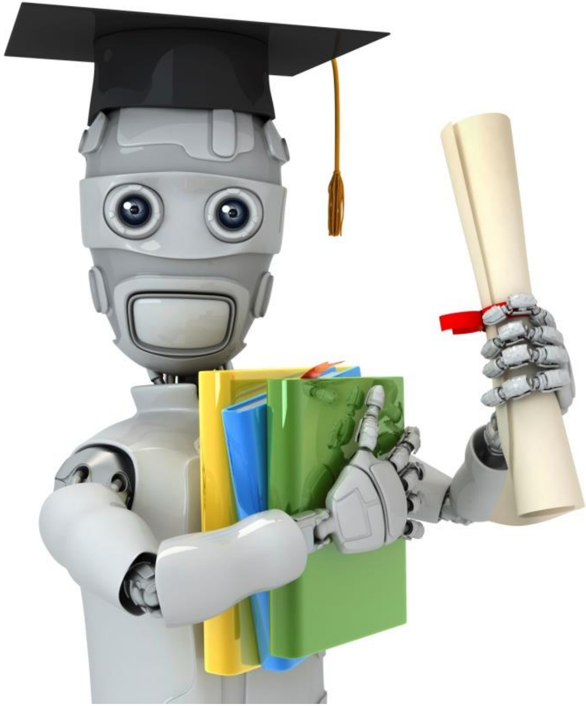


Clustering

Unsupervised learning introduction

Machine Learning

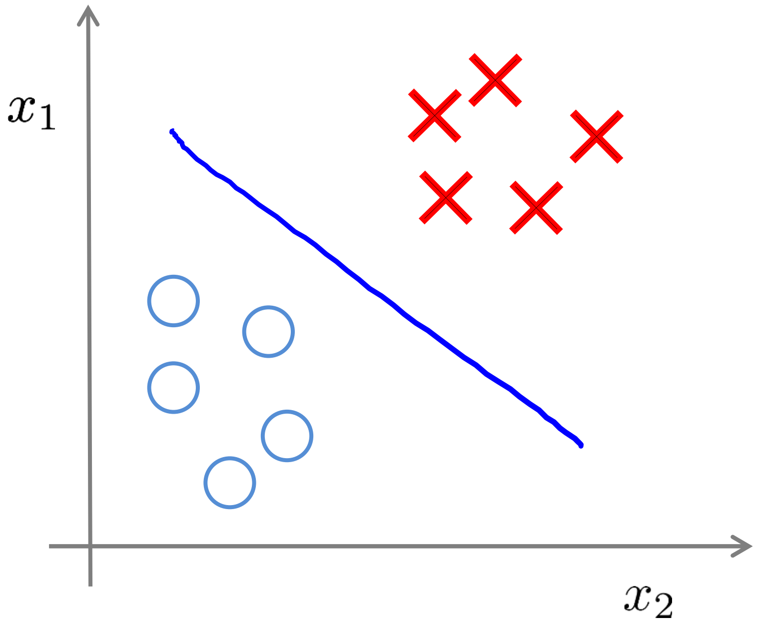
Prepared By:

Dr. Sara Sweidan

Clustering as a machine learning task

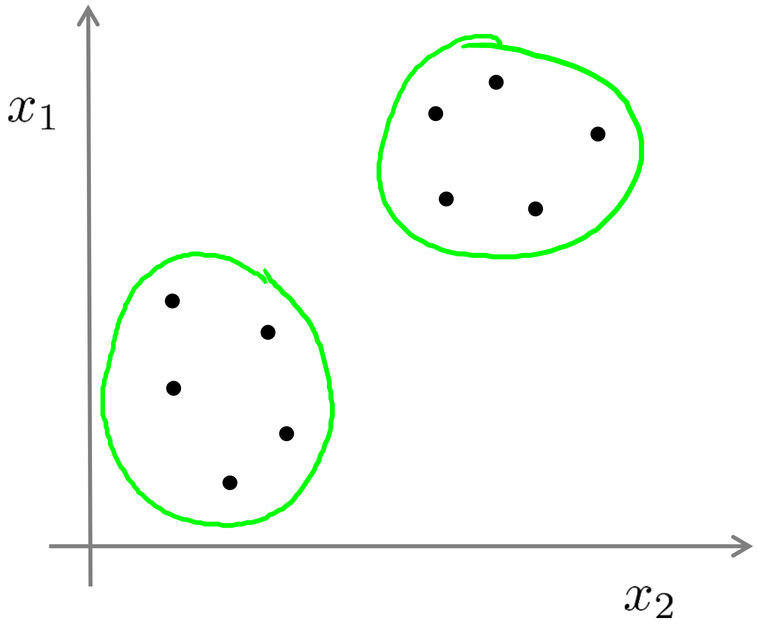
Clustering is somewhat different from the classification, numeric prediction, and pattern detection tasks we examined so far. In each of these cases, the result is a model that relates features to an outcome or features to other features; conceptually,
the model describes the existing patterns within data. In contrast, clustering creates new data. Unlabeled examples are given a cluster label that has been inferred entirely from the relationships within the data.

Supervised learning



Training set: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)}), \ldots, (x^{(m)}, y^{(m)})\}$

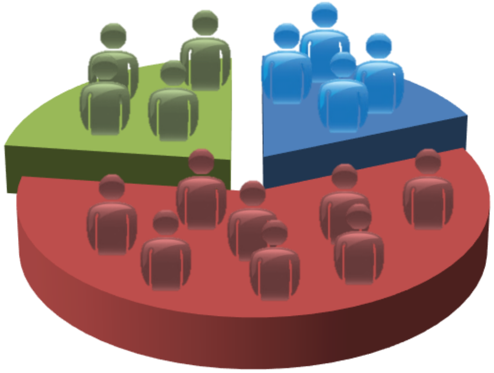
Unsupervised learning



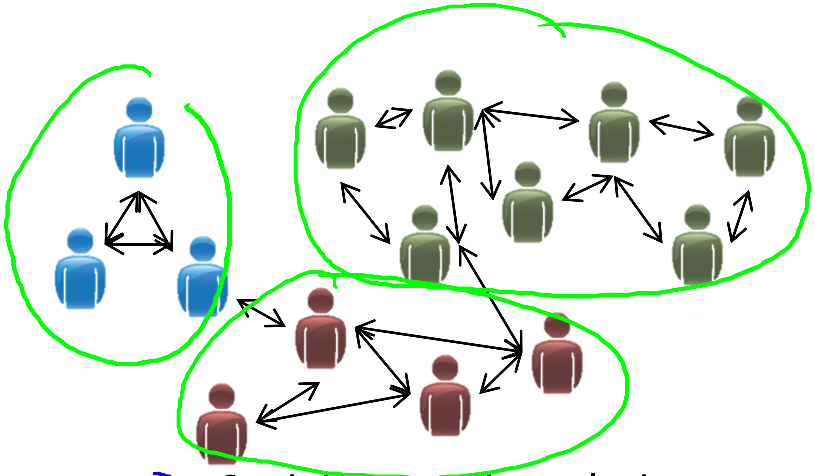
Clustering algorithm

Training set: $\{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(m)}\} \leftarrow$

Applications of clustering



Market segmentation



Social network analysis

Organize computing clusters

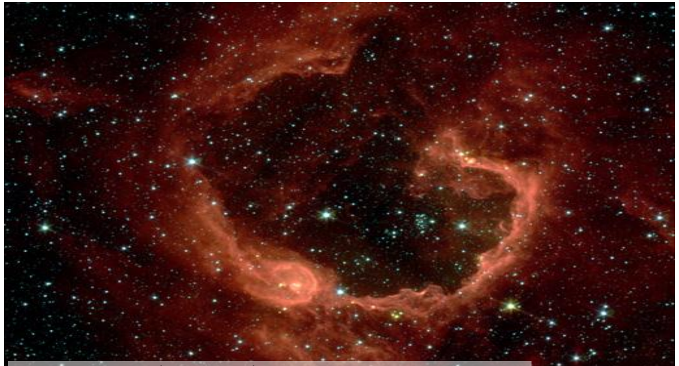
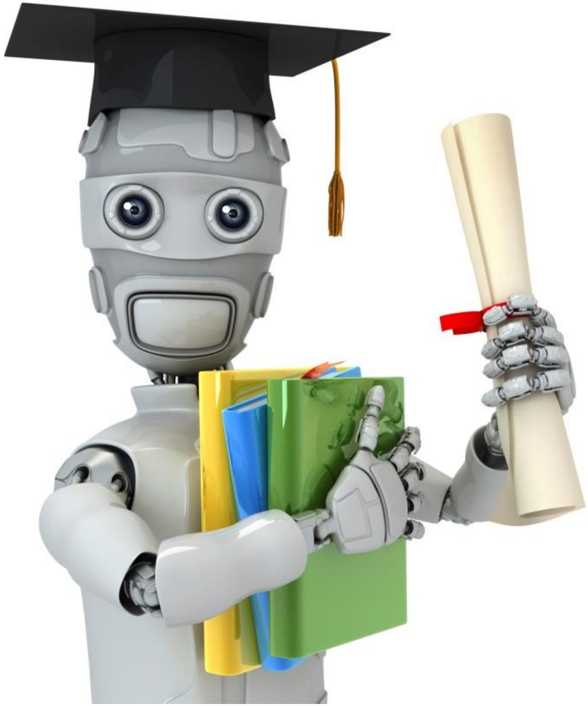


