

Machine Learning Assignment Report

By: Nadine TareK El bialy fahmy 40-0379

Table of Contents

[**Assignment 1** 2](#_Toc70221052)

[**Linear regression with one variable** 2](#_Toc70221053)

[**Linear regression with multiple variables** 3](#_Toc70221054)

[**Assignment 1 – Take 2** 4](#_Toc70221055)

[**Model Selection** 4](#_Toc70221056)

[**Assignment 2** 5](#_Toc70221057)

[**Logistic Regression** 5](#_Toc70221058)

[**Regularized Logistic Regression** 5](#_Toc70221059)

[**Assignment 2 – Take 2** 6](#_Toc70221060)

[**Regularization** 6](#_Toc70221061)

# **Assignment 1**

## **Linear regression with one variable**

Linear regression attempts to model the relationship between variables by fitting a straight line among such data.

The first part of the assignment covered linear regression for one variable (population). The dataset used contained two columns one for the population of the city and the other for the profit yielded in this city. which was the population of the city where a food truck resides. Data is plotted below in figure 1.

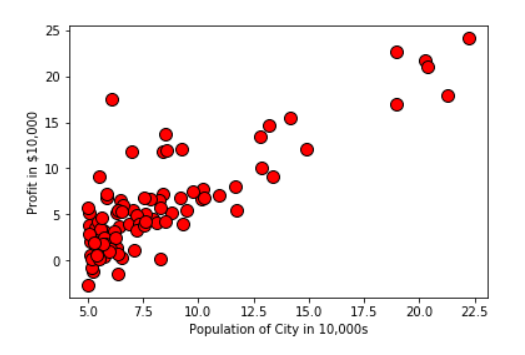


Figure 1

The cost is then calculated using the following equation shown below in *figure (2)*, where ***h*** is the hypothesis and ***J*** is the cost function calculated. Then, the gradient descent is calculated which is an optimization algorithm aiming to minimize the cost ***J*** with respect to, or, in other terms, by varying the values of theta. After covering such points, the following plot shown in *figure (3)* was yielded and prediction could be made based on such calculations.

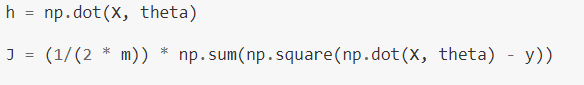


Figure 2

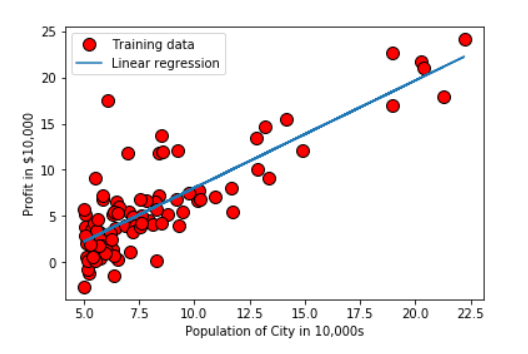


Figure 3

## **Linear regression with multiple variables**

In this part of the assignment, a different dataset was used which contained multiple variables giving details about different houses and the prices of such houses. The first step was normalization so as to make all the values within the same range which is done by setting the mean to 0 and the standard deviation to 1 using the equation in *figure (4)* below. The following steps were the same as linear regression for one variable. Learning rate is then plotted by altering the value of alpha to find the value of alpha which yields the best solution which is the learning rate which converges most quickly.

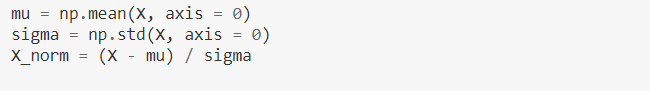


Figure 4

# **Assignment 1 – Take 2**

## **Model Selection**

For the second take on assignment 1 referred to in the GitHub repository as Take2\_Assig1, a dataset of size 17999x21 was used, which consisted of features regarding a set of houses and their relative price. In this case, linear regression and model selection were implemented.

