
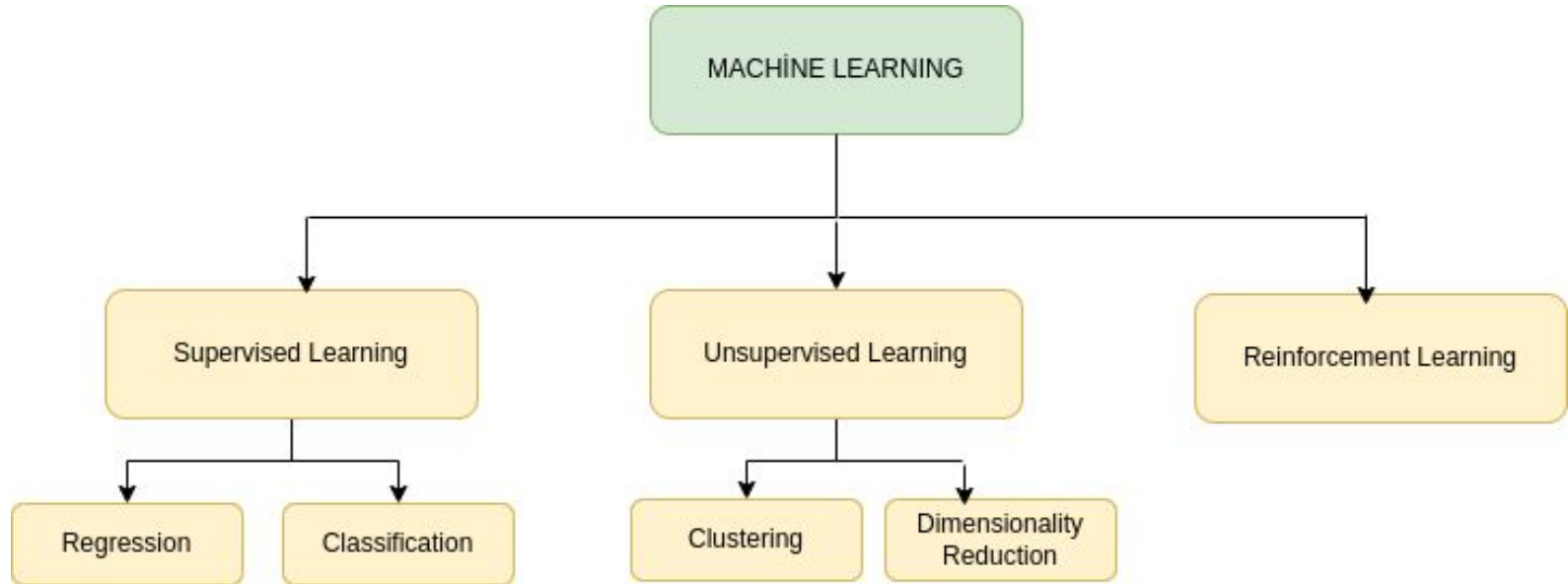


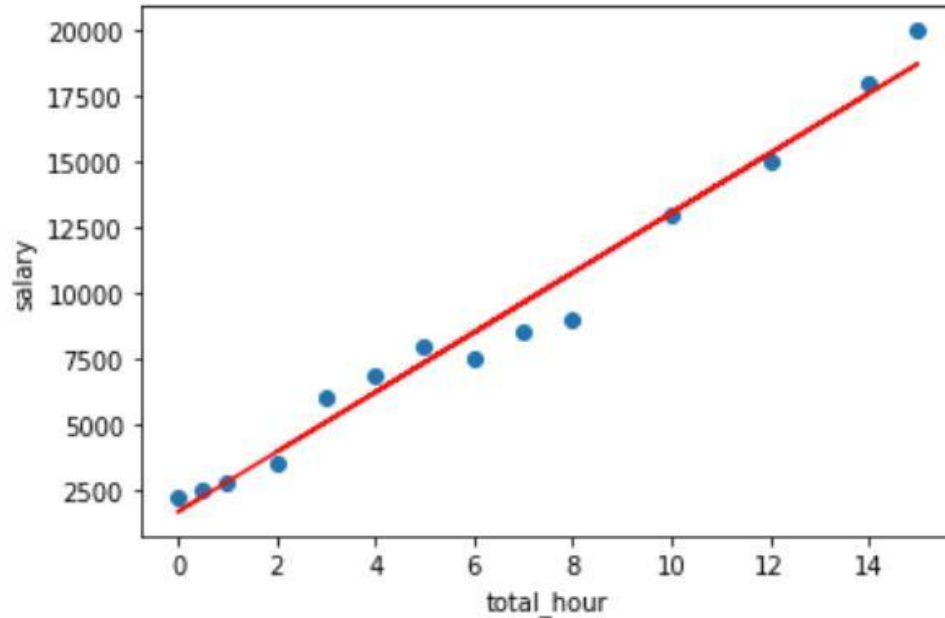
# Machine Learning



# Machine Learning



# Linear Regression



$$h_{\theta}(x^{(i)}) = \theta_0 x_0 + \theta_1 x_1 + \dots + \theta_j x_j$$

$x_0 = 1$ ,  $j = \text{the number of features}$

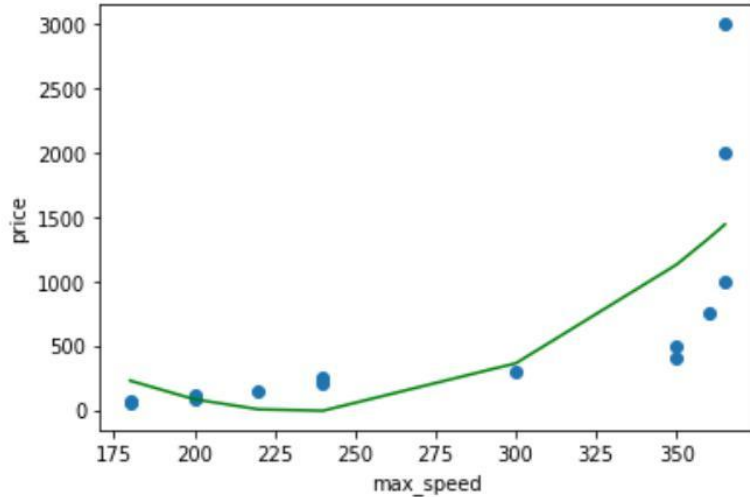
$$\text{Least Squared Error} = \frac{1}{2} (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

