Logistic Regression

Prof. G Panda,
FNAE,FNASc,FIET(UK)
IIT Bhubaneswar

Outline

- What is Regression
- Basics of Linear Regression
- Logistic Regression: what and why
- Example
- Linear vs Logistic Regression
- Use-cases

What is Regression

- **Regression analysis** is a set of statistical processes for estimating the relationships among variables.
- It is a predictive modelling technique.
- It estimates relationship between a dependent variable(target) and one or more independent variables(predictors or features).

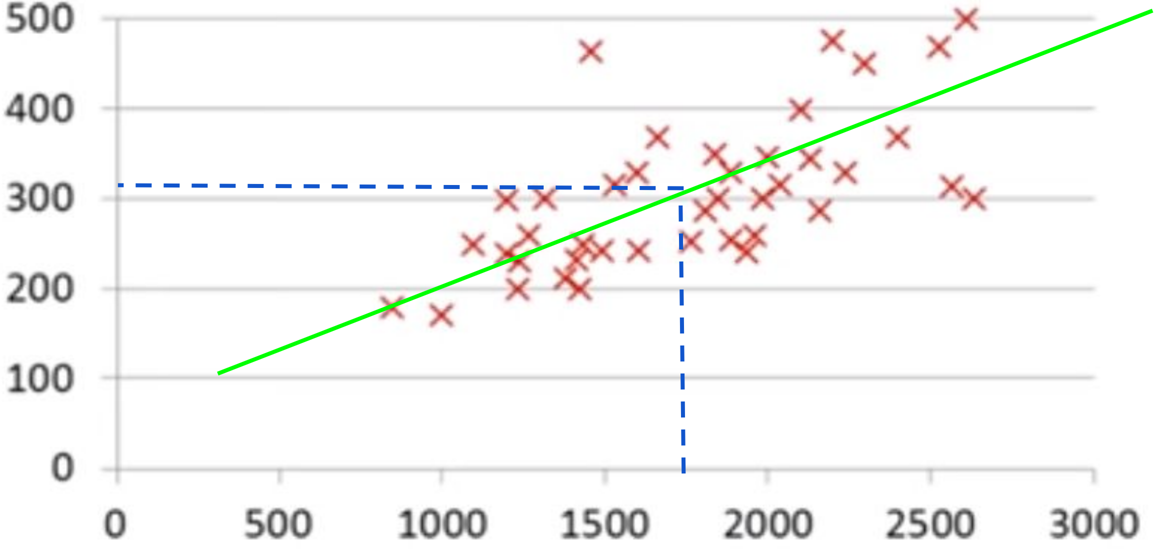
Example

- Prediction of house price from house size.
- House price is the dependent variable because based on the house size its price value varies.
- House size is independent(free) variable as it doesn’t depend on any factor.

Example

- Plot showing prediction of house price from house size.

Prices (In 1000s of dollors)



Size(feet^2)

Given, house size 1700 feet^2

Predicted value is 310K($)

Basics of Linear Regression

- It is a supervised learning algorithm used for prediction.
- Fits a straight line to data, so that the output variable (dependent variable) varies linearly based on the input variable(in-dependent variable).
- Line equation can be written as

$$y = mx + c.$$

Where, m is slope
c is y-axis intercept
x is input variable
y is output variable

Linear Regression: Example

Training data of house prices

Size in feet^2 (x)	Price($) in 1000's(y)
2104	460
1416	232
1534	315
852	178
...	...

Notation:

m = Number of input variables

x's = "input" variable / features

y's = "output variable"

$(X^{(i)}, Y^{(i)})$ is i^{th} training example

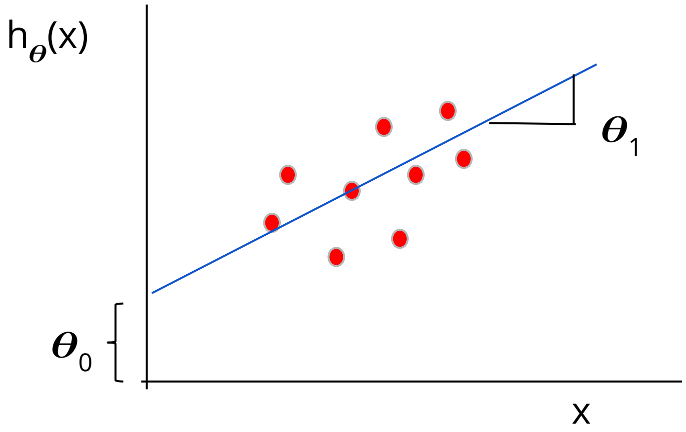
Linear Regression

- We need to come up with hypothesis which maps input variable to output variable linearly.

$h_{\theta}(x) = \theta_0 + \theta_1 x$

θ_i's are Parameters.

x is input variable.



This is Linear Regression with one variable, called **univariate Linear Regression**

Linear Regression

- **Idea:** choose θ_0 , θ_1 so that predicted value $h_\theta(x)$ is close to true value y for our training examples(x,y).
- So difference between x and y should be as small as possible.
- To achieve that, it is required to minimize mean squared error. This is called as cost function.

$$\underset{\theta_0\,,\,\theta_1}{\text{minimize}}\ \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

- ½ is multiplied to make math easier.

Simplified Linear Regression

To have best fit of the line to the training set, it is required to come up with parameters chosen by minimizing the cost function.

cost function:

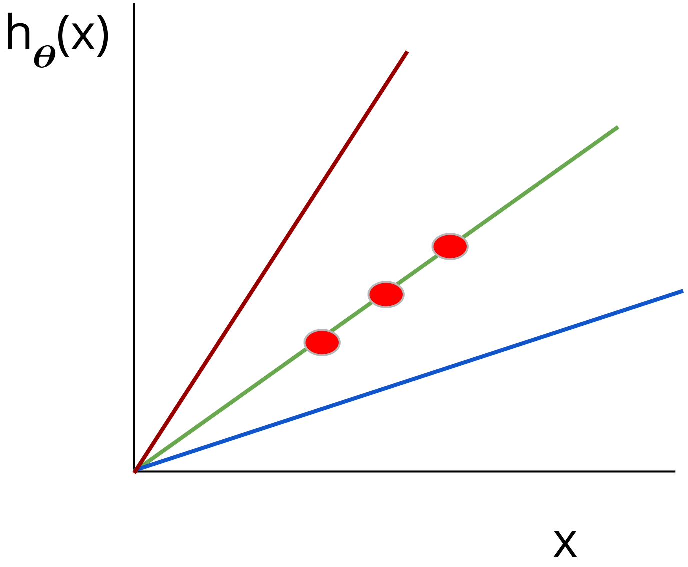
$$E(\theta_0 , \theta_1) = \frac{1}{2m}\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \quad \text{and} \quad h_\theta(x^{(i)}) = \theta_0 + \theta_1 x^{(i)}$$

To analyze these functions, consider only one parameter θ_1 taking $\theta_0 = 0$

So, $E(\theta_1) = \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$ and $h_\theta(x^{(i)}) = \theta_1 x^{(i)}$