Image credit: NASA/JPL-Caltech/E. Churchwell (Univ. of Wisconsin, Madison)

Astronomical data analysis



Machine Learning

Clustering

K-means algorithm

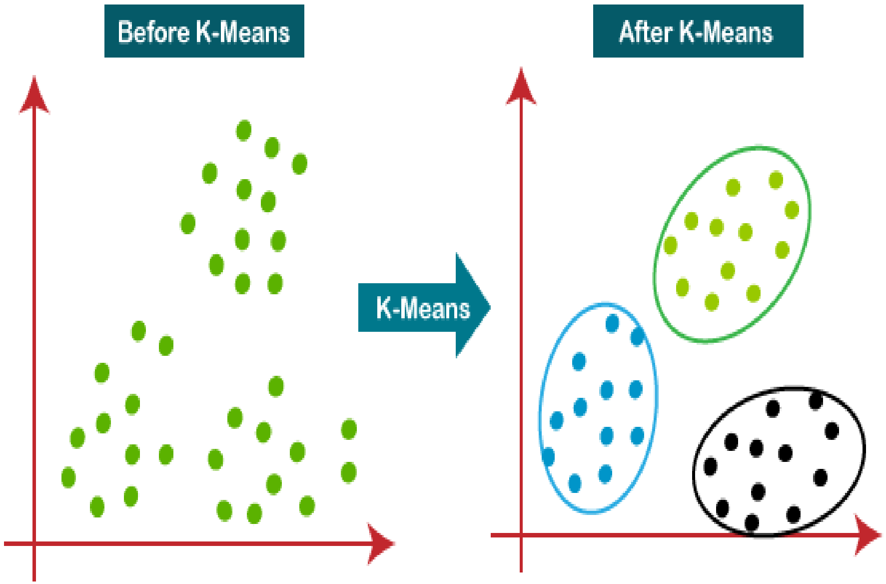
What is K-means used for?

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

The k-means clustering algorithm mainly performs two tasks:

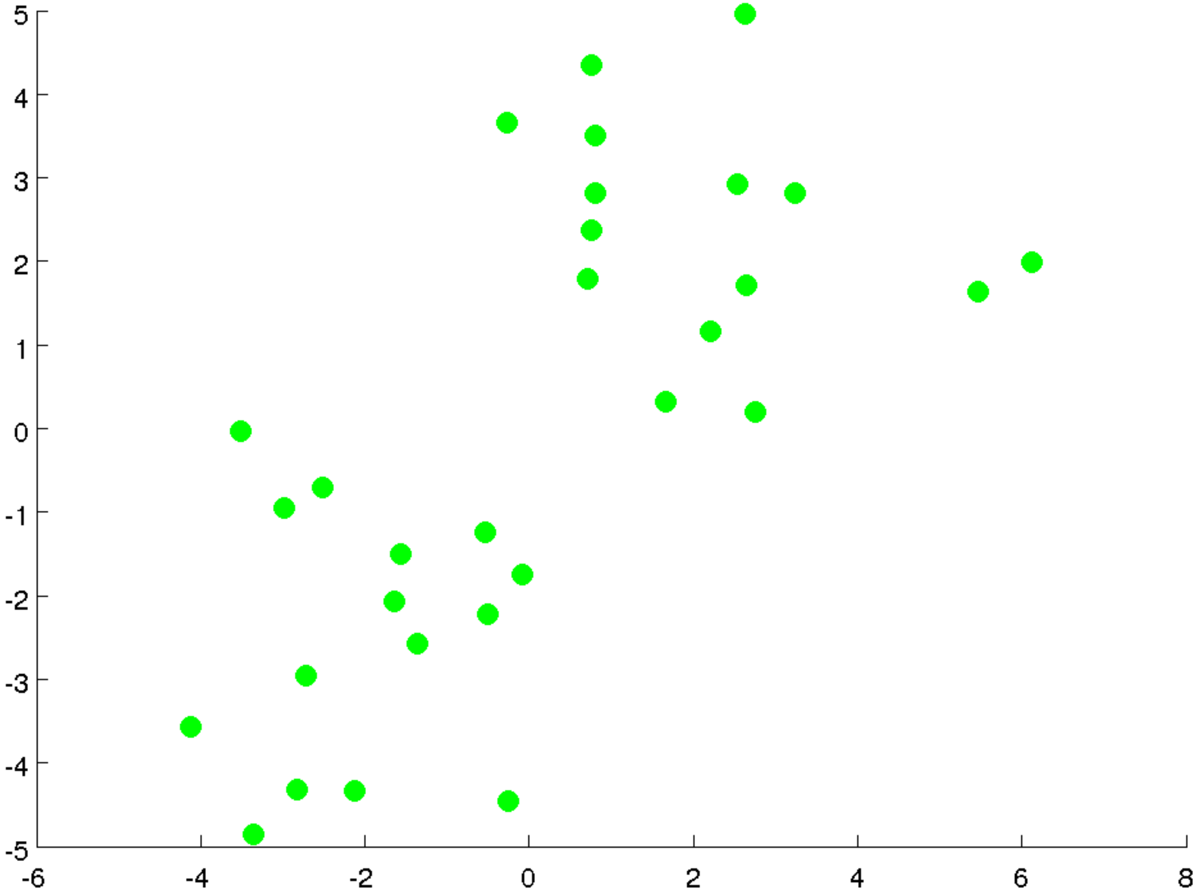


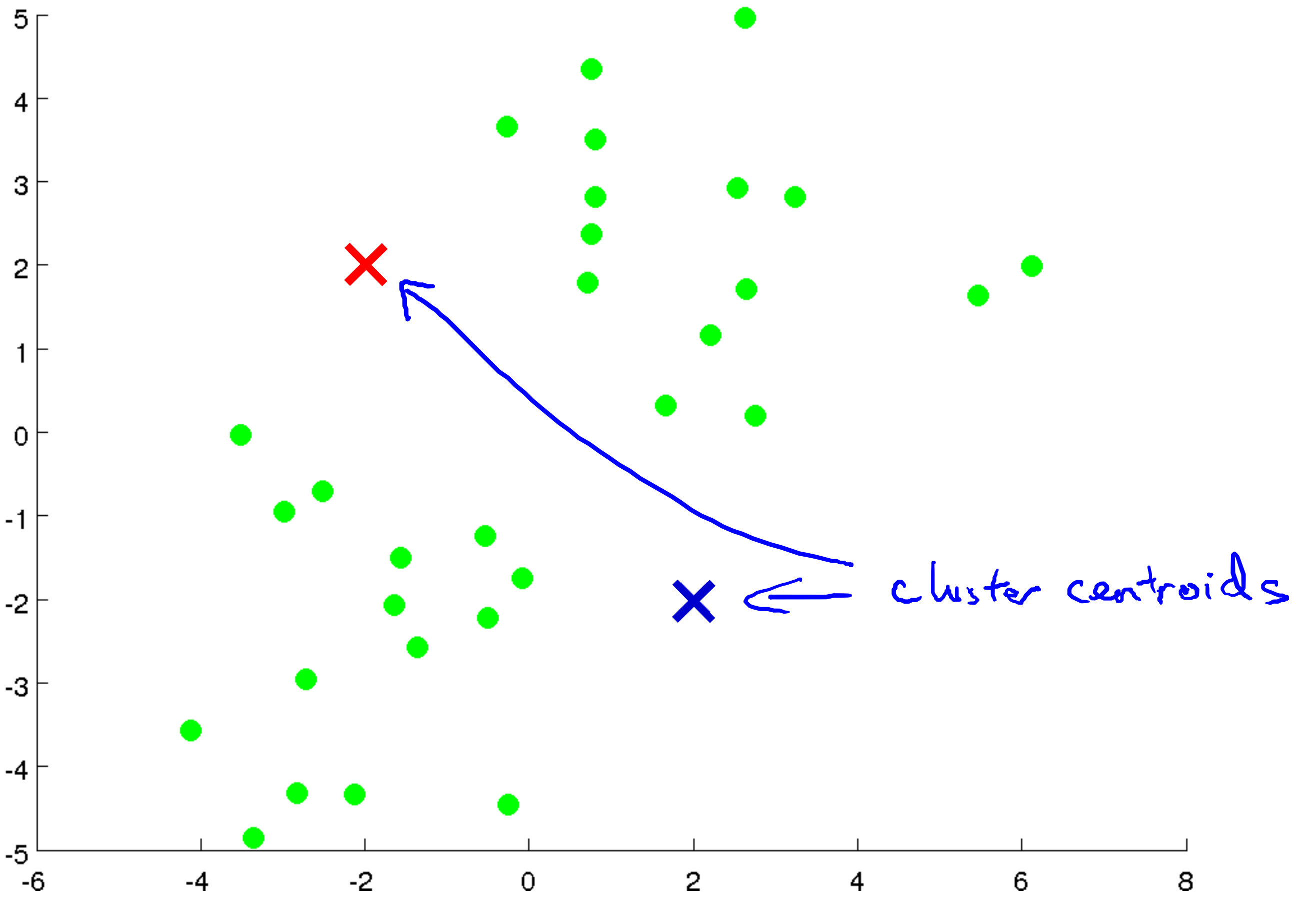
- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

K-means algorithm:

