**Problem Description: What is the task that you intend to solve in this assignment?**

Estimating house prices is essential for homeowners and investors alike with both needing to understand the value of their asset, and to understand real estate assets as part of an overall portfolios. In this task, we are intended to use machine learning techniques to predict house price using various kind of house features that effect on its price. These features values are used to train the model that can try to predict the value of house.

**Method: Brief description of the method that you follow ( if you are doing linear regression, describe the process, preferably with mathematical details.)**

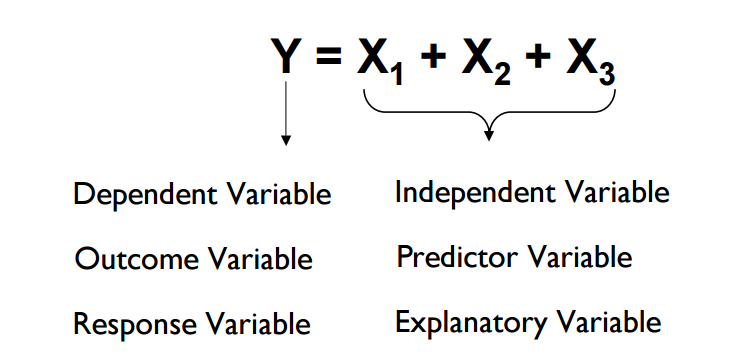
We followed these steps approaches to complete this task:

1. Data Analysis:

I have observed the dataset and check its different features by this I am able to understand different features and their correlation with price.

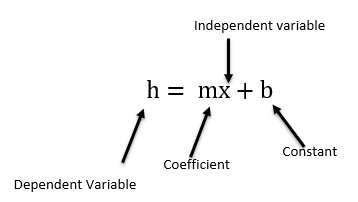
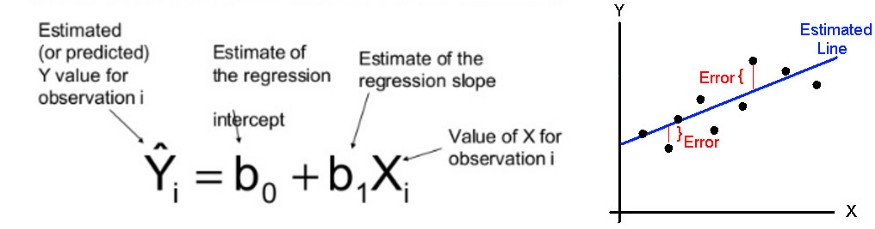
1. Data Preprocessing:
   1. Removing missing values from dataset.
   2. Dropping unique features like id, date from dataset.
   3. Performing the standard mean normalization on the dataset.
   4. Splitting the dataset in training and testing parts.
2. Linear Regression:

Linear regression is simply a linear method to model the relationship between your independent variables(X) and your dependent variables(y).



This method uses a single independent variable to predict a dependent variable by fitting

a best linear relationship. Simple linear regression is fitting a line equation on independent variables and try to predict the dependent variable.

Where

m= slope, x= data variable, b= constant, h= predicted Answer.

Given our hypothesis of linear regression , MSE (L2) is defined as

**Cost**

Where

* + : is number of examples in data set.
  + : is actual output/label in data set.
  + : is predicted output/label from our algorithm.

So our goal will be to choose weights in such a way that value of our cost function is minimum i.e

Gradient descent algorithm optimizes parameters to minimize cost function. It works in an iterative manner. The basic idea behind is to move in the direction of steepest descent by taking negative of gradients of cost function.

