Statistical Inference and Multivariate Analysis (MA324)

LECTURE SLIDES
Lecture 36

Cluster Analysis

Indian Institute of Technology Guwahati

Jan-May 2023

Cluster Analysis

- Cluster analysis or clustering is the task of **grouping a set of objects** in such a way that objects in the **same group** (called a cluster) are **more similar** (in some sense) to each other than to those in other groups (clusters).

- Clustering can therefore be formulated as a **multi-objective optimization** problem.

- Cluster analysis as such is not an automatic task, but an **iterative process** of knowledge discovery or interactive multi-objective optimization that involves trial and failure.

- The basic objective in cluster analysis is to **discover natural groupings** of the items (or variables).

Application of Cluster Analysis...

- Image analysis
- Pattern recognition
- Information Retrieval
- Data compression
- Bioinformatics
- Computer graphics
- Anomaly detection
- Medical science
- Natural language processing (NLP)
- Crime analysis
- Social science
- Robotics
- Finance
- Petroleum geology
- Food Industry

Similarity Measures : Understanding Proximity

- In cluster analysis, we must first develop a quantitative scale on which to **measure the association (similarity)** between objects.
- To understand the "**closeness**" or "**similarity**" among clusters, there can be two different methods:
  - **Distance Measure** : Distances and Similarity Coefficients for Pairs of Items.
  - **Association Measure** : Similarities and Association Measures for Pairs of Variables.

Distance Measure

Here, using this method, we try to understand or estimate the **statistical distance between two given clusters** (say two p-dimensional observations, $\mathbf{x}^{'} = [x_1, ..., x_p]$ and $\mathbf{y}^{'} = [y_1, ..., y_p]$) . For this procedure, we may using various distance metrics, namely :

- **Mahalanobis distance** (Statistical Distance) between two observations, given by,

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})' \mathbf{A} (\mathbf{x} - \mathbf{y})},$$

  where, $\mathbf{A} = \mathbf{S}^{-1}$, $\mathbf{S}$ being the matrix of sample variance and covariances.

- **Minkowski** Metric, given by,

$$d(\mathbf{x}, \mathbf{y}) = \left[\sum_{i=1}^{p} |x_i - y_i|^m \right]^{\frac{1}{m}}$$

- **Canberra** metric,

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

- **Czekanowski** coefficient,

$$d(\mathbf{x}, \mathbf{y}) = 1 - \frac{2 \sum_{i=1}^{p} min(x_i, y_i)}{\sum_{i=1}^{p} (x_i + y_i)}$$

Association Measure

When the variables are binary, the data can again be arranged in the form of a contingency table. In such a situation, it is better to get a measure of association among the variables.

A contingency table for,

Variable i	Variable k		
	1	0	Total
1	a	b	a + b
0	c	d	c + d
Total	a + c	b + d	n = a + b +c + d

Product Moment Correlation

- The usual **product moment correlation** formula applied to the binary variables in the contingency table is,

$$r = \frac{ad - bc}{[(a+b)(c+d)(a+c)(b+d)]^{\frac{1}{2}}}$$

- The above **moment correlation** can be taken as a **measure of the similarity** between the two variables.
- The **moment correlation coefficient** is **related** to the **Chi-square statistic** $\left(r^2 = \frac{\chi^2}{n}\right)$ for testing the independence of two categorical variables. Keeping n fixed, a **large similarity** (or correlation) is **consistent** with the **absence of independence**.

Cluster Creation

Now, in cluster analysis, the main aim is to **create the clusters** using any one of the two major techniques, namely

- **Hierarchical Clustering** : which mainly proceeds by either a series of successive mergers or a series of successive divisions. The two types of hierarchical clustering are :
  - **Agglomerative hierarchical** methods.
  - Divisive hierarchical methods (Self Study)
- **Non-Hierarchical Clustering** : technique is mainly designed to group items, rather than variables, into a collection of K clusters. The most common methodology is:
  - **K-Means** Method.

