Lecture 11 - 03/10/24

Cluster Analysis

- Clustering Analysis is about discovering groupings in the data.
- Clustering Methods are unsupervised methods.
  - Give an unlabeled dataset
  - Group data into clusters
- Part of the analysis is → choosing K with K being # of clusters.
- Visualizing data sometimes can be more effective.

Clustering Methods

- We seek to partition data into a # of clusters → $C_1, \dots, C_K$.
- Assumption: Each datapoint can only belong to a single cluster.
- Datapoints in the same cluster are similar & data points in ≠ clusters are dissimilar!
- Methods to cover:
  - K-means clustering
  - Hierarchical clustering

K-mean clustering

- First choose a # (K)
- Partition data into K clusters $C_1, \dots, C_K$
- All clusters are non-empty & all data points belongs to exactly one cluster!
  - Seek a partitioning that minimizes total within-cluster variation:

$$\underset{C_1,\dots,C_K}{\text{minimize}} \sum_{k=1}^{K} W(C_k)$$

  - Usually use "similarity" to correlate with distance so popular $W(C_k)$ is:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \|x_i - x_{i'}\|^2$$

What about Centroids?

- We do something a bit different:
  - We measure the within-cluster variation as the summed distance between data points & cluster centroid r_k:

$$W_{C_k}(r_k) = \sum_{i:x_i \in C_k} \|x_i - r_k\|^2$$

  - If we know which obs belong to a C_k, then the centroid is chosen as cluster mean: $r_k = \frac{1}{|C_k|} \sum_{i:x_i \in C_k} x_i$

The K-means algorithm

Initialize the K cluster centres $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_k$.
(e.g. by choosing K random points in the data)

1. Assign all data points to their nearest cluster center.

$$\boldsymbol{x}_i \in C_k \Leftrightarrow \forall j \neq k : \|\boldsymbol{x}_i - \boldsymbol{r}_k\|^2 < \|\boldsymbol{x}_i - \boldsymbol{r}_j\|^2$$

2. Move the cluster center to the mean of the cluster.

$$\boldsymbol{r}_k = \frac{1}{|C_k|} \sum_{i:\boldsymbol{x}_i \in C_k} \boldsymbol{x}_i$$

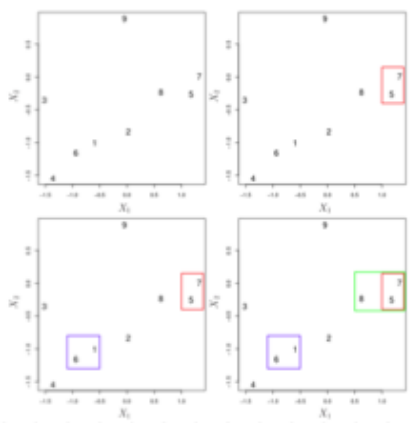
This is iterated until centres stop moving.

- THE ALGO STOPS IN A FINITE # OF ITERATIONS
- WE TEND TO STOP WHEN NOTHING CHANGES
- THE ALGO CAN CONVERGE TO LOCAL OPT
- CHOICE OF K HIGHLY IMPACTS WHICH DATA POINTS ARE CLUSTERED TOGETHER
- ANYTHING THAT IMPACTS DISTANCE, IMPACTS CLUSTERING:
  - * HIGH DIMENSIONALITY.
  - * OUTLIERS.
  - * SCALE OF VARIABLES

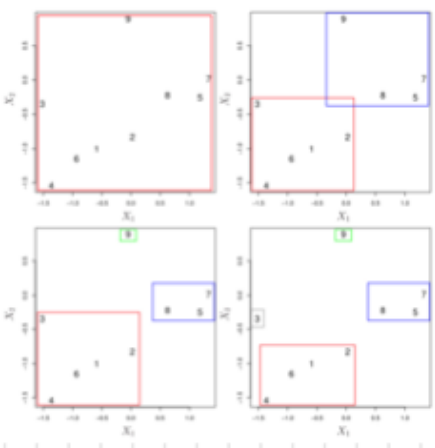
HIERARCHICAL METHODS

- HAVE THE FOLLOWING PROPERTY:
  - * CLUSTERING WILL LOOK SIMILAR IF YOU VARY THE # OF CLUSTERS BY A LITTLE.
- ALL K CLUSTERS WILL BE SUBSETS OF CLUSTERS IN A CLUSTERING WITH FEWER CLUSTERS
- TWO OBVIOUS STRATEGIES:
  - HAVING K CLUSTERS, MERGE TWO MOST SIMILAR ONES TO GET K-1 CLUSTERS.
  - HAVING K CLUSTERS, DIVIDE ONE TO GET K+1 CLUSTERS.
- HIGH DISSIMILARITY IS BEST FOR THE CLUSTERS.

Agglomerative/bottom-up/merging
Agglomerative strategies start with each observation as a separate cluster and repeatedly merge the two clusters with the smallest dissimilarity.



