

Logistic Regression

Notebook: Machine Learning

Created: 15-Apr-20 3:19 PM

Updated: 16-Apr-20 5:09 PM

Author: Patel Yas

URL: <https://www.coursera.org/learn/machine-learning/lecture/RJXfB/hypothesis-representation>

WHAT is Logistic Regression ?

it is an algorithm used to solve the classification problems,
in other words when dependent variable is categorical then it is used.

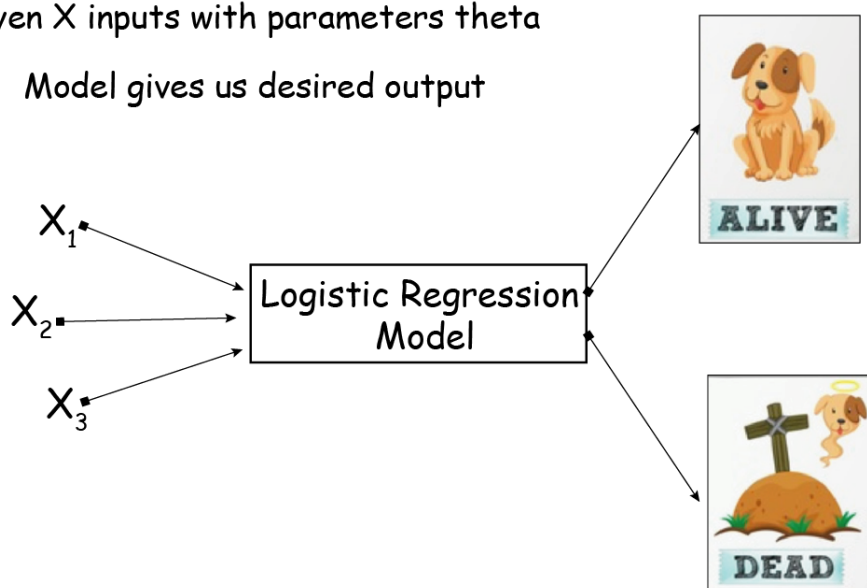
here, for example,

we trained model using logistic regression to predict whether a DOG is alive or dead. so
here,

We give X inputs in form dataset of different images of Dog ,
the trained Model will give us whether Dog is **alive or not**.

Given X inputs with parameters theta

Model gives us desired output



So here are some another examples of classification problems :

- 1) To predict whether an email is spam or not
- 2) To predict whether tumor is malignant or benign.
- 3) To predict whether human face is happy or sad.

SO What are the types of Logistic Regression ?

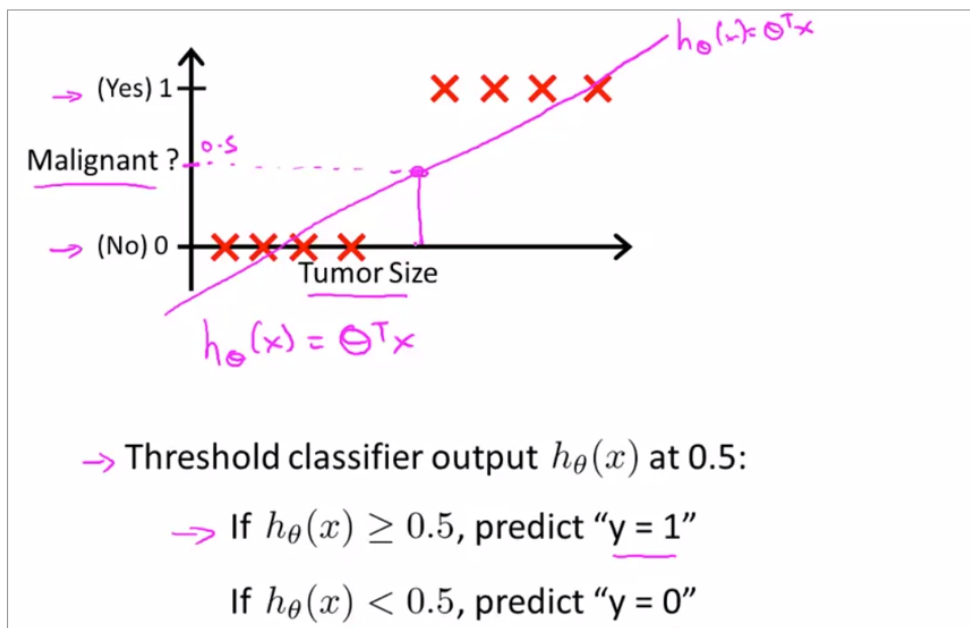
- 1) Binary Logistic Regression (any two possible outcomes ex : Email Whether spam or not)
- 2) Multinomial Logistic Regression (Three or more possible outcomes ex : email tagging - for home, work, friend)
- 3) Ordinal Logistic Regression (Three or more outcomes having categorical ordering ex- movie rating 1 to 5 (1,2,3,4,5)).

