CSE 258, Fall 2017: Homework 1

Instructions

Please submit your solution **by the beginning of the week 3 lecture (Oct 16).** Submissions should be made on **gradescope**. Please complete homework **individually**.

You will need the following files:

50,000 beer reviews : `http://jmcauley.ucsd.edu/cse258/data/beer/beer_50000.json`

Code examples : `http://jmcauley.ucsd.edu/cse258/code/week1.py` (regression) and `http://jmcauley.ucsd.edu/cse258/code/week2.py` (classification)

Executing the code requires a working install of Python 2.7 or Python 3 with the scipy packages installed. **Please include the code of (the important parts of) your solutions.**

Tasks — Regression (week 1):

In the first three questions, we'll see how ratings vary across different categories of beer. These questions should be completed on the *entire dataset*.

1. How many reviews are there for each style of beer in the dataset ('beer/style')? What is the average value of 'review/taste' for reviews from each style? (1 mark)

2. Train a simple predictor with a single binary feature indicating whether a beer is an 'American IPA':

$$\text{review/taste} \simeq \theta_0 + \theta_1 \times [\text{beer is an American IPA}]$$

   Report the values of θ_0 and θ_1. Briefly describe your interpretation of these values, i.e., what do θ_0 and θ_1 represent (1 mark)?

3. 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)?

4. Extend the model above so that it incorporates binary features for *every* style of beer with ≥ 50 reviews. Report the values of θ that you obtain, and the model's MSE on the training and on the test set (1 mark).

Tasks — Classification (week 2):

Next we'll try to train classifiers that are able to predict a beer's style from the characteristics of its review. Again, split the data so that the first half is used for training and the second half is used for testing as we did for Q3.

5. First, let's train a predictor that estimates whether a beer is an 'American IPA' using two features:

$$[\text{'beer/ABV'}, \text{'review/taste'}].$$

   Train your predictor using an SVM classifier (see the code provided in class) – remember to train on the first half and test on the second half. Use a regularization constant of $C = 1000$ as in the code stub. What is the accuracy (percentage of correct classifications) of the predictor on the train and test data? (1 mark)

6. Considering the 'American IPA' style, can you come up with a more accurate predictor (e.g. using features from the text, or otherwise)? Write down the feature vector you design, and report its train/test accuracy (1 mark).

7. What effect does the regularization constant C have on the training/test performance? Report the train/test accuracy of your predictor from the previous question for $C \in \langle 0.1, 10, 1000, 100000 \rangle$.

8. **(Hard)** Finally, let's fit a model (for the problem from Q5) using logistic regression. A code stub has been provided to perform logistic regression using the above model on `http://jmcauley.ucsd.edu/cse258/code/homework1.py` Code for the log-likelihood has been provided in the code stub (`f`) but code for the derivative is incomplete (`fprime`)

   Complete the code stub for the derivative (`fprime`) and provide your solution. What is the log-likelihood of after convergence, and what is the accuracy (on the test set) of the resulting model (1 mark)?