Jessica Braganza
F17112181
BE Comp 2

ML Practical 2

Title : Decision Tree Classification

1) Explain following terminologies related to decision tree building

a) Impurity :

It defines how well each classes are seperated. In general, the impurity measure should satisfy the most when data are split evenly for attribute values.

$$P_i = \frac{1}{\text{No. of classes}}$$

Impurity should be 0 when all data belong to the same class.

b) Entropy :

The entropy of a random variable x is defined by

Entropy $(H(x)) = -\Sigma p(x) \log p(x)$

The entropy measures the expected uncertainty in x.

It has the following properties :

$H(x) \geq 0$, entropy is always non-negative

$H(x) = 0$, if and only if x is deterministic.

c) Information Gain :

The expected information needed to classify a tuple in P is given by,

$$\text{Info}(P) = -\sum_{i=1}^{n} P_i \log_2(P_i)$$

$$\text{Info}_A(P) = \sum_{i=1}^{n} \frac{|P_i|}{|P|} \times \text{Info}(P_i)$$

$$\text{Gain}(A) = \text{Info}(P) = \text{Info}_A(P)$$

2) What is Gini Index? Explain with formula.

Ans. 1) The Gini index is used in CART.

2) The Gini index measures the impurity of P, a data partition or set of training tuples as,

$$\text{Gini}(P) = 1 - \sum_{i=1}^{n} P_i^2$$

3) The attribute that maximizes the reduction in impurity (or equivalent) has the maximum Gini Index and selected as splitting attribute.

4) Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified.

3) Solve the problem for given dataset in problem statement to explain how to find root node using entropy and information gain.

Ans.

Step 1 : Finding the entropy :

$$\text{Entropy } (E(S)) = - \Sigma\, p(x) \log_2 p(x)$$

∴ For the dataset, entropy is,

$$\text{Entropy} = -P(\text{yes}) \log_2((P(\text{yes})) - P(\text{No}) \log_2 (P(\text{No}))$$

$$= -\left(\frac{9}{14}\right) \log_2 \left(\frac{9}{14}\right) - \left(\frac{5}{14}\right) \log_2 \left(\frac{5}{14}\right)$$

$$= 0.94$$

Step 2 : Finding Information Gain of each attribute

Age

< 21		21 - 35		> 35	
Yes	No	Yes	No	yes	No
2	2	3	1	4	2

$H(AGE<21) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0.971$

$H(age\ 21\text{-}35) = -\frac{4}{4}\log_2\left(\frac{4}{4}\right) - \frac{0}{4}\log_2\left(\frac{0}{4}\right) = 0$

$H(age > 35) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) = 0.971$

$\therefore$ Information gain (age) $= -\frac{5}{14} \times 0.971 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.971$

Info (age) $= 0.693$.

Gain (Age) = E(S) − Info (age)

$= 0.94 - 0.693$

$= 0.247$

Income

High		Low		Medium	
Yes	No	Yes	No	Yes	No
2	2	3	1	4	2

$H(INCOME\ HIGH) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1$

$H(INCOME\ LOW) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = 0.811$

$H(INCOME\ MED) = -\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) = 0.918$

$\therefore$ Information Gain (Income) $= \frac{4}{14} \times 1 + \frac{4}{14} \times 0.811 + \frac{6}{14} \times 0.918 = 0.911$

Gain (Income) = E(S) − Info (Income) = 0.0291

Gender

Male		Female	
Yes	No	Yes	No
3	4	6	1

DATE:

H (Gender MALE) $= -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right) = 0{\cdot}985$

H(Gender FEMALE) $= -\frac{6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right) =$ ~~0.1885~~ 0·592

Information gain (GENDER) $= \frac{7}{14} \times 0{\cdot}985 + \frac{7}{14} \times \overset{0\cdot592}{\cancel{0{\cdot}1885}} = 0{\cdot}7885$

Gain (Gender) $= 0{\cdot}94 - 0{\cdot}7885 = 0{\cdot}1515$

Marital Status			
Single		Married	
Yes	NO	Yes	NO
5	2	4	3

H (MS SINGLE) $= -\frac{5}{7}\log_2\left(\frac{5}{7}\right) - \frac{2}{7}\log_2\left(\frac{2}{7}\right) = 0{\cdot}863$

H (M.S. MARRIED) $= -\frac{4}{7}\log_2\left(\frac{4}{7}\right) - \frac{3}{7}\log_2\left(\frac{3}{7}\right) = 0{\cdot}985$

Information gain (M·S) $= \frac{7}{14} \times 0{\cdot}863 + \frac{7}{14} \times 0{\cdot}985 = 0{\cdot}924$

Gain (marital status) $= 0{\cdot}94 - 0{\cdot}924 = 0{\cdot}016$

Attributes	Gain
Age	0·247
Income	0·0291
Gender	0·1515
marital status	0·016

Since the attribute "Age" has highest gain, it is selected as root node.

Scanned with CamScanner