|  |  |
| --- | --- |
| Faculty of Computer & Information Sciences  Ain Shams University  Subject: (SCC 234)  Statistical Analysis and Applications  Year: (2nd year) undergraduate  Academic year: 2nd term 2019-2020 |  |

**Research Topic (C)**

**Title: Using Probability to Build Decision Trees for Classification**

**What ‘s a Decision Tree ?** (Ref. 1)

A decision tree is a diagram or chart that people use to determine a course of action or show a statistical probability. It forms the outline of the namesake woody plant, usually upright but sometimes lying on its side. Each branch of the decision tree represents a possible decision, outcome, or reaction. The farthest branches on the tree represent the end results.

Individuals use decision trees to clarify and find an answer to a complex problem. Decision trees are frequently employed in determining a course of action in finance, investing, or business.

The Basics of a Decision Tree :

A decision tree is a graphical depiction of a decision and every potential outcome or result of making that decision. Individuals deploy decision trees in a variety of situations, from something simple and personal ("Should I go out for dinner?") to more complex industrial, scientific or microeconomic undertakings.

By displaying a sequence of steps, decision trees give people an effective and easy way to visualize and understand the potential options of a decision and its range of possible outcomes. The decision tree also helps people identify every potential option and weigh each course of action against the risks and rewards each option can yield.

An organization may deploy decision trees as a kind of decision support system. The structured model allows the reader of the chart to see how and why one choice may lead to the next, with the use of the branches indicating mutually exclusive options. The structure allows users to take a problem with multiple possible solutions and to display those solutions in a simple, easy-to-understand format that also shows the relationship between different events or decisions.

In the decision tree, each end result has an assigned risk and reward weight or number. If a person uses a decision tree to make a decision, they look at each final outcome and assess the benefits and drawbacks. The tree itself can span as long or as short as needed in order to come to a proper conclusion.

**Mathematical formulation of Decision Trees.** (Ref. 2)

**Information gain** IG(A) is the measure of the difference in entropy from before to after the set S is split on an attribute A. In other words, how much uncertainty in S was reduced after splitting set S on attribute A.

IG(A,S) = H(S) -

Where,

* H(S) – Entropy of set S
* T – The subsets created from splitting set S by attribute A.
* P(t) – The proportion of the number of elements in t to the number of elements in set S
* H(t) – Entropy of subset t.

In ID3, information gain can be calculated (instead of entropy) for each remaining attribute. The attribute with the largest information gain is used to split the set S o()n this iteration.

**Entropy** H(S) is a measure of the amount of uncertainty in the (data) set S (i.e. entropy characterizes the (data) set S).

H(S) =

Where,

* S – is the current (data) set for which entropy is being calculated (changes every iteration of the ID3 algorithm)
* c – Set of classes in S c = { yes, no}
* p(c) – The proportion of the number of elements in class c to the number of elements in set S.

When H(S) = 0, the set S is classified (i.e. all elements in S are of the same class ).

In ID3, entropy is calculated for each remaining attribute. The attribute with the smallest entropy is used to split the set S on this iteration.

The higher the entropy, the higher the potential to improve the classification here.

**What are the terms in the Decision Trees mean and the intuitions behind them?**

Important Terminology related to Decision Trees

* Root Node: It represents the entire population or sample and this further gets divided into two or more homogeneous sets.
* Splitting: It is a process of dividing a node into two or more sub-nodes.
* Decision Node: When a sub-node splits into further sub-nodes, then it is called the decision node.
* Leaf / Terminal Node: Nodes do not split is called Leaf or Terminal node.
* Pruning: When we remove sub-nodes of a decision node, this process is called pruning. You can say the opposite process of splitting.
* Branch / Sub-Tree: A subsection of the entire tree is called branch or sub-tree.
* Parent and Child Node: A node, which is divided into sub-nodes is called a parent node of sub-nodes whereas sub-nodes are the child of a parent node.

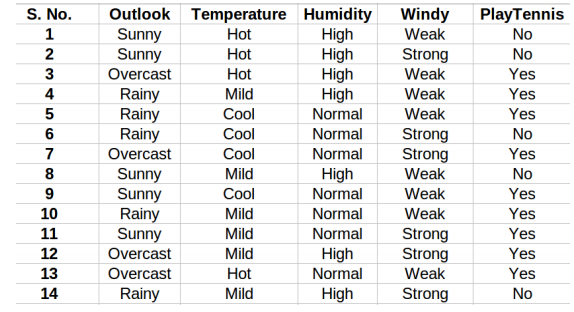
A decision tree is mainly made up of 6 different nodes (components), which are:

* Start Page: an optional introduction page
* Questions & Answers: ask your leads questions
* Decision: process the answers chosen by your leads
* Forms: to display legal disclaimers, terms & conditions, etc. and collect consent
* ESPs: email service providers such as Aweber, MailChimp and etc.
* Result Pages: display the results based on your leads answers.

**Explain the ID3 method to construct the decision tress how it is related to probability?**

**Working numerical examples of how Decision Trees are built.**

**Ex ;** Make a decision tree that predicts whether tennis will be played on the day?



**Solution in detail:**

**Step 1 : We’ll create a Root Node**

* How to choose the root node?

The attribute that best classifies the training data, use this

attribute at the root of the tree.

* How to choose the best attribute?

So from here, ID3 algorithm begins

- Calculate **Entropy** (The amount of uncertainty in dataset):

Entropy =