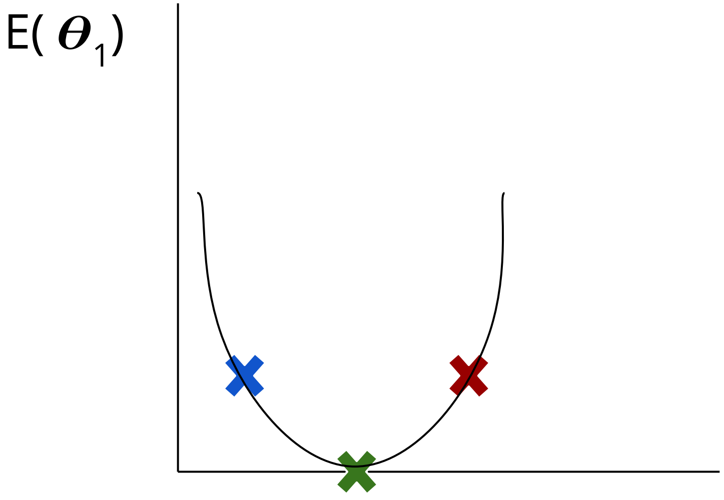
Simplified Linear Regression

Hypothesis function



Points in the plot indicate data points and the lines are different hypothesis that fit data.

Cost function



Points in the plot indicate computed cost function values for the corresponding hypothesis.

Linear Regression: Gradient Descent

It is the algorithm used to minimize the cost function(E).

Applying it to cost function(E)

Outline:

- Start with some θ_0 , θ_1 . (say $\theta_0 = 0$, $\theta_1 = 0$)
- Keep changing θ_0 , θ_1 to reduce E(θ_0 , θ_1) until we hopefully end up at minimum.

Gradient Descent

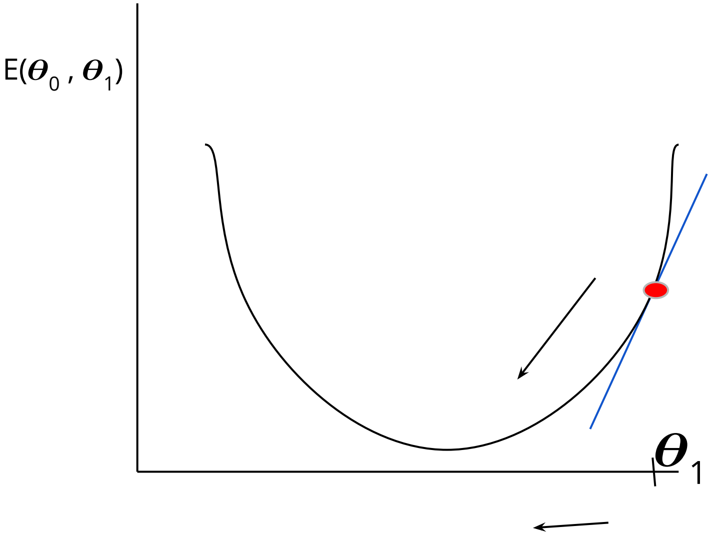
- Gradient Descent algorithm

  Repeat until convergence{

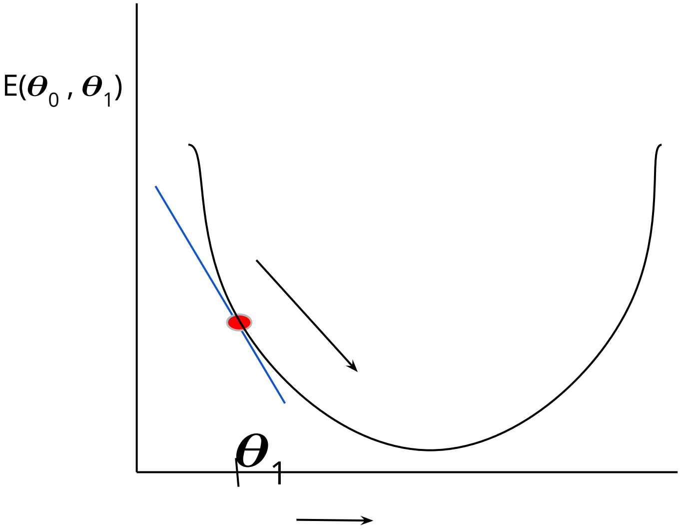
  $$\boldsymbol{\theta}_i := \boldsymbol{\theta}_i \mid \alpha \frac{\partial}{\partial \boldsymbol{\theta}_i} E(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) \quad (\text{ for i =0 and i =1})$$

  }

Gradient Descent



$$\frac{\partial}{\partial \boldsymbol{\theta}_i} E(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) \geq 0$$



$$\frac{\partial}{\partial \boldsymbol{\theta}_i} E(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) \leq 0$$

Linear Regression

$$\frac{\partial}{\partial \boldsymbol{\theta}_i} E(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = \frac{\partial}{\partial \boldsymbol{\theta}_i} \frac{1}{2m} \sum_{i=1}^{m} (h_{\boldsymbol{\theta}}(x^{(i)}) - y^{(i)})^2$$

$$= \frac{\partial}{\partial \boldsymbol{\theta}_i} \frac{1}{2m} \sum_{i=}^{m} (\boldsymbol{\theta}_0 + \boldsymbol{\theta}_1 x \ - y^{(i)})^2$$

For $i = 0$: $\frac{\partial}{\partial \boldsymbol{\theta}_0} E(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = \frac{1}{m} \sum_{i=1}^{m} (h_{\boldsymbol{\theta}}(x^{(i)}) - y^{(i)})$

$i = 1$: $\frac{\partial}{\partial \boldsymbol{\theta}_1} E(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = \frac{1}{m} \sum_{i=1}^{m} (h_{\boldsymbol{\theta}}(x^{(i)}) - y^{(i)}).x^{(i)}$

Multi-variable Linear Regression

$h_{\theta}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots\ldots\ldots + \theta_n x_n$

For convenience of notation, define $x_0 = 1$

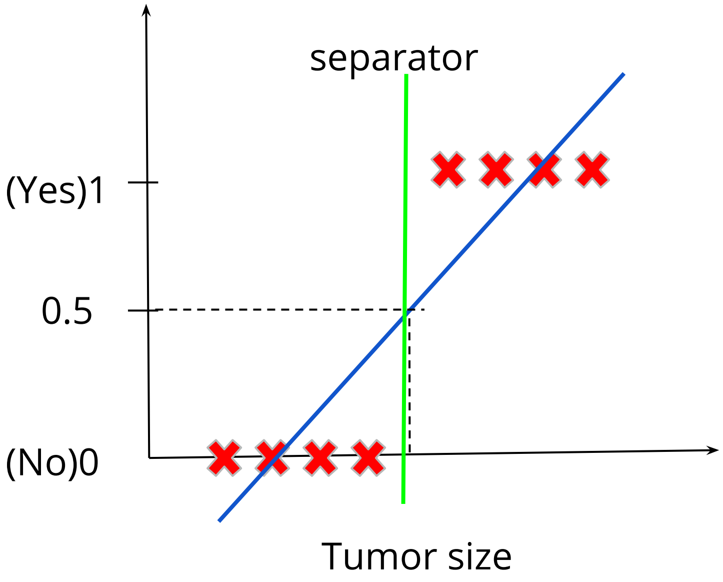
$$X = \begin{pmatrix} X_0 \\ X_1 \\ X_2 \\ . \\ . \\ X_n \end{pmatrix} \qquad \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ . \\ . \\ \theta_n \end{pmatrix}$$

Therefore, $h_{\theta}(x) = \theta^T X$

For Multi-variable Linear Regression use the above hypothesis and follow same procedure of single variable Linear Regression.