We have

**Cost**

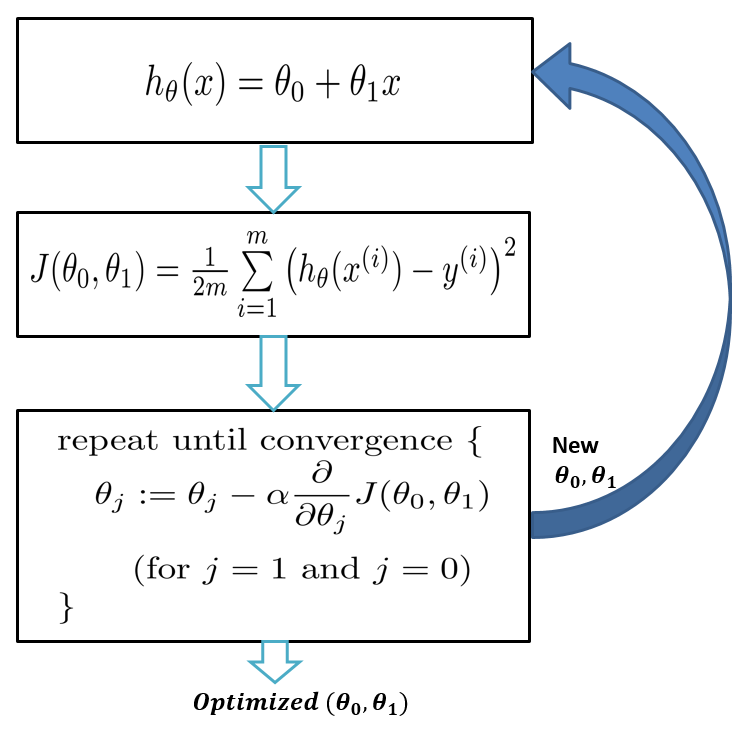
To calculate gradient descent, we need to calculate **partial derivate** with respect to parameters/weights.

Where is called learning rate, will be discussed in the coming slides.

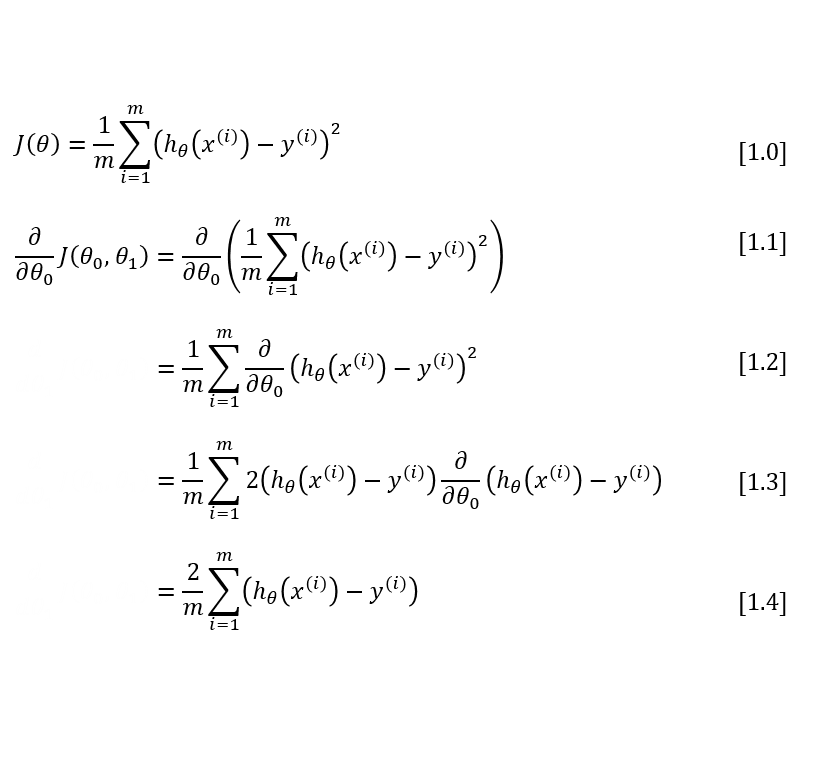
Using these two equations we can implement gradient descent algorithm to find our optimized parameters.

Reference:

