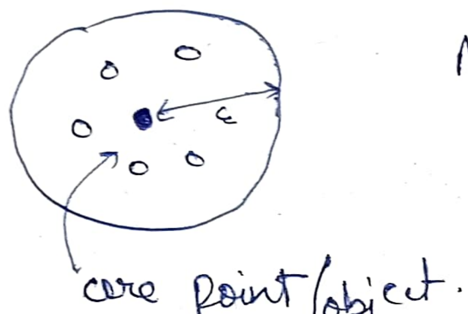
Q1.

Ans (a) (i) Core Object:

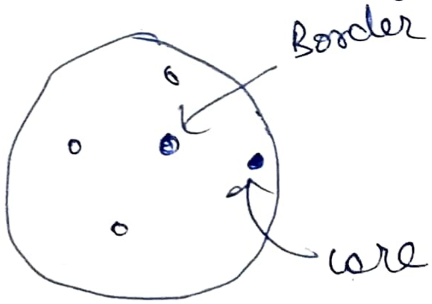
An object is a core object if it has at least MinPts number of points in its neighbourhood of radius ε called epsilon neighbourhood.



Min pts = 5

Border object:

It is an object whose neighbourhood has less than MinPts number of points and it ~~lies~~ lies in the neighbourhood of a core object



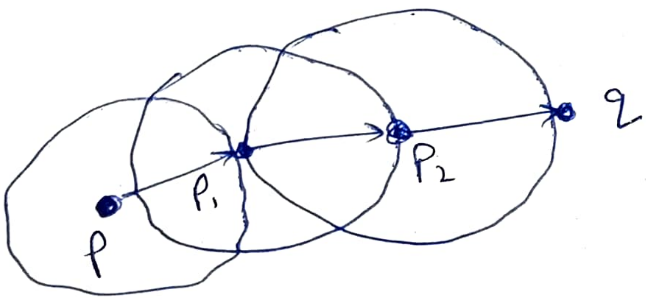
MinPts = 6

Noise Object:

It is an object which is neither a core object nor a border object i.e. its neighbourhood does not have $\geq$ minpts points and it does not lie in the neighbourhood of a core object.

(ii) Density reachability:

A point q is density reachable from a point p if there is a sequence of points $p_1, p_2, p_3 \dots p_n$. such that p_1 is directly density reachable from p, p_2 from p_1, q from p_n and so on.



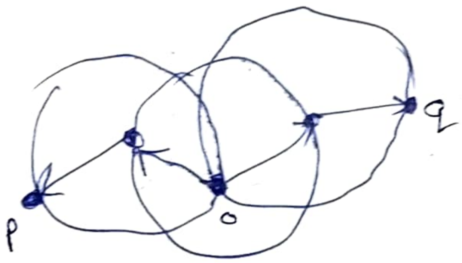
→ q is density reachable from p.

→ Asymmetric.

Density connectivity:

Two objects p and q are said to be density connected if there exists a point o such that p is density reachable from o and q is density reachable from o.

5



→ p & q density connected.

→ Symmetric.

Ans (b). (i) Maximality Condition:

It states that if there is a cluster C and an object $p \in C$ then if q is density reachable from p that means q also belongs to C.

2

(ii) Connectivity Condition:

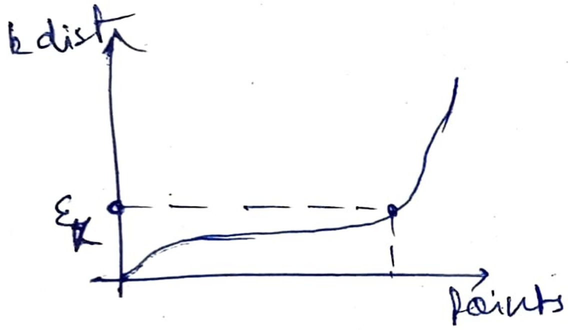
Any two point $p, q \in C$, it is true that p and q are density connected to each other.

Ans (c). The idea to find Eps and MinPts for DBSCAN algorithm is that the kth nearest neighbour is farther for noise points than the points in the cluster.

Steps:

1. Choose an appropriate k.
2. Compute distance of kth nearest neighbor, k-dist for all points.
3. Sort all points according to k dist and plot the data.

eg.



4. Mark the point of sharp change as shown.
5. MinPts $= k$

   ε = distance of kth neighbor at sharp point.

Remarks for choosing k

→ small k means noise will be classified as a cluster.

→ Large k means small clusters discarded as noise.

→ Appropriate k = expected minimum size of clusters.

Q2.