WHY Linear Regression is not used to solve classification problems?

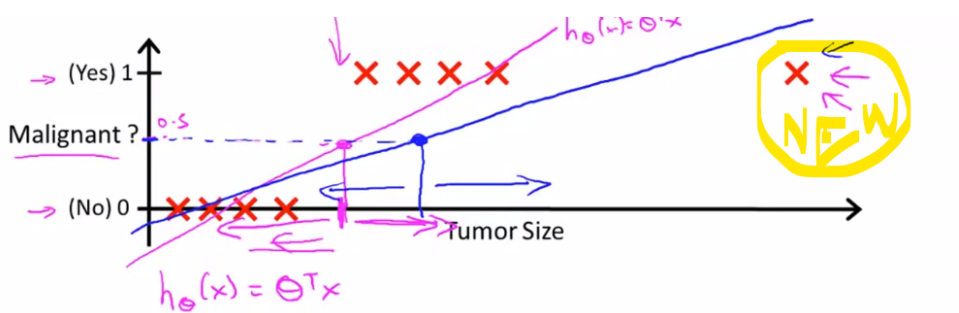
Here, first classification problem is given ! So we need to solve that problem !

Example : cancer : Malignant or Benign ?

IF we try to solve this problem with Linear regression with following dataset it will look like as Follows :



Now if we add some new dataset then predicated hypothesis well look like as follows :



So as per from above two test cases we can say that by luck sometimes the linear regression gives us GOOD hypothesis but it's just a luck.

Sometimes it gives us WORST hypothesis so for classification Problem we can't solve it with linear regression .

There is one another thing is that when we apply the linear regression algorithm for classification problem we get values for $h(\theta)$ as follows : $h(\theta) > 1$ or $h(\theta) < 0$

Though we have the 0 and 1 as our values for Tumor Classification Problem.

it should be $0 < h(\theta) < 1$ so that's why also we can't use linear regression algorithm to solve the classification problem.

Conclusion :

we can't use linear regression algorithm to solve the classification problem.

Logistic Regression :

Hypothesis Representation :

to satisfy $0 < h(\Theta) < 1$ property we use the **sigmoid** function or **logistic** function.

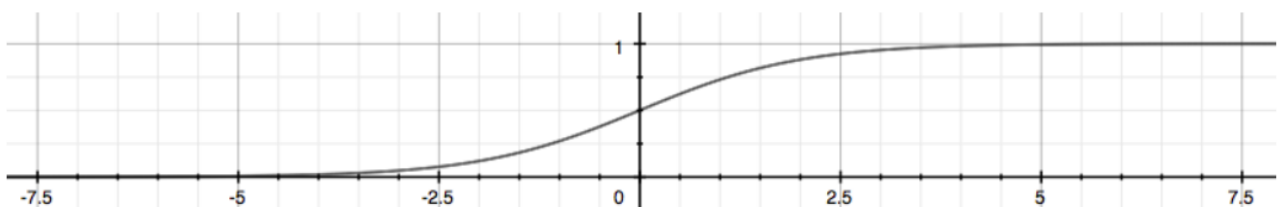
Logistic function :

$$h_{\theta}(x) = g(\theta^T x)$$

$$z = \theta^T x$$

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

Following graph shows us if we apply the logistic function to the classification problem it will look like this :



Now let's see , **interpretation of hypothesis** output :

$h(\Theta)$ = estimated probability that $y=1$ when input x
= $P(y=1 \mid x ; \Theta)$
= probability that $y=1$, given by x and parameterized Θ

Suppose we want to predict, from data x about tumor, whether it is malignant ($y=1$) or benign ($y=0$).

our logistic regression classifier gives us output for specific tumor as follows ,

$$h(\Theta) = P(y=1 \mid x ; \Theta) = 0.7 ,$$

So as per this equation we can say :

