ML

LECTURE-16

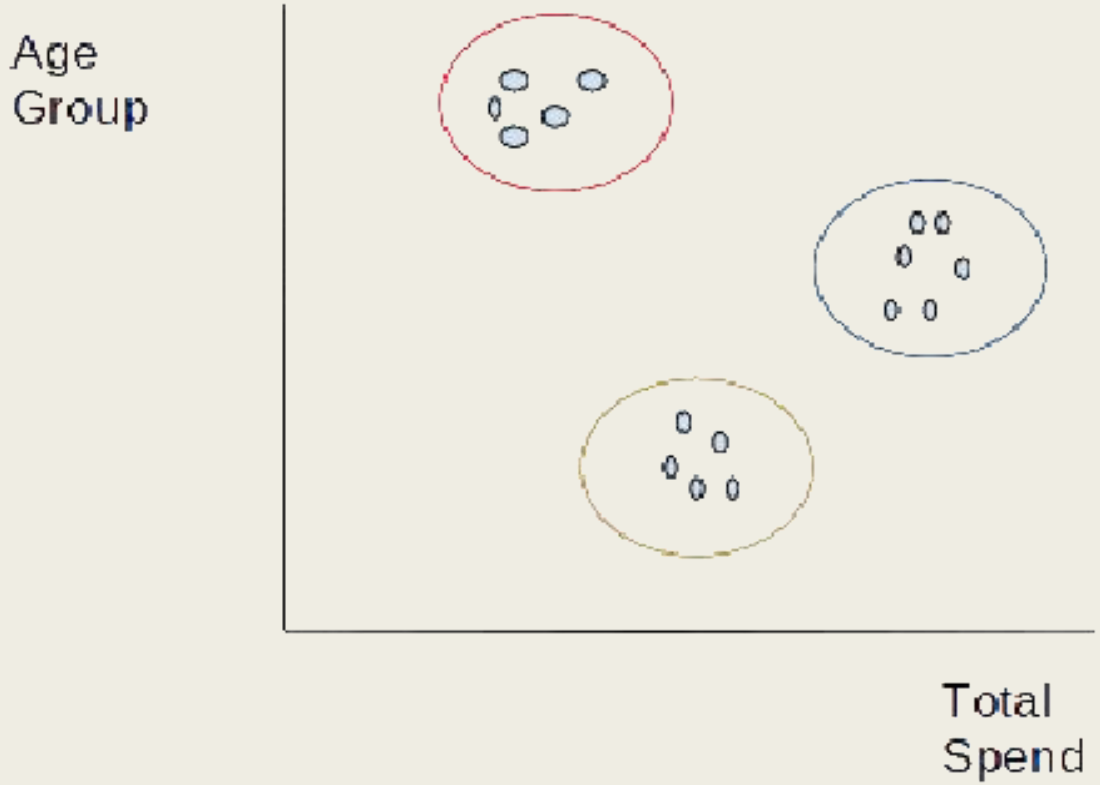
BY
Dr. Ramesh Kumar Thakur
Assistant Professor (II)
School Of Computer Engineering

Clustering

- Clustering was introduced in 1932 by H.E. Driver and A.L.Kroeber in their paper on “Quantitative expression of cultural relationship”.
- Since then this technique has taken a big leap and has been used to discover the unknown in a number of application areas eg. Healthcare.
- Clustering is a type of **unsupervised learning where the references need to be drawn from unlabelled datasets.**
- Generally, it is used to capture meaningful structure, underlying processes, and grouping inherent in a dataset.
- In clustering, the task is to **divide the population into several groups in such a way that the data points in the same groups are more similar to each other than the data points in other groups.**
- In short, it is a collection of objects based on their similarities and dissimilarities.
- There are many clustering algorithms grouped into different cluster models. Before choosing any algorithm for a use case, it is important to get familiar with the **cluster models.**
- One more thing which should be considered while choosing any clustering algorithm is the **size of your dataset.**
- Datasets can contain millions of records and not all algorithms scale efficiently. **K-Means is one of the most popular algorithms** and it is also scale-efficient as it has a **complexity of O(n).**

Clustering Example

- Let's take an example, imagine you work in a Walmart Store as a manager and would like to better understand your customers to scale up your business by using new and improved marketing strategies.
- It is difficult to segment your customers manually. You have some data that contains their age and purchase history, here clustering can help to group customers based on their spending.
- Once the customer segmentation will be done, you can define different marketing strategies for each of the groups as per target audiences.



