CSE 158/258, Fall 2020: Homework 1

Instructions

Please submit your solution **by the beginning of the week 3 lecture (Oct 19).** 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.

You will need the following files:

GoodReads Fantasy Reviews : `https://cseweb.ucsd.edu/classes/fa20/cse258-a/data/fantasy_10000.json.gz`

Beer Reviews : `https://cseweb.ucsd.edu/classes/fa20/cse258-a/data/beer_50000.json` The above is a *json* formatted dataset. Data can be read using the *json.loads* function in Python, or by using *eval*.

Code examples : `http://cseweb.ucsd.edu/classes/fa19/cse258-a/code/week1.py` (regression) and `http://cseweb.ucsd.edu/classes/fa19/cse258-a/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 whether ratings can be predicted as a function of review length, or the number of comments on the review.

Use GoodReads

1. **(CSE158 only)** What is the distribution of ratings and review lengths in the dataset? Report the number of 1-, 2-, 3-star (etc.) ratings, and show the relationship with length (e.g. via a scatterplot) (1 mark).

2. Train a simple predictor that estimates rating from review length, i.e.,

$$\text{star rating} \simeq \theta_0 + \theta_1 \times [\text{review length in characters}]$$

   Report the values θ_0 and θ_1, and the Mean Squared Error of your predictor (on the entire dataset) (1 mark).

3. Re-train your predictor so as to include a second feature based on the number of comments, i.e.,

$$\text{star rating} \simeq \theta_0 + \theta_1 \times [\text{length}] + \theta_2 \times [\text{number of comments}]$$

   Report the values of θ_0, θ_1, and θ_2, along with the MSE of the new model. Briefly explain why the coefficient θ_1 in this model is different from the one in Question 2.

4. Train a model that fits a polynomial function to estimate ratings based on review length. I.e.,

$$\text{star rating} \simeq \theta_0 + \theta_1 \times [\text{length}] + \theta_2 \times [\text{length}]^2 + \theta_3 \times [\text{length}]^3$$

   Fit polynomials up to degree five (i.e., including up to $[\text{length}]^5$) and report the MSE of each. Hint: instead of fitting length directly, you can rescale the feature to be between zero and one by dividing by the maximum length in the dataset; this may help to prevent numerical stability issues.

5. Repeat the above question, but this time split the data into a training and test set. You should split the data randomly into 50%/50% train/test fractions. Report the MSE of each model separately on the training and test sets.

6. **(CSE258 only)** Show that for a trivial predictor, i.e., $y = \theta_0$, the best possible value of θ_0 in terms of the Mean Squared Error is $\bar{y}$ (i.e., the average value of the label y). Hint: compute the derivative of the model's MSE and solve for θ_0

Tasks — Classification (week 2):

In this question we'll try to predict user gender based on users' beer reviews. Load the 50,000 beer review dataset, discarding any entries that don't include a specified gender.

7. Fit a logistic regressor that estimates gender from review length, i.e.,

$$p(\text{gender is female}) \simeq \sigma(\theta_0 + \theta_1 \times [\text{length}])$$

   Report the True Positive, True Negative, False Positive, False Negative, and Balanced Error Rates of the predictor.

8. Retrain the regressor using the `class_weight='balanced'` option, and report the same error metrics as above.

9. Improve your predictor (i.e., reduce the balanced error rate) by incorporating additional features from the data (e.g. beer styles, ratings, features from text, etc.)