$$\text{Cost Function} = J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2, \quad m = \text{number of sample data}$$

Derivative of cost function at  $\theta_j$ :

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

# Polynomial Regression



Simple  
Linear  
Regression

$$y = b_0 + b_1x_1$$

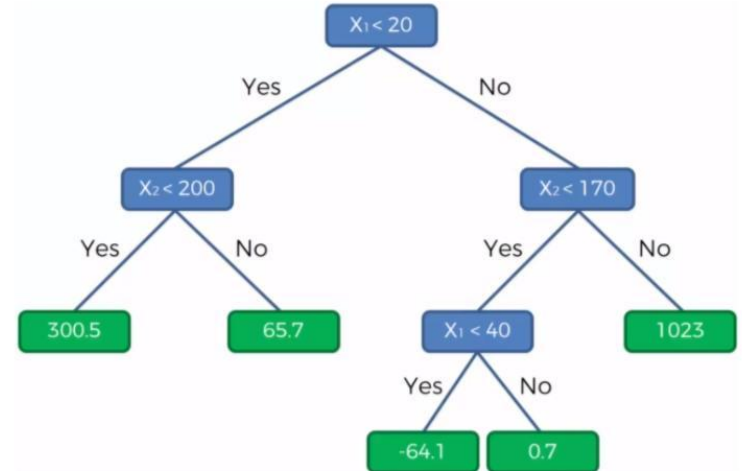
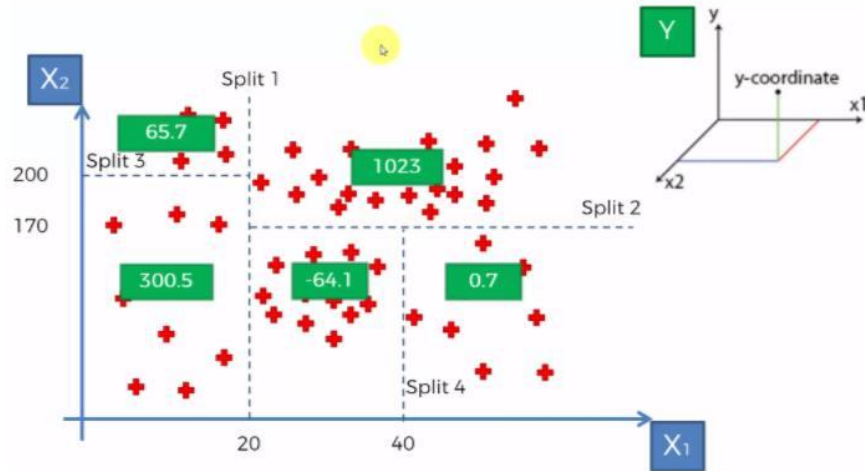
Multiple  
Linear  
Regression

