VIT®
BHOPAL
www.vitbhopal.ac.in

Introduction to clustering

(Practice Exercise on K-Means Clustering Algorithm)

Dr. Virendra Singh Kushwah

Assistant Professor Grade-II

School of Computing Science and Engineering

Virendra.Kushwah@vitbhopal.ac.in

7415869616

Consider 4 data points A,B,C,D as below

	X1	X2
A	2	3
B	6	1
C	1	2
D	3	0

K=2

K - mean → mean or avg of cluster

no. of clusters

- Choose two centroids AB and CD, calculated as

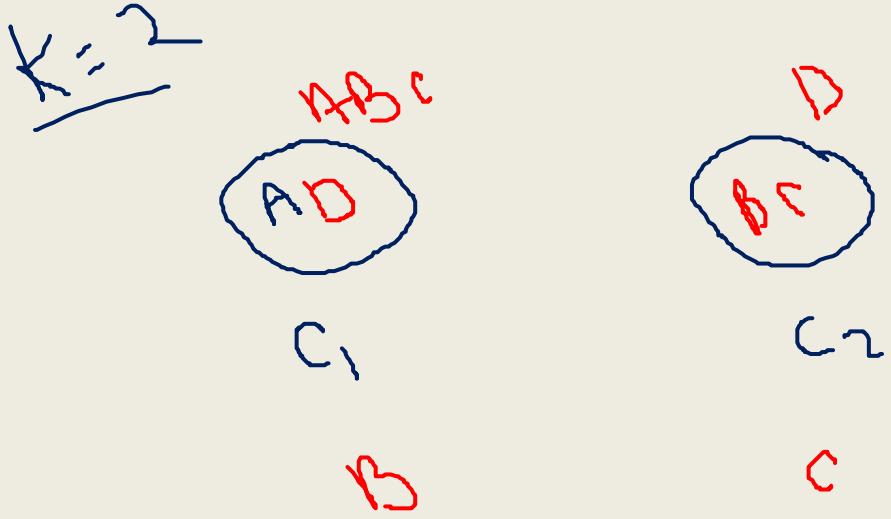
- ✓ AB = Average of A, B
- ✓ CD = Average of C,D

Centroid Table

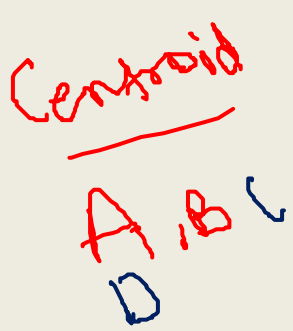
	X1	X2
AB	$\frac{1+6}{2} = 4$	2
CD	$\frac{1+3}{2} = 2$	1

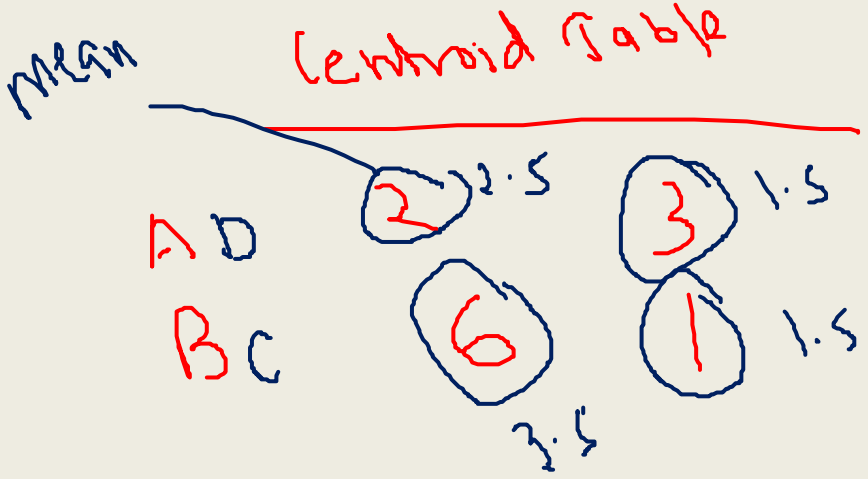
	X1	X2
A	2	3
B	6	1
C	1	2
D	3	0

Centroid = cluster mean



	X1	X2
A	2	3
B	6	1
C	1	2
D	3	0





- Calculate squared Euclidean distance between all data points to the centroids AB, CD. For example distance between A(2,3) and AB (4,2) can be given by $s = (2–4)^2 + (3–2)^2$.

