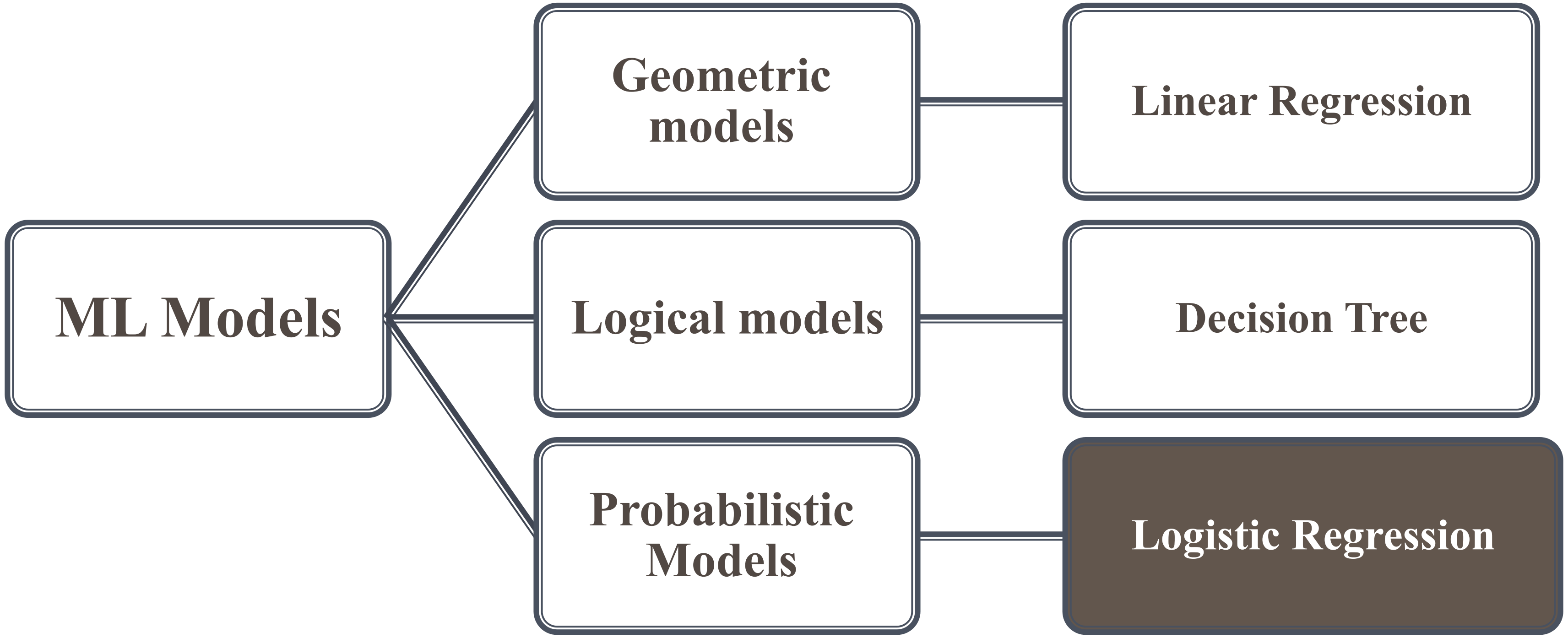
# Machine learning

Presented by : Dr. Hanaa Bayomi

Lecture 4 : Logistic Regression

# Flach talks about three types of Machine Learning models [Fla12]



# CLASSIFICATION

The classification problem is just like the regression problem, except that the values **y** we now want to predict take on only a **small number of discrete values**.

