

CLUSTERING

- K-MEANS

What is clustering?

Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often according to some defined distance measure.

- ❑ Clustering is the process of finding meaningful groups in data.

For example, customers of a company can be grouped based on the purchase behavior. In recent years, clustering has even found its use in political elections

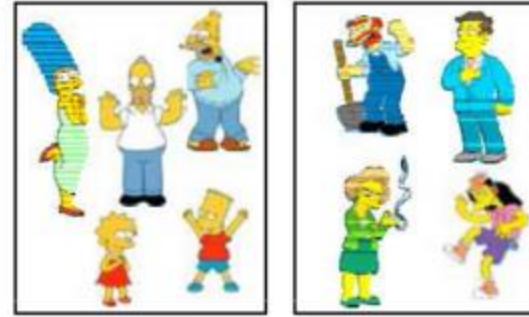
Types of clustering:

1. **Hierarchical algorithms**: these find successive clusters using previously established clusters.
 - a) **Agglomerative ("bottom-up")**: Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters.
 - b) **Divisive ("top-down")**: Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters.
2. **Partitional clustering**: Partitional algorithms determine all clusters at once. They include:
 - K-means and derivatives**
 - Fuzzy c-means clustering
 - QT clustering algorithm

Clustering algorithms

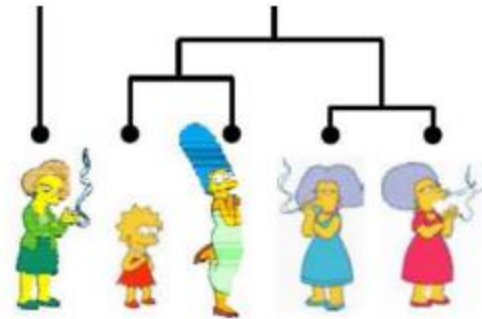
- Partition algorithms (Flat)

- K-means
- Mixture of Gaussian
- Spectral Clustering



- Hierarchical algorithms

- Bottom up – agglomerative
- Top down – divisive



Clustering examples

Image segmentation

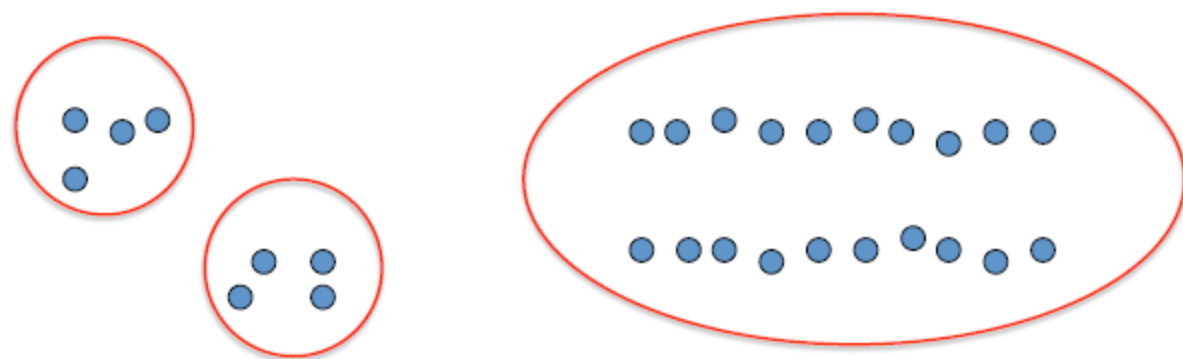
Goal: Break up the image into meaningful or perceptually similar regions



[Slide from James Hayes]

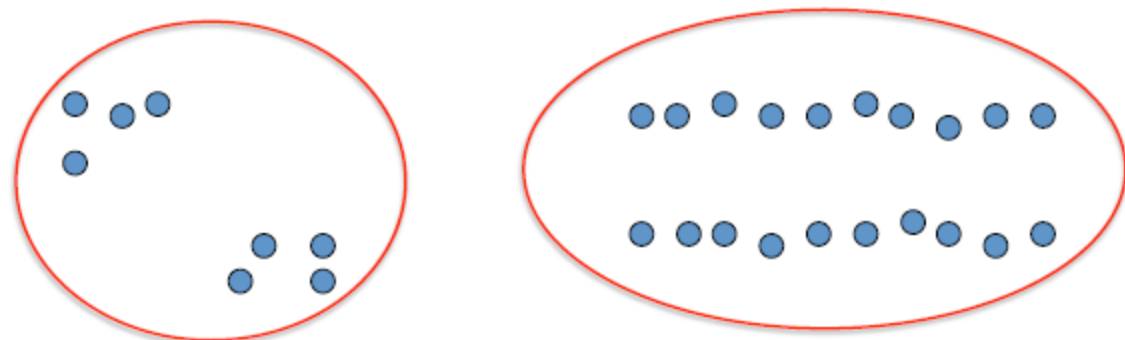
Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