What is wrong with Linear Regression

In a classification problem for example cancer Tumor(Benign/Malignant)



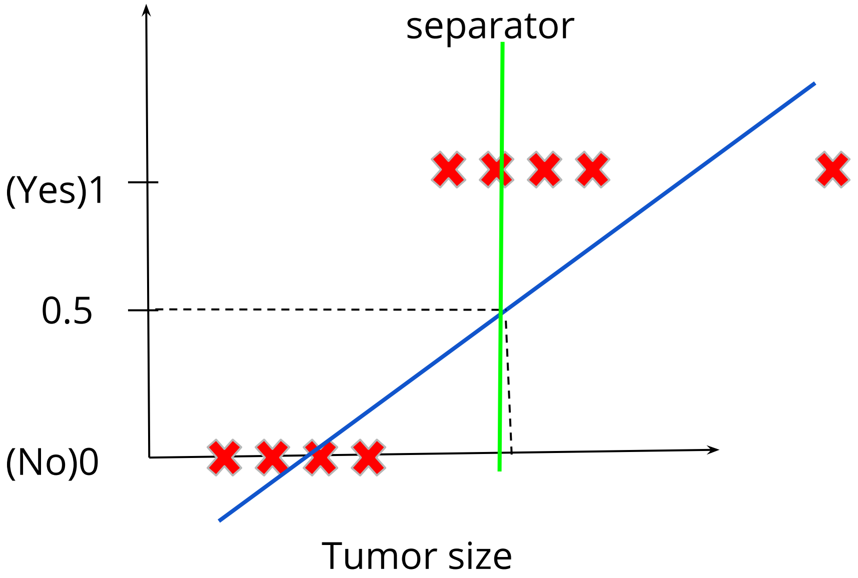
Classification is done based on threshold

If $h_\theta(x) \geq 0.5$, predict "y = 1"

If $h_\theta(x) < 0.5$, predict "y = 0"

What is wrong with Linear Regression

If a new training sample is added



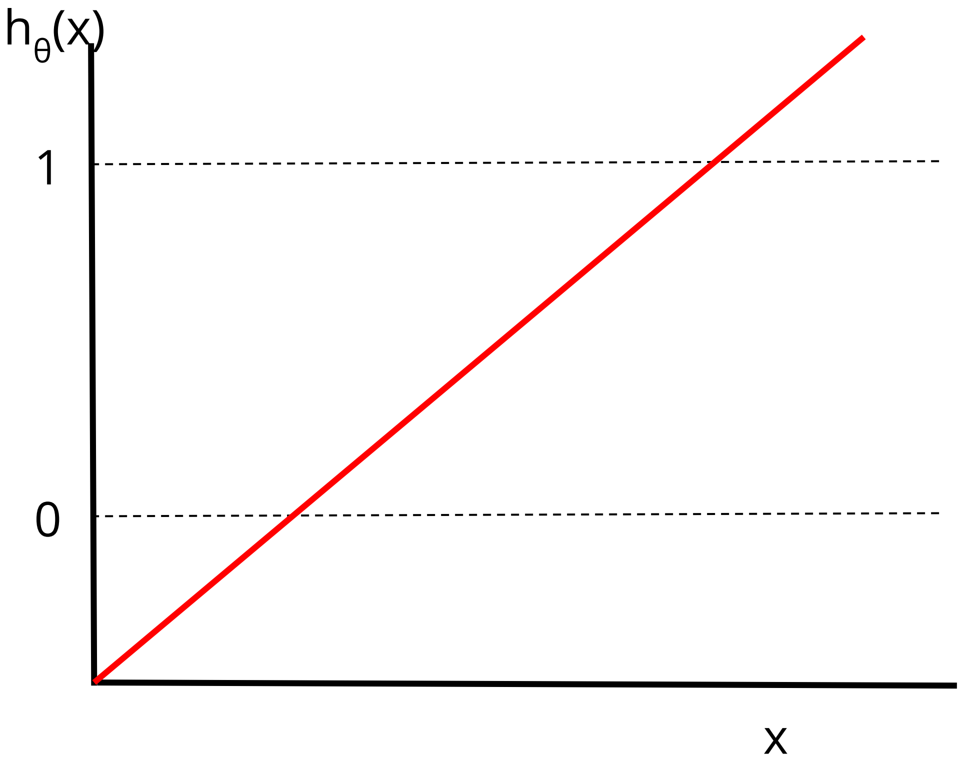
Classification is done based on threshold

If $h_\theta(x) \geq 0.5$, predict "y = 1"

If $h_\theta(x) < 0.5$, predict "y = 0"

Here Linear Regression fails to predict the actual class.

What is wrong with Linear Regression

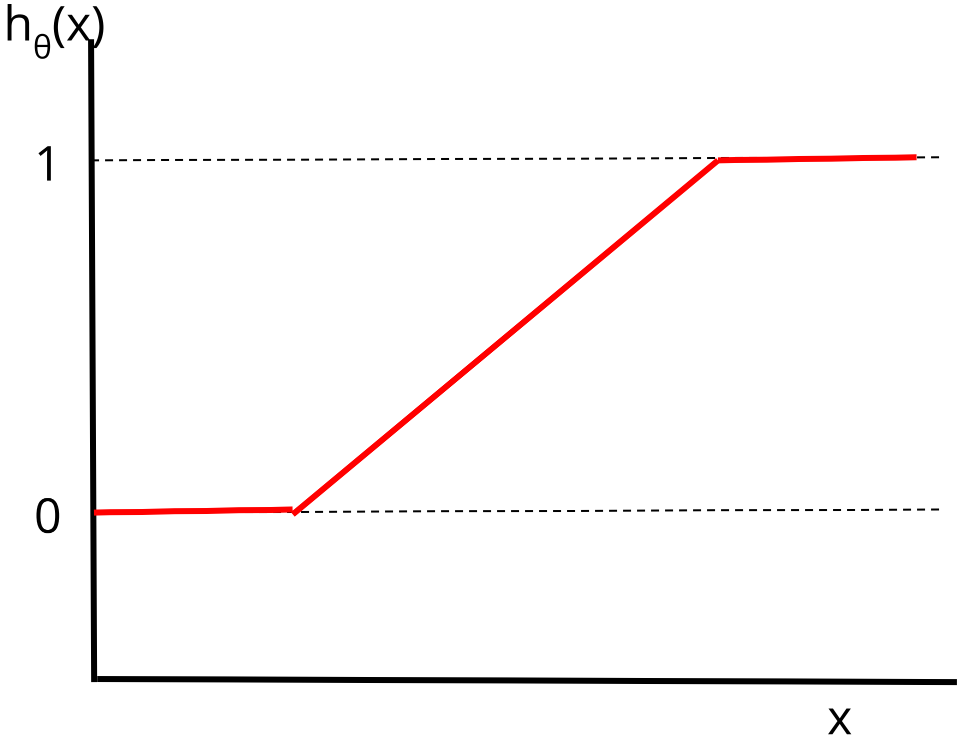


Classification y = 0 or 1

but here **$h_\theta(x)$ can be > 1 or < 0**

So Linear Regression is not good idea for classification problem.

