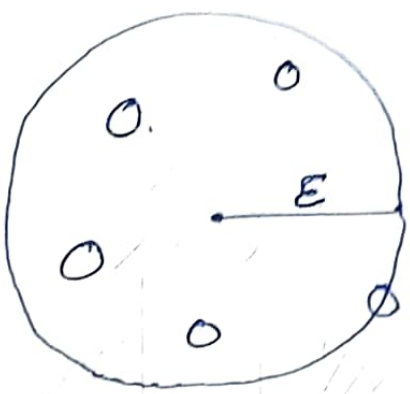
1) a) i) Core Object: Objects that are within the range of specified Radius (R_s) & are greater than or equal to a certain number ($minpts$) are known as Core objects.

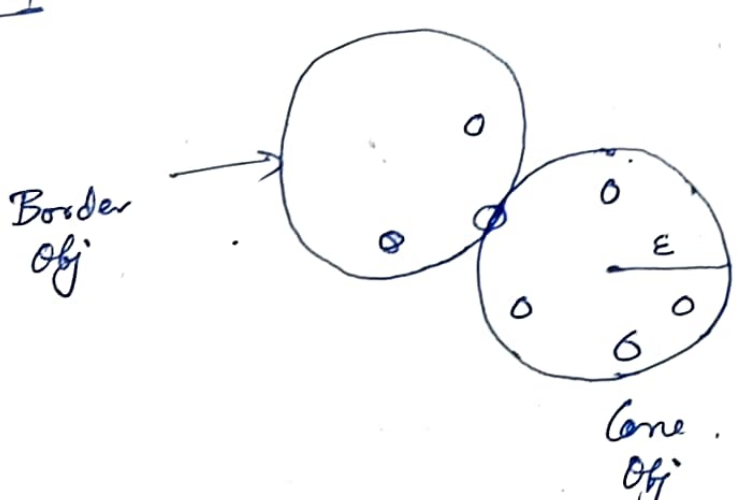
E.g. $minpts = 4$



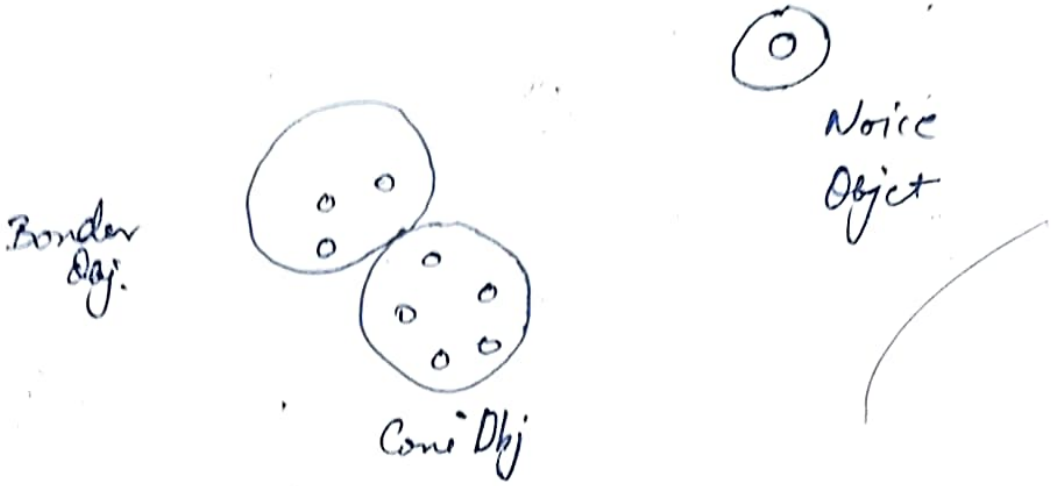
Core Object

ii) Border Object: Objects that are fewer than $minpts$ & are at ε-neighbourhood of core object are known as border objects.

