Final Project Spec

TA email: `ml2017ta@csie.ntu.edu.tw`

1 Problem Description

In this project, you are going to play with data from Book-Crossing dataset, which contains information of many user IDs, book descriptions and book ratings (integer value from 1 to 10), etc. We also provide the implicit book rating data (value 0) and hope you can make the best use of them if possible. Your goal is to predict the book rating, and you should notice that different tracks have different criteria. Besides the original Book-Crossing dataset, we additionally provide the external data crawled by TAs, which are the descriptions of some books (not all the books). **Per course policy, you are NOT allowed to use any other data**.

2 Data Description

There are several files which contain all the information about the task:

- `users.csv`: The user ids and corresponding demographic data
  - User-ID: user IDs which have been anonymized
  - Location: demographic data (may contain NULL-value)
  - Age: demographic data (may contain NULL-value)
- `book_ratings_train/test.csv`: Containing the book rating information
  - User-ID: user IDs which have been anonymized
  - ISBN: International Standard Book Number which you can find some description through this (just like the Book ID)
  - Book-Rating: the book ratings range from 1 to 10 (Note that the test data would not have this value)
- `implicit_ratings.csv`: Containing the book rating information
  - User-ID: user IDs which have been anonymized
  - ISBN: International Standard Book Number which you can find some description through this (just like the Book ID)
  - Book-Rating: all the book ratings are implicit, that is 0
- `books.csv`: Contain content information about books.
  - ISBN: International Standard Book Number
  - Book-Title: content-based information
  - Book-Author: content-based information
  - Year-Of-Publication: content-based information
  - Publisher: content-based information
  - Image-URL-S: URLs linking to cover images (small size)
  - Image-URL-M: URLs linking to cover images (medium size)
  - Image-URL-L: URLs linking to cover images (large size)
  - Book-Description: TA-crawled descriptions of books
- `submission.csv`: Your book-rating predictions for testing samples.
  - Book-Rating: the predicted book ratings (Please note that you **don't** need to provide the header in the first row)

3 Evaluation

We will have two tracks of competition, each evaluated with a different goodness measure.

- Track 1: Mean Absolute Error (MAE)

  Mean Absolute Error is a common evaluation criterion to measure the average magnitude of the errors, which doesnt consider the direction. In this track, your hypothesis should predict the **integer** book rating $\hat{y}$ and we would calculate the error through the Mean Absolute Error. That is,

  $$\text{MAE}(y_i, \hat{y}_i) = \frac{1}{n_{samples}} \sum_{n=1}^{n_{samples}} |y_i - \hat{y}_i|, \text{ where } \hat{y}_i \in \mathbb{Z}$$

  where $\hat{y}_i$ is the predicted value of the i-th sample and y_i is the corresponding true value. $n_{samples}$ means the number of testing samples

- Track 2: Mean Absolute Percentage Error (MAPE)

  In many kinds of scenario, the company that runs book selling service may focus on the books which have low rating, because sometimes the values among high ratings may have no difference. Therefore, we choose the MAPE (focus on the low value) to measure the error. In this track, your hypothesis g should predict the **float** book rating $\hat{y}$. Then, we will take

  $$\text{MAPE}(y_i, \hat{y}_i) = \frac{100\%}{n_{samples}} \sum_{n=1}^{n_{samples}} |\frac{y_i - \hat{y}_i}{y_i}|, \text{ where } \hat{y}_i \in \mathbb{R}$$

  where $\hat{y}_i$ is the predicted value of the i-th sample and y_i is the corresponding true value. $n_{samples}$ means the number of testing samples.