	A	B	C	D
AB	5			
CD	4			

$= (2-2)^2 + (3-3)^2$
$= 0 + (-2)^2$
$= 4$

$= (4-2)^2 + (2-3)^2$
$= 2^2 + (-1)^2$
$= 4 + 1 = 5$

Centroid Table

	X1	X2
AB	4	2
CD	2	1

Data Points

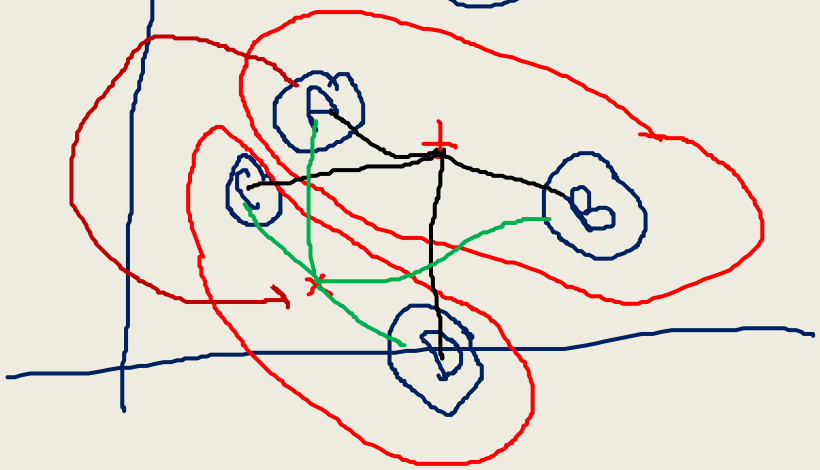
	X1	X2
A	2	3
B	6	1
C	1	2
D	3	0

Steps involved in K-Means Clustering

- The first step when using k-means clustering is to indicate the number of clusters (k) that will be generated in the final solution.
- The algorithm starts by randomly selecting k objects from the data set to serve as the initial centers for the clusters. The selected objects are also known as cluster means or centroids.
- Next, each of the remaining objects is assigned to it's closest centroid, where closest is defined using the Euclidean distance between the object and the cluster mean. This step is called "cluster assignment step".
- After the assignment step, the algorithm computes the new mean value of each cluster. The term cluster "centroid update" is used to design this step. Now that the centers have been recalculated, every observation is checked again to see if it might be closer to a different cluster. All the objects are reassigned again using the updated cluster means.
- The cluster assignment and centroid update steps are iteratively repeated until the cluster assignments stop changing (i.e until convergence is achieved). That is, the clusters formed in the current iteration are the same as those obtained in the previous iteration.

1. Centroid
2. Centroid Table
3. Distance

	A	B	C	D
AB	5		5	9	5
CD	4		16	2	2



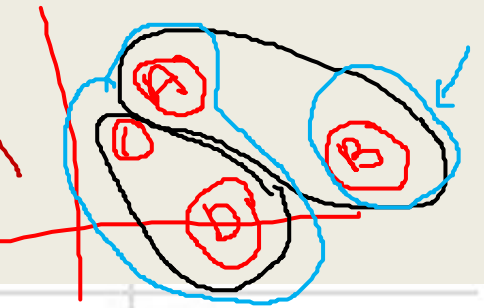
	X1	X2
A	2	3
B	6	1
C	1	2
D	3	0

- If we observe in the fig, the highlighted distance between (A, CD) is 4 and is less compared to (AB, A) which is 5. Since point A is close to the CD we can move A to CD cluster.

- There are two clusters formed so far, let recompute the centroids i.e, B, ACD similar to step 2.

- ACD = Average of A, C, D

- B = B



	X1	X2
A	2	3
B	6	1
C	1	2
D	3	0

	X1	X2
B	6	1
ACD	2	1.67

	A	B	C	D
AB	5	5	9	5
CD	4	16	2	2

- As we know K-Means is iterative procedure now we have to calculate the distance of all points (A, B, C, D) to new centroids (B, ACD).

	A	B	C	D
B				
ACD				

	X1	X2
A	2	3
B	6	1
C	1	2
D	3	0

	A	B	C	D
B	20	0	26	10
ACD	1.78	16.44	1.11	3.78

	X1	X2
B	6	1
ACD	2	1.67

- As we know K-Means is iterative procedure now we have to calculate the distance of all points (A, B, C, D) to new centroids (B, ACD).

	A	B	C	D
B	20	0		
ACD	1.76			

	X1	X2
A	2	3
B	6	1
C	1	2
D	3	0

	A	B	C	D
B	20	0	26	10
ACD	1.78	16.44	1.11	3.78

1.76

Updated Centroid table

	X1	X2
B	6	1
ACD	2	1.67

K = 2

- In the previous slide, we can see respective cluster values are minimum that A is too far from cluster B and near to cluster ACD. All data points are assigned to clusters (B, ACD) based on their minimum distance. The iterative procedure ends here. When no data point have been moved.

- To conclude, we have started with two centroids and end up with two clusters, K=2.

