Universiti Malaysia PAHANG
Engineering • Technology • Creativity

CHAPTER 4
Data Based Segmentation

DATA MINING (BSD3533)

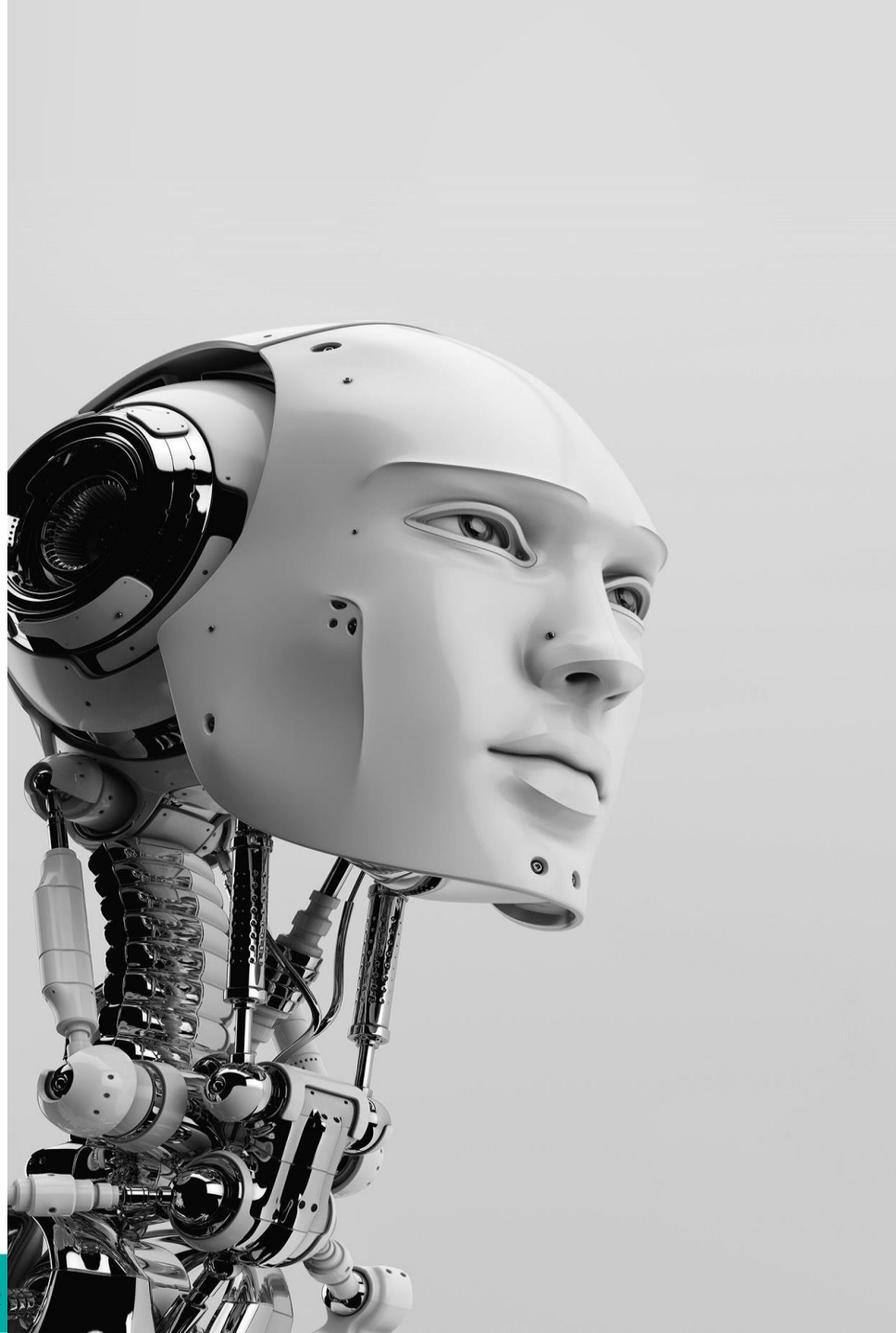
DR. KU MUHAMMAD NA'IM KU KHALIF

Content

Chapter 4.1: Cluster Analysis

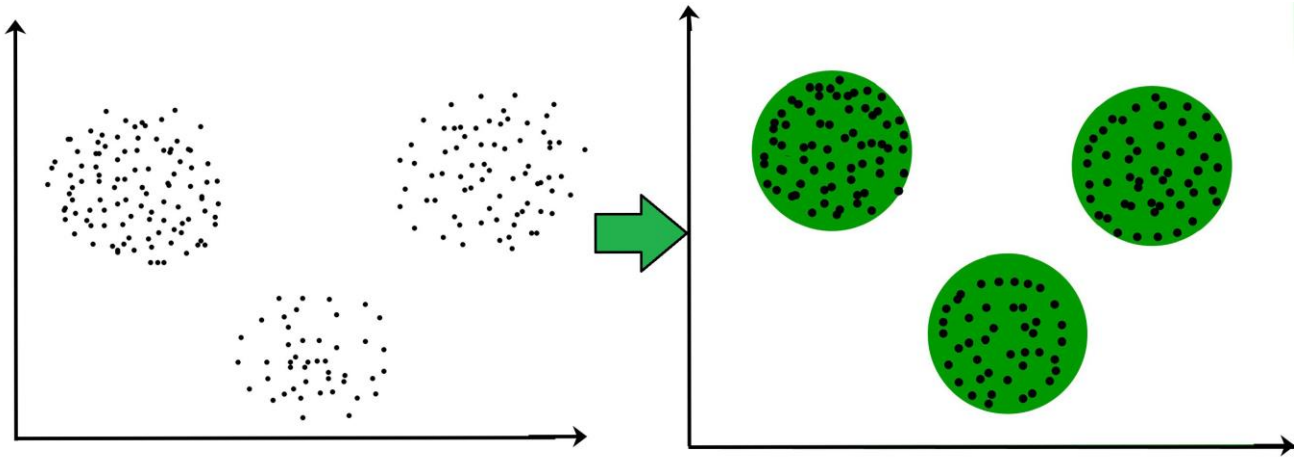
By the end of this topic, you should be able to:

- understand the concepts of clustering analysis in data mining problems.
- understand the applications of clustering analysis in data science problems.