Requirement for classification



Classification y = 0 or 1

Requires $0 \leq h_\theta(x) \leq 1$

Logistic Regression : what and why

- It is a Supervised learning algorithm used for classification.
- Classification :
  - Email : spam / not spam ?
  - Tumor : Malignant / Benign ?
  - Online Transactions : Fraudulent(Yes/No) ?

Here in two class classification output label $y \in \{0,1\}$

0 - for Negative class(e.g., benign tumor)

1 - for Positive class(e.g., malignant tumor)

Logistic Regression

Hypothesis of Logistic Regression should be such a way that

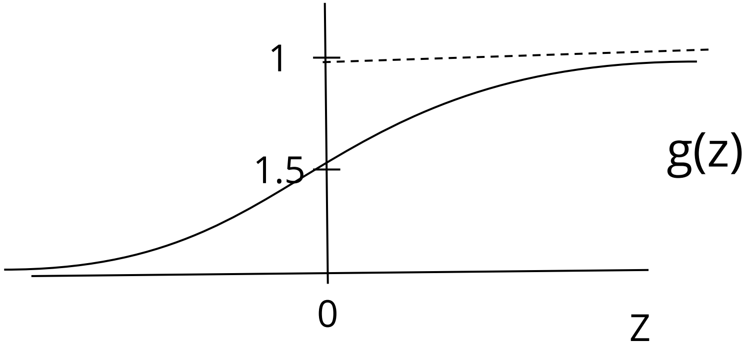
$0 \le h_\theta(x) \le 1$

Previously, $h_\theta(x) = \theta^T X$

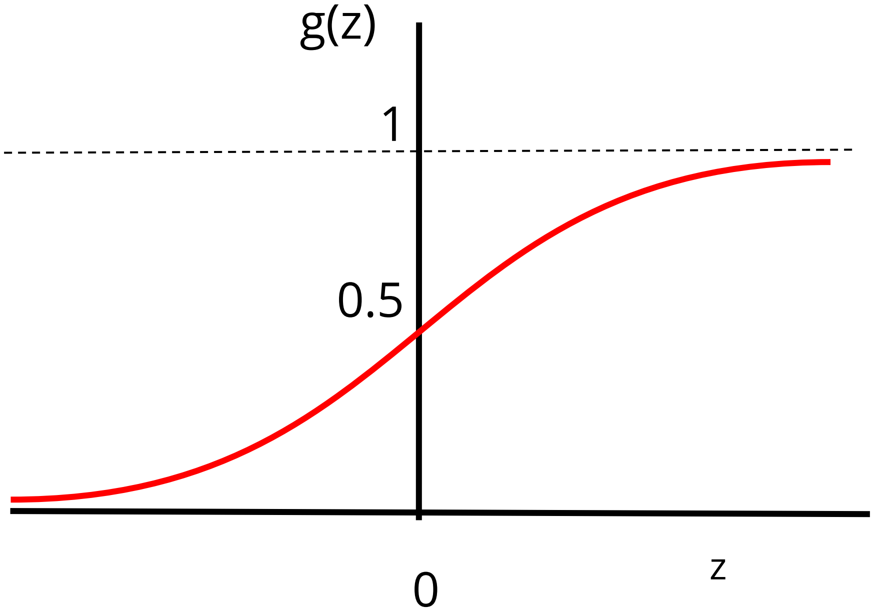
Now, $\mathbf{h_\theta(x) = g(\theta^T X)}$ and $\mathbf{g(z) = 1/(1+e^{-z})}$

where $\mathbf{z = \theta^T X}$

Where, g(z) is called **sigmoid function or Logistic function.**



Logistic Regression: Sigmoid function



It has asymptotes at 0 and 1

As

$z \longrightarrow -\infty$, $g(z) \longrightarrow 0$

$z \longrightarrow \infty$, $g(z) \longrightarrow 1$

$0 \leq g(z) \leq 1$

Logistic Regression

- Interpreting hypothesis output: Example of Tumor
- $h_\theta(x)$ = estimated probability that y=1 (tumor being malignant) on input x.
- Suppose if $h_\theta(x) = 0.7$ it implies that there is 70% chance of tumor being malignant.
- $h_\theta(x) = P(y = 1|x;\theta)$ is "probability that y =1, given x, parameterized by θ.
- We can also compute $P(y = 0|x;\theta) = 1 - P(y = 1|x;\theta)$

Logistic Regression

In Two class problem

Suppose predict "y = 1" if $h_\theta(x) \geq 0.5$

predict "y = 0" If $h_\theta(x) < 0.5$

For $\mathbf{h_\theta(x) \geq 0.5}$

=> $g(z) \geq 0.5$ (where $z = \theta^T X$)

=> $z \geq 0$

So whenever $\boldsymbol{\theta^T X \geq 0}$,
predict "y = 1"

For $\mathbf{h_\theta(x) < 0.5}$

=> $g(z) < 0.5$ (where $z = \theta^T X$)

=> $z < 0$

So whenever $\boldsymbol{\theta^T X < 0}$,
predict "y = 0"