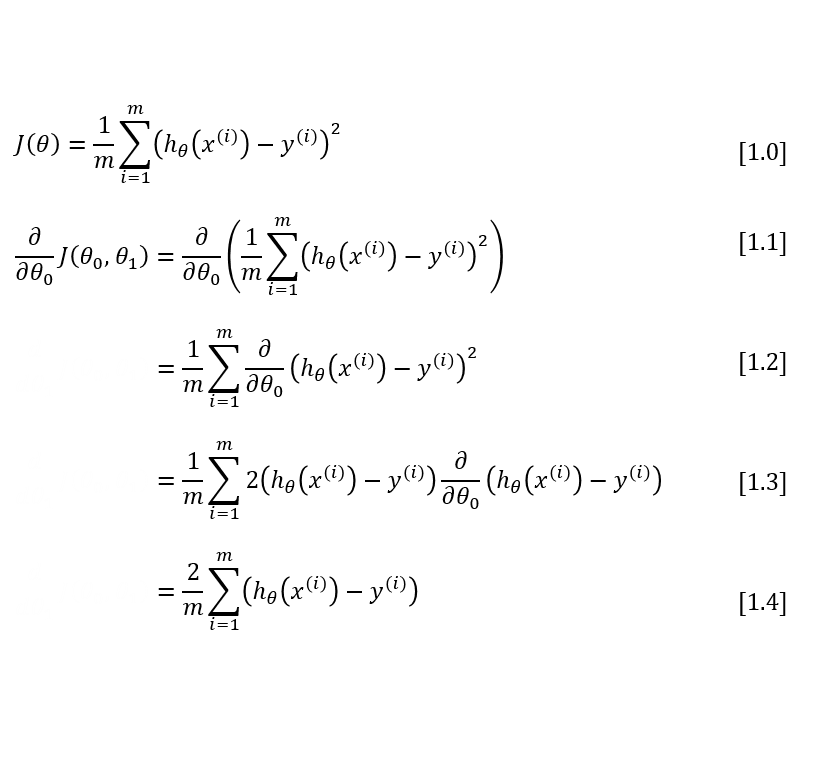
1. <https://medium.com/@lachlanmiller_52885/machine-learning-week-1-cost-function-gradient-descent-and-univariate-linear-regression-8f5fe69815fd>
2. <https://www.geeksforgeeks.org/gradient-descent-in-linear-regression/>
3. <https://www.coursera.org/learn/machine-learning>



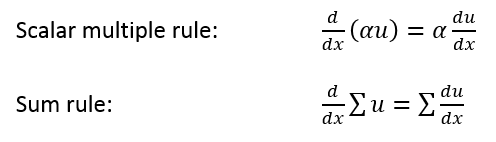
* Derivative of cost function shown below, steps are explained in next slides:

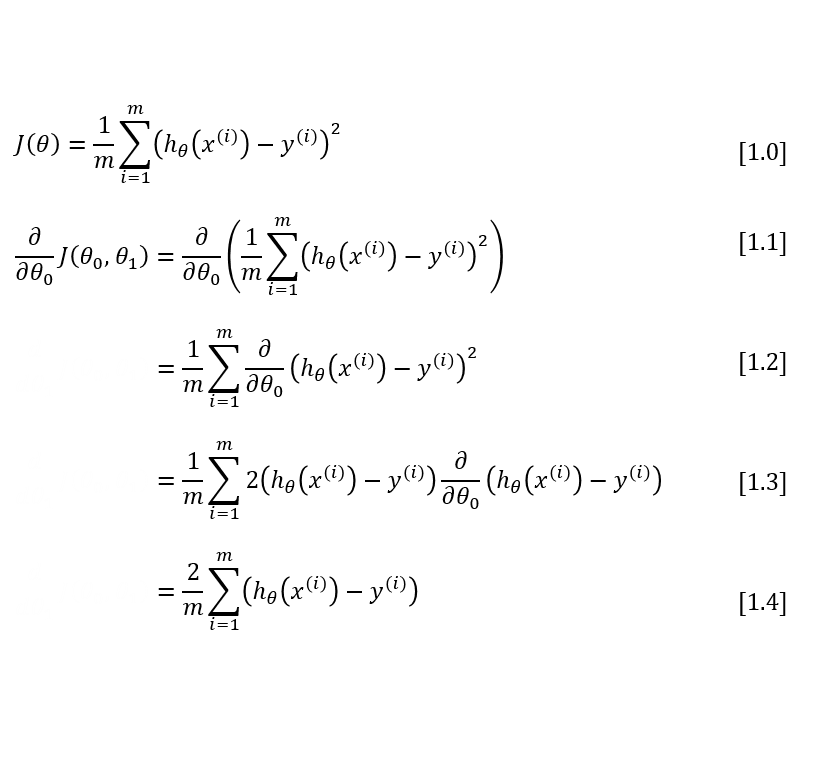


* Cost function equation show in [1.0], Derivative steps cost function shown below

