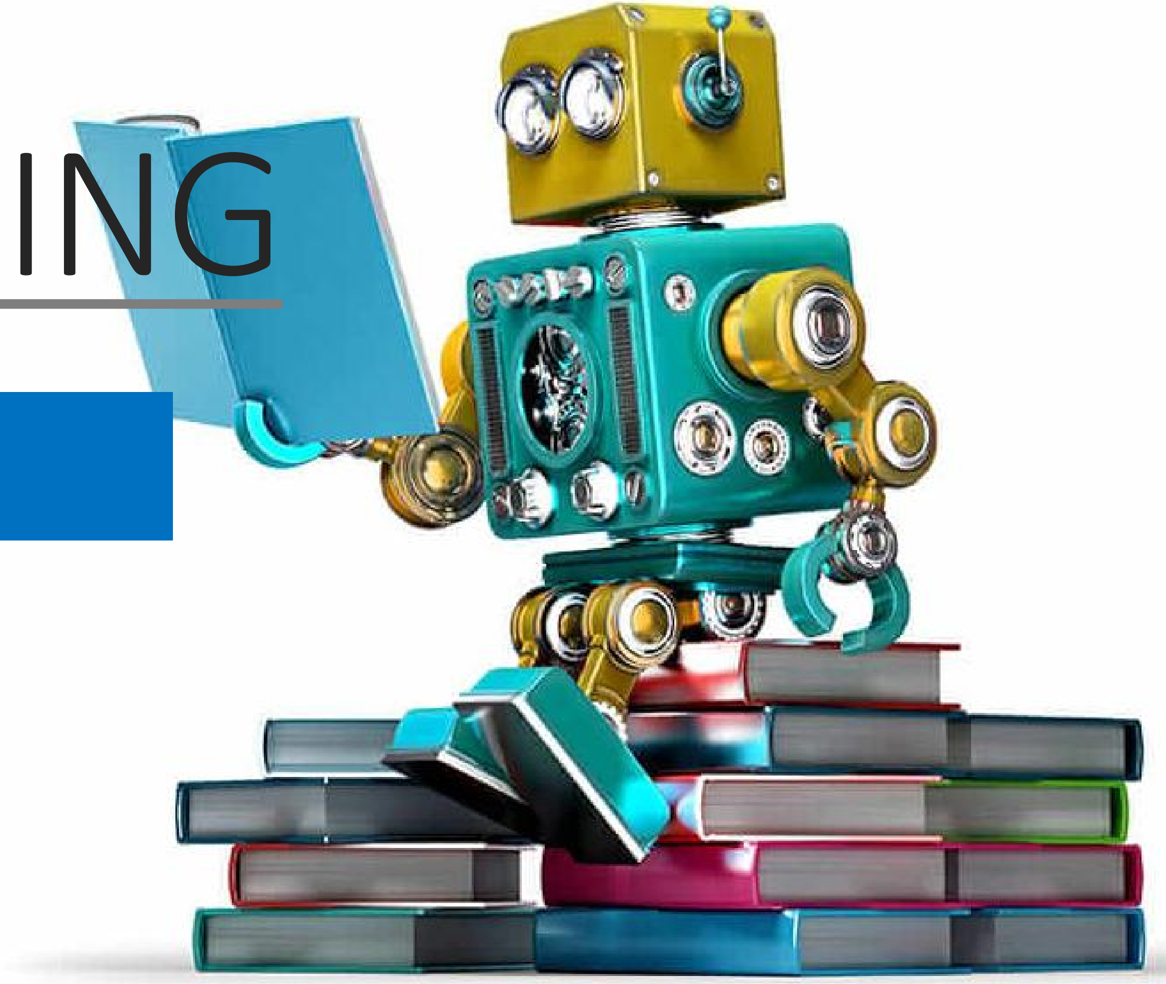
MACHINE LEARNING

LAB11 Clustering

贾艳红 Jana
Email:jiayh@mail.sustech.edu.cn

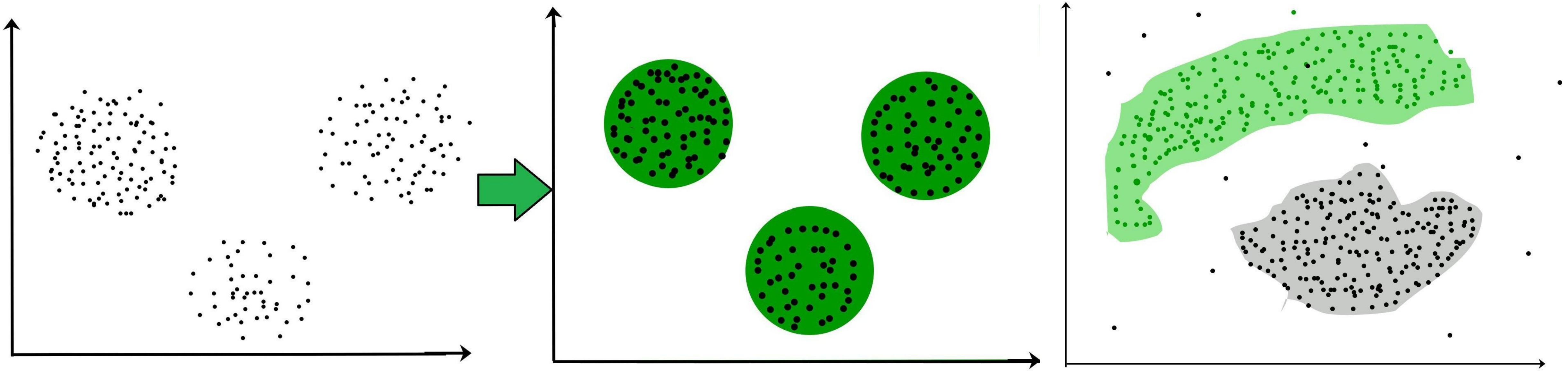


Outline

- Intro. to Clustering
- K-means Clustering
- example

Introduction to Clustering

- It is basically a type of **unsupervised** learning method .
- It is the task of dividing data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups (It is basically a collection of objects on the basis of similarity and dissimilarity between them).



Clustering Methods

- **Density-Based Methods: (DBSCAN , OPTICS)**

