(2)

(a)

(i) <u>Purity</u> : Purity of a cluster determines the goodness of a cluster. It is the ratio of max objects of a class in the cluster to the size of the cluster.

$$Purity(\omega_i) = \frac{1}{n_i} \max_j \{n_{ij}\}$$

n_i = size of ith cluster

n_{ij} = size of jth class in ith cluster.

<u>Example</u>.

3

cluster 1 cluster 2

$purity(\omega_1) = \frac{4}{6}$ $purity(\omega_2) = \frac{4}{5}$.

(ii) <u>Rand Index</u> : It determines the percentage of correct classification out of the total number of classifications by the model.

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

where, TP = True Positive, TN = True Negative

FP = False Positive, FN = False Negative.

Confusion Matrix

		Predicted	
		same class	different class
Ground truth	same class	TP	FN
	different class	FP	TN

Intercluster distance (δ)

Intercluster distance is the measure of the distance between two closest clusters generated by any clustering algorithm.

1. **Single Linkage distance**

$$\delta(S,T) = \min \left\{ \begin{matrix} d(x,y) \\ x \in S,\, y \in T \end{matrix} \right\}$$

[$d(x,y)$ = distance between two objects x and y.]

It is the closest distance between two clusters S and T which is equal to the minimum distance between an object from S and another object from T.

2. **Complete linkage distance**

$$\delta(S,T) = \max \left\{ \begin{matrix} d(x,y) \\ x \in S,\, y \in T \end{matrix} \right\}$$

It is the farthest distance between any two objects in the ~~two~~ clusters S and T respectively.

* Since, a good clustering have large δ value and small Δ value the Dunn's index helps in determining the goodness of clustering using δ and Δ values.

Intra cluster distance (Δ)

Intra cluster distance is the measure of the distance between two objects within the same cluster.

1. Complete diameter distance

4

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

It is the max distance between two objects in a cluster S.

2. Average diameter distance

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

It is defined as the average of the distances between all pairs of objects within a cluster, S.

(c) Dunn's Index

2

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

A good cluster has large value of Dunn's index.

$\delta(X_i, X_j)$ = inter cluster distance between the clusters X_i and X_j

$\Delta(X_k)$ = intra cluster distance ~~between~~ of X_k.

*

(3)

(a) Support : It is defined as the ratio of the frequency of an itemset in a transaction database to the total number of transactions in the database.

$$\text{Support} = \frac{(X \cup Y).\text{count}}{n} \quad \left[(X \cup Y) = \text{itemset}, \; n = \text{no. of transactions} \right]$$

Confidence : It is defined as the ratio of the frequency of an itemset to the frequency of a proper non-empty subset of the itemset in the transaction database.

4

$$\text{Confidence} = \frac{(X \cup Y).\text{count}}{X.\text{count}} \quad [X \subseteq X \cup Y]$$

An itemset is referred to as a frequent itemset when the ~~frequency~~ support of the itemset $\geq$ minsup. (i.e, min support).

An association rule is referred as an important rule when the confidence of the association rule $\geq$ minconf (i.e, min confidence or threshold).

(b)

Transactions	Items
T1	Bread, Butter
T2	Bread, Milk, Butter
T3	Bread, Jelly, Butter
T4	Bread, Coke
T5	Bread, Milk
T6	Milk, Coke

Let, I_1 = Bread, I_2 = Butter, I_3 = Milk, I_4 = Jelly, I_5 = Coke

then,

Transactions	Items
T_1	I_1, I_2
T_2	I_1, I_2, I_3
T_3	I_1, I_2, I_4
T_4	I_1, I_5
T_5	I_1, I_3
T_6	I_3, I_5

minsup = 30%.

