Capstone Proposal

1. Domain background

The music resource is used to be very expensive to access during 1990s. Nowadays, with the availability of new devices (such as MP3 and smart phone) and online services (Spontify), in contrast, we are having a hard time to find what we really like from ocean of online music resource.

During the 11th ACM International Conference on Web Search and Data Mining (WSDM 2018), one challenge on building a better music recommendation system using a donated dataset from KKBOX [2] was announced. The competition was hosted as a Kaggle competetion as well.

2. Problem Statement

Even though, one can listen to all kinds of music, algorithms still struggle in recommending the right music to the user. The popular algorithms include collaborative filtering based algorithm with matrix factorization.

However, without enough observed data, how would an algorithm know if listeners will like a new song or a new artist? And, how would it know what songs to recommend brand new users?

In this proposal, we try to easy this challenge by leveraging meta information of users/songs and new deep learning architecture.

3. Datasets

The dataset is from KKBOX, “Asia’s leading music streaming service, holding the world’s most comprehensive Asia-Pop music library with over 30 million tracks.” [3]

In this project we only use the following files.

3.1. train.csv

This dataset includes historical interaction data between user and songs. The user and songs are all tokenized to remove confidential information. In addition, there are three columns related to the source system where the data was collected.

The column “target” are what we’d like to forecast with machine learning models.

For detailed information, please check out the Kaggle dataset page of the KKBOX competition.

3.2. songs.csv and song_extra_info.csv

These two data set includes meta information of songs. For detailed information, please check out the Kaggle dataset page of the KKBOX competition.

3.3. members.csv

These two data set includes meta information of songs. For detailed information, please check out the Kaggle dataset page of the KKBOX competition

We will exclude test.csv since it contains no label information.

4. Solution Statement

Neural Collaborative Filtering (NCF) [1] is a deep learning model framework by which one can not only capture nonlinear interaction between user and item but also take into account the rich meta information of users and items.

5. Benchmark Models

For the benchmark, we will use standard biased Matrix Factorization (MF) and Multi-Layer Perception (MLP) models.

6. Evaluation Metrics

Since the recommendation problem can be seen as a binary classification problem, we will use the following metrics for performance evaluation.

- Precision,
- Recall
- F1 score
- area under the ROC curve

7 Project Design

- Data preprocessing
- Model implementation in PyTorch Lightning
- Local model evaluation
- AWS sage maker model training and deployment

Reference:

[1] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative Filtering. WWW 2-17, 173–182. Link

[2] KKbox challenge: https://www.kaggle.com/c/kkbox-music-recommendation-challenge

[3] KKbox dataset: https://www.kaggle.com/bvmadduluri/wsdm-kkbox