  Considering the clusters as the dense region having some similarity and different from the lower dense region of the space.

- **Hierarchical Based Methods: (CURE, BIRCH)**

  Forming a tree type structure based on the hierarchy. New clusters are formed using the previously formed one (Bottom up approach[Agglomerative] and top-dowm approach[Divisive])

- **Partitioning Methods: (K-means, GMM, CLARANS)**

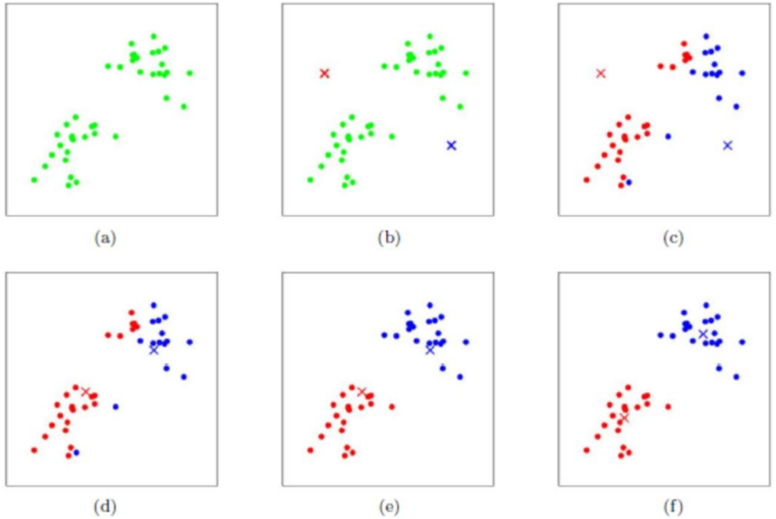
  Partitioning the objects into k clusters and each partition forms one cluster

- **Grid-based Methods: (STING, CLIQUE, wave cluster)**

  In this method the data space are formulated into a finite number of cells that form a grid-like structure.

