Assignment 2
Due on September 12

A group of 20 students studied 0 to 6 hours for the exam. Some passed and others failed. Results are given below (Data is taken from Wikipedia).

Student	Hours studied - x	Result (0 – fail, 1 – pass) - y
1	0.5	0
2	0.75	0
3	1.00	0
4	1.25	0
5	1.50	0
6	1.75	0
7	1.75	1
8	2.00	0
9	2.25	1
10	2.50	0
11	2.75	1
12	3.00	0
13	3.25	1
14	3.50	0
15	4.00	1
16	4.25	1
17	4.50	1
18	4.75	1
19	5.00	1
20	5.50	1

1. Determine the optimal linear hypothesis using linear regression to predict if a student passes or not based on the number hours studied.

Given a single feature, considering the linear hypothesis considered will be

Hypothesis

$H = \theta_0 + \theta_1 * X_1$

Cost function

$J(\boldsymbol{\theta}) = (1/2m) \sum_{i=1 \text{ to } m} (h_\theta (X_i) - y_i)^2$.

Gradient Descent

$\theta_j = \theta_j - (\alpha/m) \sum_{i=1 \text{ to } m} (h_\theta (X_i) - y_i) x_{ij}$, for j = 0. 1...n.

Fig1: plot of the training data set

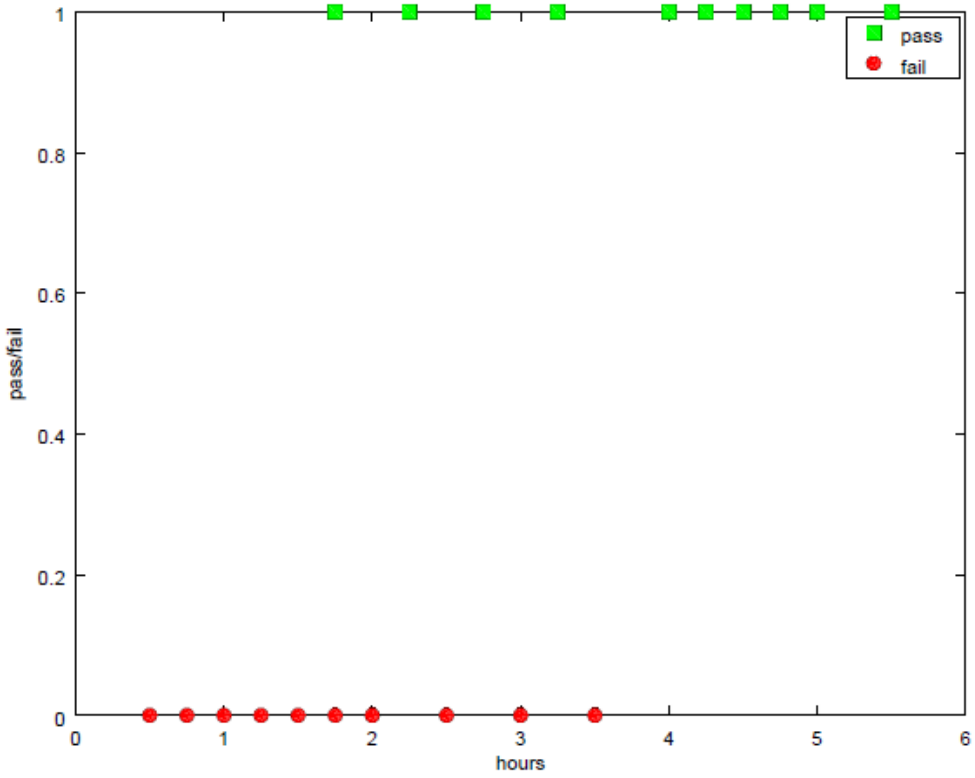

fig1: plot of training data set.

