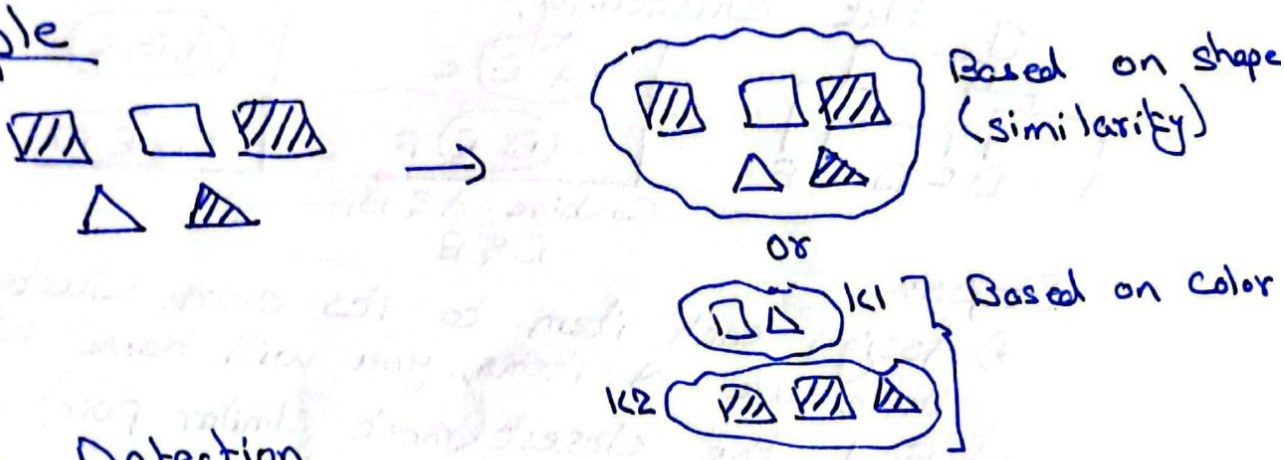
Week 2

(Unsupervised Learning)

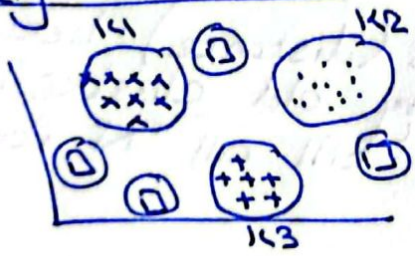
→ For data that is
- unclassified
- Unlabelled
- More complex
- Moderately accurate but reliable results.

→ Used for
- finding patterns (clustering)
- Anomaly detection
- ~~[illegible]~~

Example



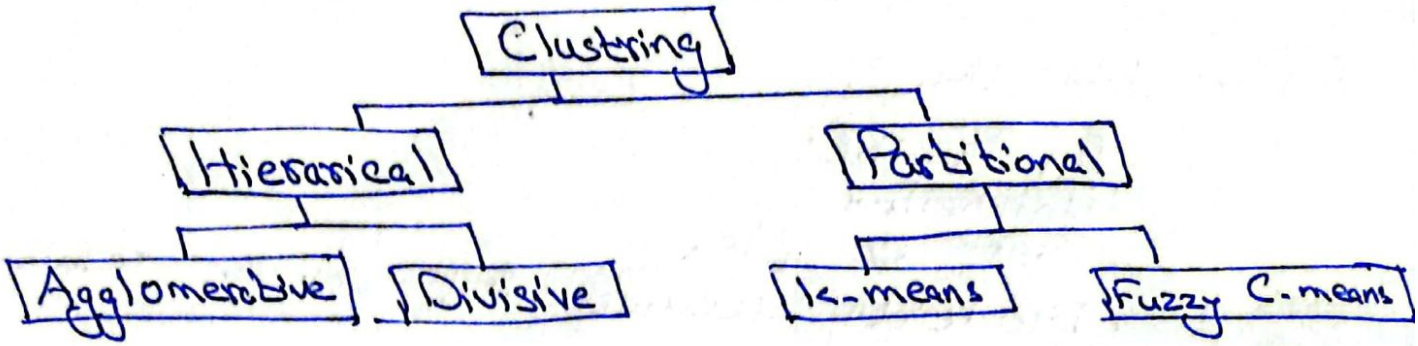
Anomaly Detection



→ Fault detection
→ Intrusion "
→ Systen fault "