Agglomerative Hierarchical Methods

- This methodology of clustering, start with the individual objects.
- Initially as many clusters as objects.
- The most similar objects are first grouped, and these initial groups are merged according to their similarities.
- As clustering progresses, similarity decreases, and all subgroups are eventually fused into a single cluster.
- One of the most common method of Agglomerative Hierarchical Method is ***Linkage Method***.

Algorithm for Agglomerative Hierarchical CLustering

Usually while doing an agglomerative clustering methodology for grouping of ***N*** objects, the below steps (or algorithm) is followed :

- Starting with ***N*** clusters, each one containing a single entity and an $N \times N$ symmetric matrix of distances (or similarities) $\mathbf{D} = d_{ik}$.
- The distance matrix for the nearest (most similar) pair of clusters are observed. Let the **distance between "most similar" clusters** *U* and *V* be d_{UV}
- After merge, clusters *U* and *V* as one newly formed cluster (UV), the entries in the distance matrix are updated:
  - **Deleting** the rows and columns corresponding to clusters *U* and *V*.
  - **Adding** a row and column for the distances between newly formed cluster *(UV)* and the remaining clusters.
- The above steps are **repeated** for a total of $N - 1$ times.

Linkage Method

In the above mentioned algorithm, different forms of metric $\mathbf{D}(d_{ik})$ gives rise to different types of linkages and hence different types of clustering methodologies. The three used, linkages are :

- **Single Linkage** : $d_{(UV)W} = min\{d_{UW}, d_{VW}\}$
- **Complete Linkage** : $d_{(UV)W} = max\{d_{UW}, d_{VW}\}$
- **Averagre Linkage** : $d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W}$, where $N_{(UV)}$ and N_W are the number of items in clusters (UV) and W, and d_{ik} is the distance between i^{th} object of cluster (UV) and k^{th} object of cluster W.

Clustering using single linkage:

$$
\begin{array}{c c}
 & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} &
\left[\begin{array}{ccccc}
0 & & & & \\
9 & 0 & & & \\
3 & 7 & 0 & & \\
6 & 5 & 9 & 0 & \\
11 & 10 & 2 & 8 & 0
\end{array}\right]
\end{array}
\longrightarrow
\begin{array}{c c}
 & \begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\
\begin{array}{c} (35) \\ 1 \\ 2 \\ 4 \end{array} &
\left[\begin{array}{cccc}
0 & & & \\
3 & 0 & & \\
7 & 9 & 0 & \\
8 & 6 & 5 & 0
\end{array}\right]
\end{array}
\longrightarrow
$$

$$
\begin{array}{c c}
 & \begin{array}{ccc} (135) & 2 & 4 \end{array} \\
\begin{array}{c} (135) \\ 2 \\ 4 \end{array} &
\left[\begin{array}{ccc}
0 & & \\
7 & 0 & \\
6 & 5 & 0
\end{array}\right]
\end{array}
\longrightarrow
\begin{array}{c c}
 & \begin{array}{cc} (135) & (24) \end{array} \\
\begin{array}{c} (135) \\ (24) \end{array} &
\left[\begin{array}{cc}
0 & \\
6 & 0
\end{array}\right]
\end{array}
$$

Reference: Applied Multivariate Statistical Analysis by Johnson and Wichern.

Dendrogram

The result of the previous single linkage clustering can be graphically observed using a **dendrogram** or a **tree diagram**. In hierarchical clustering, the dendrogram **illustrates the arrangement of the clusters** produced by the corresponding cluster analyses.

