1.
(a)

(i) Core Object

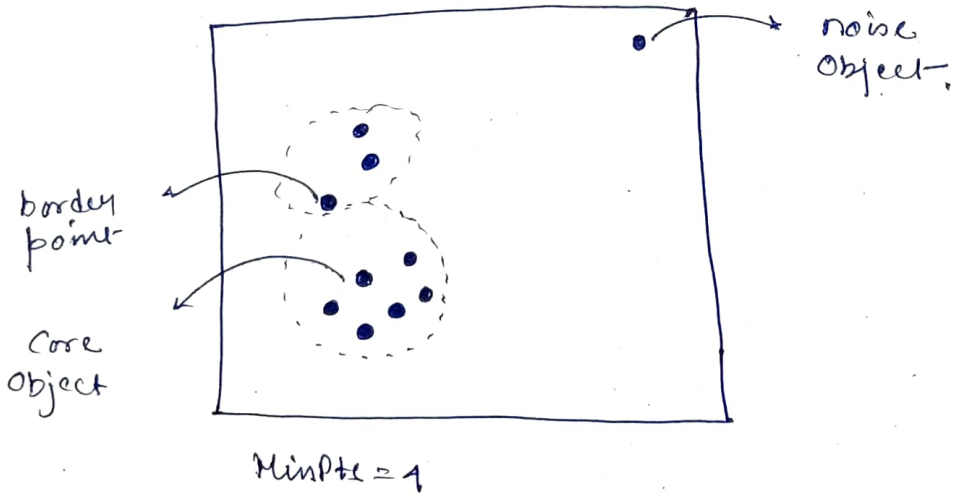
For a certain value of MinPts and ϵ, a point 'p' will be a core object if there are more or equal number of points as of MinPts within a radius of ϵ of the point p.

Border Object.

For a specific MinPts and ϵ, a border object will be having less no. of points than MinPts within a radius of ϵ, but the point is a neighbor of a core object.

Noise Object.

Neither a core nor a border object. Noise objects lie far away from core objects or border objects.



MinPts = 4

(ii) Density reachability.

A point 'q' will be density reachable from point 'p' if —

there exist some points $p_1, p_2, \ldots p_n$

where $p_1 = p$ and $p_n = q$

then p_2 is directly-density reachable from p_1

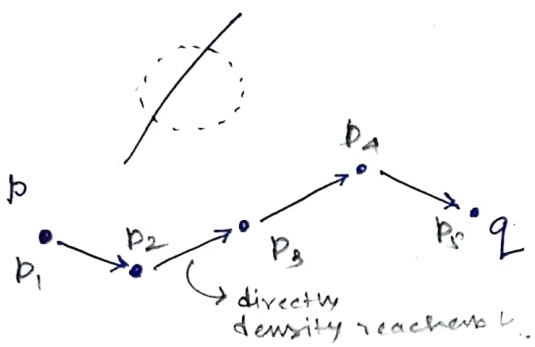
p_3 is " " " p_2

⋮

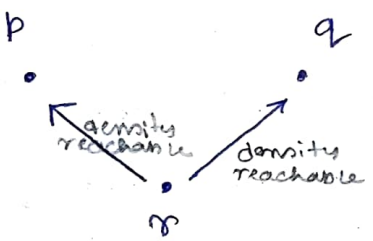
$p_n(q)$ is " " " p_{n-1}

Density Connectivity.

A point 'q' will be density connected to point 'p' if there exists another point 'r' and both 'p' and 'q' are density reachable from point 'r'.



Density reachable



Density Connectivity

(b)

(i) Maximality Condition.

It states that if a point 'p' is within a cluster 'C' (i.e $p \in C$), and another point 'q' which is density reachable from 'p' then point 'q' is also within the cluster i.e. $q \in C$.

3

(ii) Connectivity Condition.

It states that if two points 'p' and 'q' are within a cluster C, then 'p' & 'q' will be density connected to each other.

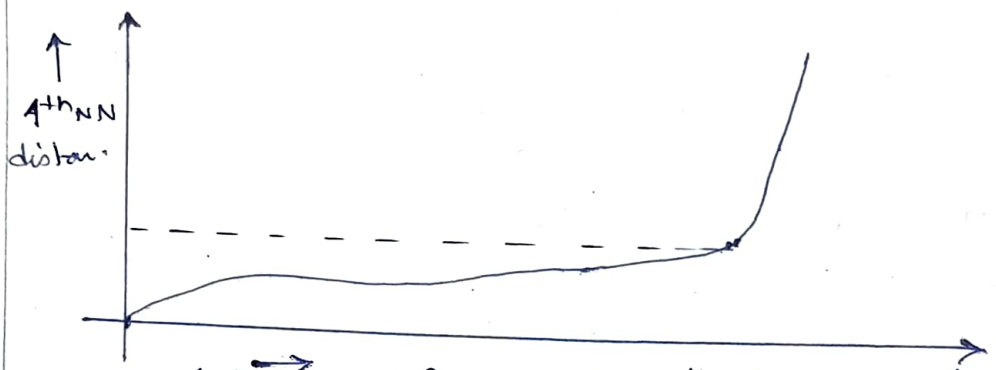
(c) The main idea of determining Eps and MinPts in DBSCAN algo is —

The K^{th} nearest neighbors of each ~~point~~ object will be approximately at the same distance.

for noisy data, the K^{th} nearest neighbors will be at a farther ~~for~~ distance.