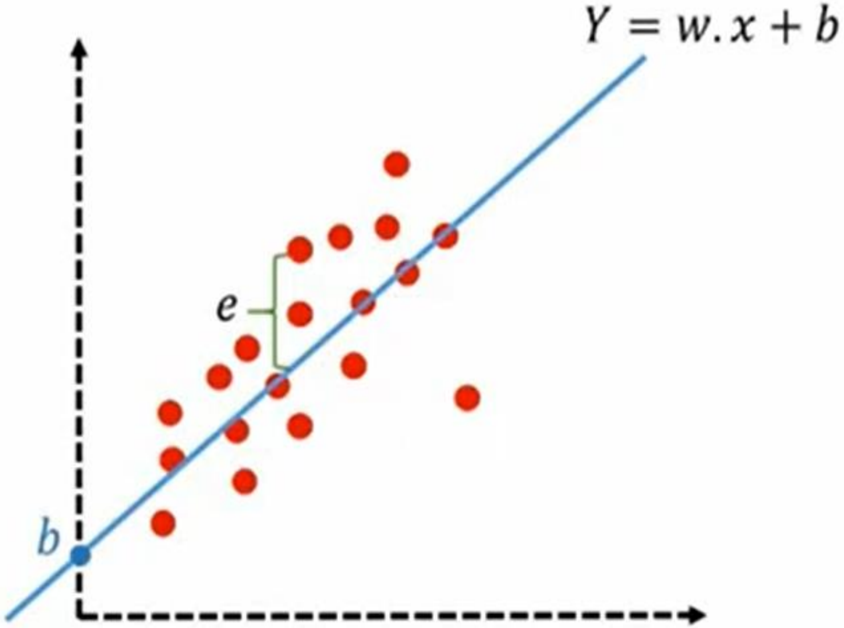
Some Example of Classification problem

- Email : Spam / Not spam
- Tumor: Malignant/ Benign

$y \in \{0, 1\}$

0: "Negative Class" (e.g., benign tumor)
1: "Positive Class" (e.g., malignant tumor)

