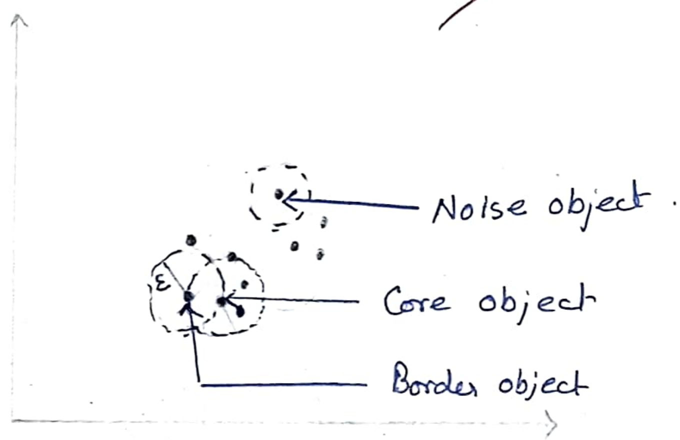
1a) i) Core object - An object with atleast MinPt number of (objects) points in it's ε-neighborhood is called a core object.

Border object - An object which does not have MinPt number of objects in it's ε-neighbourhood but is close to a core object is called border object.

Noise object - An object which is neither close to a core object nor has MinPt number of objects in it's ε-neighbourhood is called noise object.

~~ii) Density reac~~

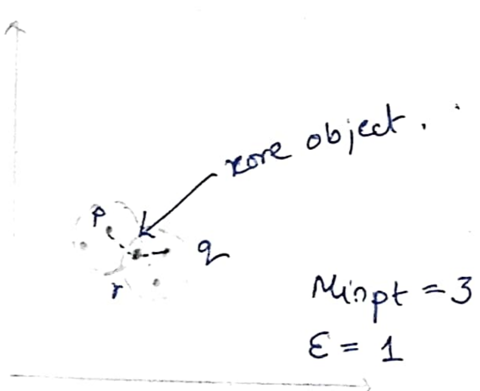
$4\frac{1}{2}$



MinPt = 4
ε = 1.

ii) Density reachability - A point q is said to be (directly) density reachable from a point p if p is a core object and q lies in the ε-neighbourhood of p.

Density connectivity - Two points p and q are said to be density connected if ~~p is directly density reachable from~~ there exists a point r such that p is directly density reachable from r and q is also directly density reachable from r.



p & q are density reachable from r.

b) In DBSCAN clustering algorithm,

i) Maximality condition states that if $p \in C$ where p is an object and C is a cluster, and q is directly density reachable from p, then q should also be in cluster C.

ii) Connectivity condition states that if $p, q \in C$ where p & q are two objects and C is a cluster, then p and q should be density connected, $\forall p, q \in C$.

c) ~~The algorithm for~~

Let the distance of the k^{th} nearest neighbour be k-dist.

If k is smaller than the size of the cluster, the k-dist will be small. If cluster size is too small, noise points may get incorrectly labeled with some cluster.

However, if ~~noise~~ ~~po~~ the clusters become too large then ~~noise~~ small ~~Points~~ clusters will be labeled as noise.

To find the optimal values of the parameters Eps and MinPts, the k-dist for all points are computed and points are sorted according to the increasing k-dist.

The point where there is a sharp change of k-dist will give the optimal size of cluster. This k-dist is takes as Eps and k becomes MinPts.

2a) Given a gold standard data with c class labels and a clustering with k ~~cla~~ clusters, denoted by $\{\omega_1, \omega_2 \ldots \omega_k\}$, the purity of the clustering is given by

$$Purity(\omega_i) = \frac{1}{n_i} \max_j (n_{ij}) \quad \text{where } j \in c$$

The ratio of the ~~size of the~~ dominant class in the i^{th} cluster, n_{ij} to the size of the cluster ω_i.

We calculate the Rand Index by calculating the ~~values~~ number of pairs that are in the same cluster and same class, that in different class, different clusters and same and different class in the following way

3

	Same clusters.	Different Clusters
Same labels	A	B
Different Labels.	C	D

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

b) Intercluster distance is the measure of dissimilarity between different clusters in a ~~clust~~ partitioning and intracluster distance is the measure of dissimilarity ~~betw~~ within a cluster in the partitioning.

4

A good partitioning will try to maximize intercluster distance and minimise intracluster distance.

Two intercluster distances are :

i) simple linkage distances : $\delta(s,t) = \min\{d(x,y) : x \in s, y \in t\}$ where s & t are two clusters.

Average linkage distance :

$$\delta(S,T) = \frac{1}{|S|\,|T|}\left\{\sum_{x\in S,\, y\in T} d(x,y)\right\}$$

where $|C|$ denotes the size of a cluster C.

