CSE 255, Winter 2015: Homework 7

Instructions

Please submit your solution **at the beginning of the next lecture (February 23)** or outside of CSE 4102 beforehand. Please complete homework **individually**. I will accept homework as late as **Wednesday morning** in 4102 in light of the assignment being due on Monday.

Download the "Video Game Reviews" data from the course webpage:
`http://jmcauley.ucsd.edu/cse255/data/homework6_7.tar.gz`

This data is equivalent to the data provided for Assignment 2, with three crucial differences:

1. It is for a different category (Video Games)
2. It is smaller (1/10th 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 2.

Tasks

1. 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).

2. In practice, we may not want to consider a review with 1/1 helpfulness votes to be 'more helpful' than one with 48/50; rather we would conclude that the 1 helpfulness vote may just have been down to luck. Suggest a scheme for ordering reviews by helpfulness that considers both the ratio and the total number of helpfulness votes in order to rank reviews (1 mark).