Clustering

Clustering is the method of grouping objects or entities so that objects in same cluster are more similar than objects in other clusters. This is usually done based on some *similarity* among the objects within a cluster.

Can you divide the state of Minnesota into 3 clusters?

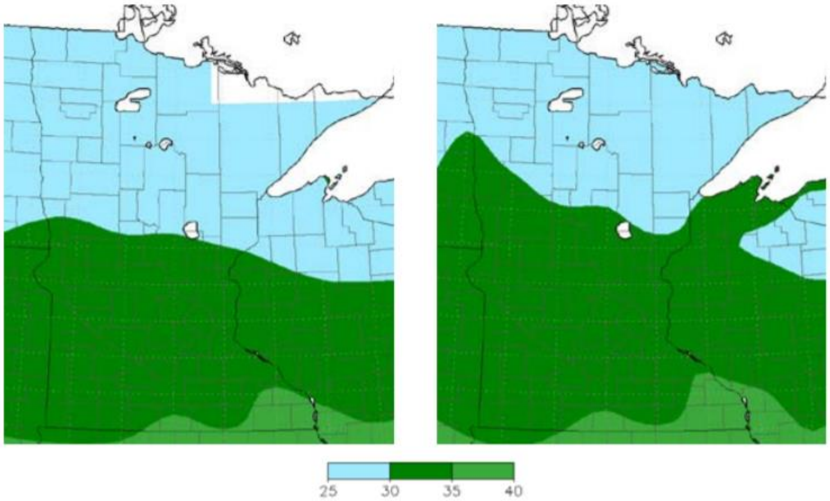

Figure 1: Annual average minimum temperature (°F) across Minnesota for 1900- 1959 (LEFT) and 1960-2013 (RIGHT). DATA AND IMAGE SOURCE: MRCC, 2014.
https://www.health.state.mn.us/communities/environment/climate/docs/mnprofile2015.pdf

We may cluster the whole geographic area based on annual average minimum temperature for those years. Here, we have formed three clusters. Cluster one (temperature 25 – 30˚F), comprises of area in northern parts of Minnesota, cluster two (temperature 30 – 35˚F) includes most of the southern Minnesota while cluster three (temperature 35 – 40˚F) can be seen at the bottom of Minnesota, covering the least area. The similarity measure that helped us in forming these clusters is the temperature magnitude.

Traditionally, an optimal clustering result is the one where we have minimized the difference/dissimilarity within a cluster and maximized it among different clusters.

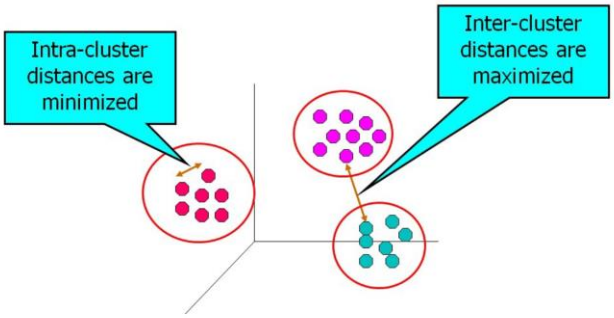
Intra-cluster distances are minimized
Inter-cluster distances are maximized

Figure 2: Inter and Intra cluster distances.

Ways to define clusters

- **Partitional** and **Hierarchical:**
  Partitional: It is a simple a division of the set of data objects into non-overlapping subsets (clusters) such that

each data object is in exactly one subset.
Hierarchical: we permit clusters to have sub clusters, then it is hierarchical clustering.

- **Exclusive** and **Overlapping (**or **fuzzy):**
  Exclusive: each object is assigned to a single cluster.
  Overlapping: an object may belong to more than one cluster.
  Fuzzy: In fuzzy clustering, a point belongs to every cluster with probabilistic weight between 0 and 1.
- **Complete** and **Partial:**
  Complete: all objects are clustered.
  Partial: not all objects are clustered. E.g., sometimes we may decide to eliminate some weird data or anomalies that might be errors.

Classify following examples by their type of clustering - *hierarchical or partitional; exclusive, overlapping, or fuzzy; and complete or partial*.

Question 1: A nutrionist asks all her patients for their water level intake. Based on the survey results, he groups all people into clusters: high level intake, medium level intake and low-level intake. This will be: *Partitional, Exclusive, Complete*.

1. You want to group all the movies on Netflix into several genre/topics, each of which can have several subtopics.

2. You want to group all your purchased items from Target / Walmart based on brand.

3. Your teacher wants to group all students based on students' affiliation to different cultural activities.

Clustering Algorithms

Most popular clustering algorithms are K-means and its variants, hierarchical clustering and density-based clustering. We will briefly discuss K – means and Hierarchical clustering.

K – means

It is a partitional clustering approach where we group data points into 'K' clusters (where K needs to be chosen).

- Based on choice of K, we select K initial centroids.
- Form K clusters by assigning each point to one of the K clusters, usually based on its distance to the centroids.
- Recompute centroids as a central tendency value for each cluster based on data points that you clustered in previous step.
- Again, assign each data point to closest centroid. Repeat this procedure, until the cluster composition stops changing.