100 data points

K = 2 - to - 99

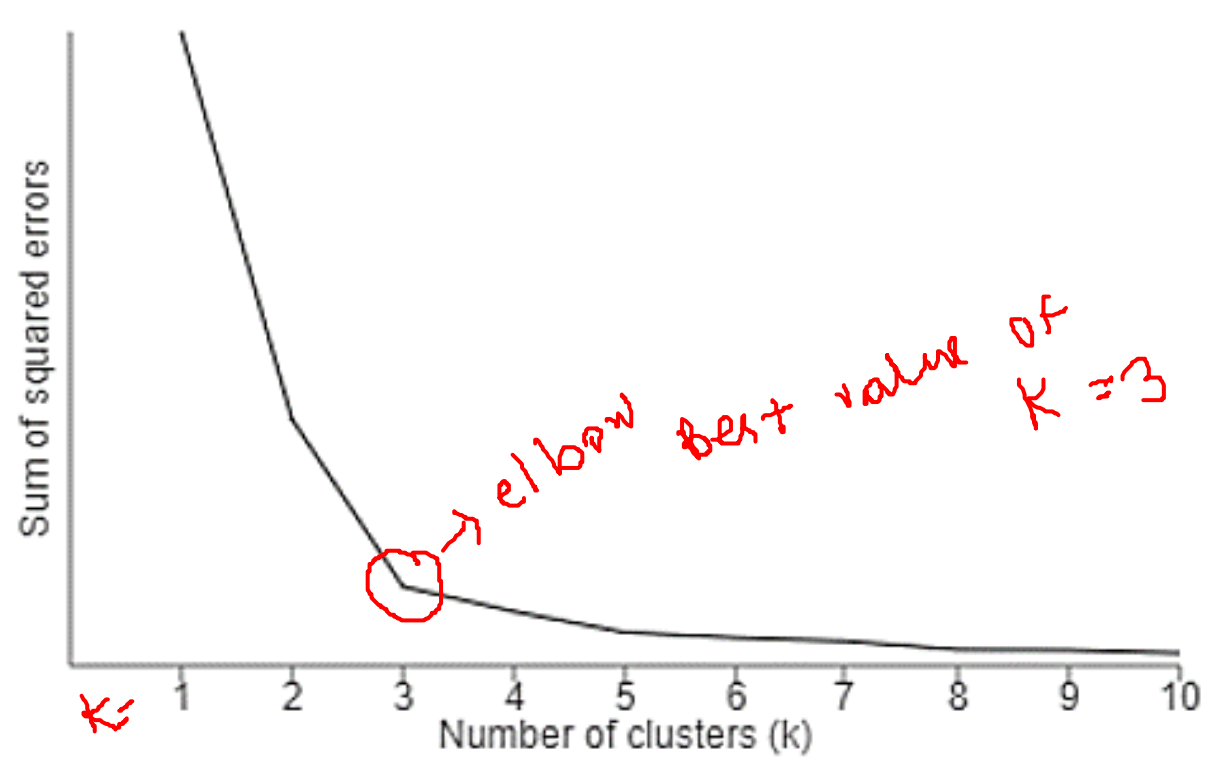
K value as optimum =

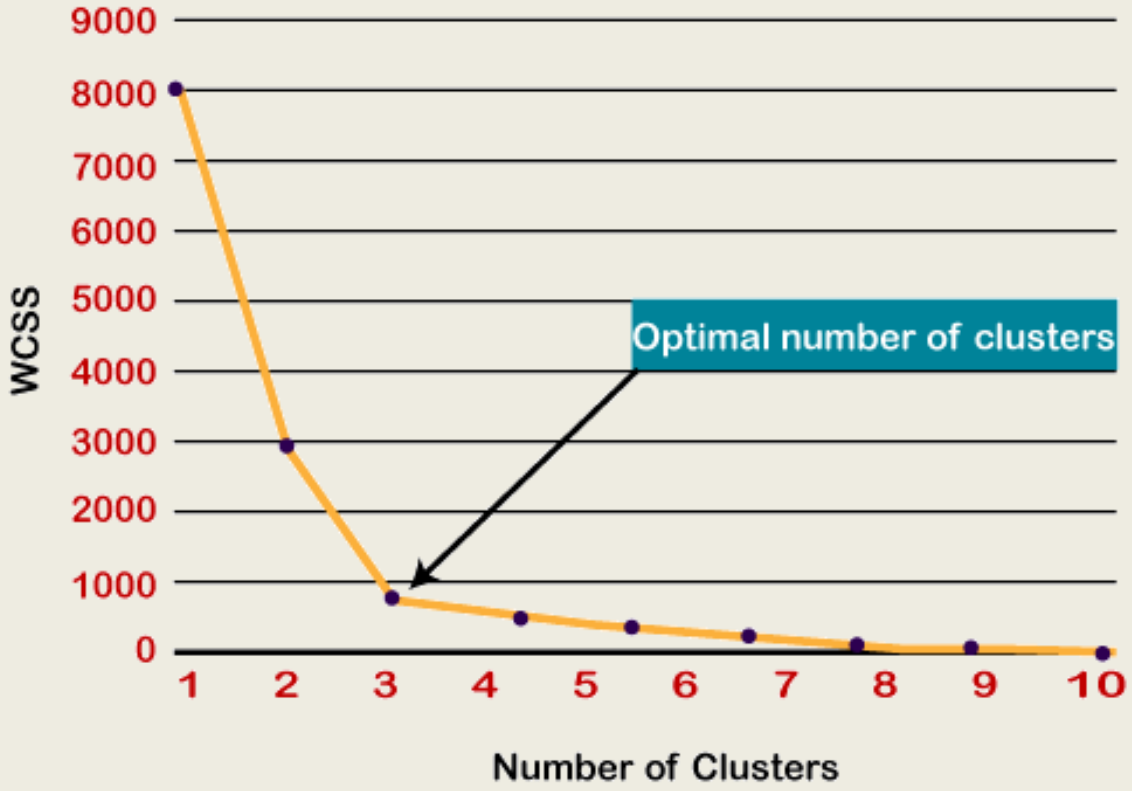
Choosing K

- One method of choosing value K is the elbow method. In this method we will run K-Means clustering for a range of K values lets say (K= 1 to 10) and calculate the Sum of Squared Error (SSE). SSE is calculated as the mean distance between data points and their cluster centroid.