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$$

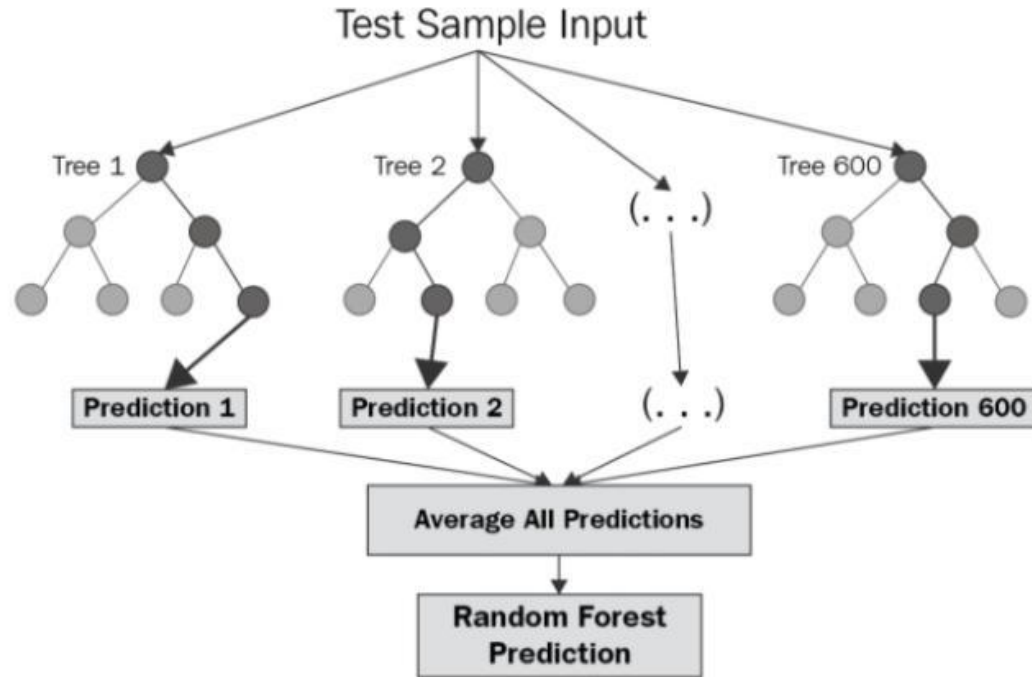
Polynomial  
Linear  
Regression

$$y = b_0 + b_1x_1 + b_2x_1^2 + \dots + b_nx_1^n$$

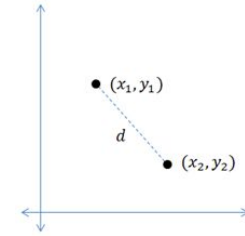
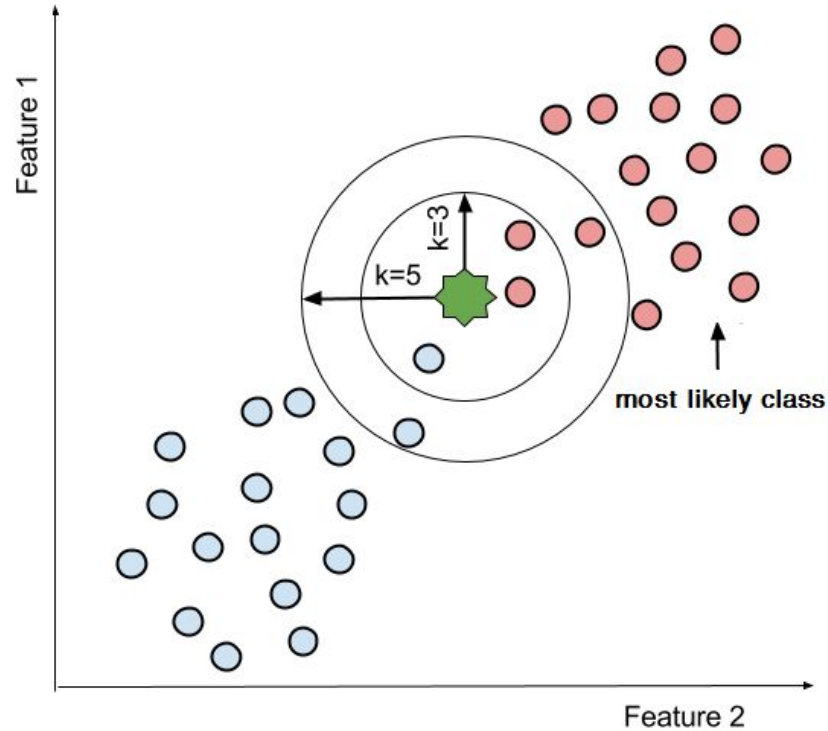
# Decision Tree Regression



