In The Name of Allah

Machine Learning (Winter 2018)

Instructor: Mahdi Yazdian

TA: Farzane Nadi

Homework#1: ***Linear & Logistic Regression***

Due Date: 1397.07.25

A. Linear Regression with one variable

I. Consider the attached file dataset1.txt. The first column of the data file shows the input data (x) and the second column shows the output value (y) for each sample.

1) Fit a linear regression model on your data using :
   a) Closed-form solution calculated by LSE method
   b) Gradient descent method in online (stochastic) mode (1500 iterations)
   c) Gradient descent method in batch mode (1500 iterations)
2) What is the equation of cost function $J(\theta)$ for linear regression?
3) Plot the dataset and superimpose the fitted models using three above methods.
4) Use each estimated parameter $\hat{\theta}$ to predict the output for x=6.2, 12.8, 22.1, 30.
5) Compare the parameter $\hat{\theta}$ estimated by each method.
6) Plot the cost function $J(\theta)$ along the epochs (plot both online & batch methods on one figure using **hold on** command).
7) To understand the cost function better, plot $J(\theta)$ in terms of θ_0 and θ_1. (Note: generate a grid of and $\theta_0 \in [-10 \quad 10]$) and $\theta_1 \in [-1 \quad 4]$ and plot the cost function, $J(\theta_0, \theta_1)$, using **surf** command).
8) Which type of G.D. (online\batch) do you prefer here? Why?

B. Analysis the effect of outlier

In this part we are going to analysis the effect of outlier on liner regression method. Here, use the dataset1 in the previous part.

1) Add some outlier by randomly generating some samples in the border of the data range. For this, generate 5 samples with $x \in [6\ 8]$ and $y \in [20\ 25]$ and 5 samples with $x \in [20\ 24]$ and $y \in [0\ 10]$.
2) Fit a model on the new data using:
   a) Closed form solution of liner regression.
   b) Batch Gradient descent (1500 iterations).
3) Plot the data and superimpose the fitted models from Part A (i.e., A-1-a & A-1-c) and in this part onto your figure.
4) Use the estimated parameter $\hat{\theta}$ to predict the output for x=6.2, 12.8, 22.1, 30.
5) How do you explain the effect of outlier on linear regression method?

C. Linear Regression with multiple variables

The data file ‘dataset2.txt’ contains a 2-dimentional data. The first two columns of the data file show the feature of each sample and the last column illustrate its corresponding output value.

1) Fit a linear regression model on your data using :
   a) Closed form solution of liner regression.
   b) Batch Gradient descent (1500 iterations).
2) Plot the cost function $J(\theta)$ along the epochs for Gradient descent method.
3) Use each estimated parameter $\hat{\theta}$ to predict the output for x = [1357, 5], x=[2500, 4].
4) Compare the parameter $\hat{\theta}$ estimated by each method.

D. Analysis the effect of feature renormalization

1) Apply feature normalization on the dataset2 used in Part C (Note: for normalization you can scale the samples such a way that the minimum and maximum values of each feature change to 0 and 1, respectively).
2) Fit a model on the new data using
   a) Closed form solution of liner regression.
   b) Batch Gradient descent (1500 iterations)
3) Use the estimated parameter $\hat{\theta}$ to predict the output for x = [1357, 5], x=[2500, 4].
4) Explain and Discuss about the results of the methods before and after feature normalization.

E. Logistic Regression

Consider the attached file dataset3.txt. The first two columns of the data file show the feature of each sample and the last column illustrates its corresponding binary level.

1) What is the cost function $J(\theta)$ in logistic regression?
2) Estimate the parameter $\hat{\theta}$ using Newton method.
3) Plot the cost function $J(\theta)$ along the epochs of the Newton method.
4) Use the learned model to classify all training example.
5) Plot the data and show the class of the sample using different colors (red for class 0 & blue for class 1)
6) (optional) Plot the boundary of the classifier.
7) Determine the accuracy of the learned method over the training data.

F. (Optional) Weighted Linear Regression

In part B, you analyzed the effect of outlier on linear regression method. In this part, you should apply weighted linear regression on the data file used in part B which includes the outliers.

1) What is the closed form solution for estimating the parameters in WLR?
2) Propose a weighting function to decrease the effect of outlier and explain the reasons that you select that function.
3) Apply weighted linear regression using your suggested weighting function and:
   a) Closed form solution of weighted liner regression.
   b) Batch Gradient descent (1500 iterations).
4) Plot the data (including the outliers) and superimpose the fitted model in part B and the fitted models in this part onto the figures.
5) Compare the results in this part with part B.
6) When it is better to use WLR method and when it is not?

Implementation Note:

- Implement your codes as a functional form.
- In your coeds, show one sample in a column vector format.
- In all parts, set the initial value of the parameters to zero.
- In G.D., set the learning rate to 0.01.

Report:

- Prepare a report in PDF format including the figures, answer to the questions and discussions mentioned in the homework.
- Make a folder including your report and you codes (Note that your code is needed to be self-comment)
- Submit all things in a zipped **folder** named as “UrName_UrFamily.rar”

Good Luck