Module 6

Clustering Methods

What is Clustering ?

* Clustering or cluster analysis is an unsupervised learning technique.
* It is the task of grouping a set of objects in such a way that objects in the same group (cluster) are more similar to each other than to those in other groups.

* Various algorithms are:-
  - → K-mean clustering
  - → Hierarchical clustering
  - → Expectation-Maximization algm.
  - → Density based clustering.

K-means clustering

In K-means clustering, the given data points are grouped into K-clusters, based on the similarity of the data-points.

Algorithm

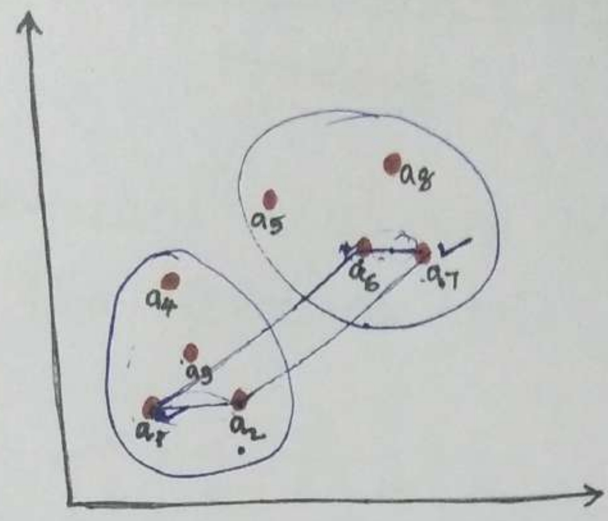
Step 1: Randomly select K cluster centers, $\bar{v}_1, \bar{v}_2, \ldots \bar{v}_K$

Step 2: Calculate the distance between each data point a_j and each cluster centers $\bar{V}_i$

Step 3: Assign each data point a_j to the cluster center $\bar{v}_i$ for which the distance $\|\bar{a}_j - \bar{v}_i\|$ is minimum.

Step 4: Recalculate each cluster center by taking the average of cluster's data points.

Step 5: Repeat from step 2 to step 5 untill the recalculated cluster centers are same as previous or No reassignment of data points happend.



$k = 2$

cluster 1 of a_1

$= \{a_2, a_1, a_3, a_4\}$

cluster 2 of a_7

$= \{a_6, a_7, a_8, a_5\}$

Distance between data points

We assume that each data point is a n-dimensional vector.

The distance between two data points

$$\bar{x} = (x_1, x_2, \dots x_n)$$

and

$$\bar{y} = (y_1, y_2, \dots y_n)$$

is defined as

$$\|\bar{x} - \bar{y}\| = \sqrt{(x_1 - y_1)^2 + \dots + (x_n - y_n)^2}$$

K-means clustering

Problem

Q.1. Use k-means clustering algorithm to divide the following data into two clusters.
K=2

x_1	1	2	2	3	4	5
x_2	1	1	3	2	3	5

Ans) step 1: Choosing randomly 2 cluster centers.
Say $V_1 = (2,1)$ $V_2 = (2,3)$

step 2: Finding the distance b/w the cluster centers and each data points.

Datapoint	Distance from V_1 (2,1)	Distance from V_2 (2,3)	Assigned center
a_1 (1,1)	1	2.24	V_1
a_2 (2,1)	0	2	V_1
a_3 (2,3)	2	0	V_2
a_4 (3,2)	1.41	1.41	V_1.
a_5 (4,3)	2.83	2	V_2
a_6 (5,5)	5	3.61	V_2

step 3: Cluster 1 of V_1 : $\{a_1, a_2, a_4\}$
Cluster 2 of V_2 : $\{a_3, a_5, a_6\}$

Step 4 : Recalculate the cluster centers.

$$V_1 = \frac{1}{3}\left[(1,1) + (2,1) + (3,2)\right]$$

$$= \frac{1}{3}(6,4)$$

$$= (2, 1.33)$$

$$V_2 = \frac{1}{3}\left[a_3 + a_5 + a_6\right]$$

$$= \frac{1}{3}\left[(2,3) + (4,3) + (5,5)\right]$$

$$= \frac{1}{3}(11, 11)$$

$$= (3.67, 3.67)$$

Step 5 : Repeat from step 2 until we get same cluster center or same cluster elements as in the previous iteration.