Divisive/top-down/splitting
Divisive strategies work in the opposite direction: they start with all observations in one cluster and recursively split one cluster.



- HOW TO MEASURE DISSIMILARITY BETWEEN CLUSTERS?
  - MOST OF THE TIME DISSIMILARITY IS BASED ON DISTANCE BETWEEN TWO DATA POINTS

Single-Link Clustering

$$D(C_i, C_j) = \min \{d(x, y) \mid x \in C_i, y \in C_j\}$$

Group-Average Clustering

$$D(C_i, C_j) = \text{mean} \{d(x, y) \mid x \in C_i, y \in C_j\}$$

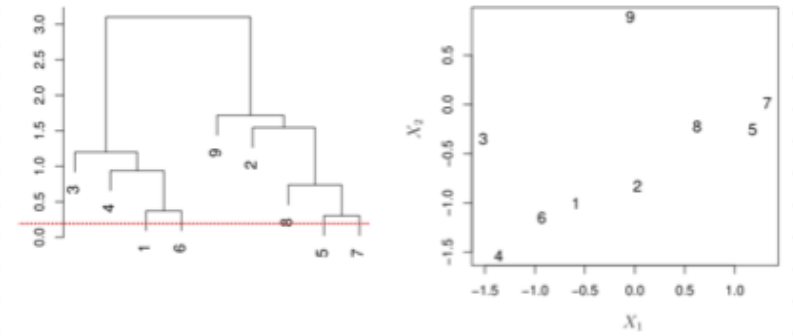
Complete-Link Clustering

$$D(C_i, C_j) = \max \{d(x, y) \mid x \in C_i, y \in C_j\}$$

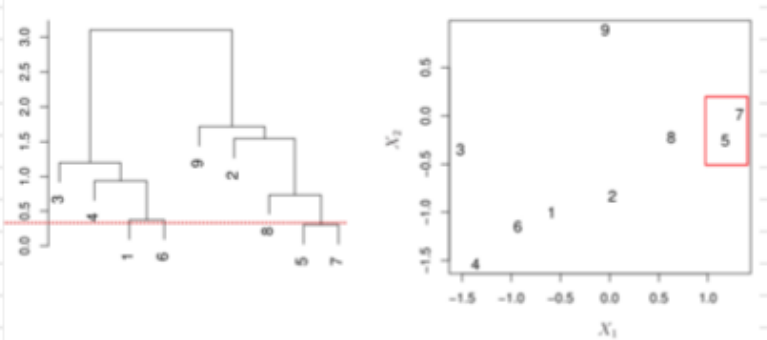
- HIERARCHICAL CLUSTERING GIVES US DENDOGRAMS

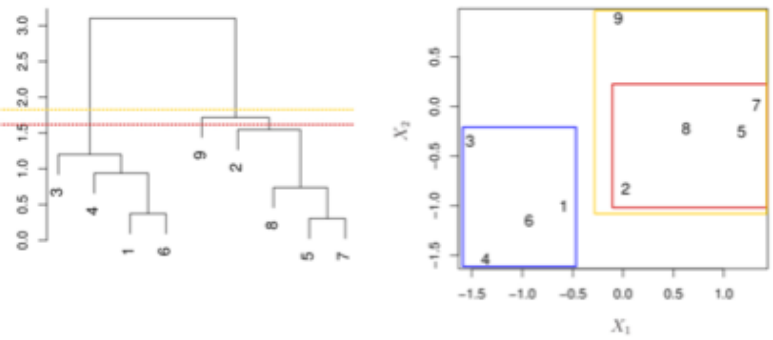
- A DENDOGRAM VISUALIZES PARTITIONS FOR CHOICES OF:
  - K=1 (TOP)
  - K=n (BOTTOM)
- MOVING FROM THE BOTTOM TOWARDS THE TOP, PAIRS OF CLUSTERS FUSE.

- A HORIZONTAL CUT IN A DENDOGRAM CORRESPONDS TO A PARTITIONING



IF A CUT INTERSECTS A LINE, ALL NODES "DOWNSTREAM" FROM THAT INTERSECTION ARE GROUPED.





DIVISIVE STRATEGY!

- AMONG ALL CLUSTERS & ALL POSSIBLE SPLITS OF THE CLUSTER:
  - CHOOSE COMB OF CLUSTER & SPLIT THAT GIVES THE GREATEST DISSIMILARITY BETWEEN THE TWO.
- MORE COMPLEX THAN AGGLOMERATIVE METHODS CAUSE:
  - DECIDE WHICH TO DIVIDE
  - DECIDE HOW TO DIVIDE

THERE ARE:

$$\frac{2^N - 2}{2} = 2^{N-1} - 1$$

WAYS OF SPLITTING A CLUSTER WITH N OBS IN TWO NON-EMPTY GR