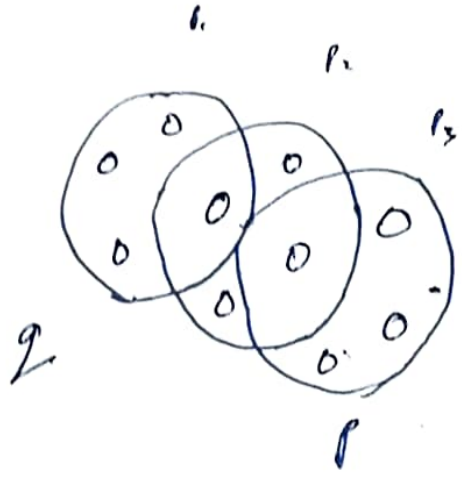
E.g. $minpts = 4$



a)i) Noice Object: Object that are not Core objects or Border Objects are Noice Objects. They are far away from Clusters.



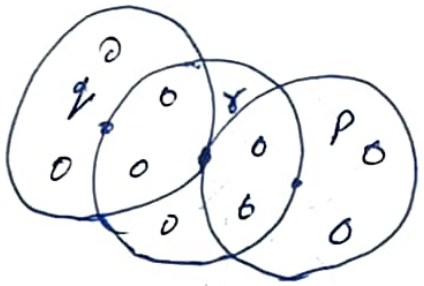
i)ii) Density Reachibility: An object q is said to be Density reachible to p if there exist $p_1, p_2 \ldots p_n$, where p_1 is q & p_n is p that are Direct Density Reachable in a series ie p_k is directly Density Reachable to p_{k-1}. Then p & q are called Density reachable.



Hence p & q are density Reachable.

1)a) iii) Density Connectivity:- Object p & q are said to be Density connected if they are directly density connected to a common object o.

4 1/2



Here r is direct density connected to both p & q

1)b) i) Maximality Condition:- If q ∈ C_1 & it is density reachable to p ∈ C_2, then both p, q ∈ C.

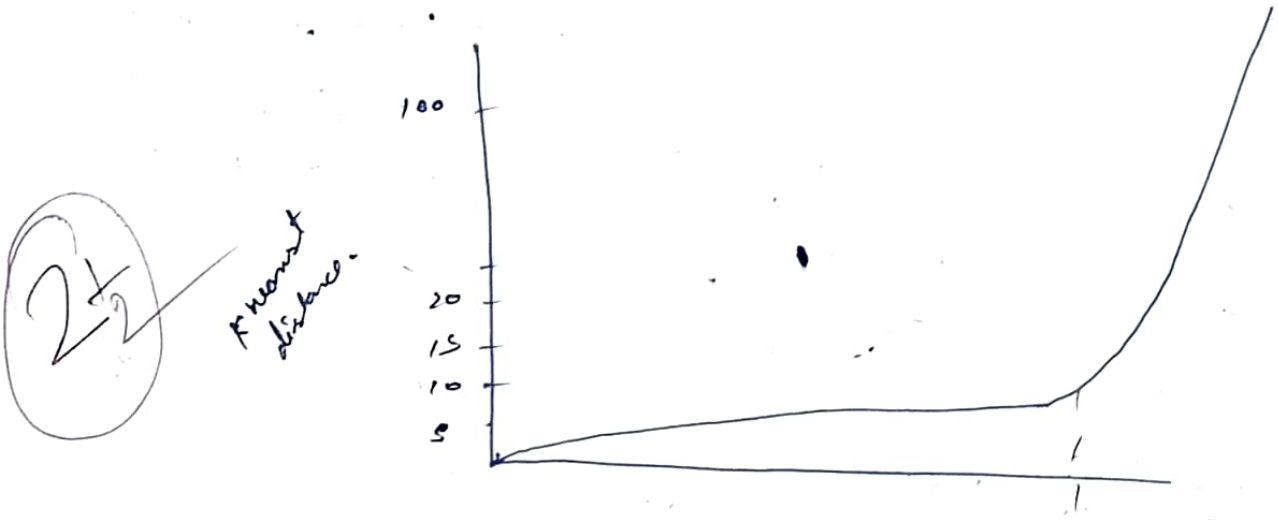
Here p & q are core objects

ie Density Reachable clusters are combined to form a single cluster, untill there is no q left that can be added to the Cluster.