# Random Forest Regression

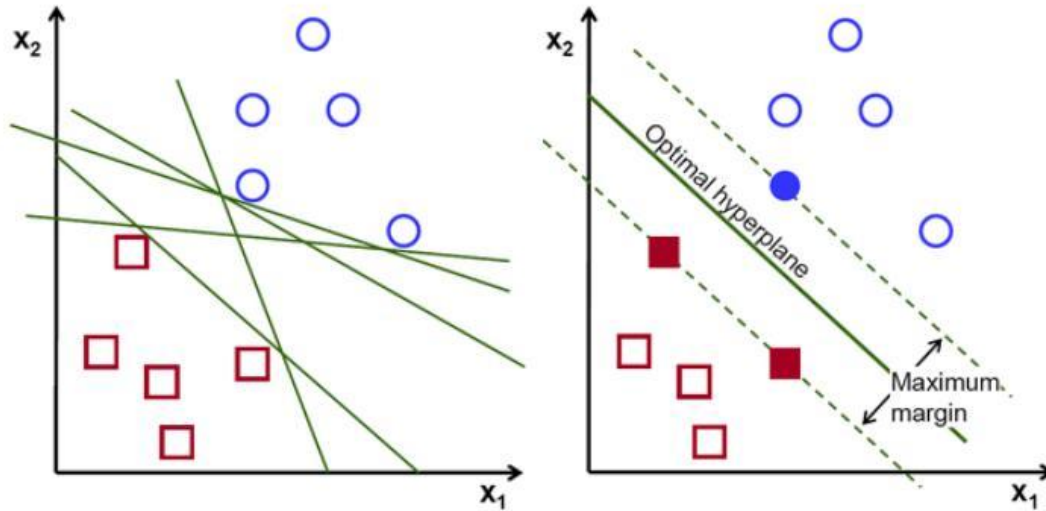


# KNN Classification



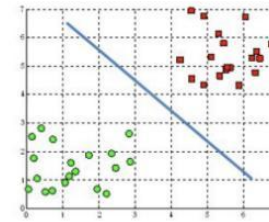
$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

# SVM

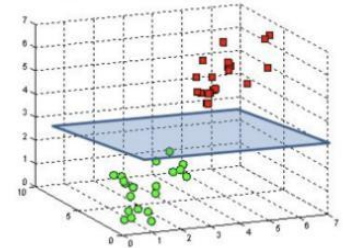


Possible hyperplanes

A hyperplane in  $\mathbb{R}^2$  is a line



A hyperplane in  $\mathbb{R}^3$  is a plane

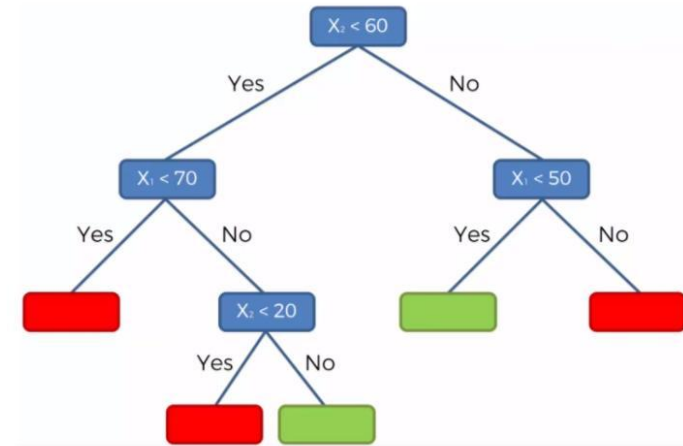
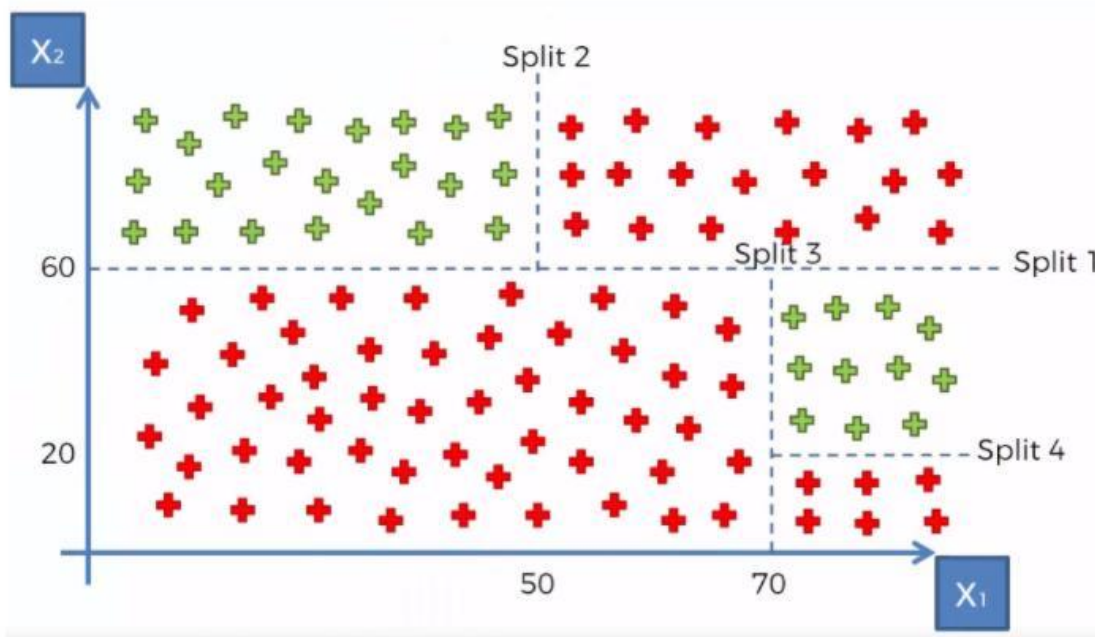


