CLUSTER ANALYSIS

Seth Howells
Concepts of Statistics II
Week# 8 Assignment – Cluster Analysis
08/29/20

TABLE OF CONTENTS

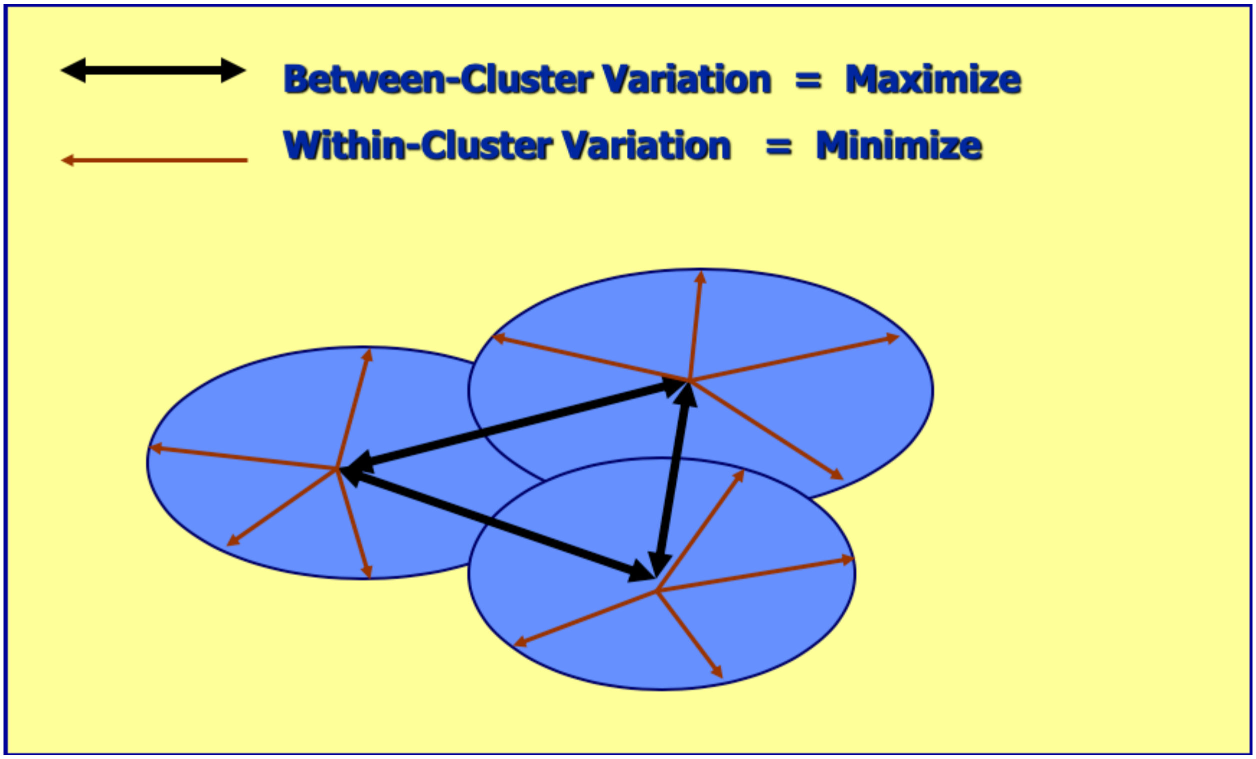
OVERVIEW

Cluster analysis is a widely used interdependence multivariate technique that aims to group objects by the characteristics they possess and as a result, the objects in the groups are more similar than the objects in other groups. This type of analysis has similar goals of factor analysis and multidimensional scaling, but it differs in its approach. The requirement of separating the observations into respective groups can be conceptually explained in two ways, as well as visually understood in *FIG 9-1*:

1. Maximize the separation of different groups by assessing the between-cluster variation
2. Minimize the range of distance from centroids by assessing the within-cluster variation

As in the figure below, the researcher's goal is to clearly define the groups so that overlapping is kept to a minimum. Since cluster analysis will create clusters even if the actual existence of structures with the data is not apparently present. Because cluster analysis completely relies on the data imported, the cluster solution is not generalizable.

FIG 9-1


Between-Cluster Variation = Maximize
Within-Cluster Variation = Minimize

QUESTIONS FOR RESEARCHERS

As in any analysis, the questions regarding the variables under investigation are critical for the proper results. Aside from the normal questions that a researcher asks before conducting a cluster analysis, such as measurement levels and variables at play, there are several other considerations:

- How to measure similarity?
- How to form clusters?
- How many groups are formed?

Given that cluster analysis is a multivariate technique, the primary goal is to partition the set of objects into <u>two or more</u> groups based on the similarity of characteristics (cluster variate). Additional considerations a researcher may want to address are:

- *Taxonomy description*: traditionally used for exploratory purpose to create classifications
- *Data simplification*: view observations as characteristics of a group as opposed to each observation being uniquely defined.
- *Relationship identification*: the cluster structure helps to identify relationships among the groups which can include qualitative examinations, which might be specifically important in industries where assessments of ideas can explain underlying structures found in the data.

