MoD 5 - Unsupervised learning

unsupervised learning.

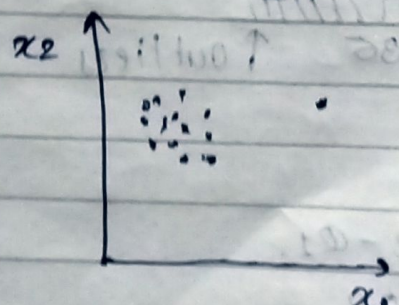
Anomaly detection — clustering — Association Rule mining.

- univariate
- multi-variate.

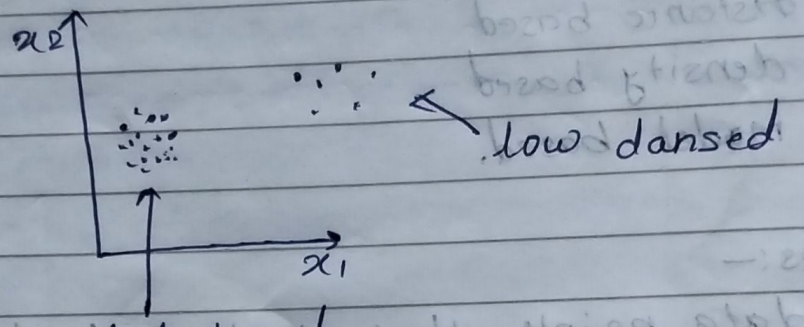
1) Anomaly detection:-
   1) distance based
   2) density based
   3) model based.

- Outliers:-
  - data points that significantly differs from data points in dataset.
  - These data points can distort statistical analyses and ml model, leading to inaccurate predictions or conclusions.

1) Distance based:-
   - this method identifies outliers based on the distance of data points from others in dataset.
   - These methods assume that outliers are located far away from the majority of data points.



2) Density based :-
- identifies outliers by assessing the density of data points within the dataset.
- These methods are based on the assumption that outliers are data points that lie in low-density regions compared to majority of data.

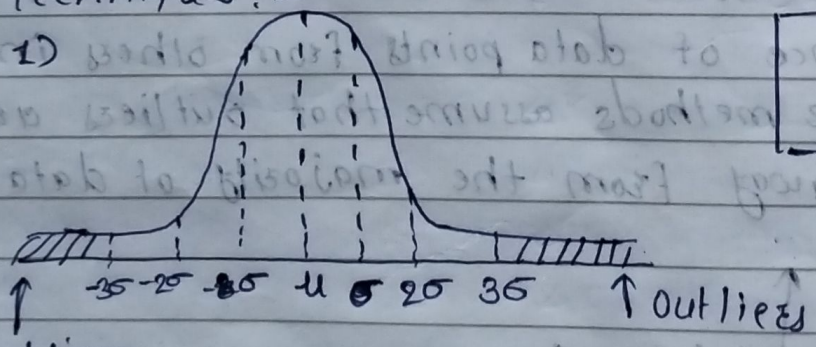


3) Model based :-
- higher ml models used to determine outliers.

ex). Decision tree (iForest), Regression.

* Techniques :-

1)



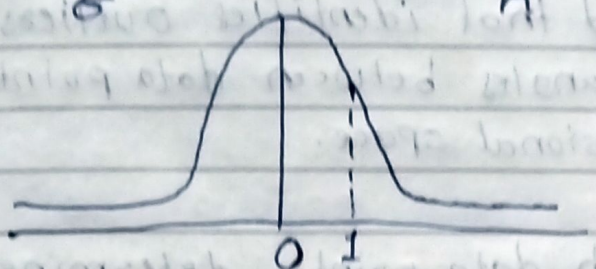
> μ+3σ
< μ-3σ
Outliers.

2) Q1, Q2, IQR = Q3 - Q1.

< Q1 - 1.5 * IQR
> Q3 + 1.5 * IQR

outliers

3) $z_i = \frac{x_i - \mu}{\sigma}$ $\mu = \frac{\sum x_i}{n}$



$z_i > 3$
$z_i < -3$
outliers.

4) Median absolute deviation (MAD) :-

$MAD = \frac{\sum_{i=1}^{n} (x_i - M)}{n}$ M = median

Given, data :-

f : 5, 10, 2, 3, 25, 45, 100

sort. 2, 3, 5, 10, 25, 45, 100
M

$\therefore MAD = \frac{(5-10) + (10-10) + (2-10) + (3-10) + (25-10) + (45-10) + (100-20)}{7}$

$= 17.14.$

$< M - C.MAD$
$> M + C.MAD$
outliers.

C = 2.5 / 3.0

5) K-NN : (kth nearesth neighbour) usually k = 3

* Techniques to find NN.