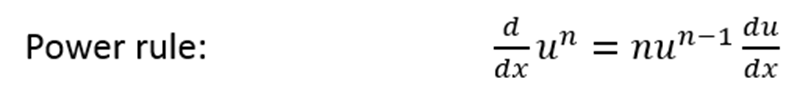


* we need to apply two basic derivative rules after that we get 1.1 and 1.2 equations:

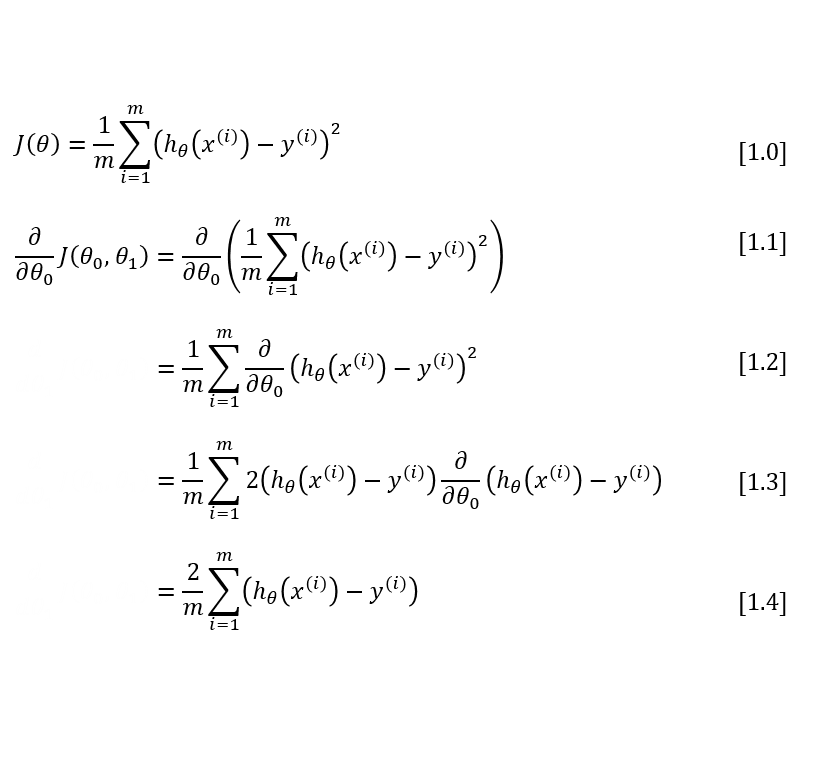
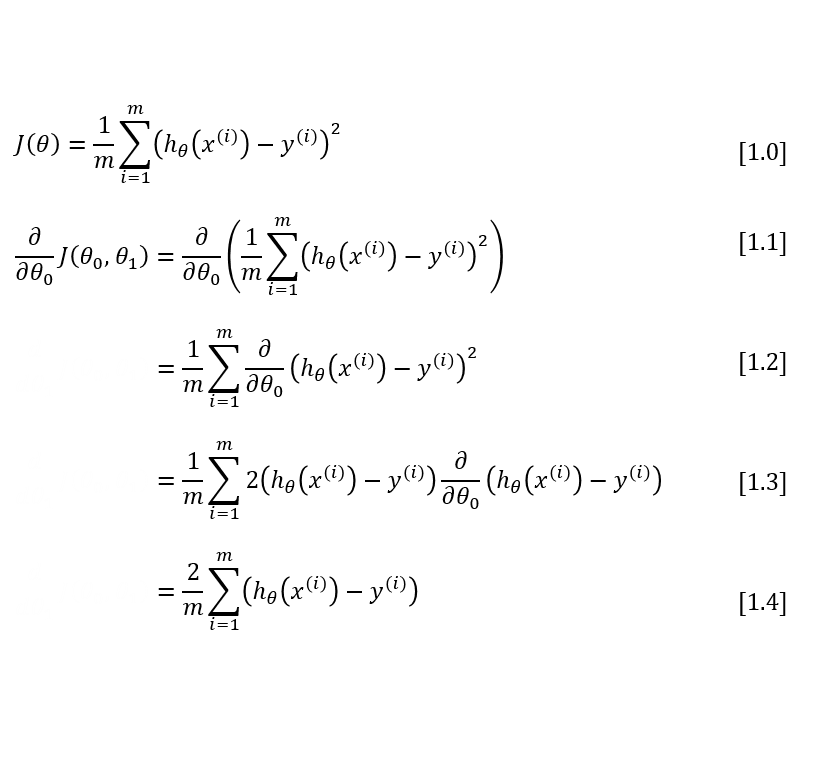
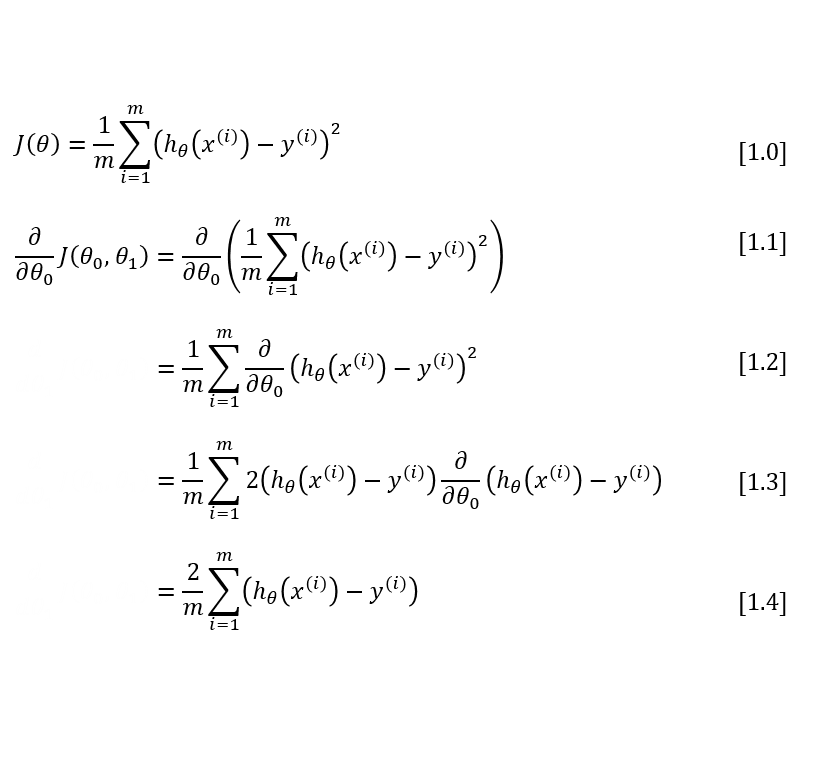