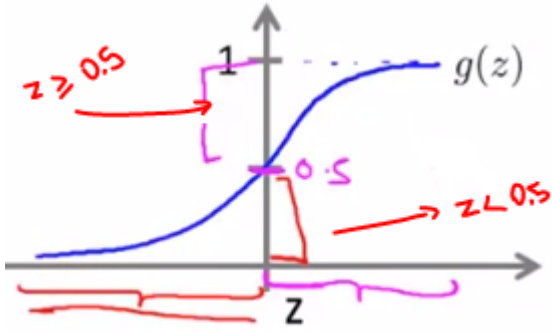
- 1) there is 70% chance of tumor being malignant.
- 2) there is 30% chance of tumor being benign.
- 3) $P(y=0 \mid x ; \Theta) = 0.3$

Conclusion :

$$P(y=1 \mid x ; \Theta) + P(y=0 \mid x ; \Theta) = 1$$

Decision Boundary:

for previous training example,

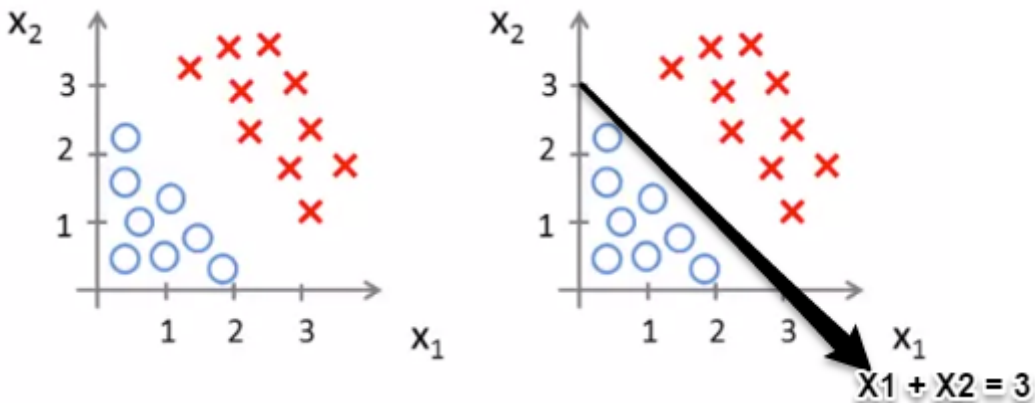


suppose predict $y=1$ if $h_{\theta}(x) \geq 0.5$ & when $\theta^T X \geq 0$ ($h_{\theta}(x) \geq 0.5$ when $z \geq 0$)

suppose predict $y=0$ if $h_{\theta}(x) < 0.5$ & when $\theta^T X < 0$ ($h_{\theta}(x) < 0.5$ when $z < 0$)

New Example :

$h_{\theta}(x) = g(\theta_0 + \theta_1 X_1 + \theta_2 X_2)$ and $[\theta_0, \theta_1, \theta_2] = [-3, 1, 1]$