Clustering Analysis

- Clustering is the task of dividing the population or 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.



Reference: https://www.geeksforgeeks.org/clustering-in-machine-learning/

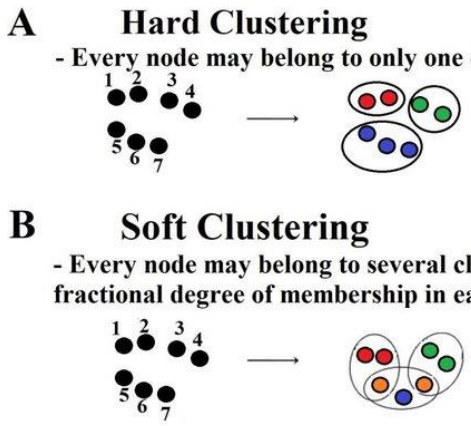
Why Clustering?

- ✓Clustering is somewhere similar to the classification algorithm, but the difference is the type of dataset that we are using. In classification, we work with the labeled data set, whereas in clustering, we work with the **unlabeled dataset**.
- ✓Clustering is very much important as it determines the intrinsic grouping among the unlabeled data present.
- ✓There are no criteria for good clustering. It depends on the user, what is the criteria they may use which satisfy their need. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding “natural clusters” and describe their unknown properties (“natural” data types), in finding useful and suitable groupings (“useful” data classes) or in finding unusual data objects (outlier detection). This algorithm must make some assumptions that constitute the similarity of points and each assumption make different and equally valid clusters.

Reference: https://www.geeksforgeeks.org/basic-concept-classification-data-mining/

Types of Clustering

- **Hard Clustering**: In hard clustering, each data point either belongs to a cluster completely or not.
- **Soft Clustering**: In soft clustering, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned.



Community Affiliation

	Nodes 1 2 3 4 5 6 7
Cluster 1	1 1 0 0 0 0 0
Cluster 2	0 0 1 1 0 0 0
Cluster 3	0 0 0 0 1 1 1

	Nodes 1 2 3 4 5 6 7
Cluster 1	1 1 0 0 1 0 0
Cluster 2	0 0 1 1 0 0 1
Cluster 3	0 0 0 0 1 1 1

Reference: https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/

Clustering Models

Since the task of clustering is subjective, the means that can be used for achieving this goal are plenty. Every methodology follows a different set of rules for defining the 'similarity' among data points. In fact, there are more than 100 clustering algorithms known. But few of the algorithms are used popularly, let's look at clustering models:

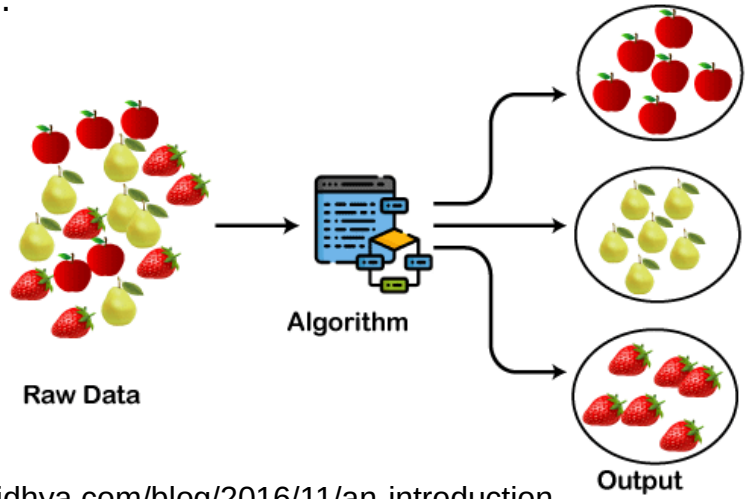
a. **Connectivity models**: As the name suggests, these models are based on the notion that the data points closer in data space exhibit more similarity to each other than the data points lying farther away. These models can follow two approaches. In the first approach, they start with classifying all data points into separate clusters & then aggregating them as the distance decreases. In the second approach, all data points are classified as a single cluster and then partitioned as the distance increases. Also, the choice of distance function is subjective. These models are very easy to interpret but lacks scalability for handling big datasets. Examples of these models are hierarchical clustering algorithm and its variants.

Reference: https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/

b. **Centroid models**: These are iterative clustering algorithms in which the notion of similarity is derived by the closeness of a data point to the centroid of the clusters. K-Means clustering algorithm is a popular algorithm that falls into this category. In these models, the no. of clusters required at the end have to be mentioned beforehand, which makes it important to have prior knowledge of the dataset. These models run iteratively to find the local optima.

