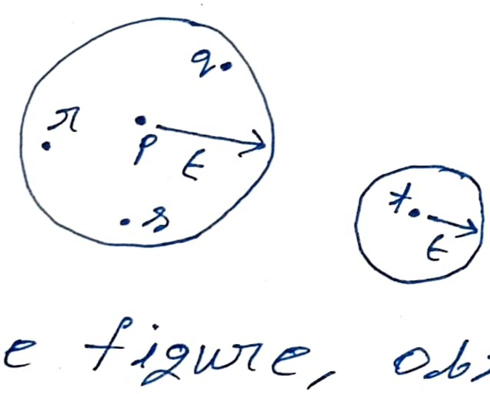
1. a) i)

core object - In DBSCAN clustering algorithm if within ϵ radius drawn from a object, there are more than or equal to minimum points number of objects, the object is referred to as core object.

Border object - The objects which are not core objects and belong within the ϵ radius of a core object, are referred to as border object.

Noise object - The objects which are neither core object nor border object are referred to as noise object in DBSCAN clustering algorithm

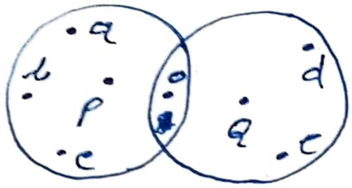
4



In the figure, object p is core object. object r, q, s are border objects and object t is noise object.

ii) <u>Density reachability</u> - ~~A~~ In DBSCAN clustering algorithm, an object p is said to be density reachable from object c if p lies inside the circle of radius ε drawn centering c.

<u>Density connectivity</u> - In DBSCAN algorithm, two objects p and q are said to be density connected if ~~they~~ another object o is density reachable from p and q.



object a is density reachable from p. and object ~~o~~ p and q are density connected through object o.

b) i) <u>Maximality condition</u> - In DBSCAN clustering algorithm, an object p is considered to form a cluster if it is density reachable from another object q. A cluster should include atleast MinPoints number of objects excluding the core object.

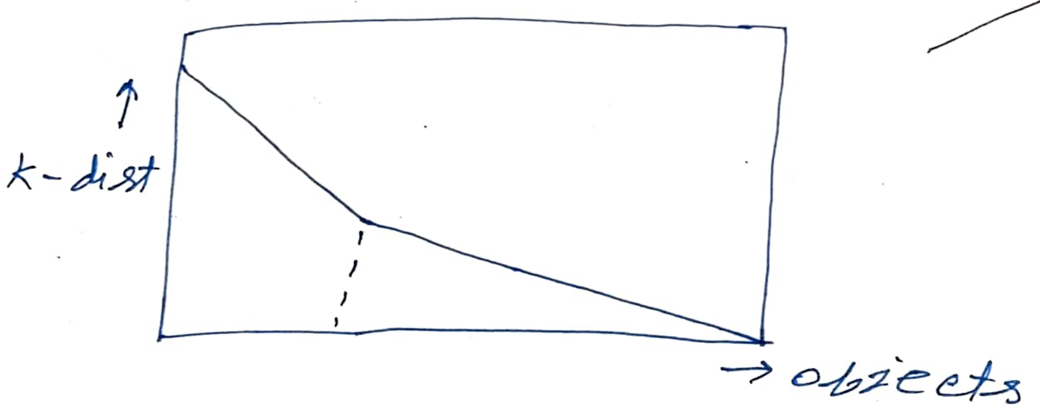
ii) Connectivity condition - In DBSCAN clustering algorithm, if two object p and q are density connected they are supposed to fall inside a single large cluster.

2

c) Determining the parameters of DBSCAN algorithm:

- Initially, for all the objects in the dataset, distance to k th nearest neighbour are determined.
- The distances to k th nearest neighbour are sorted and a corresponding graph is drawn.

2½



- The sharp change in curve i.e the elbow point is considered as the eps value.

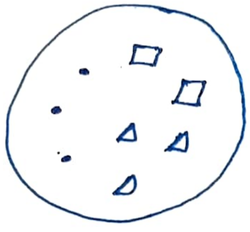
- The corresponding value of K is considered as the Mintt-s value.

2. a) i) <u>Purity</u> - The value of purity of a cluster is determined as, Purity (P) $= \frac{1}{n} \max(c_1, c_2, \dots c_n)$

n is the number of clusters

c_1, c_2 are number of objects in corresponding cluster.

