MACHINE INTELLIGENCE

UNIT - 4

Clustering

feedback/corrections: vibha@pesu.pes.edu

VIBHA MASTI

CLUSTERING

- Unsupervised ML
- Group objects such that all objects in a group are similar to each other
- Intention:
  - min intra-cluster distance
  - max inter-cluster distance
- Eg: doc segmentation, recommendation systems, customer grouping

Types

1. Hierarchical
2. Partitional k-means
3. Density-based DBSCAN

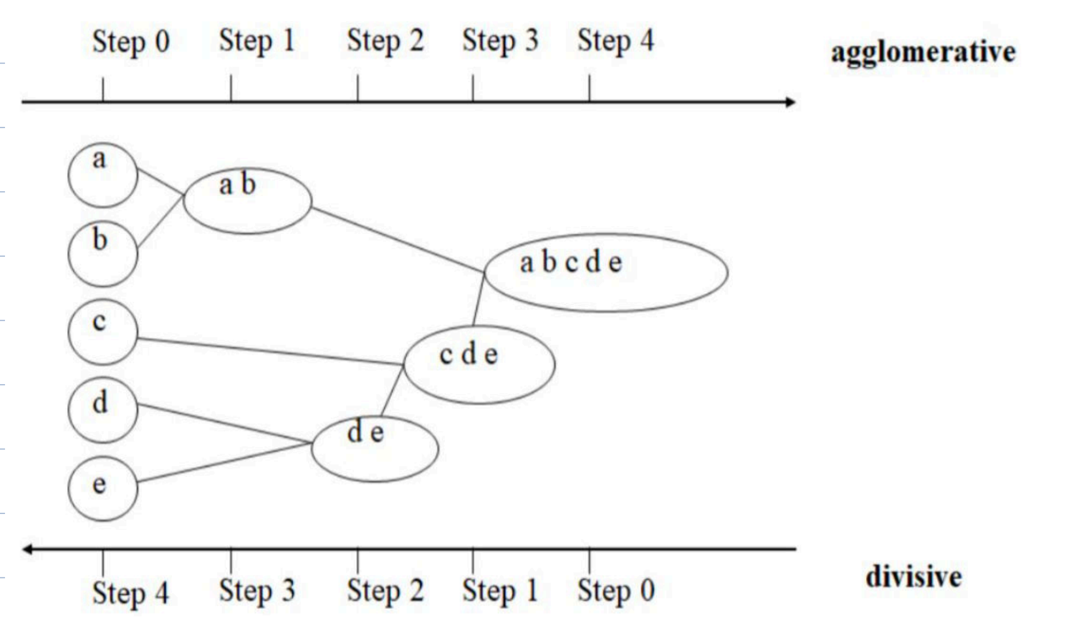
1. Hierarchical

(a) Agglomerative

- start with no. of clusters = no. of instances
- Group together similar points based on similarity measure
- Each step, only cluster one instance at a time (not in parallel)
- Designer decides no. of clusers

(b) Divisive

- Initally assume one single cluster for all the instances
- Divide into 2 clusters (using centroids)



2. Partional Clustering

(a) Hard Clustering

- Assign every point to exactly one cluster
- K-means

(b) Soft Clustering

- Point belongs to a cluster with some likelihood between 0 and 1
- Eg: GMM

3. Density-Based

- DBScan

DISTANCE METRICS

1. Minkowski

$$d_{x,y} = \left(\sum_{i=1}^{n} |x_i - y_i|^r \right)^{1/r}$$

(i) Inter-Cluster Distance

- Minimum distance
  - distance b/w pair of points from two clusters closest to one another
  - single linkage

- Maximum distance
  - distance b/w pair of points from two clusters farthrest from each other
  - complete link

- Average distance
  - group average

- Centroid distance
  - distance b/w cluster centroids

Dendogram

- Diagrammatic representation of hierarchical clustering
- Two points/clusters that are part of a larger cluster are represented by a single branch of the dendogram
- The height