To separate the two classes of data points, there are many possible hyperplanes that could be chosen.

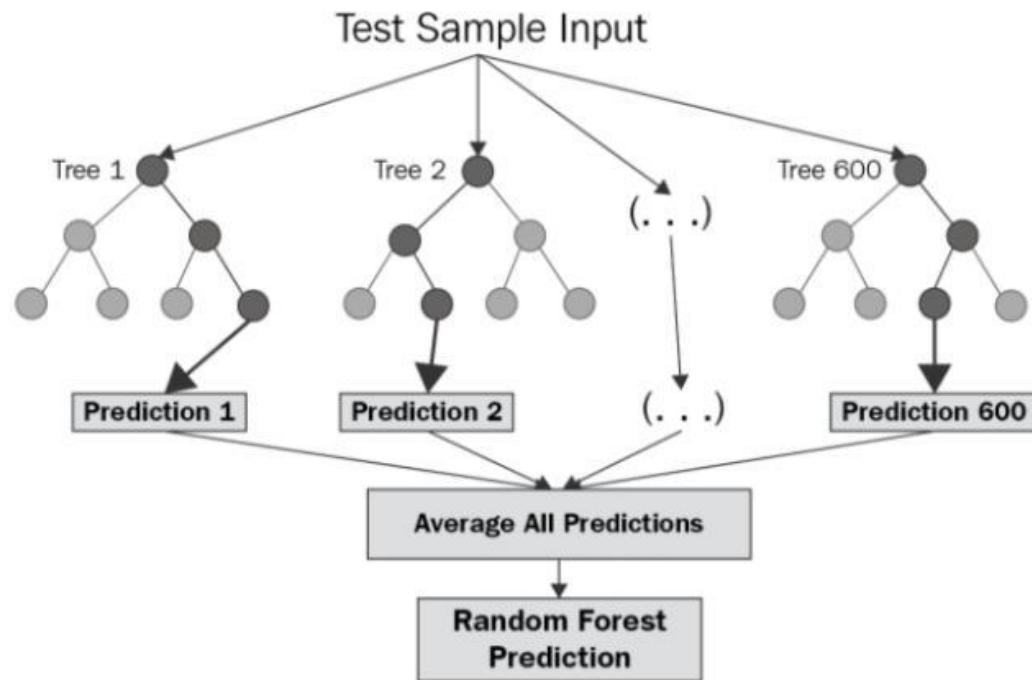
Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes



