Machine Learning (Unsupervised Learning and K-Means Clustering)

Prof K R Chowdhary

CSE Dept., MBM University

February 27, 2025
Lecture #9

Introduction to Clustering

⇒ The process of clustering partitions a set of data, according to some similarity measure, into several groups such that "similar" records are in the same group, so that each group represents a similar subpopulations in the data.

⇒ As an example, each cluster could be a group of customers, which have similar purchase histories or interactions or some other factors or the combinations [1].

Table 1: Data groupings of similar objects

Cluster No.	<Qty, Unit Price>
Cluster 1	<2, 1800>
	<3, 2050>
	<5, 2270>
Cluster 2	<15, 1800>
	<18, 2200>
	<12, 2380>
Cluster 3	<3, 250>
	<4, 180>
	<4, 200>

Introduction to Clustering...

⇒ Clustering is based on common properties of items in each group/cluster, as follows: customers in cluster 1 purchased few high-priced items, customers in cluster 2 purchased many high-priced items, and customers in cluster 3 purchased few low priced items.

Definition

Cluster feature(CF). Collective summarized representation of a cluster to optimize space as well as to facilitate faster access. □

⇒ CF is a triple: cluster centroid, cluster radius, and number of points in the cluster.
⇒ CF based approach is efficient due: 1. they consume less space as all objects in a cluster are not required, and 2. they constitute sufficient information for computing all intra-cluster and inter-cluster. So distances can be computed fast.

Introduction to Clustering...

$\Rightarrow$ Some points in clusters can be discarded, while the others can be compressed, as defined below.

Definition

Discardable Point. A point is considered discardable, if its membership can be ascertained with high confidence. □

Definition

Compressible Point. A point that is not discardable, but belongs to a tight subcluster consisting of a set of points that always move between clusters simultaneously, is called a compressible point. □

$\Rightarrow$ Clustering applications:
Information Retrieval
Biology
Business

Summarization
Nearest neighbors
Compression

Introduction to Clustering ...

$\Rightarrow$ Consider a group of 12 sales records each indicating sales price, and have been sorted in ascending order as: 5, 8, 11, 13, 15, 35, 45, 55, 72, 92, 201, 215. It is required to partition these into three clusters. The partitions finally formed are shown in Table:

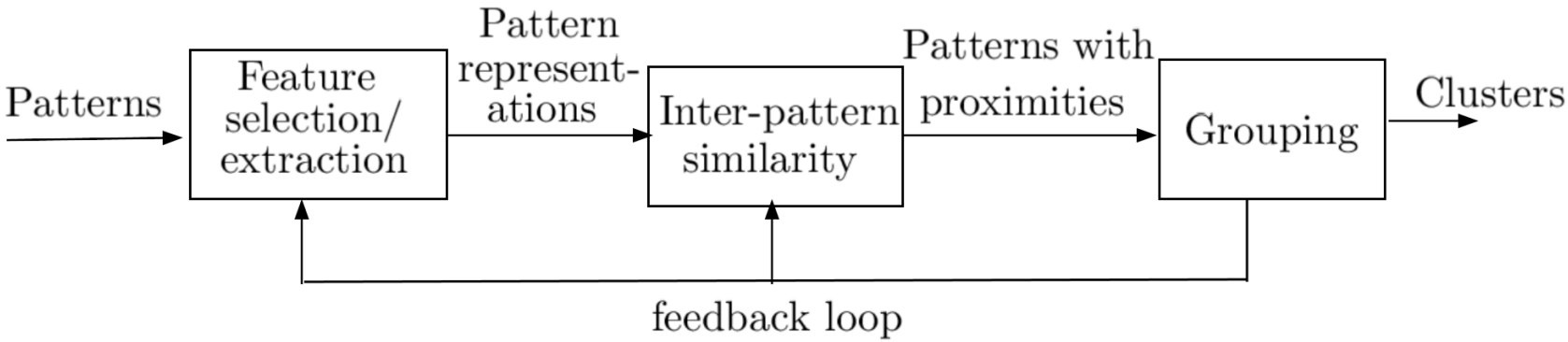
Cluster-1	Cluster-2	Cluster-2
5,8,11, 13,15	35,45,55, 72,92	201,215

Also, applications are: image segmentation, pattern classification, and data mining.