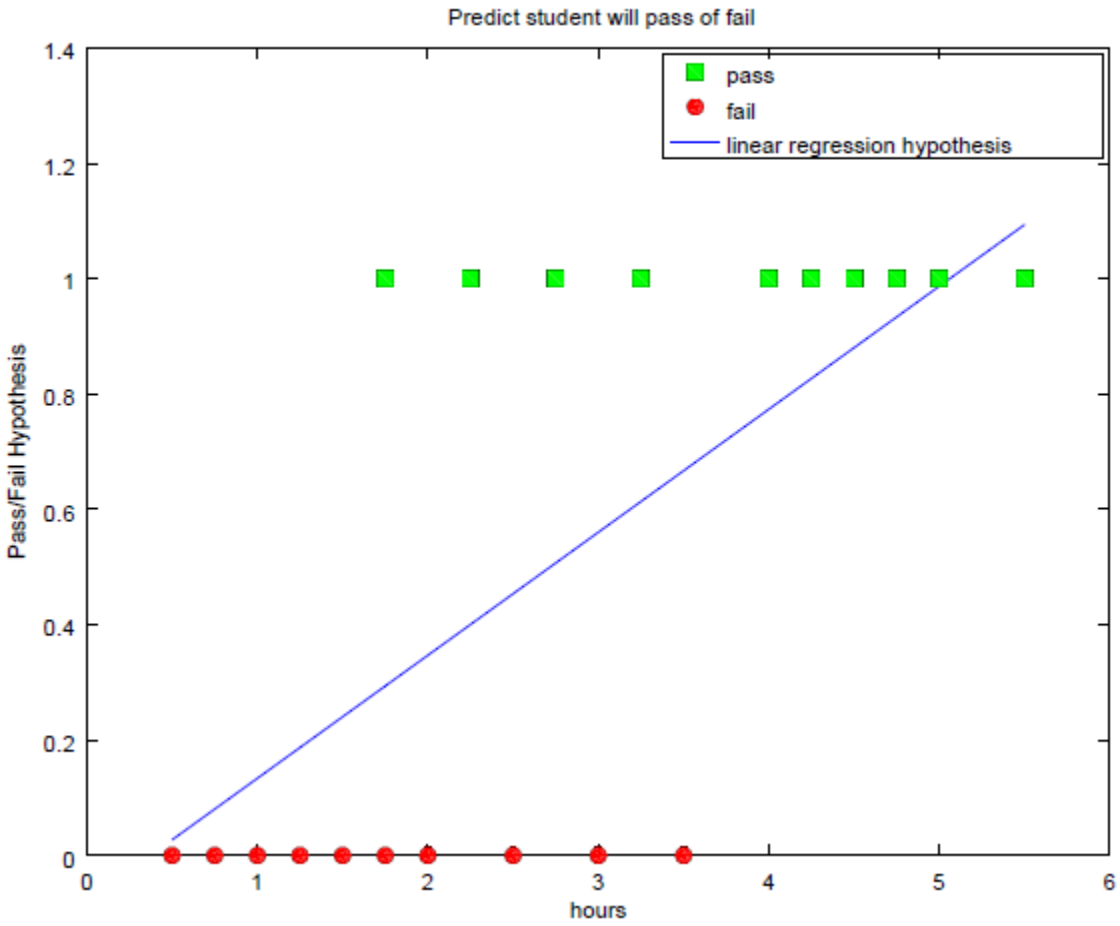

fig2: Plot of linear Hypothesis and the training Data set. .

Considering the Prediction concept for the classification problem. Having threshold as 0.5.

Predicts pass, if $H(x) > 0.5$
Predicts fail, if $H(x) < 0.5$

Dividing the Training data set into two classes, class 1 Y = 0 (fail) , class 2 Y = 1 (pass).
The decision boundary is found. For the study hours greater than 2.75, the student will pass.
For study hours less than 2.75, student will fail.