K-means Clustering

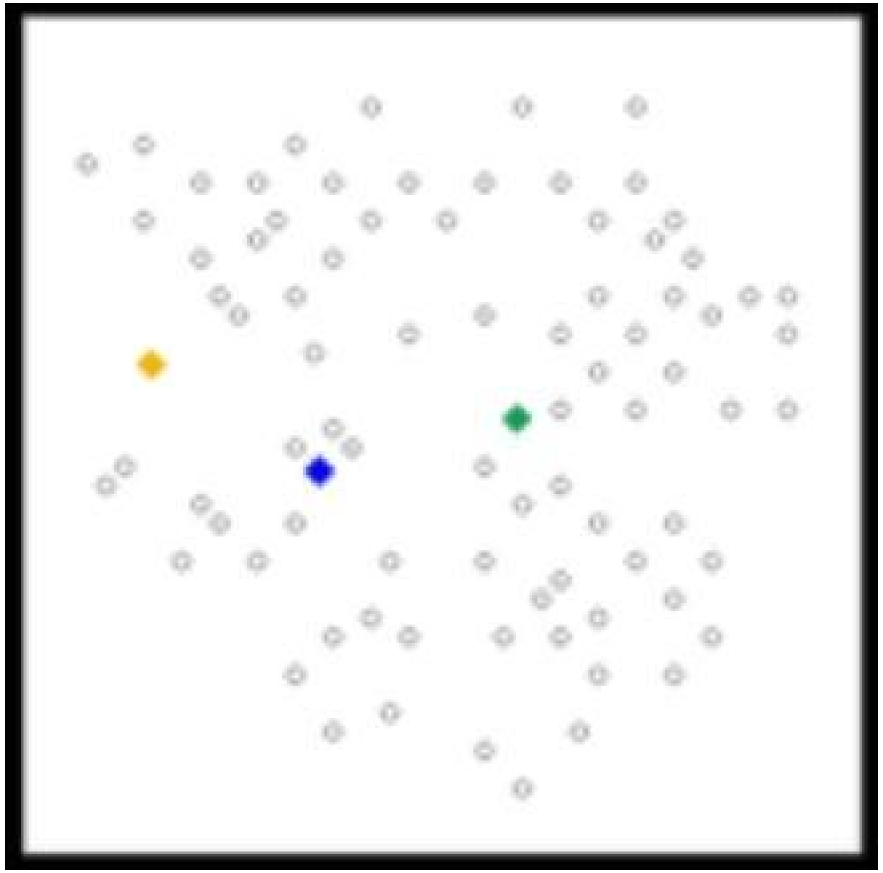
- k-means is a an unsupervised learning algorithm, and k is the number of clusters. k-means groups points into k clusters by minimizing the distances between points and their cluster's centroid. The centroid of a cluster is the mean of all the points in the cluster.
- To cluster data into k clusters, k-means follows four steps
  - Centroids initialization
  - Clusters initialization
  - Recomputation of centroids
  - Clusters reassignments



