CSE 190, Spring 2015: Homework 1

Instructions

Please submit your solution **at the beginning of the week 3 lecture (April 14)** or outside of CSE 4102 beforehand. Please complete homework **individually**.

You will need the following files:

50,000 beer reviews : `http://jmcauley.ucsd.edu/cse190/data/beer/beer_50000.json`. You may also use the non-alcoholic beer data if you prefer, though please mention it if you did: `http://jmcauley.ucsd.edu/cse190/data/beer/non-alcoholic-beer.json`

Book descriptions : `http://jmcauley.ucsd.edu/cse255/data/amazon/book_descriptions_50000.json`

Code examples : `http://jmcauley.ucsd.edu/cse190/code/week1.py` (regression) and `http://jmcauley.ucsd.edu/cse190/code/week2.py` (classification)

Executing the code requires a working install of Python 2.7 with the scipy packages installed.

Tasks (Regression, week 1):

1. Compute the following statistics about the data: (1) number of unique items ('beer/beerId'), (2) number of unique users ('user/profileName'), (3) mean for each of the five ratings ('review/appearance', 'review/palate', 'review/overall', 'review/aroma', 'review/taste'), (4) mean ABV ('beer/ABV') (1 mark).
2. What is the variance of the 'review/taste' scores in the data? What is the Mean Squared Error (MSE) obtained when predicting the 'review/taste' score using the mean value obtained above (1 mark)?
3. Using ordinary linear regression, train a predictor that uses the ABV ('beer/ABV') to predict the taste rating ('review/taste'), i.e.,

$$\text{review/taste} \simeq \theta_0 + \theta_1 \times \text{beer/ABV}.$$

   You may use Python libraries to do so. What are the fitted values of θ_0 and θ_1? What are the interpretations of these fitted values (1 mark)?
4. Split the data into two equal fractions – the first half for training, the second half for testing (based on the order they appear in the file). Train the same model as above *on the training set only.* What is the model's MSE on the training and on the test set (1 mark)?
5. Suppose you want to incorporate the time of day into your predictor of the 'taste' rating (e.g. because you believe people may evaluate the taste more positively or negatively at certain times of the day). How would you construct a feature vector to represent this quantity? Using the same train/test split as above, train a model using this feature and report its performance (MSE) on the train/test sets. What conclusions can you draw (if any) can you draw from the fitted values (2 marks)?

Tasks (Classification, week 2):

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")

   (1 mark)
2. Following the naïve Bayes assumption, compute the value of

$$\frac{p(\text{has category “Romance”} \mid \text{mentions “love” in description} \wedge \text{mentions “beaut” in description})}{p(\text{doesn't have category “Romance”} \mid \text{mentions “love” in description} \wedge \text{mentions “beaut” in description})}.$$

   Why might the string 'beaut' be more effective than 'beauty'/'beautiful' (1 mark)?
3. Implement a naïve Bayes classifier using the above two features. Report the above probability for the four possible feature combinations (mentions "love" and "beaut"; mentions "love" and *doesn't* mention "beaut" etc.). On the 50,000 books report its true-positive, false-positive, true-negative, and false-negative rate (2 marks).