A general pattern clustering process has following steps:

1. pattern representation, which may also includes feature selection and extraction,
2. defining proximity measures patterns specific to data domains,
3. grouping of patterns (clustering),
4. optionally, abstraction of data, and
5. optionally, assessment of output.

Clustering stages



⇒ The Fig. above is a typical case of clustering. Feedback path indicates that the grouping process output could affect feature extraction and similarity computations in the next iteration.

⇒ The *pattern representation* depends on: available patterns, classes and their number, feature types and their scale for clustering algorithm. Feature selection helps in identifying the most effective subset of the original features to use in clustering.

⇒ Through *feature extraction*, one or more transformations of input features is carried out.

Clustering stages...

$\Rightarrow$ A *Pattern* set is denoted by $\mathscr{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. The ith pattern in $\mathscr{X}$ is: $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,d})$. Often, a pattern set to be clustered is viewed as an $n \times d$ pattern matrix.

$\Rightarrow$ *Pattern proximity* or closeness of one pattern to other, is measured by a distance function defined on a pair of patterns. A distance measuring function is *Euclidean distance*: used to reflect dissimilarity between two patterns. And, if the Euclidean distance is zero...

$\Rightarrow$ The *grouping* step can be carried out in many ways. The output of clustering can be *crisp* or *fuzzy*. When clustering at output is crisp (*hard*), the data is partitioned into groups, where as when it is fuzzy partition, each pattern has a variable degree of membership of $[0, 1]$, in each of the output clusters.

Clustering stages...

$\Rightarrow$ All clustering algorithms produce clusters when presented with data, irrespective of whether the data really contain clusters or not.

$\Rightarrow$ It is not necessary that every set of data contains some clusters. For example, the continuous sequence $1, 2, ..., 100$, in no way represents a cluster. Only what we can have is all these numbers as clusters, each of size one.

$\Rightarrow$ To evaluate the cluster quality, we should know what characterizes a "good" clustering result. If data contain no clusters, some clustering algorithms may obtain "better" clusters than others.

$\Rightarrow$ Criteria for comparing clustering algorithms are based on: 1. how the clusters are formed, 2. their data structure, and 3. how sensitive is the clustering technique to changes, which does not affect the data structure.

Data Clustering and Cluster Analysis

⇒ The table 1 (page no. 2) is a sample database as clusters, here clusters of customers' database based on three purchase behavior: quantity, unit price, and their combination.

⇒ We organize a collection of patterns into groups based on their similarity. Intuitively, the patterns in same cluster are more similar to each other than those in different clusters. See Fig. 1 (a) and (b). Measurement of the patterns (x, y) coordinate values.