# Decision Tree Classification



# Random Forest Classification



# Unsupervised Learning

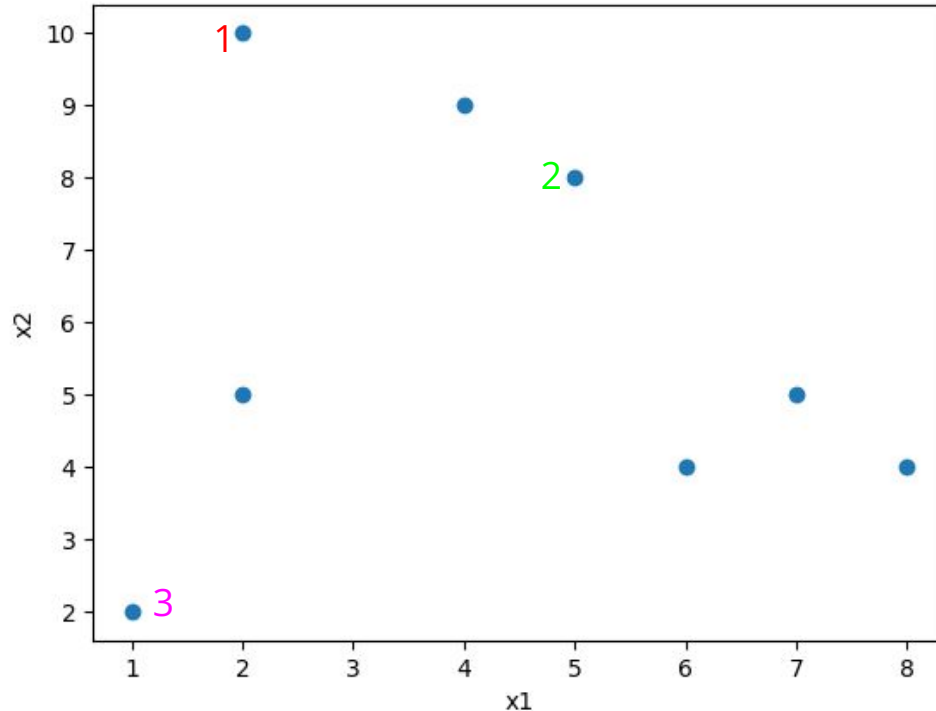
K-means Clustering

Knn clustering

Hierarchical Clustering

PCA

# K-means Clustering



Iteration 1

		(2, 10)	(5, 8)	(1, 2)	
	Point	Dist Mean 1	Dist Mean 2	Dist Mean 3	Cluster
A1	(2, 10)	0	5	9	1
A2	(2, 5)	5	6	4	3
A3	(8, 4)	12	7	9	2
A4	(5, 8)	5	0	10	2
A5	(7, 5)	10	5	9	2
A6	(6, 4)	10	5	7	2
A7	(1, 2)	9	10	0	3
A8	(4, 9)	3	2	10	2

Cluster 1  
(2, 10)

Cluster 2  
(8, 4)  
(5, 8)  
(7, 5)  
(6, 4)  
(4, 9)

Cluster 3  
(2, 5)  
(1, 2)

# K-means Clustering

- For Cluster 1, we only have one point  $A_1(2, 10)$ , which was the old mean, so the cluster center remains the same.
- For Cluster 2, we have  $((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6)$
- For Cluster 3, we have  $((2+1)/2, (5+2)/2) = (1.5, 3.5)$

# K-means Clustering

		(2, 10)	(6, 6)	(1.5, 3.5)	
	Point	Dist Mean 1	Dist Mean 2	Dist Mean 3	<b>Cluster</b>
A1	(2, 10)	0	8	7	1
A2	(2, 5)	5	5	2	3
A3	(8, 4)	12	4	7	2
A4	(5, 8)	5	3	8	2
A5	(7, 5)	10	2	7	2
A6	(6, 4)	10	2	5	2
A7	(1, 2)	9	9	2	3
A8	(4, 9)	3	5	8	1

# K-means Clustering

- In Cluster 1, we have points 1 and 8. Therefore the centroid is:  $((2+4)/2, (10+9)/2) = (3, 9.5)$
- In Cluster 2, we have points 3, 4, 5 and 6. Therefore, the centroid is:  $((8+5+7+6)/4, (4+8+5+4)/4) = (6.5, 5.25)$
- For Cluster 3, we have points 2 and 7. Therefore, the centroid is:  $((2+1)/2, (5+2)/2) = (1.5, 3.5)$

# K-means Clustering

		(3, 9.5)	(6.5 ,5.25)	(1.5, 3.5)	
	Point	Dist Mean 1	Dist Mean 2	Dist Mean 3	<b>Cluster</b>
A1	(2, 10)	1.5	9.25	7	1
A2	(2, 5)	5.5	4.75	2	3
A3	(8, 4)	10.5	2.75	7	2
A4	(5, 8)	3.5	4.25	8	1
A5	(7, 5)	8.5	0.75	7	2
A6	(6, 4)	8.5	1.75	5	2
A7	(1, 2)	9.5	8.75	2	3
A8	(4, 9)	1.5	6.25	8	1