Firstly, the dataset was read using the pandas library and the columns of ***id*** and ***date*** were dropped as they were found to be redundant and serve no purpose. The data was then split into a training set, a validation set and a testing set with a percentage of 60%, 20% and 20% respectively. Similar to assignment 1, each set was then normalized.

The first phase was the ***training phase***. For the training phase, firstly, x and y where defined, where ***x*** is the set of all features columns, which were 18 features after dropping the date and id columns of the training set and ***y*** is the corresponding ***price*** column of the training set. A loop was then generated where all degrees from 1 to 18 were generated, given that we have 18 features, and the corresponding theta coverage and J (cost) were calculated per degree. The thetas (theta coverage) of the degree with minimum cost J were saved in an array for future use.

The second phase was the ***validation phase***. For the validation phase, the thetas of least cost J generated in the training phase were used by the computeCostMulti function to calculate the cost J for each degree through the use of a loop. The results showed that the degree that yielded the least cost J was the highest degree same as in training.

The third and last phase was the ***testing phase***. For the testing phase the theta of the minimum cost was used in the computeCostMulti function to calculate the error.

In conclusion, the training phase gave us the theta values, the validation phase gave us the degree of the hypothesis and the testing phase is used to test for accuracy and errors.

Lastly, a plot was implemented to show the convergence of the cost J with respect to the different degree values for the both the cross validation set and the training set as seen in *figure (5)* below that the cost significantly decreases as the degree value increases.

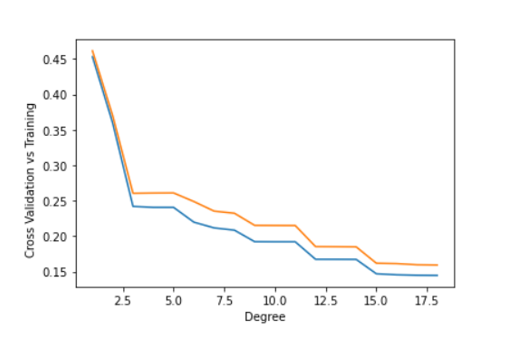


Figure 5

# **Assignment 2**

## **Logistic Regression**

Logistic regression is a classification technique that deals with discrete values. The logistic regression hypothesis is defined by the equation below in *figure (6),* where g is the sigmoid function. Therefore, the sigmoid function is implemented initially. Then, the cost function and gradient were calculated and tested to show that the calculated values and expected values matched.



Figure 6

The learning rates part is then used to optimize the cost function by finding the minimum theta. The decision boundary is then plotted as shown in *figure (7)* below.

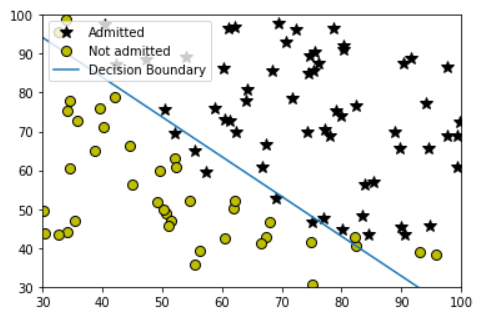


Figure 7

## **Regularized Logistic Regression**

In regularized logistic regression, the same process as the logistic regression shown above is implemented but the idea of feature mapping is added which helps make the existing features and data more expressive.

# **Assignment 2 – Take 2**

## **Regularization**

In this assignment, the same datasets were used as the ones in the previous assignment. For the first part of the assignment, firstly, the dataset is split using the *train\_test\_split* function into a training set and a testing set where the training set comprises of 70% of the dataset and the testing set of 30%.

For the regularized logistic regression part, the dataset is split into training, validating and testing sets with percentages of 60%, 20% and 20% respectively. For the training phase, a set of lambda values are used which were [0,0.01,0.02,0.04,0.08]. This yielded the theta values, then the validating phase was implemented. Finally, the testing phase which tested for accuracy was implemented and showed an accuracy of 60% which is low due to the fact that the dataset only comprised of 100 rows which when split into training, validating and testing sets, they were too small.