K-means Clustering

Centroids initialization

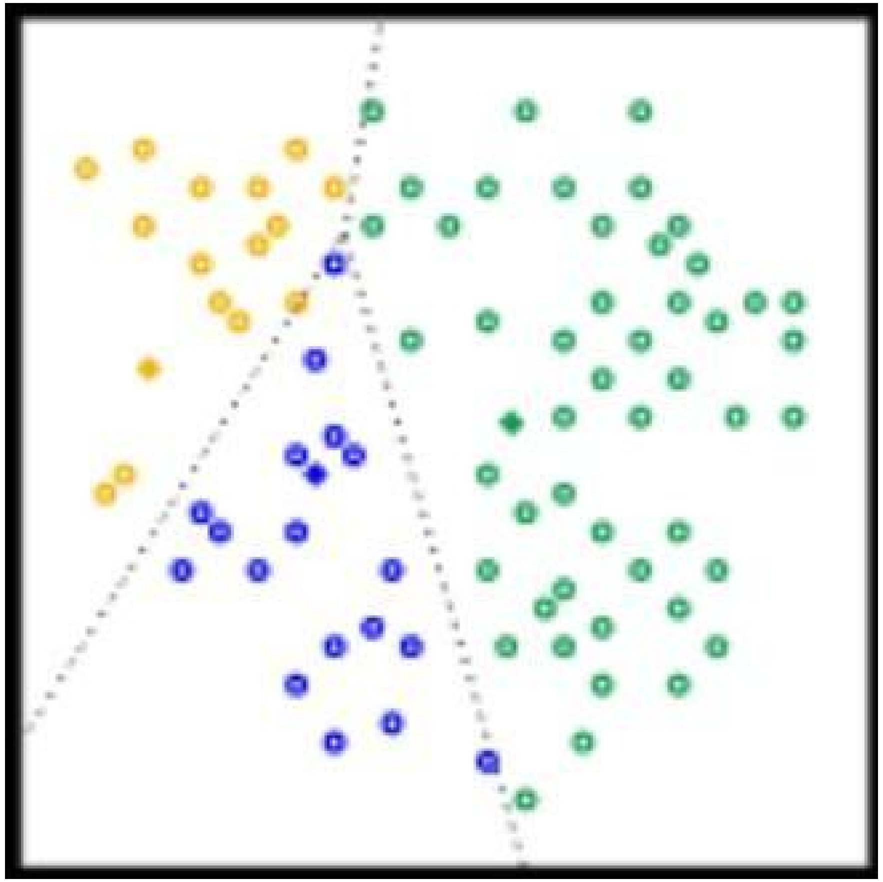
- Before running k-means, you must choose the number of clusters, k . Initially, start with a guess for k.
- The algorithm randomly chooses a centroid for each cluster. In our example, we choose a k of 3, and therefore the algorithm randomly picks 3 centroids.



K-means Clustering

Clusters initialization

- Assigns each point to the closest centroid to get initial clusters.