3



• → objects of c_1
□ → objects of c_2
△ → objects of c_3

$$P = \frac{1}{8} \max(3, 2, 3) = \frac{3}{8}$$

ii) <u>Rand Index:</u>

	Same class in cluster	Diff class in cluster
Same class in ground truth	A	B
Diff class in ground truth	C	D

$$\text{Rand Index (RI)} = \frac{A+D}{A+B+C+D} \quad -(1)$$

Rand Index is defined by formula
1 given the matrix.
For example,

2	4
2	1

$$RI = \frac{3}{9} = \frac{1}{3}$$

b) <u>Intercluster distance</u> - The distance of two different clusters is referred to as intercluster distance (S). Generally, higher the intercluster distance, better the clustering algorithm.

<u>Intracluster distance</u> - The distance of objects within a given cluster is referred to as intracluster distance (D). Generally, lower the intracluster distance, better the clustering algorithm.

<u>Two Intercluster distance</u> -

<u>Single linkage distance</u> - The single linkage distance is the ~~maximum~~ minimum

distance of two objects belonging to two different clusters.

$$S_1 = \min_{\substack{x \in S \\ y \in T}} \{d(x,y)\}$$

<u>complete linkage distance</u> - It is the maximum distance of two objects belonging to two different clusters.

$$S_2 = \max_{\substack{x \in S \\ y \in T}} \{d(x,y)\}$$

4

<u>Two Intracluster distance</u>

~~centroid distance~~

• <u>complete diameter distance</u> - The maximum distance of two points belonging to same cluster is defined as complete diameter distance.

$$\Delta_1 = \max_{\substack{x \in S \\ y \in S}} \{d(x,y)\}$$

• Average centroid distance - The average of distances of all points from centroid of a cluster is defined as average centroid distance.

$$D_2 = \frac{1}{|S|(|S|-1)} \sum_{x \in S} d(v_S, x)$$

c) Dunn's cluster validation Index →

$$DI = \min_{\substack{1 \le k \le c \\ i \ne j}} \left(\sum \min_{\substack{1 \le i \le c \\ i \ne j}} \left(\sum \frac{\delta(x_i, x_j)}{\max(\Delta x_k)} \right) \right)$$

3

where, δ refers to the intercluster distance and Δ refers to the intracluster distance.

Dunn's Index helps determine the usefullness of a cluster algorithm. A clustering algorithm is considered good if it has high intercluster distance and low intra cluster index. Thereby high value of Dunn's Index is considered as good clustering algorithm whereas low Dunn's Index value is considered as bad clustering algorithm.

3. a) Support - Support of an association rule A → B is defined as,

$$S(A \rightarrow B) = \frac{n(A \cup B)}{n_d}$$

~~where,~~ i.e $\dfrac{\text{no. of documents containing both A and B}}{\text{Total number of documents.}}$

4

Confidence - Confidence of an association rule is defined as,

$$C(A \rightarrow B) = \frac{S(A \cup B)}{S(A)}$$ where S is the support.

An item set is referred to as frequent item set if it's ~~support a~~ the support of all items in the set is greater than or eq. minimum support threshold value.

An association rule is referred to as an important rule if it's confidence is greater ~~or~~ than or eq. minimum confidence threshold value and it's support greater or equal to minimum support.

b) Given,

$min\ s = 0.3$

$min\ c = 0.8$

Let,

I_1 = Bread

I_2 = Butter

I_3 = Milk

I_4 = Jelly

I_5 = coke

~~I_6 = Milk~~

$$S(I_1) = \frac{5}{6} = 0.83$$

$$S(I_2) = \frac{3}{6} = 0.5$$

$$S(I_3) = \frac{3}{6} = 0.5$$

$$S(I_4) = \frac{1}{6} \approx 0.17 \quad (<0.3)$$

$$S(I_5) = \frac{2}{6} = 0.33$$

$$\cancel{S(I_6) = \frac{3}{6} = 0.5}$$

Frequent set $C_1 = \{I_1, I_2, I_3, I_5\}$

excluding < min. support item. 312356

$S(I_1 I_2) = \frac{3}{6} = 0.5$

$S(I_1, I_3) = \frac{2}{6} = 0.33$

$S(I_1, I_5) = \frac{1}{6} = 0.17$ (<0.3)

