①

Data mining
clustering

★ it is mainly use for data visualization

★ clustering means groups, grouping similar properties, similar characteristics.
to find homogeneous group

example Suppose
if I have dataframe

1	12	20
2	15	23
3	16	43
4	17	60

Now I have to group the datapoints
So how we will do this by techique
called clustering

★ how to find a good cluster

① similar records should belong to the same cluster

② Dissimilar records should belong to different cluster.

* typesclustering algorithms

- (1) Hietachical clustering
- (2) K-means clustering
- (3) DBScan clustering

(1) Hierarchical clustering Algorithms

- (a) Agglomerative clustering (bottoms up approach)
- (b) Division clustering (Top to bottom aproach)

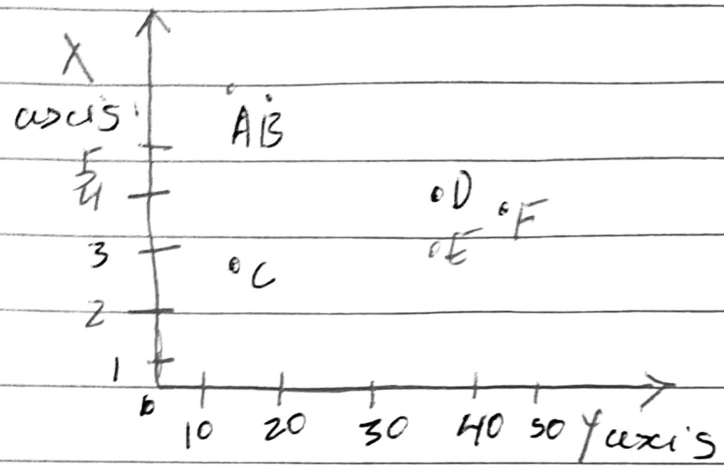
(a) Agglomeradive clustering

To explain Agglometatic clustering we will take an example.

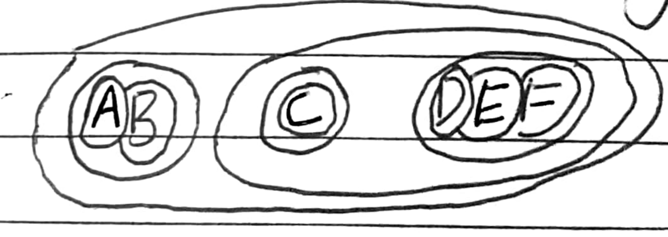
Let say we have datapoints called

A,B,C,D,E,F

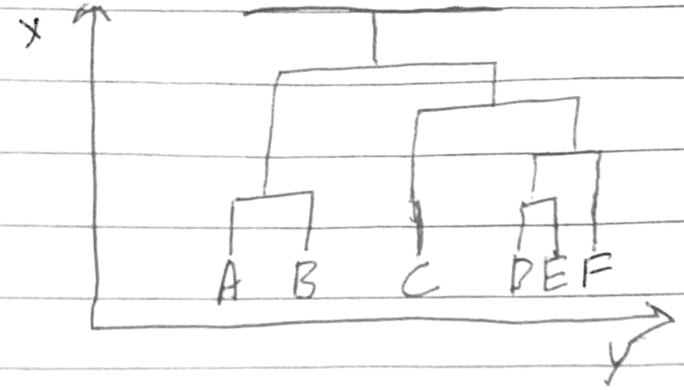
Now we will plot the graph.



Aglomerative is an iterative process in First iteration everydata point will be treated as signal cluster and then 2nd iteration whoever is near of datapoint will be-it will group them suppose in 1st iteration ABCDEF will be signal cluster then in 2nd iteration it found AB and DEF is close to each other so they will group them and lastly c is to the cluster of DEF so it make clucter with them and lastly the cluster of AB and DFFC will merge together will form one cluster to represent this cluster For better data visualisation we will use always dendogram.