(s = support)

C_1 : $\{I_1\}$, $\{I_2\}$, $\{I_3\}$, $\{I_4\}$, $\{I_5\}$

$s = 5/6$ ✓ $s = 3/6$ ✓ $s = 3/6$ ✓ $s = 1/6$ ✗ $s = 2/6$ ✓

F_1 : $\{I_1\}$, $\{I_2\}$, $\{I_3\}$, $\{I_5\}$

C_2 : $\{I_1, I_2\}$ $s = 3/6$ ✓, $\{I_1, I_3\}$ $s = 2/6$ ✓, $\{I_1, I_5\}$ $s = 1/6$ ✗

$\{I_2, I_3\}$ $s = 1/6$ ✗, $\{I_2, I_5\}$ $s = 0/6$ ✗

$\{I_3, I_5\}$ $s = 1/6$ ✗

F_2 : $\{I_1, I_2\}$, $\{I_1, I_3\}$ ✓

C_3 : $\{I_1, I_2, I_3\}$

$s = 1/6$ ✗

F_3 : ϕ

$\therefore$ Frequent itemset $F = F_1 \cup F_2 \cup F_3$

$F = \{I_1\}, \{I_2\}, \{I_3\}, \{I_5\}, \{I_1, I_2\}, \{I_1, I_3\}$

Association rules

For $\{I_1\}$, $\{I_2\}$, $\{I_3\}$, $\{I_5\}$ itemsets there are no association rules as they are singleton sets.

For $\{I_1, I_2\}$, and $\{I_1, I_3\}$.

mincanf = 80%
c = confidence

A	→	B	
I_1	→	I_2	$c = 3/5 = 0.6$ ✗
I_2	→	I_1	$c = 3/3 = 1$ ✓
I_1	→	I_3	$c = 2/5 = 0.4$ ✗
I_3	→	I_1	$c = 2/3 = 0.66$ ✗

Important association rules :

$I_2 \rightarrow I_1$

Butter → Bread.

(c) Drawbacks of a-priori algorithm :

1. The space complexity for generating all association rules is exponential i.e. $O(2^m)$ where m is the number of items in the itemset I.

2. The algorithm exploits the sparseness of data, high minsup and high minconf values.

3. Single minsup value problem : The algorithm considers all itemsets of same nature and similar frequency greater than a single minsup value. In practical, this is not always true as the nature and frequency of items in dataset may vary.

2/2

4. The algorithm generates high number of association rules which are difficult to interpret.

(4)

(a) Attributes: Outlook, Humidity, Wind.

Output variable: PlayTennis.

The output variable has 2 class: Yes and No.

Let, p = Yes count = 6

n = No. count = 4 $\therefore$ p+n = 10.

$\therefore$ Information for the whole table,

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

$$= -\frac{6}{10}\log_2\left(\frac{6}{10}\right) - \frac{4}{10}\log_2\left(\frac{4}{10}\right)$$

$$= -0.6(\log_2 3 - \log_2 5) - 0.4(\log_2 2 - \log_2 5)$$

$$\boxed{I(p,n) = 0.94}$$

Now entropy for each attribute,

$$E(\text{Outlook}) = \sum_{i=1}^{3} \frac{p_i + n_i}{p+n} \cdot I(p_i, n_i)$$

Sunny	Overcast	Rain
p=2	p=2	p=2
n=2	n=0	n=2

$$= \frac{4}{10}\cdot I(2,2) + \frac{2}{10}\cdot I(2,0) + \frac{4}{10}\cdot I(2,2)$$

$$= \frac{4}{10} + 0 + \frac{4}{10} = 0.8$$

$[I(2,0) = 0,\ I(2,2) = 1]$

$$\boxed{E(\text{Outlook}) = 0.8}$$

$$E(\text{Humidity}) = \sum_{i=1}^{2} \frac{p_i + n_i}{p+n} \cdot I(p_i, n_i)$$

Normal	High
$p = 4$	$p = 2$
$n = 1$	$n = 3$

$$= \frac{5}{10} \cdot I(4,1) + \frac{5}{10} \cdot I(2,3)$$

$$= 0.5\left[-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}\right] + 0.5\left[-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right]$$

