Date :

Name: Deepak Kumar P Honakeri
Class/Branch: CSE - 6A
Reg No.: 1PE17CS044

Program No. / Experiment No.

Title : DM Assignment

Q1 what is a descision tree? Explain working of hunts algorithm

Ans A descision tree is a decision support tool and is a flow chart like tree structure, where each internal node denotes a test on a attribute, each branch represents a test on an attribute and an outcome of test

Hunt's Algorithm

let D_t be the set of training records that reach a node 't'

General Procedure.

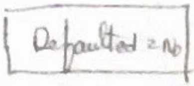
1. If D_t contains records that belong to the same class y_t, then t is a leaf node labelled as y_t

2. If D_t contains records that belong to more than one class, use an attribute test to split the data into smaller subsets Recursively apply procedure to each subset.

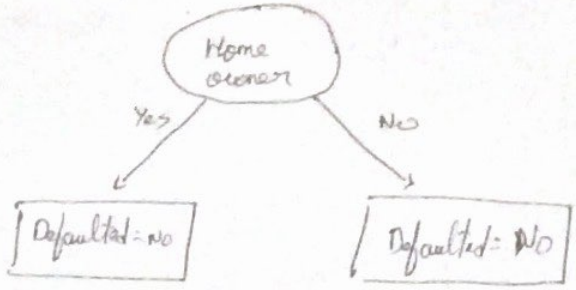
Marks

Staff Signature

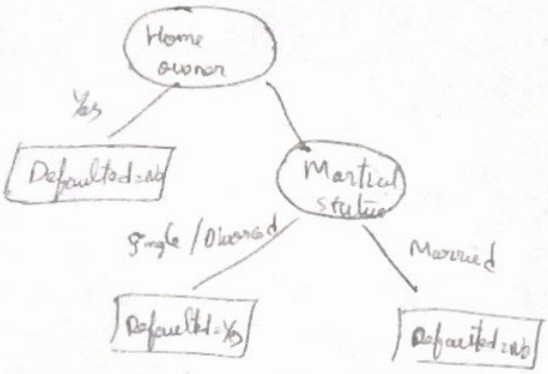
1PE17CS044



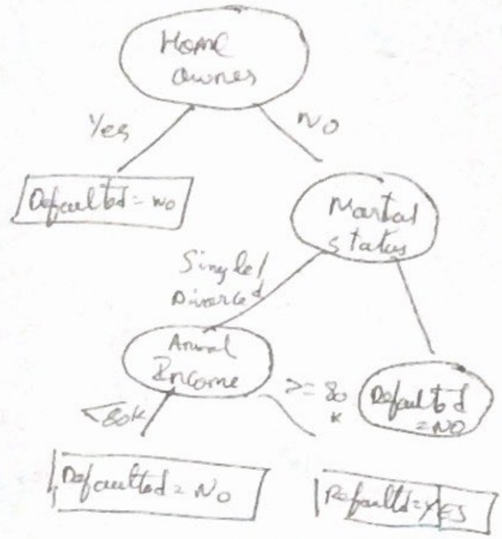
(a)



(b)



(c)



Q2 Ans

Case 1 Based on Home Owner

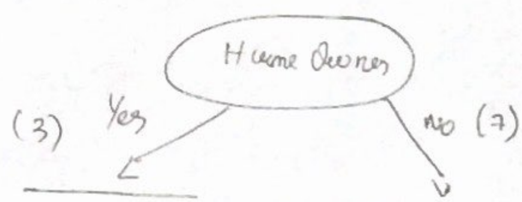
Defaults = $\frac{7}{10}$ for the class No

$= \frac{3}{10}$ for the class Yes

Gini Index $= 1 - \sum_{i=0}^{c-1} [p(i/t)]^2$

$= 1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2$

$= 0.42$



Yes (3)	No (7)
(+): 0	(+): 3
(−): 3	(−): 4
$= 1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2$	$= 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2$
$= 0$	$= 0.49$

∴ Weighted

Summed GI $= \frac{3}{10} \times 0 + \frac{7}{10} \times (0.49)$

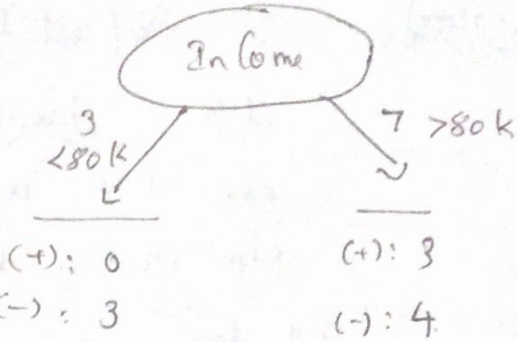
$\boxed{\text{Summed GI} = 0.343}$

1PE17S044

Case 2 : Annual Income

Defaults : $\frac{7}{10}$ Class No

$\frac{3}{10}$ Class Yes



∴ Weighted Summed

$$GI = \left(\frac{7}{10}\right)\times 0.49 + \frac{3}{10}(0)$$

$$= 1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2 = 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2$$