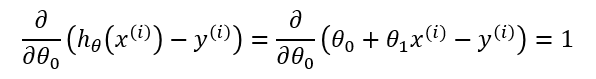
* we apply both the power rule and the chain rule on [1.2] and got [1.4]:



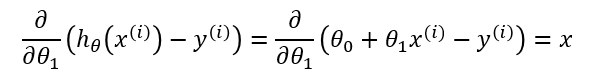
Chain Rule



* we must evaluate the partial derivative as follows. Recall again that when taking this partial derivative all letters except are treated as constants



* Equation [1.4] gives us the partial derivative of the MSE cost function with respect to one of the variables . Now we must also take the partial derivative of the MSE function with respect to . The only difference is in the final step, where we take the partial derivative of the error:



Reference:

1. <https://mccormickml.com/2014/03/04/gradient-descent-derivation/>

**Dataset description: size, training /testing distribution, features etc.**

The dataset consists of 21 variables and 21613 examples,some variables are unique respective to specific house like id, date. Some variables are binary features that defined that specific features exists or not like waterfront etc. Variable are explained below:

* id: Unique id of each house.
* date: date of recording price
* price: price of house
* bedrooms: No. of bedrooms
* bathrooms: No. of bathrooms
* sqft\_living: living building area
* sqft\_lot: Total area
* floors: No. of floors
* waterfront: waterfront exist or not
* view: House view
* condition: House condition from 1 to 5.
* grade: House grade
* sqft\_above: area of above building
* sqft\_basement: area of basement
* yr\_built: House building year
* yr\_renovated: House renovation year
* zipcode: zip code of house area
* lat: latitude coordinate of house
* long: longitude coordinate of house
* sqft\_living15: living15 area
* sqft\_lot15: lot15 area

The dataset is split into 90%training and 10% testing parts.

