MLC1-Yelp Recommendation Model

Anshika, Vishnu, Niegil, Sakthisree

Problem Statement

To create a recommendation system for restaurants using collaborative filtering (CF). You will be using the Yelp Dataset for this.

The general structure of a recommendation system is that there are users and there are items. Users express explicit or implicit preferences towards certain items. CF thus relies on users' past behavior.

The goal of this project is to compare different methodologies for recommending local business to users. This involves predicting rating values of business that users have not visited before based on their historical rating records. The performance of our models is mainly measured by Mean Square Error (MSE).

01 **Data Loading, Preprocessing & EDA**

02 **Baseline Matrix Factorization & SGD CF Model**

03 **Deep Learning CF Model**

Data & Labels

yelp_academic_dataset_business.json 118.62 MB

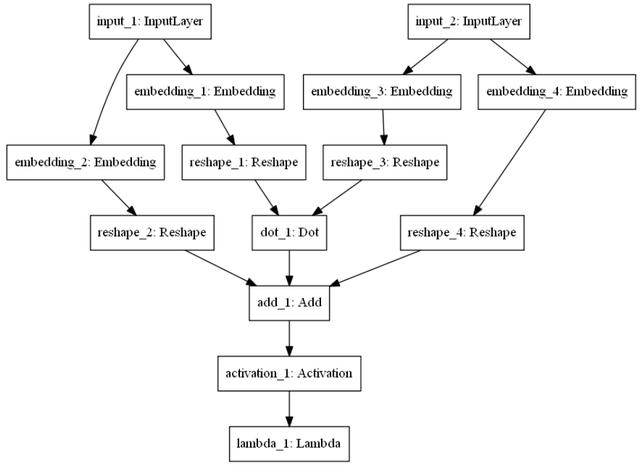
yelp_academic_dataset_review.json 6.46 GB

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 17732 entries, 8 to 209390
Data columns (total 14 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   business_id   17732 non-null  object
 1   name          17732 non-null  object
 2   address       17732 non-null  object
 3   city          17732 non-null  object
 4   state         17732 non-null  object
 5   postal_code   17732 non-null  object
 6   latitude      17732 non-null  float64
 7   longitude     17732 non-null  float64
 8   stars         17732 non-null  float64
 9   review_count  17732 non-null  int64
 10  is_open       17732 non-null  int64
 11  attributes    17326 non-null  object
 12  categories    17732 non-null  object
 13  hours         15650 non-null  object
dtypes: float64(3), int64(2), object(9)
memory usage: 2.0+ MB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8021122 entries, 0 to 8021121
Data columns (total 9 columns):
 #   Column       Dtype
---  ------       -----
 0   review_id    object
 1   user_id      object
 2   business_id  object
 3   stars        float64
 4   useful       int64
 5   funny        int64
 6   cool         int64
 7   text         object
 8   date         object
dtypes: float64(1), int64(3), object(5)
memory usage: 550.8+ MB
```

Models Used

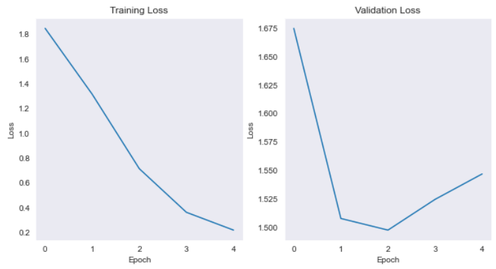
Deep Learning based recommender



Bias SGD Baseline Model

$$\min_{Q*,P*} \sum_{(u,i)\in K} (r_{ui} - P_u^T Q_i)^2 + \lambda(||Q_i||^2 + ||P_u||^2)$$

Results



MSE Value of our Baseline Model

2.71

MSE Value of our Neural Network Model

1.50

Conclusion

- Clearly, the neural network worked better than SGD model.
- Neural Network has a chance of over-fitting and can be furthered looked at.
- Trend of data were analyzed in the EDA phase - most of the reviews were centered around 3-4.
- Cosine similarity based models are hard computationally to scale up but Neural Networks and SGD seem to be computationally better.

Further Work

- Trying out ALS - alternating least squares method.
- Spending more time on tuning model & increasing metric rates.
- Figuring out ways to further work well with sparse matrix.
- Trying out ensemble modelling for better predictions.
- Making a user interface for the model - using streamlit.
- Figuring out ways to effectively handle new user recommendations - handling cold start problems

References

Music artist Recommender System using Stochastic Gradient Descent | Machine...

Learn how to build a Recommender System for music artists by implementing Stochastic Gradient Descent from scratch

Ed — Digital Learning Platform

Ed is the next generation digital learning platform that redefines...

edstem.org

HegdeChaitra/Yelp-Recommendation-System

github.com