Logistic Regression : Decision Boundary

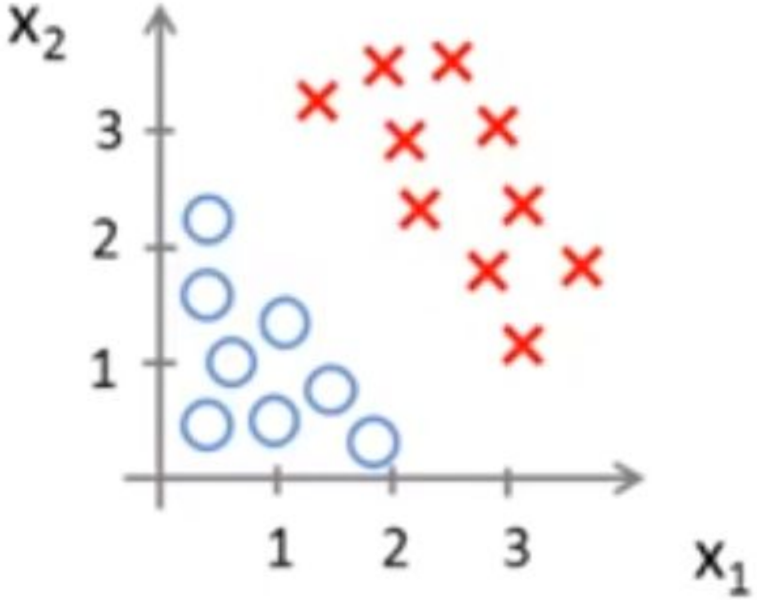
Linear decision boundaries

Let , $h_{\theta(x)} = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$

If we choose parameters

$\theta_0 = -3, \; \theta_1 = 1$ and $\theta_2 = 1$

So ,$\theta^T X = -3 + x_1 + x_2$



Logistic Regression : Decision Boundary

Predict y = 1 if $\boldsymbol{\theta^T X \geq 0}$

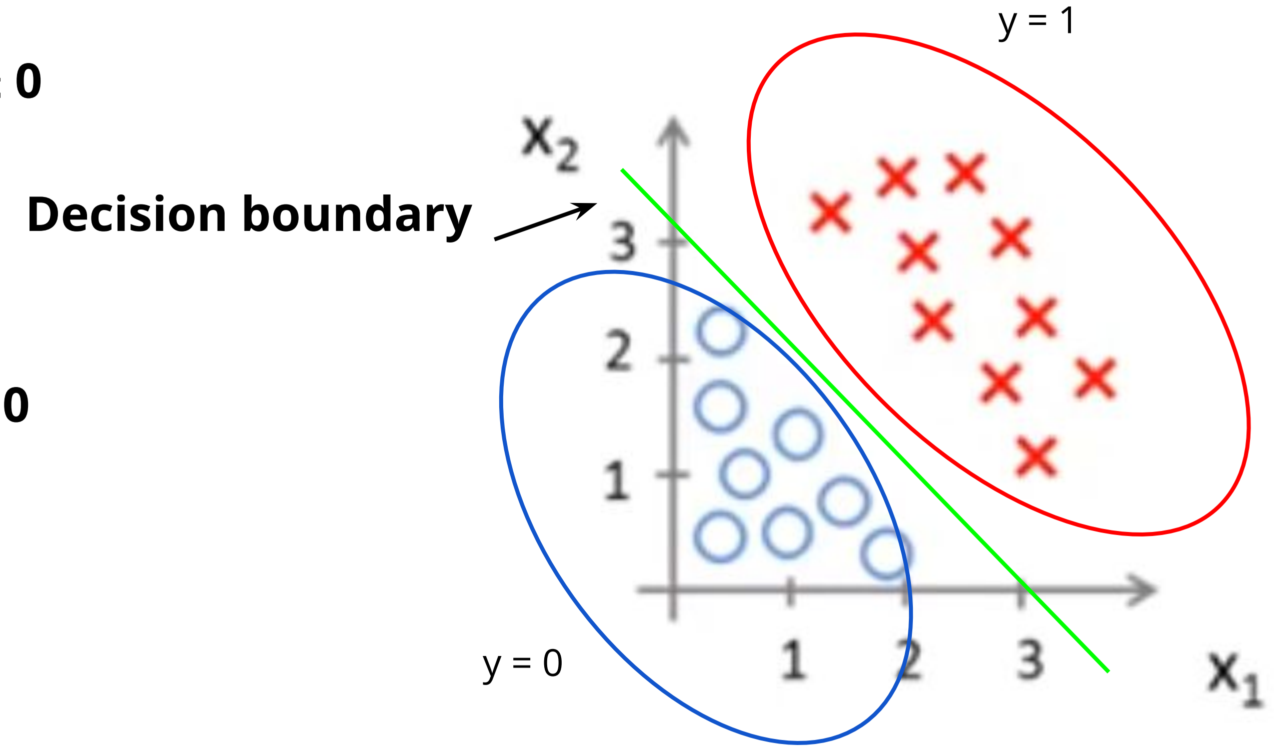
$-3+x_1 + x_2 \geq 0$

$\Rightarrow x_1 + x_2 \geq 3$

Predict y = 0 if $\boldsymbol{\theta^T X < 0}$

$-3+x_1 + x_2 < 0$

$\Rightarrow x_1 + x_2 < 3$



Logistic Regression : Decision Boundary

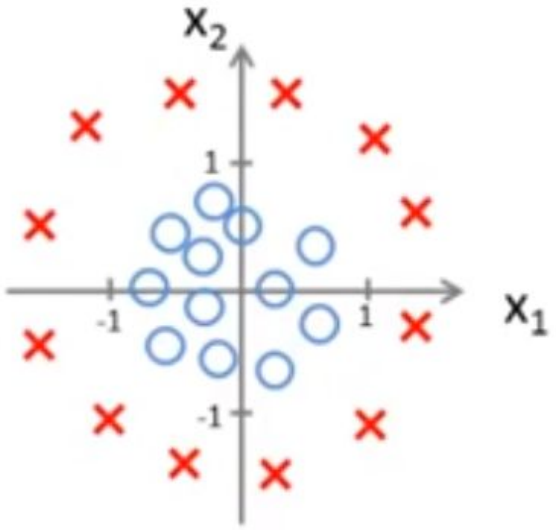
Non-linear decision boundaries

Let $h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$

If we choose parameters

$\theta_0 = -1, \ \theta_1 = 0, \ \theta_2 = 0, \ \theta_3 = 1, \ \theta_4 = 1$

So , $\theta^T X = -1 + x_1^2 + x_2^2$ (Equation of circle)



Logistic Regression : Decision Boundary

Non-linear decision boundaries

Predict y = 1 if $\boldsymbol{\theta^T X \geq 0}$

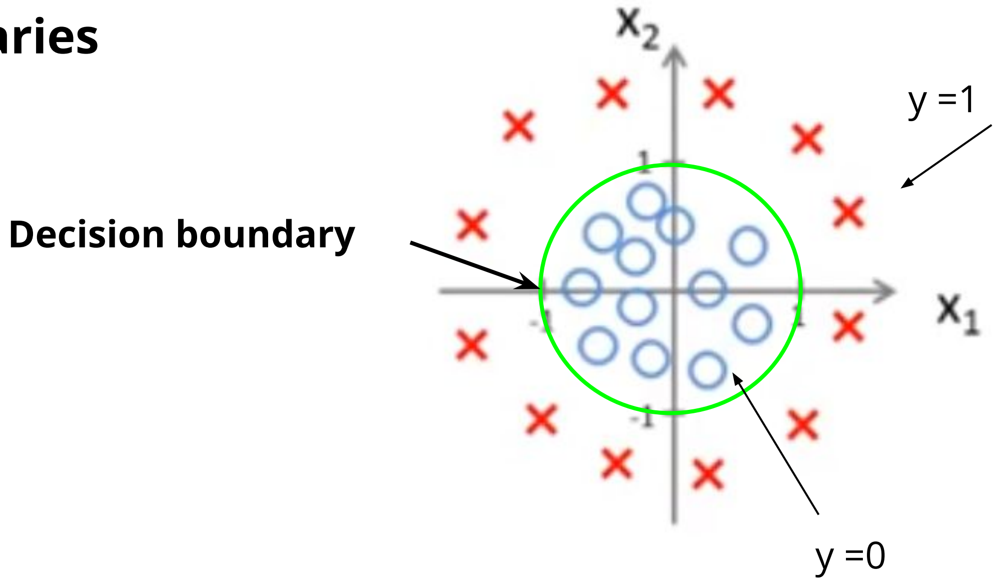
$-1 + x_1^2 + x_2^2 \geq 0$

$\Rightarrow x_1^2 + x_2^2 \geq 1$

Predict y = 0 if $\boldsymbol{\theta^T X < 0}$

$-1 + x_1^2 + x_2^2 < 0$ 2

$\Rightarrow x_1^2 + x_2^2 < -1$



Logistic Regression

Training set: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(m)}, y^{(m)})\}$