3

Hence, for a particular value of 'K', we plot the graph of sorted points according to distance of K^{th} nearest neighbor with it the 'K' K^{th} nearest neighbor distance.



Let's assume $K = 4$.

There will be a sharp point in the plot due to the noisy data.

At that point the value of 'K' can be considered as MinPts and and the K^{th} nearest neighbor will be considered as Eps.

2.

(a)

(i) Purity.

purity is defined as —

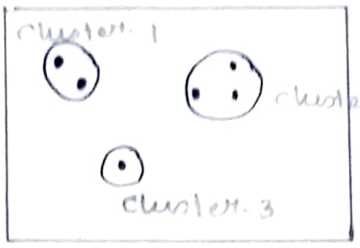
$$\text{purity} = \frac{\text{maximum number of points in a cluster.}}{\text{total number points (including all clusters)}}$$

for example, there are total 6 points.

2 points form cluster-1

3 points form cluster-2

1 point form cluster-3



the purity will be = $\frac{\max(2,3,1)}{6} = \frac{3}{6} = \frac{1}{2}$

(ii) Rand Index.

Here we compare the clusters formed by the algorithm with the ground truth data.

There can be 4 different possibilities.

3

Prediction.

Ground Truth	Same Class in Clustering	Different Class in Clustering
Same class in ground truth	A	B
Different class in ground truth	C	D

Rand Index = $\frac{A+D}{A+B+C+D}$.

(b) Intercluster distance.

It refers to the distance between two different clusters. It is denoted by 'δ'. There are several ways to calculate intercluster distance —

(i) Single linkage distance.

for any two clusters S, T; it refers to the minimum distance between two points of S and T.

$$\delta = \min(d(x,y))$$

$x \in S$, $y \in T$, $d(x,y)$ distance between points x & y.

(ii) Complete linkage distance.

for any two cluster S, T; it refers to the maximum distance between two points in S and T.

$$\delta = \max(d(x,y))$$

$x \in S$, $y \in T$, $d(x,y)$ distance between points x & y.

Intra-cluster distance.

It refers to the distance between two points within the same cluster. It is denoted by 'Δ'. There are several ways to calculate intra-cluster distance.

(i) Complete diameter distance.

It is the farthest distance of any two nodes within the same cluster, S.

$$\Delta = \max(d(x,y))$$

$x, y \in S$ and $d(x,y)$ distance between two points x and y.

5

(ii) Average diameter distance.

It is the average distance of all points with other points.

$$A = \frac{\sum d(x,y)}{|S|(|S|-1)}$$

$x, y \in S$ but $x \neq y$.

$|S|$ is the number of points in cluster S.

$d(x,y)$ distance between two points x & y.

(c) Dunn's Index

$$= \max_{1 \le i \le c} \left\{ \max_{\substack{1 \le j \le c \\ j \neq i}} \left\{ \frac{\delta(x_i, x_j)}{\max_{1 \le k \le c} (\Delta(x_k))} \right\} \right\}$$

$\delta(x_i, x_j)$ is the inter-cluster distance between two clusters x_i & x_j.

$\Delta(x_k)$ is intra-cluster distance of the cluster x_k.

A cluster will be 'good' cluster if its intra-cluster distance is less and inter-cluster distance with other clusters is more.

3½ Hence, in Dunn's index, the aim will be to ~~decrease~~ increase $\delta(x_i, x_j)$ and decrease $\Delta(x_k)$.

Hence, if the Dunn's index value is more, the cluster is good cluster, the more Dunn's index value, better the cluster is.

3.

(a) support.

~~An~~ In any set of transactions 't', an itemset $\{x, y\}$. The support of the itemset will be

$$= \frac{\text{number of transactions where } x, y \text{ appears.}}{\text{total number of transactions.}}$$

$$= \frac{(X \cup Y).\text{count}}{n} \qquad [n \rightarrow \text{total no. of transaction}]$$

confidence:

In any set of transactions 't', an association rule $X \rightarrow Y$ is there.

The confidence of the rule will be

$$= \frac{\text{no. of transactions where } \{X, Y\} \text{ appears.}}{\text{no. of transaction where } X \text{ appears.}}$$

$$= \frac{(X \cup Y).\text{count}}{X.\text{count}}$$

- for a specific minsup (minimum ~~sp~~ support) and minconf (minimum confidence) value, when the support of itemset will be greater than the minsup value, it will be refferred to as a frequency item set.

Similarly, for an association rule, if it confidence value is greater than the ~~min~~ ~~min~~ minconf, then the association rule will be an important rule.

