DATE
PAGE

Unit - 5. Unsupervised learning.

Clustering - grouping of data points having similarity
→ No. training

Soft clustering -> a data point can be part of 2 clusters or more than 2

hard clustering → only one cluster at a time.

K mean - hard clustering Algorithm

Steps ① Decide K = no. of clusters
② Select K random data points as centroids
③ find distance d b/w all and each points w.r.t each centroids.
$$d = \sqrt{(x_1-x_2)^2 + (y_1-y_2)^2}$$
④ Recompute centroid.
$$\left(x_c\ y_c = \frac{\sum x_i}{m}, \frac{\sum y_i}{m}\right)$$
⑤ Repeat 3 & 4 until
① centroid do not change
② point remain in same cluster
③ Reached max. iteration.

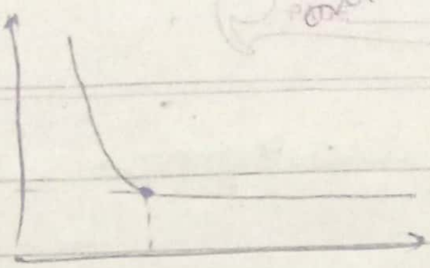
Elbow Method

graph b/w WCSS and no. of cluster K.

WCSS => within cluster sum of square or sum of the sq. distance b/w points in a cluster &

Scanned with CamScanner

compile Run
overload overriding



cluster centroid.

K medoids → partioning Algorithm.
the center of the subset is a member of the subset called a medoid.

medoid are robust to outliers.

PAM - partioning around medoids.

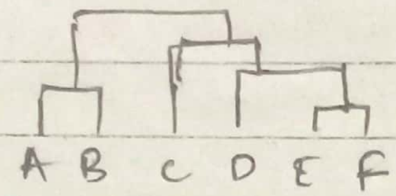
(1). select K,
(2) Select random k points as medoids,
(3) calculate cost $|x_1 - x_2| + |y_1 - y_2|$ for remaining points for all K points.
(4) then assign that point to minimum cost.
(5) Then we calculate total cost
(6) Swap → select one of non medoids O
(7) Find the same thing cost and assign points to cluster.
(8) find the the total cost. if total cost is max then it is bad idea to swap, if found min then it is good idea.

DATE
PAGE

Hierarchical clustering (HC or HCA)

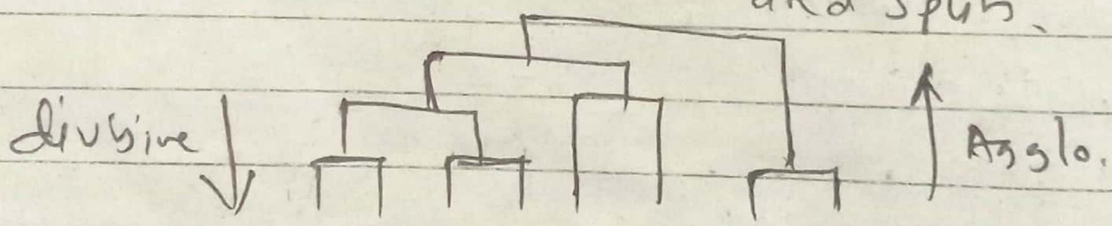
→ data points are arranged in a hierarchy of clusters

<u>dendogram</u> → a diagram representing hierarchy



height indicate the order

HCA Algorithm
(1) Agglomerative - bottom up → merging
(2) divisive - top-down - Start with one cluster and split.



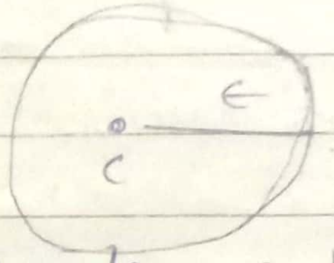
<u>DBSCAN</u>

Density Based Spatial clustering of Application with noise.

→ <u>unsupervised</u> - clustering.
→ based on density.