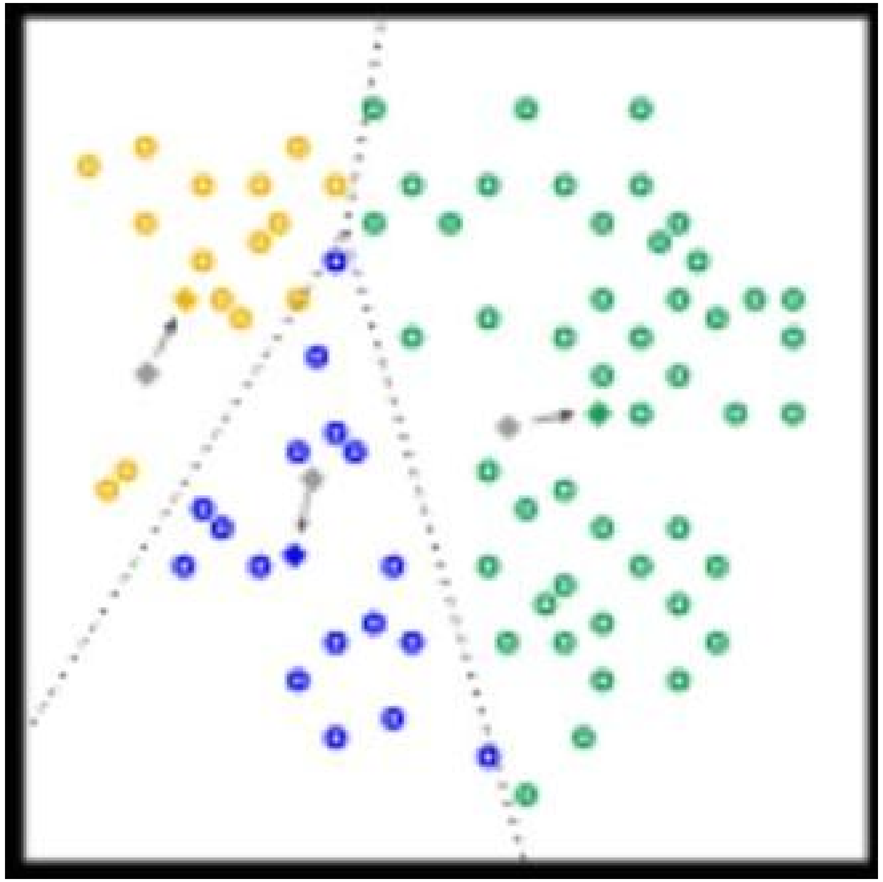
For every i, set

$$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2.$$

K-means Clustering

Recomputation of centroids

- For every cluster, the algorithm recomputes the centroid by taking the average of all points in the cluster.
- The changes in centroids are shown in Figure by arrows. Since the centroids change, the algorithm then re-assigns the points to the closest centroid.



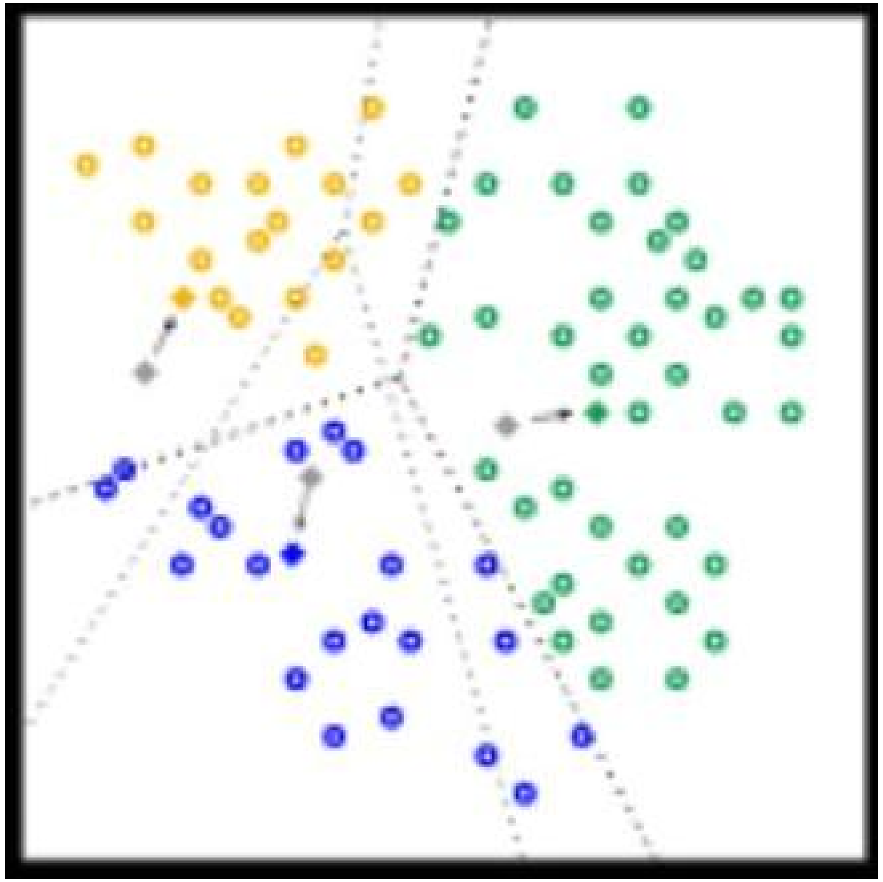
For each j, set

$$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}.$$

K-means Clustering

Clusters reassignment

- The algorithm repeats the calculation of centroids and assignment of points until points stop changing clusters. When clustering large datasets, you can stop the algorithm before reaching convergence, using other criteria instead.