Distance table :-

Data point	Distance from V_1 (2, 1.33)	Distance from V_2 (3.67, 3.67)	Assigned center
a_1 (1,1)	1.05	3.78	V_1
a_2 (2,1)	0.33	3.15	V_1
a_3 (2,3)	1.67	1.8	V_1
a_4 (3,2)	1.204	1.8	V_1
a_5 (4,3)	2.605	0.75	V_2
a_6 (5,5)	4.74	1.88	V_2

Cluster 1 of $V_1 = \{a_1, a_2, a_3, a_4\}$

Cluster 2 of $V_2 = \{a_5, a_6\}$.

Recalculating the cluster centers

$$V_1 = \frac{1}{4}[a_1 + a_2 + a_3 + a_4]$$

$$= \frac{1}{4}[(1,1) + (2,1) + (2,3) + (3,2)]$$

$$= \frac{1}{4}[8,7] = \underline{(2, 1.75)}$$

$$V_2 = \frac{1}{2}[a_5 + a_6]$$

$$= \frac{1}{2}[(4,3) + (5,5)]$$

$$= \frac{1}{2}(9,8) = \underline{(4.5, 4)}$$

So clusters elements and centers are not same as in the previous.

Distance ~~table~~ between cluster centers and data-points :-

Data points	Distance from V_1 (2, 1.75)	Distance from V_2 (4.5, 4)	Assigned center
a_1 (1,1)	1.25	4.61	V_1
a_2 (2,1)	0.75	3.9	V_1
a_3 (2,3)	1.25	2.69	V_1
a_4 (3,2)	1.03	2.5	V_1
a_5 (4,3)	2.36	1.12	V_2
a_6 (5,5)	4.42.	1.12.	V_2

cluster 1 of v_1 : $\{a_1, a_2, a_3, a_4\}$.

cluster 2 of v_2 : $\{a_5, a_6\}$.

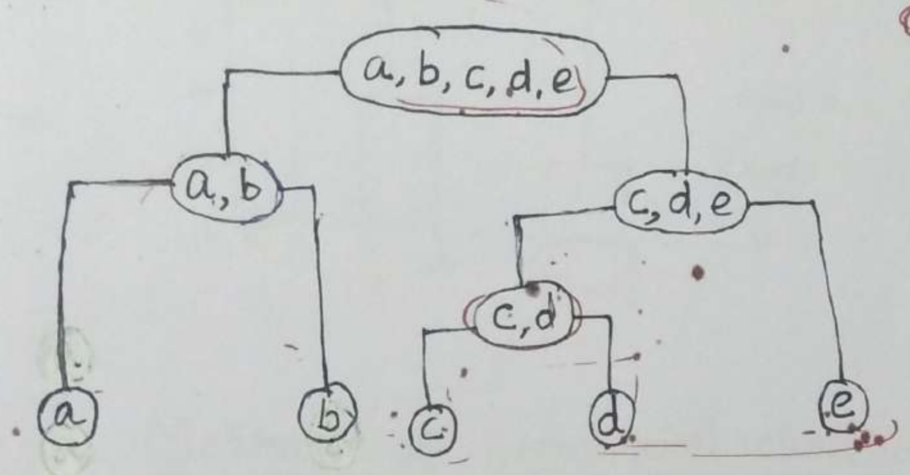
Cluster elements are same as in the previous iteration..

cluster 1 : $\{(1,1), (2,1), (2,3), (3,2)\}$

cluster 2 : $\{(4,3), (5,5)\}$.

Hierarchical Clustering

→ Hierarchical clustering or hierarchical cluster analysis or HCA is a method of clustering which seeks to build a hierarchy of clusters in a given dataset.



