CS 430/530 - Assignment 1

Due date – Wednesday, September 13th, 2023, on or before 11:59 pm.

Assigned - Sunday, September 3rd, 2023

Please answer all the questions. No late submissions will be accepted. The solutions for all questions except the programming project should be neatly typed or written on a document and submitted through Canvas on or before the deadline. It can also be submitted in class in hard copy before the deadline.

Please write your name legibly and staple the sheets before submitting.

Problem 01 [5 +5+5+5= 20 points]

a) In linear regression the cost function is expressed as:

$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$

Please explain the reason why a squared error is used in the cost fuction instead of an absolute error.

The main reason is that using the squared error makes the cost twice differentiable. This allows us to use a variety of different methods to solve our regression. Additionally, being twice differentiable allows us to use a function f(x) in the place of x for many functions. This is not possible if we use absolute error since the absolute value of a function is not always differentiable.

b) Use the least square method to solve for the parameters θ_0 & θ_1 for linear regression and compute the values of the parameters for given data set. Compare the values you got using least square method with the values that you will get using gradient descent (see the programming assignment below). How close are they?

By using Excel to compute the Liner regression, I arrived at parameters

$\theta_0 = -3.89578,\ \theta_1 = 1.193034$

When running the Gradient Descent Model with an epsilon of 1E-10, I arrived at the parameters

$\theta_0 = -3.89523,\ \theta_1 = 1.192978$

When running the Gradient Descent Model with an epsilon of 0.01, I arrived at the parameters

$\theta_0 = 0.05782,\ \theta_1 = 0.79135$

The values are very close when using a very small epsilon, but when using an epsilon like 0.01, the values are far enough away that a graph would look substantially different.

Problem 02 (Programming project) [80 points]

This problem needs to be submitted separately through email. See submission section below.

Problem statement: Linear regression with one variable

In this exercise, you will implement linear regression with one variable to predict profit for a retail store. Suppose you are the CEO of a retail store chain and are considering different cities for opening a new outlet. The chain already has stores in various cities, and you have data for profits for each of those stores and the population of the corresponding city.

The file `data.txt` contains the dataset for our linear regression problem. The first column is the population of a city, and the second column is the profit of the retail store in that city. A negative value for profit indicates a loss.

Plotting the Data

Before starting on any task, it is often useful to understand the data by visualizing it. For this dataset, you can use a scatter plot to visualize the data, since it has only two properties to plot (profit and population). Now, when you plot the data, the result should look like Figure 1, with the same red "x" markers and axis labels.

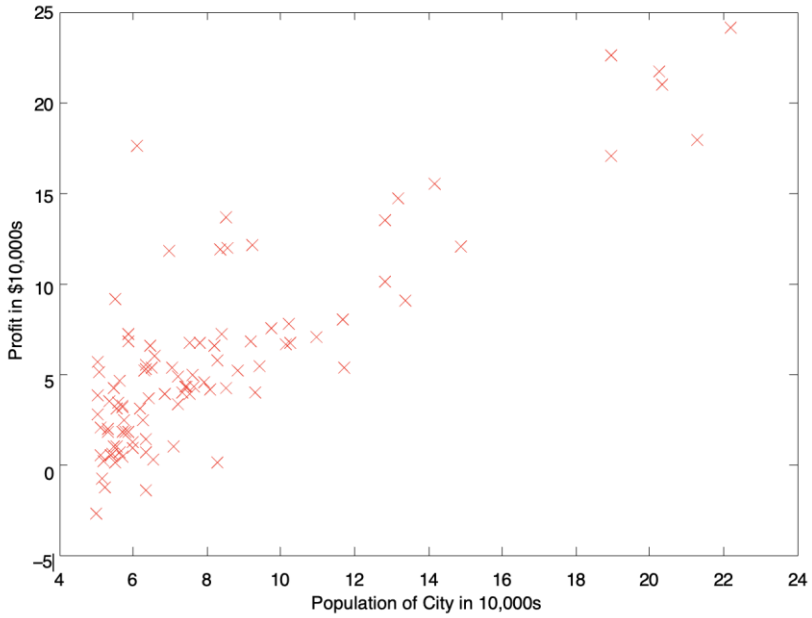


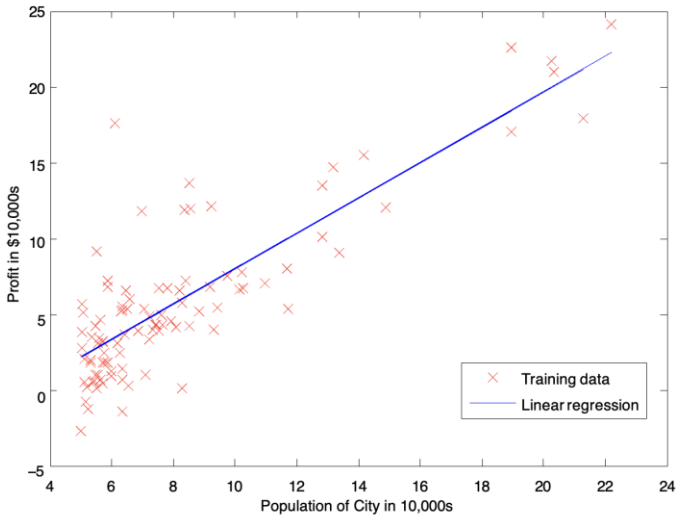
Figure 1: Scatter plot of training data

Gradient Descent

As discussed in class, model the problem as a linear regression problem with a single variable and estimate the parameters using gradient descent. Hence, you will fit the linear regression parameters θ_0 & $\theta_1$to our dataset using gradient descent. Use the cost function and the update equations (of the parameters) as discussed in class and are available in slides. you will use batch gradient descent to solve the problems.

Initialize the initial parameters to 0 and the learning rate `alpha` to 0.01. A good way to verify that gradient descent is working correctly is to look at the value of J(θ_0, θ_1) and check that it is decreasing with each step.Your program should stop after reaching an acceptable accuracy, say $\epsilon = 0.01$ or less.

After you are finished, use your final parameters to plot the linear fit. The result should look something like Figure 2:



Finally, use the final values for θ_0 & θ_1 to make predictions on profits in areas of 35,000 and 70,000 people.

Programming languages and formatting

You should use preferably python or MATLAB and should implement the algorithms yourself. Please do NOT use any library function which provides an implementation of linear regression or gradient descent. You will get ZERO otherwise.

Submission Guidelines

Your submission must have the following:

A README file that describes how the code can be compiled and run. Also list any external dependencies that need to be satisfied for compiling and running the code. If your code fails to compile YOU GET A ZERO.

Please e-mail your submission to hs0111@uah.edu by 11:59 PM on Wednesday September 13th, 2023. DO NOT SUBMIT THROUGH DROPBOX or CANVAS. They will not be accepted and will result in late penalty. Put all your materials for the PROGRAMMING project ONLY in a folder with your name and then create a zip file out of it. Email the zip file to the TA.

Point Breakdown

For loading the data: 10 points

For plotting the data: 10 points

For correct implementation of gradient descent: 40 points

For plotting the regression line: 10 points

For predictions: 5 +5 = 10 points