K - means Clustering

- K-means is a **centroid-based clustering algorithm**, where we calculate the **distance between each data point and a centroid to assign it to a cluster.**
- The goal is to identify the K number of groups in the dataset.
- It is an iterative process of assigning each data point to the groups and slowly data points get clustered based on similar features.
- The objective is to minimize the sum of distances between the data points and the cluster centroid, to identify the correct group each data point should belong to.
- Here, we **divide a data space into K clusters and assign a mean value to each.**
- The **data points are placed in the clusters closest to the mean value of that cluster.**
- There are several distance metrics available that can be used to calculate the distance but **euclidean distance is most commonly used.**

K - means Clustering Steps

- **1. Choosing the number of clusters :-** The first step is to define the K number of clusters in which we will group the data.
- **2. Initializing centroids :-** Centroid is the center of a cluster but initially, the exact center of data points will be unknown so, we select random data points and define them as centroids for each cluster.
- **3. Assign data points to the nearest cluster :-** Now that centroids are initialized, the next step is to assign data points X_n to their closest cluster centroid C_k. In this step, we will first calculate the distance between data point X and centroid C using Euclidean Distance metric.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- And then choose the cluster for data points where the distance between the data point and the centroid is minimum.
- **4. Re-initialize centroids :-** Next, we will re-initialize the centroids by calculating the average of all data points of that cluster.

$$Ci = \frac{1}{|Ni|} \sum xi$$

- **5. Repeat steps 3 and 4 :-** We will keep repeating steps 3 and 4 until we have optimal centroids and the assignments of data points to correct clusters are not changing anymore.

K - means Clustering Example Problem

- Use k-means clustering algorithm to divide the following data into two clusters and also compute the the representative data points for the clusters.

x1 1 2 2 3 4 5

x2 1 1 3 2 3 5

- Solution:-

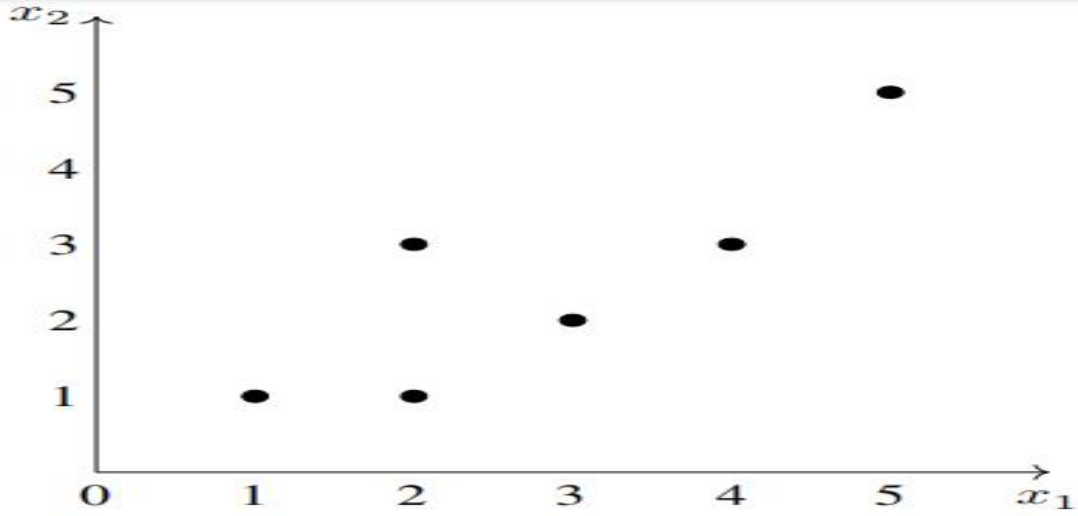


Figure 13.1: Scatter diagram of data in Table 13.1

1. In the problem, the required number of clusters is 2 and we take $k = 2$.
2. We choose two points arbitrarily as the initial cluster centres. Let us choose arbitrarily (see Figure 13.2)