Where minimisation of SSE is started

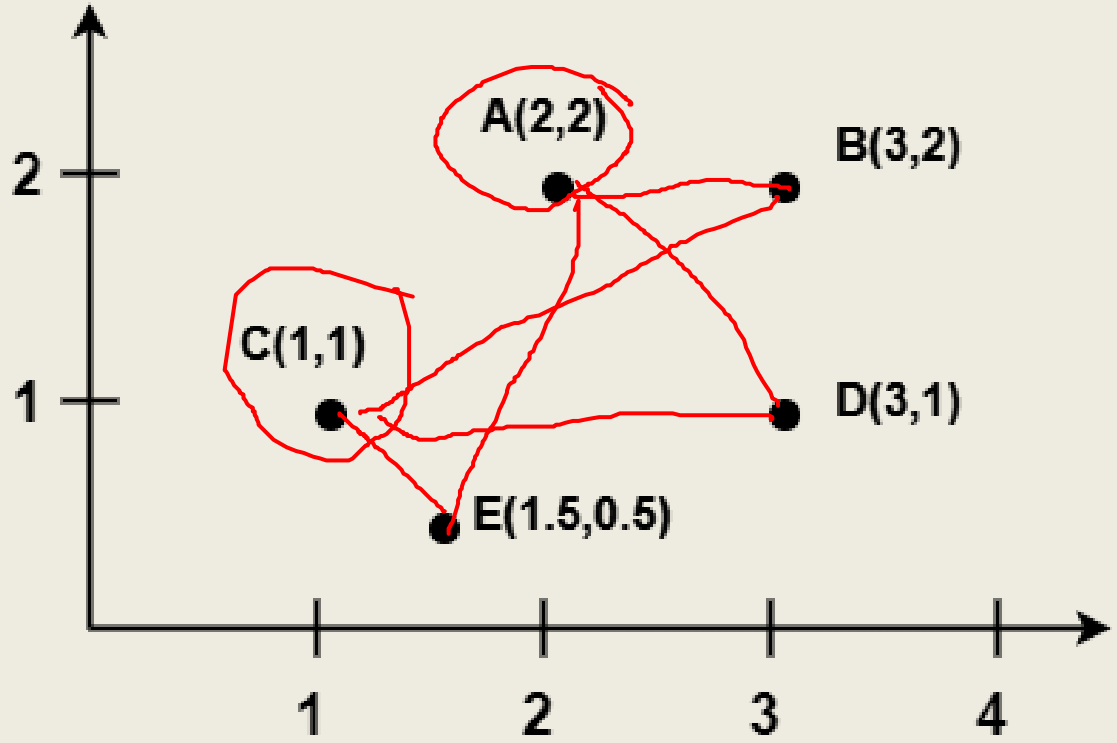
- Then plot a line chart for SSE values for each K, if the line chart looks like an arm then the elbow on the arm is the value of K that is the best.





Solve it

- Use K-Means Algorithm to create two clusters-



Assume A(2, 2) and C(1, 1) are centers of the two clusters.

K = 2

VIT®

BHOPAL

www.vitbhopal.ac.in