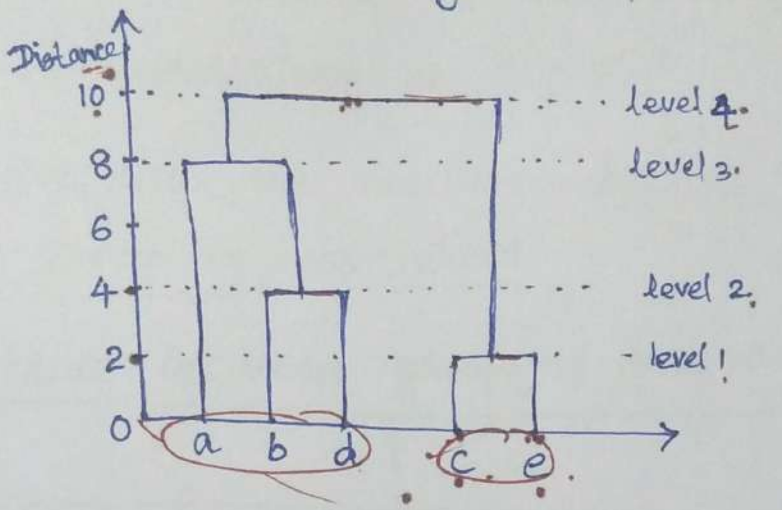
→ The clusters at each level of the hierarchy are created by merging clusters at the next lower level.

→ At the lowest level, each cluster contains a single observation and at the highest level, there is only one cluster containing all the data.

→ The decision regarding whether two clusters are to be merged or not is taken based on the measure dissimilarity between the clusters.

Dendrogram -

A dendrogram is a tree diagram used to illustrate the arrangement of the clusters produced by hierarchical clustering.



N

N-1

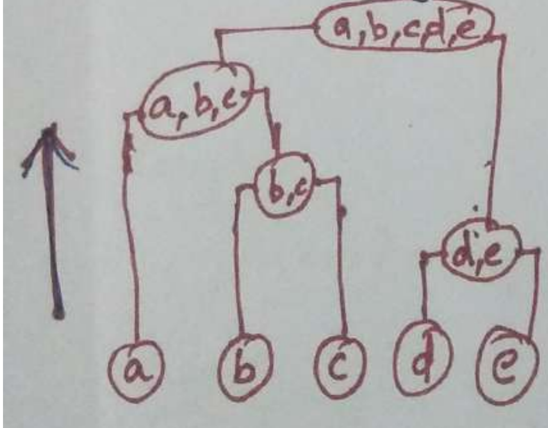
N=5

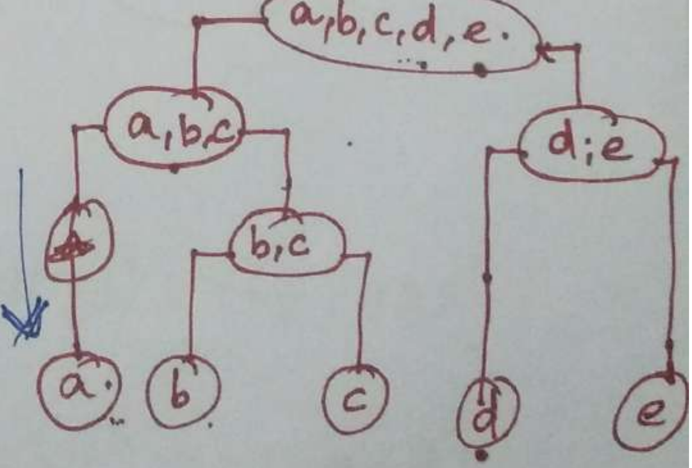
4

Methods of Hierarchical clustering

a, b, c, d, e

Agglomerative clustering	Divisive Analysis (DIANA)