$$\vec{v}_1 = (2, 1), \quad \vec{v}_2 = (2, 3).$$

3. We compute the distances of the given data points from the cluster centers.

K - means Clustering Example Problem

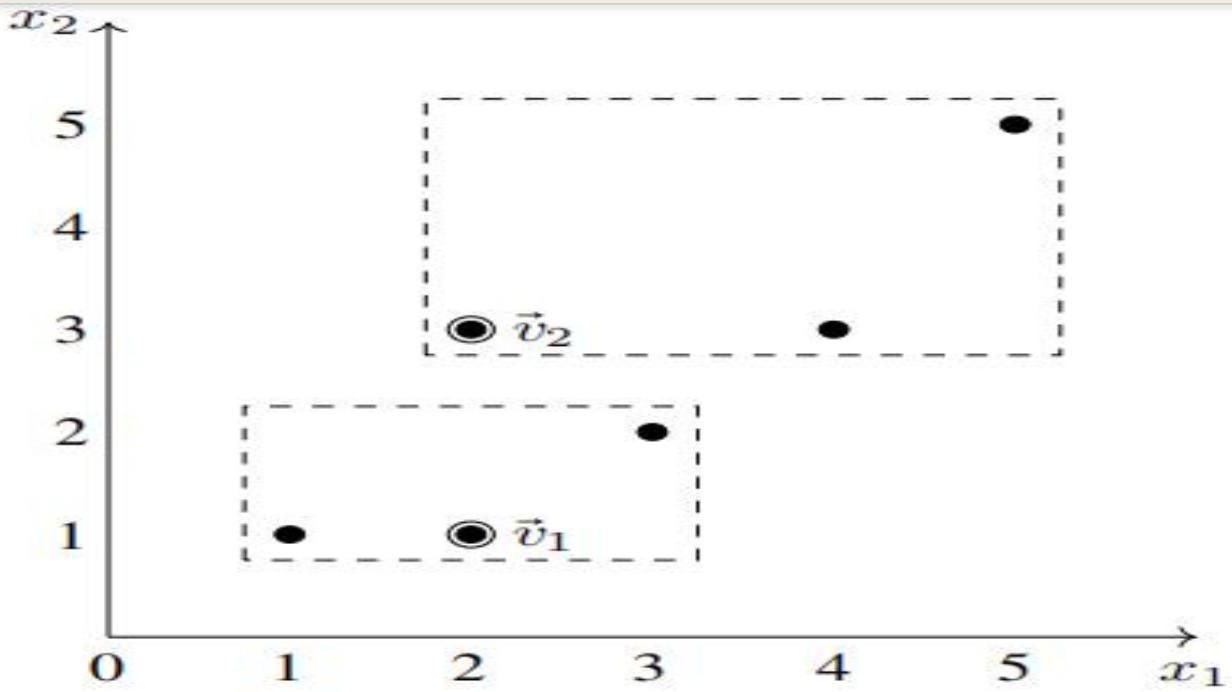


Figure 13.2: Initial choice of cluster centres and the resulting clusters

$\vec{x}_i$	Data point	Distance from $\vec{v}_1 = (2, 1)$	Distance from $\vec{v}_2 = (2, 3)$	Minimum distance	Assigned center
$\vec{x}_1$	$(1, 1)$	1	2.24	1	$\vec{v}_1$
$\vec{x}_2$	$(2, 1)$	0	2	0	$\vec{v}_1$
$\vec{x}_3$	$(2, 3)$	2	0	0	$\vec{v}_2$
$\vec{x}_4$	$(3, 2)$	1.41	1.41	1.41	$\vec{v}_1$
$\vec{x}_5$	$(4, 3)$	2.82	2	2	$\vec{v}_2$
$\vec{x}_6$	$(5, 5)$	5	3.61	3.61	$\vec{v}_2$

K - means Clustering Example Problem

(The distances of $\vec{x}_4$ from $\vec{v}_1$ and $\vec{v}_2$ are equal. We have assigned $\vec{v}_1$ to $\vec{x}_4$ arbitrarily.)