Scanned with CamScanner

DBSCAN : (1) Epsilon (eps) ϵ, measure of neighbourhood, or radius of circle



(2) min_sample → How many Minimum data point required to consider an area dense

(3) Density → no. of data point.

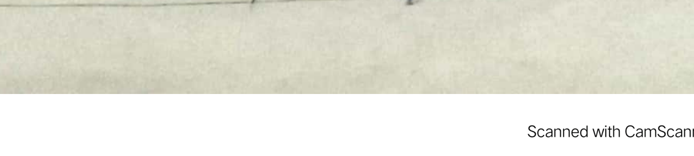
(4) core point → if atleast a specified no. of neighbouring points falls within specified radius ϵ.

(5) Border point - which has atleast a core point

(6) Noise point → not core, nor Border.

(7) Density edge - two points p & q and distance b/w 2 points is less than ϵ then is called density edge

(8) Density Connected points



Algorithm ① distance is Calculated.
② neighbourhood consider ϵ.
③ min_sample wit make cluster
④ repeat until categorized

Advantages ① does not assume that the cluster have Spherical Shape
② No need of finding K
③ high efficient
④ remove noise

Disadvantage.

→ It fail when have multiple density
→ Not work well for high dimension.

Spectral clustering (graph cuts)
Steps ① pre-processing
→ Construct a Matrix representation of the graph
② Decomposition
→ Compute eigen value & eigen vectors
→ Map each point to a lower-dimensional representation based on one or more eigenvectors
③ Grouping
→ Assign points to two or more clusters, based on new representation.

DATE
PAGE

~~Outlier Analysis~~

* Spectral clustering is unsupervised,
* we find clusters in graph
* we find subgraph and that will cluster

Scanned with CamScanner

Unit - 5 (Part 2)

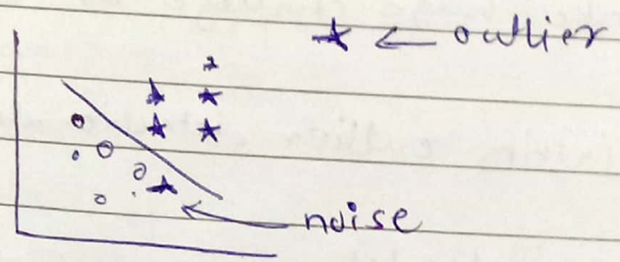
Outlier Analysis

outlier. Outlier are those datapoint that are significantly different from the rest of the dataset.

→ why to detect outlier

They are abnormal, and their presence can often skeaw the result of Statistical analyses on the dataset

- This could lead the less effective and less useful models.



noise is random error.

Types of outlier

① Global outlier
- If a data point is far away from rest of the data set.

② Contextual (conditional) outlier
- If a data point is deviates with respect to a specific context.

eg. 28°C is outlier in Shimla in winter but not in summer.

Page No.
Date

(3) Collective.
A Subset of data point are deviates significantly from the entire data set.

Challenge in outlier detection.

* The boundary b/w normality & abnormality is no clear.
* differect application may have very diff. requirement, Application Specific detection
* Noise can make huge challege for outlier detection.
* we have to justify outlier detection also.

Outlier Detection Model.
(1) Supervised. :- training + testing,
- classifier
- imbalanced handling.

(2) Semisupervised :- Having ondy some labelled data. then we apply both supervised & semi supervised.

(3) Unsupervised, - clustering

(4) Statistical Method for Outlier Detection.
- Statistical model make assumption of data normality (Z-Score or Standard deviation, IQR (percentile

(5) proximity based Method for outlier Detection.
+ object is an outlier if the nearest of the object are far away in feature Space.

