Technique Assignment 3: Linear regression
Cogs 109 Spring 2020
Due: Tuesday May 12
100 points total

Submit your completed assignment on Gradescope as a pdf report. Additionally, include all Python code you generated either as part of a Jupyter notebook or as an appendix to your report. Include comments for clarity.

1. **(15 points) Datasets and variables**

   (9) Find a dataset from the UCI machine learning repository (http://archive.ics.uci.edu/ml/index.php) that is suitable for linear regression. Provide a link to your chosen dataset and briefly describe its content.

   List the following:

   - number of variables
   - number of samples
   - labels (what is the label?)

   (6) Create and report a research question that you could answer using this dataset and some or all of the variables.

2. **(20 points) Arrays and numpy**

   a. (10) Use `arange()` and `reshape()` from numpy to create a 4x5 array containing the integers 1 through 20. Append a column of ones to the left of your array to create a 4x6 array. Multiply every element of the array by 2. Print **only** the resulting array.

   b. (10) Use `linspace()` and `reshape()` to create a 20x20 array that contains a smooth range of values between 0 and 1, inclusive.

      Create a scatter plot using the first (leftmost) column of your array as the x values and the last (rightmost) column of your array as the y values. Use x limits 0 and 1 and y limits 0 and 1 for your plot.

3. **(40 points)** Univariate linear regression

Note: Solutions to this problem must follow the method described in class and the linear regression handout. There is some flexibility in how your solution is coded, but you may not use special functions that automatically perform linear regression for you.

Load in the BrainBodyWeight.csv dataset. Perform linear regression using two different models:

M1: brain_weight = w0 + w1 x body_weight

M2: brain_weight = w0 + w1 x body_weight + w2 x body_weight^2

   a. (15) For each model, follow the steps shown in class to solve for w. Report the model, including w values and variable names for both models.

   b. Use subplots to display two graphs, one for each model. In each graph, include:

- Labeled x and y axes
- Title
- Scatterplot of the dataset
- A smooth line representing the model

   c. For each model, calculate the sum squared error (SSE). Show your 2 SSE values together in a bar plot.

   d. Which model do you think is better? Why? Is there a different model that you think would better represent the data?

4. (25 points) Multivariate regression with cross validation

   Using the dataset found in Housing.csv, build a multivariate model to predict house price using lot size and the number of bedrooms as predictors.

   Hint: You may use this as your model:

   Price = w0 + w1 x Lot size + w2 x Bedrooms

   First, split your data into a training set (80%) and a test set (20%). Then perform linear regression using the **training data** only.

   Report your model and show the mean squared error (MSE) for your **training** and **test** data using a bar graph.

   MSE can be found by dividing SSE by the number of samples in your data.

Extra credit (up to 10 points)

   What model(s) you would use to compare to the model used in Question 4? Explain why you would choose this model (or models).

   Redo the analysis in Question 4 using a different model and explain your results, comparing them to the results in Question 4.