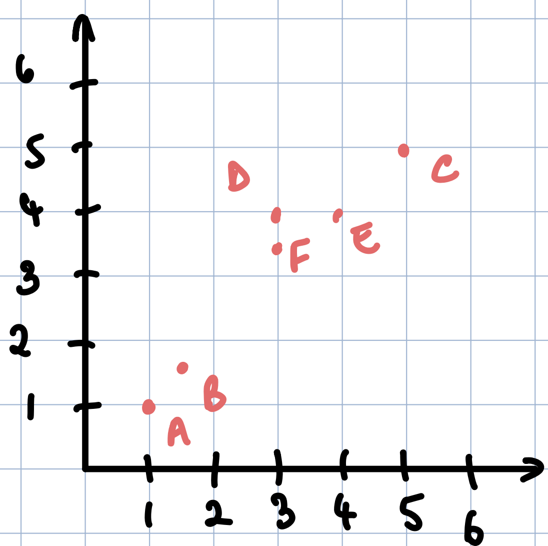
Agglomerative Clustering

- AGNES: Agglomerative Nesting
- Decide no. of clusters using
  - domain knowledge
  - elbow method

- compute distance matrix of all points
- Find pair with min dist
- combine into cluster, recompute distances

Q: Use Euclidean Distance, Single linkage distance
Data: 6 points, 2 attrs each

	X1	X2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5



Distance matrix

	A	B	C	D	E	F
A						
B	0.71					
C	5.66	4.95				
D	3.61	2.92	2.24			
E	4.24	3.54	1.41	1.00		
F	3.20	2.50	2.50	0.50	1.12	

	A	B	C	D,F	E
A					
B	0.71				
C	5.66	4.95			
D,F	3.20	2.50	2.24		
E	4.24	3.54	1.41	1.00	

	A,B	C	D,F	E
A,B				
C	4.95			
D,F	2.50	2.24		
E	3.54	1.41	1.00	

Time & Space Complexity

- Space: $O(N^2)$
- Time: $O(N^3)$
  - N steps (N clusters to 1 cluster)
  - N^2 to update proximity matrix

Limitations

- Cannot undo a clustering step
- Sensitivity to noise / outliers

- Difficultly handling different sized clusters
- Breaking large clusters

K-Means Clustering

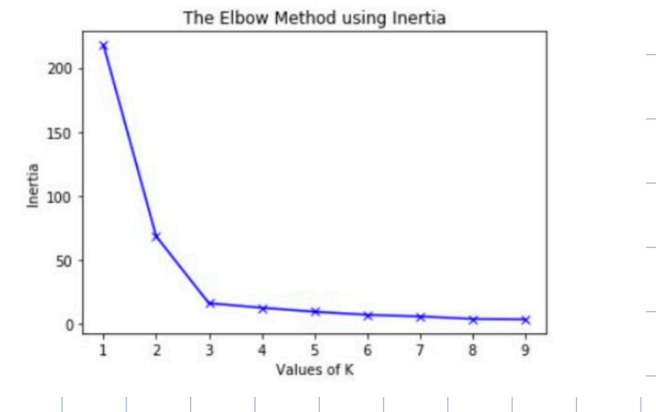
- Partitional clustering
- WCSS : within cluser sum of squares (inertia)

$$J = \sum_{i=1}^{N} \sum_{k=1}^{k} w_{ik} \left\| x_i - \mu_k \right\|^2$$

w_{ik} = indicator var for presence of x_i in cluster k

- For each cluster, SS distances from points in that cluster to cluster centroid

Value of K



elbow method

Q: KMC

Object	X (weight index)	Y (pH)
Medicine A	1	1
B	2	1
C	4	3
D	5	4

- Choose $k=2$
- choose 2 random centroids

$$c_1 = (1, 1)$$
$$c_2 = (2, 1)$$

Distance matrix M_1

	c_1	c_2
A	0	1
B	1	0
C	3.61	2.83
D	5	4.24

Hard clustering

	C_1	C_2
A	1	0
B	0	1
C	0	1
D	0	1

New C_1 & C_2

$$C_1 = (1, 1)$$

$$C_2 = \left(\frac{2+4+5}{3}, \frac{1+3+4}{3} \right) = \left(\frac{11}{3}, \frac{8}{3} \right)$$

Time Complexity

- $O(K\,I\,N\,D)$

Points to Remember

- Depends on initial points
- spherical (can use kernel trick)
- same size, density
- slow
- can use MR
- scale sensitive

Bisecting k-Means

- k means + divisive hierarchical clustering
- All in single cluster
- 2-means
- Pick one with larger WCSS
- Perform 2 means
- Efficient when k large
- Time : $O(\log(k)\, I N D)$