This divides the data into two clusters as follows (see Figure 13.2):

Cluster 1: $\{\vec{x}_1, \vec{x}_2, \vec{x}_4\}$ represented by $\vec{v}_1$

Number of data points in Cluster 1: $c_1 = 3$.

Cluster 2 : $\{\vec{x}_3, \vec{x}_5, \vec{x}_6\}$ represented by $\vec{v}_2$

Number of data points in Cluster 2: $c_2 = 3$.

4. The cluster centres are recalculated as follows:

$$\begin{aligned}
\vec{v}_1 &= \frac{1}{c_1}(\vec{x}_1 + \vec{x}_2 + \vec{x}_4) \\
&= \frac{1}{3}(\vec{x}_1 + \vec{x}_2 + \vec{x}_4) \\
&= (2.00, 1.33) \\
\vec{v}_2 &= \frac{1}{c_2}(\vec{x}_3 + \vec{x}_5 + \vec{x}_6) \\
&= \frac{1}{3}(\vec{x}_3 + \vec{x}_5 + \vec{x}_6) \\
&= (3.67, 3.67)
\end{aligned}$$

5. We compute the distances of the given data points from the new cluster centers.

K - means Clustering Example Problem

$\vec{x}_i$	Data point	Distance from $\vec{v}_1 = (2,1)$	Distance from $\vec{v}_2 = (2,3)$	Minimum distance	Assigned center
$\vec{x}_1$	$(1,1)$	1.05	3.77	1.05	$\vec{v}_1$
$\vec{x}_2$	$(2,1)$	0.33	3.14	0.33	$\vec{v}_1$
$\vec{x}_3$	$(2,3)$	1.67	1.80	1.67	$\vec{v}_1$
$\vec{x}_4$	$(3,2)$	1.20	1.80	1.20	$\vec{v}_1$
$\vec{x}_5$	$(4,3)$	2.60	0.75	0.75	$\vec{v}_2$
$\vec{x}_6$	$(5,5)$	4.74	1.89	1.89	$\vec{v}_2$

This divides the data into two clusters as follows (see Figure 13.4):

Cluster 1 : $\{\vec{x}_1, \vec{x}_2, \vec{x}_3, \vec{x}_4\}$ represented by $\vec{v}_1$

Number of data points in Cluster 1: $c_1 = 4$.

Cluster 2 : $\{\vec{x}_5, \vec{x}_6\}$ represented by $\vec{v}_2$

Number of data points in Cluster 1: $c_2 = 2$.

6. The cluster centres are recalculated as follows:

$$\vec{v}_1 == \frac{1}{c_1}(\vec{x}_1 + \vec{x}_2 + \vec{x}_3 + \vec{x}_4)$$
$$= \frac{1}{4}(\vec{x}_1 + \vec{x}_2 + \vec{x}_3 + \vec{x}_4)$$
$$= (2.00, 1.33)$$
$$\vec{v}_2 = \frac{1}{2}(\vec{x}_5 + \vec{x}_6) = (3.67, 3.67)$$