* Bottom-up approach

* Top-down approach

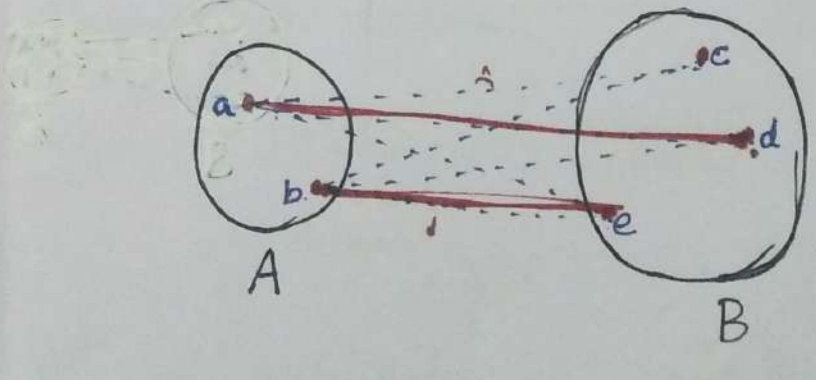
Measures of dissimilarity

In order to decide which clusters should be combined or where a cluster should be split, a measure of dissimilarity between sets of observations is required.

We use the measure - distance between the group of observations.

Distance between groups of data points (clusters)

- Complete linkage
- Single linkage
- Average linkage



Complete

$$d(A,B) = \max\{d(x;y) : x \in A, y \in B\}$$

Single

$$d(A,B) = \min\{d(x;y) : x \in A, y \in B\}$$

Average

$$d(A,B) = \frac{1}{|A||B|} \sum_{x \in A, y \in B} d(x,y)$$

Distance beteveen data points

Consider two data points $\bar{x} = (x_1, x_2, \dots x_n)$ and $\bar{y} = (y_1, y_2, \dots y_n)$.

Numeric data

Euclidean distance $= \sqrt{(x_1-y_1)^2 + \dots + (x_n-y_n)^2}$

Squared Euclidean distance $= (x_1-y_1)^2 + \dots + (x_n-y_n)^2$

Manhattan distance $= |x_1-y_1| + \dots + |x_n-y_n|$

Maximum distance $= \max\{|x_1-y_1|, |x_2-y_2|, \dots |x_n-y_n|\}$

Non-numeric data (text or word)

Levenshtein distance

kitten sitting.

distance $= 3$

- substitution of s for k
- substitition of i for e.
- insertion of g at the end.

Agglomerative Clustering

Algm & Example

Problem 1

Given the dataset {a, b, c, d, e} and the following distance matrix, construct a dendogram by complete-linkage hierarchical clustering using the agglomerative method.

	a	b	c	d	e
a	0	9	3	6	11
b	9	0	7	5	10
c	3	7	0	9	2
d	6	5	9	0	8
e	11	10	2	8	0

Ans) **Step 1:** Assigning each data item to its own cluster, so that we have (N=5) clusters, each containing just one item.

Data set = {a, b, c, d, e}

Initial cluster set C_1 : {a}, {b}, {c}, {d}, {e}

Table which give the distance between the various clusters in C_1:

	$\{a\}$	$\{b\}$	$\{c\}$	$\{d\}$	$\{e\}$
$\{a\}$	0	9	3	6	11
$\{b\}$	9	0	7	5	10
$\{c\}$	3	7	0	9	2
$\{d\}$	6	5	9	0	8
$\{e\}$	11	10	2	8	0

Step 2: Find the closeset pair of clusters and merge them into a single cluster so that now we have one less cluster.

Minimum distance is between $\{c\}$ and $\{e\}$

$$d(\{c\},\{e\}) = 2$$

New set of clusters $C_2 : \{a\}, \{b\}, \{d\}, \{c, e\}$

d(

Step 3: Compute the distance between the new cluster and each of the old clusters.

$$d(\{c,e\},\{a\}) = \max(d(c,a), d(e,a))$$
$$= \max(3, 11) = 11$$
$$d(\{c,e\},\{b\}) = \max(d(c,b), d(e,b))$$
$$= \max(7, 10) = 10$$
$$d(\{c,e\},\{d\}) = \max(d(c,d), d(e,d))$$
$$= \max(9, 8) = 9$$

Distance table of C_2

	{a}	{b}	{d}	{c,e}
{a}	0	9	6	11
{b}	9	0	5	10
{d}	6	5	0	9
{c,e}	11	10	9	0

Step 4: Repeat step 2 and 3 untill all items are clustered into a single cluster of size N.

