CSE 255, Winter 2015: Homework 2

Instructions

Please submit your solution **at the beginning of the next lecture (January 26)** or outside of CSE 4102 beforehand. Please complete homework **individually**.

Download the "50,000 book reviews" data, as well as the "5,000 book images" data from the course webpage:
`http://jmcauley.ucsd.edu/cse255/data/amazon/book_descriptions_50000.json`
`http://jmcauley.ucsd.edu/cse255/data/amazon/book_images_5000.json`

You will also need the code stub from `http://jmcauley.ucsd.edu/cse255/code/homework2.py`

Code is provided on the course webpage showing how to load and perform simple processing on the data. Executing the code requires a working install of Python 2.7 with the scipy packages installed.

Tasks

1. Download the book descriptions data. For this and the next question we will consider identifying "Romance" novels based on words in their descriptions. Based on all 50,000 descriptions, write down
   (a) p(has category "Romance")
   (b) p(mentions "love" in description | has category "Romance")
2. Implement a naïve Bayes classifier to predict p(has category "Romance"|mentions "love" in description $\wedge$ mentions "relationship" in description). Following the naïve Bayes assumption, compute the value of

$$\frac{p(\text{has category "Romance"} \mid \text{mentions "love" in description} \wedge \text{mentions "relationship" in description})}{p(\text{doesn't have category "Romance"} \mid \text{mentions "love" in description} \wedge \text{mentions "relationship" in description})}$$

3. Download the code stub from `http://jmcauley.ucsd.edu/cse255/code/homework2.py`. This code is an implementation of logistic regression using gradient **ascent**. Functions are provided for the log-likelihood and its derivative. What is the log-likelihood obtained (on the training data) by setting the 4096-dimensional parameter vector $\theta = \mathbf{0}$?
4. (2 marks) Code for the log-likelihood has been provided in the code stub (`f`), but code for the derivative is incomplete (`fprime`).
   (a) Complete the code stub for the derivative (`fprime`) and provide your solution (1 mark).
   (b) What is the log-likelihood obtained (on the training data) when the model converges (1 mark)?

   Hint: to debug your code, set the regularization parameter to zero – the log-likelihood should approach 0 (i.e., you should be able to overfit very badly) if your implementation of the derivative is correct.