CSE 255, Fall 2015: Assignment 1

Instructions

In this assignment you will build **recommender systems** to make predictions related to reviews of Books on *Amazon*.

Solutions will be graded on Kaggle (see below), with the competition closing at **midnight, November 17** (note that the time reported on the competition webpage is in UTC!).

You will also be graded on a brief report, to be submitted electronically via an online form (`https://goo.gl/frJ5eB`) on or before November 18 (the day after the competition closes). Your grades will be determined by your performance on the predictive tasks as well as your written report about the approaches you took.

This assignment should be completed **individually**.

To begin, download the files for this assignment from:
`http://jmcauley.ucsd.edu/data/assignment1.tar.gz`

Files

train.json.gz 1,000,000 reviews to be used for training. It is not necessary to use *all* reviews for training, for example if doing so proves too computationally intensive. While these files are one-json-per-line (much as we have seen so far in class), you may find it useful to represent them more concisely in order to produce a more efficient solution. The fields in this file are:

- **itemID** The ID of the item. This is a hashed product identifier from Amazon.
- **reviewerID** The ID of the reviewer. This is a hashed user identifier from Amazon.
- **helpful** Helpfulness votes for the review. This has two subfields, 'nHelpful' and 'outOf'. The latter is the total number of votes this review received, the former is the number of those that considered the review to be helpful.
- **reviewText** The text of the review. It should be possible to successfully complete this assignment *without* making use of the review data, though an effective solution to the 'helpfulness prediction' task will presumably make use of it.
- **summary** Summary of the review.
- **unixReviewTime** Time of the review in seconds since 1970.
- **reviewTime** Plain-text representation of the review time.
- **category** Category labels of the product being reviewed.

pairs_Helpful.txt Pairs on which you are to predict helpfulness votes. A third column in this file is the total number of votes, from which you should predict how many were helpful.

pairs_Rating.txt Pairs (userIDs and itemIDs) on which you are to predict ratings.

helpful.json.gz The review data associated with the helpfulness prediction *test* set. The 'nHelpful' field has been removed from this data, since that is the value you need to predict above. This data will only be of use for the helpfulness prediction task.

baselines.py A simple baseline for each task, described below.

Please do not try to crawl these products from Amazon, or to reverse-engineer the hashing function I used to anonymize the data. I assure you that doing so will not be easier than successfully completing the assignment.

Tasks

You are expected to complete the following tasks:

Helpfulness prediction Predict whether a user's review of an item will be considered helpful. The file 'pairs_Helpful.txt' contains (user,item) pairs, with a third column containing the number of votes the user's review of the item received, you must predict how many of them were helpful. Accuracy will be measured in terms of the total *absolute error*, i.e., you are penalized one according to the difference |nHelpful − prediction|, where 'nHelpful' is the number of of helpful votes the review actually received, and 'prediction' is your prediction of this quantity.

Rating prediction Predict people's star ratings as accurately as possible, for those (user,item) pairs in 'pairs_Rating.txt'. Accuracy will be measured in terms of the *(root) mean-squared error* (RMSE).

These error measures are described on *Kaggle*:

Absolute error `https://www.kaggle.com/wiki/AbsoluteError`

RMSE `https://www.kaggle.com/wiki/RootMeanSquaredError`

A competition page will be set up on Kaggle to keep track of your results compared to those of other members of the class. The leaderboard will show your results on *half of* the test data, but your ultimate score will depend on your predictions across the *whole* dataset.

Grading and Evaluation

This assignment is worth 25% of your grade. You will be graded on the following aspects. Each of the two tasks is worth 10 marks (i.e., 10% of your grade), plus 5 marks for the written report.

- Your ability to obtain a solution which outperforms the baselines on *the unseen portion of* the test data (5 marks for each task). Obtaining full marks requires a solution which is substantially better (i.e., at least several percent) than baseline performance.
- Your ranking for each of the tasks compared to other students in the class (3 marks for each task).
- Obtain a solution which outperforms the baselines on *the seen portion of* the test data (i.e., the leaderboard). This is a consolation prize in case you overfit to the leaderboard. (2 mark for each task).

Finally, your written report should describe the approaches you took to each of the tasks. To obtain good performance, you should not need to invent new approaches (though you are more than welcome to!) but rather you will be graded based on your desicion to apply reasonable approaches to each of the given tasks (5 marks total).

Baselines

Simple baselines have been provided for each of the tasks. These are included in 'baselines.py' among the files above. These baselines operate as follows:

Helpfulness prediction Multiply the number of votes by the global average helpfulness rate, or the user's rate if we saw this user in the training data.

Rating prediction Return the global average rating, or the user's average if we have seen them before in the training data.

Running 'baselines.py' produces files containing predicted outputs. Your submission files should have the same format.

Kaggle

We have set up a Kaggle page to help you evaluate your solution (please contact Sheeraz if you didn't receive the invitation). The Kaggle pages for each of the tasks are:
`https://inclass.kaggle.com/c/cse-190-255-fa15-assignment-1-task-1-helpfulness-prediction/`
`https://inclass.kaggle.com/c/cse-255-fa15-assignment-1-task-2-rating-prediction/`