CSE 255, Winter 2015: Homework 1

Instructions

Please submit your solution **at the beginning of the next lecture (January 12)** or outside of CSE 4102 beforehand. Please complete homework **individually**.

Download the "50,000 beer reviews" data from the course webpage: `http://jmcauley.ucsd.edu/cse255/data/beer/beer_50000.json`. Code is provided on the course webpage showing how to load and perform simple processing on the data. Executing the code requires a working install of Python 2.7 with the scipy packages installed.

Tasks

Each task is worth 1 (out of 5) points. Report solutions to six significant figures.

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').

2. What is the Mean Squared Error (MSE) obtained when predicting the 'review/overall' score using the mean value obtained above?

3. Using ordinary linear regression, train a predictor that uses the ABV ('beer/ABV') to predict the overall rating ('review/overall'), i.e.,

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

   What are the fitted values of θ_0 and θ_1?

4. Split the data into two equal fractions – reviews 1 to 25,000 for training, and reviews 25,001 to 50,000 for testing. 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?

5. Suppose you want to train a linear predictor that uses the month to predict people's ratings ('review/timeStruct'/'mon'). How would you represent the month of the year using a linear model? Using your representation, write down the feature vectors of the first five reviews.