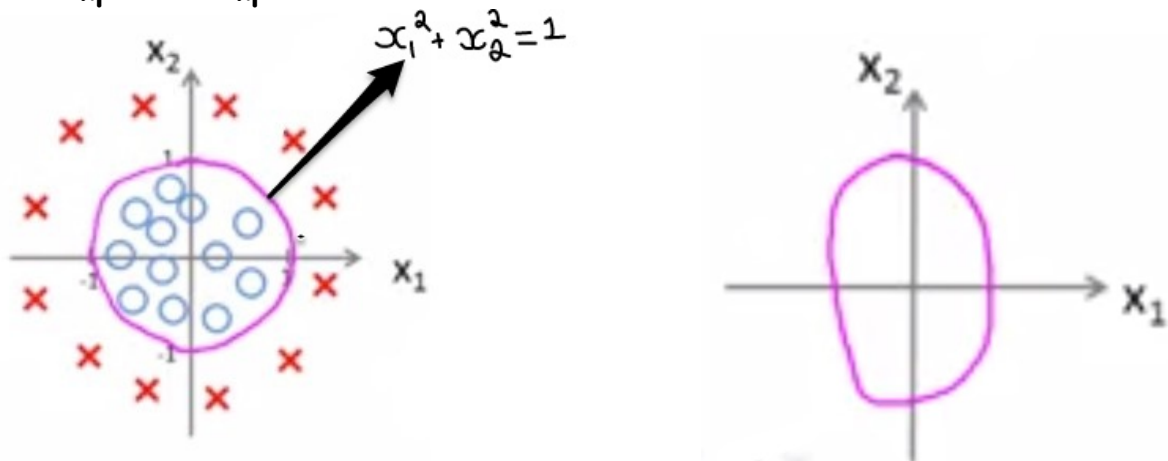


suppose predict $y=1$ when $-3 + X_1 + X_2 \geq 0$

so $-3 + X_1 + X_2 \geq 0 \Rightarrow X_1 + X_2 \geq 3$

SO here in this example $X_1 + X_2 = 3$ is decision boundary for that data set

Other complex examples :



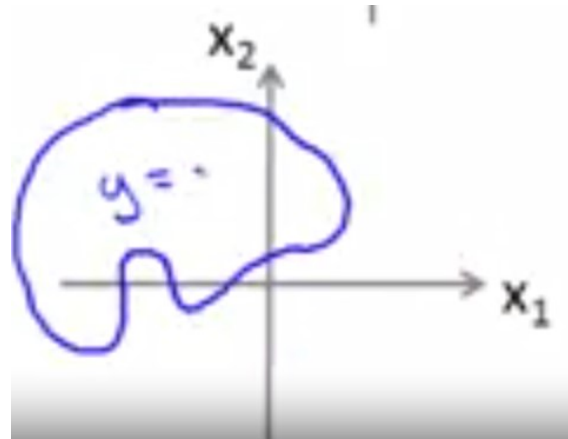
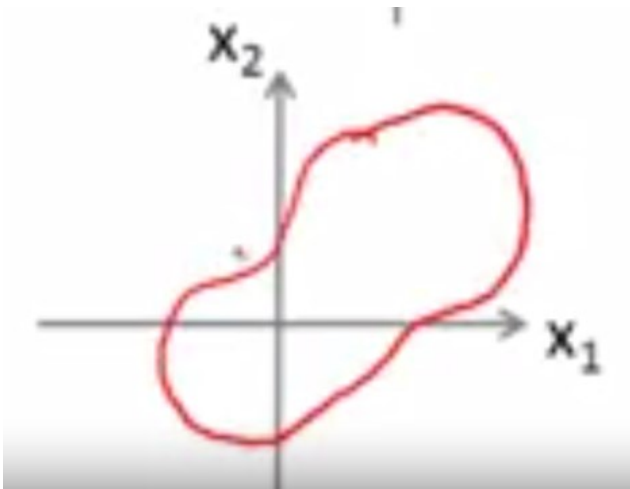
$h_{\theta}(x) = g(\theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_1^2 + \theta_4 X_2^2)$ & $[\theta_0, \theta_1, \theta_2, \theta_3, \theta_4] = [-1, 0, 0, 1, 1]$

suppose predict $y=1$ when $-1 + X_1^2 + X_2^2 \geq 0$ so $-1 + X_1^2 + X_2^2 \geq 0 \Rightarrow X_1^2 + X_2^2 \geq 1$

SO here in this example $X_1^2 + X_2^2 = 1$ is decision boundary for these data set.

complex hypothesis

-- > sometimes complex hypothesis look like these also



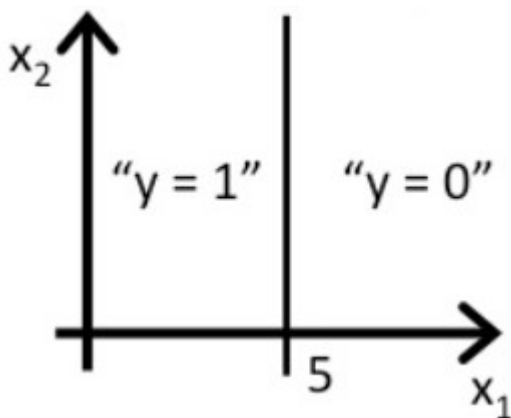
SO what is decision boundary ?

The **decision boundary** is the line that separates the area where $y = 0$ and where $y = 1$. It is created by our hypothesis function.

follow below example :

Consider logistic regression with two features X_1 & X_2 . suppose $[\theta_0, \theta_1, \theta_2] = [5, -1, 0]$ so that

$h_{\theta}(x) = g(5 - X_1)$. so below shows the decision boundary of $h_{\theta}(x)$.