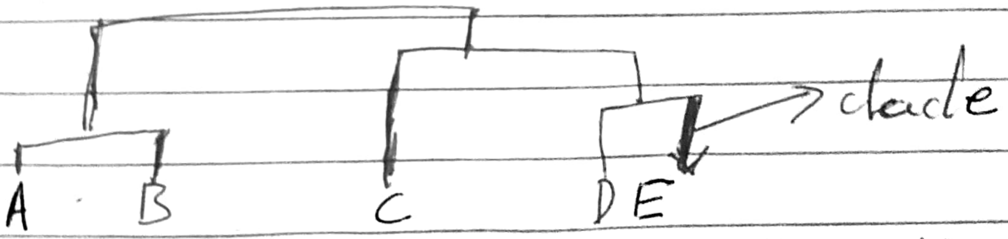
Dedogram visualization.



cutting the detogram

before going to this topic we will learn 2 New word called
① clade (intra)
② intracluster distance.

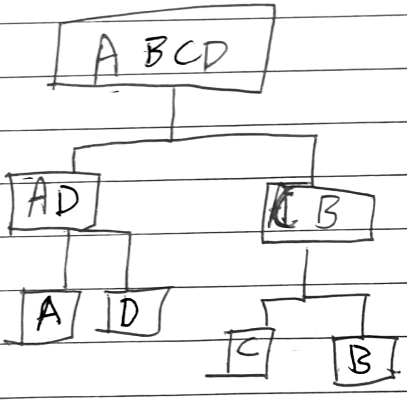
① clade - its branch in dentogram which represents the similarity between the datapoint it tells you the similarity



when the distance is more then there will be disimilarity between datapoint So which is the best cluster is AB because the clade measure is less.

(b) Division clustering

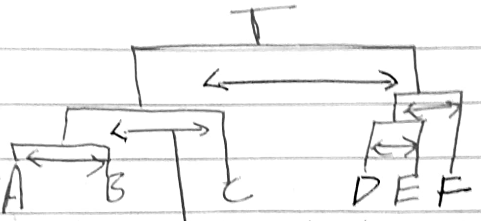
its same as Agglomerative clustering but instead of going bottoms up approach it use top down approach it means firstly it calculate the cluster as whole then it will divide the cluster which has longest distand and iterative process goes on till it reaches singal datapoint



(2) intraecluster distance

→ its distance between datapoint of cluster and different cluster

For example

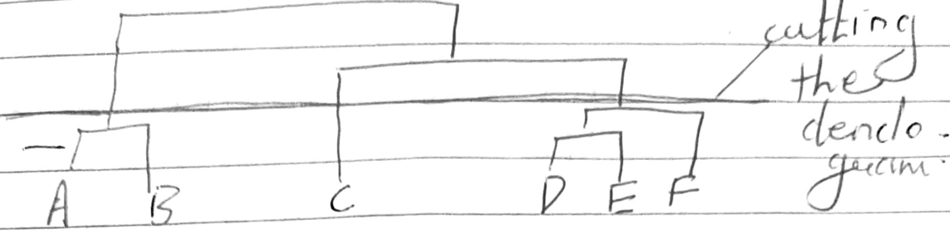


intraacluster distance

So intracluster distance should always grows simuntaneouly when its going bottoms up otherwise its really a bad dedogram.

Now comes the part of cutting the deodogram in order to get best cluster (Note:- we have to this visually)

Lets take example of same previous dataset.



* **Euclidean Distance** $(d_{ij} = \sqrt{(x_{ip} - X_{jp})^2})$

d_{ij} = distance between observaction of i and j

* it length of a line segment between two points.

So how will measure the length who how should be the method. inorder to do that we have.

* linkage method

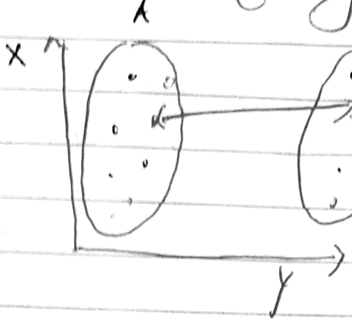
minimum (singal linkage)

maximum (complete linkage)

Average linkage.