K - means Clustering Example Problem

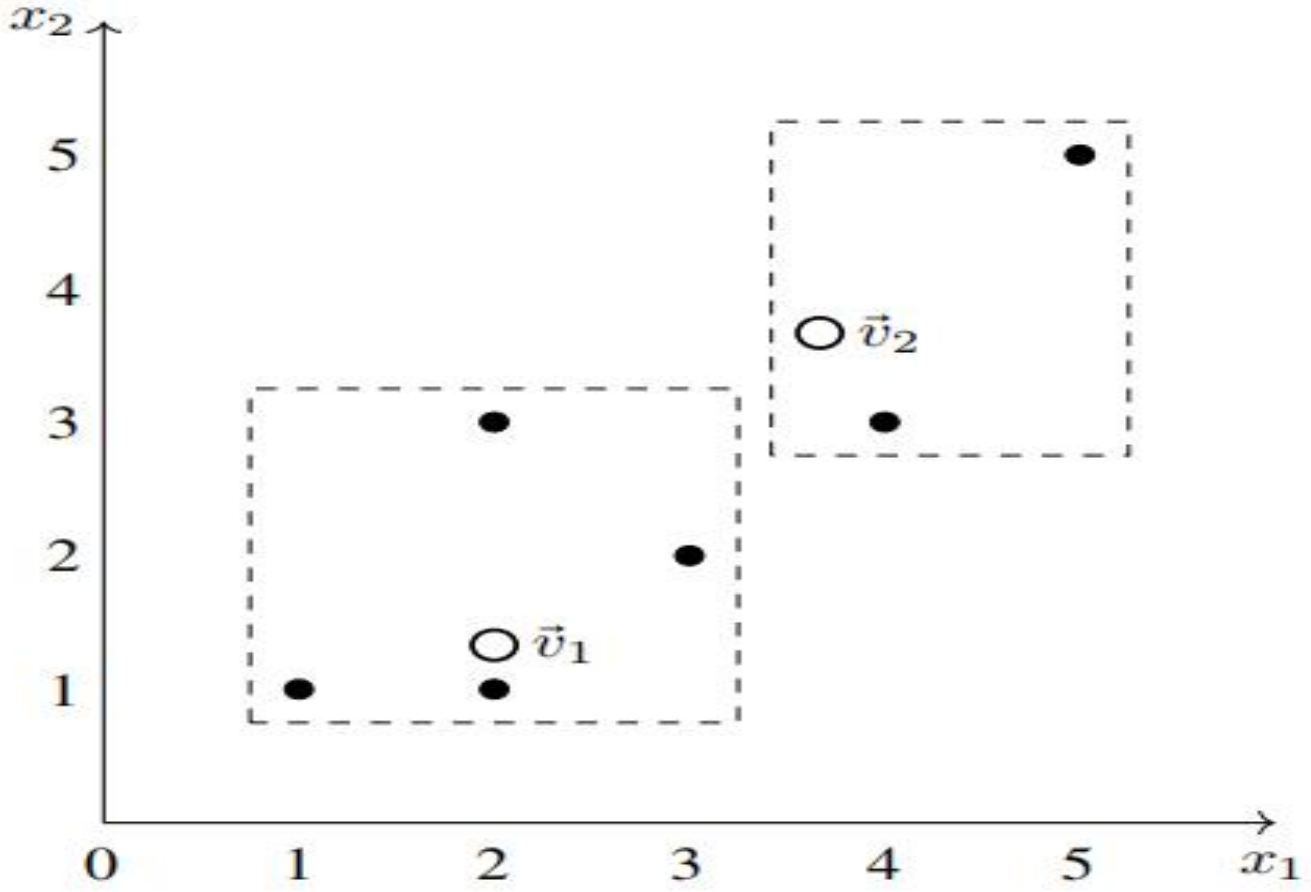


Figure 13.3: Cluster centres after first iteration and the corresponding clusters

7. We compute the distances of the given data points from the new cluster centers.
   4.609772 3.905125 2.692582 2.500000 1.118034 1.118034

