Latakia University
Information Engineering Faculty
Artificial Intelligence Department
Fourth Year

Artificial Neural Networks

Lecture 4 : Logistic Regression

Dr. Wisam Ibrahim

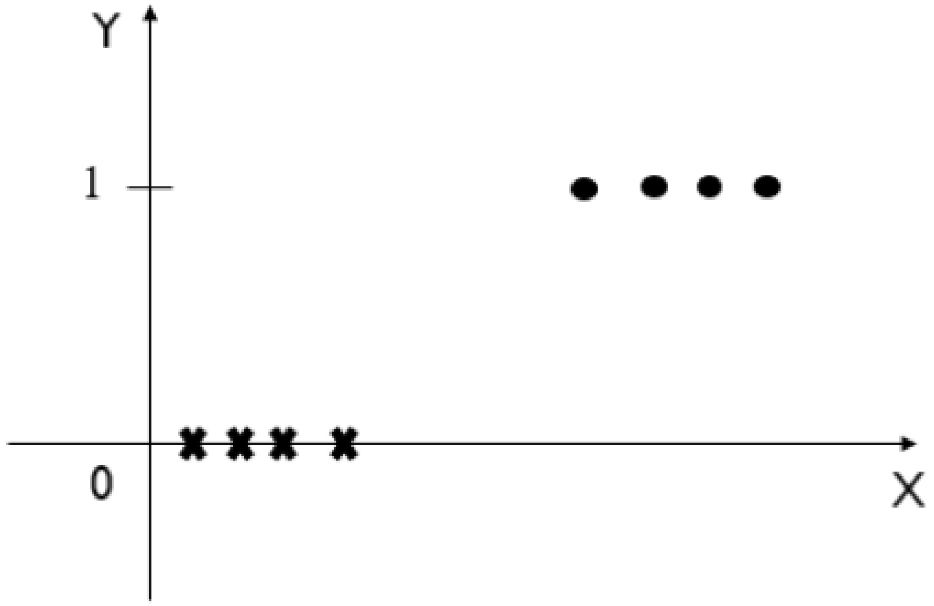
2025-2026 Semester 1

BINARY CLASSIFICATION

- We have studied Linear Regression and polynomial Regression to predict the continuous values (house prices).
- But in the case of discontinuous data such as the weather: (sunny, rainy, cloudy) this dataset has specific and countable values, so we use the classification and Logistic Regression.
- The simplest type of the classification is the binary classification (Two classes).
- Foe example:
- we have a dataset that classifies whether the patient has cancer or not according to the size of the tumor as follows:

BINARY CLASSIFICATION

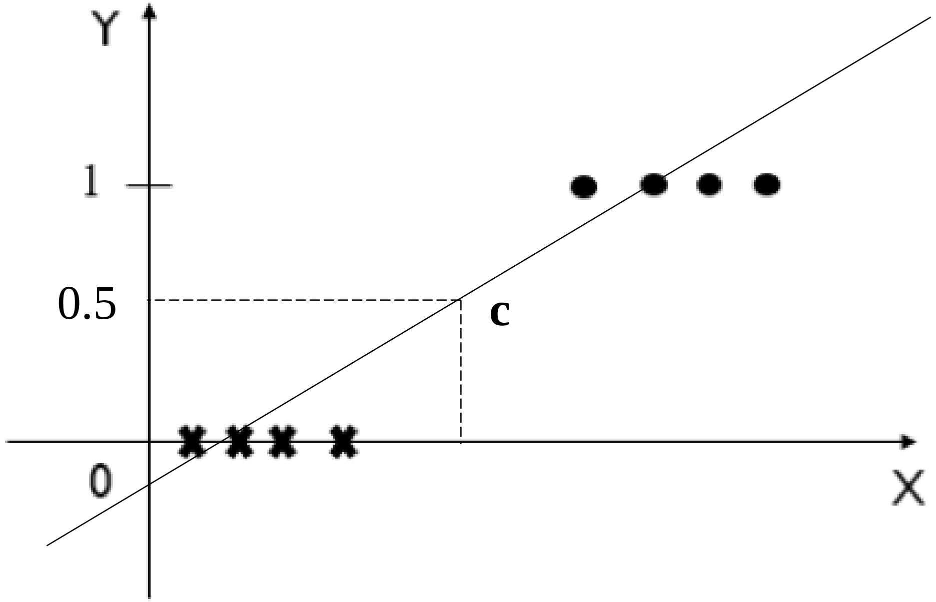
- Where value 1 means that patient is infected and value 0 means that he is healthy.
- The data can be plotted as follows:



