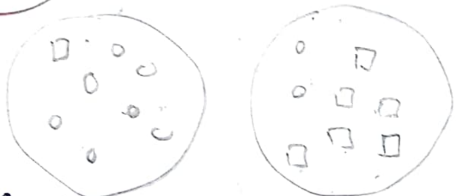
Q2.

a) i) Purity : simple measure of cluster validation checking for the percentage of a particular class in a certain cluster (which may have misclassified data).

2

$$\text{Purity} = \frac{1}{n} \cdot (\max(C_i))$$



$$= \frac{1}{n} (\max(6,2)).$$

c_i = circle.
$= \frac{1}{8} \cdot 7 = \frac{7}{8}$.

c_i = square
$= \frac{1}{8} \cdot 6 = \frac{3}{4}$

ii) Rand Index : refers to the number of correct ~~@~~ classifications made. In a confusion matrix,

Predicted \ Actual	C	¬C
C	A = 10	B = 5
¬C	C = 12	D = ~~20~~ 8

$$RI = \frac{A+D}{A+B+C+D} = \frac{10+8}{35}$$

$$= 0.514.$$

b) Intercluster distance : distance ~~between items ~~of~~~~ of different clusters.

* Types : i) single linkage distance : defined as the minimum distance between items of 2 ~~diff~~ different clusters.

consider 2 cluster S, T.

$$\delta_1 d = \min_{\substack{x_i \in S \\ x_j \in T}} \{d(x_i, x_j)\}$$

ii) complete linkage : defined as the maximum distance between any 2 points in the different clusters.

$$\delta_2 = \max_{\substack{x_i \in S \\ x_j \in T}} \{d(x_i, x_j)\}$$

iii) average linkage distance : average distance between all points in 2 different clusters

$$\delta_3 = \frac{1}{|S||T|} \cdot \sum_{\substack{x \in S \\ y \in T}} d(x, y).$$

Intracluster Distance: distance between 2 points (items) in the same cluster.

Types

consider cluster S.

i) Complete Diameter: furthest distance b/w 2 points in the same cluster.

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

ii) Average diameter: the average distance between all pairs of points in a given cluster.

$\Delta_2 = \frac{1}{|S|(|S|-1)} \cdot \sum_{x,y \in S} d(x,y)$

3

c) Dunn's cluster evaluation index.

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

Usefulness

~~Features~~

- based on maximising intercluster distance
- and minimizing intracluster distance
- larger value of Dunn's index implies well-clustered while lower values mean there will be misclassified data

2

Q3

a) **Support**: defined as the probability of having item B and item A together in all transactions.

$$sup = \frac{(A \cup B).count}{N}$$

Confidence: defined as the probability of having B in the transactions that also contain item A.

$$conf = \frac{(A \cup B).count}{A.count}$$

2

An itemset is referred to as a frequent itemset if all of its (K-1) subsets are also frequent itemsets.

An association rule is considered an important rule if it contains inferences about all items in the data set.

An association rule is important if $conf(A \rightarrow B) > minconf$.

b) ~~minsup = 30% = 0.3~~
~~minconf = 80% = 0.8~~

No 6.

b) minsup = 30% = 0.3

minconf = 80% = 0.8

Item	Freq.	Sup
Bread	5	0.8
Butter	3	0.5
Milk	3	0.5
Jelly	1	0.16 ← discarding, < minsup
Coke	2	0.33

F_2

Item	Freq	Sup
Bread, Butter	3	0.5
Bread, Milk	2	0.33
Bread, Coke	1	discarded
Butter, Milk	1	
Butter, Coke,	0	
Milk, Coke	1	

F_3 = {Bread, Butter, Milk}. all subsets exist in F_2

Frequent-itemset = {bread, butter, milk}.

Association Rules

$A \to B$

$\frac{(A \cup B).count}{A.count}$

		conf.	
bread →	butter, milk	1/5	20%.
butter →	bread, milk.	1/3	33.3%
milk →	bread, butter.	1/3	33.3%.
bread, butter →	milk	1/3	33.3%.
bread, milk →	butter	1/2	50%
butter, milk →	bread	1/1	100%.

butter, milk → bread is important association rule with conf. > min conf.

e) Major drawbacks of apriori algorithm

i) takes $O(2^m)$ space complexity (exponential complexity)

ii) generates arbitrarily large number of rules even for small dataset

2

iii) considers all items equally important whereas in real life there would be priority of items.

---

4) a). $Info(D) = -P_y \log_2(P_y) - P_n \log_2(P_n)$.