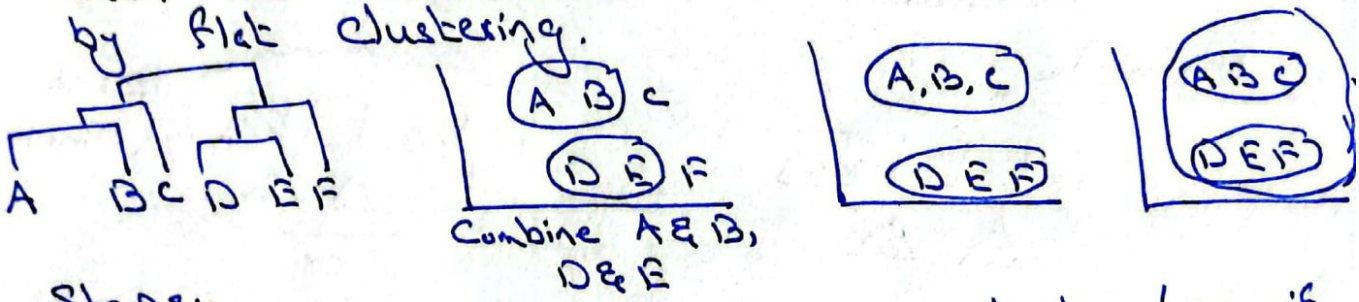
→ **Need for clustring**
in unlabelled, structured & unstructured data
- To determine intrinsic grouping.
- To organize data into clustering showing the internal structure of data.
- To partition the data points.
- To understand and exhibit value from large sets of structured & unstructured data.

→ Types of Clustering:



→ Hierarichal Clustering

- Occupys hierarichy, a structure more informative than the unstructured set of clusters returned by flat clustering.



Steps:-

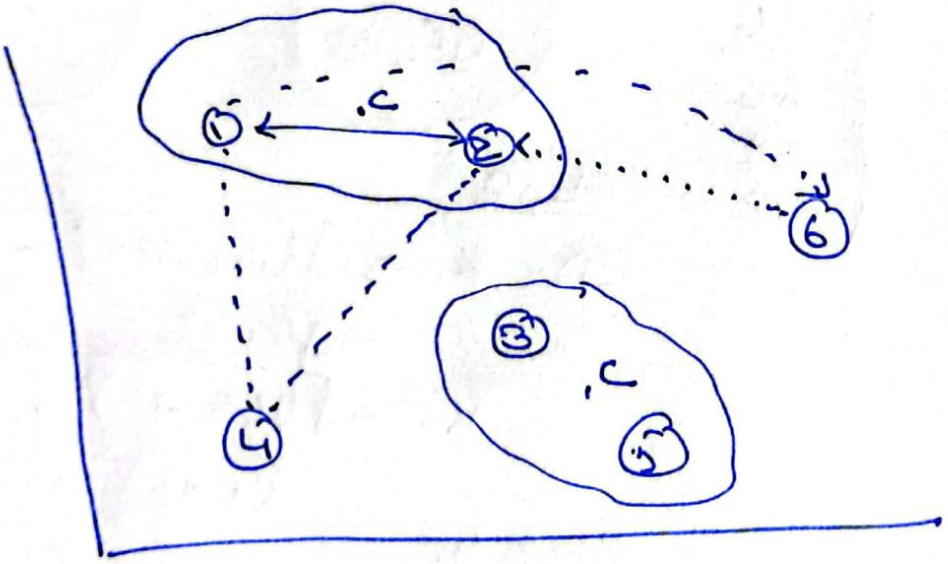
1) Assign each item to its own cluster, (eg. if there are N items, you will have N clusters)
2) Find the closest (most similar pair) of clusters & combine them.
3) Compute similarities (distance) between the new clusters & every old cluster. Then combine
4) Repeat step 2 & 3 till all 'N' items are in single cluster.

→ Partitional Clustering

- Division of data into non-overlapping clusters where a data object is only in one set (cluster).

→ Distance Measure

1) Complete Linkage Clustering
   ↳ maximum possible distance between points.
2) Single Linkage Clustering
   ↳ Minimum possible distance betw points.
3) Mean Linkage Clustering
   ↳ find all possible pair-wise distance between pair-wise two clusters & then calculate the average distance.
4) Centroid Linkage clustering
   ↳ Find centroids of each cluster & calculate the distance between them.



→ K-Means Clustering

Step 1: Choose clusters ($K=2$ e.g. $K1, K2$) centroids.

Step 2: Calculate Euclidean Distance of each point(item)

$$ED = \sqrt{(x_p - x_c)^2 + (y_p - y_c)^2}$$

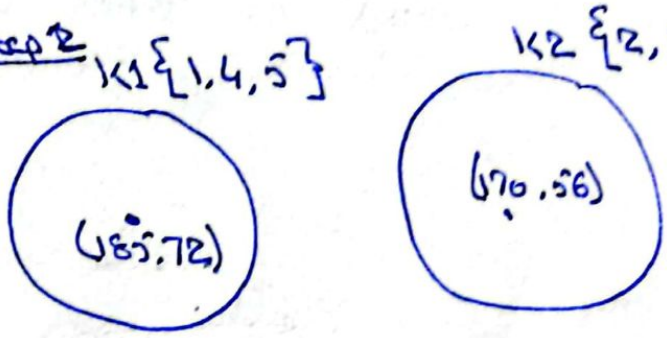
Step 3: Put the point(item) with smallest (nearest) ED in respective cluster.

Step 4: Recalculate the respective cluster's centroid with new addition.

Step 5: Repeat step 2 ~ 4.

