COMP4388: **MACHINE LEARNING**

Unsupervised learning – K-Means

Dr. Radi Jarrar
Department of Computer Science
Birzeit University

جامعة بيرزيت
BIRZEIT UNIVERSITY

Unsupervised learning

- Supervised learning maps instances from an instance space x to an output space y using a set of labelled input instances.
- Unsupervised learning, on the other hand, is the task of inferring/describing hidden structures or patterns from unlabelled data.
- It is unsupervised as there is no class labels attached to the input instances.

Unsupervised learning

- Classification (or prediction or pattern detection) tasks result in a model that relates a set of input features to an output feature (i.e., target class). These models relate features to features and identify patterns within data.
- Clustering (unsupervised) creates new data by assigning a cluster label from the set of unlabelled input feature vectors.
- The label assigned to the cluster is inferred from the relationships within the data.

Clustering

- Clustering is an unsupervised machine learning task with the aim to divide data into clusters.
- Clustering entails grouping data with similar properties together.
- Used for Knowledge Discovery rather than prediction.
- Can be seen as *Learning a new labelling function from unlabelled data.*

Clustering (2)

- Clustering is based on the concept that similar observations should have similar properties to each other and should be different from the observations outside that cluster (group).
- Related elements are grouped together.

Clustering (3)

- Useful to exemplify diverse data into much smaller number of groups.
- This results in meaningful structures within data, which reduces complexity and provides insight into patterns of relationships among the groups.

Applications for Clustering

- Customer segmentation
  - Group customers with similar behaviours or similar demographics or even buying patterns for targeted marketing campaigns.
- Anomaly detection
  - Detecting illegal or unauthorised intrusions into computer networks by identifying patterns outside the known patterns.

Applications for Clustering (2)

- Social media
  - Clustering is used to determine communities of users. This is used, such as in Facebook, to refine advertising so that some ads go to certain groups of users.
- Data simplification
  - Large datasets can be simplified by grouping large number of features with similar values into smaller number of homogeneous categories.

K-means clustering

- The most common clustering algorithm.
- The basis of many more complicated clustering algorithms.
- The '*k*' in the name is similar to the 'k' in kNN classifier!
- It assigns each of n input examples to k clusters.
- k is the number of clusters (predefined; set by users).

K-means clustering (2)

- Goal: Minimise the differences within each cluster and maximise the differences between clusters.
- For each feature vector i, k-means assigns i to a cluster (initial guess) and then modifies the assignment to see changes in the homogeneity within clusters.

K-means clustering (3)