2 1/2

ii) Connectivity Condition:- If q is border object to the core object p ∈ C, then, q is combined to the cluster C ie q, p ∈ C & the Noices that are far away are discarded.

a. 1) c) Eps :- is specified radius E, & Minpts is the minimum number of points within that radius. The Eps & Minpts are determined using distance of k-nearest neighbours.



The sharp turn in graph decides the optimal distance radius E. Minpts value is decided using above. Higher value of Minpts will isolate many cluster & lower values of Minpts will make Noise a cluster. So, deciding E & Minpts is crucial part of Density-Based ~~spatial~~ Clustering of Application with Noise (DBSCAN).

3) a) Let I be the item set $(i_1, i_2, i_3 \ldots i_n)$ & T be the transaction $(t_1, t_2, t_3 \ldots t_n)$.

Support: It is the ratio number of transactions X occured in to the total Transactions number.

$$\text{Sup}(X) = \frac{(X \cup Y) \cdot \text{count}}{n}$$

; n is total no. of transaction.

3

Confidance: Confidance is ratio of number of times X occurs in diffent transaction to the support of X.

$$\text{Conf}(X) = \frac{(X \cup Y) \cdot \text{Count}}{X \cdot \text{Count}}$$

3a) A k-item set is referred to a frequent itemset if it is found in at least k-1 transations.

3a) An association rule is referred as important if its confidance is 100%. A higher value of Confidance indicats high chance of association rule to be valid.

3)b) Min Support 30%,

Min Confidance 80%,

Transactions	Items (I1) (I2)
T_1	Bread, Butter
T_2	Bread, Milk, Butter (I3)
T_3	Bread, Jelly, Butter (I4)
T_4	Bread, Coke (I5)
T_5	Bread, Milk
T_6	Milk, Coke

Support $I_1 = \frac{5}{6}$

$I_2 = \frac{3}{6}$

$I_3 = \frac{3}{6}$

$I_4 = \frac{1}{6}$

$I_5 = \frac{2}{6}$

Confidence $\{I_1, I_2\} : 3$, $\{I_1, I_3\} : 2$, $\{I_1, I_4\} : 1$

$\{I_1, I_5\} : 1$, $\{I_2, I_3\} : 1$, $\{I_2, I_4\} : 1$

$\{I_3, I_5\} : 1$,

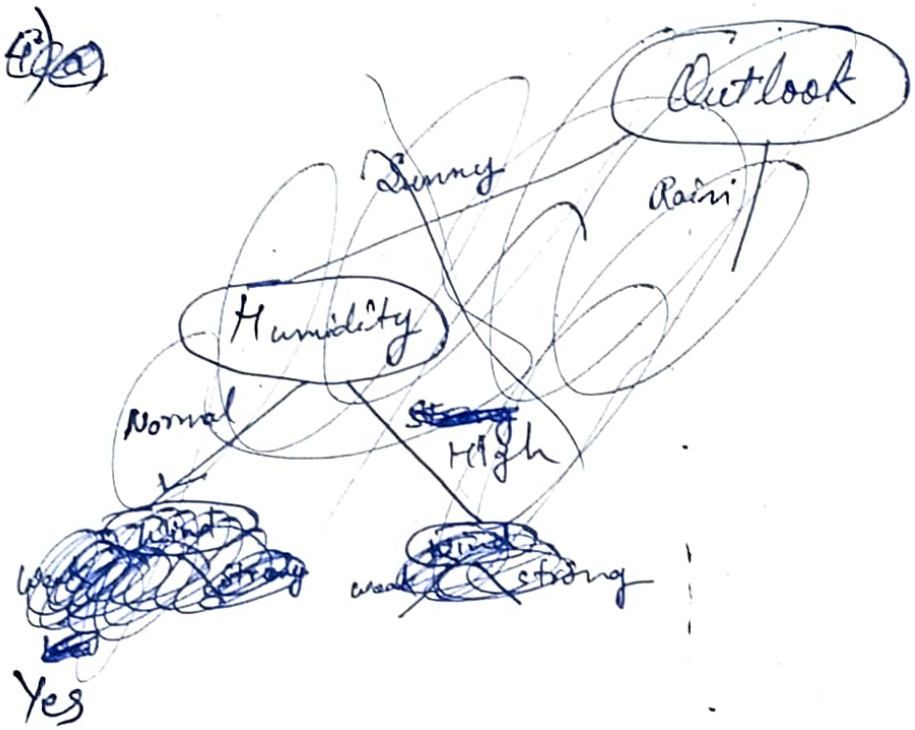
Frequent one $\{I_1, I_2\}$, $\{I_1, I_3\}$,