Logistic Regression Model

Training set : $\{ (x^1, y^1), (x^2, y^2), \dots, (x^m, y^m) \}$ & m examples , how to choose parameters ?

cost function :

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

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

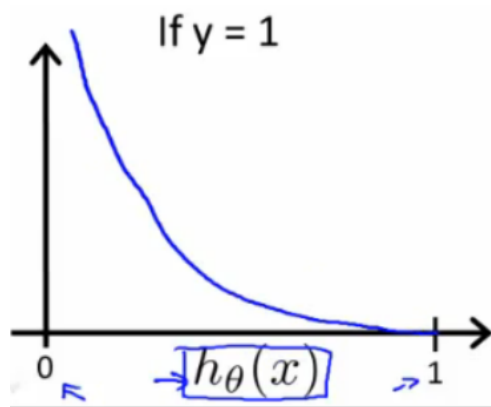
$$\text{Cost}(h_{\theta}(x), y) = -\log(1 - h_{\theta}(x)) \quad \text{if } y = 0$$

When $y=1$, we get following plot for $J(\Theta)$ vs $h_{\Theta}(x)$

we can say that

if $h_{\Theta}(x) = 1$ then $\text{cost}() = 0$,

if $h_{\Theta}(x) = 0$ then $\text{cost}() = \text{infinity}$.

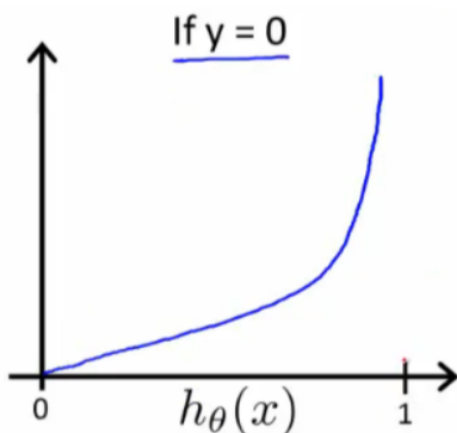


When $y=0$, we get following plot for $J(\Theta)$ vs $h_{\Theta}(x)$

we can say that

1) if $h_{\Theta}(x) = 0$ then $\text{cost}() = 0$,

2) if $h_{\Theta}(x) = 1$ then $\text{cost}() = \text{infinity}$.



Gradient Descent for logistic regression:

our cost function :

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(h_{\theta}(x^{(i)}), y^{(i)})$$
$$\begin{aligned} \text{Cost}(h_{\theta}(x), y) &= -\log(h_{\theta}(x)) && \text{if } y = 1 \\ \text{Cost}(h_{\theta}(x), y) &= -\log(1 - h_{\theta}(x)) && \text{if } y = 0 \end{aligned}$$

we can rewrite this equation as follows :

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

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

$$\text{when } y = 0 \text{ then } \text{Cost}(h_{\theta}(x), y) = -\log(1 - h_{\theta}(x))$$

so $J(\theta)$:

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

Gradient Descent :

General form of Gradient Descent :

$$\begin{aligned} &\text{Repeat} \{ \\ &\quad \theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \\ &\} \end{aligned}$$

replacing $J(\theta)$ in that general form :

Gradient descent look like this :

$$\begin{aligned} &\text{Repeat} \{ \\ &\quad \theta_j := \theta_j - \frac{\alpha}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)} \\ &\} \end{aligned}$$

A vectorized implementation of this form :

$$\theta := \theta - \alpha \frac{1}{m} \sum_{i=1}^m [(h_{\theta}(x^{(i)}) - y^{(i)}) \cdot x^{(i)}]$$

WHAT we do with this algorithm ? we use cost function to min($J(\theta)$).