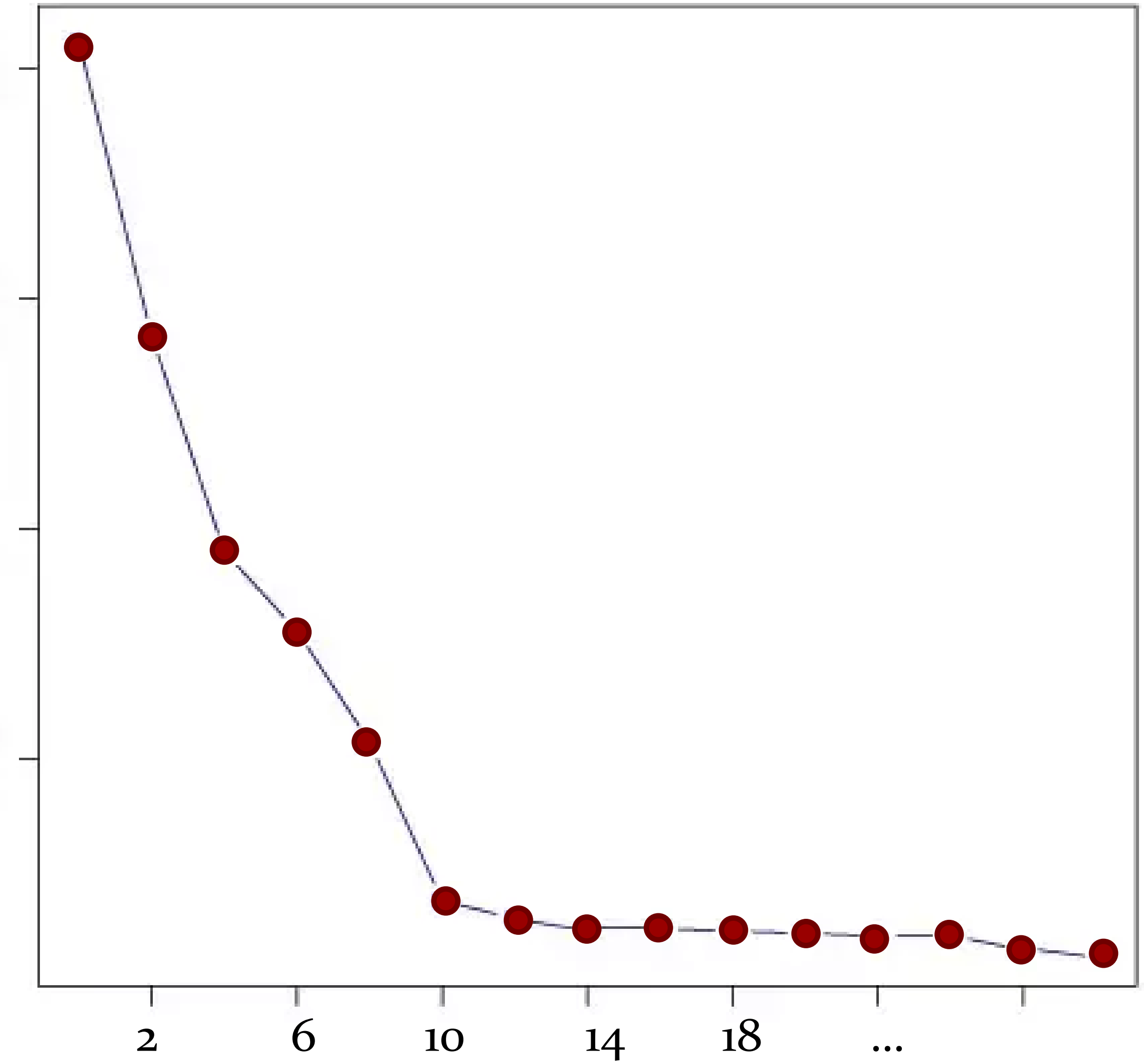
- Consists of two phases:
  - Assigns example to an initial set of k clusters.
  - Updates the assignments by adjusting the cluster boundaries of each cluster based on the examples fall into each cluster.
- This process is repeated until no improvement on the cluster.
- The process is stopped and clusters are finalised.

K-means clustering (4)

- Similarly to kNN, k-means deals with data in multidimensional feature space.
- The first step is to define the number of clusters
  - A quick rule-of-thump method is to select the sqrt(n/2) where n is number of data points in the dataset.
  - One method to decide on the number of clusters is the Elbow Method.
  - Select the number of clusters in which the Sum of Squared Error rate changes abruptly.

K-means clustering (4)

- Elbow Method



K-means clustering (5)

- Each cluster has a centroid (referred to as mean as well).
- The centroid is a point to which the distance of the objects will be calculated.
- Often, the points are chosen by selecting k random examples from the training set.

K-means clustering (6)

- Having chosen the initial cluster centres, new examples are assigned to the cluster centre that is nearest according to a distance function.
- For a new input feature vector, the distance is computed with the centroids of all clusters and the new instance is assigned to the cluster with the minimal distance.

K-means clustering (7)