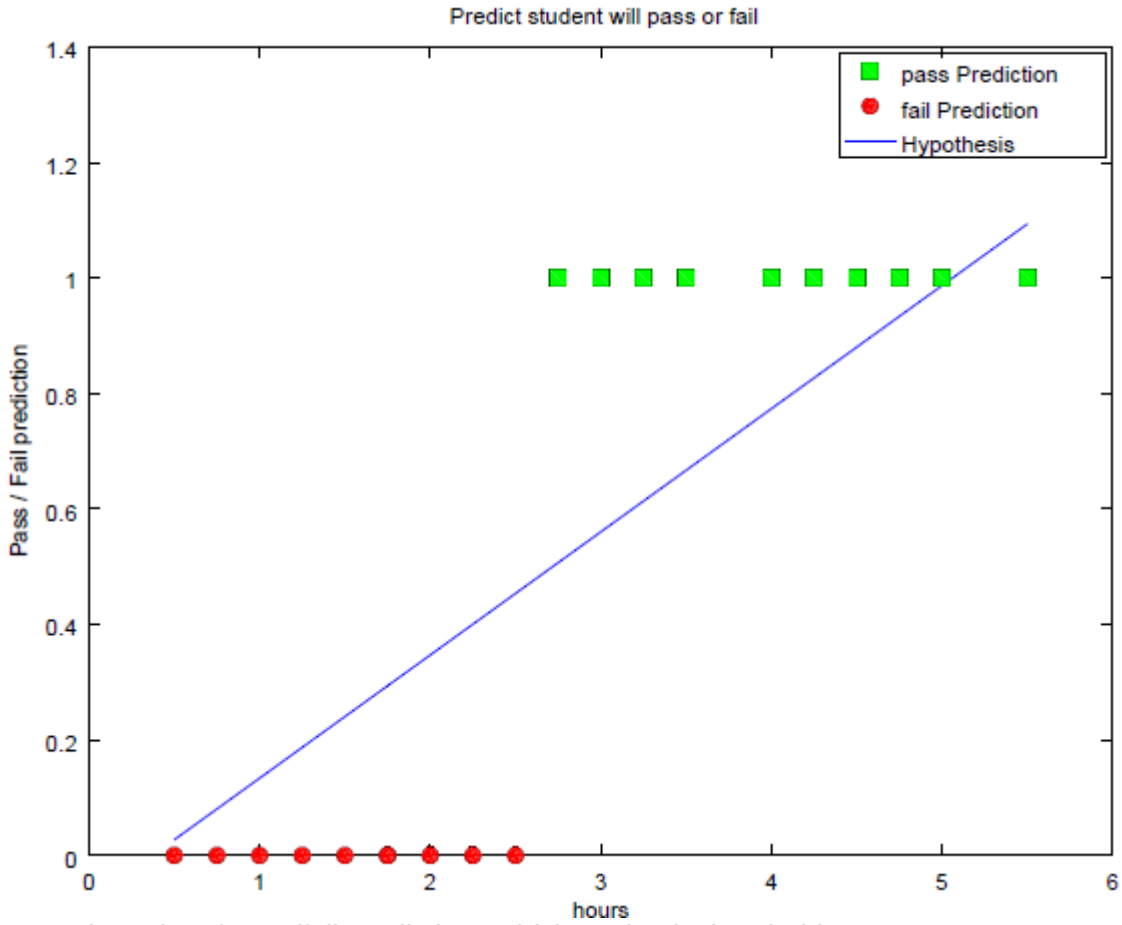

Fig3: plot of pass/fail prediction, H(x) hypothesis threshold as 0.5.

2. Determine the optimal logistic hypothesis using logistic regression to predict if a student passes or not based on the number hours studied.

Hypothesis

$z = X*\theta$

$h_\theta(x) = g(z)$ where g(z) is a sigmoid function

$g(z) = 1/(1+ e^{-z})$

Cost function

$\text{cost}(h_\theta(x), y) = - y \log(h_\theta(x)) - (1 - y) \log(1 - h_\theta(x))$.

Minimization of J

$J(\theta) = (-1/m) [\sum_{i=1 \text{ to } m} (y_i \log(h_\theta(x_i)) + (1 - y_i) \log(1 - h_\theta(x_i))$