INTER-OBJECT SIMILARITY

Inter-object similarity is referred to as an empirical measure of correspondence, or resemblance, between objects to be clustered. The similarity can be measured through three different approaches:

- *Correlational Measures*: correlation coefficients between a pair of objects measured on several variables. This is achieved by inverting the date matrix so that the columns represent the objects while the rows represent the variables.
  - High correlation: indicates a strong degree of similarity
  - Low correlation: indicates a lack of similarity

- *Distance Measures*: because observations are in groups, the group variation affects the group size, and consequently the likelihood of overlapping areas of the different groups. There are several ways to measure distance with each being more applicable to some scenarios or analyses than others. Contrasting with *correlational measures,* the lower the distance indicates a stronger similarity while larger distances indicates a lack of similarity. All *distance measures* are proximity based as shown in *FIG 9-1.*
  - **Euclidean distance**: derived as proof of the Pythagorean Theorem, measures a straight line of coordinates on a plane. The distance of the points is the length of the hypotenuse of a right triangle. See *FIG 9-2.*

  - **Squared (or absolute) Euclidean distance**: similar to the Euclidean distance above, the absolute distance differs in that it does not have the square root as seen in the equation of *FIG 9-2* which allows the researcher to make quicker computations. It is typically applicable to measure the distance of the centroid.

  - **City-block (Manhattan) distance**: not based on Euclidean distance, this measurement uses the sum of absolute differences of variables, as shown in *FIG 9-3.* While Manhattan distance is the easiest to compute, it also yields less accurate measurements as shown in *FIG 9-4.*

  - **Mahalanobis distance D^2**: measures the correlation among variables in a way that weights each variable equally. This measurement of distance relies on standardized variables. This is applicable when variables are highly intercorrelated.

- *Association Measures*: used to compare similarities of objects that are nonmetric (nominal or ordinal). Because of the nonmetric feature, this measurement is applied to qualitative data that tend to compare ideas or theoretical questions.

HIERARCHICAL VS NON-HIERARCHICAL CLUSTERING

Hierarchical: involves a treelike structure based on *n-1* clustering*,* where *n* is the number of observations. There are two basic types of hierarchical cluster producers:

- *Agglomerative methods*: each observation starts out as its own cluster and joins the two most similar clusters as more observations are entered.
  - Single Linkage (nearest neighbor)
  - Complete Linkage (farthest neighbor)
  - Average Linkage
  - Centroid Method
  - Ward's Method

- *Divisive methods:* in contradistinction to agglomerative methods, each observation starts out in the same cluster and is divided into different clusters based the differences until all observations are in its own cluster.

Non-hierarchical: does not produce treelike constructions, unlike hierarchical clustering, and rather assigns observations into a specified number of clusters.

  - Sequential Threshold
  - Parallel Threshold
  - Optimization

FIG 9-2: Euclidean distance

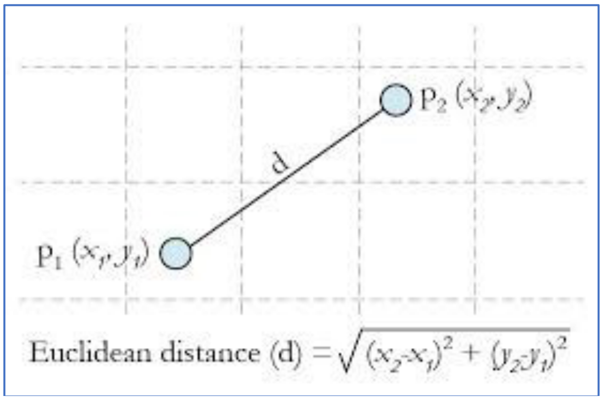


FIG 9-3: City-Block (Manhattan) distance

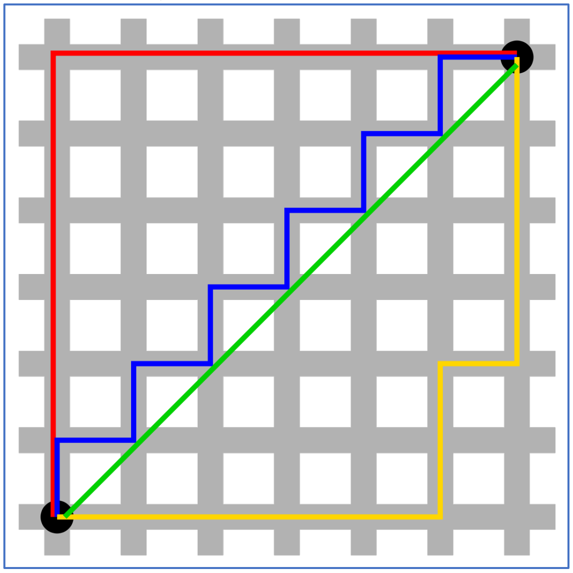
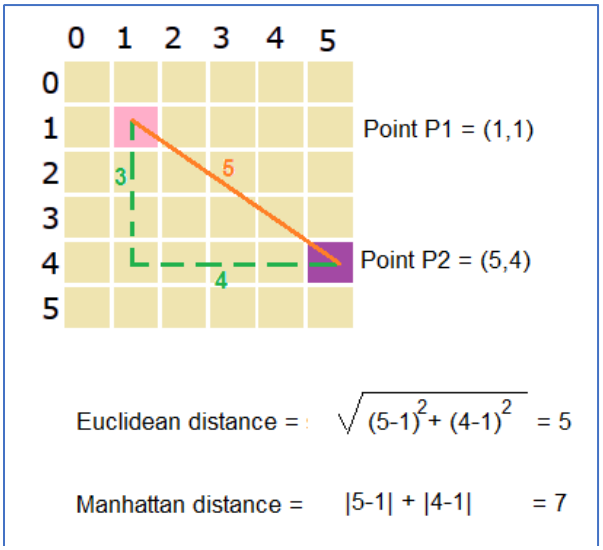


FIG 9-4: Euclidean vs Manhattan distance


0 1 2 3 4 5
0 1 2 3 4 5
3
5
4
Point P1 = (1,1)
Point P2 = (5,4)
Euclidean distance = $\sqrt{(5-1)^2 + (4-1)^2}$ = 5
Manhattan distance = |5-1| + |4-1| = 7