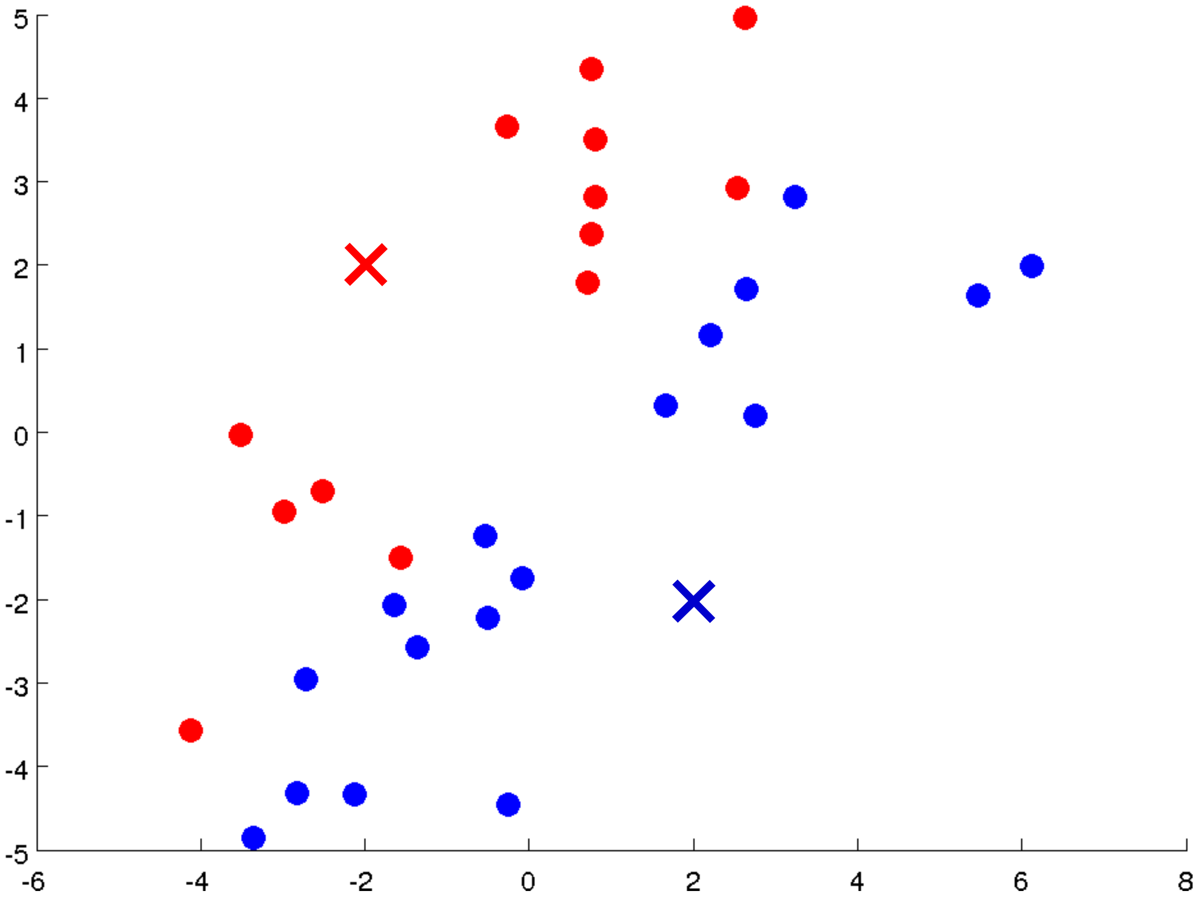
- **Step-1:** Select the number K to decide the number of clusters.
- **Step-2:** Select random K points or centroids. (It can be other from the input dataset).
- **Step-3:** Assign each data point to their closest centroid, which will form the predefined K clusters.
- **Step-4:** Calculate the mean-value and place a new centroid of each cluster.
- **Step-5:** Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.
- **Step-6:** If any reassignment occurs, then go to step-4 else go to FINISH.
- **Step-7**: The model is ready.

Advantages	Limitations
- Uses simple principles that can be explained in non-statistical terms • Highly flexible, and can be adapted with simple adjustments to address nearly all of its shortcomings • Performs well enough under many real-world use cases	- Not as sophisticated as more modern clustering algorithms • Because it uses an element of random chance, it is not guaranteed to find the optimal set of clusters • Requires a reasonable guess as to how many clusters naturally exist in the data • Not ideal for non-spherical clusters or clusters of widely varying density