Minimum distance is between {d} and {b}

$$d(\{b\},\{d\}) = 5$$

New set of clusters C_3: {a}, {b,d}, {c,e}

$$d(\{b,d\},\{a\}) = \max(d(b,a), d(d,a))$$
$$= \max(9, 6) = \underline{\underline{9}}$$

$$d(\{b,d\},\{c,e\}) = \max(d(b,c), d(b,e), d(d,c), d(d,e))$$
$$= \max\{7, 10, 9, 8)$$
$$= \underline{\underline{10}}$$

Distance table of C_3

	{a}	{b,d}	{c,e}
{a}	0	9	11
{b,d}	9	0	10
{c,e}	11	10	0

Minimum distance is b/w {a} and {b,d}

$$d(\{a\},\{b,d\}) = 9$$

New set of clusters C_4: {a,b,d}, {c,e}

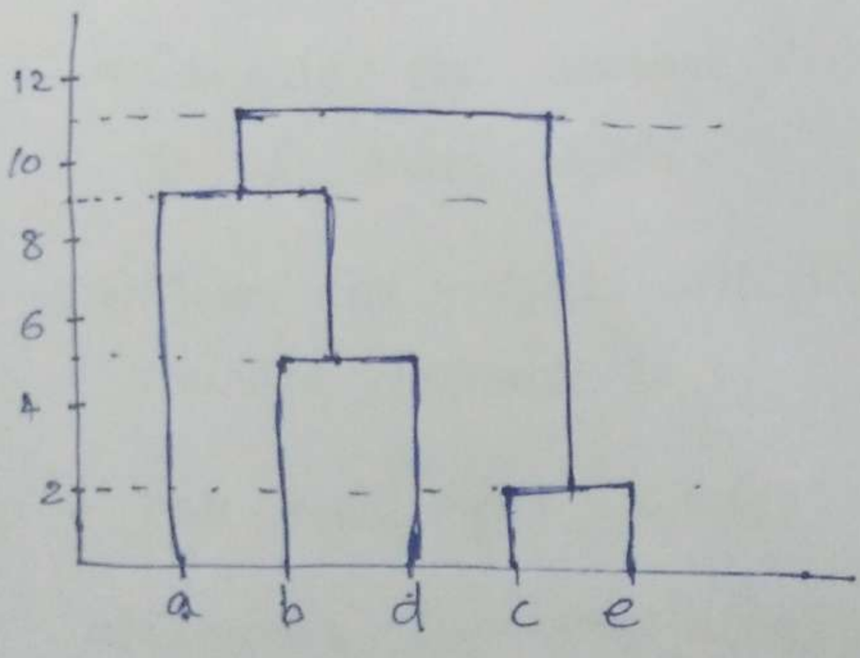
$$d(\{a,b,d\},\{c,e\}) = \max(d(a,c), d(a,e), d(b,c), d(b,e), d(d,c), d(d,e))$$

$$= \max(3, 11, 7, 10, 9, 8)$$

$$= \underline{\underline{11}}$$

New ~~clust~~ cluster C_5: $\{a, b, d, c, e\}$

Dendogram.



Divisive Analysis (DIANA)

Algorithm

Step 1: Initialy we have a single cluster C_l. Suppose the cluster C_l is going to be split into clusters C_i and C_j.

Step 2: Let $C_i = C_l$ and $C_j = \emptyset$.

Step 3: For each object $x \in C_i$:

a) Compute the average distance of x to all other objects.

b) Move the object with the maximum average distance to C_j.

Step 4: For each object $x \in C_i$:

a) Compute $D_x = \text{average}\{d(x,y) : y \in C_i\} - \text{average}\{d(x,y) : y \in C_j\}$

b) Find an object x in C_i for which D_x is largest. If $D_x > 0$ then move x to C_j.

Step 5: Repeat step 4 until all differences D_x are negative. Then C_l is split into C_i and C_j.

Step 6: Select the cluster with the larget diameter. (The diameter of a cluster is the largest dissimilarity between any two of its objects). Then divide this cluster, following steps 1-5.