K-means Clustering

Detail of Similarity Measure

- In clustering, the choice of Similarity Measure is important. To calculate the similarity between two examples, you need to combine all the feature data for those two examples into a single numeric value.
- In general, your similarity measure must directly correspond to the actual similarity. If your metric does not, then it isn’t encoding the necessary information

K-means Clustering

Mathematical proof

Given n examples assigned to k clusters, minimize the sum of distances of examples to their centroids. Where:

- $A_{nk} = 1$ when the nth example is assigned to the kth cluster, and 0 otherwise
- θ_k is the centroid of cluster k

We want to minimize the following expression:

$$\min_{A,\theta} \sum_{n=1}^{N} \sum_{k=1}^{K} A_{nk} ||\theta_n - x_n||^2$$

subject to:

$$A_{nk} \in \{0, 1\} \forall n, k$$

and

$$\sum_{k=1}^{K} A_{nk} = 1 \forall n$$

K-means Clustering

Mathematical proof

To minimize the expression with respect to the cluster centroids θ_k, take the derivative with respect to θ_k and equate it to 0.

$$f(\theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} A_{nk} ||\theta_k - x_n||^2$$

$$\frac{\partial f}{\partial \theta_k} = 2 \sum_{n=1}^{N} A_{nk} (\theta_k - x_n) = 0$$

$$\implies \sum_{n=1}^{N} A_{nk} \theta_k = \sum_{n=1}^{N} A_{nk} x_n$$

$$\theta_k \sum_{n=1}^{N} A_{nk} = \sum_{n=1}^{N} A_{nk} x_n$$

$$\theta_k = \frac{\sum_{n=1}^{N} A_{nk} x_n}{\sum_{n=1}^{N} A_{nk}}$$

K-means Clustering

Mathematical proof

The numerator is the sum of all example-centroid distances in the cluster. The denominator is the number of examples in the cluster. Thus, the cluster centroid θ_k is the average of example-centroid distances in the cluster. Hence proved.

Because the centroid positions are initially chosen at random, k-means can return significantly different results on successive runs. To solve this problem, run k-means multiple times and choose the result with the best quality metrics. (We'll describe quality metrics later in this course.) You'll need an advanced version of k-means to choose better initial centroid positions.

K-means Clustering

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

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)

$$d(m_1,2) = \sqrt{|1.0-1.5|^2 + |1.0-2.0|^2} = 1.12$$

$$d(m_2,2) = \sqrt{|5.0-1.5|^2 + |7.0-2.0|^2} = 6.10$$

K-means Clustering

Machine Learning

Individual	Centroid 1	Centroid 2
1	0	7.21
2 (1.5, 2.0)	1.12	6.10
3	3.61	3.61
4	7.21	0
5	4.72	2.5
6	5.31	2.06
7	4.30	2.92

Step 2:

Thus, we obtain two clusters containing:

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

Their new centroids are:

$$m_1 = \left(\frac{1}{3}(1.0+1.5+3.0), \frac{1}{3}(1.0+2.0+4.0)\right) = (1.83, 2.33)$$

$$m_2 = \left(\frac{1}{4}(5.0+3.5+4.5+3.5), \frac{1}{4}(7.0+5.0+5.0+4.5)\right)$$

$$= (4.12, 5.38)$$

K-means Clustering

Individual	Centroid 1	Centroid 2
1	1.57	5.38
2	0.47	4.28
3	2.04	1.78
4	5.64	1.84
5	3.15	0.73
6	3.78	0.54
7	2.74	1.08

Step 3:

- Now using these centroids of step2, 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)

K-means Clustering

Individual	Centroid 1	Centroid 2
1	0.56	5.02
2	0. 56	3.92
3	3.05	1.42
4	6.66	2.20
5	4.16	0.41
6	4.78	0.61
7	3.75	0.72