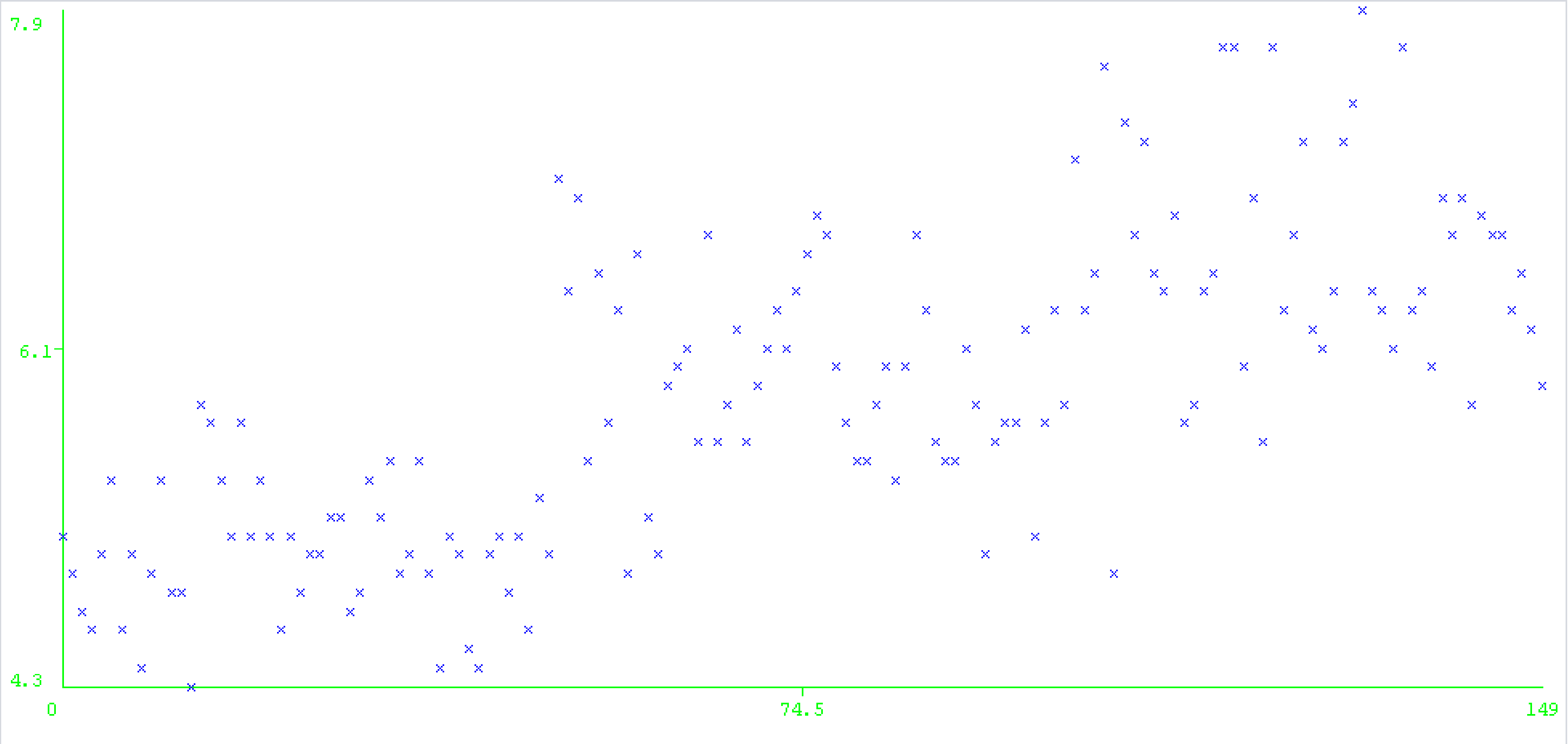
- Update step: the centroids of each cluster are re-calculated.
- The new centroids are calculated as the average of the objects that belong to the cluster.
- This is carried iteratively until there is no change in clusters.

Example: Iris Dataset

- The iris dataset consists of 150 number of training examples.
- These are of three classes: Iris-setosa, Iris-verginica, Iris-versicolor.
- Each feature vector consists of the following features: sepal width, sepal length, petal width, petal length.