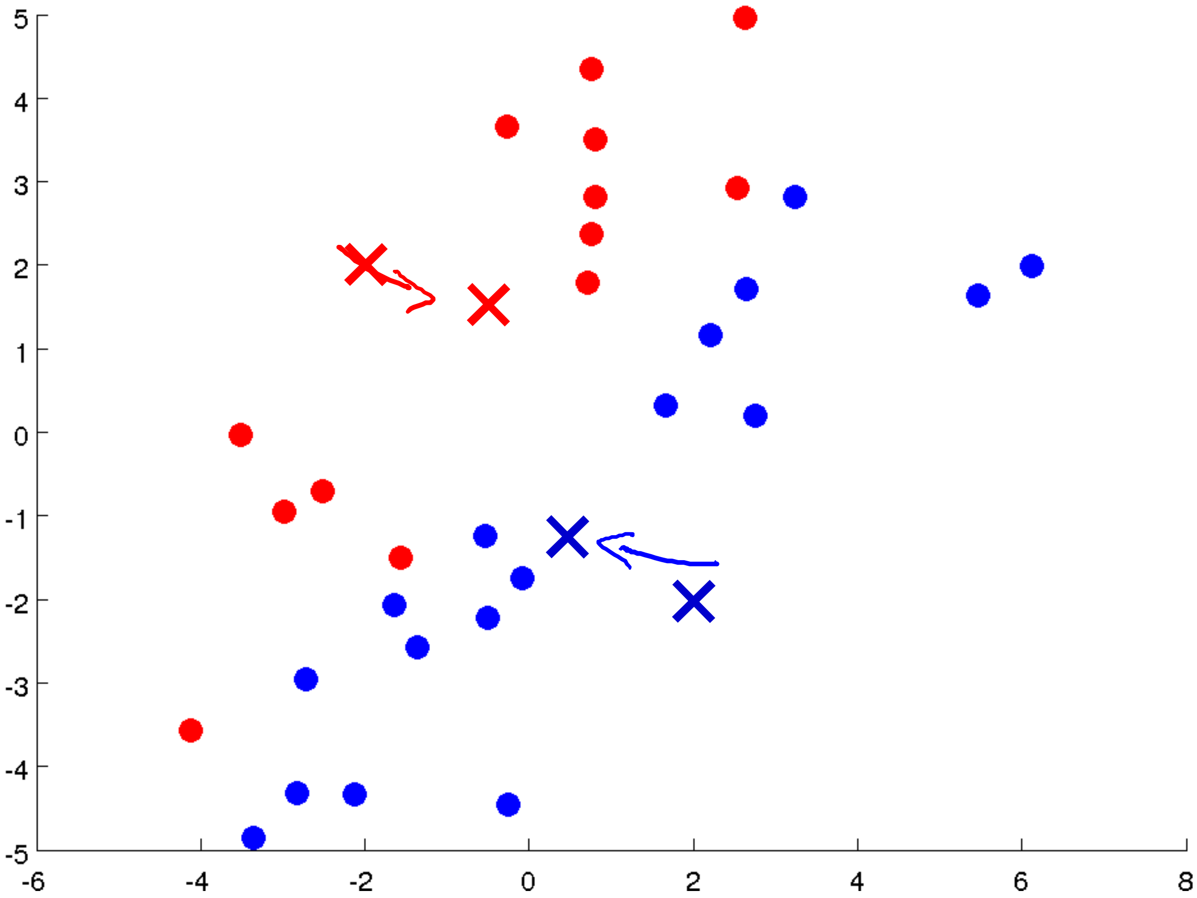

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8



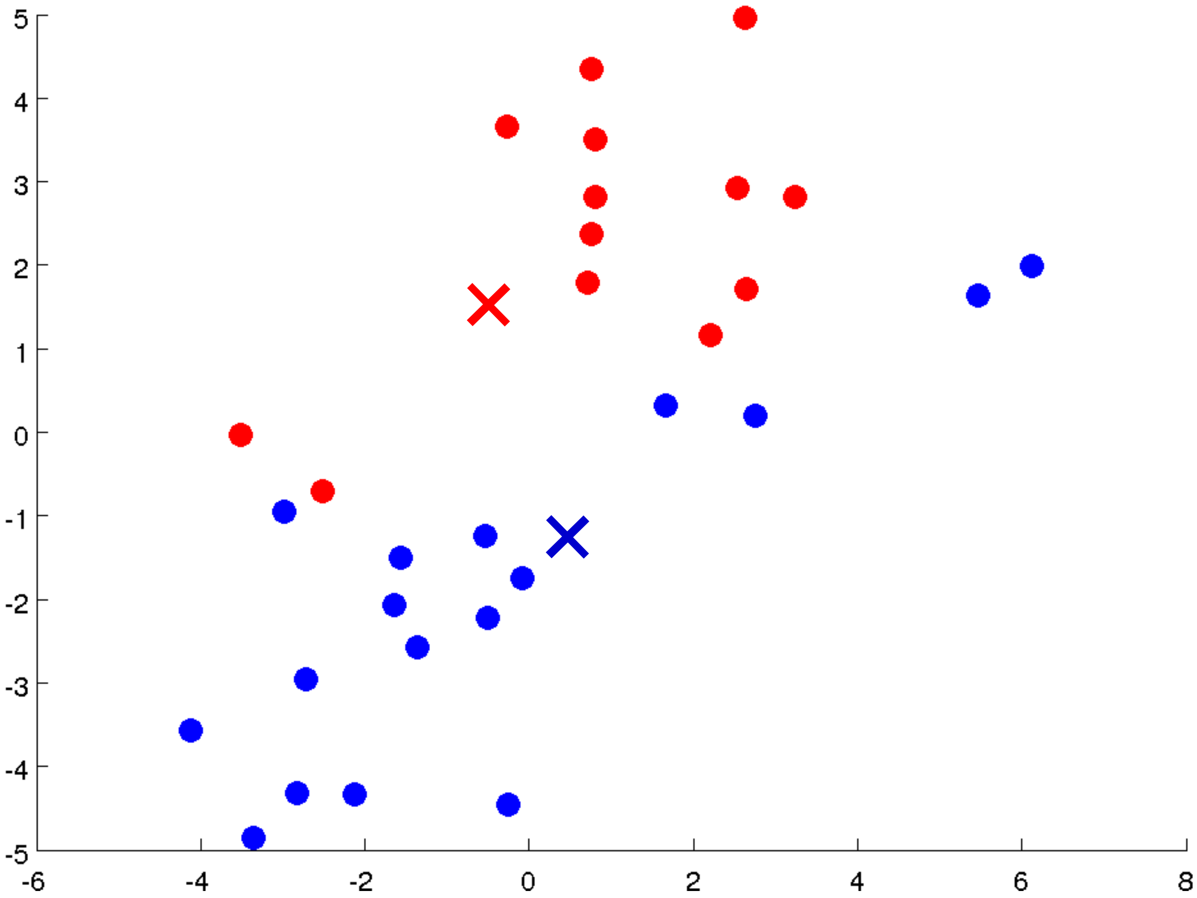
Andrew Ng

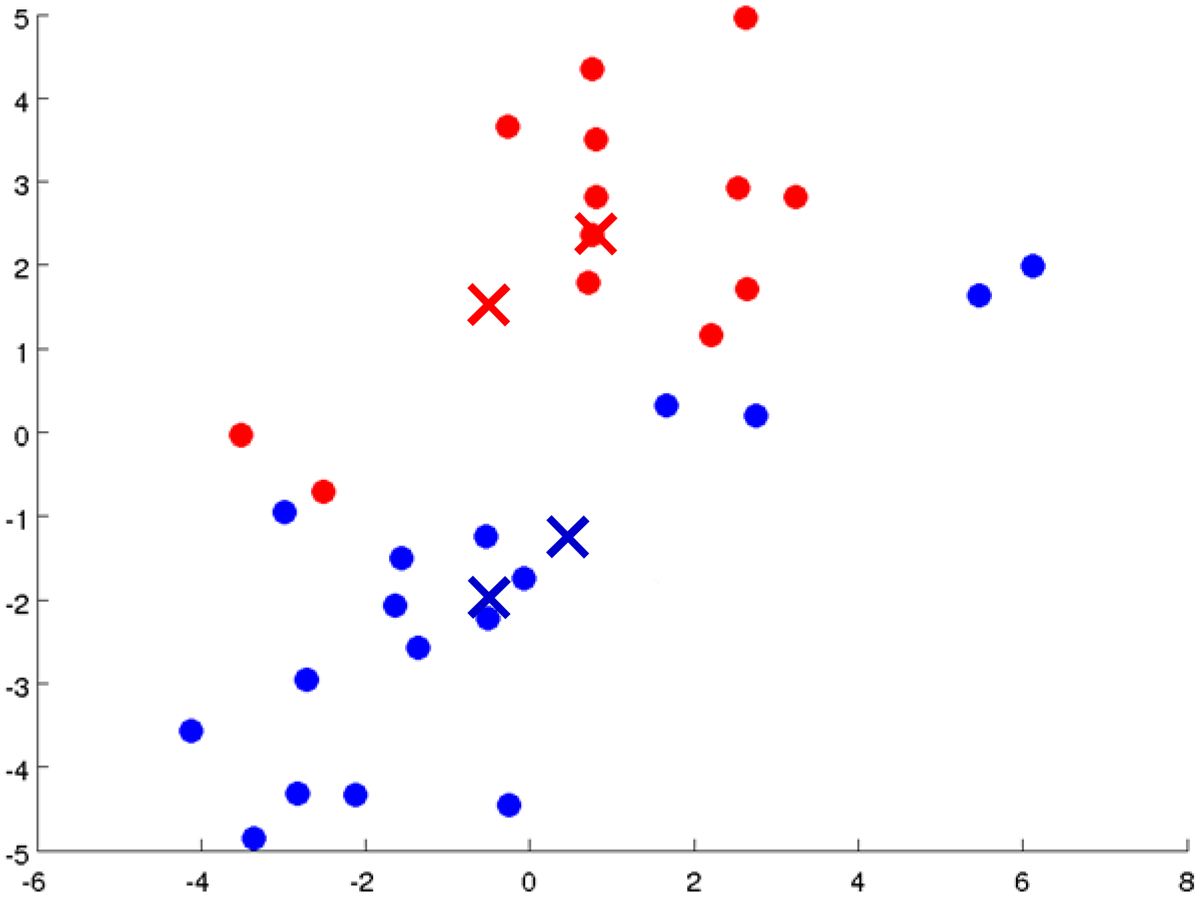


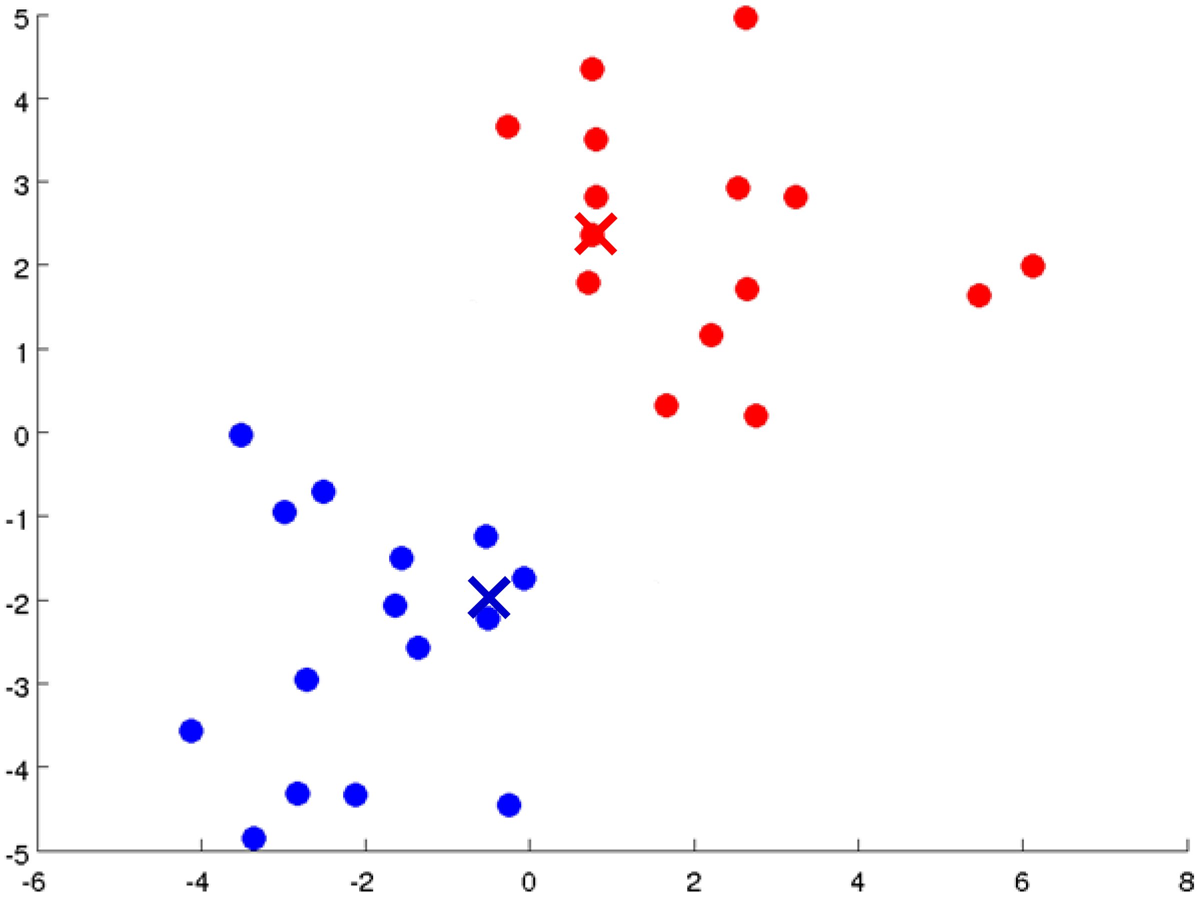
Andrew Ng

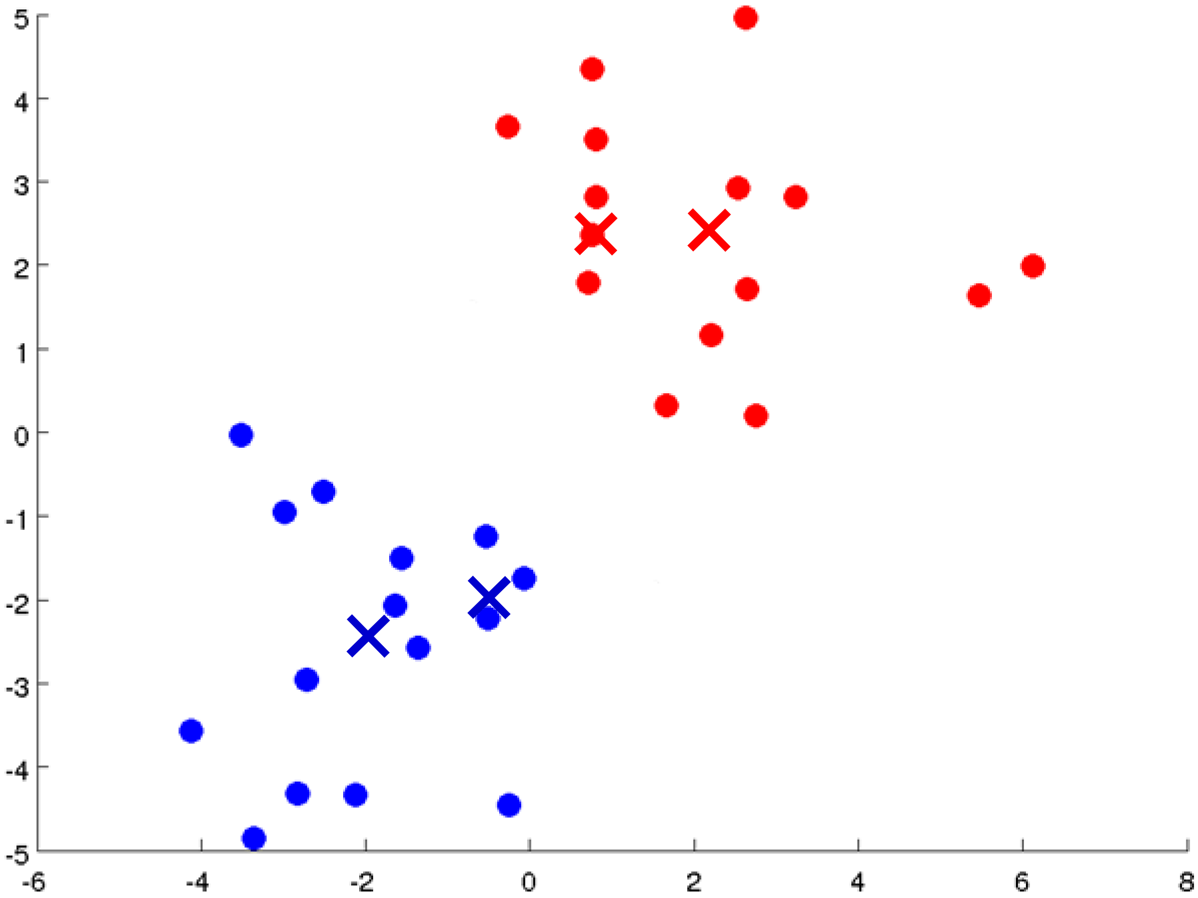


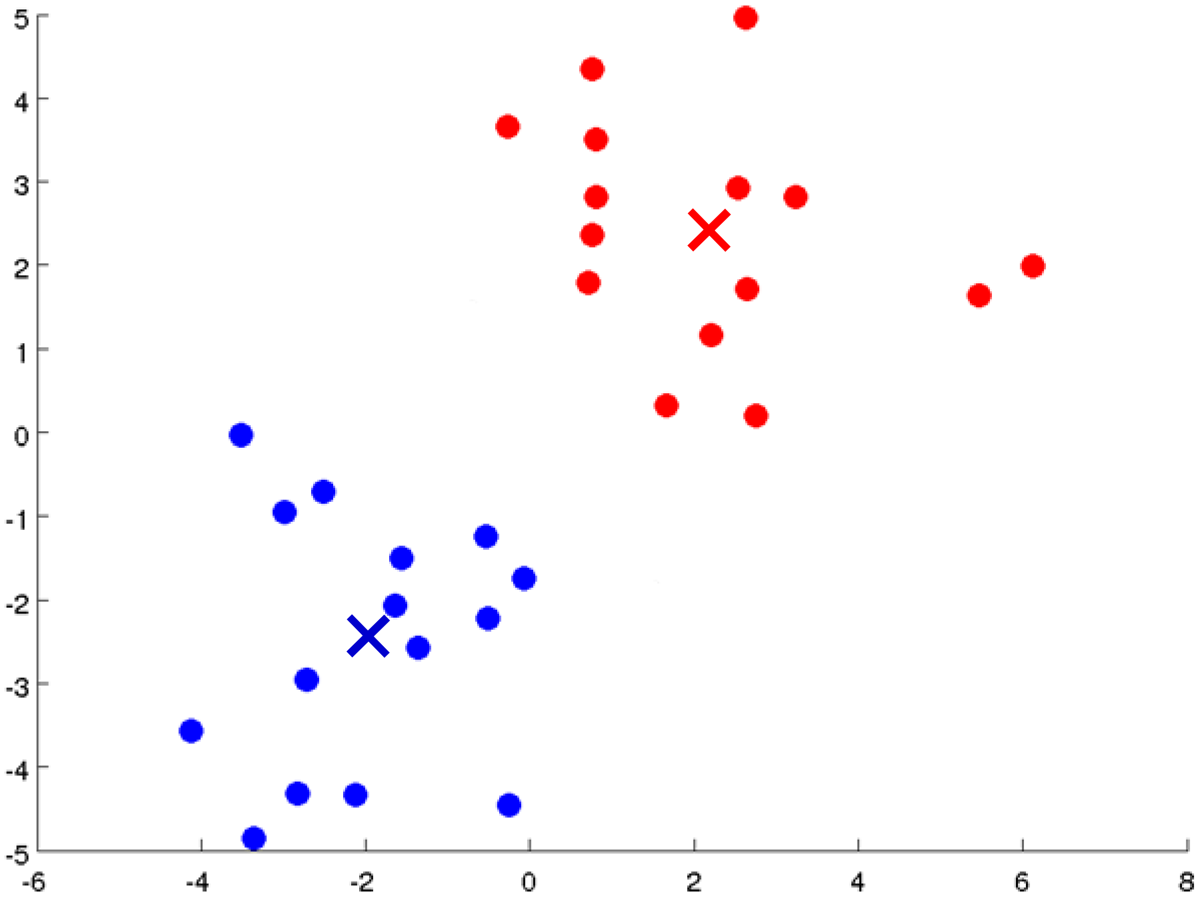
Andrew Ng


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8



5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

K-means algorithm

Input:

- K (number of clusters)
- Training set $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$

$x^{(i)} \in \mathbb{R}^n$ (drop $x_0 = 1$ convention)

K-means algorithm