K - means Clustering Example Problem

$\vec{x}_i$	Data point	Distance from $\vec{v}_1 = (2,1)$	Distance from $\vec{v}_2 = (2,3)$	Minimum distance	Assigned center
$\vec{x}_1$	$(1,1)$	1.25	4.61	1.25	$\vec{v}_1$
$\vec{x}_2$	$(2,1)$	0.75	3.91	0.75	$\vec{v}_1$
$\vec{x}_3$	$(2,3)$	1.25	2.69	1.25	$\vec{v}_1$
$\vec{x}_4$	$(3,2)$	1.03	2.50	1.03	$\vec{v}_1$
$\vec{x}_5$	$(4,3)$	2.36	1.12	1.12	$\vec{v}_2$
$\vec{x}_6$	$(5,5)$	4.42	1.12	1.12	$\vec{v}_2$

This divides the data into two clusters as follows (see Figure **??**):

Cluster 1 : $\{\vec{x}_1, \vec{x}_2, \vec{x}_3, \vec{x}_4\}$ represented by $\vec{v}_1$

Number of data points in Cluster 1: $c_1 = 4$.

Cluster 2 : $\{\vec{x}_5, \vec{x}_6\}$ represented by $\vec{v}_2$

Number of data points in Cluster 1: $c_1 = 2$.

8. The cluster centres are recalculated as follows:

$$\begin{aligned}
\vec{v}_1 &= \frac{1}{c_1}(\vec{x}_1 + \vec{x}_2 + \vec{x}_3 + \vec{x}_4) \\
&= \frac{1}{4}(\vec{x}_1 + \vec{x}_2 + \vec{x}_3 + \vec{x}_4) \\
&= (2.00, 1.75) \\
\vec{v}_2 &= \frac{1}{c_2}(\vec{x}_5 + \vec{x}_6) \\
&= \frac{1}{2}(\vec{x}_5 + \vec{x}_6) \\
&= (4.00, 4.50)
\end{aligned}$$