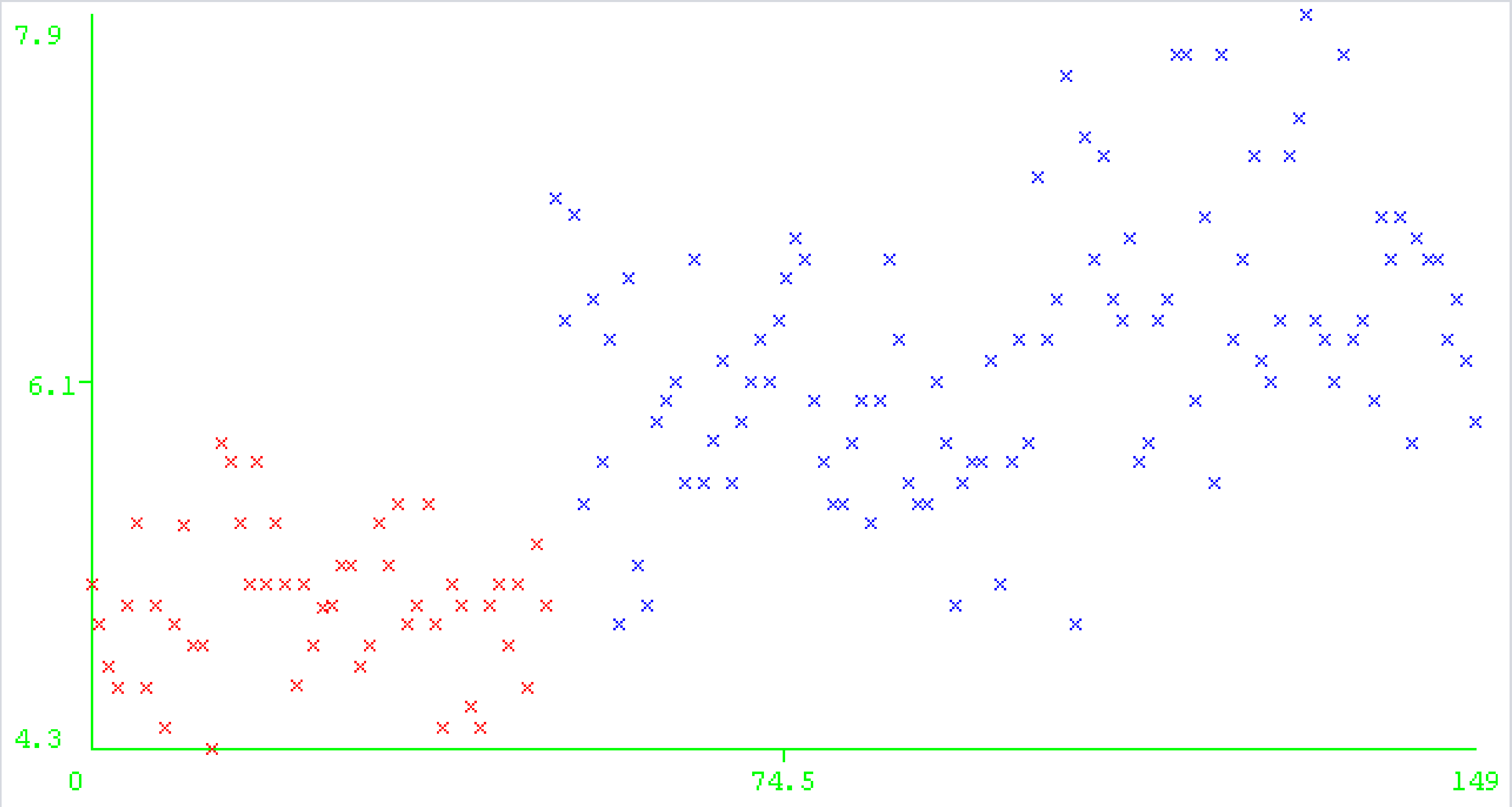
Example: Iris Dataset

- Original Data



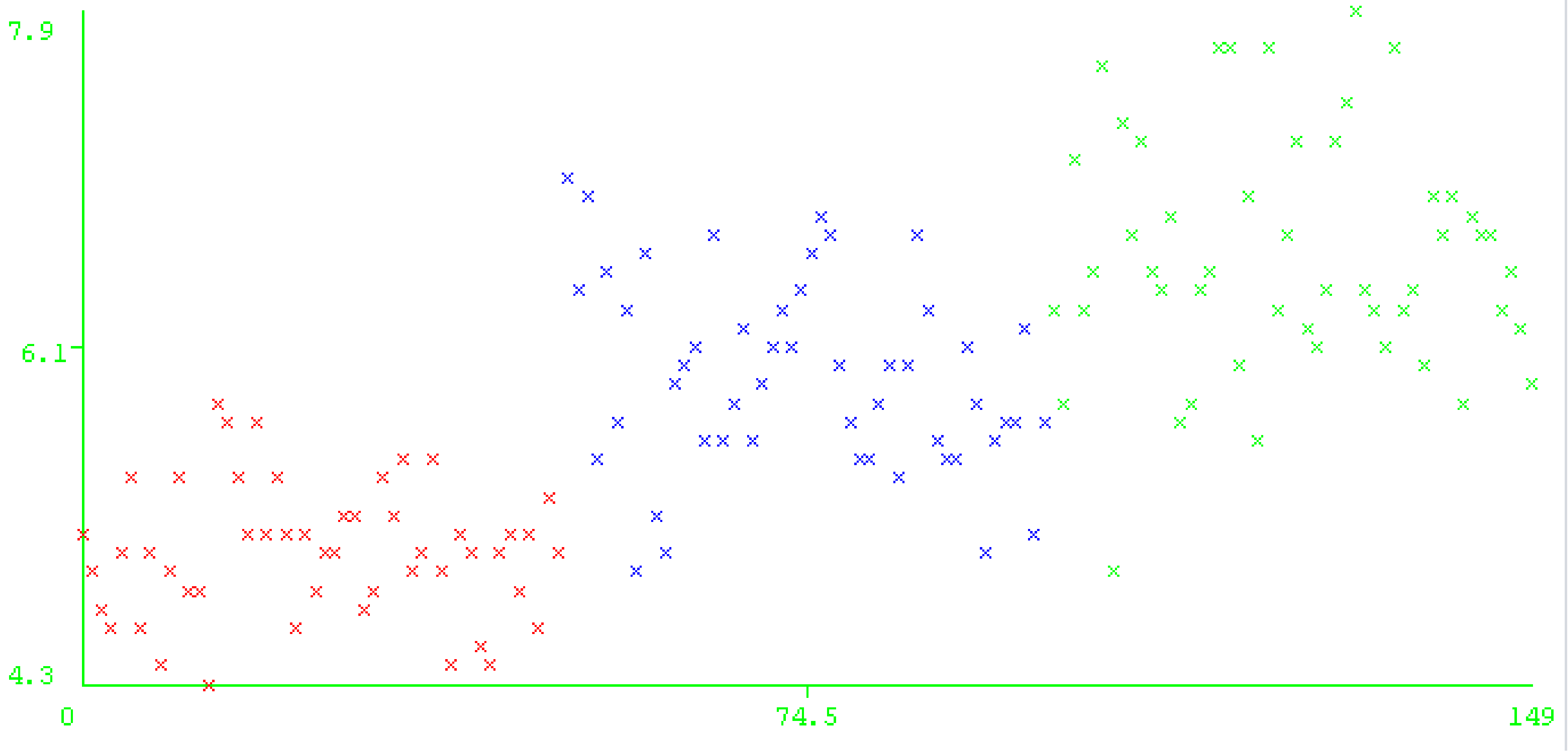
Example: Iris Dataset

- Two clusters



Example: Iris Dataset

- Three clusters



Strengths of k-means

- Based on an easy to understand and simple principle for identifying clusters.
- Flexible and adjustable.
- Efficient and performs well at dividing the data into useful clusters.

Weaknesses of k-means

- Less sophisticated than more recent clustering algorithms.
- Based on a random chance, which does not guarantee the algorithm to find the optimal set of clusters.
- Requires reasonable guesses as to how many clusters naturally exist in the data.

Example

ID	A	B
1	1.0	1.0
2	2.0	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