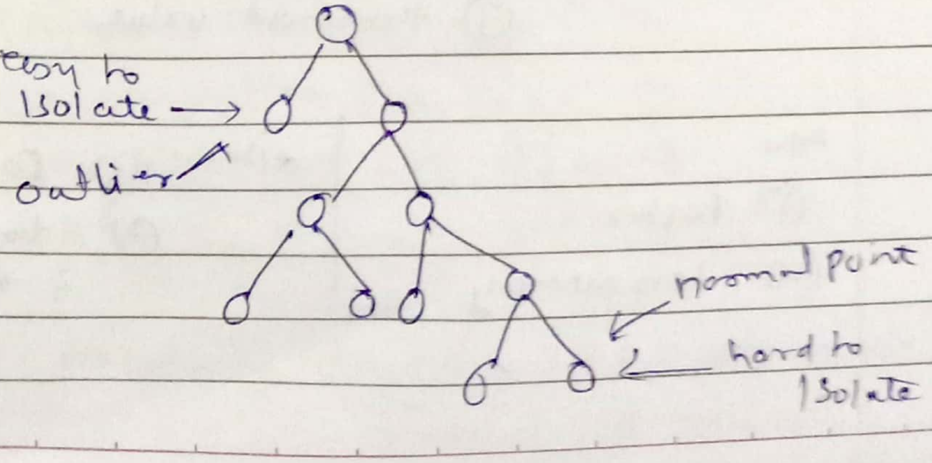
+ neighbour is far away.

(6) Clustering based Method for outlier detection.
- normal data belong to large & dense cluster
- Outlier belong to Small or Sparse cluster.

Introduction to Isolation forest (iForest)

→ variation of Random forest
→ Unsupervised learning
→ outlier are "few and different"
- To make position such that each data point is isolated

→ Isolation Tree is used (Binary tree)
(a) a regular / normal point is much harder than outlier



Page No.
Date

Algorithm.

1) Trainning: Building a forest of isolation trees
→ Randomly select feature
→ Randomly partition

2) Prediction Compute outlier score for newpoint.
let m: Sample size

$$S(x, m) = 2^{-\frac{E(h(x))}{c(m)}}$$

$E(h(x))$: Average Search height for x from the itree

$c(m)$ = Average value of $h(x)$.

$E(h(x)) << c(m) \Rightarrow S(x, m) \cong 1$ (outlier)

$E(h(x)) \cong c(m) \Rightarrow S(x, m) \cong 0.5$ (Normal)

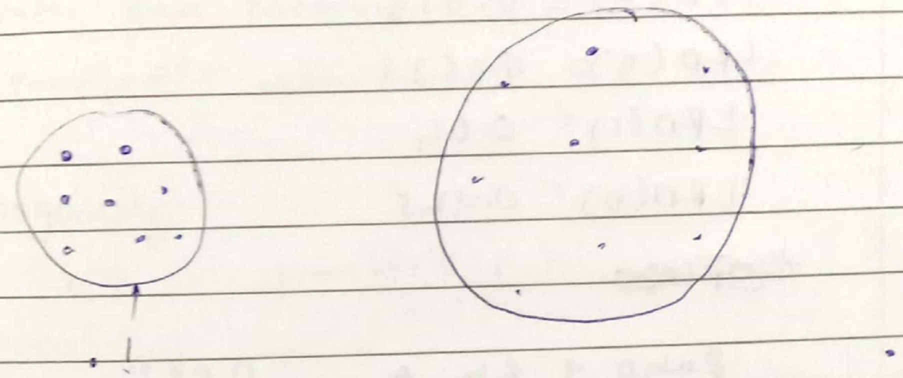
hyper parameter ① No. of itree
② Sampling size
③ threshold value

Adv.
(1) faster
(2) less memory.

disadv (1) Can suffer bias
(2) threshold value is not clear.

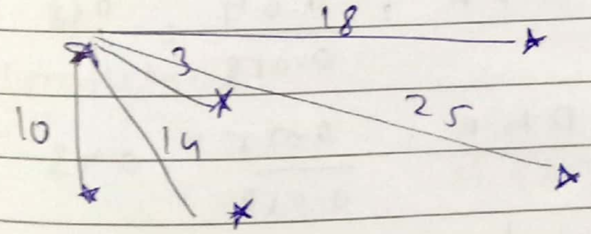
LOF - Local outlier factor.