m examples $\quad x \in \begin{bmatrix} x_0 \\ x_1 \\ \cdots \\ x_n \end{bmatrix} \qquad x_0 = 1, y \in \{0, 1\}$

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

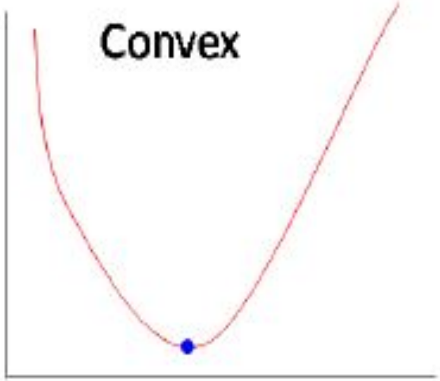
How to choose parameters θ ?

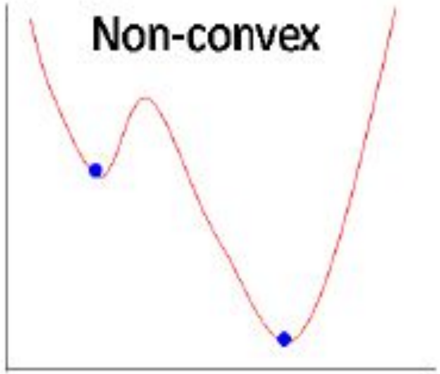
Andrew Ng

Cost function - Convergence

Convex function : Function on which Gradient descent when applied will converge to global minimum.

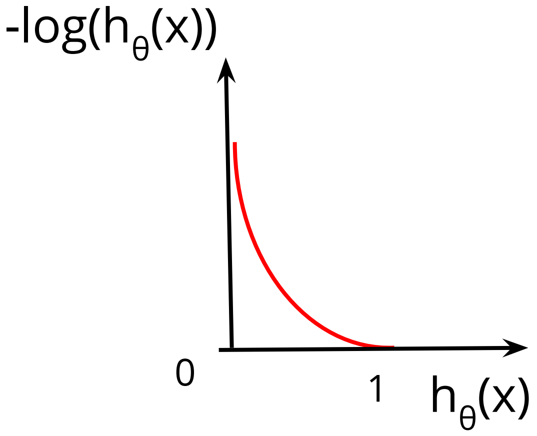
Non-convex function : Function on which Gradient descent when applied may or may not converge to global minimum.





Logistic Regression : Cost function

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$



Cost $= 0$ if $y = 1$, $h_\theta(x) = 1$

But as $h_\theta(x) \longrightarrow 0$

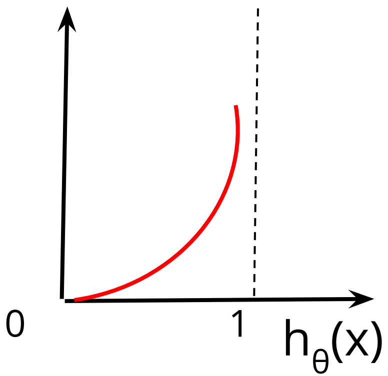
Cost $\longrightarrow \infty$

Captures the intuition that if $h_\theta(x) = 0$, (predict $P(y = 1 | x;\theta) = 0$) but $y=1$, it will penalize the learning algorithm by a very large cost.

Logistic Regression : Cost function

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

$-\log(1 - h_\theta(x))$



Cost $= 0$ if $y = 0$, $h_\theta(x) = 0$

But as $h_\theta(x) \longrightarrow 1$

Cost $\longrightarrow \infty$

Captures the intuition that if $h_\theta(x) = 1$, (predict $P(y = 0 | x;\theta) = 0$) but $y=0$, it will penalize the learning algorithm by a very large cost.

Why this cost function

Let

$\hat{y} = P(y = 1|x)$ {probability that y=1 given x}

$1 - \hat{y} = P(y = 0|x)$ {probability that y=0 given x}

$P(y\ |x) = \hat{y}^{y} + (1 - \hat{y})^{(1-y)}$

If y=1 => $P(y|x) = \hat{y}$

$\log(P(y\ |x)) = y\log(\hat{y}) + (1-y)\log(1-\hat{y})$

$= -L(\hat{y}, y)$ {Loss or cost function}

Why this cost function

Therefore,

$\log(P(y \mid x)) = -L(\hat{y}, y)$

- This negative function is because when we train, we need to maximize the probability by minimizing loss function.
- Decreasing the cost will increase the maximum likelihood $(P(y|x))$ assuming that samples are drawn from an identically independent distribution.
- This cost function is convex function and Gradient descent on this cost function will reach global minimum.

Cost function - Convergence

- Cost function of Linear Regression which is mean squared error is non-convex function. It may or may not converge to optimal solution. So it is generally not used.
- Cost function of Logistic Regression which is called as cross entropy is convex function. It will converge to optimal solution. So it is most widely used algorithm

Logistic Regression

Now we need to fit the parameters θ using the hypothesis of Logistic Regression. Cost function J(θ):

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

Note: $y = 0$ or 1 always

Cost can be written as:

$\mathrm{Cost}(h_\theta(x),y) = -y\log(h_\theta(x)) - (1-y)\log(1-h_\theta(x))$

Logistic Regression

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$
$$= -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)}\log h_\theta(x^{(i)}) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))]$$

To fit parameters θ:

$$\min_\theta J(\theta)$$

To make a prediction given new x:

Output $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$

Logistic Regression

Gradient Descent

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))]$$

Want $\min_\theta J(\theta)$:

Repeat $\{$

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

$\}$ (simultaneously update all θ_j)

Logistic Regression

Gradient Descent

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))]$$

Want $\min_\theta J(\theta)$:

Repeat $\{$

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

(simultaneously update all θ_j)

$\}$

Logistic Regression Example

Example classification problem of cancer Tumor : Benign/Malignant.

Data :

Tumor size	Tumor
1	Benign
2	Benign
3	Malignant
4	Malignant
5	Malignant

For Benign label y = 0 and for Malignant label y = 1.