Update rule: $\theta_j = \theta_j - \alpha \text{ grad}_j [J(\theta)]$

$= \theta_j - \alpha [\sum_{i=1 \text{ to } m} ((h_\theta(x_i) - y_i) x_{ij}]$

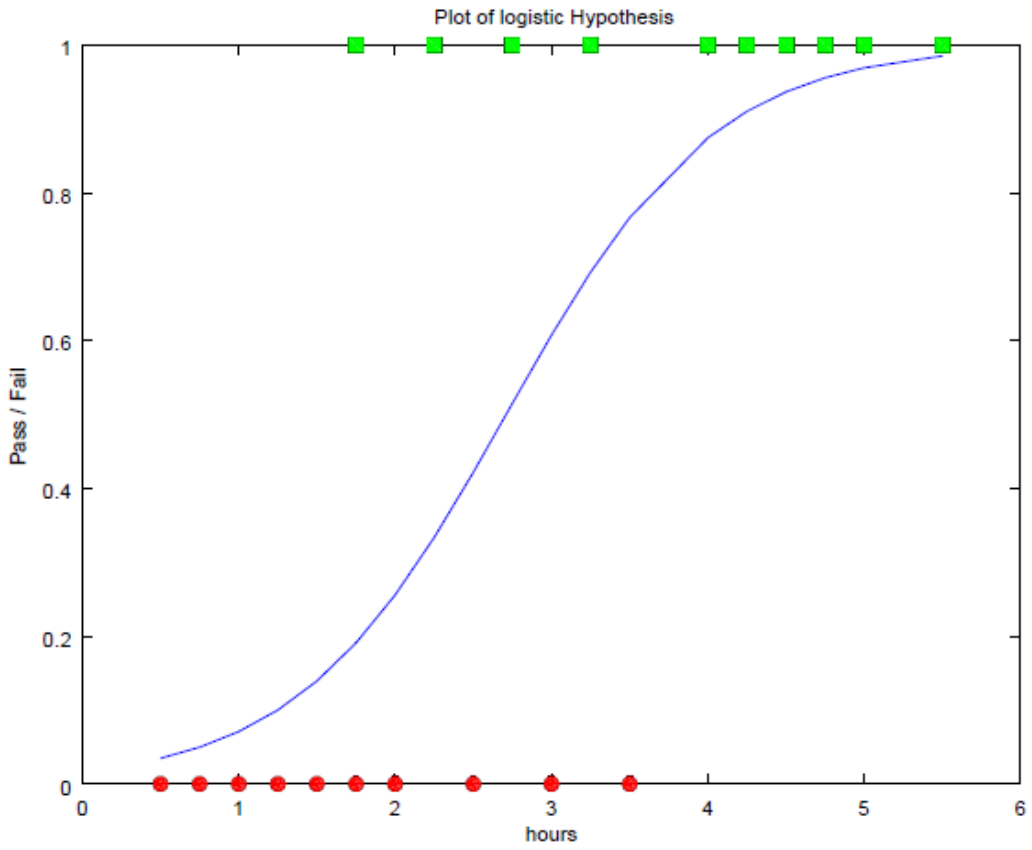

Fig4: plot of Logistic Hypothesis for given training dataset.

Considering the Prediction concept for the classification problem. Having threshold as 0.5.

Predicts pass, if $H(x) > 0.5$
Predicts fail, if $H(x) < 0.5$

Dividing the Training data set into two classes, class 1 Y = 0 (fail) , class 2 Y = 1 (pass).
The decision boundary is found. For the study hours greater than 2.75, the student will pass.
For study hours less than 2.75, student will fail.

