Homework#2-part 1: Linear Regression

Part 1

A. Linear Regression with one variable

Consider the attached file dataset1.txt. The first column of the data file shows the input data (x), and the second column shows each samples' output value (y).

1. What is the cost function $J(\theta)$ equation for linear regression?

2. Fit a linear regression model on your data using:
   a. Closed-form solution calculated by MSE method
   b. Gradient descent method in online (stochastic) mode (1500 iterations)
   c. Gradient descent method in batch mode (1500 iterations)

3. Plot the dataset and superimpose the fitted models using the three above methods.

4. Use each estimated parameter θ (for each method) to predict the output for x=6.2, 12.8, 22.1, 30.

5. Compare the parameter θ estimated by each method by plotting them in one figure.

6. Plot the cost function $J(\theta)$ along the epochs (plot both online & batch methods on one figure using hold on command).

7. Which type of G.D. (online\batch) do you prefer here? Why?

B. Multiple variable Regression

In this problem, we want to estimate the medical cost for each person due to their traits by applying linear regression. The required train and test data are provided in train.csv and test.csv. Each data sample contains six features(x) and one output (y) related to medical cost. You could use the Pandas library to read csv files.

Note: As it can be observed by reviewing the data samples, each sample contains some categorical features such as sex, region, if a person smokes or not. To encode sex and smoking variables, apply integer encoding, and for region variable, apply one-hot encoding (OHE). These methods are well explained here.

Q: Why is OHE better at encoding regional features compared to integer encoding?

1. Calculate w by closed-form solution calculated by MSE method.

Note: set the basis function for feature bmi to x^2. (use bmi^2 instead of bmi feature).

2. At first, assume that the total training data is 10, then gradually increase the training data (with step size equals to 10) until reach out to 1000. In each step, report the test error (MSE) in a diagram.

3. Implement batch gradient descent and stochastic gradient descent to solve this problem. Report the train and test error in each case.

4. Now, add a L2 regulator to the cost function and solve the problem using closed-form. Specify the best value for regularizer parameter (λ) between 10^{-4} , 10^{-3} , ..., 1, ..., 10^{3}, 10^{4} . Plot both train and test errors concerning the logarithmic amount of λ.

5. compare the results with and without regularization.

Report:

- Prepare a PDF format report including the figures and answers to the questions.

- Make a folder, including your report and your code.

- Submit all your files in a zipped folder named as “HW2_student name_studentnumber.zip”

- Dataset link: https://www.dropbox.com/s/5pw3hmzzosqkoid/HW2_Dataset.zip?dl=0