$\cancel{S(I_1 I_6) = \frac{2}{6}}$

$S(I_2 I_3) = \frac{1}{6} = 0.17$ (<0.3)

$S(I_2 I_5) = 0$ (<0.3)

$\cancel{S(I_2 I_6) = \frac{1}{6}}$

$S(I_3 I_5) = \frac{1}{6} = 0.17$ (<0.3)

$\cancel{S(I_3 I_6)}$

$\cancel{S(I_5 I_6)}$

4+3
+2+1

$I_1 I_2 I_3$

Frequent set $C_2 = \{I_1 I_2, I_1 I_3\}$

$\cancel{S(I_1 I_2 I_3)}$

as there are no more sets other than (I_1, I_2, I_3).

$S(I_1 I_2 I_3) = \frac{1}{6} = 0.17$ (<0.3)

so, generating rules from C_2,

$R_1 = I_1 \rightarrow I_2$

$R_2 = I_2 \rightarrow I_1$

$R_3 = I_1 \rightarrow I_3$

$R_4 = I_3 \rightarrow I_1$

$s(R_1) = 0.5 \qquad c(R_1) = \frac{s(I_1 I_2)}{s(I_1)} = \frac{0.5}{5/6}$

$= 0.6 \; (< 0.8)$

so, R_1 is not important Rule. ↳ less than min confidence

$s(R_2) = 0.5 \qquad c(R_2) = \frac{s(I_1, I_2)}{s(I_2)} = \frac{0.5}{0.5} = 1$

5

(> 0.8)

so, R_2 is important.

$s(R_3) = 0.33 \qquad c(R_3) = \frac{s(I_1, I_3)}{s(I_1)} = \frac{0.33}{5/6}$

$= 0.4$

(< 0.8)

so, R_3 is not important.

$s(R_4) = 0.33 \qquad c(R_4) = \frac{s(I, I_3)}{s(I_3)} = \frac{0.33}{3/6}$

$= 0.67$

(< 0.8)

so, R_4 is not important.

so, only important rule is $I_2 \rightarrow I_1$

or, Butter → Bread.

c) ~~Drawbacks of apriori~~

c) Apriori algorithm mainly considers frequency of item in a dataset to generate association rules, which in many cases, exclude the other major correlation factors which lead to exclusion of important rules.

6. a) confusion matrix:

	Predicted class yes	Predicted class NO
True class yes in dataset	True positive (TP) 1	False negative (FN) 1
True class NO in dataset	False positive (FP) 1	True negative (TN) 2

~~a.toz~~

i) Precision:

$$P = \frac{TP}{TP+FP} = \frac{1}{1+1} = \frac{1}{2} = 0.5$$

ii) Recall:

$$R = \frac{TP}{TP+FN} = \frac{1}{2} = 0.5$$

iii) F score:

4

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$$= \frac{2 \times \frac{1}{2} \times \frac{1}{2}}{\frac{1}{2} + \frac{1}{2}} = \frac{1}{2} = 0.5$$

6) i) Hold out method- In hold out method, a dataset is put on hold and tested with classifier multiple times ~~used~~ to test the goodness of the classifier.

2½

ii) Cross-validation- A dataset is divided in k folds. k-1 folds are used for training and kth fold for testing. This process is continued multiple times to estimate goodness.

iii) Bootstrap - A dataset is divided into n instances. for each classification, n samples with replacement are picked from dataset.

c) The main purposes of ensembles of classifiers →

i) Ensembling multiple classifiers increases the overall accuracy of the model.

ii) In case of large datasets, there may be multiple hypothesis applicable to multiple parts. But a single hypothesis may not cover them. In that case Ensembling helps.

iii) If multiple classifiers are ensembled, it helps reduce the overall error of classification.

Ada Boost algorithm

Given a dataset d and t classifiers.

Generation:

- Equal weight is assigned to all instances.
- If ~~the~~ after applying the classifier, e is error →

  If e is equal to 0 or e greater than equal to 0.5, exclude the classifier.

  Else, multiply the weights by e/1-e.

2½

- ~~Repeat~~ save the classification model.

Classification:

- Apply the weight 0 to all instances.
- Apply the classifier, add $\log\left(\frac{1-e}{e}\right)$ to the weights.
- Return the class label with maximum weight.