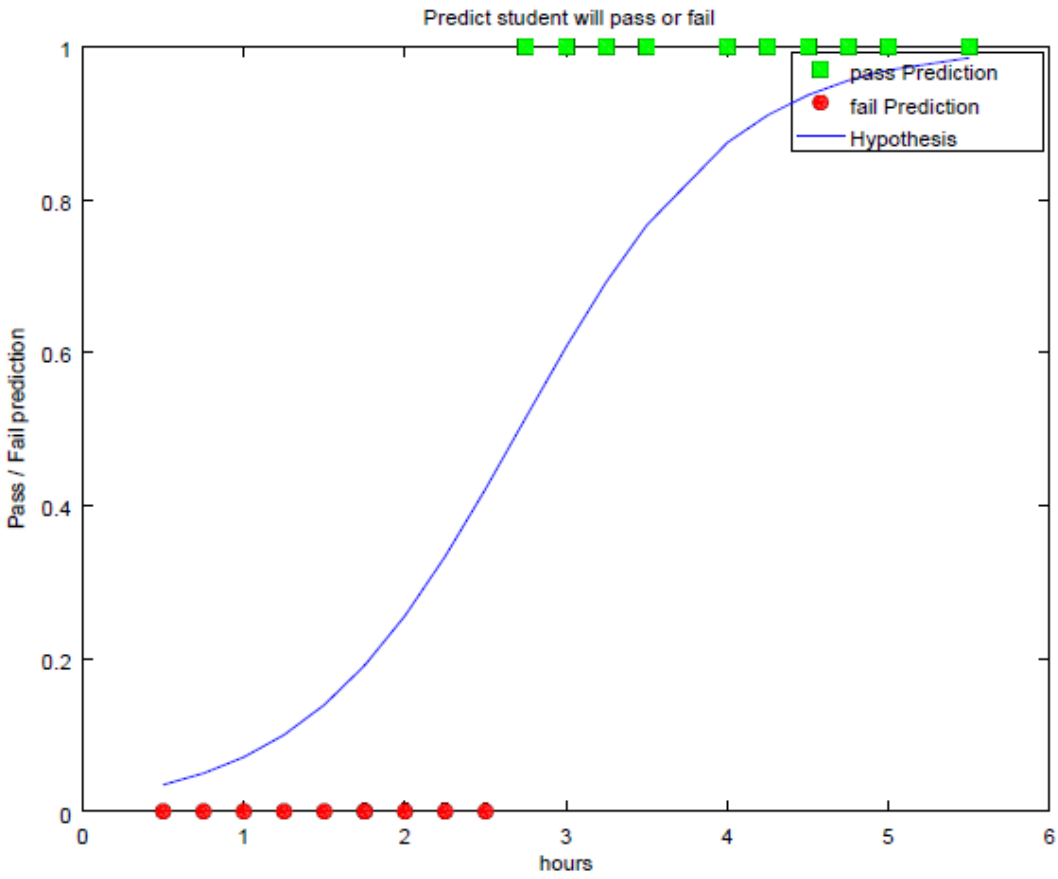

Fig5: plot of Logistic Hypothesis for given training dataset.

3. Plot both hypothesis function $0 < x < 6$. Compare and explain the two results obtained.

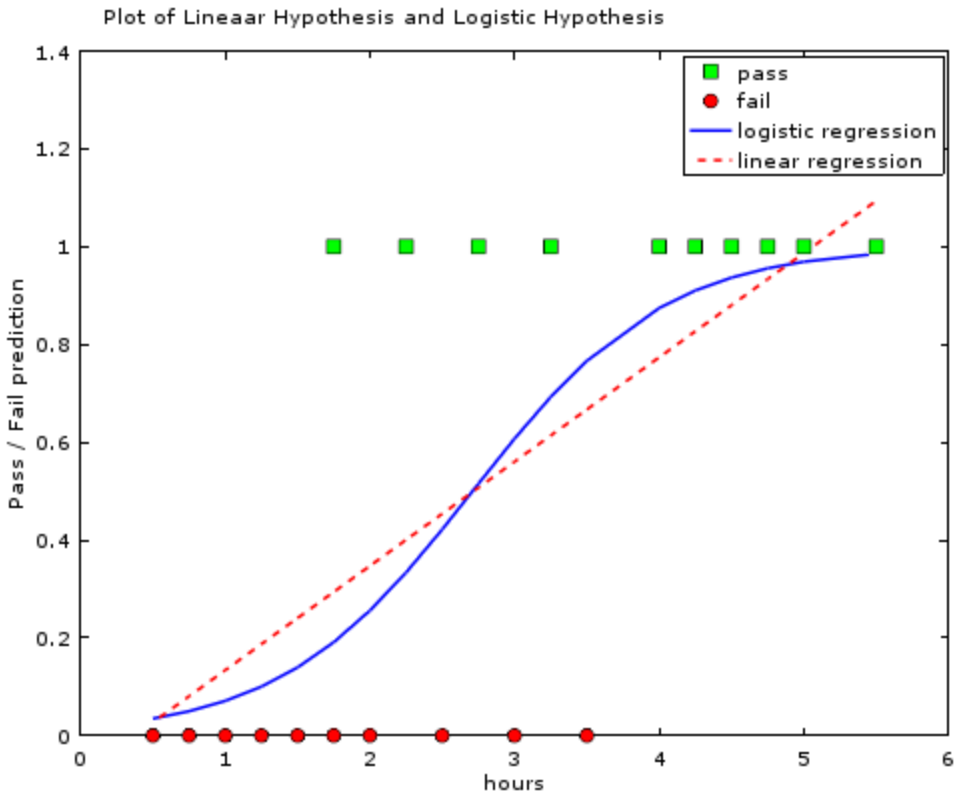

Fig5: plot of Logistic and linear Hypothesis for given training dataset.