Ans(a). External cluster validation checks the data with ground truth labels.

(i) Purity:

Purity is a validation measure which checks the maximum proportion of the cluster which belongs to the same class. for w_i cluster

$$\text{purity}(w_i) = \frac{1}{n_i} \max_j (n_{ij})$$

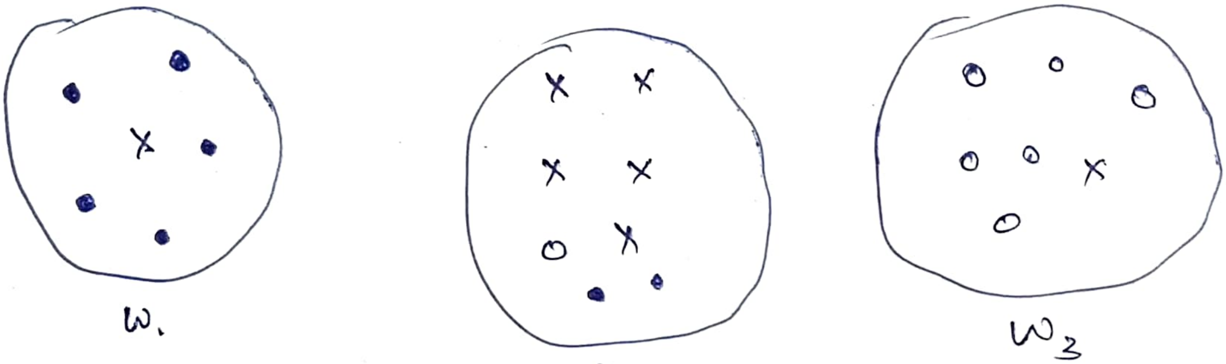
n_i → number of objects in i^{th} cluster w_i

n_{ij} → number of objects in w_i belonging to j^{th} ground truth class.

example:

Consider a ground truth with three classes $\{ \bullet : 7, \circ : 7, \times : 7 \}$.

And after clustering



ω_1 ω_2 ω_3

$Purity(\omega_1) = \frac{5}{6}$ $Purity(\omega_2) = \frac{5}{8}$ $Purity(\omega_3) = \frac{6}{7}$

(ii) Rand Index:

Rand index uses the ground truth and predicted ~~classes~~ clusters to generate a confusion matrix to compute accuracy of model.

Actual/Clustered	Same as ground truth	Different
same as clustered	A	C
Different	B	D

$$RI = \frac{A+D}{A+B+C+D}$$

Example: consider a dataset with 12 points which after clustering generates the following matrix

(3)

Actual/predicted	Same	Different
Same	4	1
Different	1	6

$$RI = \frac{4+6}{4+6+1+1} = \frac{10}{12}$$

Ans (b). Inter cluster distance:

It is the distance between any two clusters ~~gene~~ of objects

It measures the spatial seperation aspect of clustering.

Intra Cluster Distance

It is the distance of the diameter of the cluster.

It measures the compactness aspect of clustering.

→ Two intercluster distances:

i) Single linkage distance:

Distance between the closest points among the clusters.

$$\delta(X_i, X_j) = \min_{\substack{p \in X_i \\ q \in X_j}} \{d(p, q)\}$$

d → any distance measure.

(ii) Complete linkage distance

Distance between two farthest points among clusters.

$$\delta(X_i, X_j) = \max_{\substack{p \in X_i \\ q \in X_j}} \{d(p, q)\}$$

→ two intracluster distances:

i) Complete diameter distance:

Distance between farthest objects within the same cluster.

$$\Delta(X) = \max_{p,q \in X} \{d(p,q)\}$$

5

ii) Average diameter distance:

Average of distances between pair of objects in the cluster.

$$\Delta(X) = \frac{1}{|X|(|X|-1)} \left\{ \sum_{\substack{p,q \in X \\ p \neq q}} d(p,q) \right\}$$

Ans (C) Dunn's Index is an internal cluster validation index defined as

$3\frac{1}{2}$

$$\text{DIndex}(U) = \min_{1 \le i \le c} \left\{ \min_{\substack{1 \le j \le c \\ j \neq i}} \left\{ \frac{\delta(X_i, X_j)}{\max_{1 \le k \le c} \{\Delta(X_k)\}} \right\} \right\}$$

where U is a set of clusters with k clusters

The Dunn's index have intercluster distance in the neumerator and the intracluster distance in the denominator. A high intercluster distance and a low intracluster ~~distand~~ distance is ~~require~~ desired from a ~~clustering~~ clustering.

A high value of the Dunn's index represents a good clustering ~~with~~ and a low, bad clustering. In this way DI is used in cluster evaluation.