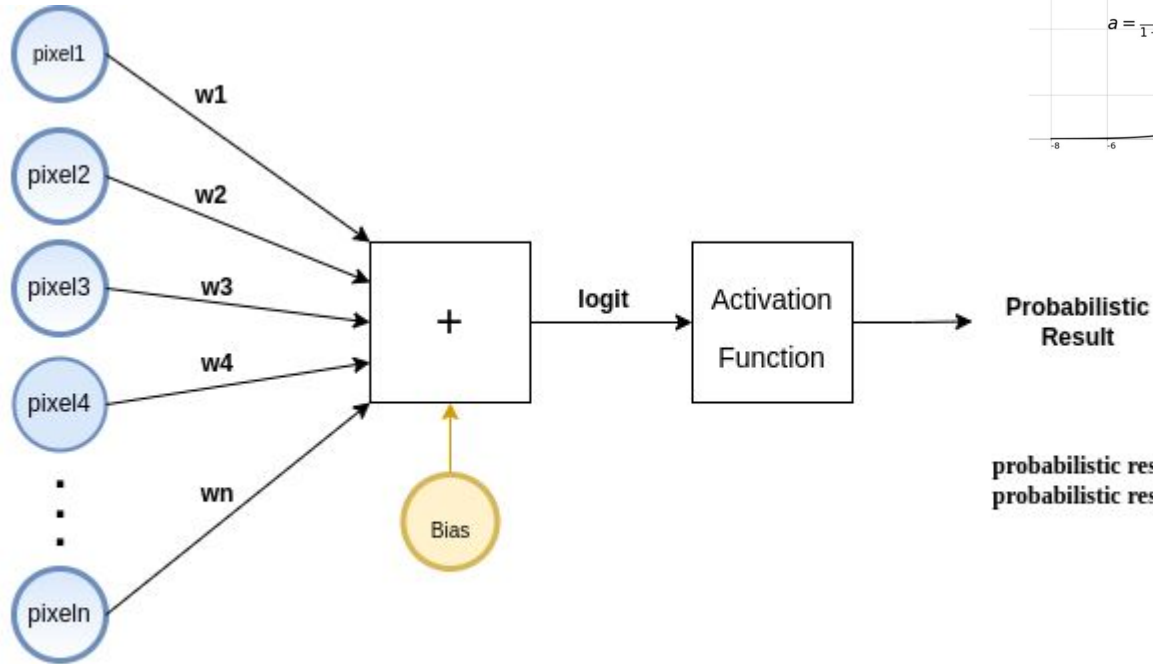
# K-means Clustering

- In Cluster 1, we have points 1, 4, and 8. Therefore the centroid is:  $((2+5+4)/2, (10+8+9)/2) = (3.67, 9)$
- In Cluster 2, we have points 3, 5 and 6. Therefore, the centroid is:  $((8+7+6)/4, (4+5+4)/4) = (7, 4.3)$
- For Cluster 3, we have points 2 and 7. Therefore, the centroid is:  $((2+1)/2, (5+2)/2) = (1.5, 3.5)$

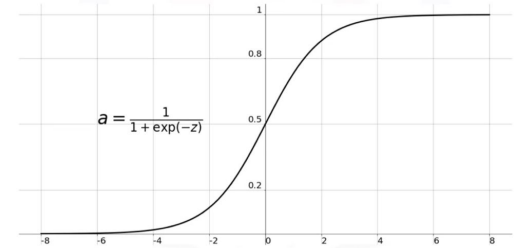
# K-means Clustering

		(3.67, 9)	(7 ,4.3)	(1.5, 3.5)	
	Point	Dist Mean 1	Dist Mean 2	Dist Mean 3	<b>Cluster</b>
A1	(2, 10)	2.67	10.7	7	<b>1</b>
A2	(2, 5)	5.67	5.7	2	<b>3</b>
A3	(8, 4)	9.33	1.3	7	<b>2</b>
A4	(5, 8)	2.33	5.7	8	<b>1</b>
A5	(7, 5)	7.33	0.7	7	<b>2</b>
A6	(6, 4)	7.33	1.3	5	<b>2</b>
A7	(1, 2)	9.67	8.3	2	<b>3</b>
A8	(4, 9)	0.33	7.7	8	<b>1</b>

# Logistic Regression



## Sigmoid Function



probabilistic result < 0.5 --> first class  
probabilistic result >= 0.5 --> second class