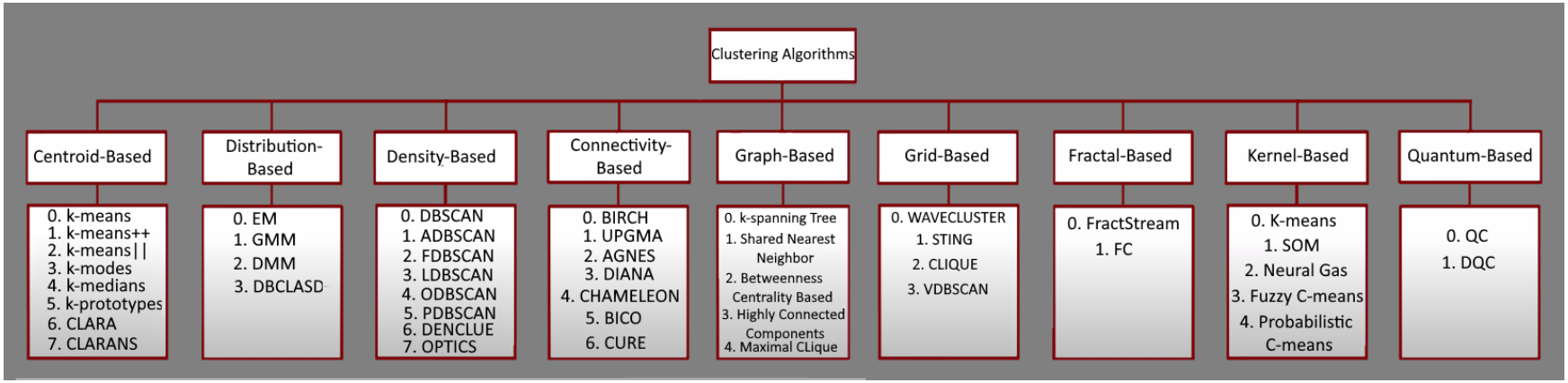
c. **Distribution models**: These clustering models are based on the notion of how probable is it that all data points in the cluster belong to the same distribution (For example: Normal, Gaussian). These models often suffer from overfitting. A popular example of these models is Expectation-maximization algorithm which uses multivariate normal distributions.

d. **Density Models**: These models search the data space for areas of varied density of data points in the data space. It isolates various different density regions and assign the data points within these regions in the same cluster. Popular examples of density models are DBSCAN and OPTICS.



Reference: https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/

Clustering Algorithms



Reference: https://www.javatpoint.com/clustering-in-machine-learning

Applications

- In identification of cancer cells: The clustering algorithms are widely used for the identification of cancerous cells. It divides the cancerous and non-cancerous data sets into different groups.
- In search engines: Search engines also work on the clustering technique. The search result appears based on the closest object to the search query. It does it by grouping similar data objects in one group that is far from the other dissimilar objects. The accurate result of a query depends on the quality of the clustering algorithm used.

Reference: https://www.javatpoint.com/clustering-in-machine-learning

- Customer segmentation: It is used in market research to segment the customers based on their choice and preferences.
- In biology: It is used in the biology stream to classify different species of plants and animals using the image recognition technique.
- In land use: The clustering technique is used in identifying the area of similar lands use in the GIS database. This can be very useful to find that for what purpose the particular land should be used, that means for which purpose it is more suitable.

Reference: https://www.javatpoint.com/clustering-in-machine-learning

Features of Clustering

- The desired feature of an ideal clustering technique is that intra-cluster distances should be minimized and inter-cluster distances should be maximized.
- Following are the other important features that an ideal cluster analysis method should have:
    1. Scalability: Clustering algorithms should be capable of handling small as well as large datasets smoothly.
    2. Ability to handle different types of attributes: Clustering algorithms should be able to handle different kinds of data such as binary, categorical and interval-based (numerical) data.
    3. Independent of data input order: The clustering results should not be dependent on the ordering of input data.
    4. Identification of clusters with different shapes: The clustering algorithm should be capable of identifying clusters of any shape

- Following are the other important features that an ideal cluster analysis method should have:
    5. Ability to handle noisy data: Usually, databases consist of noisy, erroneous or missing data, and algorithms must be able to handle these.
    6. High performance: To have a high-performance algorithm, it is desirable that the algorithm should need to perform only one scan of the dataset. This capability would reduce the cost of input-output operations.
    7. Interpretability: The results of clustering algorithms should be interpretable, logical and usable
    8. Ability to stop and resume: For a large dataset, it is desirable to stop and resume the task as it can take a huge amount of time to accomplish the full task and breaks may be necessary.
    9. Minimal user guidance: The clustering algorithm should not expect too much supervision from the analyst, because commonly the analyst has limited knowledge of the dataset.
- In clustering, distance metrics play a vital role in comprehending the similarity between the objects. In the next section, we will discuss different distance metrics that play an important role in the process of clustering objects.

Distance Metrics

- A distance metric is a function d(x, y) that specifies the distance between elements of a set as a non-negative real number. Two elements are equal under a particular metric if the distance between them is zero. Distance functions present a method to measure the closeness of two elements. Here, elements can be matrices, vectors or arbitrary objects and do not necessarily need to be numbered.
- In the following subsections, important distance metrics used in measuring similarity among objects have been illustrated.