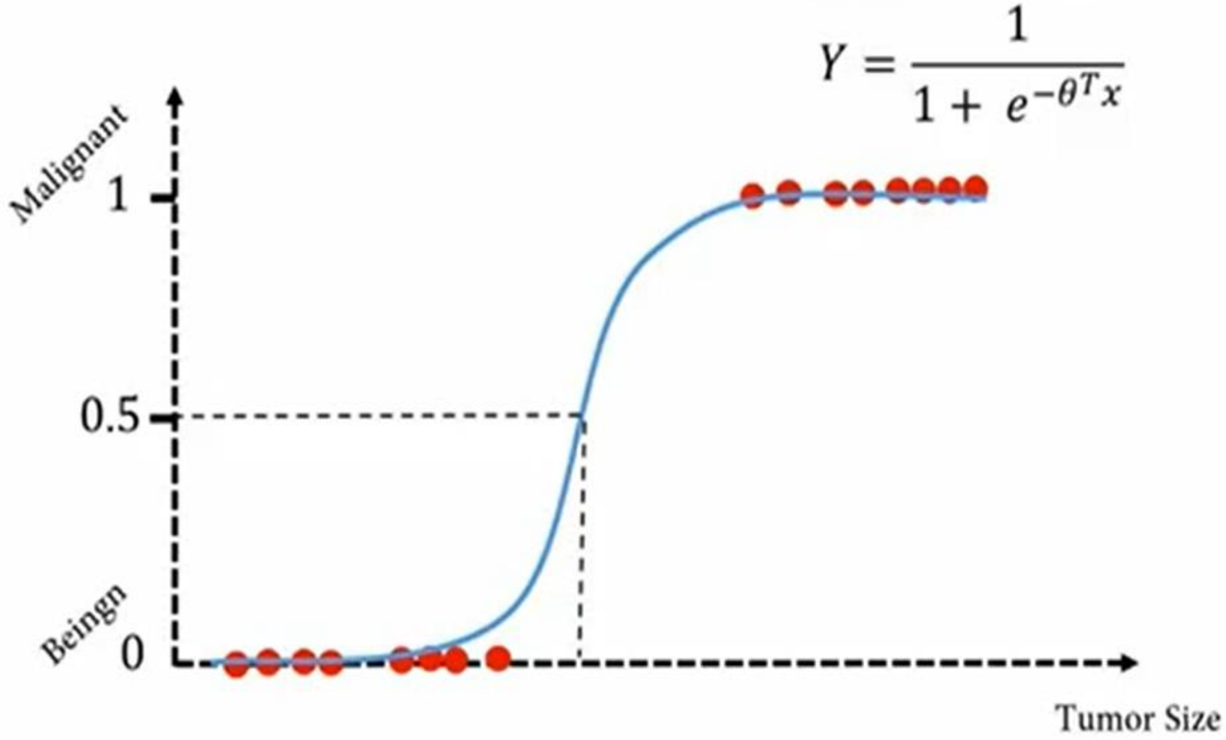
# Logistic Regression




Linear regression

Regression Probelm: Continous

- Stock prices




Logistic regression

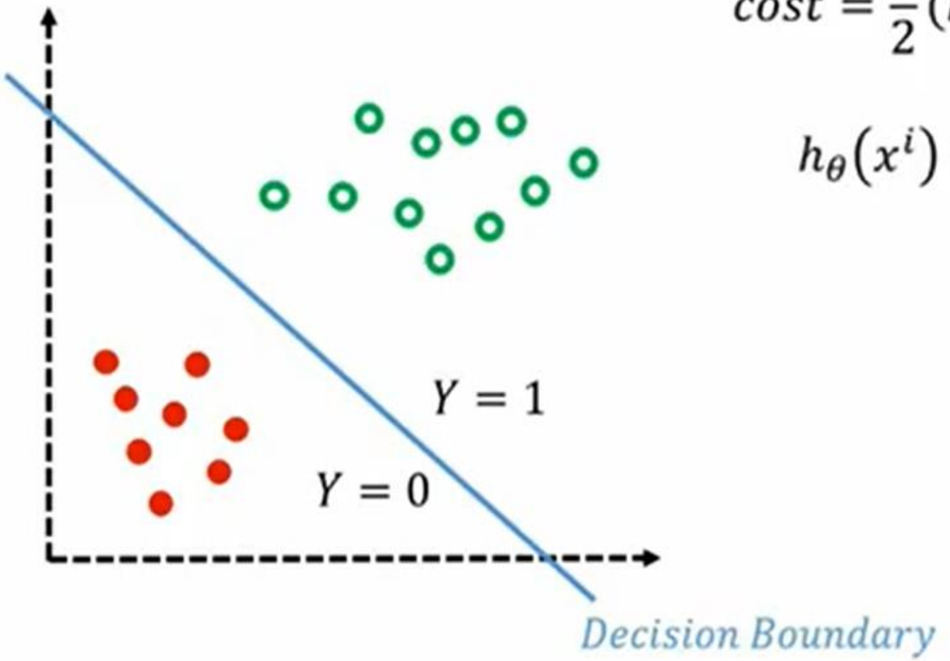
Classification Probelm: Discrete

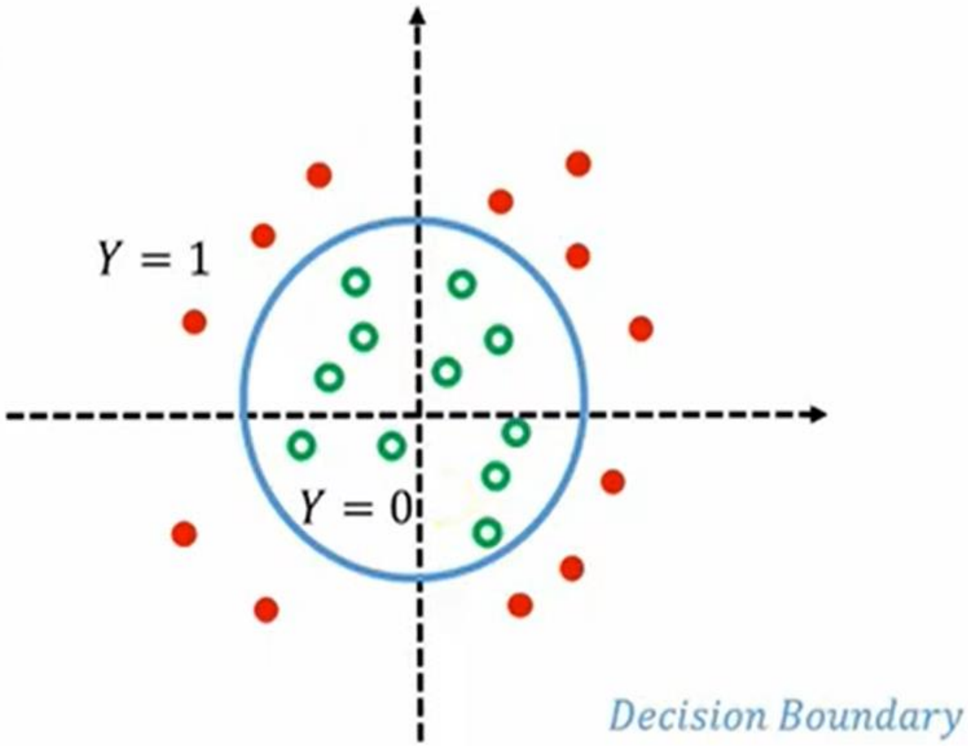
- Malignant or benign tumor

# Logistic Regression

$$cost = \frac{1}{2}(h_\theta(x^i) - y^i)^2$$

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




$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

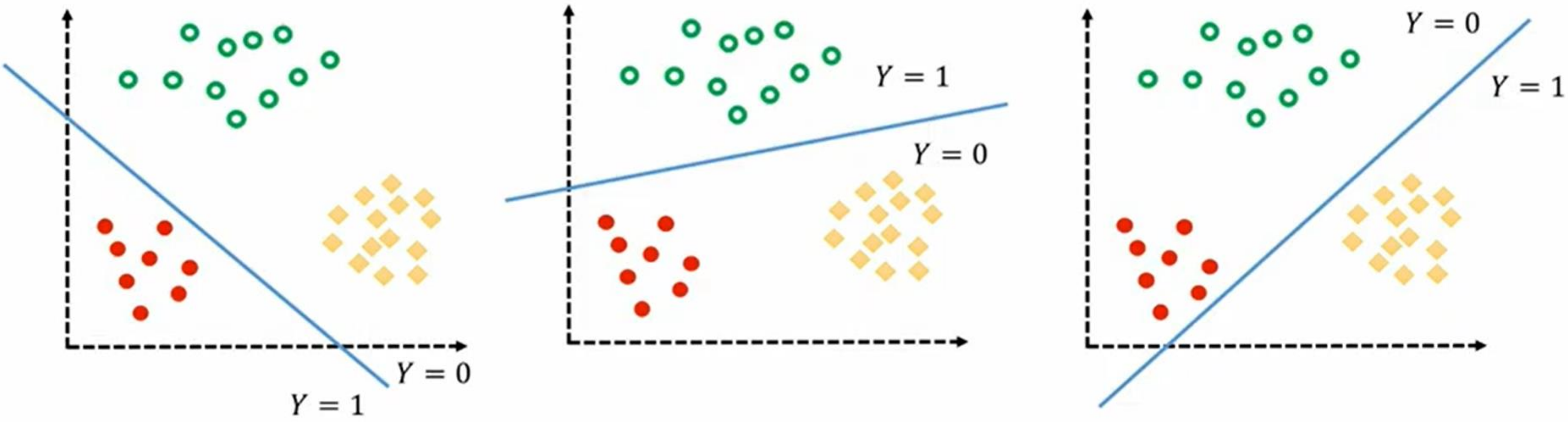
$$h_\theta(x) = -3 + x_1 + x_2$$