Tumor	Cancer
140	1
190	1
10	0
5	0
4	0
160	1
6	0
200	1

BINARY CLASSIFICATION

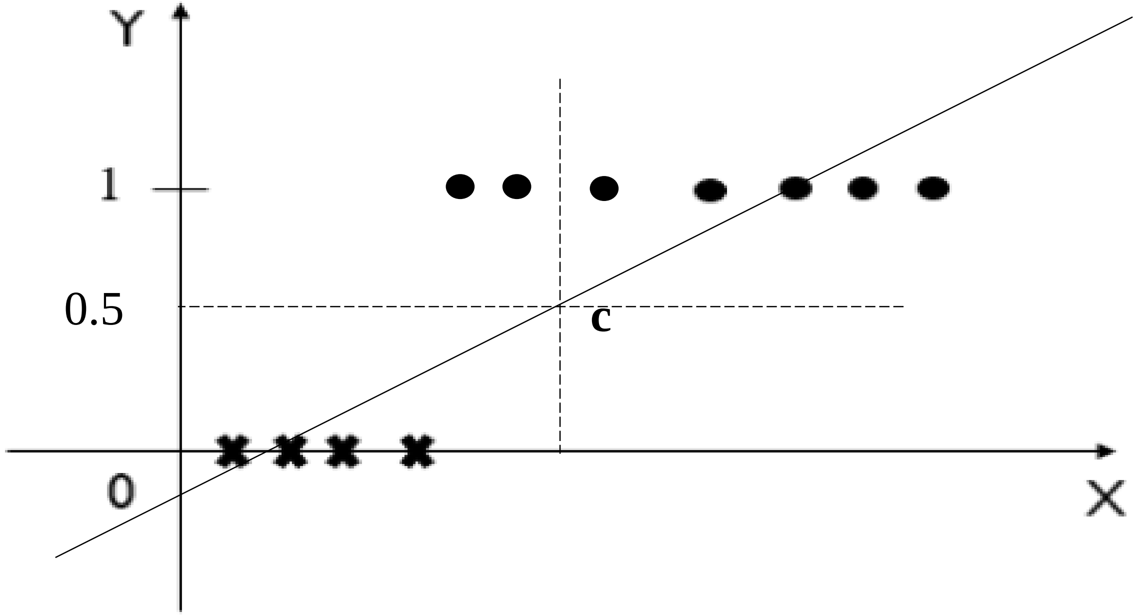
- We will try to use Linear Regression to predict new data .
- The form of the previous function was as follows:
- ***h= w x + b***
- we plot the line of this function.



- We will choose a certain threshold, c = 0.5 . When we test a certain point, we find its output, if it is greater than 0.5 the patient is infected, but if it is less than 0.5 the patient does not have cancer.

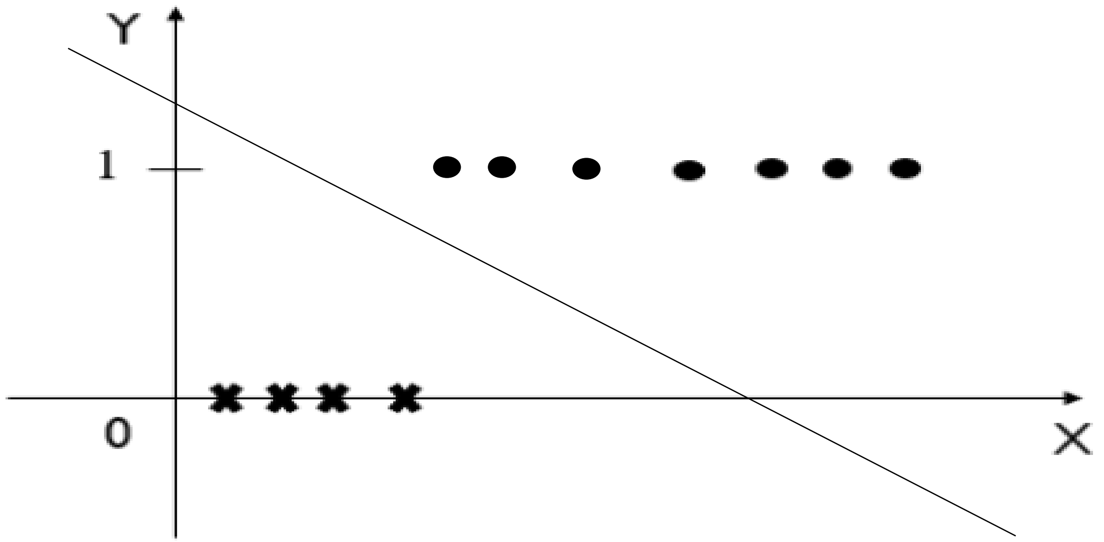
BINARY CLASSIFICATION

- In the following figure, we notice that there are two samples outside the threshold, therefore we have to change the threshold or change the position of the straight line.



BINARY CLASSIFICATION

- We can change the position of the line as follows:



- We notice that the line separated the samples into two classes. Samples above the line belong to one class and the samples below the line belong to another class.

