coeficiente de correlação de Pearson . Que serão avaliados no meu modelo proposto e foram avaliados no modelo de benchmark citado anteriormente. As métricas de avaliação poderão ser obtidas através da funções ***accuracy.rmse***, ***accuracy.mae*** e ***similarities.pearson*** da biblioteca Surprise. RMSE e MAE serão utilizados para medir a performance do modelo. E o Coeficiente de correlação de Pearson será utilizado para saber se dois filmes ou dois usuários estão relacionados. Os mesmos têm a mesma função que os das biblioteca Sklearn e Scipy.

Design do projeto

Na fase de análise de dados pretendo verificar se há filmes sem pontuações fornecidas por usuários e removê-los se necessário, pois um filme sem classificação de usuários talvez não seja relevante para recomendação.
Farei uma Busca de Grade nos algoritmos SVD e K-NN para ajustar aquele que obtiver maior pontuação RMSE ou MAE. Além disso, uma medição do tempo de execução pode ser relevante, pois sistemas de recomendação geralmente fazem treinamento em tempo real, e escolher o algoritmo com menor tempo de execução pode otimizar o sistema como um todo.
A fase final será testar as recomendações com base no Coeficiente de correlação de Pearson. Elas podem ser feitas tanto baseadas em similaridades entre usuários quanto de outros itens.