μ_1 $\times$ μ_2 $\times$

Randomly initialize K cluster centroids $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^n$

Repeat {

Cluster assignment step

for i = 1 to m

$c^{(i)}$:= index (from 1 to K) of cluster centroid closest to $x^{(i)}$

$\min_k \|x^{(i)} - \mu_k\|^2 \rightarrow c^{(i)}$

Move centroid

for k = 1 to K

$\rightarrow$ μ_k := average (mean) of points assigned to cluster k

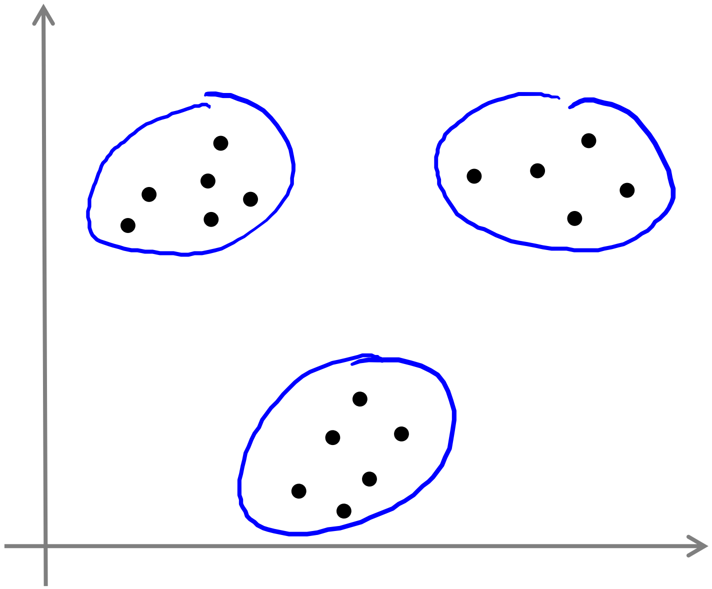
$x^{(1)}, x^{(5)}, x^{(6)}, x^{(10)}$ $\rightarrow$ $c^{(1)}=2,\ c^{(5)}=2,\ c^{(6)}=2,\ c^{(10)}=2$

$$\mu_2 = \frac{1}{4}\left[x^{(1)} + x^{(5)} + x^{(6)} + x^{(10)}\right] \in \mathbb{R}^n$$