Euclidean Distance

- Euclidean distance is mainly used to calculate distances. The distance between two points in the plane with coordinates (x, y) and (a, b) according to the Euclidean distance formula is given by:

$$\text{Euclidean dist}((x, y), (a, b)) = \sqrt{(x - a)^2 + (y - b)^2}$$

For example, the (Euclidean) distance between points (-2, 2) and (2, -1) is calculated as

$$\text{Euclidean dist}((-2, 2), (2, -1)) = \sqrt{(-2 - (2))^2 + (2 - (-1))^2}$$
$$= \sqrt{(-4)^2 + (3)^2}$$
$$= \sqrt{16 + 9}$$
$$= \sqrt{25}$$
$$= 5$$

Manhattan Distance

- Manhattan distance is also called L1-distance. It is defined as the sum of the lengths of the projections of the line segment between the two points on the coordinate axes.
- For example, the distance between two points in the plane with coordinates (x, y) and (a, b) according to the Manhattan distance formula, is given by:

$$\text{Manhattan dist}((x, y), (a, b)) = |x - a| + |y - b|$$

Manhattan dist((30, 70), (40, 54)) = |30 - 40| + |70 – 54|

= |- 10 | + | 16| = 10+16

= 26

Chebyshev Distance

- It is also called as chessboard distance because, in a game of chess, the minimum number of moves required by a king to go from one square to another on a chessboard equals Chebyshev distance between the centres of the squares. Chebyshev distance is defined on a vector space, where the distance between two vectors is the maximum value of their differences along any coordinate dimension. Formula of Chebyshev distance is given by:

$$\text{Chebyshev dist}((r1, f1), (r2, f2)) = \max(|r2-r1|, |f2-f1|)$$

Object A coordinate = {0,1,2,3}

Object B coordinate = {6,5,4,-2}

According to Chebyshev distance formula

$$D = \max(|r2-r1|, |f2-f1|)$$

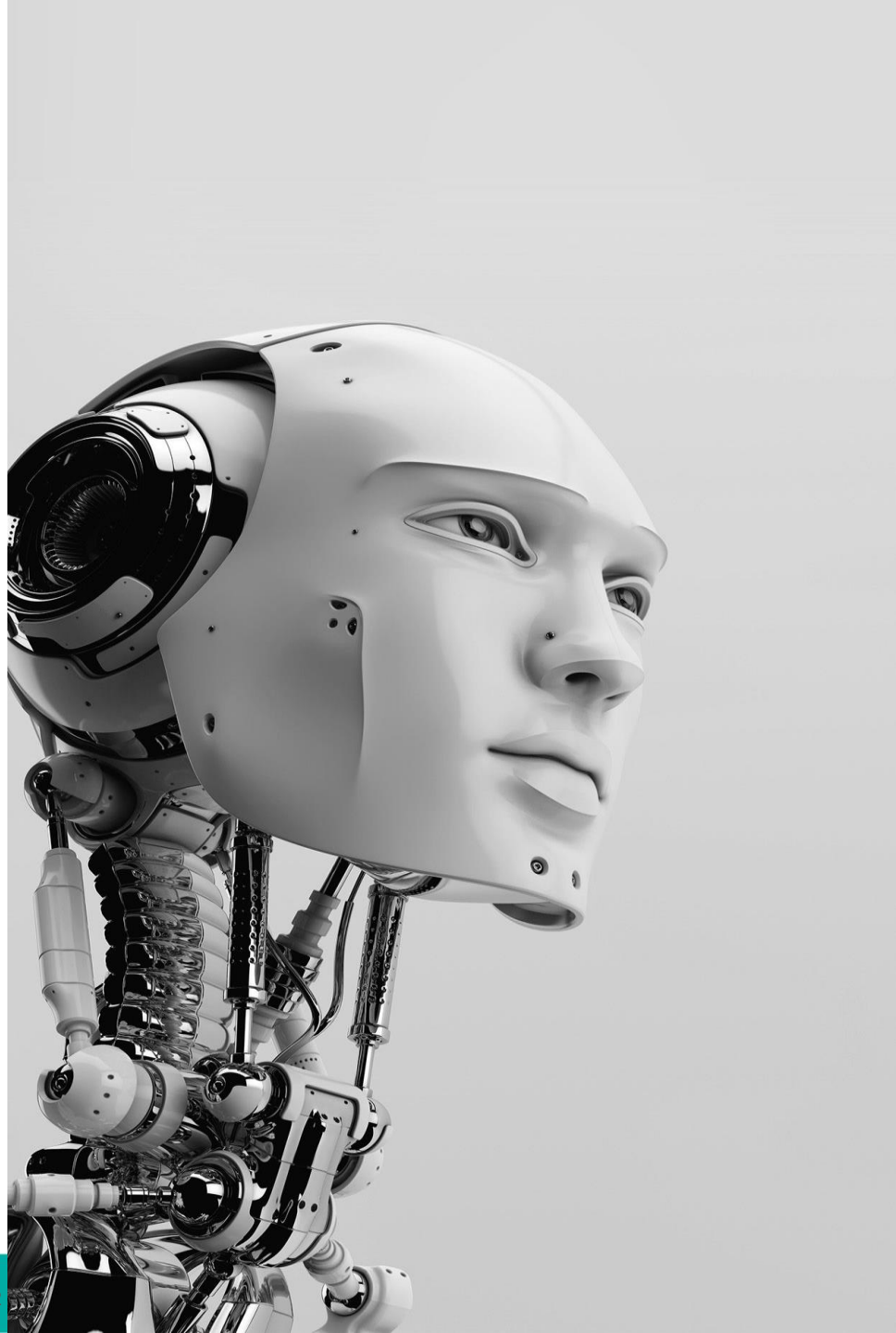
$$= \max(|6-0|, |5-1|, |4-2|, |-2-3)|)$$

