CSE 190, Spring 2015: Homework 3

Download the "Video Game Reviews" data from the course webpage:
`http://jmcauley.ucsd.edu/cse190/data/homework3.tar.gz`

This data is equivalent to the data provided for Assignment 1, with three crucial differences:

1. It is for a different category (Video Games)
2. It is smaller (1/5th the size)
3. Labels for the test data **have been provided** ('labeled_*.txt')

These homework exercises are intended to help you get started on potential solutions to Assignment 1.

Tasks

1. Using the *training* data ('train.json.gz') fit a simple predictor of the form

$$\text{rating}(\text{user}, \text{item}) \simeq \alpha.$$

   Report α and the MSE of your predictor against on the *test* data ('labeled_Rating.txt') (2 marks).

2. Fit a predictor of the form

$$\text{rating}(\text{user}, \text{item}) \simeq \alpha + \beta_{\text{user}} + \beta_{\text{item}},$$

   by fitting the mean and the two bias terms as described in the lecture notes (with the regularization parameter $\lambda = 1$). Report the item bias $\beta_{\text{I102776733}}$ and the user bias $\beta_{\text{U566105319}}$, and the MSE of the predictor against the test data (2 mark).

3. The following experiments should be conducted on the *training data* ('train.json.gz'):

   (a) Compute the Jaccard similarity between the users 'U229891973' and 'U622491081' in terms of the sets of items they have reviewed in the training set (1 mark).

   (b) Find the user (or users) with the highest Jaccard similarity to 'U622491081' (if there are multiple users with the same maximum similarity, list all of them) (1 mark).

4. Using the *training* data ('train.json.gz') fit a simple predictor of the form

$$\frac{\text{nHelpful}}{\text{outOf}} \simeq \alpha$$

   (see the 'helpful' field in each review).

   (a) What is the fitted value of α (1 mark)?

   (b) The four columns in the test file 'labeled_Helpful.txt' correspond to (user,item,outOf,nHelpful) quadruples. Predict 'nHelpful' by multiplying your predictor (α) above by 'outOf' for each quadruple. What is the MSE of these predictions, and what is the Absolute error (see `https://www.kaggle.com/wiki/AbsoluteError`) (1 mark)?

   (c) To fit the same quantity, train a predictor of the form

$$\frac{\text{nHelpful}}{\text{outOf}} \simeq \alpha + \beta_1(\#\text{ words in review}) + \beta_2(\text{review's rating in stars}).$$

   Report the fitted parameters (1 mark).

   (d) To compute the error of the above predictor on the test data, you will need to use the file 'helpful.json.gz'. This file contains all of the *features* associated with the test pairs, but not their labels (i.e., not 'nHelpful'). Using the features from this file, compute the MSE and the Absolute error (1 mark).