BINARY CLASSIFICATION

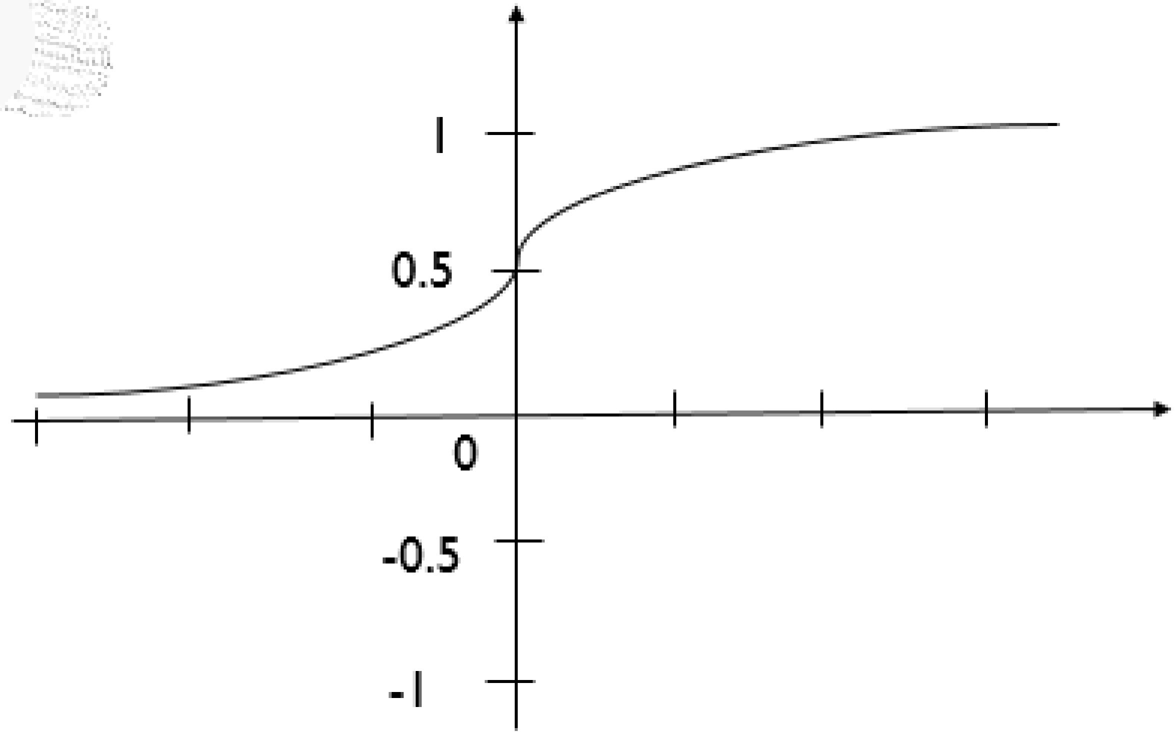
- A point can be determined while it was above or below a line by substituting the point in the equation of the line.
- If the result is negative, the point is under the line, so Y=0 , but if it is positive, the point is above the line and Y =1.
- If the result is zero the point is on the line, then we can consider Y=1.
- We call this line **Decision Boundary,** because it separates the space into classes (above and below).

LOGISTIC REGRESSION

- The problem of regression with binary classification is that function values are real and much greater than 1 or much smaller than 0 but the values in binary classification problem are 0 and 1.
- Therefore we change the function

form to be:

- ***h= g(wx +b) where 0<h<1***
- g: is sigmoid function
- ***g(z)= 1/(1+e^-z^)*** $g(z) = 1/(1+e^{-z})$
- we can plot this function as follows:



LOGISTIC REGRESSION

- We notice that the function values range from 0 to 1:
- *if* $z = -\infty$:
- $g(-\infty)= 1/(1+e^{\infty}) = 1/(1+ \infty)= 1/ \infty=0$
- *if* $z = 0$:
- $g(0)= 1/(1+e^{0}) = 1/(1+ 1)= 1/ 2=0.5$
- *if* $z = +\infty$:
- $g(+\infty)= 1/(1+e^{-\infty}) = 1/(1+ 0)= 1/ 1=1$