Step 7: Repeat step 6 until all clusters contain only a single object.

Expectation-Maximisation Algorithm

The expectation-maximisation algorithm (EM algm) is a method to find MLE of the parameters of a statistical model in cases where the equations cannot be solved directly.

Gaussian mixture is a kind of statistical model which involves latent variables and hence cannot be solved directly using MLE method.

Outline of EM algorithm

Step 1: Initialise the parameters θ to be estimated.

Step 2: Expectation step (E-step) - Using the observed available data of the dataset, estimate (guess) the values of the missing data.

Step 3: Maximization step (M-step) - Complete data generated after the expectation step is used to update the parameters, θ, by maximizing likelihood function.

Step 4: Repeat step 2 and 3 until converge.

In machine learning, Clustering is an example for missing data problem. Here the missing data are the cluster labels.

Gaussian mixture models can be used to cluster unlabeled data points.

That is, not knowing what samples came from which class, our goal is to use Gaussian mixture models to assign the data points to the appropriate cluster.

Since Gaussian mixture model contains latent variables, we apply EM algorithm to solve the problem.

EM algorithm for Gaussian Mixture

Problem :-

Suppose we are given a set of N observations $\{x_1, x_2, \ldots, x_N\}$ of a numeric variable X.

Let X be a mix of k normal distributions and let the probability density function of X be

$$f(x) = \pi_1 f_1(x) + \pi_2 f_2(x) + \cdots\cdots + \pi_k f_k(x)$$

where

$$\pi_i \geq 0 \quad , \quad i = 1, 2, \ldots k$$

$$\pi_1 + \pi_2 + \cdots + \pi_k = 1$$

$$f_i(x) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} \quad , \text{for } i = 1, 2, \ldots, k$$

Estimate the parameters $\mu_i, \mu_2, \ldots, \mu_k$, $\sigma_1, \sigma_2, \ldots, \sigma_k$ and $\pi_1, \pi_2, \ldots, \pi_k$.

Log-likelihood function

Let θ denote the set of parameters μ_i, σ_i, π_i (for $i = 1, 2, \ldots, k$). The log-likelihood function for the above problem is given below:

$$L(\theta) = \log\left[f(x_1) + f(x_2) + \cdots + f(x_N)\right]$$

$$= \sum_{i=1}^{N} \log\left[f(x_i)\right]$$

$$= \sum_{i=1}^{N} \log\left[\pi_1 f_1(x_i) + \pi_2 f_2(x_i) + \cdots\cdots + \pi_k f_k(x_i)\right]$$

$$= \sum_{i=1}^{N} \log\left[\frac{\pi_1}{\sigma_1\sqrt{2\pi}} e^{-\frac{(x_i-\mu_1)^2}{2\sigma_1^2}} + \cdots + \frac{\pi_k}{\sigma_k\sqrt{2\pi}} e^{-\frac{(x_i-\mu_k)^2}{2\sigma_k^2}}\right]$$

Algorithm

Step 1: Initialize the means μ_i's, the variance σ_i^2's and the mixing coefficients π_i's.

Step 2: Calculate the following for

$n = 1, 2, \dots, N$ and $i = 1, 2, \dots, k$

$$\gamma_{in} = \frac{\pi_i f_i(x_n)}{\Sigma \pi_i f_i(x_n)}$$

$$N_i = \gamma_{i1} + \gamma_{i2} + \dots + \gamma_{iN}$$

Step 3: Recalculate the parameters using the following:

$$\mu_i = \frac{1}{N_i}\left(\gamma_{i1} x_1 + \dots + \gamma_{iN} x_N\right)$$

$$\sigma_i^2 = \frac{1}{N_i}\left[\gamma_{i1}(x_1 - \mu_i)^2 + \dots + \gamma_{iN}(x_N - \mu_i)^2\right]$$

$$\pi_i = \frac{N_i}{N}$$

Step 4: Evaluate the log-likelihood function and check for convergence of either the parameters or the log-likelihood function. If coverge then stop; Else goto step 2.