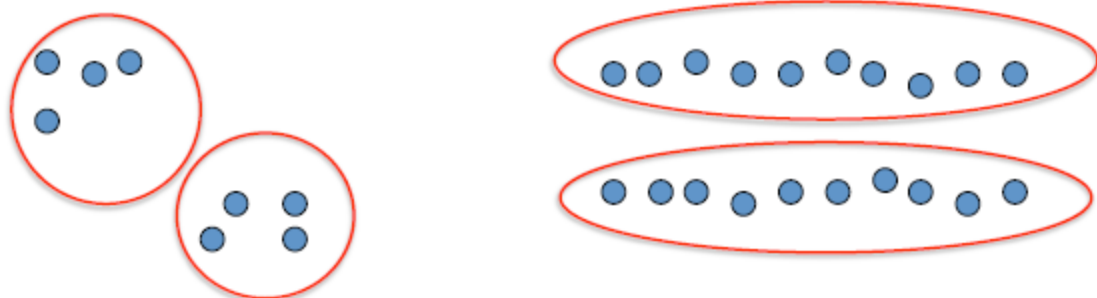
Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



- What could “similar” mean?
 - One option: small Euclidean distance (squared)
$$\text{dist}(\vec{x}, \vec{y}) = ||\vec{x} - \vec{y}||_2^2$$
 - Clustering results are crucially dependent on the measure of similarity (or distance) between “points” to be clustered

Common Distance measures:

Distance measure will determine how the *similarity* of two elements is calculated and it will influence the shape of the clusters.

They include:

1. The [Euclidean distance](#) (also called 2-norm distance) is given by:

$$d(x, y) = \sum_{i=1}^p |x_i - y_i|$$

2. The [Manhattan distance](#) (also called taxicab norm or 1-norm) is given by:

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

Common Distance measures:

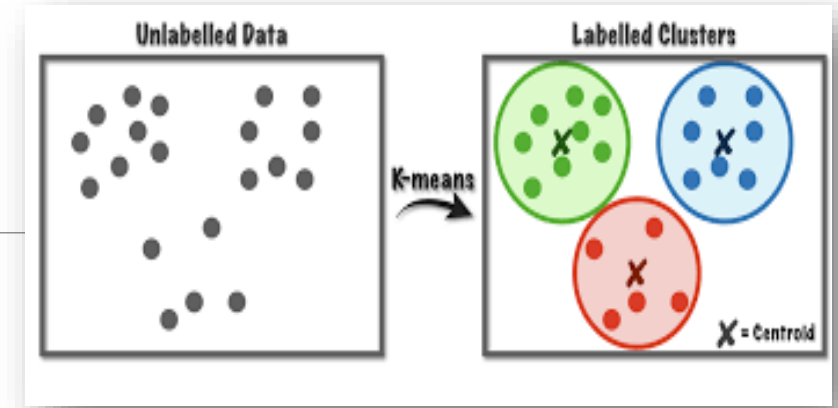
3. The maximum norm is given by:

$$d(x, y) = \max_{1 \leq i \leq p} |x_i - y_i|$$

4. The Mahalanobis distance corrects data for different scales and correlations in the variables.

5. Inner product space: The angle between two vectors can be used as a distance measure when clustering high dimensional data

6. Hamming distance (sometimes edit distance) measures the minimum number of substitutions required to change one member into another.



K-MEANS CLUSTERING

k-Means

k-means clustering is a prototype-based clustering method where the data set is divided into k clusters.

Objective: find a prototype data point for each cluster; all the data points are then assigned to the nearest prototype, which then forms a cluster

What is k-means clustering algorithm in machine learning?

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.