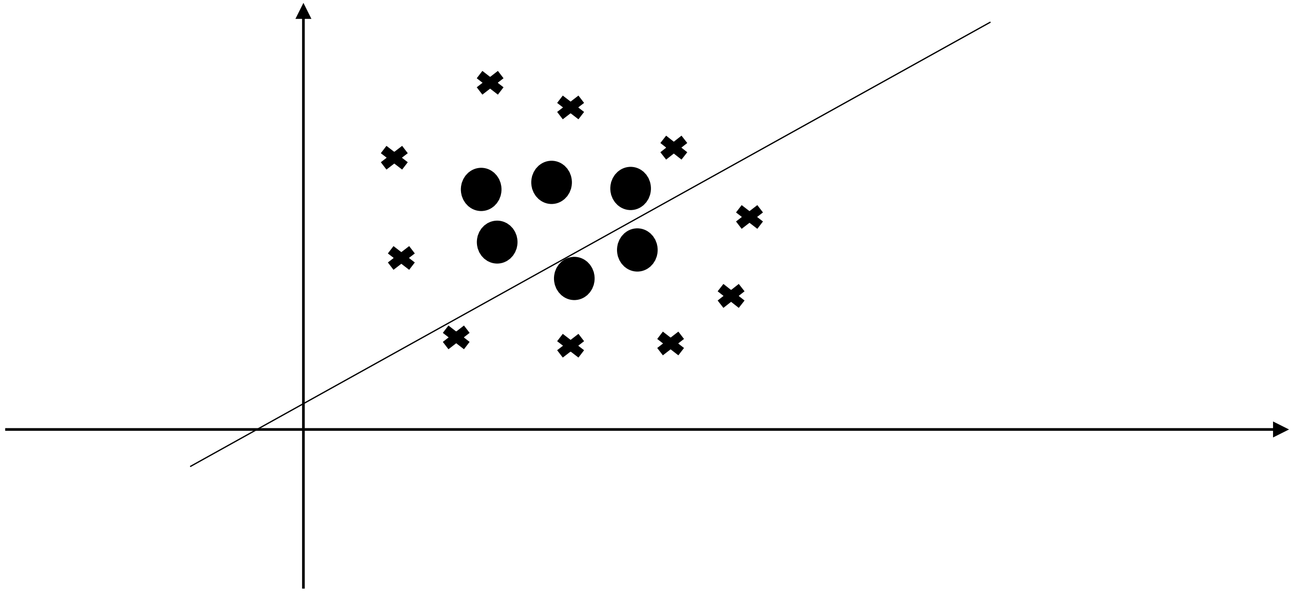
LOGISTIC REGRESSION

- The sigmoid function is a logistic function
- the model of this function ***h= g(wx +b)*** is more suitable for the binary classification problem:

- *if* $z=wx+b \geq 0$ *then* $h \geq 0.5$ *so* $Y=1$
- *if* $z=wx+b \leq 0$ *then* $h \leq 0.5$ *so* $Y=0$

LOGISTIC REGRESSION

- Suppose we have the following figure



- Where the circular samples represent a class, and the other samples represent another class. Therefore, no line can’t completely separate the two classes, so we have to look for a polynomial model that the samples inside the curve are one class and ones outside the curve another class.

LOGISTIC REGRESSION- COST FUNCTION

- We have learned that the Target values Y are continuous in Linear Regression algorithm and discontinuous in Logistic algorithm.
- The cost function for Linear regression is:
- $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_i (h_\theta(x_i) - y_i)^2$

- We have used this cost function in the linear gradient descent algorithm, but this function can’t be used in the logistic algorithm, because when the sigmoid function is squared, we can’t get convex function, so it is possible we can’t reach to a local minimum.

