Machine Learning Engineer Nanodegree

Capstone Proposal

Mariusz Kowalewski
August 25, 2019

Proposal

Domain Background

Goodreads is a social cataloging website that allows individuals to search its database of books, quotes and reviews. Users can sign up and register books to generate library catalogs and reading lists. They can also create their own groups of book suggestions, surveys, polls, blogs, and discussions.

Goodreads is the world's largest site for book readers. Being a bookish myself, I think it would be very interesting to explore differences in people's tastes and find recommendations for individual users.

Problem Statement

According to some statistics people read only 12 books per year on average and this figure is probably inflated. Taking into consideration the number of books available in shops and ebooks in the Internet, it is very hard to find one that will be consistent with reader's taste. Moreover, lists of bestsellers are not neceserilly useful for everybody. Having a big database of books, reviews and ratings, it's possible to build an engine able to recommend readers similar books adjusted to their preferences.

Datasets and Inputs

The dataset is provided on Kaggle website. It was obtained using Goodreads API and contains details about the books. The structure of books.csv is as follows:

- bookID - A unique Identification number for each book.
- title - The name under which the book was published.
- authors - Names of the authors of the book. Multiple authors are delimited with -.
- average_rating - The average rating of the book received in total.
- isbn - Another unique number to identify the book, the International Standard Book Number.

- isbn13 - A 13-digit ISBN to identify the book, instead of the standard 11-digit ISBN.
- language_code - Helps understand what is the primary language of the book. For instance, eng is standard for English.
- # num_pages - Number of pages the book contains.
- ratings_count - Total number of ratings the book received.
- text_reviews_count - Total number of written text reviews the book received.

In order to categorize books by genres I'll try to use Goodreads API to create additional dataset.

Solution Statement

One of possible solutions is to use a clustering algorithm like k-means to group readers according to their ratings which in result will give us clusters of readers with similar ratings and generally similar preferences in books. Basing on this, we can calculate the ratings for certain books that weren't read by averaging ratings of other readers in the same cluster.

Benchmark Model

For a benchmark model we can split the data to training and testing datasets and check if some of recommended books are already rated by particular readers. If the rating is at least 4 of 5 score for recommended books, we can draw a conclusion that the model works fine.

We can also try to use Goodreads built-in recommendation system and see if some of the titles are in line with recommendations given by the model.

Evaluation Metrics

In order to evaluate the accuracy of rating prediction we can use Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE). MAE is the most popular and commonly used error function. It is a measure of deviation of recommendation from reader's actual rating. MAE and RMSE are computed as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \qquad \mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

The lower the MAE and RMSE, the more accurate are predictions of user ratings.

As regards the relevance of recommendation lists, we can use Precision and Recall metrics. They help readers select items that are more similar among available set of books.

Project Design

In the first step I will collect the data. As I mentioned above, I will obtain main dataset from Kaggle website. The provided file is in CSV format which will be easy to process.

Next, I'd like to obtain additional genre data using Goodread API and normalize features if necessary.

Then, I will analyze the achieved data and choose the most relevant needed for the recommendation system.

I am considering to apply a few clustering algorithms, including k-means and Gaussian Mixture Model. In each strategy I'd like to test different hyparameters.

After evaluating the performance of each strategy and checking possibilities of combining them I want to create an optimal model achieving the best results and then present the results.

Reference

- Kaggle
- Goodreads API
- Article: How Many Books Does the Average Person Read?
- Article: Recommendation Systems — Models and Evaluation