Question 2: Your class has four students and they are scored in two subjects. Your teacher wants to group the students into 2 clusters.

S.No.	Subject 1	Subject 2
1	1	1.5
2	1.2	2.5
3	4	2
4	3.5	4.6

Centroids are initialized at (1,1) for cluster 1 and (3,3) for cluster 2.
Points have been visualized on next page.

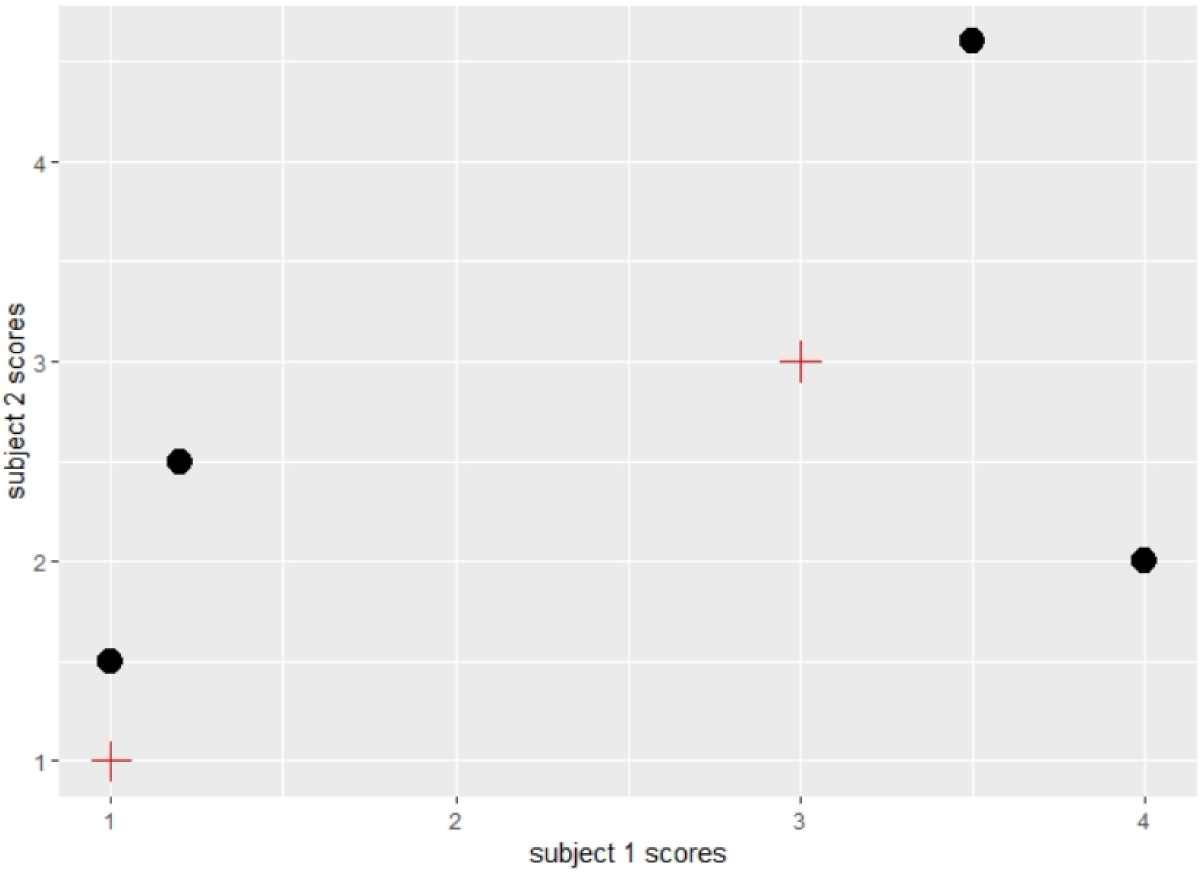

Figure 3: Question 2 Points Visualization. Centroids (red plus signs) and data points (solid black circles)

Answer the following questions:

1. After looking at the above plot (without implementing K-means), how will you cluster the datapoints into two groups (each cluster having equal number of data points)? Where would the cluster centroids be?

2. Implement K means, where you choose Euclidean distance (= $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$) to be the proximity measure. Repeat till the cluster centers converge. Draw a graphical representation of your results. First table for Euclidean distances are given:

	Point 1	Point 2	Point 3	Point 4
Centroid 1 (1,1)	0.5	1.513275	3.162278	4.382921
Centroid 2 (3,3)	2.5	1.868154	1.414214	1.676305

3. Did your solutions in question 1 and 2 match? Why or Why not?

Hierarchical Clustering

Hierarchical clustering, as we discussed earlier, is the one where we allow clusters to have sub clusters. We visualize such clusters in a tree called dendrogram. Here, the total number of clusters depend on where you decide to cut the dendrogram. Lets, work through an example for better understanding.

Question 3: We have three points, P1, P2 and P3, whose similarity is quantified in the 'similarity matrix' as shown:

	P1	P2	P3
P1	1.00	0.10	0.41
P2	0.10	1.00	0.64
P3	0.41	0.64	1.00

Draw a dendrogram. At what point do you cut to get two clusters?