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

$$h_\theta(x) = -1 + x_1^2 + x_2^2$$

# Multiclass Classification

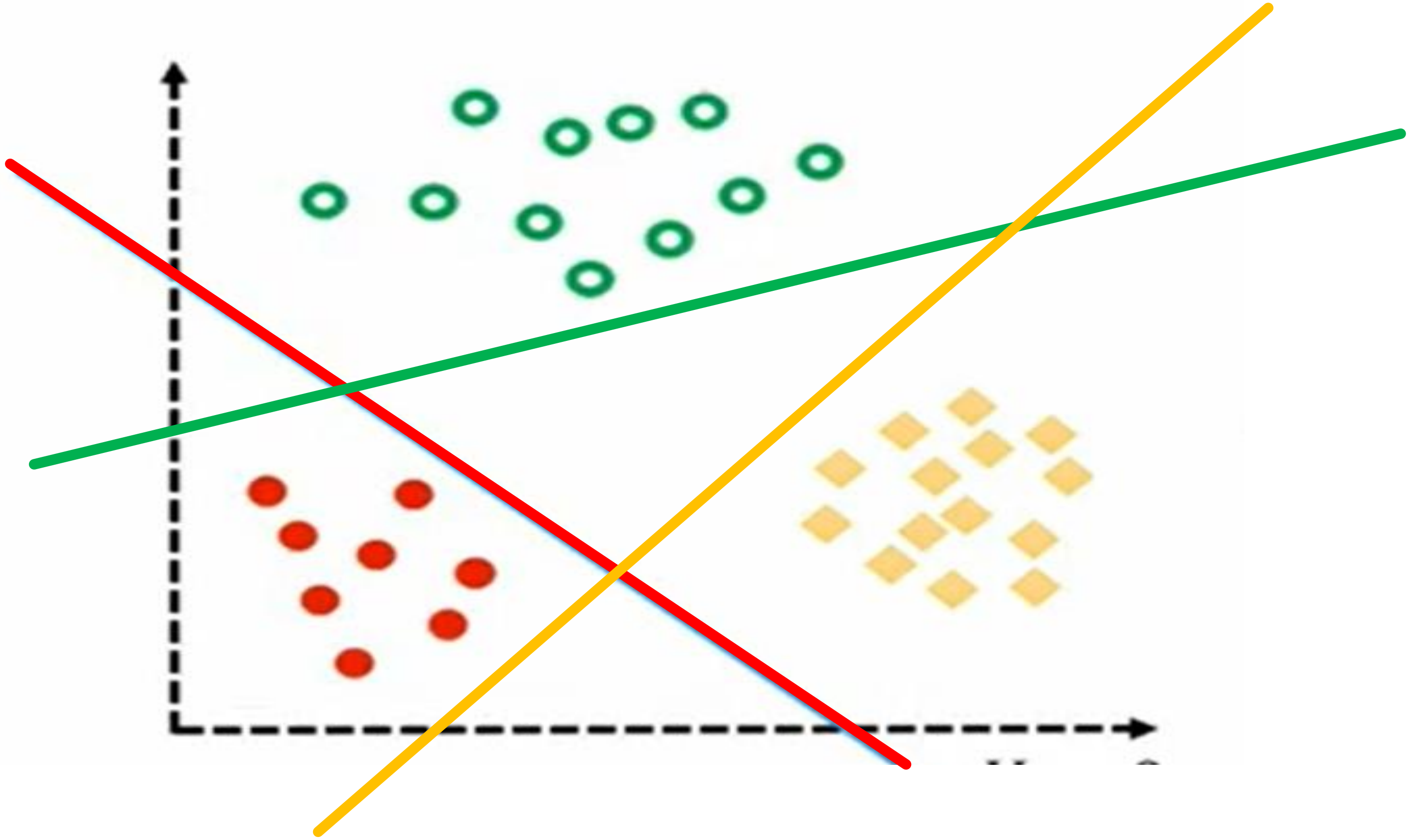
**One-vs-all**

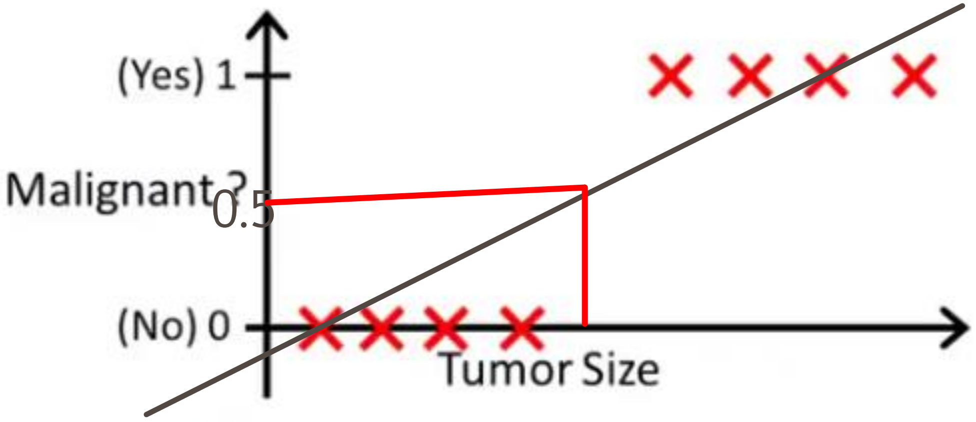



$h_\theta^1(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots$ $h_\theta^2(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots$ $h_\theta^3(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots$

# Multiclass Classification

**One-vs-all**






Threshold classifier output $h_\theta(x)$ at 0.5:

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

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

# DEFINITION

- Binary Logistic Regression

  - We have a set of feature vectors X with corresponding binary outputs

$$X = \{x_1, x_2, \ldots, x_n\}^T$$

$$Y = \{y_1, y_2, \ldots, y_n\}^T, where \quad y_i \in \{0,1\}$$

  - We want to model p(y|x)

$$p(y_i = 1 \mid x_i, \theta) = \sum_j \theta_j x_{ij} = x_i \theta$$

By definition $p(y_i = 1 \mid x_i, \theta) \in \{0,1\}$ . We want to transform the probability to remove the range restrictions, as $x_i\theta$ can take any real value.