5

$$= 0.5(0.24 + 0.46) + 0.5(0.52 + 0.42)$$

$$= 0.82$$

$$\boxed{E(\text{Humidity}) = 0.82}$$

$$E(\text{Wind}) = \sum_{i=1}^{2} \frac{p_i + n_i}{p+n} \cdot I(p_i, n_i)$$

Strong	Weak
$p = 2$	$p = 4$
$n = 3$	$n = 1$

$$= \frac{5}{10} \cdot I(2,3) + \frac{5}{10} \cdot I(4,1)$$

$$= 0.82$$

$$\boxed{E(\text{Wind}) = 0.82}$$

Therefore, information gain for each attributes are,

$$G(\text{Outlook}) = I(p,n) - E(\text{Outlook}) = 0.94 - 0.8 = 0.14$$

$$G(\text{Humidity}) = I(p,n) - E(\text{Humidity}) = 0.94 - 0.82 = 0.12$$

$$G(\text{Wind}) = I(p,n) - E(\text{Wind}) = 0.94 - 0.82 = 0.12$$

Since attribute 'Outlook' has the maximum information gain among all attributes, the root of the decision tree is the 'Outlook' attribute.

(b) The decision tree algorithms terminates when all instances of the dataset are classified using the attributes and conditions (classes).

0

(c) The main cause of model overfitting is noise learning which results in high accuracy in the training data samples and low accuracy in the test data samples. In overfitting, the model train itself with outliers or noise which have very low correlation coefficient.

In decision tree, the overfitting problem can be solved by,

1. Prepruning (Early stop): The growth of the decision tree is stopped before the complition of the whole tree.

2. Postpruning: The tree is pruned after the decision tree is completely built.

(5)

(a) X : <Outlook = Overcast, Humidity = High, Wind = Weak>

According to Naive Bayes Classification algorithm,

$$P(C_i/X) = P(C_i) \cdot P(X/C_i)$$

where, $$P(X/C_i) = \prod_{j=1}^{k} P(x_j/C_i)$$

C_i = i^{th} class of output variable

x_j = j^{th} ~~class of~~ term in test sample.

k = number of terms in the test sample.

$P(C_1) = P(C = Yes) = \frac{6}{10} = 0.6$

$P(C_2) = P(C = No) = \frac{4}{10} = 0.4$

C = Yes	C = No
$P(\text{Outlook = Overcast}/C = Yes)$	$P(\text{Outlook = Overcast}/C = No)$
$= \frac{2}{6} = 0.33$	$= 0/4 = 0$
$P(\text{Humidity = High}/C = Yes)$	$P(\text{Humidity = High}/C = No)$
$= \frac{2}{6} = 0.33$	$= 3/4 = 0.75$
$P(\text{Wind = Weak}/C = Yes)$	$P(\text{Wind = Weak}/C = No)$
$= 4/6 = 0.66$	$= 1/4 = 0.25$

$\therefore P(X/C_1) = 0.33 \times 0.33 \times 0.66 = 0.0718$

$P(X/C_2) = 0 \times 0.75 \times 0.25 = 0$

4½

$P(C_1/x) = 0.6 \times 0.0718 = 0.043$

$P(C_2/x) = 0.4 \times 0 = 0$

$\therefore$ According to Naive Bayes algorithm, the predicted class of the sample X is Yes, as, $P(C_1/x)$ is greater, i.e. PlayTennis = Yes, for sample X.

(b) Yes, there is an error in the prediction of the sample class, as P(Outlook = Overcast / C = No) = 0.

The error can be corrected using Laplacian correction.

2½

1. We inject a tuple with 'Outlook' attribute set to Overcast and 'PlayTennis' variable set to No, This will make the corresponding probabilities non zero.

2. The new probabilities (corrected) will be,

P(Outlook = Overcast / C = No) = $\frac{1}{5}$ = 0.2

P(Humidity = High / C = No) = $3/5$ = 0.6

P(Wind = Weak / C = No) = $1/5$ = 0.2

3. $P(C_2/X) = P(C_2) \cdot P(X/C_2)$

5

$= 0.4 \times (0.2 \times 0.6 \times 0.2)$

$= 0.0096.$

(c) For any continuous value attribute, the algorithm will work after discretization of the attribute. The gini-index will help in finding the prediction of a particular sample only after the attribute is discretized.

Discretization can be done in different ways including quartile divisions and range classification methods.

The humidity attribute can be ~~clas~~. discretized and then the sample can be predicted using Naive Bayes algorithm.

1