P.T.O

(b) Let Bread = I_1, Butter = I_2, Milk = I_3, Jelly = I_4, Coke = I_5.

Transactions	Items
T_1	I_1, I_2
T_2	I_1, I_3, I_2
T_3	I_1, I_4, I_2
T_4	I_1, I_5
T_5	I_1, I_3
T_6	I_3, I_5

minsup = 30%.
min conf = 80%.

C_1: $\gamma I_1 \gamma = 5$, $\gamma I_2 \gamma = 3$, $\gamma I_3 \gamma = 3$, $\gamma I_4 \gamma = 1$, $\gamma I_5 \gamma = 2$

support $\gamma I_1 \gamma = \frac{5}{6} > 30\%$. support $\gamma I_3 \gamma = \frac{3}{6} > 30\%$.

support $\gamma I_2 \gamma = \frac{3}{6} > 30\%$. support $\gamma I_4 \gamma = \frac{1}{6} < 30\%$.

support $\gamma I_5 \gamma = \frac{2}{6} > 30\%$.

$\therefore f_1$: $\gamma I_1 \gamma$, $\gamma I_2 \gamma$, $\gamma I_3 \gamma$, $\gamma I_5 \gamma$

C_2: $\gamma I_1, I_2 \gamma = 3$ $\gamma I_1, I_3 \gamma = 2$ $\gamma I_1, I_5 \gamma = 1$

$\gamma I_2, I_3 \gamma = 1$ $\gamma I_2, I_5 \gamma = 0$ $\gamma I_3, I_5 \gamma = 1$

support $\gamma I_1, I_2 \gamma = \frac{3}{6} > 30\%$. support $\gamma I_2, I_3 \gamma = \frac{1}{6} < 30\%$.

support $\gamma I_1, I_3 \gamma = \frac{2}{6} > 30\%$. support $\gamma I_2, I_5 \gamma = 0 < 30\%$.

support $\gamma I_1, I_5 \gamma = \frac{1}{6} < 30\%$. support $\gamma I_3, I_5 \gamma = \frac{1}{6} < 30\%$.

$\therefore f_2$: $\gamma I_1, I_2 \gamma$, $\gamma I_1, I_3 \gamma$

$\therefore C_3 = \{I_1, I_2, I_3\}$

but this cannot be considered as $\{I_2, I_3\} \notin f_2$.

$\therefore C_3 = \emptyset$.

Hence no important rules can be generated.

Drawbacks of a-priori algorithm.

- Setting the minsup, minconf value incorrectly may not lead to a optimal solution.
- The time taken to train the model is relatively higher.

4.

(a) Total Yes (p) = 6

Total No (n) = 4

Entropy of entire table —

$$I(p,n) = -\frac{p}{p+n}\log\left(\frac{p}{p+n}\right) - \frac{n}{p+n}\log\left(\frac{n}{p+n}\right)$$

$$= -\frac{6}{10}\log\left(\frac{6}{10}\right) - \frac{4}{10}\log\left(\frac{4}{10}\right)$$

$$= 0.971$$

$$E(A) = \sum \frac{p_i+n_i}{p+n} I(p_i, n_i)$$

Outlook	p_i	n_i	$I(p_i, n_i)$
Sunny	2	2	1
Overcast	2	0	0
Rain	2	2	1

$$\therefore E(\text{Outlook}) = \frac{4}{10} \times 1 + 0 + \frac{3}{10} \times 1 = 0.7$$

Humidity	p_i	n_i	$I(p_i, n_i)$
Normal	4	1	0.722
High	2	3	0.971

$$\therefore E(\text{Humidity}) = \frac{5}{10} \times 0.722 + \frac{5}{10} \times 0.971$$

$$= 0.8465$$

Wind.	p_i	n_i	$I(p_i, n_i)$
Strong	2	3	0.971
Weak	4	1	0.722

$$E(\text{Humidity}) = \frac{5}{10} \times 0.971 + \frac{5}{10} \times 0.722$$

$$= 0.8465$$

$$\therefore \text{Information Gain (outlook)} = I(p,n) - E(\text{Outlook})$$

$$= 0.971 - 0.7$$

$$= 0.271$$

Similarly, Gain (Humidity) = $0.971 - 0.8465 = 0.1245$

Gain (Wind) = $0.971 - 0.8465 = 0.1245$

∴ Outlook has highest information gain. It will be the root node.

(b) possible terminating Criteria.

3

(i) All samples of a given node belong to the same class.

(ii) There is no remaining attribute for further partitioning.

(iii) There is no more samples left.

(c) Model Overfitting may occur due to several reasons —

(i) The model can become more complex than required.

(ii) The model can train noisy data or outliers which will lead to a overfitting ~~seee~~ scenario.

2/2

In decision-tree we solve overfitting by mainly two approaches —

(i) Pre-pruning

Here, we don't let the tree to grow at its fullest. At a threshold we prune the tree.

(ii) Post-pruning

Here, we let the ~~grow~~ tree to grow at its fullest & then start the pruning.