Example

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

Example

	Cluster 1		Cluster 2	
Step	ID	Mean Vector (centroid)	ID	Mean Vector (centroid)
1	1	(1.0, 1.0)	4	(5.0, 7.0)
2	1, 2	(1.5, 1.5)	4	(5.0, 7.0)
3	1, 2, 3	(1.8, 2.3)	4	(5.0, 7.0)
4	1, 2, 3	(1.8, 2.3)	4, 5	(4.2, 6.0)
5	1, 2, 3	(1.8, 2.3)	4, 5, 6	(4.3, 5.7)
6	1, 2, 3	(1.8, 2.3)	4, 5, 6, 7	(4.1, 5.4)

Example

	ID	Mean Vector (centroid)
Group 1	1, 2, 3	(1.8, 2.3)
Group 2	4, 5, 6, 7	(4.1, 5.4)

Example

ID	Distance to mean (centroid) of Cluster 1	Distance to mean (centroid) of Cluster 2
1	1.5	5.4
2	0.4	4.3
3	**2.1**	**1.8**
4	5.7	1.8
5	3.2	0.7
6	3.8	0.6
7	2.8	1.1

Example

	ID	Mean Vector (centroid)
Group 1	1, 2	(1.3, 1.5)
Group 2	3, 4, 5, 6, 7	(3.9, 5.1)