$$= \max(6,4,2,5) = 6$$

Chapter 4.2:
Partitioning Clustering

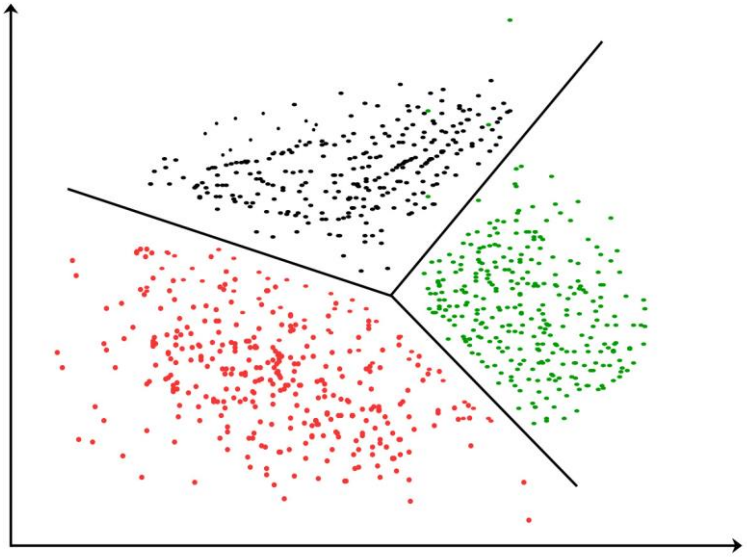
By the end of this topic, you should be able to:

- understand the concepts of partitioning clustering analysis in data mining problems.
- understand the applications of k-means and fuzzy c-means clustering analysis in data science problems.



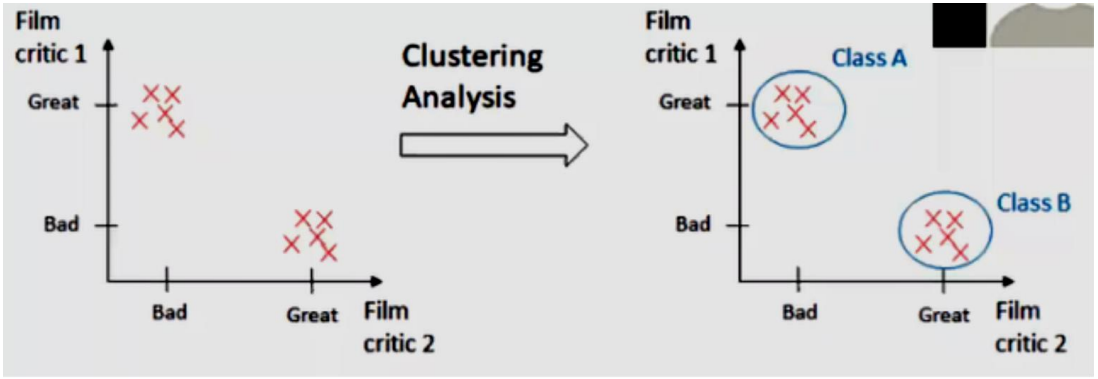
K-Means

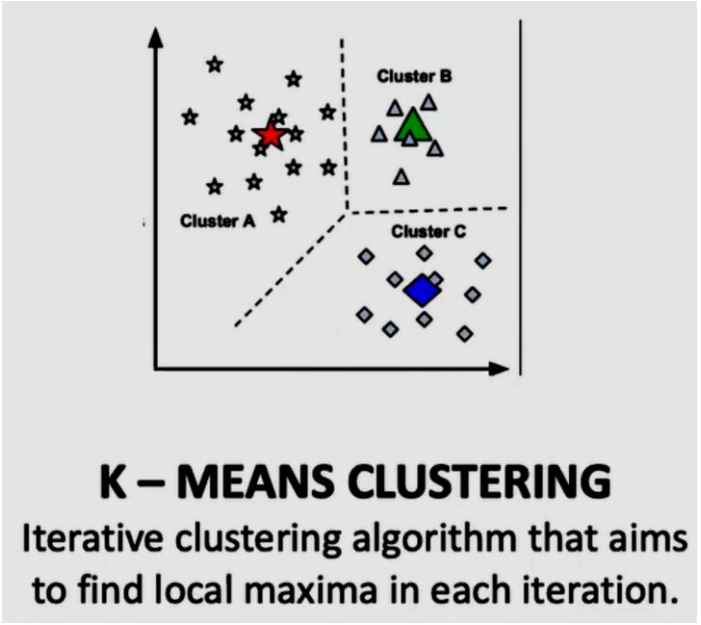
- K-means algorithm partitions *n* observations into *k* clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster.



Reference: https://www.geeksforgeeks.org/clustering-in-machine-learning/

- Clustering is a method to identify different of the elements in the data.
- Used to find the structure in the dataset so that element of the same cluster are more similar to each other.



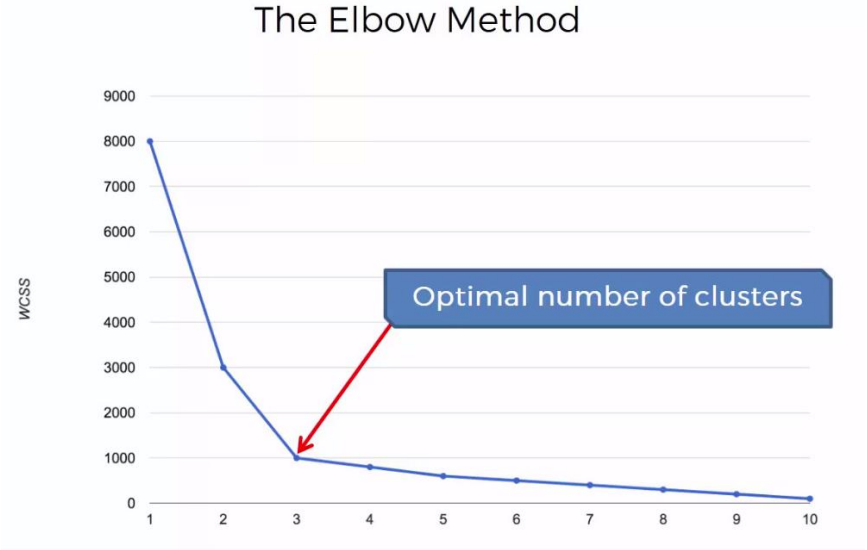


- It learned from input value from dataset without referring to known or labelled output.
- Partitioning or clustering the data points into *k* number of distinct clusters, also called centroid.
- Each data point will match the nearest distance of located centroid based on Square Euclidean Distance.