# USING ODDS

- Odds

  p : probability of an event occurring

  1 – p : probability of the event not occurring

  The odds for event i are then defined as

$$odds_i = \frac{p_i}{1 - p_i}$$

Taking the ***log*** of the odds removes the range restrictions.

$$\log\left(\frac{p_i}{1 - p_i}\right) = \sum_j \theta_j x_{ij} = x_i \theta$$

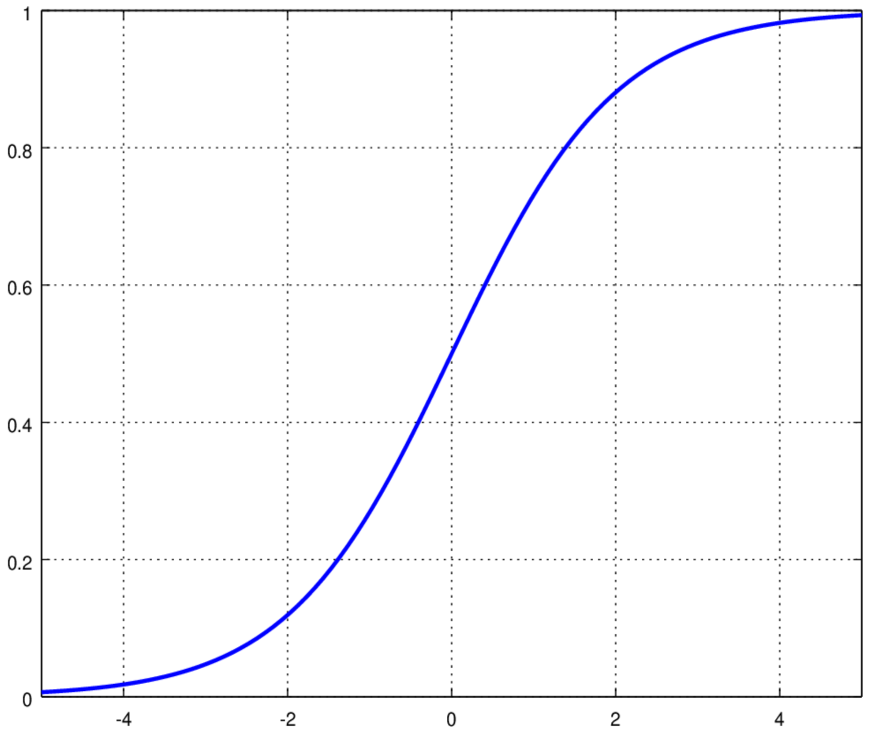
This way we map the probabilities from the [0; 1] range to the entire number line (real value).

# HYPOTHESIS FUNCTION

$$\log\left(\frac{p_i}{1-p_i}\right) = x_i\theta$$

$$\frac{p_i}{1-p_i} = e^{x_i\theta}$$

$$p_i = \frac{e^{x_i\theta}}{1+e^{x_i\theta}} = \frac{1}{1+e^{-x_i\theta}}$$




Standard logistic sigmoid function

$$h_\theta(x) = \theta^T x$$

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

$$p_i = g(\theta^t x) = \frac{1}{1+e^{-\theta^t x}}$$

# LOGISTIC REGRESSION MODEL

Linear Regression

$$h_\theta(x) = \theta^t x$$

Logistic Regression

$$g(\theta^t x) = \begin{cases} 1, & \dfrac{1}{1+e^{-\theta x}} \geq 0.5 \\ 0, & 1-\dfrac{1}{1+e^{-\theta x}} < 0.5 \end{cases}$$

$$p(y_i = 1 \mid x_i, \theta) = \frac{1}{1+e^{-\theta^t x}}$$

$$p(y_i = 0 \mid x_i, \theta) = 1 - \frac{1}{1+e^{-\theta^t x}}$$

$$p(y_i \mid x_i : \theta) = \left(\frac{1}{1+e^{-\theta^t x}}\right)^{y_i} \left(1 - \frac{1}{1+e^{-\theta^t x}}\right)^{1-y_i}$$