S.No	Height	Weight
1.	185	72
2.	170	56
3.	168	60
4.	179	68
5.	182	72

Step 2



Step 3

$$ED = K_1 \rightarrow \sqrt{(168-185)^2 + (60-72)^2}$$

$$\rightarrow \boxed{20.8}$$

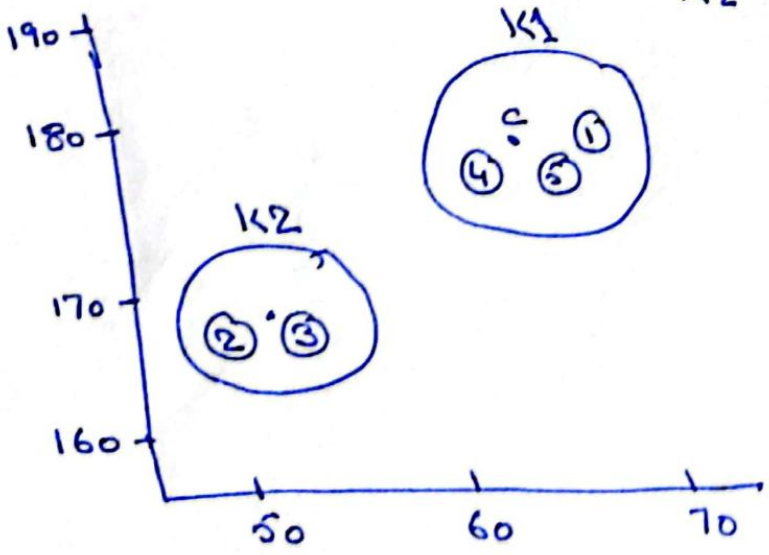
$$K_2 \rightarrow \sqrt{(168-170)^2 + (60-56)^2}$$

$$\boxed{4.48}$$

Step 4:

K_2 Centroid $= \left(\frac{170+168}{2}, \frac{60+56}{2}\right)$

$= (169, 58)$



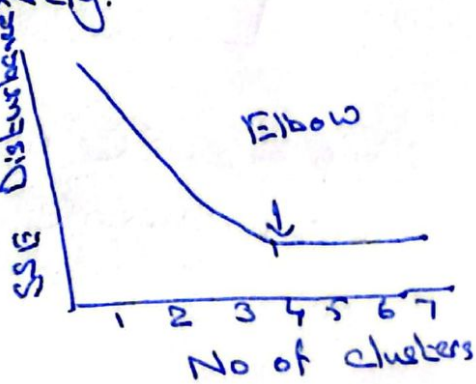
→ Optimum Cluster Number:

1) With-in-sum-of-Square (WSS): Total distance of data points from their respective centroids.
2) Total-sum-of-Squares (TSS): Total distance of data points from global mean for a given data point. This is constant.
3) Between-Sum-of-Squares (BSS): Total weighted distance of various cluster centroids to the global mean of data.
4) R^2: R-square is the total variance (BSS/TSS)
5) Sum-of-Square-Error (SSE): is Euclidean distance of each point to it's closest centroid.

→ How many clusters:

- It is a fundamental issue in k-means clustering.
- If SSE plotted against k, you will see the error decreases as k increases beez their size decreases & hence distortion is also small.
- The goal of the 'Elbow Method' is to choose k, where SSE decreases abruptly.

from k1 ..... k20
(WSS for each value of k)



→ Three most popular
1) Silhouette Coefficient
2)
3)

Silhouette Coefficient (SC)

→ We have to compute

Step 1) SC of each point

$$1 - \frac{a}{b}$$

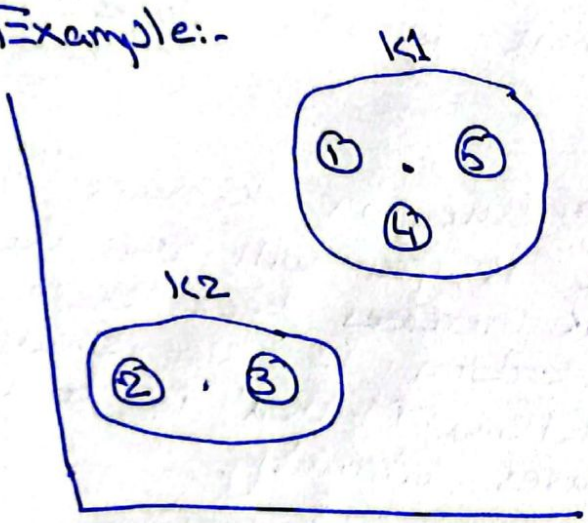
(a: avg dist of a point to all other points in a cluster)

(b: Minimum avg dist of a point to all points in another cluster)

Step 2) SC of each cluster

Step 3) of all clusters.

→ Example:-



Step 1

$$a = \frac{\{(1 \to 5) + (4 \to 5)\}}{2^{(\text{No. of points})}}$$

$$b = \frac{\{(1 \to 2) + (1 \to 3)\}}{2}$$

SC of (1) = $1 - \frac{a}{b}$

Step 2: SC of each cluster

Let's suppose SC of (2) & (3) is x & y. repectively

SC of K2 = $\frac{x+y}{2}$

Step 3: Overall SC

SC ~~of K1 & K2~~ = $\frac{(\text{SC of K1}) + (\text{SC of K2})}{2}$