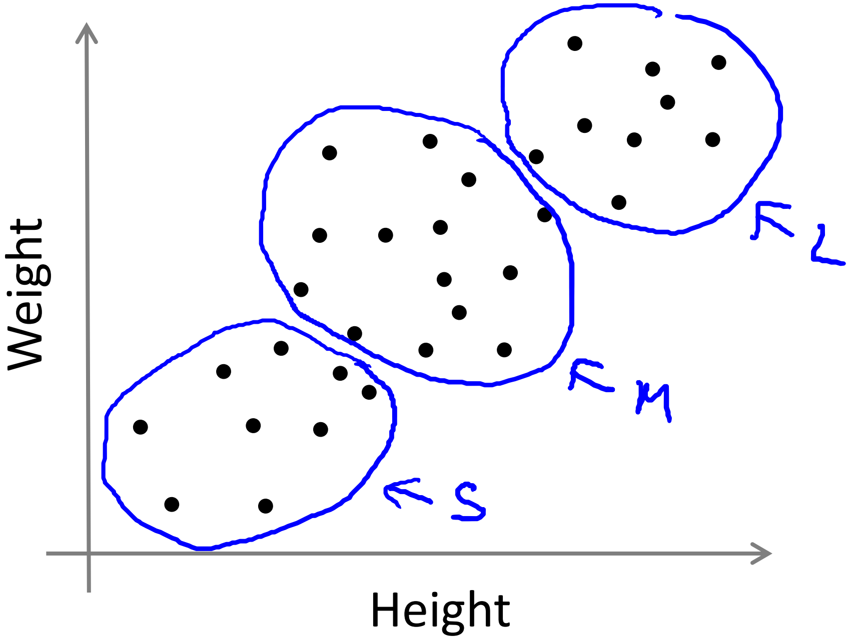
}

K-means for non-separated clusters

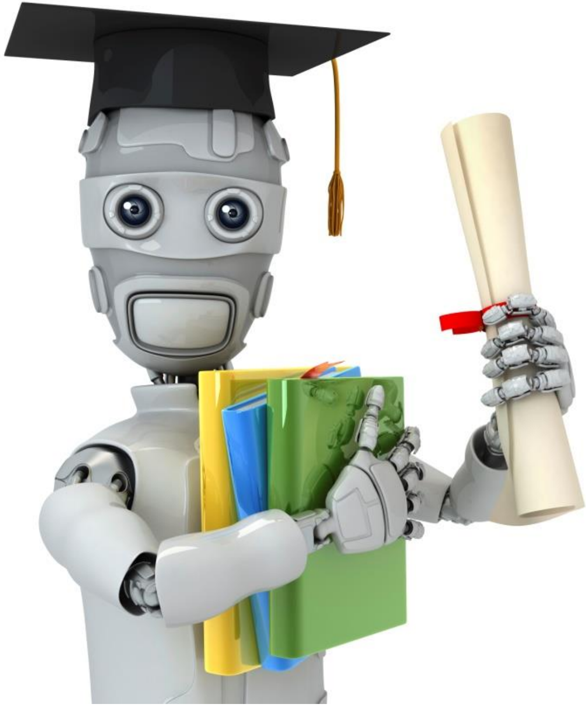
S, M, L



T-shirt sizing



Andrew Ng



Machine Learning

Clustering

Optimization objective

K-means optimization objective

$c^{(i)}$ = index of cluster (1,2,…,K) to which example $x^{(i)}$ is currently assigned

μ_k = cluster centroid $\underline{k}$ ($\mu_k \in \mathbb{R}^n$)   K $k \in \{1, 2, \ldots, K\}$

$\mu_{c^{(i)}}$ = cluster centroid of cluster to which example $x^{(i)}$ has been assigned

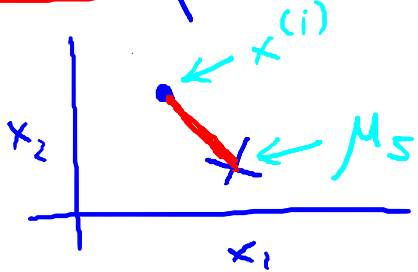
$x^{(i)} \rightarrow \underline{5}$ $\underline{c^{(i)} = 5}$ $\underline{\mu_{c^{(i)}}} = \mu_5$

Optimization objective:

$$J(\underline{c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K}) = \frac{1}{m} \sum_{i=1}^{m} \boxed{||x^{(i)} - \mu_{c^{(i)}}||^2}$$

$$\min_{\substack{c^{(1)}, \ldots, c^{(m)}, \\ \mu_1, \ldots, \mu_K}} J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K)$$

Distortion



K-means algorithm

Randomly initialize K cluster centroids $\mu_1, \mu_2, \dots, \mu_K \in \mathbb{R}^n$

Cluster assignment step

Minimize $J(\dots)$ wrt $\boxed{c^{(1)}, c^{(2)}, \dots, c^{(m)}}$ ←

(holding $\mu_1, \dots, \mu_K$ fixed)

Repeat {

for i = 1 to m

$c^{(i)}$:= index (from 1 to K) of cluster centroid closest to $x^{(i)}$

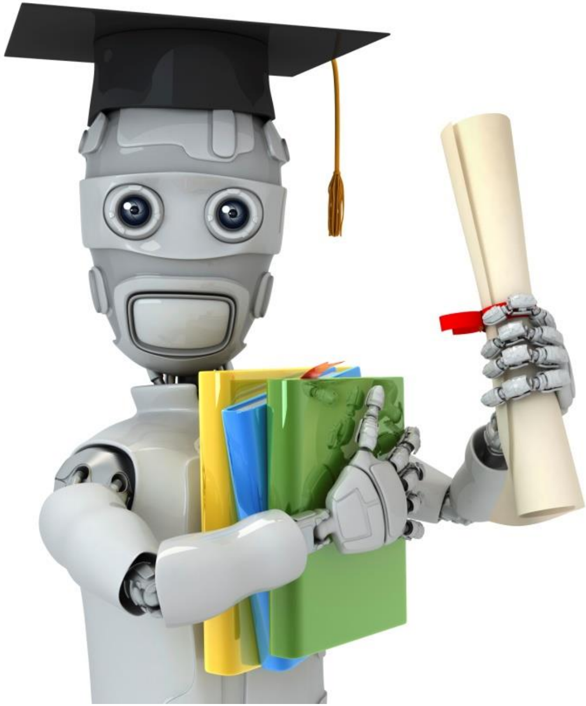
move centroid

for k = 1 to K

μ_k := average (mean) of points assigned to cluster k

}

minimize $J(\dots)$ wrt $\boxed{\mu_1, \dots, \mu_K}$



Machine Learning

Clustering

Random initialization

K-means algorithm

Randomly initialize K cluster centroids $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^n$

Repeat {

for i = 1 to m

$c^{(i)}$:= index (from 1 to K) of cluster centroid closest to $x^{(i)}$

for k = 1 to K

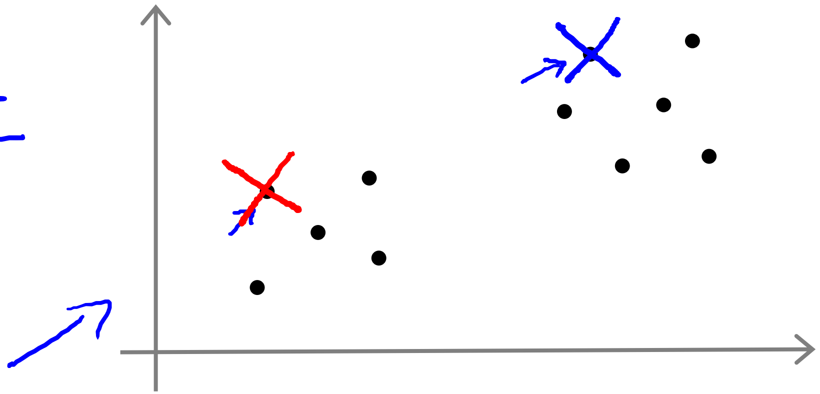
μ_k := average (mean) of points assigned to cluster k

}

Random initialization

$K=2$

Should have $K < m$



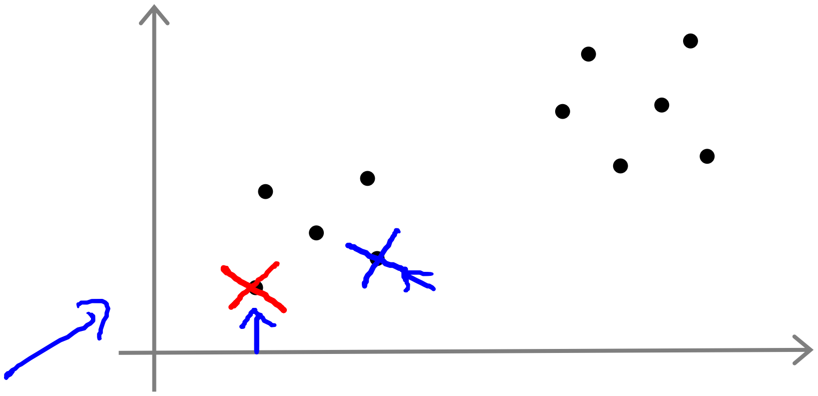
Randomly pick K training examples.

Set $\mu_1, \ldots, \mu_K$ equal to these K examples.