Q4.

Ans(a) For identifying the root of the decision tree, we need to compute the information gain after splitting with each attribute and take the maximum one. Yes - p , No - n

Initial entropy:

~~p n I(p,n)~~

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

p	n	I (p, n)
6	4	0.94

$$I(6,4) = \frac{-6}{10}\log_2\frac{6}{10} - \frac{4}{10}\log_2\frac{4}{10}$$

$$= -\frac{3}{5}\left[\log_2 3 - \log_2 5\right] - \frac{2}{5}\left[\log_2 2 - \log_2 5\right]$$

$$= -\frac{3}{5}\left[1.6 - 2.3\right] - \frac{2}{5}\left[1 - 2.3\right]$$

$$= 0.94$$

Splitting w.r.t 'Outlook':

Outlook	p	n	I (p, n)
Sunny	2	2	1
Overcast	2	0	0
Rain	2	2	1

$$I(2,2) = \frac{-2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}$$

$$= \frac{-1}{2}\left[\log_2 1 - \log_2 2\right] - \frac{1}{2}\left[\log_2 1 - \log_2 2\right]$$

$$= \frac{-1}{2}\left[0 - 1\right] - \frac{1}{2}\left[0 - 1\right]$$

$$= 1$$

$$I(2,0) = 0$$

Entropy for Attribute

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

$\Rightarrow$ $E(\text{outlook}) = \frac{4}{10} \times 1 + \frac{2}{10} \times 0 + \frac{4}{10} \times 1$

$= 0.8$

$\text{Gain}(\text{Outlook}) = I(P,n) - E(\text{Outlook})$

$= 0.94 - 0.8$

$= 0.14$

→ For splitting wrt "Humidity"

Humidity	P	n	I(P,n)
Normal	4	1	0.7
High	2	3	0.94

$$I(4,1) = -\frac{4}{5}\log_2 \frac{4}{5} - \frac{1}{5} \log_2 \frac{1}{5}$$

- $I(4,1) = -\frac{4}{5}[\log_2 4 - \log_2 5] - \frac{1}{5}[\log_2 1 - \log_2 5]$

$= -\frac{4}{5}[2 - 2.3] - \frac{1}{5}[0 - 2.3]$

$= 0.7$

$I(2,3) = -\frac{2}{5}\log_2 \frac{2}{5} - \frac{3}{5}\log_2 \frac{3}{5}$

$= -\frac{2}{5}[\log_2 2 - \log_2 5] - \frac{3}{5}[\log_2 3 - \log_2 5]$

$= -\frac{2}{5}[1 - 2.3] - \frac{3}{5}[1.6 - 2.3]$

$= 0.94$

$E(\text{Humidity}) = \frac{5}{10} \times 0.7 + \frac{5}{10} \times 0.94$

$= 0.82$

$\text{Gain}(\text{Humidity}) = 0.94 - 0.82 = 0.12$

→ For splitting wrt "Wind"

Wind	P	n	I(P, n)
Strong	2	3	0.94
Weak	4	1	0.7

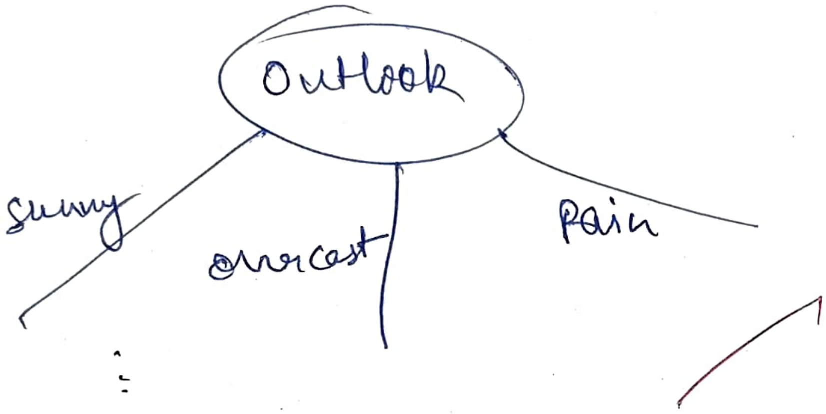
$I(2,3)$ & $I(4,1)$ already calculated

$E(\text{wind}) = \frac{5}{10} \times 0.94 + \frac{5}{10} \times 0.7 = 0.82$

Gain (Wind) = 0.94 - 0.82 = 0.12

=> Gain (Outlook) is maximum, Hence outlook will be the root of decision tree for this dataset.

