CSE 158, Fall 2018: Homework 1

Instructions

Please submit your solution **by the beginning of the week 3 lecture (Oct 15).** 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/cse158/code/week1.py` (regression) and `http://jmcauley.ucsd.edu/cse158/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. What is the distribution of ratings in the dataset (for 'review/taste')? That is, how many 1-star, 2-star, 3-star (etc.) reviews are there? You may write out the values or include a simple plot (1 mark).

2. Among beers with ≥ 5 reviews, which has the highest average rating? (Report multiple if there is a tie) (1 mark)

3. Train a simple predictor to predict a beer's 'taste' score using two features:

$$\text{review/taste} \simeq \theta_0 + \theta_1 \times [\text{beer is a Hefeweizen}] + \theta_2 \times \text{beer/ABV}$$

   Report the values of θ_0, θ_1, and θ_2. Briefly describe your interpretation of these values, i.e., what do θ_0, θ_1, and θ_2 represent (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. Using the first half for training and the second half for testing may lead to unexpected results (e.g. the training error could be higher than the test error). Repeat the above experiment by using a random 50% split of the data (i.e., half for training, half for testing, after first shuffling the data). Report the MSE on the train and test set, and suggest one possible reason why the result may be different from the previous experiment (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.

6. First, let's train a predictor that estimates whether a beer is a 'Hefeweizen' using five features describing its rating:

$$[\text{'review/taste'}, \text{'review/appearance'}, \text{'review/aroma'}, \text{'review/palate'}, \text{'review/overall'}].$$

   Train your predictor using an SVM classifier (see the code provided in class). Use a *random* split of the data as we did in Question 5. 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)

7. Considering same prediction problem as above, 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).

8. 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 $C \in \langle 0.1, 10, 1000, 100000 \rangle$.