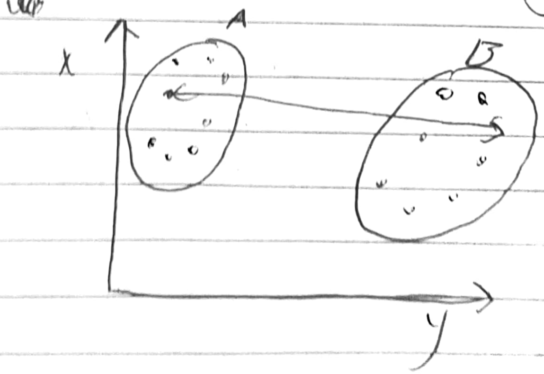
centroid linkage

① minimum (singal linkage)



in singal linkage Suppose there are two cluster called A and B and they have a datapoint which is nearest to the other cluster of datapoint so that segment we can use in Euclidean distance.

② maximum (complete linkage)

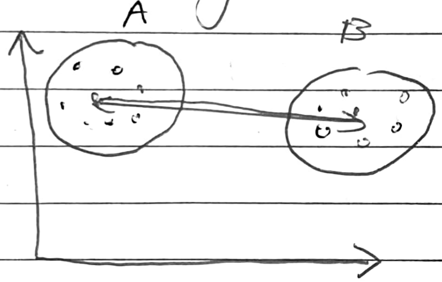


the cluster is same as the above example Just difference ~~unstean~~ unstead of minimum datapoints as segment we took maximum datapoint distance.

(3) Average linkage

we will take same graph example as precisous Now just the different is we will calaulate Average distance between them and then we will use in eculodian distance.

(4) centroid linkage



So every ~~e~~cluster has centre point we will take that point segment of each cluster and take it as a segment for two calculate eculodian distance

PAGE No.
DATE / /

Point to remember

☆ Standardization and Normalization in terms of clustering while calculating Euclidean distance.

So whenever we have dataset so the unit of each coloum will be different So this will effect the measurement of Euclidean distance.

So solution is

1. when dataset is Normally distributed we will use standardization (it ca be positive or Negative values).
2. when dataset is not Normally distributed So we will use Normalization in our dataset (it will always be 0's and 1s values).

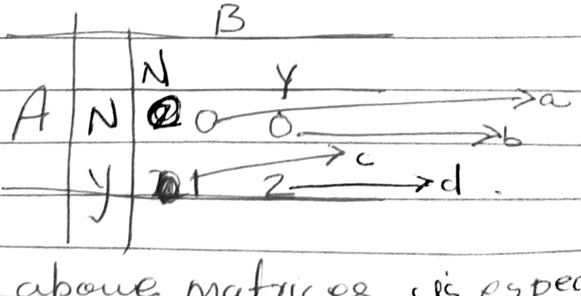
Distance for Binary data

Suppose what if ~~we~~ we have categori dataset columns and we have to calcul ate eculodian distance

lets take an example.

	Married	Smoker	manager
S	Yes	yes	yes
A	No	yes	No
B	yes	No	yes

so we have x variable called S,A,B let assume they are person. and these column of married, smoker, manager they ~~they~~ have got yes or No. Now we will put this above dataset in 2X2 Matrix columns.



so above matrices is especial for B person so he have got 1 Nos and 2 yes

So the measure will be

① Binary Euclidean Distance

$(b+c)/(a+b+c+d)$

② Simple matching coefficient

$(a+d)/(a+b+c+d)$

⑥ Jaquard coefficient:

$d/(b+c+d)$.

✯ for (Numerical + categorical) data

step ① Normalize the data to [0,1]

step ② use Grower's General Dissimilarity coefficient.