1) L1 / ~~Mahatam~~ Manhatan / city block.

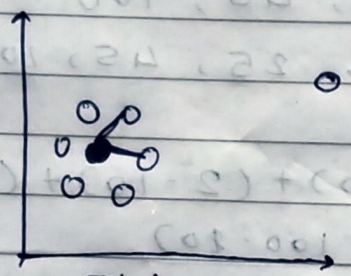
→ $L1 = |x_1 - x_2| + |y_1 - y_2| \ldots$

2) L2 / distance / Eucledian :-

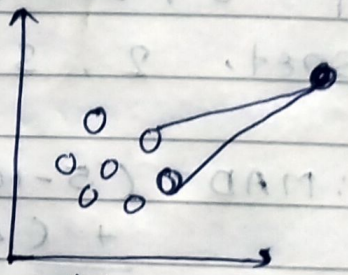
→ $L2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

* Angle based anomaly detection:-
  - method that identifies outliers by measuring the angles between data points in high-dimensional space.

steps:-
1) For each data point, determine the angle it makes with all pairs of other data points
2) Calculate the variance of this angle.
3) Points for which the variance is below predetermined threshold are anomalies.



This points represent normal point



This point represent anomaly.

* Example on k-NN :-
Determine the outlier for threshold for distance kth neighbour as 22. Use L1 measure. k=3.

→

ID	F1	F2	F3
1	2	22	48
2	11	23	48
3	3	21	47
4	4	22	46
5	3	24	49
6	12	19	51
7	11	38	55
8	10	19	56
9	13	20	49
10	14	20	49

(2, 3, 4, 5 Marks (1))

$L_1 = |x_2 - x_1| + |y_2 - y_1| + |z_2 - z_1|$

	1	2	3	4	5	6	7	8	9	10
1	0	10	3	4	4	16	32	19	14	15
2	10	0	11	10	10	8	22	13	6	7
3	3	11	0	3	5	15	38	18	13	14
4	4	10	3	0	6	16	32	19	14	15
5	4	10	5	6	0	16	28	19	14	15
6	16	8	15	16	16	0	24	7	4	5
7	32	22	33	32	28	24	0	21	26	27
8	19	13	18	19	19	7	21	0	11	12
9	14	6	13	14	14	4	26	11	0	1
10	15	7	14	15	15	5	27	12	1	0

dist[1][2] = |2 - 11| + |22 - 23| + |48 - 48|

= 9 + 1 + 0 = 10.

Given k = 8, Now kth neighbours of each point

1 → 10 ∵ (3, 4, 4, 10, 16, 15, 16, 19, 32)

↑ kth

2 → 8

3 → 11

4 → 6

5 → 6

6 → 7

7 → 24 → outlier because 24 > threshold (22)

8 → 12

9 → 6

10 → 7

Note:—

Another way of detecting outliers:—

- consider top 'k' highest distances from kth neighbour distances.

∴ For above examples:—

24, 12, 11 are outliers.

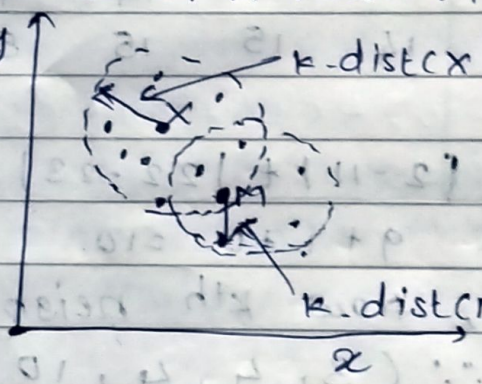
* Local outlier Factor (LOF) :- (unsupervised)
- density based technique of outliers in ml.
- LOF compares the local density of point with the densities of its neighbours.
- points that have a substatially lower density than their neighbors are considered to be outliers.

1) Reacability_distance :-

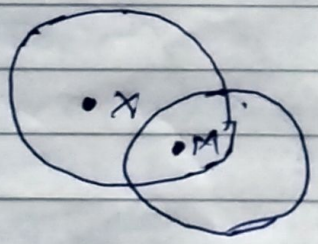
$reach\_dist_k(M,X) = max(\{k\_dist(X), dist(M,X)\})$

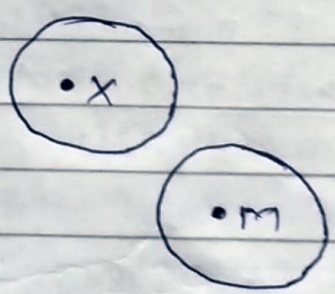
Here,

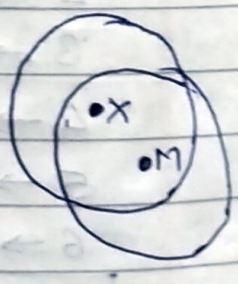
(M,X) = distance of X from M.