**Interpretation of Hypothesis Output**

$h_\theta(x)$ = estimated probability that y = 1 on input x

Example: If $x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{tumorSize} \end{bmatrix}$

$$h_\theta(x) = 0.7$$

Tell patient that 70% chance of tumor being malignant

$$h_\theta(x) = p(y = 1 \mid x; \theta)$$

"probability that y = 1, given x, parameterized by $\theta$"

$$P(y=0|x;\theta) + P(y=1|x;\theta) = 1$$
$$P(y=0|x;\theta) = 1 - P(y=1|x;\theta)$$

# ESTIMATION OF COEFFICIENTS Θ

- Maximum Likelihood Estimation (MLE)

1. Step 1 : get the probability for all observations

$$p(y \mid X:\theta)=\prod_{i=1}^{m}\left(\frac{1}{1+e^{-\theta^t x}}\right)^{y_i}\left(1-\frac{1}{1+e^{-\theta^t x}}\right)^{1-y_i}$$

2. Step 2 : Express this is a function of θ, where X and y are fixed parameters $L(\theta)=p(y \mid X:\theta)$

3. Step 3 : Maximize $L(\theta)$ likelihood function

$$l(\theta) = \prod_{i=1}^{m} \left( \frac{1}{1+e^{-\theta^t x}} \right)^{y_i} \left( 1 - \frac{1}{1+e^{-\theta^t x}} \right)^{1-y_i}$$

We can simplify L(θ) by taking its ***log*** and then differentiate to get the gradient.

$$j(\theta) = l(\theta) = \sum_{1}^{m} \left[ y_i \log\left( \frac{1}{1+e^{-\theta^t x}} \right) + (1-y_i)\log\left( 1 - \frac{1}{1+e^{-\theta^t x}} \right) \right]$$

$$\frac{d}{d\theta} j(\theta) = \frac{d}{d\theta} \sum_{1}^{m} \left[ y_i \log\left( \frac{1}{1+e^{-\theta^t x}} \right) + (1-y_i)\log\left( 1 - \frac{1}{1+e^{-\theta^t x}} \right) \right]$$

$$= \sum_{\mathrm{i=1}}^{\mathrm{m}} \left( \mathrm{y_i} - \frac{1}{1+e^{-\theta^t x_i}} \right) x_i$$

# GRADIENT DESCENT

## ▪ in Linear Regression

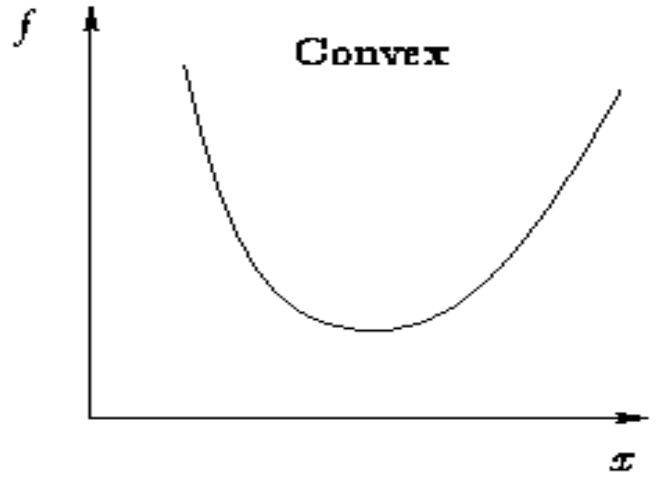
Want $\min_\theta J(\theta)$:

Repeat $\{$

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

(simultaneously update all $\theta_j$)

$\}$




## ▪ in Logistic Regression

We can now use **<u>gradient ascent</u>** to maximize j(θ) The update rule will be:
repeat until convergence

$\{$

$$\theta_j = \theta_j + \alpha \sum_{i=1}^{m} \left( y_i - \frac{1}{1+e^{-\theta^t x_i}} \right) x_{ij}$$

$\}$