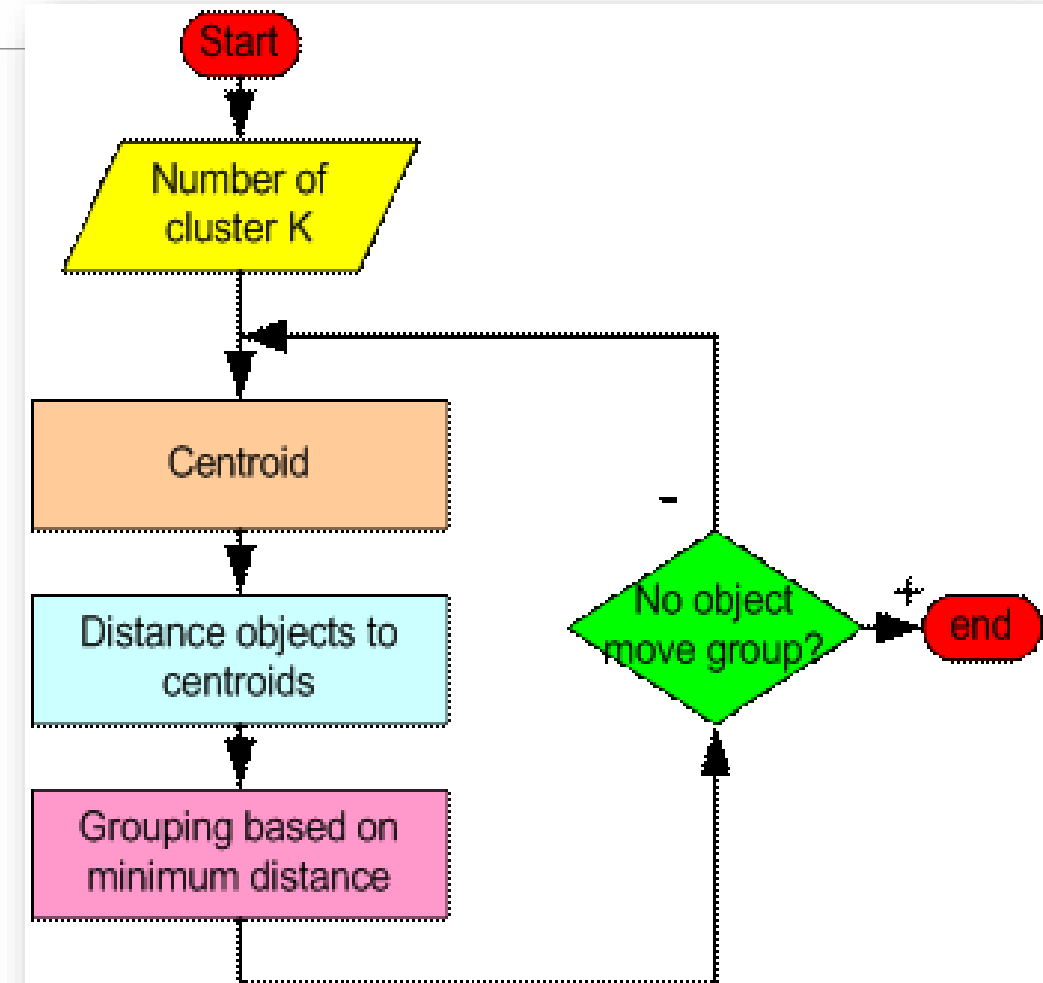


How the K-Mean Clustering algorithm works?

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.

Algorithm 1 k -means algorithm

- 1: Specify the number k of clusters to assign.
- 2: Randomly initialize k centroids.
- 3: **repeat**
- 4: **expectation:** Assign each point to its closest centroid.
- 5: **maximization:** Compute the new centroid (mean) of each cluster.
- 6: **until** The centroid positions do not change.



How the K-Mean Clustering algorithm works?

Step 1: Begin with a decision on the value of k = number of clusters .

Step 2: Put any initial partition that classifies the data into k clusters. You may assign the training samples randomly, or systematically as the following:

1. Take the first k training sample as single- element clusters
2. Assign each of the remaining $(N-k)$ training sample to the cluster with the nearest centroid. After each assignment, recomputed the centroid of the gaining cluster.

