CSE 158/258, Fall 2019: Homework 1

Instructions

Please submit your solution **by the beginning of the week 3 lecture (Oct 14).** Submissions should be made on **gradescope**. Please complete homework **individually**.

This specification includes both questions from the undergraduate (CSE158) and graduate (CSE258) classes. You are welcome to attempt questions from both classes but will only be graded on those for the class in which you are enrolled. MGTA495 students should attempt CSE258 questions.

You will need the following files:

Amazon Gift Card data : `https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Gift_Card_v1_00.tsv.gz` The above is a *TSV* formatted dataset, including reviews from one of the smaller Amazon categories. Data can be read using the Python *csv.reader* library.

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):

First, let's see how ratings can be predicted as a function of (a) whether a review is a 'verified purchase', and (b) the length of the review (in characters).

1. What is the distribution of ratings in the dataset? 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. **(CSE158 only)** Repeat the above question, but generate the distribution (a) only for reviews that are 'verified,' and (b) only for reviews that are *not* verified. Write out the values or generate a plot to show the *difference* between these distributions (1 mark).

3. Train a simple predictor to predict the star rating using two features:

$$\text{star rating} \simeq \theta_0 + \theta_1 \times [\text{review is verified}] + \theta_2 \times [\text{review length}]$$

   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? Explain these in terms of the features and labels, e.g. if the coefficient of 'review length' is negative, what would that say about positive versus negative reviews (1 mark)?

4. Train another predictor that only uses *one* feature:

$$\text{star rating} \simeq \theta_0 + \theta_1 \times [\text{review is verified}]$$

   Report the values of θ_0 and θ_1. Note that coefficient you found here might be quite different (i.e., much larger or smaller) than the one from Question 3, even though these coefficients refer to the same feature. Provide an explanation as to why these coefficients might vary so significantly (1 mark).[1]

5. Split the data into two fractions – the first 90% for training, and the remaining 10% testing (based on the order they appear in the file). Train the same model as in Question 4 *on the training set only.* What is the model's MSE on the training and on the test set (1 mark)?

6. **(CSE158 only)** Using the test set from Question 5, report the Mean Absolute Error (MAE) and R^2 coefficient for your predictor (on the test set) (1 mark).

7. **(CSE258 only)** Repeat the above experiment, varying the size of the training and test fractions between 5% and 95% for training (using the complement for testing). Show how the training and test error vary as a function of the training set size (again using a simple plot or table). Does the size of the training set make a significant difference in testing performance? Comment on why it might or might not make a significant difference in this instance (2 marks).

[1] Hint: you should consider *both* of the features from Question 3 in your explanation.

Tasks — Classification (week 2):

In this question we'll alter the prediction from our regression task, so that we are now classifying whether a review is verified. Continue using the 90%/10% training and test sets you constructed previously, i.e., *train on the training set and report the error/accuracy on the testing set.*

8. First, let's train a predictor that estimates whether a review is verified using the rating and the length:

$$p(\text{review is verified}) \simeq \sigma(\theta_0 + \theta_1 \times [\text{star rating}] + \theta_2 \times [\text{review length}])$$

   Train a logistic regressor to make the above prediction (you may use a logistic regression library with default parameters, e.g. *linear_model.LogisticRegression()* from *sklearn*). Report the classification accuracy of this predictor. Report also the proportion of *labels* that are positive (i.e., the proportion of reviews that are verified) and the proportion of predictions that are positive (1 mark).

9. Considering same prediction problem as above, can you come up with a more accurate predictor (e.g. using features from the text, timestamp, etc.)? Write down the feature vector you design, and report its train/test accuracy (1 mark).