Logistic Regression Example

Since it is two class and single variable problem, hypothesis can be written as

$h_\theta(x) = g(\theta_0 + \theta_1 x)$ and $g(z) = 1/(1+e^{-z})$

Initialize θ_0 ,θ_1

1) Random initialization: Take a random number between [0,1] and and subtract with 0.5.

2)Initialize to zeros.

Logistic Regression Example

For simplicity consider initialization to zeros.

So, $\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ $\quad X = \begin{pmatrix} x_0 \\ x_1 \end{pmatrix} = \begin{pmatrix} 1 \\ x \end{pmatrix}$

For the first training sample (x,y) = (1,0)

For input x = 1, $h_\theta(x) = g(\theta^T X) = g(0 + 0x1) = g(0) = 1/(1+ e^{-0}) = 0.5$

Similarly for x=2,3,4,5 for all $h_\theta(x) = 0.5$

Logistic Regression Example

Finding cost

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))\right]$$

Substituting parameters (θ_0 ,θ_1) = (0,0)

Loss = 1.0

update parameters(weights) using gradient descent.

Repeat {

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$$

(simultaneously update all θ_j)

}

Logistic Regression Example

$\theta_{0\ new} = \theta_0 - \alpha((0.5 - 0) + (0.5 - 0) +(0.5 - 1) + (0.5 - 1) + (0.5 - 1))x1$

$= 0 - 0.1x(-0.5)\ x1$

$= 0.05$

$\theta_{1\ new} = \theta_1 - \alpha((0.5 - 0) + (0.5 - 0) +(0.5 - 1) + (0.5 - 1) + (0.5 - 1))x1$

$= 0 - 0.1x(-0.5)\ x1$

$= 0.05$

Logistic Regression Example

For the second training sample (x,y) = (2,0)

compute outputs with updated parameters.

For input x = 1, $h_\theta(x) = g(\theta^T X) = g(0.05 + 0.05x1) = 1/(1+e^{-0.1}) = 0.525$

input x = 2, $h_\theta(x) = g(0.05 + 0.05x2) = 1/(1+e^{-0.15}) = 0.537$

input x = 3, $h_\theta(x) = g(0.05 + 0.05x3) = 1/(1+e^{-0.2}) = 0.549$

input x = 4, $h_\theta(x) = g(0.05 + 0.05x4) = 1/(1+e^{-0.25}) = 0.562$

input x = 5, $h_\theta(x) = g(0.05 + 0.05x5) = 1/(1+e^{-0.3}) = 0.574$

Logistic Regression Example

Loss = 0.935

Parameters update:

$\theta_{0\ new} = \theta_0 - \alpha(\ (0.525 - 0) + (0.537 - 0) + (0.549 - 1) + (0.562 - 1) + (0.574 - 1))x1$

$= 0.05 - 0.1x(-0.251)\ x1 = 0.0751$

$\theta_{1\ new} = \theta_1 - \alpha(\ (0.525 - 0) + (0.537 - 0) + (0.549 - 1) + (0.562 - 1) + (0.574 - 1))x2$

$= 0.05 - 0.1x(-0.251)\ x2 = 0.100$

Logistic Regression Example

For the second training sample (x,y) = (3,1)

compute outputs with updated parameters.

For input x = 1, $h_\theta(x) = g(\theta^T X) = g(0.0751 + 0.100\ x1) = 1/(1 + e^{-0.175}) = 0.543$

input x = 2, $h_\theta(x) = g(0.0751 + 0.100x2) = 1/(1 + e^{-0.275}) = 0.568$

input x = 3, $h_\theta(x) = g(0.0751 + 0.100x3) = 1/(1 + e^{-0.375}) = 0.592$

input x = 4, $h_\theta(x) = g(0.0751 + 0.100x4) = 1/(1 + e^{-0.476}) = 0.616$

input x = 5, $h_\theta(x) = g(0.0751 + 0.100x5) = 1/(1 + e^{-0.576}) = 0.640$

Logistic Regression Example

Loss = 0.887

Parameters update:

$\theta_{0\ new} = \theta_0 - \alpha((0.543 - 0) + (0.567 - 0) +(0.591- 1) + (0.615- 1) + (0.638 - 1))x1$

$= 0.0751 - 0.1x(-0.038) x1 = 0.0789$

$\theta_{1\ new} = \theta_1 - \alpha((0.543 - 0) + (0.567 - 0) +(0.591- 1) + (0.615- 1) + (0.638 - 1))x3$

$= 0.100 - 0.1x(-0.038) x3 = 0.112$

Logistic Regression Example

Repeat the procedure till convergence.

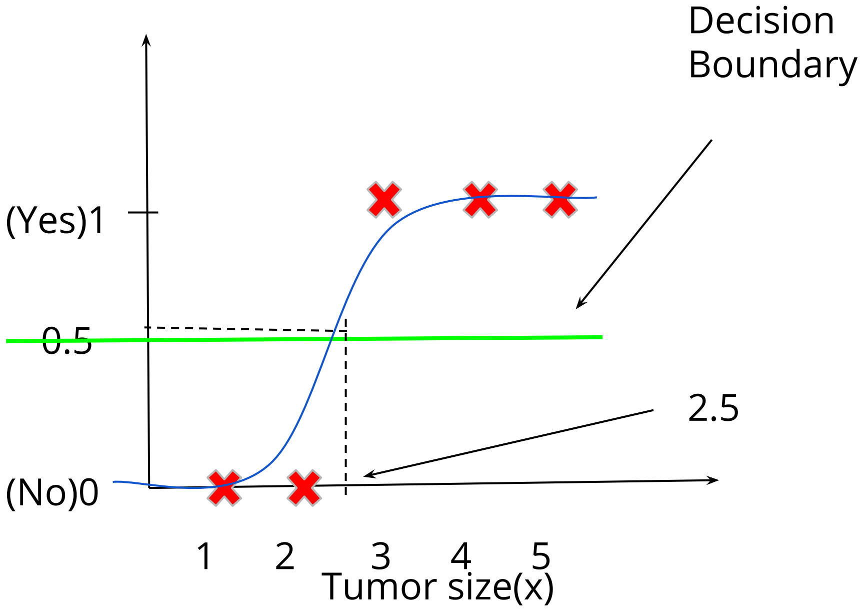
After convergence learned parameters(weights) are

$[\theta_0,\theta_1]$ = [0.0781, 0.1099]

For a tumor size = 2.5 the hypothesis $h_\theta(x) = 1/(1+e^{-0.0781+ 2.5x0.1099}) = 0.587$

Since, $h_\theta(x) > 0.5$ we can classify it as **Malignant (y =1)**

Logistic Regression Example



Classification is done based on threshold

If $h_\theta(x) \geq 0.5$, predict "y = 1"

If $h_\theta(x) < 0.5$, predict "y = 0"

For x = 2.5, 0. $h_\theta(x) = 0.587$

Logistic Regression: Multi-class

Multi class classification:

Email foldering/tagging :work, Friends, Family, Hobby

Medical diagram : Not Nil, cold , flu

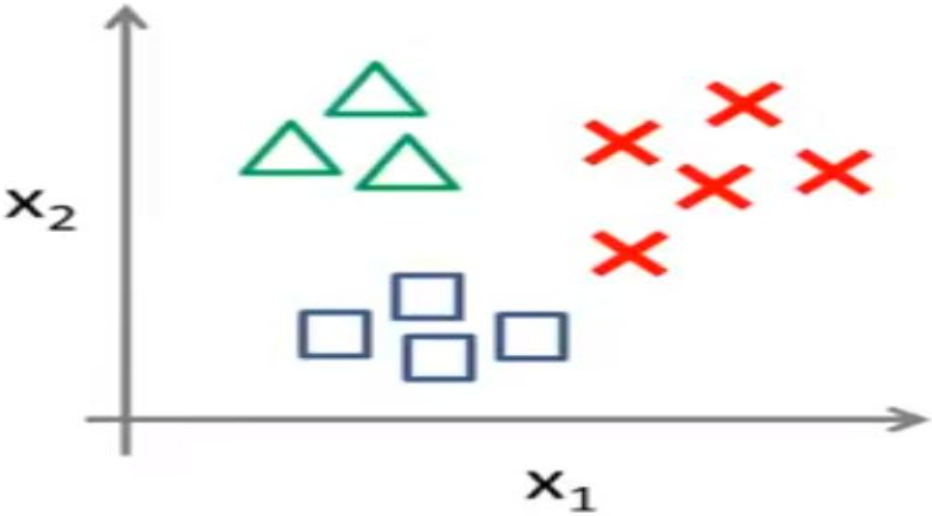
Weather : Sunny, Cloudy, Rain, Snow

Here output label y $\in$ 0,1,2, ...

Logistic Regression: Multi-class

Consider a three class classification problem with two input variables

x_1 and x_2.



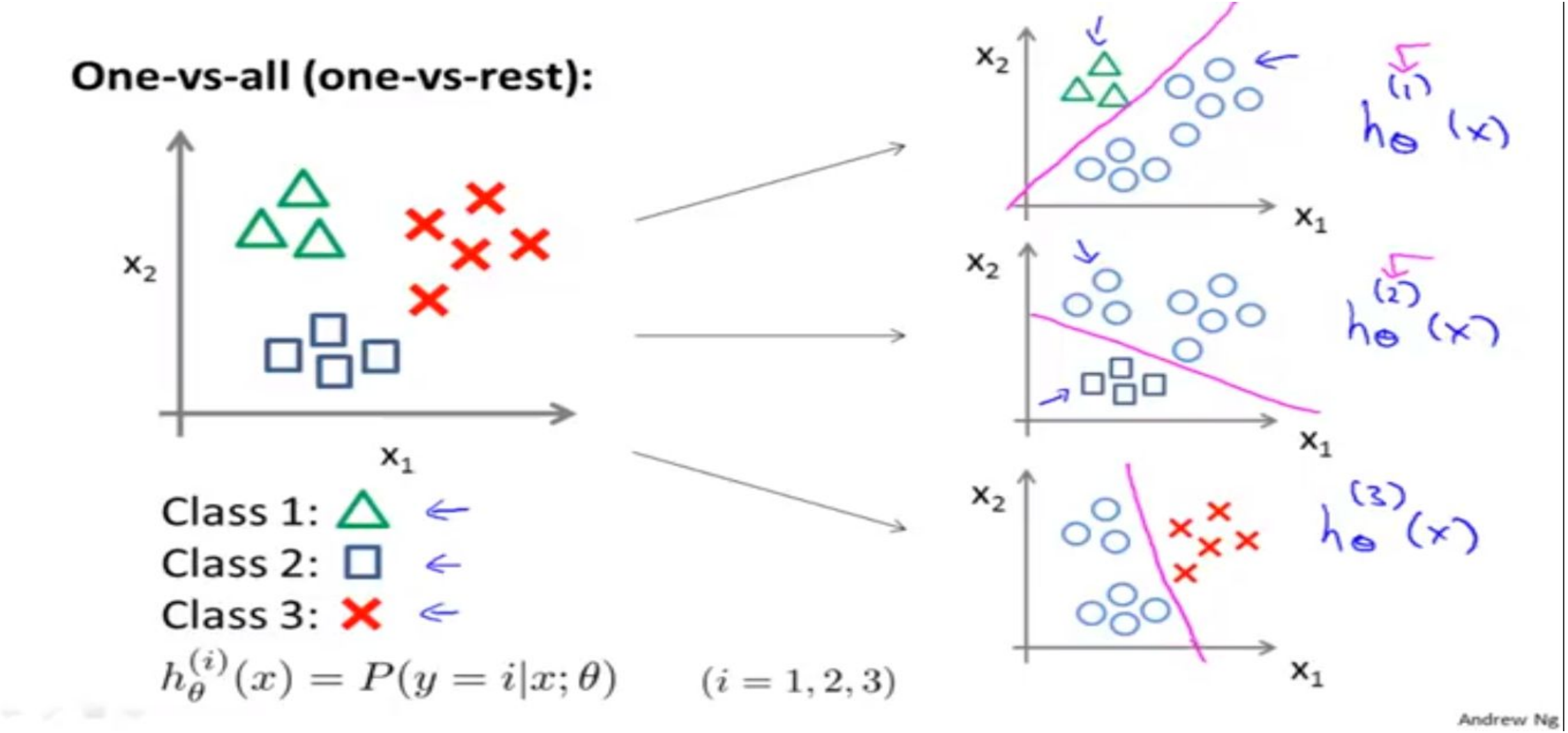
Logistic Regression: Multi-class

Procedure for n-class classification:

- We take the training set and turn the problem into ‘n’ separate binary classification problem.
- Create a fake training set with class ‘i’ assign label 1 and other classes with label 0.
- Train a Logistic Regression classifier $h_\theta^{(i)}(x)$ for each class i to predict the probability that y = i.
- On a new input x to make a prediction pick the class i that maximizes

$$\max_i h_\theta^{(i)}(x)$$

Logistic Regression: Multi-class



Linear vs Logistic Regression

Linear Regression

1. Continuous Variables.
2. Solves Regression Problems.
3. Straight line.
4. Non- convex cost function.
5. Outcome values are predicted in range.

Logistic Regression

1. Categorical / Discrete Variables.
2. Solves classification Problems.
3. Sigmoid - curve.
4. Convex cost function
5. Outcome is Discrete/Categorical

Use-Cases

- Weather Prediction.
  - Rainy or not / Sunny or not / cloudy or no)
- Temperature prediction(Linear Regression).
- Determine illness.
  - Multi-class classification problem.(Age, previous medical history.)
- Geographic Image Processing.

References

- Applied logistic regression ,31 July 1989,Book by David W. Hosmer.
- Applied Logistic Regression Analysis, 1995,Book by Scott Menard.
- Logistic Regression: A Primer,26 May 2000 Book by Fred Pampel.
- Machine Learning-Logistic Regression by Andrew Ng.

Thank You