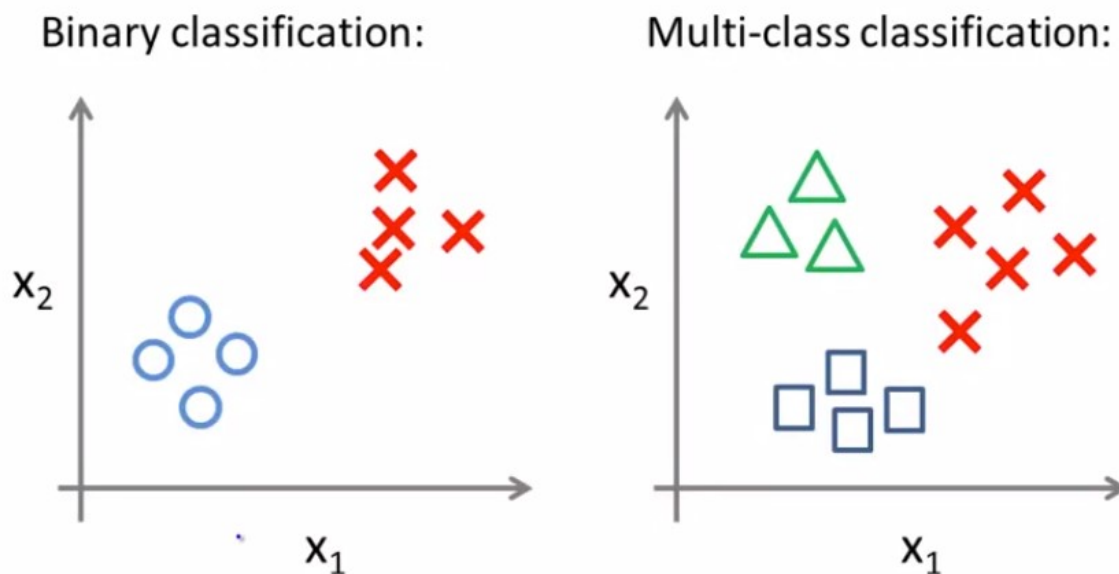
If Multiclass Classification is given, What we can do ?

what is multiclass classification ?

examples :

- 1) Email tagging : Work, Friends, Family, Hobby
- 2) Medical diagrams : not ill, cold, flu
- 3) Weather : sunnny, Cloudy, Rain, Snow

Here, is the comparison between binaryclass and multiclass classification :



How, we can solve this ?

here in above example there are three different clusters are given, imagine it is Medical daigrams with not ill, cold & flu clusters.

Clusters :

- 1) Green Triangles - Not ill Cluster
- 2) Blue Rectangles - cold Cluster
- 3) Red Cross - flu Cluster

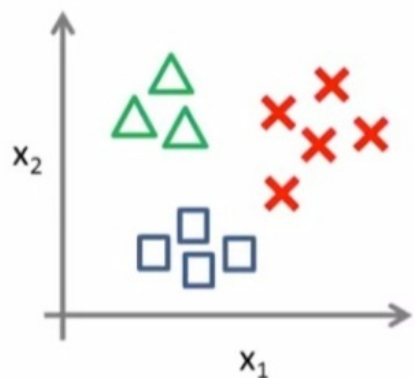
In order to solve this problem, we will divide this multi class problem into 3 different binary class problem



for division to binary class one cluster behave alone cluster and other two make one cluster by combined.

for example in below Binary class(1) is made by Cluster 1 : **Not ill cluster** and combined Cluster 2 : (**Combination of Cold and ful cluster**)

So there are **three Binary class** we can make and using older equation we can solve this problem,

One-vs-all (one-vs-rest):



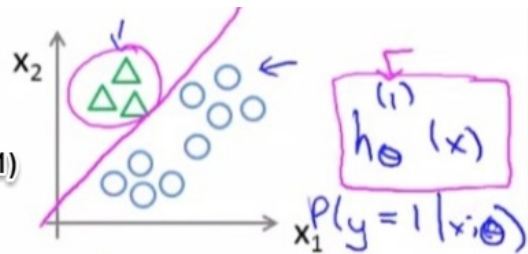
Class 1:  

Class 2:  

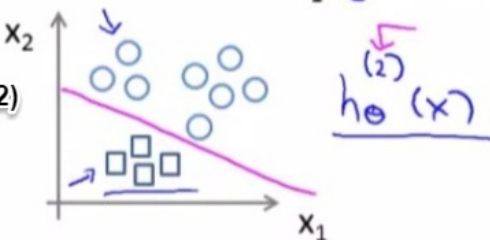
Class 3:  

$$h_{\theta}^{(i)}(x) = P(y = i|x; \theta) \quad (i = 1, 2, 3)$$

Binary CLass (1)



Binary CLass (2)



Binary CLass (3)

