STATS 507 Project Proposal:

The title

Change title and running title

TRONG DAT DO [1,*] and SIMON FONTAINE [1,**]

[1] *University of Michigan, Department of Statistics. West Hall, 1085 South University, Ann Arbor, MI, U.S.A., 48109. E-mail:* [*]dodat@umich.edu; [**]simfont@umich.edu

Todo list

▨ Change title and running title 1
▨ Check spelling of the section 1

Contents

Todo list 1
1 The MovieLens dataset 1
2 Research questions 2
3 Methodology 2
4 Preliminary repot 2
References 2

1. The MovieLens dataset

Check spelling of the section

The MovieLens datasets (Harper and Konstan, 2015) contains user ratings of a variety of movies continuously collected starting from 1998. In addition to the user-movie-rating pairings, the datasets contains information about movie genres, word tagging of movies provided by users and user demographic information.

We will consider the MovieLens 100K Dataset[1], which is one of the multiple datasets provided by GroupLens[2]. We will be interested in this particular dataset because of it contains additional demographic information about the users in the dataset. To include tagging information, we also consider the MovieLens Tag Genome Dataset[3]. Here is a summary of the contents of the datasets that will be used[4]:

[1] Available at https://grouplens.org/datasets/movielens/100k/

[2] Organization website: https://grouplens.org/

[3] Available at https://grouplens.org/datasets/movielens/tag-genome/

[4] From the README.txt file attached to the datasets (http://files.grouplens.org/datasets/movielens/ml-100k-README.txt, http://files.grouplens.org/datasets/tag-genome/README.html)

`MovieLens 100K Dataset` The dataset was collected from the `MovieLens` website (`movielens.umn.edu`) between September 19th, 1997 through April 22nd, 1998. It has been pre-processed and cleaned to include only examples where the users have made at least 20 ratings during the collection period and where demographic information are complete. In the `u.data` file, there are 100,000 ratings on the scale of 1 to 5, taking only integer values. It contains the following entries: `user id`, `item id`, `rating`, `timestamp`. In the `u.item` file, there are 1,682 movies with the following information: `movie id`, `movie title`, `release date`, `IMDb URL` and 19 columns indicating movie genre with 0-1 encoding where 1 denotes that the movie is of the corresponding genre. In the `u.user` file, there are 943 users with the following information: `user id`, `age`, `gender`, `occupation` (see `u.occupation` file for details) and `zip code`.

`MovieLens Tag Genome Dataset` This dataset contains tagging information of 9,734 movies and 1,128 tags. In particular, the `tag_relevance` file contains the relevance of all tags for all movies reported on a continuous scale from 0 to 1, where 1 indicates strong relevance.

2. Research questions

3. Methodology

We will

4. Preliminary repot

We will

References

Harper, F. M. and Konstan, J. A. (2015). The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems* **5** 1–19.