$$\boxed{= 0.343} \qquad = 0.0 \qquad = 0.49$$

03 Explain the decision tree induction algorithm with the characteristics in details.

Ans <u>Algorithm</u> : A skeleton descision tree induction algorithm

```
TreeGrowth (E, F)
    if  Stopping_Cond (E,F) = true then
        leaf = Create Node()
        leaf. label = Classify (E)
        return leaf
    else
        root = Create Node()
        root.test_cond = find_best_split (E,F)
        let V = {v | v is a possible outcome of root.test_cond}
        for each v ∈ V do
```

1PE17CS044

$E_v = \{ e \mid root.test\_cond(e) = v$ and $e \in E \}$

child = TreeGrowth (E_v, F)

add child as descendant of root and label the edge (root → child) as v

end for

end if

return root.

CreateNode ( ): function extends to the decision tree by creating a new node.

find_best_split( ): function determining which attribute should be selected as the test condition for splitting the training set.

Classify( ): function determines the class label to be assigned to a leaf node.

Characteristics of Decision Tree Induction

1. Decision tree induction is a non parametric approach for building classification models.

2. Finding an optimal decision tree is an NP-Complete problem. Many algorithms a heuristic-based approach to guide their search in the vast hypothesis space.

3. Techniques developed for constructing decision trees are computationally inexpensive, making it possible to quickly construct

And its worst case is $O(w)$, w is maximum depth of tree.

4. Decision trees, especially smaller - sized trees, are relatively easy to interpret.

5. Decision tree. algorithm are quite robust to presence of noise, especially when methods for avoiding overfitting.

Q4 Write a note on Baysean Network

Ans. A Bayesian belief network (BBN) or Bayesian Network provides a graphical representation of the probabilistic relationship among a set of random variables.

Two key elements of a Baysean Network:

1. A directed acyclic graph encoding the dependence relationships among a set of variables
2. A probability table associating each node to its immediate parent nodes.

Q5)

a) $P(B = good, F = empty, S = yes)$

Sol) Given

$P(G = empty \mid B = good, F = empty) = 0.8$

and

$P(S = Yes \mid B = good, F = empty) = 1 - P(S = No \mid B = Good, F = empty)$

$= 1 - 0.8$

$= 0.2$

IPE17CS044

10EI75044

$\therefore$ $P(B = good; F = empty, G = empty, S = yes) = 0.8 \times 0.2$

$= 0.16$

05 b)

$P(B = Bad, F = Empty, G = not\ empty, S = no)$

Sol

Given,

$P(G = notEmpty \mid B = bad, F = empty) = 1 - P(notempty \mid bad\ empty)$

$= 1 - 0.9$

$= 0.1$

$P(S = no \mid B = bad, F = empty) = 1.0$

$\therefore$ $P(B = bad, F = empty, G = notempty, S = no) = 1 \times 0.1 = 0.1$

05

c) Given $B = bad$,

To find, $P(Start = Yes)$

$\therefore$ $P(S = no \mid B = bad) = P(S = no \mid B = bad, F = notempty) \times P(F = not\ empty)$

$+ P(S = no \mid B = bad, F = empty) \times P(F = empty)$

$= 0.9 \times 0.8 + 1.0 \times 0.2$

$= 0.92$

$\therefore$ The probability that the car will start given battery is bad

$= 1 - P(S = no \mid B = bad)$

$= 1 - 0.92$

$= 0.08$

1PE17SO44

06 Explain Rule Based classifiers with examples

Ans • A rule based classifier is a technique for classifying records using a collection of "if.... then... " rules

• The rules for the model are represented in a disjunctive normal form $R = (r_1 \vee r_2 \vee \ldots r_k)$, where R is known as the rule set and r_i's are the classification rules or disjuncts.

Examples:

r1: (Gives Birth = no) ∧ (Aerial Creature = Yes) → Birds

r2: (Gives Birth = no) ∧ (Aquatic Creature = Yes) → Fishes

r3: (Gives Birth = yes) ∧ (Body Temprature = warm-Blooded) → Mammals

r4: (Gives Birth = no) ∧ (Aerial Creature = no) → Reptiles

r5: (Aquatic Creature = semi) → Amphibians.

Each classification rule can be expressed as

$$r_i : (\text{Condition } i) \rightarrow y_i$$

The LHS of rule is rule antecedent or preconditions

[1PS17SO44]

07 What are different characteristics of lazy learners. Explain any one in details with a practical application

Ans

Characteristics

- Lazy learners such as nearest neighbor classifiers do not require model building. However, classifying a test example can be quite expensive because we need to compute the proximity values individually between the test and training examples.

- Nearest-neighbor classifiers make their predictions on local information, eager learner like descision tree and rule based classifiers attempt to find a global model that fits the model space.