The Linear Hypothesis is the Linear line passing the fail and pass training set.The Logistic Hypothesis is the exponential sigmoid curve passing the fail and pass training set The value of H(x) = 0.5 for both logistic and linear regression for Hour = 2.75. The prediction says student will pass if he/she studies more than 2.75 hours.

Suppose let's add [12,1] data to the Training data set. The result plot is shown below.

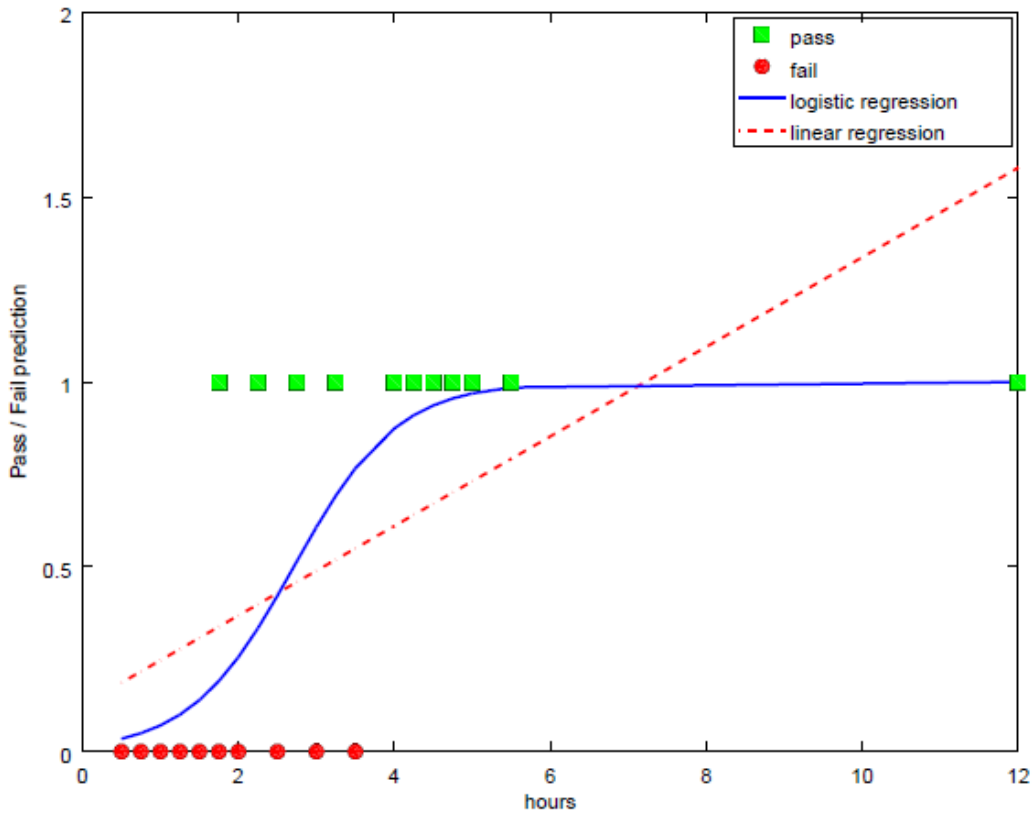

Fig6: plot of Logistic Hypothesis and linear for added training dataset.

The linear hypothesis line with respect to number of hours has changed. From the graph we find that the hypothesis value is 0.5 at hour value 3. Actual hypothesis value should be 0.5 for hour value 2.75. Add one more sample to the training set, thresholding $h_\theta(x)$ at 0.5 does not work. The problem is that the value of the hypothesis function used in linear regression can range from $-\infty$ to $+\infty$. There is a need to use a hypothesis function whose value is restricted to remain between two finite limits (0 and 1 or -1 and +1). Hence Linear regression is not optimal for classification problem.

The Logistic hypothesis curve doesn't change, even though new training data set is added. The prediction is more optimal as the hypothesis function value is restricted to remain between two finite limits (0 and 1 or -1 and +1).