- **Within Cluster Sums of Squares :** $$\mathrm{WSS} = \sum_{i=1}^{N_C} \sum_{x \in C_i} d(\mathbf{x}, \bar{\mathbf{x}}_{\mathbf{C_i}})^2$$
- **Between Cluster Sums of Squares:** $$\mathrm{BSS} = \sum_{i=1}^{N_C} |C_i| \cdot d(\bar{\mathbf{x}}_{\mathbf{C_i}}, \bar{\mathbf{x}})^2$$

C_i = Cluster, N_c = # clusters, $\bar{x}_{c_i}$ = Cluster centroid, $\bar{x}$ = Sample Mean

- The elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters.
- By default, the distortion score is computed, the sum of square distances from each point to its assigned centre.

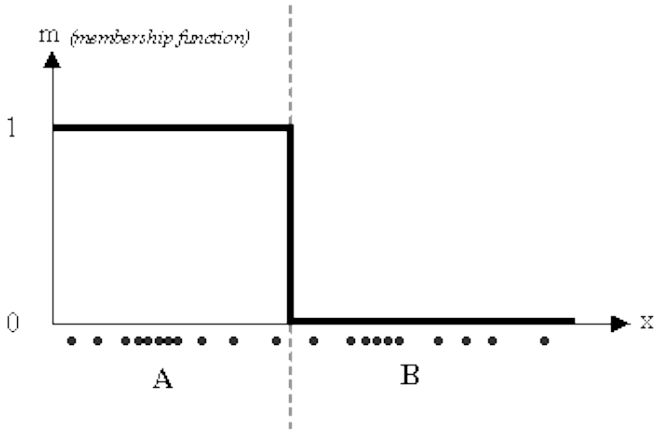


Fuzzy C-Means

- Fuzzy clustering is a type of soft method in which a data object may belong to more than one group or cluster. Each dataset has a set of membership coefficients, which depend on the degree of membership to be in a cluster.
- Data are bound to each cluster by means of a Membership Function, which represents the fuzzy behaviour of this algorithm. To do that, we simply have to build an appropriate matrix named U whose factors are numbers between 0 and 1, and represent the degree of membership between data and centers of clusters. For a better understanding, we may consider this simple mono-dimensional example. Given a certain data set, suppose to represent it as distributed on an axis. The figure below shows this:

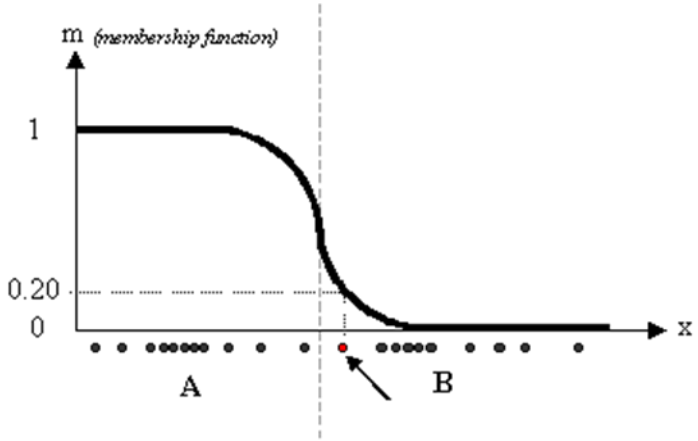
Reference: https://matteucci.faculty.polimi.it/Clustering/tutorial_html/cmeans.html

- Looking at the picture, we may identify two clusters in proximity of the two data concentrations. We will refer to them using ‘A’ and ‘B’. In the first approach shown in this tutorial - the k-means algorithm - we associated each datum to a specific centroid; therefore, this membership function looked like this:



Reference: https://matteucci.faculty.polimi.it/Clustering/tutorial_html/cmeans.html

- In the FCM approach, instead, the same given datum does not belong exclusively to a well defined cluster, but it can be placed in a middle way. In this case, the membership function follows a smoother line to indicate that every datum may belong to several clusters with different values of the membership coefficient.



Reference: https://matteucci.faculty.polimi.it/Clustering/tutorial_html/cmeans.html

- In the figure above, the datum shown as a red marked spot belongs more to the B cluster rather than the A cluster. The value 0.2 of ‘m’ indicates the degree of membership to A for such datum. Now, instead of using a graphical representation, we introduce a matrix U whose factors are the ones taken from the membership functions:

$$U_{MC} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ .. & .. \\ 0 & 1 \end{bmatrix} \qquad U_{MC} = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \\ 0.6 & 0.4 \\ .. & .. \\ 0.9 & 0.1 \end{bmatrix}$$

A B

Reference: https://matteucci.faculty.polimi.it/Clustering/tutorial_html/cmeans.html

- Main objective of fuzzy c-means algorithm is to minimize:

$$\mu_{ij} = 1 / \sum_{k=1}^{c} (d_{ij} / d_{ik})^{(2/m-1)}$$

$$v_j = (\sum_{i=1}^{n} (\mu_{ij})^m x_i) / (\sum_{i=1}^{n} (\mu_{ij})^m), \forall j = 1, 2, ...c$$