5



Ans (b): Termination criteria for decision tree algorithm :

3

1) All items of the set belong to the same class. In this case class label is assigned to the node and terminated.

2) No more attributes left to split. In this case majority voting decides class.

3) No more items left to split in the set.

Ans (c). Causes of overfitting:

1) Too many data points of similar criteria.

2) Inconsistent data present in the dataset.

$2\frac{1}{2}$

Methods to avoid overfitting:

1) Stopping the splitting of the node if it hampers the goodness evaluation. This is difficult to implement.

2) Pruning the tree based on validation dataset

3) Using statistical methods like χ^2 test to determine whether the split is viable.

Q5.

Ans (a) ~~For categorizing~~ ~~According to given dataset~~

For classifying using naive Bayes, we need to maximize

$$P(X \mid C_k)\, P(C_k)$$

where $C_k \rightarrow$ class label

$X \rightarrow$ evidence, $\{x_1, x_2, \dots x_n\}$

and $P(X \mid C_k) = P(x_1 \mid C_k) \times P(x_2 \mid C_k) \dots P(x_n \mid C_k)$

According to given data, Two classes.

$$P(\text{PlayTennis} = \text{Yes}) = \frac{6}{10}$$

$$P(\text{Play Tennis} = \text{No}) = \frac{4}{10}$$

$$P(\text{Outlook} = \text{overcast} \mid \text{PlayTennis} = \text{Yes}) = \frac{2}{6}$$

$$P(\text{Outlook} = \text{overcast} \mid \text{PlayTennis} = \text{No}) = \frac{0}{4}$$

Using laplacian correction for the error (explained in (b))

$$P(\text{Outlook} = \text{overcast} \mid \text{Playtennis} = \text{No}) = \frac{1}{7}$$

Because outlook has three labels.

- $P(\text{Humidity} = \text{High} \mid \text{PlayTennis} = \text{Yes}) = \frac{2}{6}$

$P(\text{Humidity} = \text{High} \mid \text{PlayTennis} = \text{No}) = \frac{3}{4}$

$P(\text{Wind} = \text{weak} \mid \text{PlayTennis} = \text{Yes}) = \frac{4}{6}$

$P(\text{Wind} = \text{weak} \mid \text{PlayTennis} = \text{No}) = \frac{1}{4}$

$4\frac{1}{2}$

B $X = \langle \text{Outlook} = \text{Overcast}, \text{Humidity} = \text{High}, \text{Wind} = \text{weak} \rangle$

$P(X \mid \text{PlayTennis} = \text{Yes}) = \frac{2}{6} \times \frac{2}{6} \times \frac{4}{6} = 0.074$

$P(X \mid \text{PlayTennis} = \text{No}) = \frac{1}{7} \times \frac{3}{4} \times \frac{1}{4} = 0.027$

$\Rightarrow P(\text{PlayTennis} = \text{Yes} \mid X) = P(X \mid \text{PlayTennis} = \text{Yes})\, P(\text{PlayTennis} = \text{Yes})$

$= 0.074 \times 0.6$

$= 0.0444$

Similarly

$P(\text{PlayTennis} = \text{No} \mid X) = 0.027 \times 0.4 = 0.0108$

$\Rightarrow \text{PlayTennis}(\langle \text{Outlook} = \text{Overcast}, \text{Humidity} = \text{High}, \text{Wind} = \text{weak} \rangle) = \text{Yes}$

Ans (b) Yes there was an error in computation of $P(\text{Outlook} = \text{Overcast} \mid \text{play Tennis} = \text{No})$
Due to the assumptions of independent attribute probability

4 $P(X|C_i) = P(x_1|C_i) \times \ldots P(x_n|C_i)$

if any of the $P(x_j|C_i)$ becomes 0 then the entire $P(X|C_i)$ will become 0 which is incorrect.

To resolve this, laplacian correction is used. In laplacian correction, if any attribute has a 0 probability class then 1 is added to all the classes' count and say there are k classes and n total objects then after laplacian correction there will be n+k objects.
Such the for $P(x_j)$ which was 0 now becomes $\frac{1}{n+k}$. These give close

to original probabilities and solve the 0 error.

Ans (C). If any feature/attribute A_k has continuous values then we use gaussian distribution to calculate the probability

3

$$p(x,\sigma,\mu) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where $\mu = \frac{1}{N_k} \sum x_k$

$x_k \rightarrow$ an item of A_k

$$\sigma = \sqrt{\frac{\sum (x_k-\mu)^2}{N_k}}$$

$N_k \rightarrow$ total no. of tuples with A_k

Rest of the algorithm proceeds in the same way