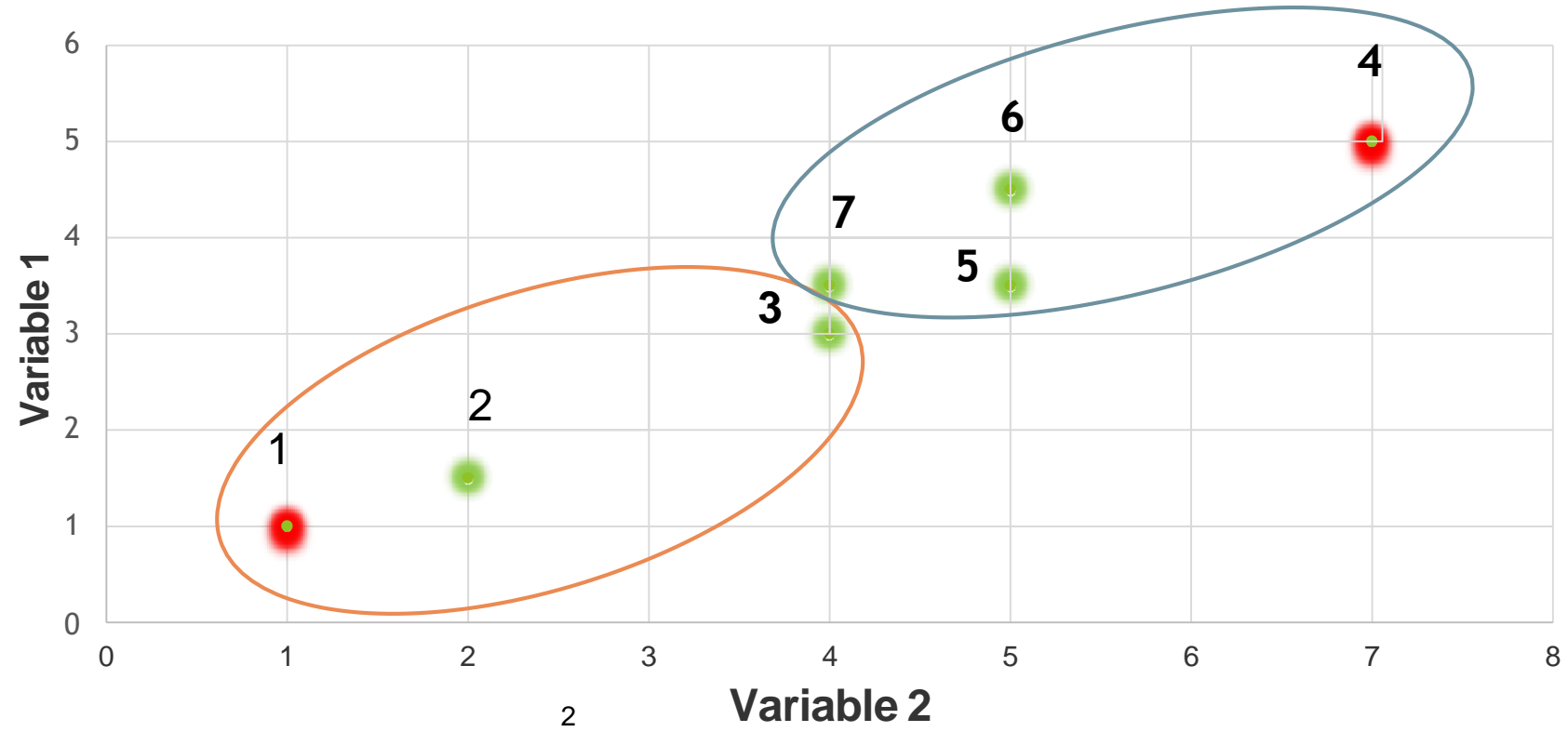
Step 3: Take each sample in sequence and compute its [distance](#) from the centroid of each of the clusters. If a sample is not currently in the cluster with the closest centroid, switch this sample to that cluster and update the centroid of the cluster gaining the new sample and the cluster losing the sample.

Step 4 . Repeat step 3 until convergence is achieved, that is until a pass through the training sample causes no new assignments.

A Simple example k-means (using $K=2$)

Individual	Variable 1	Variable 2
1	1	1
2	1.5	2
3	3	4
4	5	7
5	3.5	5
6	4.5	5
7	3.5	4.5

$K = 2$



Step 1:

Initialization: Randomly we choose following two centroids ($k=2$) for two clusters. In this case the 2 centroid are: $m1=(1.0,1.0)$ and $m2=(5.0,7.0)$.


Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

	Individual	Mean Vector
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

Step 2:

	Centroid 1	Centroid 2
1	$\sqrt{(1 - 1)^2 + (1 - 1)^2} = 0$	$\sqrt{(5 - 1)^2 + (7 - 1)^2} = 7.21$
2	$\sqrt{(1 - 1.5)^2 + (1 - 2)^2} = 1.12$	$\sqrt{(5 - 1.5)^2 + (7 - 2)^2} = 6.10$
3	$\sqrt{(1 - 3)^2 + (1 - 4)^2} = 3.61$	$\sqrt{(5 - 3)^2 + (7 - 4)^2} = 3.61$
4	$\sqrt{(1 - 5)^2 + (1 - 7)^2} = 7.21$	$\sqrt{(5 - 5)^2 + (7 - 7)^2} = 0$
5	$\sqrt{(1 - 3.5)^2 + (1 - 5)^2} = 4.72$	$\sqrt{(5 - 3.5)^2 + (7 - 5)^2} = 2.5$
6	$\sqrt{(1 - 4.5)^2 + (1 - 5)^2} = 5.31$	$\sqrt{(5 - 4.5)^2 + (7 - 5)^2} = 2.06$
7	$\sqrt{(1 - 3.5)^2 + (1 - 4.5)^2} = 4.30$	$\sqrt{(5 - 3.5)^2 + (7 - 4.5)^2} = 2.92$

Step 2:

- Thus, we obtain two clusters containing: 
 $\{1, 2, 3\}$ and $\{4, 5, 6, 7\}$.
- Their new centroids are:

$$\text{Group 1} = \left(\frac{1+1.5+3}{3} \right), \left(\frac{1+2+4}{3} \right) = (1.83, 2.33)$$

$$\text{Group 2} = \left(\frac{5+3.5+4.5+3.5}{4} \right), \left(\frac{7+5+5+4.5}{4} \right) = (4.12, 5.38)$$

Step 3:

- Now using these centroids we compute the Euclidean distance of each object, as shown in table.
- Therefore, the new clusters are:
 $\{1,2\}$ and $\{3,4,5,6,7\}$
- Next centroids are: $m1=(1.25,1.5)$ and $m2 = (3.9,5.1)$

Individual	Centroid 1	Centroid 2
1	1.57	5.38
2	0.47	4.28
3	2.04	1.78
4	5.84	1.84
5	3.15	0.73
6	3.78	0.54
7	2.74	1.08

Step 3:

	Centroid 1	Centroid 2
1	$\sqrt{(1.83 - 1)^2 + (2.33 - 1)^2} = 1.57$	$\sqrt{(4.12 - 1)^2 + (5.38 - 1)^2} = 5.38$
2	$\sqrt{(1.83 - 1.5)^2 + (2.33 - 2)^2} = 0.47$	$\sqrt{(4.12 - 1.5)^2 + (5.38 - 2)^2} = 4.29$
3	$\sqrt{(1.83 - 3)^2 + (2.33 - 4)^2} = 2.04$	$\sqrt{(4.12 - 3)^2 + (5.38 - 4)^2} = 1.78$
4	$\sqrt{(1.83 - 5)^2 + (2.33 - 7)^2} = 5.64$	$\sqrt{(4.12 - 5)^2 + (5.38 - 7)^2} = 1.84$
5	$\sqrt{(1.83 - 3.5)^2 + (2.33 - 5)^2} = 3.15$	$\sqrt{(4.12 - 3.5)^2 + (5.38 - 5)^2} = 0.73$
6	$\sqrt{(1.83 - 4.5)^2 + (2.33 - 5)^2} = 3.78$	$\sqrt{(4.12 - 4.5)^2 + (5.38 - 5)^2} = 0.54$
7	$\sqrt{(1.83 - 3.5)^2 + (2.33 - 4.5)^2} = 2.74$	$\sqrt{(4.12 - 3.5)^2 + (5.38 - 4.5)^2} = 1.08$

Therefore, the new clusters are:

{1,2} and {3,4,5,6,7}

$$\text{Group 1} = \left(\frac{1+1.5}{2} \right), \left(\frac{1+2}{2} \right) = (1.25, 1.5)$$

$$\text{Group 2} = \left(\frac{3+5+3.5+4.5+3.5}{5} \right), \left(\frac{4+7+5+5+4.5}{5} \right) = (3.9, 5.1)$$