3-Itemsets $\{I_1, I_3, I_2\} : 1$; $\{I_1, I_4, I_2\} : 1$ { $\{I_2, I_4\}$ are not frequt so, Not considered }

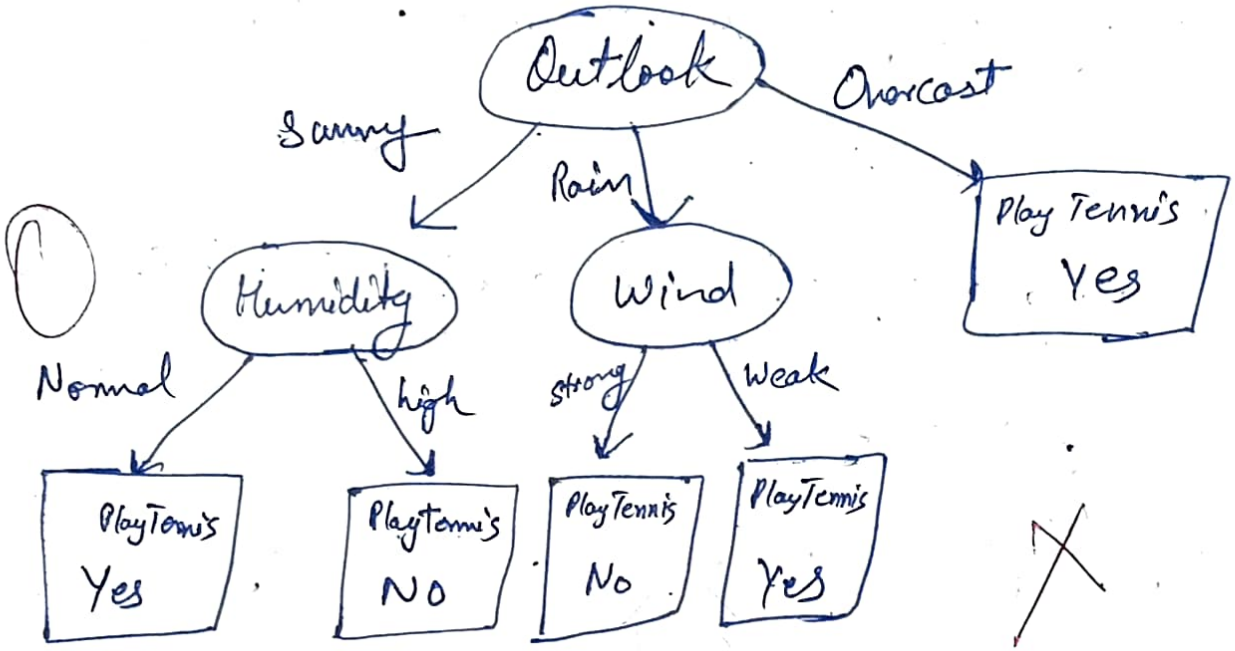
3C2

X	Y	Confidence
I_1	$\rightarrow I_2, I_3$	$\frac{1}{5}$
I_2	$\rightarrow I_1, I_3$	$\frac{2}{3}$
I_3	$\rightarrow I_1, I_2$	$\frac{3}{3}$
$I_2 I_3$	$\rightarrow I_1$	$\frac{1}{1}$
I_1, I_3	$\rightarrow I_2$	$\frac{1}{2}$
I_1, I_2	$\rightarrow I_3$	$\frac{3}{3}$

3)(c) Drawback of a-priori algorithm is there can be different minimum support for different items in the item set, a-priori does not consider the multiple minimum support. To overcome this problem we use MSpriori algorithm.



4) a)



Root of a decision Tree is Outlook.