- Based on local density
- ups: unsupervised outlier detection.
- It compute local density deviation of a given data point wrt to its neighbour.



k-distance = Maximum distance b/w K neighbour.

k=3



k-distance = 14

reaciability distance = max (k distant, achal distance)

RD (A, B) = max (14, 10) = 14

RD (A, B) = Max (14, 3) = 14

RD (A, D) = Max (14, 14) = 14

RD (A, E) = Max (14, 18) = 18

RD (A, F) = Max (14, 25) = 25

Local reacubility density LRD = inverse of Mean of its reacuability between towards other in dataset

Page No.
Date

$$LRD = \frac{1}{\frac{14+14+14+18+25}{5}} = 0.058$$

LOF = Mean of the ratio of each k-nearest to point of interest.

LRD (A) = 0.058
LRD (B) = 0.062
LRD (C) = 0.04
LRD (D) = 0.025

~~LOF(A)~~

$$\text{Ratio of B to A} = \frac{0.062}{0.058} = 1.06$$

$$\text{Ratio of C to A} = \frac{0.04}{0.058} = 0.68$$

$$\text{Ratio of D to A} = \frac{0.025}{0.058} = 0.43$$

$$LOF(A) = \frac{1.06 + 0.68 + 0.43}{3} = 0.72$$

~~If LOF~~

LOF ≤ 1 not outlier

LOF > 1 outlier

Here A is not outlier

Page No.
Date

Evaluation Metrics and Score.

2 method

(1) Extrinsic Method.

- ground truth is available, you can compare ~~it with~~ ~~the~~ clustering against the group truth and Measure

- Supervised method

(a) Homogeneity

$$h = 1 - \frac{H(C|K)}{H(C)}$$

It measure how the sample in a cluster are similar.

(b) Completness.

$$c = 1 - \frac{H(K|C)}{H(K)}$$

It measure how much similar sample are put together by clustering algorithm

(2) Intrinsic Method

- when ground truth ~~is~~ of a data is not available,
- It evaluate the clustering by examining how well cluster are seperated and how compact the cluster are.

- unsupervised

+ Sihoutte coefficient

~~2 types of~~ Sihoutte function is used.

Scanned with CamScanner

$$\left\{ s(o) = \frac{b(o) - a(o)}{\text{Max}(a(o),\ b(o))} \right\}$$

two type of data required

1. Collection of all distance b/w object.
2. partition obtained by Clustering

Silhouette value is calculated

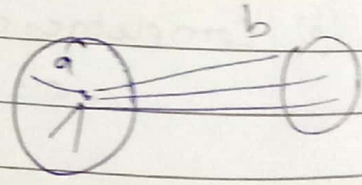
if 1 = well clustered

-1 = poorly clustered

a = Average distance of i to the points in the same cluster

b = min (average distance of i to points in another cluster)

$s = 1 - a/b$ if $a < b$

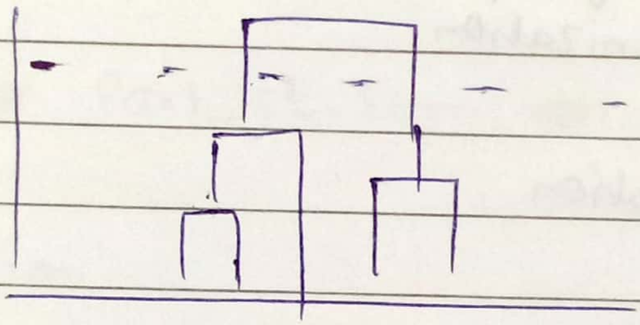


How should we choose the Number of clusters in Hierachical clustering

→ using dendogram.

→ whenever two cluster are merged, we will join them in this dendogram and height of the join will be distance b/w these points.

Now we will set threshold value and draw horizontal line (It should cut tallest line)



No. of cluster will be no. of intersection.