Date 25/03/2025

Classification

- → DT
- → BC

- → BP
- → SVM
- → LR
- → KNN

} — Test (next week)

Clustering

Boosting algorithms

Feature selection

Data clustering : UnSupervised learning.

↳ Class label not known

Partition based clustering

- → K-mean
- → PAM
- → ~~KLARA~~ Clara
- → Clarans

} ~~Variation~~ in K-mean as k-mean is not good for large set of data.

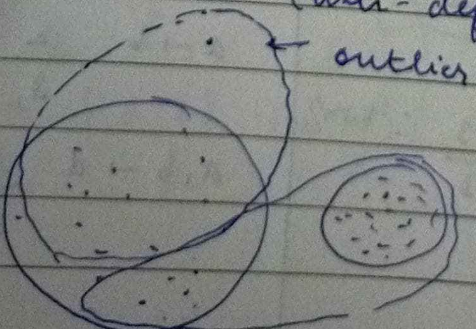
↓

- → might not give proper clusters
- → lacks accuracy & efficiency
- → Sparse matrix

Problem: Outliers are also part of a cluster

k: number of clusters

(user-defined parameter)



Due to presence of outlier means converge to a value impacted largely by the outlie

Page No.

Binary variables:

a	11	Both objects for a particular instance have 1.
b	10	
c	01	
d	00	

Symmetric:

$$d(i,j) = \frac{b+c}{a+b+c+d} = \frac{\overbrace{10}^{\text{1 for } X_1, \text{ 0 for } X_2} + 01}{11+01+10+00}$$

Asymmetric : $d(i,j) = \dfrac{b+c}{a+b+c}$

How to know if variable is sym / asym.

If $\to 0 \to 1$
& $\to 1 \to 0$] → symmetric binary variable.

If $\to 0 \nrightarrow 1$ then asymmetric binary variable.

Eg: No : N → 0
Yes : Y | Presence : P → 1

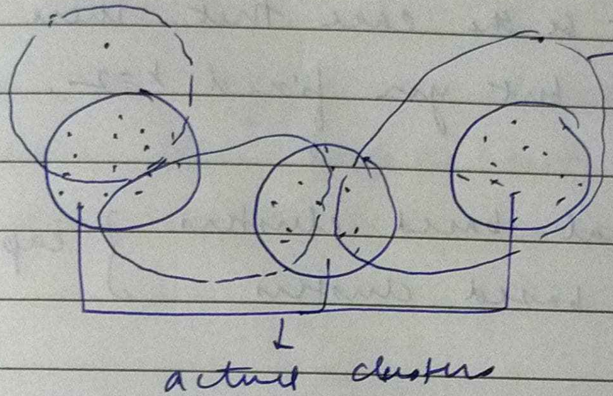
$$d(\text{Jack}, \text{Mary}) = \frac{10+01}{11+01+10 \; \cancel{00}} = \frac{\overset{\text{no instance for 10.}}{0}+1}{2+1 \; \cancel{+}} = \frac{1}{3} = 0.33$$

The points which were initially part of cluster 1 can now be part of cluster 2.

Large data ⇒ Large noise

⇒ Mean affected largely

∴ The actual clusters may not be identified which impacts accuracy.



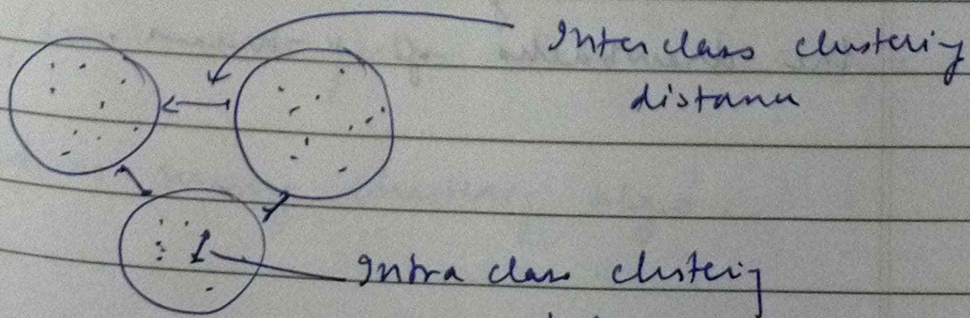
k-means + k-means ++ versions also available.

To apply k-means on a large data, remove noise from the data.

- can be used in Text mining, etc.

Data clustering: grouped based on similarity measure (could be distance). Can be on multi-dimensional data.

→ OPTIMISATION PROBLEM


Interclass clustering distance
Intraclass clustering distance.

Data types

→ Interval based

→ Ratio based

→ Binary

→ Ordinal

→ Nominal

} → Real data has to be transformed into a data of one of these types before applying clustering

Data from kaggle is already pre-processed & hence this step is not needed. But raw data must be pre-processed first.

Data matrix : attributes in columns
entities in rows.

$$\begin{array}{cc} & \begin{array}{ccccc} A_1 & A_2 \cdot & & & A_n \end{array} \\ \begin{array}{c} E_1 \\ \vdots \\ \\ \\ E_n \end{array} & \begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix} \end{array}$$

Object by variable.

Dissimilarity matrix:

Object by object : Distance b/w all objects & all other objects.

→ upper/lower triangular matrix is enough.

→ Used in many clustering algo.

Clustering requires 2 optimisations at the same time:

(1) Inter class distance of clusters should be maximised

(2) Minimise intra class data distance ~~of clusters~~.

Problem with k-means :- Fixing k

It might be the case that there are actually 4 clusters but you fixed k = 2.

→ Hierarchical based clusters } capable of doing

→ Density based clusters }

Within hierarchical based clustering we have :

Single link

Heavy link

Complete link

Birch

Density based : → CURE

→ DBSCAN

→

3 researchers ↳ BFR: Primarily uses k-means & overcomes the drawbacks of k-means.

Date ___ / ___ / _____

Smaller unit → larger variable range

To select basketball players based on height, we want clear distances in the difference due to closeness in height.

∴ Lower down the unit.

Give more weightage to height.

∴ Convert Euclidean into weighted distance.

$$w\,(h_1 - h_2)$$

weight assigned to assign higher weight to height.

Suppose 3 attributes: A_i W_i B_i

assign higher weight to the one with highest importance

$$w_1 > w_2 > w_3$$

→ Mean absolute deviation better than standard deviation due to absolute values.

Assigning weights to Minkowski distance would give weighted distance:

$$\sqrt[1/m]{w_1 \, |x_1 - \bar{x}|^m + \dots}$$

Not necessary to assign weight to all attributes