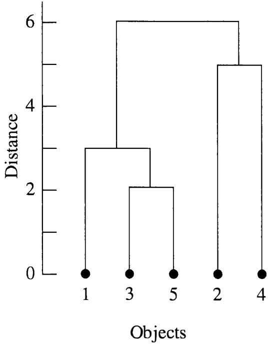
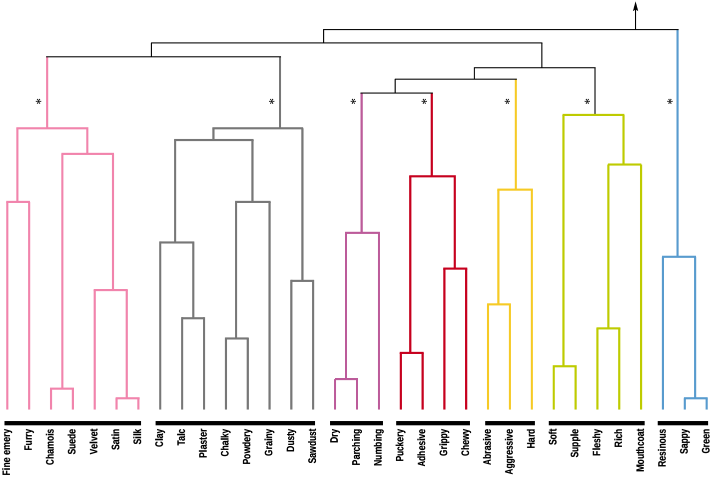


Figure 12.4 Single linkage dendrogram for distances between five objects.

Reference: Applied Multivariate Statistical Analysis by Johnson and Wichern.

- A dendrogram showing **proximity in terminology** as assessed by a combined panel of experienced **wine-tasters and wine-makers**. The asterisks show that this methodology reveals a number of logically consistent sub-groupings of terms.



Reference: Gawel, R., Oberholster, A., & Francis, I. L. (2000). A 'Mouth-feel Wheel': terminology for communicating the mouth-feel characteristics of red wine. Australian Journal of Grape and Wine Research, 6(3), 203-207.

Non-Hierarchical Clustering Methods

- These techniques are commonly designed to **group items, rather than variables**, into a collection of K clusters.
- The **number of clusters**, K, may either be **specified in advance or determined as part** of the clustering procedure.
- The Nonhierarchical clustering methods usually can start from any one of the two points :
  - an **initial partition** of items into groups.
  - an **initial set of seed points**, which will form the main nuclei of clusters.
- One of the **unbiased way** to start the clustering procedure is to, **randomly select seed points** from among the items or to randomly partition the items into initial groups.

K-Means Clustering

K-means is used to describe an algorithm that **assigns each item** to the **cluster having the nearest centroid (mean)**. The process mainly comprises of three steps :

- First of all, **partition** the items into K **initial** clusters. [Or, specify K initial centroids (seed points)]
- Now, for each of list of items, **assigning an item** to the **cluster** whose **centroid** (mean) is **nearest**. (It has to be observed, distance is usually computed using Euclidean distance with either standardized or unstandardized observations.).
- Further, **recalculate the centroid** for the cluster receiving the new item and also for the cluster losing the item.
- The above two steps are **repeated until no further reassignments** take place.

Clustering using K-means method:

We measured two variables X_1 and X_2 for each of four items A, B, C, and D. The data are given in the following table. The **objective** is to **divide these items** into $K = 2$ **clusters** such that the items **within a cluster are closer** to one another than they are to the items in different clusters.

Item	Observations x_1	x_2
A	5	3
B	−1	1
C	1	−2
D	−3	−2

→

Cluster	Coordinates of centroid $\bar{x}_1$	$\bar{x}_2$
(AB)	$\frac{5+(-1)}{2} = 2$	$\frac{3+1}{2} = 2$
(CD)	$\frac{1+(-3)}{2} = -1$	$\frac{-2+(-2)}{2} = -2$

→

Cluster	Coordinates of centroid $\bar{x}_1$	$\bar{x}_2$
A	5	3
(BCD)	−1	−1

→

	Squared distances to group centroids			
	Item			
Cluster	A	B	C	D
A	0	40	41	89
(BCD)	52	4	5	5

Reference: Applied Multivariate Statistical Analysis by Johnson and Wichern.

Good Luck for End-Sem !!