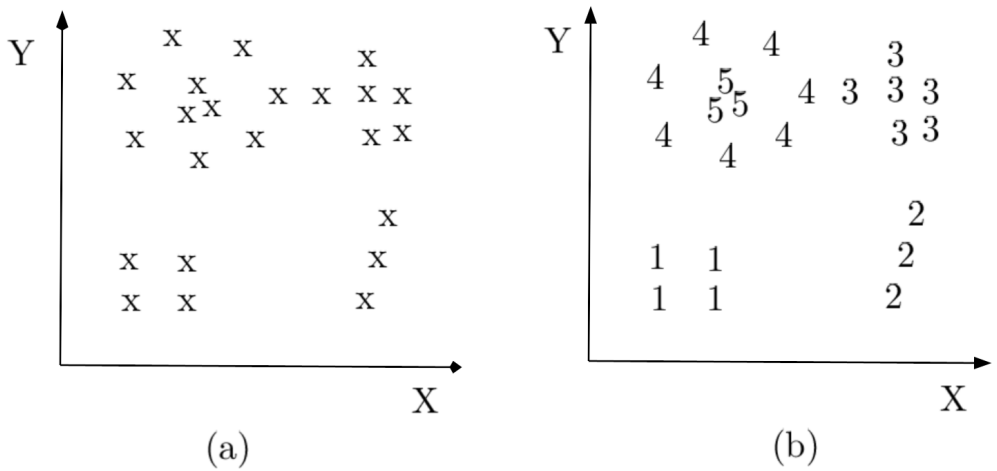


Figure 1: Data Clustering: (a) Input patterns, (b) Clusters formed

Pattern representation and feature extraction

$\Rightarrow$ A better quality pattern representation results to simple and easily understood clustering. E.g., In Cartesian coordinates, many clustering algorithms may fragment data into two or more clusters; In polar coordinates, radius coordinate causes tight clustering and a one-cluster solution can be easily obtained.

$\Rightarrow$ A pattern can be for a physical object or an abstract notion. Physical: a chair, table, book, house, abstract: a style of writing, attitude, belief. Both can be represented as multidimensional vectors. Features of pattern can be quantitative /qualitative: weight, color, $(black, 5)$ is black object with 5 units of weight, or degree of blackness.

Clustering Algorithms

$\Rightarrow$ In clustering, where it lacks class labels, feature selection is an ad hoc, but a necessity. As it lacks class labels, there can only be a trial-and-error process for selection of features.

$\Rightarrow$ The resultant patterns are clustered, and output is evaluated using a *validity index*. Popular feature extraction processes: principal components analysis (PCA), it does not depend on labeled data. Patterns having smaller number of features are beneficial [1].

$\Rightarrow$ For clustering: first requirement is to find out similarities, and more similar patterns are clubbed together.

$\Rightarrow$ Dissimilarity between two patterns is the feature space using the distance measure. The popular metric for continuous features is *Euclidean distance*:

$$d_2(\mathbf{x}_i, \mathbf{x}_j) = \Big(\sum_{k=1}^{d}(x_{i,k} - x_{j,k})^2\Big)^{1/2}$$
$$= \| \mathbf{x}_i - \mathbf{x}_j \|_2 . \qquad (1)$$

$\Rightarrow$ Approach based on Euclidean distance, the method is used to evaluate proximity of objects in 2D/3D spaces.

$\Rightarrow$ Set of 2D data points (Table 2) and a data point, $x = (2.5, 2.9)$ as a query, rank these database points based on similarity with query.

Table 2: 2D data

	A_1	A_2
x_1	1.9	1.7
x_2	2.1	2.1
x_3	2.6	3.0
x_4	2.2	2.5
x_5	1.8	2.0

$\Rightarrow$ Using equation (1), we compute the Euclidean distance for the two dimensional data points $x_1, \ldots, x_5$ with respect to the query $x = (2.5, 2.9)$. The result are shown in Table 3.

Table 3: Euclidean Distances

Data pt.	Euclid. dist. with x
x_1	1.341
x_2	0.894
x_3	0.141
x_4	0.500
x_5	1.140

The distance matrix shows that query $(2.5, 2.9)$ is nearest to x_3, having distance 0.141.

$\Rightarrow$ **Nearest Neighbor Clustering (NN)**: An iterative algorithm assigns each unlabeled pattern to the cluster of its nearest labeled neighbor pattern. Condition is: distance to that nearest pattern is below threshold.

$\Rightarrow$ This process continues until all the input patterns are labeled.

K-Means Clustering

⇒ The k-means tries to find k number of clusters, the count is specified by the user. These are represented by their centroids. It is simplest and most commonly used algorithm that uses *squared error* criterion.

⇒ The k-means algorithm starts with a random initial partition and keeps reassigning the patterns to clusters based on the similarity between the pattern and the cluster centers (centroid distances) until a convergence condition is reached.