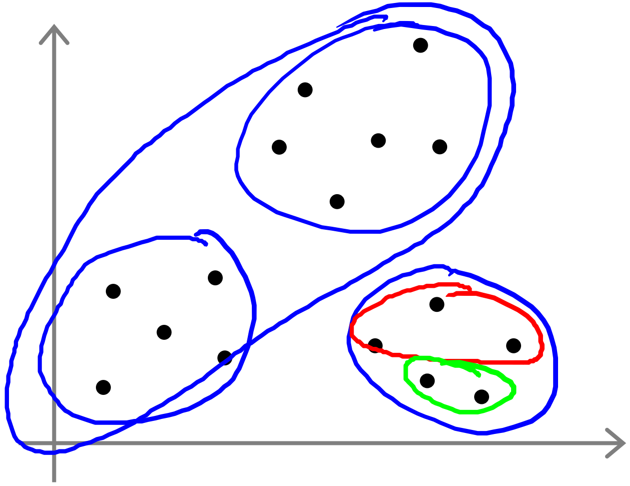
$\mu_1 = x^{(i)}$

$\mu_2 = x^{(j)}$

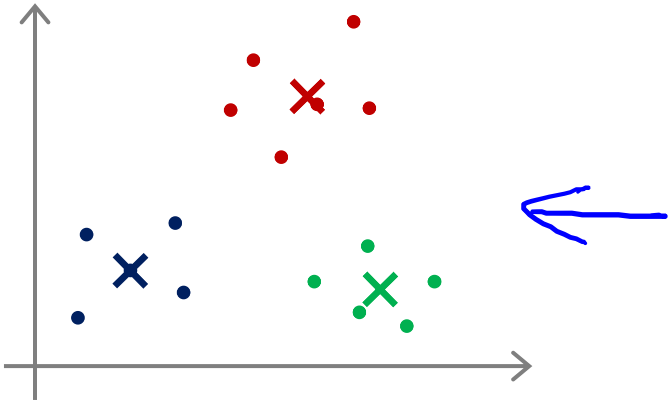
$\vdots$

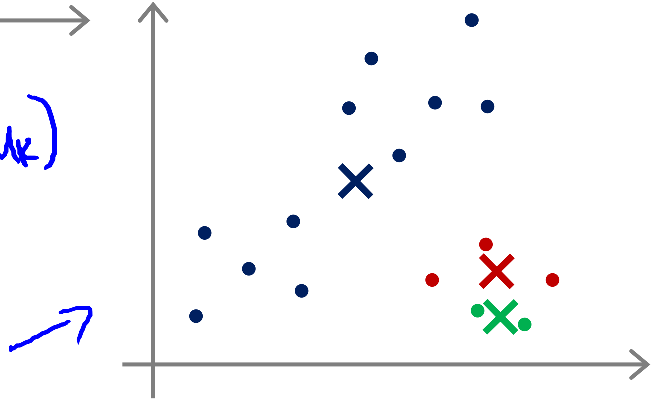


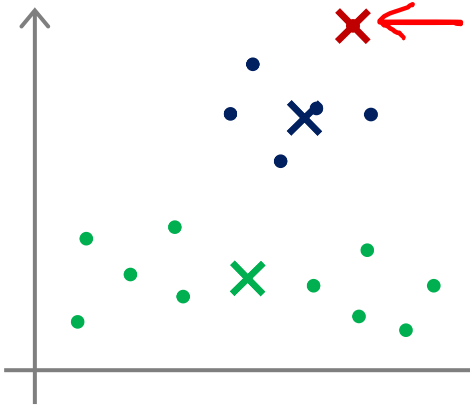
Local optima



$$J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K)$$







Random initialization

For i = 1 to 100 {

Randomly initialize K-means.

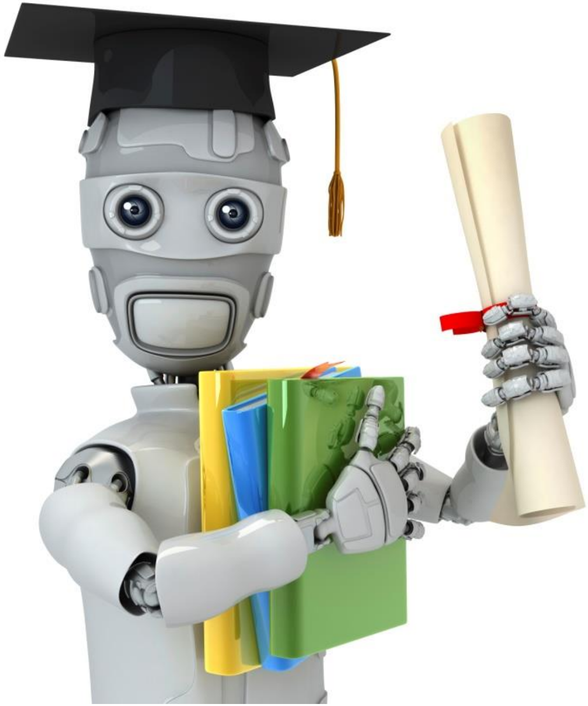
Run K-means. Get $c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K$.

Compute cost function (distortion)

$$J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K)$$

}

Pick clustering that gave lowest cost $J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K)$

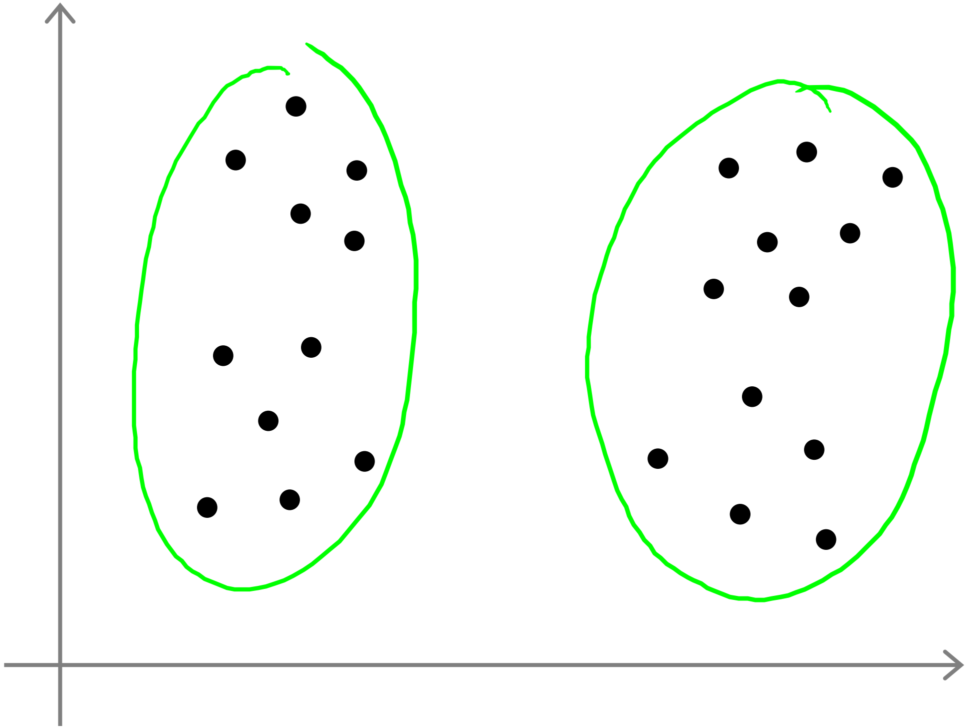


Machine Learning

Clustering

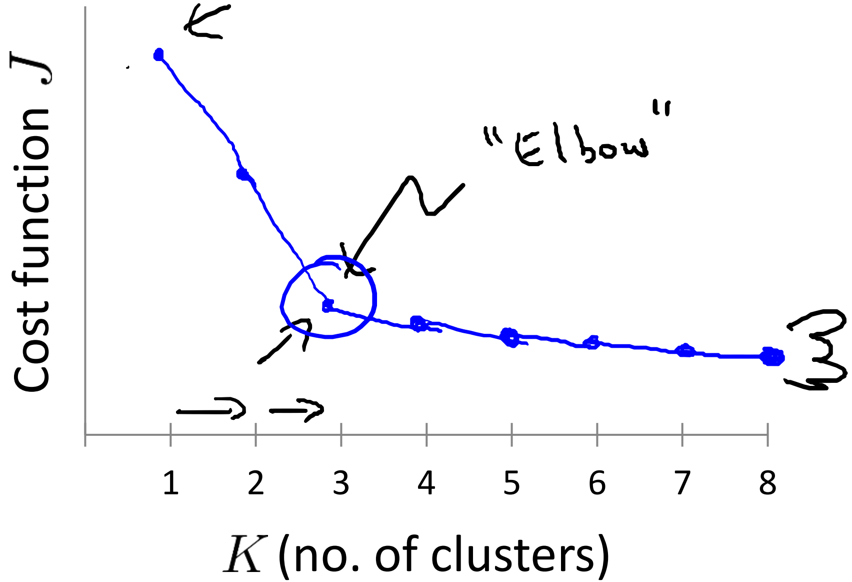
Choosing the number of clusters

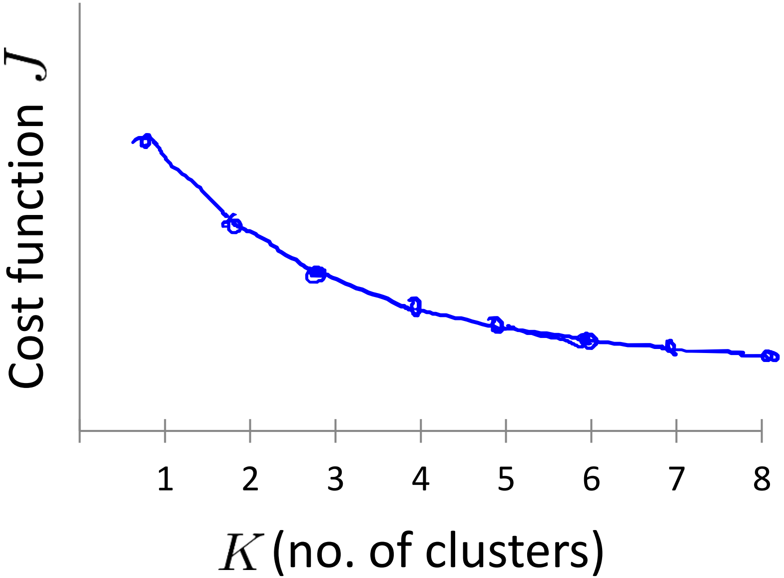
What is the right value of K?