4) b) Possible Terminating criteria for Decision Tree algorithm:

- If a node can decide the output class, then there is no need for further branching
- If the all the decision nodes are used no further nodes can be made & only leaves will be the output Class.
- If A decision node is branched but their output is same, then the succer decision node can be removed.

4) c) Causes of Overfitting occurs when the cases for which the model was not trained gives wrong output. or inconsistant decision nodes are formed.

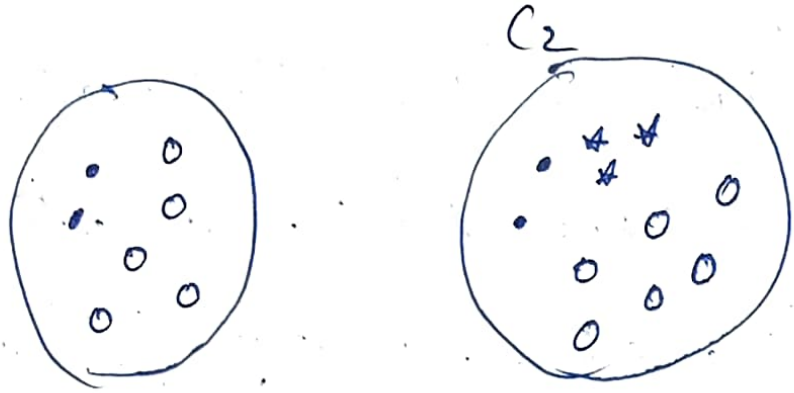
The overfitting can be solved in 2 ways

Pre Pruning:- Removing the branch of tree that is not required & is not able to perform well classification.

Post Pruning:- Adding a decision node in the tree that will be correctly ~~classify the output~~ classify the output.

2) a) i) Purity:- is a measure of validation of the cluster formed.

$$\text{Purity} = \frac{\max(w_1, w_2, \ldots w_n)}{\text{Number of points in Cluster}}$$



$$P(C_1) = \frac{\max(2,5)}{7} = \frac{5}{7}$$

$$P(C_2) = \frac{\max(2,3,6)}{11} = \frac{6}{11}$$

3

2)e)ii) Rand Index :-

Ground truth \ Predicted Output	C	~C
C	TP	FN
~C	FP	TN

$$\text{Rand Index} = \frac{TP + TN}{TP + FN + FP + TN}$$

TP & TN represent correctly classified clusters.

2) b) Inter Cluster Distance:- Inter Cluster distance is the distance between the 2 clusters. A good classifying algorithm will try to increase Inter Cluster Distance.

Intra Cluster Distance:- Intra Cluster distance is the distance between 2 points in the same Cluster. A good classification algorithm will try to reduce the intra Cluster distance.

These distances are calculated using different methodologies.

Inter Cluster Distances:-

$$\text{Mean Cluster Distance} = \frac{\text{Summation of All possible inter Cluster distances}}{\text{No. of Possible Cluster distances}}$$

$$\text{root mean Square} = \frac{\sqrt{\sum_{i,j=1}^{k} (\mathcal{E}(x_i, x_j))^2}}{N}$$

Intra Cluster Distances:-

$$\text{Mean Intra Cluster Distance} = \frac{\sum_{K=1}^{k} \Delta x_K}{N}$$

$$\text{root mean Square Intra Cluster Distance} = \frac{\sqrt{\sum_{K=1}^{k} (\Delta x_K)^2}}{N}$$

2) c) Dunn's Cluster validation Index: Let $\delta(x_i, x_j)$ be the inter cluster distance & Δx be the intra cluster distance then Dunn's Index is defined as

$$\text{Dunn's Index} = \min_{1 \le i \le c} \left\{ \min_{\substack{1 \le j \le c \\ j \ne i}} \left\{ \frac{\delta(x_i, x_j)}{\max_{1 \le k \le c} (\Delta x_k)} \right\} \right\}$$

A higher value of Dunn's index indicate ~~more~~ high inter cluster distance & low intra Cluster distance. This is used in Cluster evalation to decide how well a cluster classification performed. A lower Dunn's Index will indicate low inter Cluster distance & high intracluster distance which indicates the clusters formed are not well classified as the aim of a good classification Algorithm is to maximize interCluster distance & minimize intra-Cluster distance.