Two intracluster distances are

i) complete diameter distance given by

$$\Delta S = \max_{x,y\in S}\{d(x,y)\}$$

ii) Average diameter distance given by

$$\Delta S = \frac{1}{|S|(|S|-1)} \sum_{\substack{x,y\in S\\ x\neq y}}\{d(x,y)\}$$

c) Dunn's cluster validation index is a cluster validation technique that combines intercluster and intracluster distances. and for a partitioning U is given by

$$DIndex(U) = \min_{1\leq i\leq c}\left\{\min_{\substack{1\leq j\leq c\\ i\neq j}}\left\{\frac{\delta(x_i,x_j)}{\max_{1\leq k\leq c}(\Delta X_k)}\right\}\right\}$$

3

where $\delta(x_i,x_j)$ is the ~~intr~~ intercluster distance between x_i and x_j such that x_i ~~and~~ , $x_j \in C$ and ΔX_k is the intracluster distance of a cluster X_k.

The larger the value of Dunn's cluster validation index, ~~the~~ better the ~~clustering~~ partitioning.

3a) Support is the ratio of the number of transactions that ~~ae~~ contains ~~item~~ the itemset ~~{A∪B}~~ {A ∪ B} ~~appears~~ to the total number of transactions, for an association rule $A \to B$

$sup(A \to B) = Pr(A \cup B)$

Confidence of the association rule, $A \to B$ is given by the ratio of the number of transactions containing ~~A∪B~~ the itemset $\{A \cup B\}$ to the number of transactions containing A.

$confidence(A \to B) = Pr(B/A)$.

3

An itemset is referred to as ~~a~~ a frequent item set is ~~to~~ ~~as~~ it's support is more than the minimum support and an association rule is refered to as an important rule if it's confidence is more than the minimum confidence value.

b) minsup = 0.3

min conf = 0.8

Items:

I_1 :	Bread	$sup(I_1) = 5/6$	= 0.833
I_2 :	Butter	$sup(I_2) = 3/6$	= 0.5
I_3 :	Milk	$sup(I_3) = 3/6$	= 0.5
I_4 :	Jelly	$sup(I_4) = 1/6$	= 0.166
I_5 :	Coke	$sup(I_5) = 2/6$	= 0.33

The frequent 1-itemset is {Bread, Butter, Milk, Coke} since their support values are greater than minsup.

F_1 = {{Bread}, {Butter}, {Milk}, {Coke}}

C_2 = { {Bread, Butter}, {Bread, Milk}, {Bread, Coke}, {Butter, Milk}, {Butter, Coke}, {Milk, Coke} }

Sup({Bread, Butter}) = 3/6 = 0.5

sup({Bread, Milk}) = 2/6 = 0.33

$sup(\{Bread, Coke\}) = 1/6 = 0.166 < 0.3$ $sup(\{Butter, Coke\}) = 0 < 0.3$
$sup(\{Butter, Milk\}) = 1/6 = 0.166 < 0.3$ $sup(\{Milk, Coke\}) = 1/6 = 0.166 < 0.3$

$F_2 = [\{Bread, Butter\}, \{Bread, Milk\}]$

$C_3 = [\{Bread, Butter, Milk\}]$

$sup(\{Bread, Butter, Milk\}] = 1/6 = 0.166 < 0.3$

$F_3 = \phi$

$\therefore$ The frequent itemset is {{Bread}{Butter},{Milk}{Coke}, {Bread, Butter}, {Bread, Milk}, }.

5

$conf(Bread \rightarrow Butter) = 3/5 = 0.6 < 0.8$
$conf(Bread \rightarrow Milk) = 2/5 = 0.4 < 0.8$
$conf(Milk \rightarrow Bread) = 2/3 = 0.66 < 0.8$
$conf(Butter \rightarrow Bread) = 3/3 = 1. > 0.8$

$\therefore$ The association rule is {Butter → Bread}.

c) The major drawback of the a-priori algorithm is that it does not take importance factor of an itemset into account.

0

6a) i)

Outlook	Humidity	Wind	PlayTennis (class variable) ground truth	PlayTennis (predicted label)
Sunny	Normal	Strong	Yes	Yes
Overcast	Normal	Strong	No	Yes
Rain	High	Strong	Yes	No
Sunny	High	Weak	No	No
Rain	High	Strong	No	No

Confusion Matrix :

Actual class / Predicted class	PlayTennis	¬ PlayTennis
PlayTennis	1 (True Positive)	1 (False Negative)
¬ PlayTennis	1 (False Positive)	2. (True Negative)

4

i) Precision $= \dfrac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} = \dfrac{1}{2} = 0.5$

ii) Recall $= \dfrac{\cancel{\text{False}}\ \cancel{\text{Ate}}\ \text{True Positive}}{\text{True Positive} + \text{False Negative}} = \dfrac{1}{2} = 0.5$

iii) F-score $= \dfrac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \dfrac{2 \times 1/2 \times 1/2}{1/2 + 1/2} = 1/2 = 0.5$

b) Holdout method - The training and test set is sampled uniformly with the training data being $2/3^{rd}$ of the total number of data samples and testing data being $1/3^{rd}$ of

The model is trained on the training set and evaluated ~~this~~ on the test set t times and the accuracy reported is the average of all the t observed accuracies.

ii) Cross-validation - The total number of data samples is divided into k ~~sets~~ subsets and in each iteration the model is trained on k-1 subsets (leave one out) and tested on the remaining subset.