K - means Clustering Example Problem

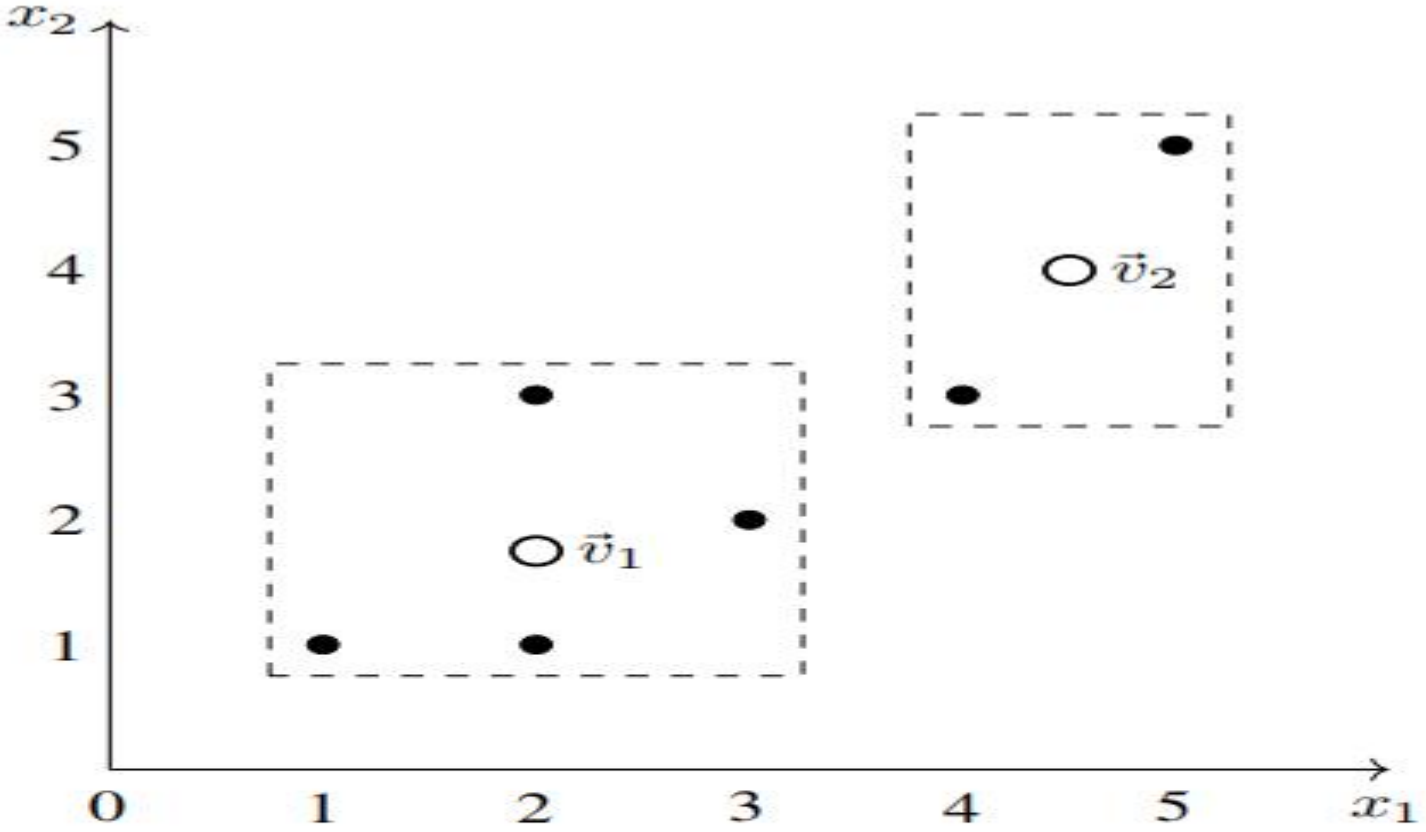


Figure 13.4: New cluster centres and the corresponding clusters

9. This divides the data into two clusters as follows (see Figure **??**):
   Cluster 1 : $\{\vec{x}_1, \vec{x}_2, \vec{x}_3, \vec{x}_4\}$ represented by $\vec{v}_1$
   Cluster 2 : $\{\vec{x}_5, \vec{x}_6\}$ represented by $\vec{v}_2$

K - means Clustering Example Problem

10. The cluster centres are recalculated as follows:

$$\vec{v}_1 = \frac{1}{4}(\vec{x}_1 + \vec{x}_2 + \vec{x}_3 + \vec{x}_4) = (2.00, 1.75)$$

$$\vec{v}_2 = \frac{1}{2}(\vec{x}_5 + \vec{x}_6) = (4.00, 4.50)$$

We note that these are identical to the cluster centres calculated in Step 8. So there will be no reassignment of data points to different clusters and hence the computations are stopped here.

11. Conclusion: The k means clustering algorithm with $k = 2$ applied to the dataset in Table 13.1 yields the following clusters and the associated cluster centres:

Cluster 1 : $\{\vec{x}_1, \vec{x}_2, \vec{x}_3, \vec{x}_4\}$ represented by $\vec{v}_1 = (2.00, 1.75)$

Cluster 2 : $\{\vec{x}_5, \vec{x}_6\}$ represented by $\vec{v}_2 = (2.00, 4.75)$

Disadvantages of K - means Clustering

- Even though the k-means algorithm is fast, robust and easy to understand, there are several disadvantages to the algorithm.
- The learning algorithm requires apriori specification of the number of cluster centers.
- The final cluster centres depend on the initial vi's.
- With different representation of data we get different results (data represented in form of cartesian co-ordinates and polar co-ordinates will give different results).
- Euclidean distance measures can unequally weight underlying factors.
- Randomly choosing of the initial cluster centres may not lead to a fruitful result.