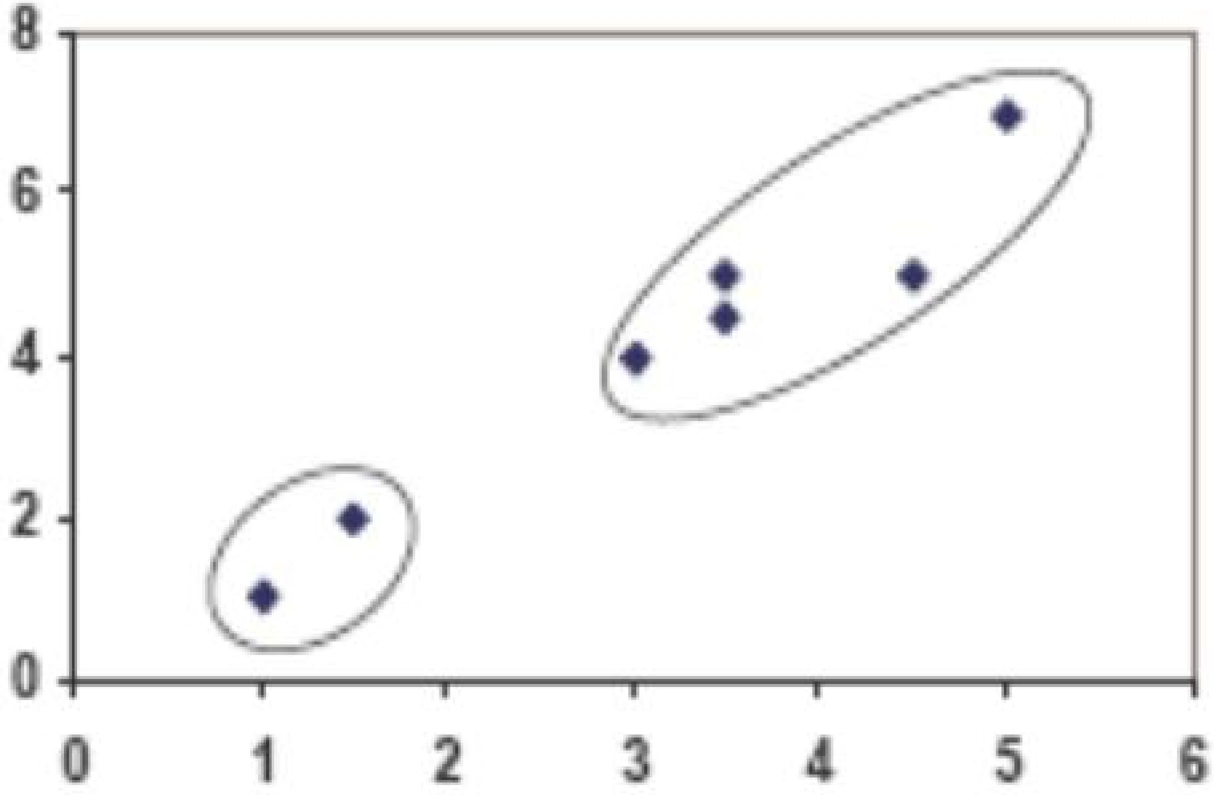
Step 4 :

- The clusters obtained are:
  {1,2} and {3,4,5,6,7}

- 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}.

K-means Clustering

Plot For k = 2

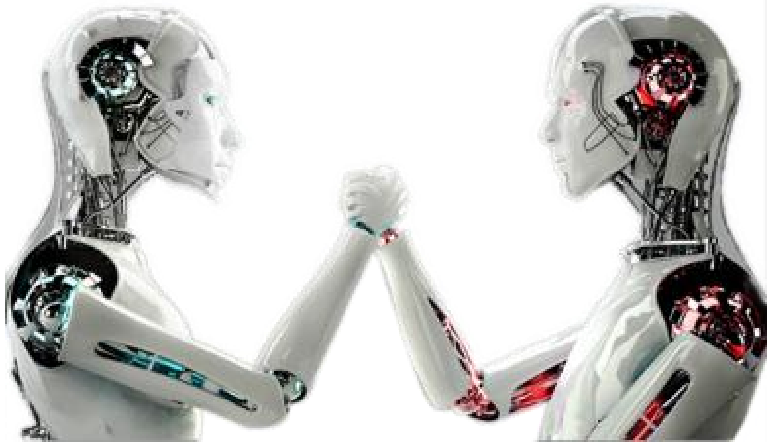


K-means Clustering

Algorithm 1 k-means algorithm

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

Lab Assignment



Lab Task

Complete the exercises and questions in the Lab11

Thanks

贾艳红 Jana
Email:jiayh@mail.sustech.edu.cn