**How is your prediction task defined on this dataset specifically? linear/logistic regression, and what is the meaning of the output variable?**

As this task is consisting of continuous numbers of values prediction that’s why we use regression method and used linear regression. Predictions are make through training weights on this dataset of houses features, that’s why they are trained specifically for this task of house price prediction. Output variable is the price of house.

**Did you process the features in any way?**

Yes, I performed these methods on dataset:

* 1. Removing missing values from dataset.
  2. Dropping unique features like id, date from dataset.
  3. Performing the standard mean normalization on the dataset.
  4. Splitting the dataset in training and testing parts.

**Did you try to tune the hyper-parameters of the learning algorithm, and in that case how?**

Yes, I have try to tuned hyper parameters like learning rate and number of iterations for training the model. As, when learning rate alpha is low then model converge slowly as small change in gradient steps so it required more number of iterations for to converge. When learning rate alpha is high value then it converges fast in low number of iterations as gradient take bigger steps. So we choose optimal value between it.

**How do you evaluate the quality of your system?**

By testing the trained model on testing dataset, which were remain hidden from model and then measuring the Root mean square error between prediction and actual values.

**If you want to use the regression result for classification, what is the optimal theta ( i.e. what is the best decision boundary based on the training data distribution? How did you arrive at your decision?**

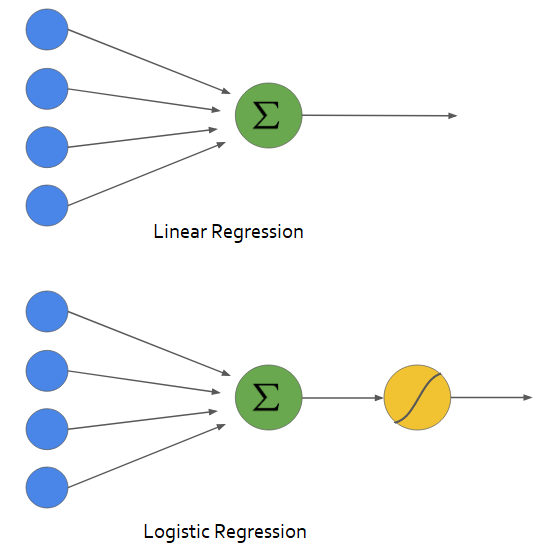
If I want to use regression results for classification then my theta has to be some nonlinear function, like sigma. It has to draw answer to some classes. The best decision boundary is between waterfront features who has only two values, defined it exist or not.

**Is it possible to say something about which features the model considers important?**

yes, those features that are more correlated with outputs values are more important than those who don’t, like in our case area based features are highly correlated with price.

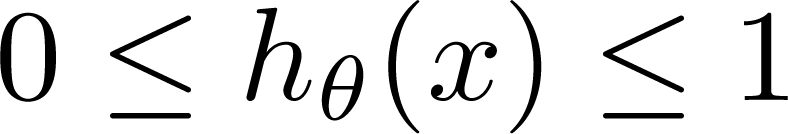
**Experimental Details A similar set of details should be added for Logistic Regression.**

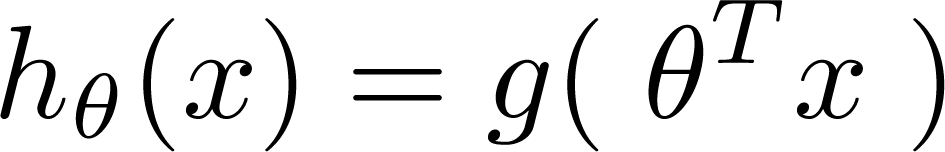
Logistic Regression is generally used for classification purposes. Unlike Linear Regression, the dependent variable can take a limited number of values only When the number of possible outcomes is only two it is called Binary Logistic Regression.