- Then

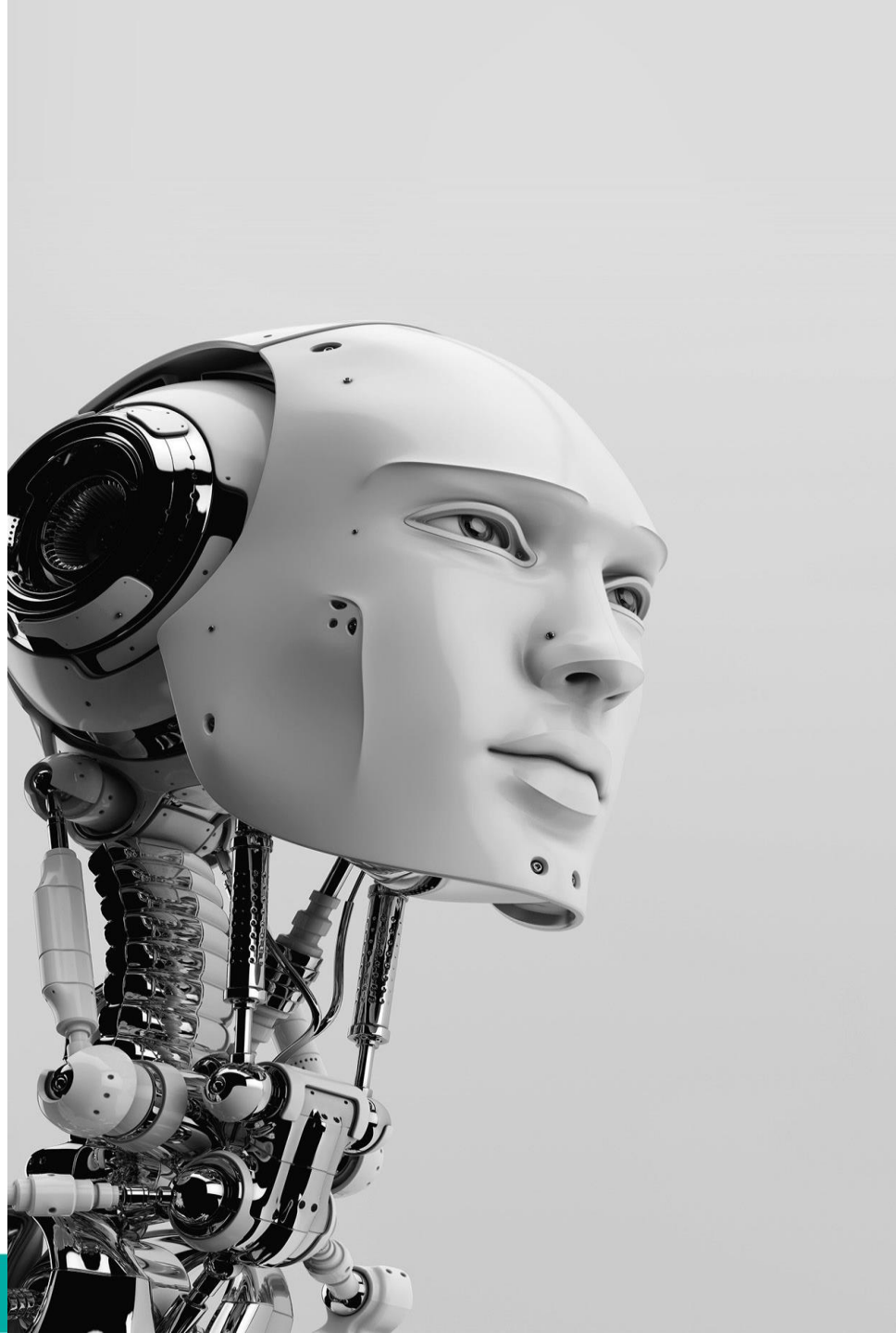
$$J(U,V) = \sum_{i=1}^{n} \sum_{j=1}^{c} (\mu_{ij})^m \left\| x_i - v_j \right\|^2$$

Reference: https://sites.google.com/site/dataclusteringalgorithms/fuzzy-c-means-clustering-algorithm

Chapter 4.3: Hierarchical Clustering

By the end of this topic, you should be able to:

- understand the concepts of hierarchical clustering in clustering algorithm.



Preview

- Hierarchical clustering is one of the popular and easy-to-understand clustering techniques.
- This clustering technique is divided into two types:

Agglomerative Hierarchical clustering

In this technique, initially, each data point is considered as an individual cluster. At each iteration, the similar clusters merge with other clusters until one cluster or K clusters are formed.

Divisive Hierarchical clustering

Since the Divisive Hierarchical clustering Technique is not much used in the real world, I'll give a brief of the Divisive Hierarchical clustering Technique.

Limitations of Hierarchical clustering Technique:

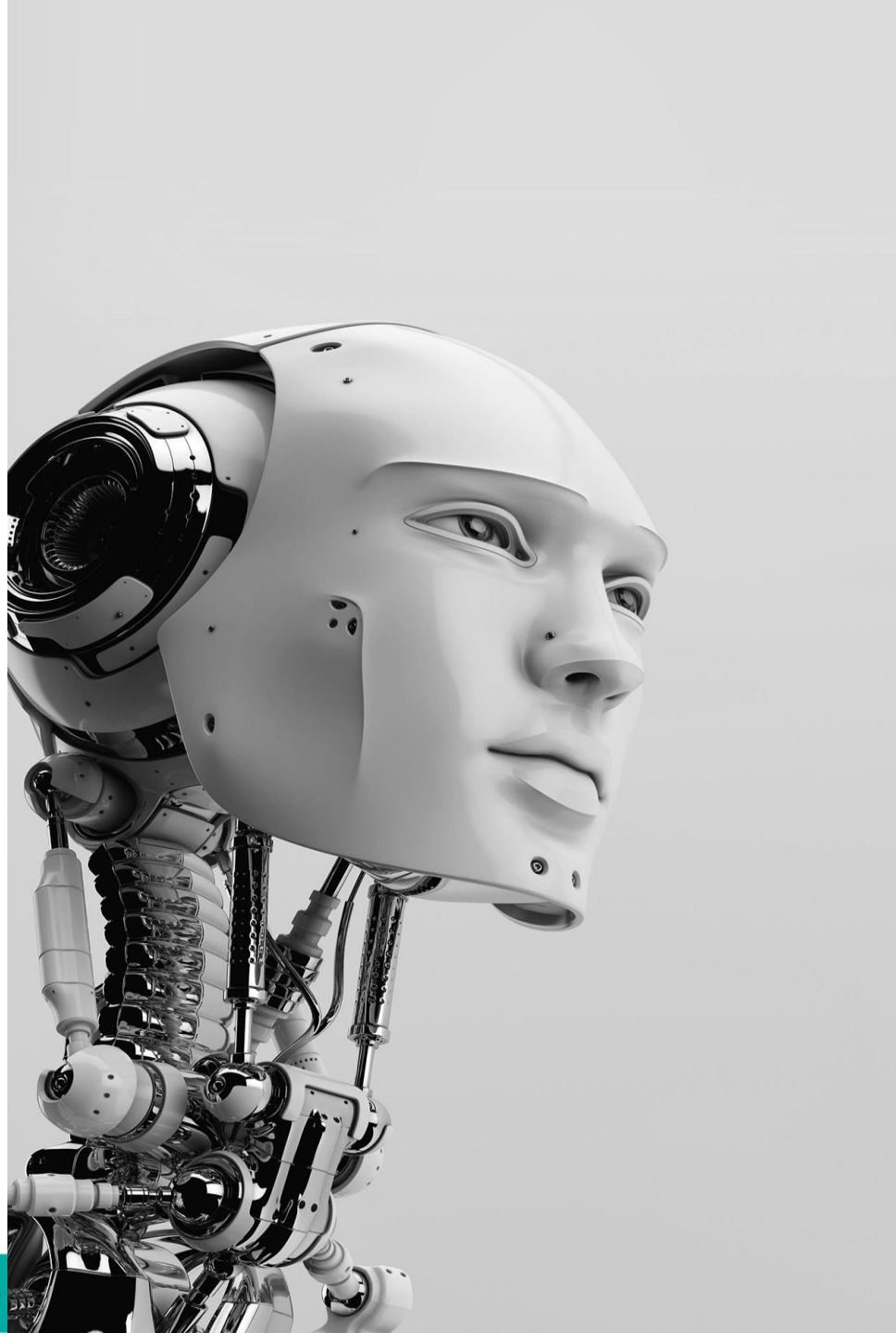
a. There is no mathematical objective for Hierarchical clustering.
b. All the approaches to calculating the similarity between clusters has their disadvantages.
c. High space and time complexity for Hierarchical clustering. Hence this clustering algorithm cannot be used when we have huge data.

https://www.analyticsvidhya.com/blog/2019/05/beginners-guide-hierarchical-clustering/

Chapter 4.4: Density Based Clustering

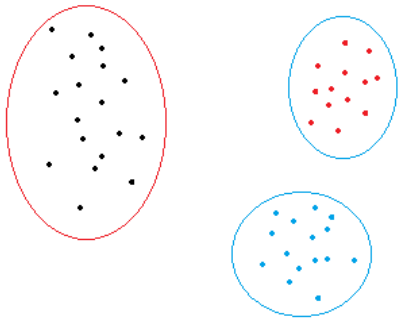
By the end of this topic, you should be able to:

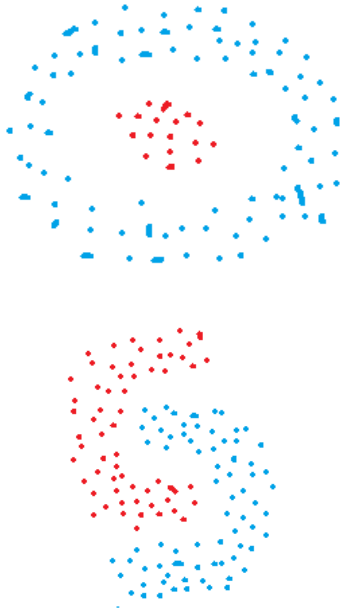
- understand the concepts of density based clustering.



Preview

- Partition-based and hierarchical clustering techniques are highly efficient with normally shaped clusters. However, when it comes to arbitrarily shaped clusters or detecting outliers, density-based techniques are more efficient.





- The data points in these figures are grouped in arbitrary shapes or include outliers.
- Density-based clustering algorithms are very efficient at finding high-density regions and outliers. It is very important to detect outliers for some tasks, e.g. anomaly detection.

Pros:

a. Does not require specifying the number of clusters beforehand.

b. Performs well with arbitrary shapes clusters.

c. DBSCAN is robust to outliers and able to detect the outliers.

Cons:

a. In some cases, determining an appropriate distneighbourhood (eps) is not easy and it requires domain knowledge.

b. If clusters are very different in terms of in-cluster densities, DBSCAN is not well suited to define clusters. The characteristics of clusters are defined by the combination of eps-minPts parameters. Since we pass in one eps-minPts combination to the algorithm, it cannot generalize well to clusters with much different densities.

https://www.analyticsvidhya.com/blog/2020/09/how-dbscan-clustering-works/