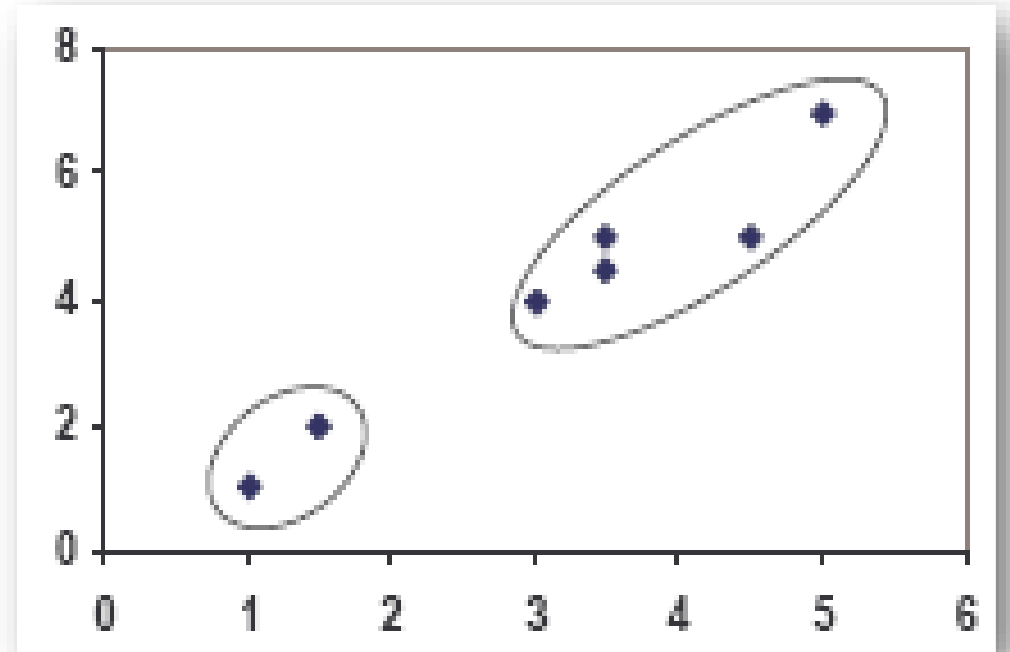
Step 4:

	Centroid 1	Centroid 2
1	$\sqrt{(1.25 - 1)^2 + (1.5 - 1)^2} = 0.58$	$\sqrt{(3.9 - 1)^2 + (5.1 - 1)^2} = 5.02$
2	$\sqrt{(1.25 - 1.5)^2 + (1.5 - 2)^2} = 0.56$	$\sqrt{(3.9 - 1.5)^2 + (5.1 - 2)^2} = 3.92$
3	$\sqrt{(1.25 - 3)^2 + (1.5 - 4)^2} = 3.05$	$\sqrt{(3.9 - 3)^2 + (5.1 - 4)^2} = 1.42$
4	$\sqrt{(1.25 - 5)^2 + (1.5 - 7)^2} = 6.66$	$\sqrt{(3.9 - 5)^2 + (5.1 - 7)^2} = 2.20$
5	$\sqrt{(1.25 - 3.5)^2 + (1.5 - 5)^2} = 4.16$	$\sqrt{(3.9 - 3.5)^2 + (5.1 - 5)^2} = 0.41$
6	$\sqrt{(1.25 - 4.5)^2 + (1.5 - 5)^2} = 4.78$	$\sqrt{(3.9 - 4.5)^2 + (5.1 - 5)^2} = 0.61$
7	$\sqrt{(1.25 - 3.5)^2 + (1.5 - 4.5)^2} = 3.75$	$\sqrt{(3.9 - 3.5)^2 + (5.1 - 4.5)^2} = 0.72$

Step 4:

- ▶ Therefore, there is no change in the cluster.
- ▶ Thus, the algorithm comes to a halt here and final result consist of 2 clusters $\{1,2\}$ and $\{3,4,5,6,7\}$.

PLOT



Pros and cons

Advantages of k-means

1. Relatively simple to implement.
2. Scales to large data sets.
3. Guarantees convergence.
4. Easily adapts to new examples.

Disadvantages of k-means

1. Choosing k manually.
2. Being dependent on initial values.
3. Scaling with number of dimensions.

Thank You