- nearest neighbour classifiers can produce arbitrarily shaped descision boundries, such boundries provide a more flexible model representation compared to descision tree. These boundaries of nearest neighbor classifiers also have high variability because they depend on composition of training examples.

- Nearest neighbour classifiers can produce wrong predictions unless appropriate proximity measure and data preprocessing steps are taken.

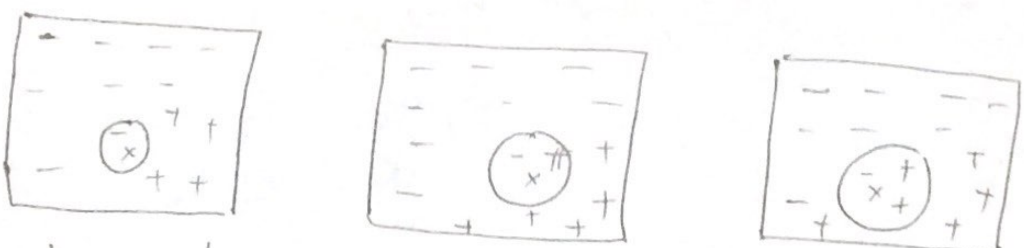
1PE17CS044

Example: Suppose we want to classify a group of people based on attributes such as height and weight. The height attribute has a low variability, ranging from 1.5 m to 1.85 m, whereas weight attribute may vary from 90 lb to 250 lb. If the scale of attributes are not taken into consideration, the proximity measure may be dominated by differences in the weights of a person.

Q8 What is KNN? Explain majority vote & weighted vote with example.

Ans The K-nearest neighbors of a given example x refer to the k point that are closest to x

Eg.



1 - nearest neighbor

2 - nearest neighbor

3 - nearest neighbor

K-NN of a record x are data point from ~~for~~ nearest neighbor list and compute distance between between two points.

$$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$

1PE175044

Once the nearest Neighbor list is obtained, test example is classified based on majority class of it's nearest neighbors.

$$\boxed{\text{Majority Voting}, \; y' = \underset{v}{\operatorname{argmax}} \sum_{(x_i, y_i) \in D_z} I(v = y_i)}$$

where v is class label

y_i is class label for one of the nearest neighbors

$I(\,)$ - indicator function that returns 1 if argument is true otherwise false.

Using distance - weighted voting scheme, the class label can be determined as

Distance - Weighted Voting,

$$\boxed{y' = \underset{v}{\operatorname{argmax}} \sum_{(x_i, y_i) \in D_z} w_i \times I(v = y_i)}$$

09

Record	A	B	C	Class
1	0	0	0	T
2	0	0	1	-
3	0	1	1	-
4	0	1	1	-
5	0	0	1	+
6	1	0	1	+
7	1	0	1	-
8	1	0	1	-
9	1	1	1	+
10	1	0	1	+

(1PE17CS044)

Estimate the Conditional Probabilities

$P(A|+)$ $P(A|-)$
$P(B|+)$ $P(B|-)$
$P(C|+)$ $P(C|-)$

Ans

$P(A) = 5/10$ $P(\bar{A}) = 5/10$

$P(B) = 3/10$ $P(\bar{B}) = 7/10$

$P(C) = 9/10$ $P(\bar{C}) = 1/10$

We know that,

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(A \cap B)}{P(B)}$$

$$P(A|+) = \frac{P(A,+)}{P(+)} = \frac{3/10}{5/10} = \frac{3}{5} = 0.6$$

$$P(B|+) = \frac{P(B,+)}{P(+)} = \frac{1/10}{5/10} = 0.2$$

$$P(C|+) = \frac{P(C,+)}{P(+)} = \frac{4/10}{5/10} = \frac{4}{5} = 0.8$$

$$P(A|-) = \frac{P(A,-)}{P(-)} = \frac{2/10}{5/10} = \frac{2}{5} = 0.4$$

$$P(B|-) = P(B,-) = \frac{2/10}{5/10} = \frac{2}{5} = 0.4$$

$$P(C|-) = \frac{P(C,-)}{P(-)} = \frac{5/10}{8/10} = \frac{5}{5} = 1.0$$

10) Explain rule Ordering Schemes in details. 1PE17CS044

Sol

Rule Ordering:

After generating the rule set, C4.5 rules uses the class-based ordering scheme to order the extracted Rules that predict the Same class are grouped together into the Same Subset. The total description length for each Subset is Computed and the classes are arranged in increasing order of their total description the class that has smallest description length is given the highest priority because it is expected to Contain the best set of rules

- The total description length for a class is given by

Lexception + g* Lmodel .

→ Lexception is no. of bits needed to encode the misclassified examples.

→ Lmodel is no. of bits needed to encode the model

→ g is tuning parameter with default = 0.5

The value of the Tuning parameter is Small if the model contains many redundant attributes.