SNS COLLEGE OF TECHNOLOGY

Coimbatore-35
An Autonomous Institution
Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A++'
Grade Approved by AICTE, New Delhi & Affiliated to Anna University,
Chennai

DEPARTMENT OF COMPUTER APPLICATIONS

23CAT702 – MACHINE LEARNING

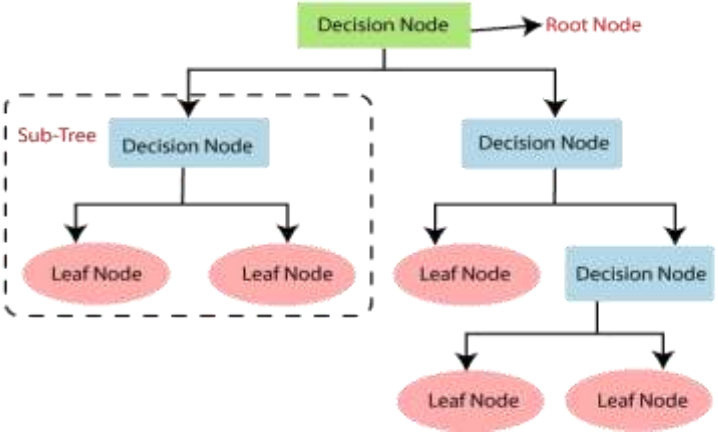
II YEAR III SEM

UNIT IV – TREE AND RULE MODELS

TOPIC 27 – Decision Tree

What’s Decision Tree?

- Decision Tree is a **Supervised learning technique** that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems.
- Collection of Node.

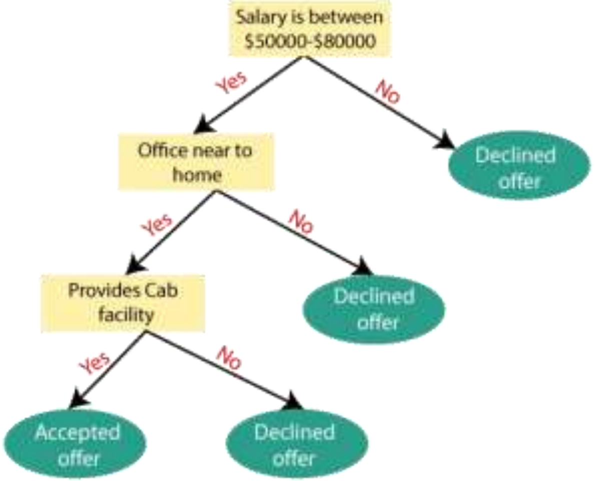


Why use Decision Tree?

- Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
- The logic behind the decision tree can be easily understood because it shows a tree-like structure.

Decision Tree Terminologies

- Root Node
- Leaf Node
- Splitting
- Branch/Sub Tree
- Pruning
- Parent/Child node



How does the Decision Tree algorithm Work?

- ✓ **Step-1:** Begin the tree with the root node, says S, which contains the complete dataset.
- ✓ **Step-2:** Find the best attribute in the dataset using Attribute Selection Measure (ASM).
- ✓ **Step-3:** Divide the S into subsets that contains possible values for the best attributes.
- ✓ **Step-4:** Generate the decision tree node, which contains the best attribute.
- ✓ **Step-5:** Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

Attribute Selection Measures

- The main issue arises that how to select the best attribute for the root node and for sub-nodes.
- So, to solve such problems there is a technique which is called as **Attribute selection measure or ASM.**
  - There are two popular techniques for ASM,
    - Information Gain
    - Gini Index

Information Gain:

- ❑ Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute.
- ❑ It calculates how much information a feature provides us about a class.

- Information Gain= Entropy(S)- [(Weighted Avg) *Entropy(each feature)

Entropy :

Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:

Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no)

Where,

S= Total number of samples

P(yes)= probability of yes

P(no)= probability of no

2. Gini Index:

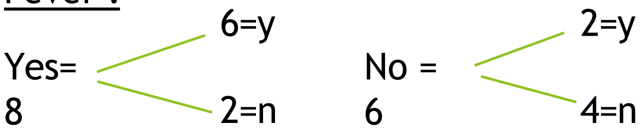
- ➢ Gini index is a measure of impurity or purity used while creating a decision tree in the CART(Classification and Regression Tree) algorithm.
- ➢ An attribute with the low Gini index should be preferred as compared to the high Gini index.

Gini Index= 1- ∑jPj2

Row= 14 , Y = 8 , N = 6

Σ (14) = -(8/14 * log2(8/14))-(6/14)*log2(6/14)
=0.985

Fever :



Yes= 8

6=y

2=n

No = 6

2=y

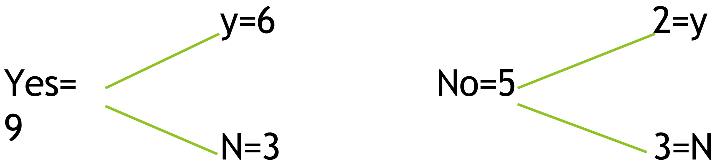
4=n

|S|=14 v = yes |Sv| = 8

Σ(Sv) = -(6/8)*log2(6/8)-(2/8)*log2(2/8) = 0.811 Σ(S) = -(2/6)*log2(2/6)-(4/6)*log2(4/6) = 0.918

IG(S,A)= ((S)-(|Sy|/|S|)* Σ(Sv))
IG (s,f) = 0.99-(8/14)*0.81-(6/14)*0.91 = 0.137

Cough :


Yes= 9

y=6

N=3

No=5

2=y

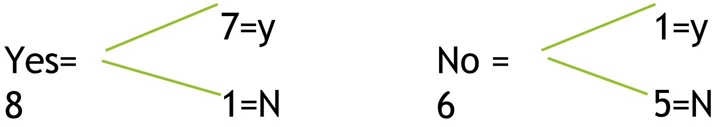
3=N

$\Sigma(Sv) = -(6/9)*\mathrm{log2}(6/9)-(3/9)*\mathrm{log2}(3/9)=0.91$

$\Sigma(Sv) =-(2/5)*\mathrm{log2}(2/5)-(3/5)*\mathrm{log2}(3/5) = 0.97$

$IG(s,c)=0.99-(8/14)*0.98-(6/14)*0.97 = 0.049$

Breathing:


Yes= 8

7=y

1=N

No = 6

1=y

5=N

$\Sigma(Sv) = -(7/8)*\mathrm{log2}(7/8)-(1/8)*\mathrm{log2}(1/8) = 0.543$

$\Sigma(Sv) = -(1/6)*\mathrm{log2}(1/6)-(5/6)*\mathrm{log2}(5/6) = 0.650$

$IG(S.B) = 0.99-(8/14)*0.543-(6/14)*0.650 = 0.401$