Choosing the value of K

Elbow method:





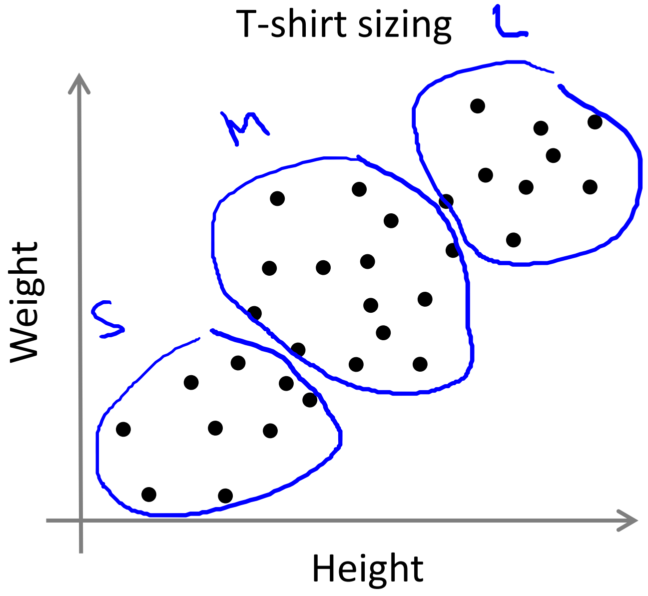
Choosing the value of K

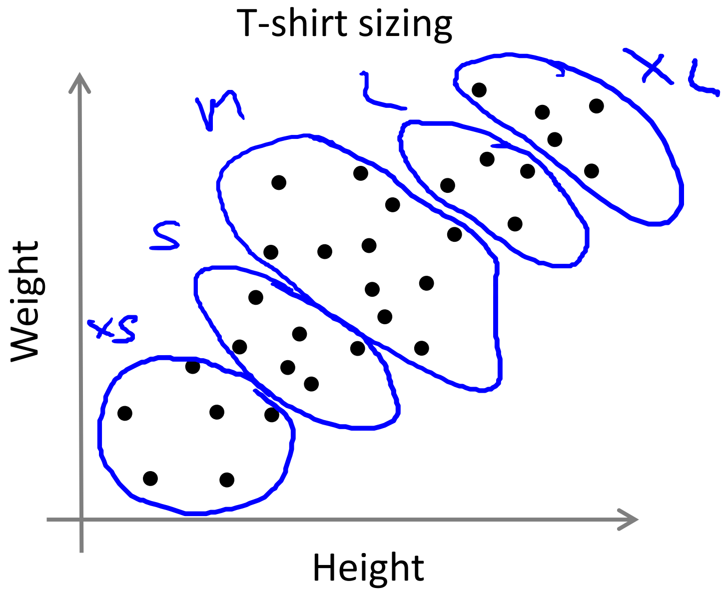
Sometimes, you're running K-means to get clusters to use for some later/downstream purpose. Evaluate K-means based on a metric for how well it performs for that later purpose.

K=3 S, M, L

K=5 XS, S, M, L, XL

E.g.





Andrew Ng

Id	Height	weight
y_1	185	72
y_2	170	56
y_3	168	60
y_4	179	68
y_5	183	84
y_6	180	67
y_7	177	76
y_8	166	55

Divide the given sample data in two (2) clusters using k-mean algorithm [Euclidean distance].

Solution:

K=2

initialize centroid $\mu_1 = (185,72), \mu_2 = (170, 56)$

	μ_1	μ_2
1	$\sqrt{(\mathbf{185-185})^2+(\mathbf{72-72})^2}$ = 0	$\sqrt{(\mathbf{185-170})^2+(\mathbf{72-56})^2}$ = 21.9
2	$\sqrt{(\mathbf{170-185})^2+(\mathbf{56-72})^2}$ = 21.9	$\sqrt{(\mathbf{170-170})^2+(\mathbf{56-56})^2}$ = 0
3	$\sqrt{(\mathbf{168-185})^2+(\mathbf{60-72})^2}$ = 20.8	$\sqrt{(\mathbf{168-170})^2+(\mathbf{60-56})^2}$ = 4.47
4	$\sqrt{(\mathbf{179-185})^2+(\mathbf{68-72})^2}$ = 7.2	$\sqrt{(\mathbf{179-170})^2+(\mathbf{68-56})^2}$ = 15
5	$\sqrt{(\mathbf{183-185})^2+(\mathbf{84-72})^2}$ = 16.4	$\sqrt{(\mathbf{183-170})^2+(\mathbf{84-56})^2}$ = 30.87
6	$\sqrt{(\mathbf{180-185})^2+(\mathbf{67-72})^2}$ = 7.07	$\sqrt{(\mathbf{180-170})^2+(\mathbf{67-56})^2}$ = 14.8
7	$\sqrt{(\mathbf{177-185})^2+(\mathbf{76-72})^2}$ = 8.9	$\sqrt{(\mathbf{177-170})^2+(\mathbf{76-56})^2}$ = 21.18
8	$\sqrt{(\mathbf{166-185})^2+(\mathbf{55-72})^2}$ = 25.4	$\sqrt{(\mathbf{166-170})^2+(\mathbf{55-56})^2}$ = 4.12

C1=$\{y_1, y_4, y_5, y_6, y_7\}$ C2= $\{y_2, y_3, y_8\}$

To calculate the new centroid from the members of each cluster:

$$\mu_1 = \left(\frac{185+179+183+180+177}{5}, \frac{72+68+84+67+76}{5}\right) = (180.8, 73.4)$$

$$\mu_2 = \left(\frac{170 + 168 + 166}{3}, \frac{56 + 60 + 66}{3}\right) = (168, 57)$$

New centroid= $\mu_1 = (180.8, 73.4)$, $\mu_2 = (168, 57)$

New centroid= $\mu_1 = (180.8, 73.4), \mu_2 = (168, 57)$

	μ_1	μ_2
1	4.427188724	22.6715681
2	20.4792578	2.236067977
3	18.53105502	3
4	5.692099788	15.55634919
5	10.82589488	30.88689042
6	6.449806199	15.62049935
7	4.604345773	21.02379604
8	23.61355543	2.828427125

C1= $\{y_1, y_4, y_5, y_6, y_7\}$ C2= $\{y_2, y_3, y_8\}$

Suppose we have a dataset with 6 data points in 2 dimensions: $x^{(1)} = (2, 2)$, $x^{(2)} = (2, 4)$, $x^{(3)} = (4, 2)$, $x^{(4)} = (4, 4)$, $x^{(5)} = (8, 6)$, and $x^{(6)} = (8, 8)$.

We want to cluster these points into K=3 clusters by minimizing the optimization objective J.

Let's initialize the cluster centroids randomly as: $\mu_1 = (2, 3)$, $\mu_2 = (5, 5)$, $\mu_3 = (7, 7)$

The optimization objective J is calculated as:

$J = \Sigma(i=1 \text{ to } 6) ||x^{(i)} - \mu_{c^{(i)}}||^2$

Where $c^{(i)}$ is the index of the closest centroid to $x^{(i)}$.

Initially, let's assign each point to its closest centroid: $c^{(1)} = 1$, $c^{(2)} = 1$, $c^{(3)} = 2$, $c^{(4)} = 2$, $c^{(5)} = 3$, $c^{(6)} = 3$

Then J becomes: $J = ||x^{(1)} - \mu_1||^2 + ||x^{(2)} - \mu_1||^2 + ||x^{(3)} - \mu_2||^2 + ||x^{(4)} - \mu_2||^2 + ||x^{(5)} - \mu_3||^2 + ||x^{(6)} - \mu_3||^2 = (2-2)^2 + (2-3)^2 + (4-2)^2 + (4-3)^2 + (4-5)^2 + (2-5)^2 + (8-7)^2 + (6-7)^2 + (8-7)^2 + (8-7)^2 = 0 + 1 + 4 + 1 + 1 + 9 + 1 + 1 + 1 + 1$
$= 20$

Now we update the centroids by taking the mean of points assigned to each centroid:

$\mu_1 = (2+2)/2, (2+4)/2 = (2, 3)$

$\mu_2 = (4+4)/2, (2+4)/2 = (4, 3)$

$\mu_3 = (8+8)/2, (6+8)/2 = (8, 7)$

With these new centroids, we reassign points and recalculate J.


THANK
YOU!