4) Develop a logistic regression-like algorithm for the following cost function.
Y = 1 - Cost function goes from 100 to 0 linearly as hypothesis function goes from 0 to 1
Y = 0 - Cost function goes from 0 to 100 linearly as hypothesis function goes from 0 to 1
Compare results with those of the standard logistic algorithm.

Hypothesis
$z = X*\theta$
$h_\theta(x) = g(z)$ where g(z) is a sigmoid function
$g(z) = 1/(1+ e^{-z})$

If y = 1 then cost (h_θ (x) , y) = **100(1-h_θ (x))**. Otherwise, cost (h_θ (x) , y) = **100 * h_θ (x)**.

Cost function

cost (h_θ (x) , y) = 100* y *(1-h_θ (x)) + 100*(1 - y) * h_θ (x).

cost (h_θ (x) , y) = 100 [y *(1-h_θ (x)) + (1 - y) * h_θ (x)].

Minimization of J

$J(\theta) = (-1/m) [\sum_{i=1 to m} (y_i \log (h_\theta (x_i)) + (1 - y_i) \log (1 - h_\theta (x_i))$

Update rule: $\theta_j = \theta_j - \alpha \text{ grad}_j [J(\theta)]$

$= \theta_j - \alpha [\sum_{i=1 to m} ((h_\theta (x_i) - y_i) x_{ij}]$

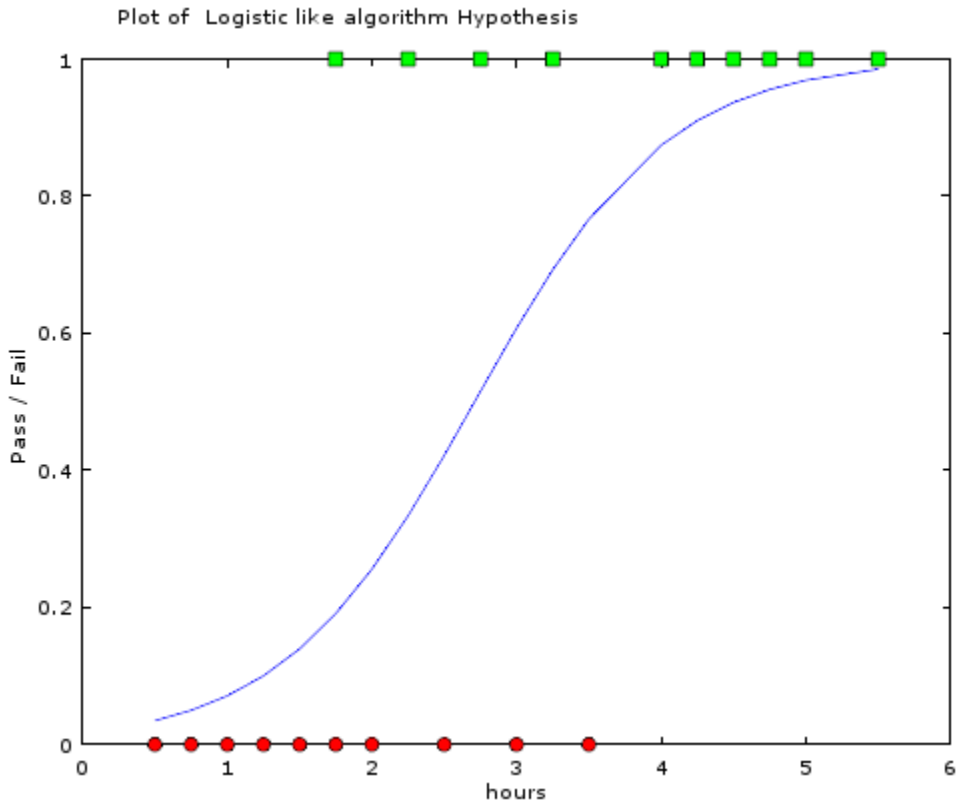

Fig6: plot of Logistic like Hypothesis for given training dataset.

The Hypothesis function is same as the logistic hypothesis. In logistic regression cost (h(x) is in range of ∞ to 0 for y= 1 and 0 to ∞ for y= 0.

In this algorithm, Cost (h(x)) is linear function of h(x) such that, the value is in range of 100 to 0 for y = 1 and 0 to 100 for y = 0. The prediction threshold is 0.5. H(x) is 0.5 at hour value 2.75.