K-Means Clustering..

$\Rightarrow$ The k-means is a partitional clustering technique that tries to find a k number of clusters (count is given by the user). These are represented by their centroids. It is simplest and commonly used algorithm that uses *squared error* criterion.

$\Rightarrow$ k-means algorithm starts with a random initial partition and keeps reassigning the patterns to clusters based on the similarity between the pattern and the cluster centers (centroid distances) until a convergence condition is reached.

$\Rightarrow$ In clustering process, there is no reassignment of any pattern from one cluster to another, this gives it a property of *linear time complexity*.

$\Rightarrow$ Advantages of k-means: 1). It is easy to implement, 2. Its time complexity is $O(n)$, where n is total number of patterns. Disadvantage: sensitive to selection of the initial partition – if not properly selected, it may converge to a *local minima* of the criterion function value.

K-Means Clustering: Example

$\Rightarrow$ Using the k-means approach to perform clustering.

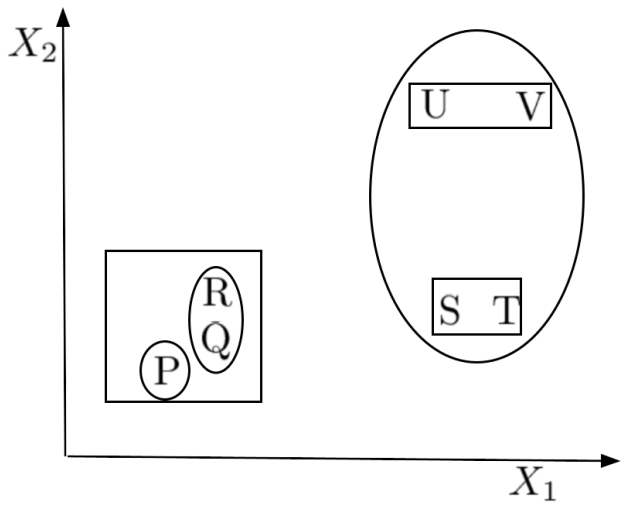


Figure 2: The k-means clustering is sensitive to initial partition

$\Rightarrow$ Fig. shows 2D patterns P, Q, R, S, T, U, V. Process is started with initial patterns P, Q, R. Around these, three (given $k = 3$) clusters are to be constructed. We we end up with partition $\{\{P\}, \{Q, R\}, \{S, T, U, V\}\}$, where three clusters are ellipses.

$\Rightarrow$ The *squared error* criterion value turns out to be very large for this partition (see eqation (2)). This will happen, for the centroid vs. the patterns in the largest ellipse.

$\Rightarrow$ "Squared error function" approach is most intuitive concept for partitional clustering, it ideally suited for compact and isolated clusters.

For an input set of $\mathcal{X}$ patterns, the "squared error" for clustering $\mathcal{C}$, consisting K clusters $(C_1, ..., C_K)$, expressed as:

$$e^2(\mathcal{X}, \mathcal{C}) = \sum_{j=1}^{K} \sum_{i=1}^{m_j} \| \mathbf{x}_i^{(j)} - \mathbf{c}_j \|^2 . \quad (2)$$

$\Rightarrow$ One technique for clustering is called *supervised learning*, while other is called *unsupervised learning*.

$\Rightarrow$ In supervised technique, a collection of labels, i.e., pre-classified patterns, are already provided (training patterns), and the task is to label newly encountered unlabeled patterns.

Data Clustering and Cluster Analysis

⇒ In the case of unsupervised learning, the task is to group a given collection of unlabeled patterns into some meaningful clusters. In fact, some kind of labels are there, associated with the clusters this time also, but this category of labels are data driven – obtained solely from the data, and not predefined.

[1] Chowdhary, K.R. (2020). Data Mining. In: Fundamentals of Artificial Intelligence. Springer, New Delhi. https://doi.org/10.1007/978-81-322-3972-7_17 pp. 519-534.