TRAIN DATA

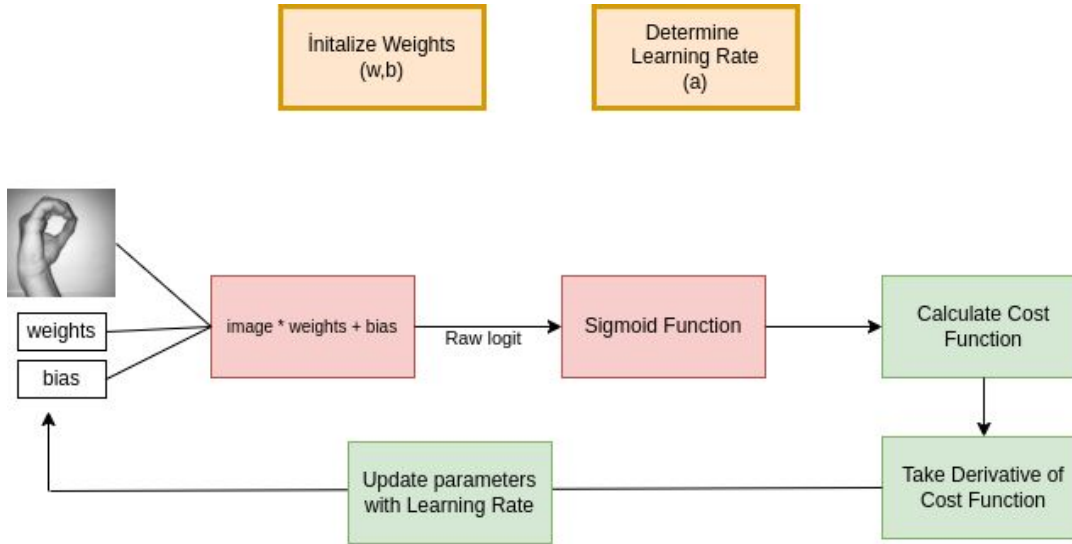
# Step by step

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

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

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

$m = \text{number of samples}$



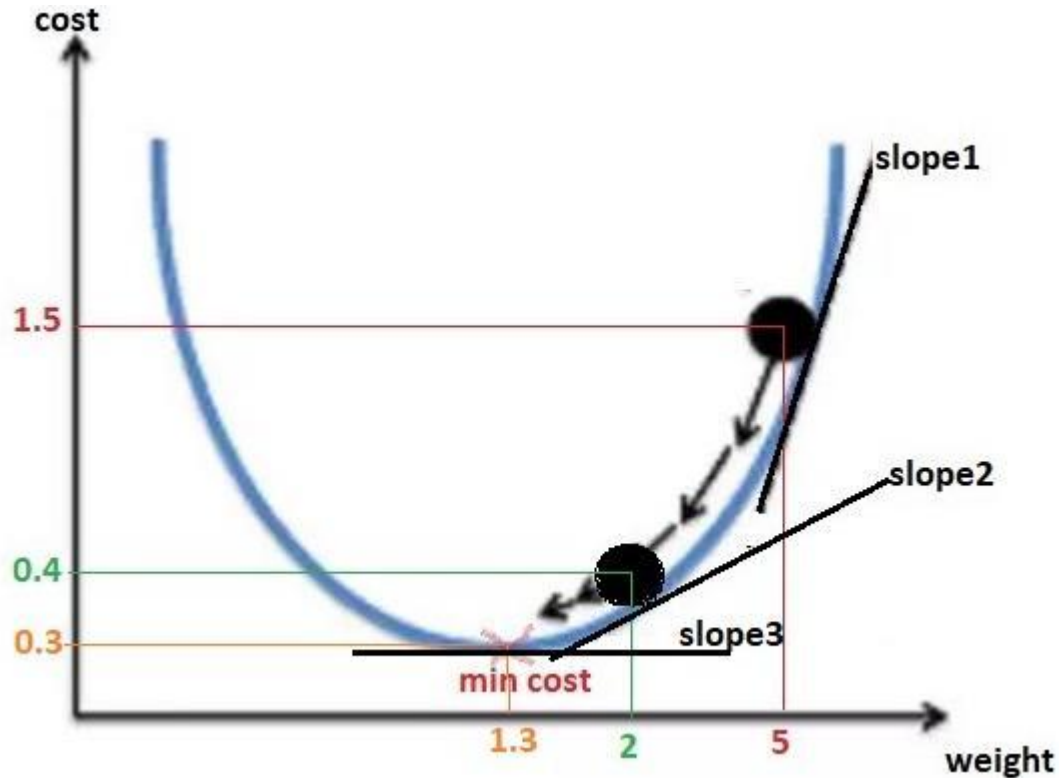
$$\frac{\partial J}{\partial w} = \frac{1}{m} x(y_{\text{head}} - y)^T$$

$$\frac{\partial J}{\partial b} = \frac{1}{m} \sum_{i=1}^m (y_{\text{head}} - y)$$

Forward Propagation

Backward Propagation

# What does it mean to take derivative?



$$w := w - \alpha \frac{\partial J(w, b)}{\partial (w, b)}$$