* cases:-

1) 

2) 

3) 

2) Local Rechability density (LRD) :-

$$LRD_k(M) = \frac{1}{\left[\frac{\sum_{X \in N_k(M)} reach\_dist(X)}{|N_k(M)|}\right]}$$

3) $LOF_k(M) = \frac{1}{|N_k(M)|} \cdot \sum_{X \in N_k(M)} \frac{LRD_k(X)}{LRD_k(M)}$

* Example :-

index	x	y
0	2	3
1	4	5
2	1	2
3	2	4
4	3	4
5	7	5
6	3	5
7	8	9
8	2	1
9	3	6

use distance matrix given in ex. pdf. $k = 2$

Finding k-dist & N_k.

index	k_dist	$N_k(M)$	$\|N_k(M)\|$
0	1.41	(2, 3, 4)	3
1	1.41	(4, 6, 9)	3
2	2.24	(0, 3, 8)	3
3	1.41	(0, 4, 6)	3
4	1.41	(0, 1, 3, 6)	4
5	4	(1, 6)	2
6	1.41	(1, 3, 4, 9)	4
7	5.66	(1, 5)	2
8	2	(0, 2)	2
9	1.41	(1, 6)	2

* Reachability distance:-

reach-dist$_k$(0,2) = max(k-dist(2), dist(0,2))

= max(2.24, 1.41)

= 2.24

∴ LRD$_k$(0) :-

X	k-dist(x)	d(x,m)	reach-dist(m,x)
2	2.24	1.41	2.24
3	1.41	1	1.41
4	1.41	1.41	1.41

Now, LRD$_k$(0) = $\frac{1}{\frac{\text{reach-dist}_k(0,2) + \text{reach-dist}(0,3) + \text{reach-dist}(0,4)}{3}}$

= $\frac{1}{\frac{2.24 + 1.41 + 1.41}{3}}$

= 0.59

LRD$_k$(2) =

X	k-dist(x)	d(x,m)	reach-dist(m,x)
0	1.41	1.41	1.41
3	1.41	2.24	~~1.41~~ 2.24
8	2	1.41	2

LRD$_4$(2) = $\frac{1}{\frac{1.41 + \text{~~1.41~~} 2.24 + 2}{3}}$

= ~~0.62~~ 0.53

LRD(2)

Rk(2,x)

LRDk(3):-

x	k-dist(x)	d(x,m)	reach-dist(m,x)
0	1.41	1.41	1.41
4	1.41	1.41	1.41
6	1.41	1.41	1.41

$$LRDk(3) = \frac{1}{\frac{1.41+1.41+1.41}{3}}$$

$$= 0.70$$

LRDk(4):-

x	k-dist(x)	d(x,m)	reach-dist(m,x)
0	1.41	1.41	1.41
1	1.41	1.41	1.41
3	1.41	1	1.41
6	1.41	1	1.41

$$LRDk(4) = \frac{1}{\frac{1.41+1.41+1.41+1.41}{4}}$$

$$= 0.70$$

$$LOFk(0) = \frac{1}{|Nk(m)|} \cdot \sum_{x \in Nk(m)} \frac{LRDk(x)}{LRDk(m)}$$

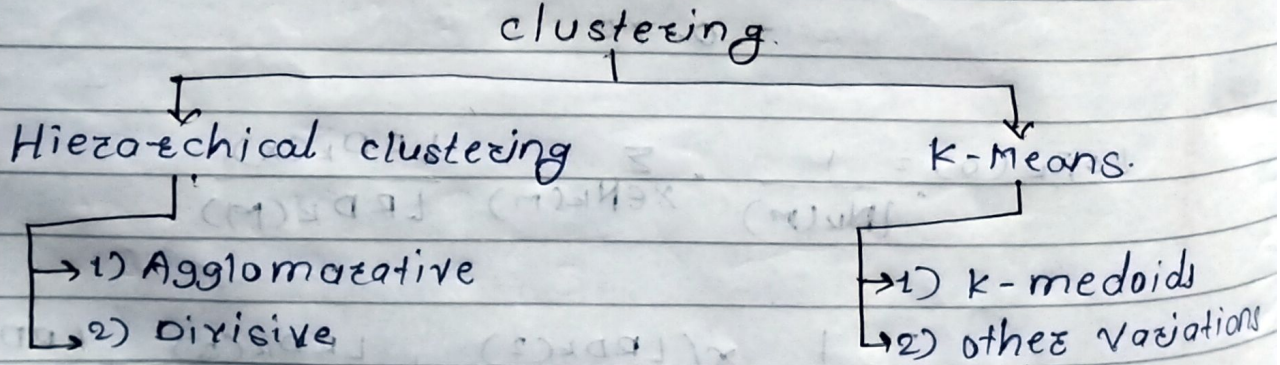
$$= \frac{1}{3} \times \left(\frac{LRDk(2)}{LRDk(0)} + \frac{LRDk(3)}{LRDk(0)} + \frac{LRDk(4)}{LRDk(0)} \right)$$