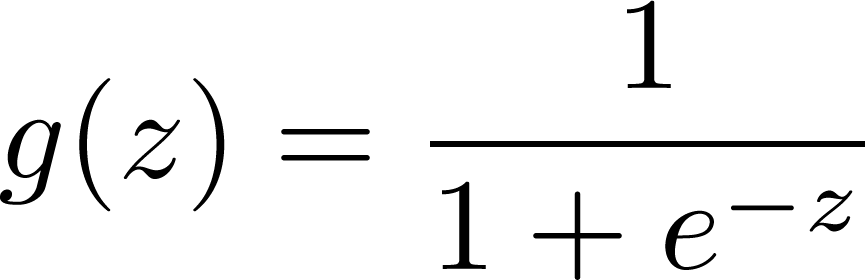


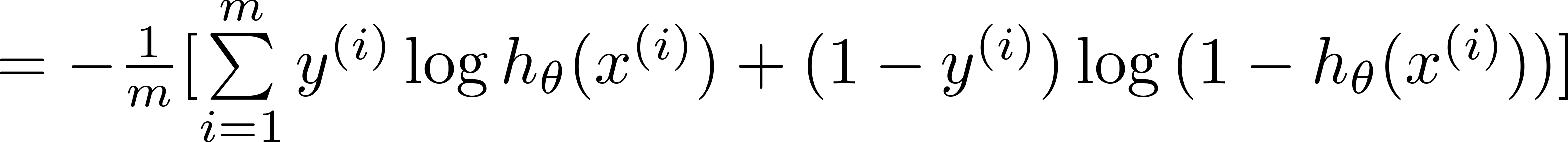
As we are doing logistic regression, so we need to fix number of classes for this house data. we have follow an approach to define that a house is costlier or not, for example if an house value is larger than respective value 450000 it is costlier, so we give it 1 tag, or other to 0.

Core idea: force hypotheses between 0 and 1

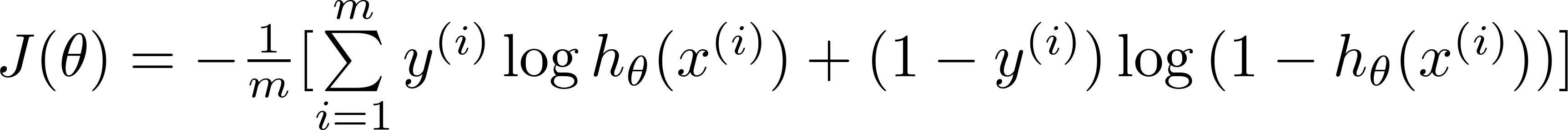
Logistic Regression: 

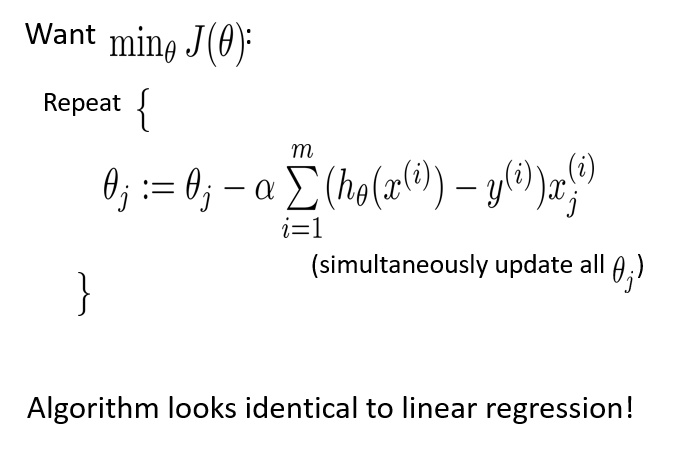
Hypothesis: 

Sigmoid Function: 

Cost Function: 

Gradient Descent:





Reference:

1. <https://medium.com/deep-math-machine-learning-ai/chapter-2-0-logistic-regression-with-math-e9cbb3ec6077>

**Comparative Study between two methods:**

1. The Linear regression models data using continuous numeric value. As against, logistic regression models the data in the binary values.
2. Linear regression requires to establish the linear relationship among dependent and independent variable whereas it is not necessary for logistic regression.
3. In the linear regression, the independent variable can be correlated with each other. On the contrary, in the logistic regression, the variable must not be correlated with each other.

**Conclusion:**

We have tried to predict the house prices in linear regression and try to classify costlier house and cheap houses in the logistic regression. We have gain Linear Regression RMSE: 0.5204, and logistic regression accuracy is 70.9%. By using these two methods not only we can predict the house price but also can classify different types of house on the base of features, which can help real estate investors to choose the compatible to his requrements.