Eq: $N = 10$
$P_y = \frac{6}{10} = 0.6$
$P_N = 0.4$

$= -0.4 \log_2(0.4) - 0.6 \log_2(0.6)$

$= \frac{1}{10}\left(4(\log 4 - (\log 2 + \log 5)) - 6(\log_2 6 - (\log 2 + \log 5))\right)$ $2 + \log 3$

$= \frac{1}{10}\left(4(2 - (1 + 2.3)) - 6(1 + 1.6 - (1 + 2.3))\right)$

$= \frac{1}{10}\left(4(-1.3) - 6(0.7)\right) = 0.94.$

$Info_{wind}(D) = \frac{strong}{N} Info(D) + \frac{weak}{N} Info(D)$

wind → strong, weak

$= \frac{5}{10} \times 0.94 + \frac{5}{10} \times 0.94$

$= 0.94$

$Info_{humidity}(D) =$ Nor

$Info_{wind}(D) = \sum \frac{|D_j|}{|D|} info(D).$

$= \frac{strong + yes}{strong} Info(D) + \frac{strong - no}{strong} Info(D).$

$= \frac{2}{10} \times 0.94 + \frac{3}{10} \times 0.94 = 0.47.$

94 → a
done on last page after all answers. Sorry!

P.T.O

Q4 b) Terminating criteria of decision tree

i) If at some level no more splitting can be done based on attributes.

ii) If all classes at some level are the same

iii) If there is no more data in the dataset to train on.

2

c) Model overfitting can be caused by

i) Too many branches in decision tree can cause overfitting as it also handles outliers and noise.

ii)

The problem is solved ~~b~~ via pruning, which has 2 types

i) Pre pruning: actively trim branches during formation if the tree falls below a certain goodness factor.
- problem: hard to ~~determing~~ determine goodness factor.

ii) Post pruning: ~~prune prune b,~~ prune branches after decision tree formation by running through some sample data set.

2

P.T.O

Q6.
A.

Outlook	Humidity	Wind	Play Tennis	Actual
sunny	normal	strong	Y	Y
overcast	normal	strong	N.	Y.
Rain	high	strong	Y	N
Sunny	high	weak	N	N.
Rain	high	strong	N	N.

Actual ↓ \ Predicted →	P	¬P.
	1.	
P	1	1.
¬P	1	2.

i) Precision $= \frac{A}{A+B} = \frac{1}{2} = 0.5$

ii) Recall $= \frac{A}{A+C} = 0.5$.

iii) F-score $= \frac{2 \times P \times R}{P+R} = \frac{2 \times 0.5 \times 0.5}{1} = 0.5$

4

b. i) Holdout :

ii) ~~Cross~~. Cross Validation : In this method, the dataset is run through n iterations.
At the i^{th} Iteration D_i is the training data while remaining are test data.

0

iii) Bootstrap : dataset divided into 2 parts - 63.2% training data & 36.8% test data and accuracy becomes a combination of both. (.632 bootstrap).

c) Main purpose of ensemble classifiers
→ get different sets of errors for different classifiers so that error can be minimized when combined.
→ possible to get higher accuracy than other algorithms even if ~~all~~ base classifiers have low accuracy.

4

P.T.O.

Adaboost algorithm . Classifier generation

(1) initialize weight of all classes to $\frac{1}{d}$ (d = items in dataset).
(2) for $i = 1$ to K, do
(3) generate D_i from D.
(4) create M_i from D_i using learning scheme and store
(5) compute error$(M_i) = e$.
(6) if $e > 0.5$ or $e = 0$
(7) goto step (3) and recreate D_i
(8) endif
(9) for each tuple in D_i that is correctly classified :
(10) multiply the weight by $\frac{e}{1-e}$.
(11) endfor
(12) normalize weights (by multiplying with $\frac{\text{old weight}}{\text{new wt}}$)
(13) end for

Classification algorithm

(1) initialize wt. of class to 0.
(2) for $i = 1$ to K do.
(3) $w = \log\left(\frac{1-e}{e}\right)$.
(4) $c = M_i(x)$ predict class of x, using M_i
(5) add weight w to class c.
(6) endfor
(7) return class with highest weight for result.

The class returned is the predicted class.

(Q4)
(a). Info (D) $= 0.94 = \left(-\frac{4}{10}\log\frac{4}{10} - \frac{6}{10}\log\frac{6}{10}\right)$

2

$\text{Info}_w(D) = I(2,3) + I(4,1)$

$= -\left(\frac{2}{14}\log\left(\frac{2}{5}\right) + \frac{3}{14}\log\left(\frac{3}{5}\right) + \frac{4}{14}\log\left(\frac{4}{5}\right) + \frac{1}{14}\log\left(\frac{1}{5}\right)\right)$

$= \left(\frac{2}{14}\times 1.3 + \frac{3}{14}\times 0.7 + \frac{4}{14}\times 0.3 + \frac{1}{14}\times 2.3\right)$

$= 0.585$

$\text{Info}_h(D) = I(4,1) + I(2,3)$

$= 0.585$

$\text{Info}_o(D) = I(2,0) + I(2,2) + I(2,2)$

$= 0 + \left(\frac{2}{14}\times\log\left(\frac{2}{4}\right) + \frac{2}{14}\times\log\left(\frac{2}{4}\right)\right)\times 2.$

$= 0.571$

$\text{Gain}_o(D) = 0.94 - 0.571$

$= 0.368$

$\text{Gain}_w(D) = \text{Gain}_h(D) = 0.355.$

~~Since gain overc~~

Gain $_{\text{outlook}}$ (D) is largest,
so, outlook
will be root.

—— x ——