$$= \frac{1}{3} \times \left(\frac{0.53}{0.59} + \frac{0.70}{0.59} + \frac{0.70}{0.59} \right)$$

$$= 1.09$$

* Isolation Forest (iForest):— unsupervised.
  - main idea is to isolate anomalies by creating binary trees & using them to partition the dataset into smaller subsets.
  - *steps:-
    - Branching of tree starts by selecting a random feature first. And then branching is done on random threshold.
    - If value is less than threshold, it goes to the left branch else to right branch.

* clustering:— unsupervised.
  - task of grouping data points based on their similarity with each other.
  - aims at forming group of homogenous data points from a heterogenous dataset.
  - It evalutes the similarity based on a metric like Euclidean distance, manhattan distance etc. & then group the points with highest similarity score together.



Example :— K=2, cluster them with centroids (2,1) & (4,5). Determine new centroids.
Use eucledian distance.

eucledian distance = $\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$

c1 = cluster1 (2,1) c2 = cluster2 (4,5) (min(c1,c2) cluster.

x	y	C1(2,1)	C2(4,5)	cluster.
2	3	2.	2.82.	c1.
4	5	4.47	0	c2
1	2	1.41	4.24	c1
2	4	3	2.23	c2.
8	4	3.16	1.41	c2
7	5	6.40	3	c2
3	5	4.12	1	c2.
8	9	10	5.65	c2
2	1	0	4.47	c1
3	6	5.09	1.41	c2.

(2,1)

1) $\sqrt{(2-2)^2+(3-1)^2} = \sqrt{(2)^2} = 2$.

(4,5)

$\sqrt{(2-4)^2+(3-5)^2} = \sqrt{4+4} = \sqrt{8} = 2.82$.

∴ c1 = (2,3), (1,2), (2,1)

c2 = (4,5), (2,4), (3,4), (7,5), (3,5), (8,9)
(3,6)

Now, new centroids:-

For c1:

$x = \frac{\Sigma x_i}{|N|} = \frac{2+1+2}{3} = 1.67$.

$y = \frac{\Sigma y_i}{|N|} = \frac{3+2+1}{3} = 2$

∴ **centroid = (1.67, 2)**

For c2:

$$x = \frac{\Sigma x_i}{|N|} = \frac{2+3+7+3+8+3+4}{7} = 4.29.$$

$$y = \frac{\Sigma y_i}{|N|} = \frac{4+4+5+5+9+6+5}{7} = 5.43.$$

$\therefore$ centroid = (4.29, 5.43)

* Agglomerative algorithm:-

Given distance matrix:-

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	[2]	8	0

smallest-distance high-similarities. (2) combine (3,5).

New matrix,

	(3,5)	1	2	4
(3,5)	0			
1	11	0		
2	10	9	0	
4	9	6	[5]	0.

dist(1, (3,5)) :-

dist(1, 3) = 3 dist(1,5) = 11

(3,11), we have 3 strategies to choose.

1) single linkage = min(3,11) = 3

2) complete linkage = max(3,11) = 11

4) Average linkage = Avg(3,11) = 7

Throughout the problem we will use complete linkage.

dist (2, (3,5)) :-

dist (2,3) = 7 max (7, 10) = 10.

dist (2,5) = 10

dist (4, (3,5)) :-

dist (4,3) = 9 max (9,8) = 9.

dist (4,5) = 8

Next smallest distance = 5.

combine (2,4)

	(3,5)	(2,4)	1
(3,5)	0		
(2,4)	10	0	
1	9	9	0

dist ((2,4), (3,5))

dist (2,3) = 7 dist (4,3) = 9

dist (2,5) = 10 dist (4,5) = 8

max = 10.

dist (1, (2,4)):

dist (1,2) = 9 dist (1,4) = 6

max = 9.

New smallest dist = 9.

combine (1, (2,4))

	(3,5)	(1,2,4)
(3,5)	0	
(1,2,4)	11	0

dist ((3,5), (1,2,4)) :-

dist (3,1) = 3 dist (5,1) = 11

dist (3,2) = 7 dist (5,2) = 10

dist (3,4) = 9 dist (5,4) = 8

max = 11

Drawing Dendrogram :-