LOGISTIC REGRESSION- COST FUNCTION

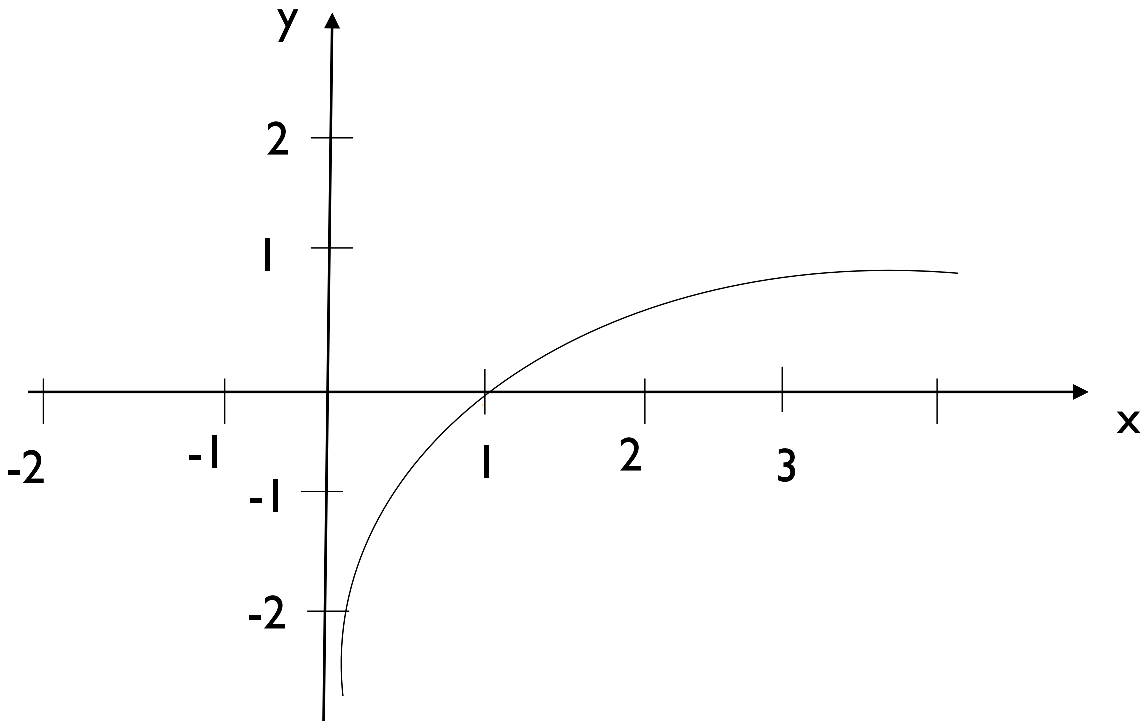
- To solve this problem we select the following function as cost function for Logistic Regression:
- $cost = \begin{cases} -y\, ln(h) & if\ y = 1 \\ -(1-y)\, ln(1-h) & if\ y = 0 \end{cases}$
- let's test this function:
- *y=1, h=1: cost (y=1, h=1) = -1*ln(1) = -1*0=0*
- *y=1, h=0: cost (y=1, h=0) = -1*ln(0) = -1*(-∞)=* + ∞
- *y=0, h=0: cost (y=0, h=0) = -(1-0)*ln(1-0) = -1*0=0*
- *y=0, h=1: cost (y=0, h=1) = -(1-0)*ln(1-1) = -1*(-∞)=* + ∞

LOGISTIC REGRESSION- COST FUNCTION

- This means that if y and $h(x)$ values are close, the cost is small, but if y and h(x) values are disparate, the cost will be large.
- We'll sum the function in one line:
- $cost = -y\,ln(h) - (1-y)\,ln(1-h)$
- for all samples in the dataset the cost function is:
- $cost = -\frac{1}{m}\sum_i (y_i * ln(h(xi)) + (1-yi) * ln(1-h(xi)))$

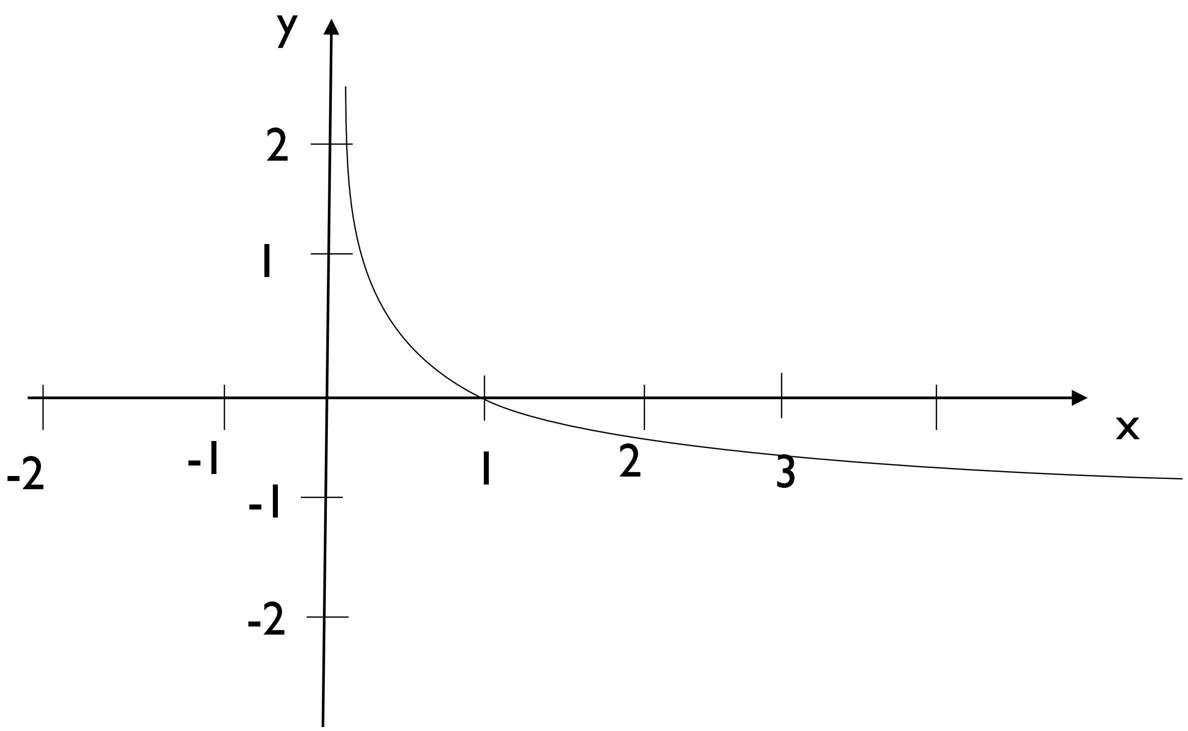
LOGISTIC REGRESSION- COST FUNCTION

- Let's see if this function is convex.
- The graph of the function *y=ln(x)* is:



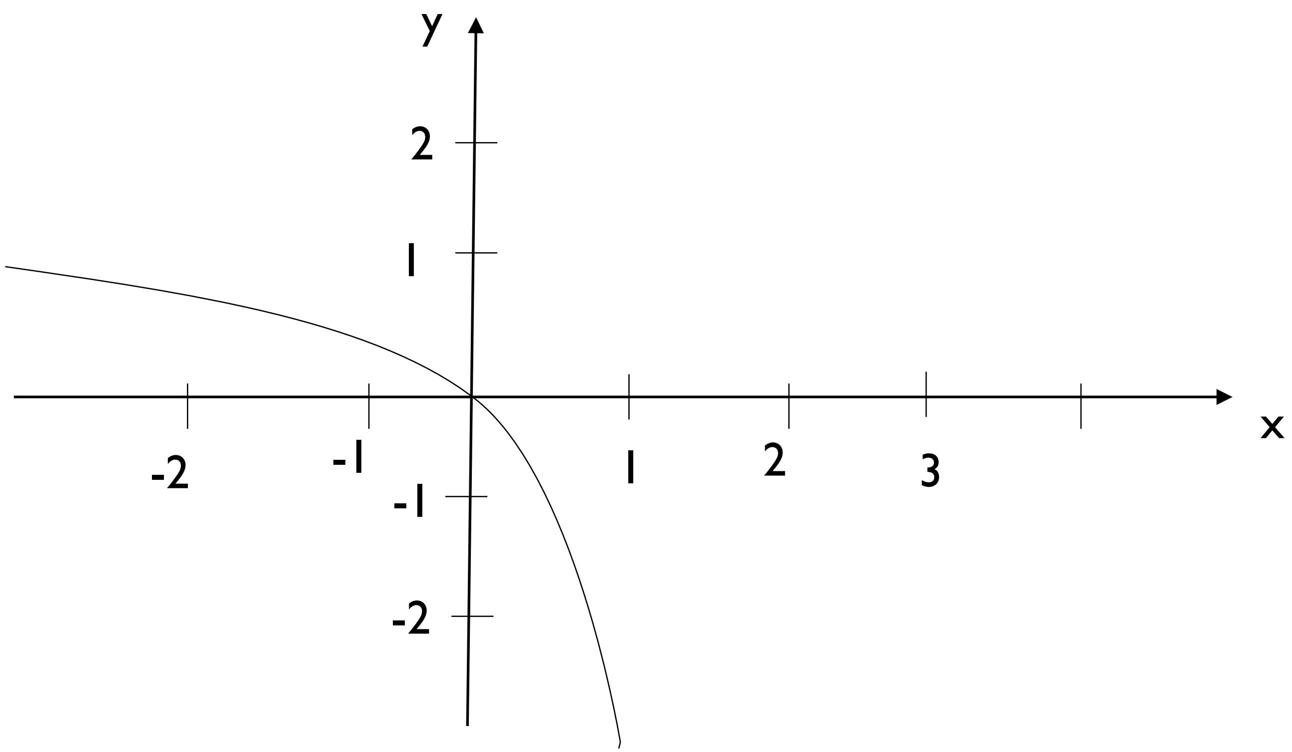
LOGISTIC REGRESSION- COST FUNCTION

- The graph of the function $y = -ln(x)$ is:



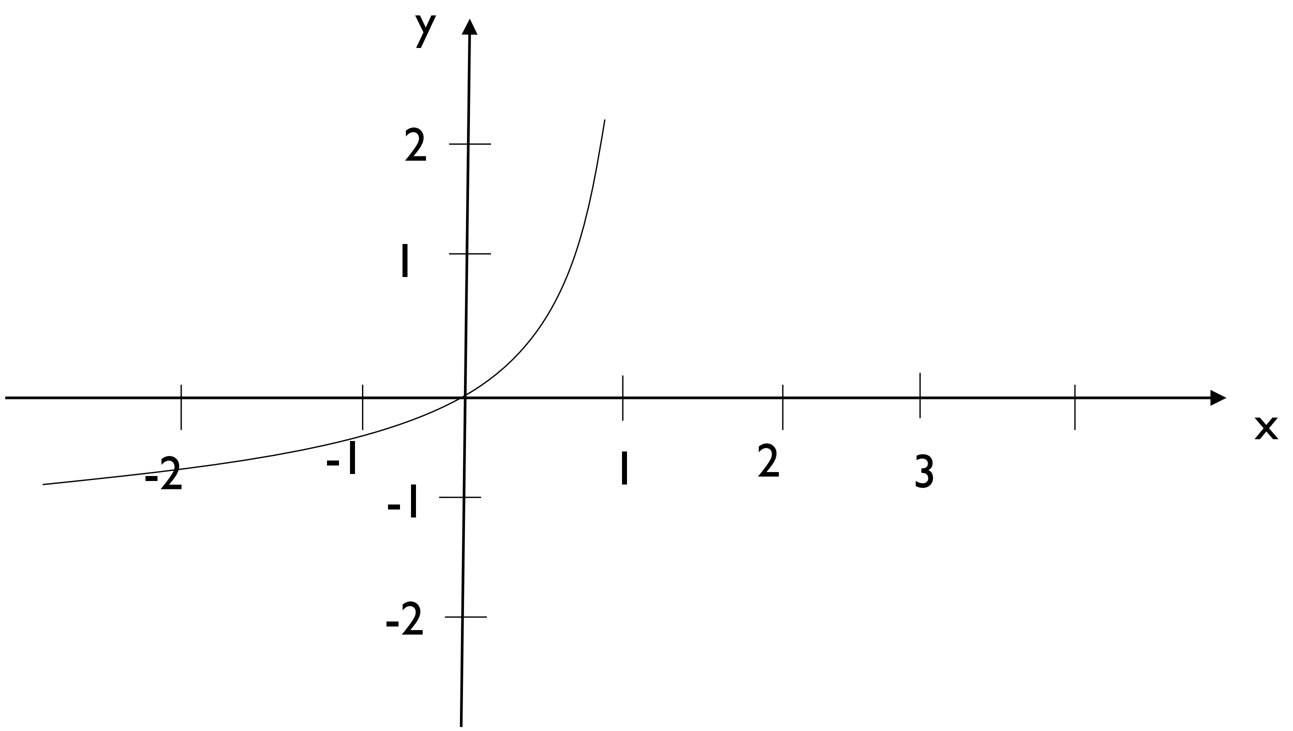
LOGISTIC REGRESSION- COST FUNCTION

- The graph of the function *y=ln(1-x)* is:



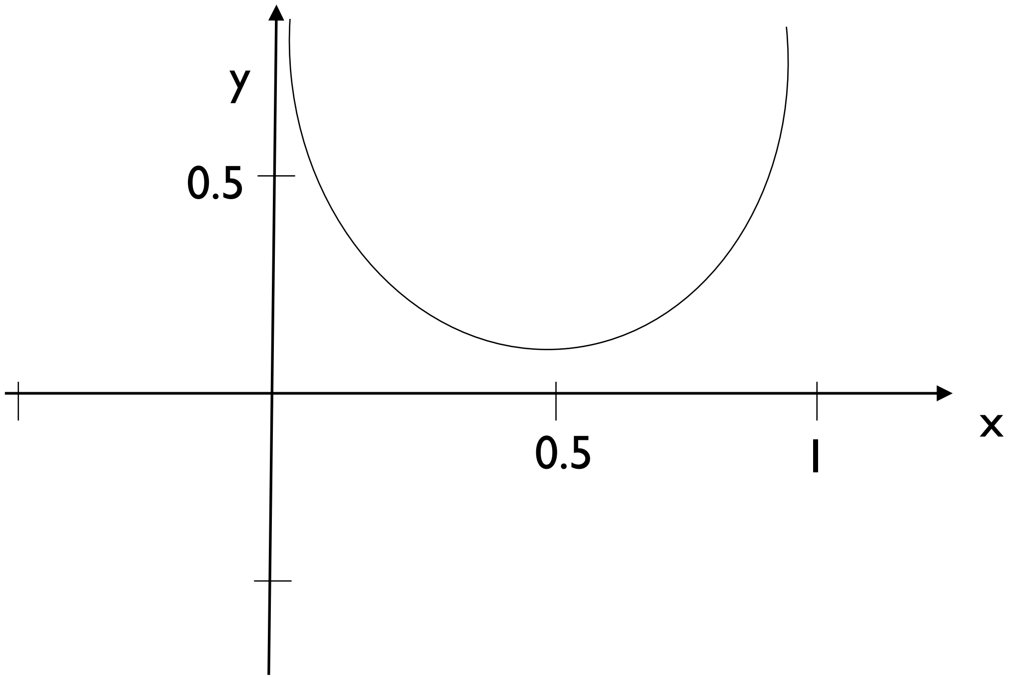
LOGISTIC REGRESSION- COST FUNCTION

- The graph of the function $y = -\ln(1-x)$ is:



LOGISTIC REGRESSION- COST FUNCTION

- The graph of the function $y= -ln(1-x) - ln(x)$ is:



- we notice that we have obtained a convex parabola function in the domain]0,1[.
- Therefore we can apply Gradient Decent on this function.