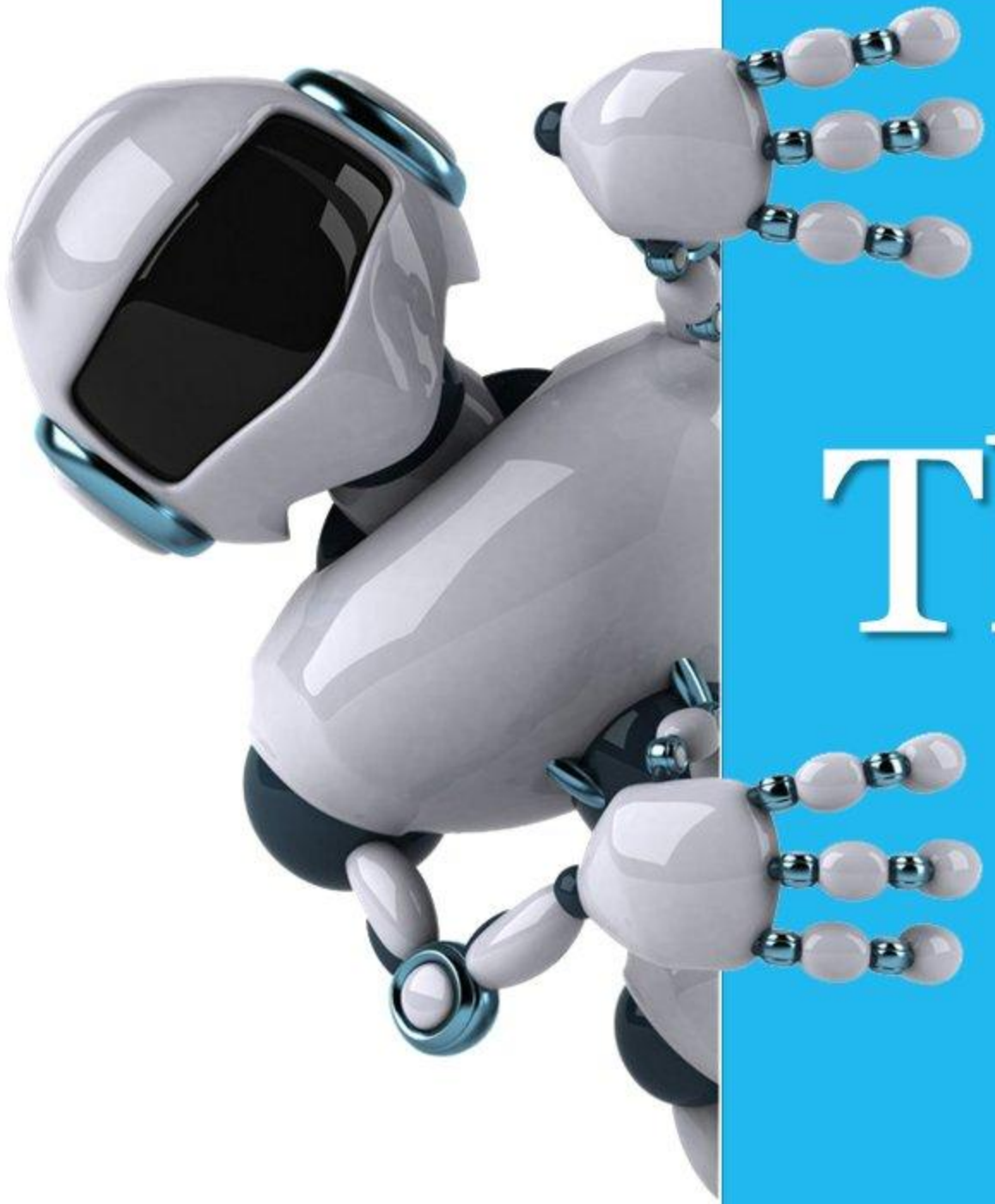
K - medians Clustering

- K-median clustering is similar to K-means. Only three differences are there:-
- 1. K-median clustering uses actual distance (which is also called the **L1 norm or the Manhattan or Taxicab distance**) to the center, instead of the square of the distance. So the distance formula is:-

$$d(p,q)=\sum_{i=1}^{n}|p_i-q_i|=\|.\|^1$$

- 2. K-median clustering choose the **median instead of the mean for the centers.**
- 3. K-median clustering need to optimize the following problem :-

$$argmin_C=\sum_{i=1}^{n}\sum_{i\in C_i}|x-median(C_i)|$$



Thank you