In stratified cross validation ensures that the distribution of the classes in the training ~~sample~~ set is same as that in the original data.

iii) Bootstrap is the process of uniformly sampling ~~~~ the training ~~data~~set with replacement ~~and~~ from the given data samples and training the classifier with it iteratively. It is observed that 63.2 % data goes to the training set and 36.8 % data is unseen by the model.

c) Ensembling classifiers reduce the error rate.
For example, if we ensemble 25 classifiers, each with error rate $\epsilon = 0.38$, by ensembling or combining different classifiers by averaging or voting etc reduces the error rate.
The error rate of the ensembled classifier is given by

$$\sum_{i=13}^{25} \binom{25}{i} \epsilon^{i} (1-\epsilon)^{25-i} = 0.06.$$

~~Ensembling ensures that the~~
The main purpose of ensembles of classifiers is not to find a highly accurate model instead to combine base models which differ in the type of misclassification or errors. It also overcomes the problem of statistical problem of a larger hypothesis space and small number of samples, representational problem etc.

The Adaboost algorithm works as follows :

Designing :

~~Boost a~~ i) Assign ~~Randomly~~ equal weight, 1/N to all classes.

ii) Build complementary classifiers ~~and predict the label of the~~ given samples.

3½

iii) Calculate the error rate ε.

iv) ~~Assign~~ If the error rate ε is 0 or ~~ε~~ is greater than or equal to 0.5 ~~then~~ then terminate the process.

v) Assign the weight $\varepsilon/(1-\varepsilon)$ to all correctly classified labels.

vi) Return the weights of the labels.

Classification :

Initialize weight of all labels with equal weight

For each classifier, ~~the predicted~~ add $-\log\left(\varepsilon/1-\varepsilon\right)$ to the weight of the ~~predicted labels~~ misclassified label.

Return the label with highest weight.

5a) C_1 = ~~Play Tennis~~ Yes.

C_2 = ~~Play Tennis~~ No

~~R~~ Naive Bayes classification algorithm makes the assumption that all the attributes are conditionally independent.

Hence $P(X|C_i) = \prod_{x_k \in X} P(x_k|C_i)$

Total number of samples = 10

~~P(Outlook = Overcast | Yes) = 2/10 = 0.2~~

$P(C_1) = 6/10$

~~P(Outlook = Overcast | No) =~~

$P(C_2) = 4/10$

Since there is no such sample for which outlook is overcast and the class variable is No, we add two samples such that

4

$P(\text{Outlook} = \text{Overcast} | \text{Yes}) = 3/12$.

$P(\text{Outlook} = \text{Overcast} | \text{No}) = 1/12$

$P(\text{Humidity} = \text{High} | \text{Yes}) = 2/10$

$P(\text{Humidity} = \text{High} | \text{No}) = 3/10$

$P(\text{Wind} = \text{Weak} | \text{Yes}) = 4/10$

$P(\text{Wind} = \text{Weak} | \text{No}) = 1/10$

$P(\text{Outlook} = \text{Overcast}, \text{Humidity} = \text{High}, \text{Wind} = \text{Weak} | \text{Yes})$

$$= \frac{3}{12} \times 2/10 \times 4/10 \times 6/10$$

$P(\text{Outlook} = \text{Overcast}, \text{Humidity} = \text{High}, \text{Wind} = \text{Weak} | \text{No})$

$$= \frac{1}{12} \times \frac{3}{10} \times \frac{1}{10} \times 4/10$$

Since, the probability of ~~R~~ the class label being Yes is more, the class label will be predicted as Yes.

b) Since Naive Bayes classification algorithm ~~assumes that~~ the ~~attributes of~~ the conditional independence assumption i.e the ~~occurrence~~ value of an attribute does not depend on the value of ~~another~~ any other attribute. However, this assumption is never true in practise. Hence, there can be some error in such prediction.

0

The error can be corrected using Bayesian Belief Networks to predict the label instead.

c) If a feature has continuous values, it is either discretized using different data preprocessing techniques like binning or the Gaussian distribution is predicted using the following formula.

2

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

where σ is the standard deviation and μ is the mean of the distribution of the values of the feature.