LOGISTIC REGRESSION- COST FUNCTION

- Gradient Descent algorithm for Logistic Regression:

Initialization

Repeat {

$w_j = w_j - \alpha \frac{\partial J}{\partial w}$

$b = b - \alpha \frac{\partial J}{\partial b}$

}

until converge

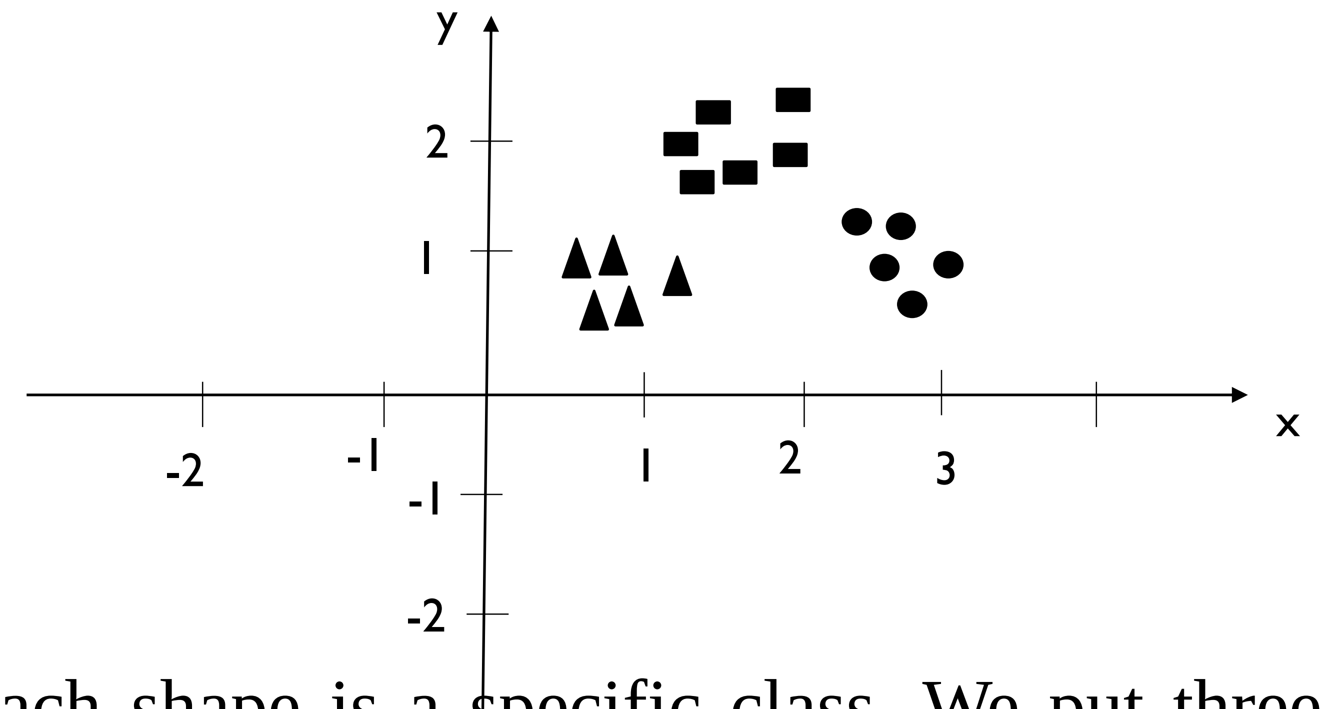
where: $J(w, b) = -\frac{1}{m} \sum_i (y_i * ln(h(xi)) + (1 - yi) * ln(1 - h(xi)))$

$h(x) = g(wx + b) = 1/(1 + e^{-(wx + b)})$

$$\frac{\partial g(z)}{\partial z} = g(z) * (1-g(z))$$

MULTICLASS CLASSIFICATION

- Let us have the data set shown in the following graph:



- Where each shape is a specific class. We put three models $h_1(x)$, $h_2(x)$, $h_3(x)$, each model for a specific class. So that the label of the triangle is 1 and the other label of the rest of the classes is 0 . For the model of the circle, the label is 1, and the rest of the classes is 0. for model square model, the label is 1 and rest of the classes is 0. Each model will be trained as binary classifier.

MULTICLASS CLASSIFICATION

- When a new point is tested for which class it belongs, each model gives a certain probability within the range [0,1]. The answer that has the greatest probability is the final answer.
- for example, if the probabilities are:
- *h1=0.3*, *h2= 0.8*, *h3= 0.4*
- the point is circle
- This classification is called ***one- versus – all*** classification **(multiclass)**