DECISION TREE LEARNING.

is one of the most widely used methods for inductive inference. It is a method for approximating discrete valued functions from a given set of training data.

* It works well for noisy data
* It is capable of learning disjunctive expressions.

Definition:-

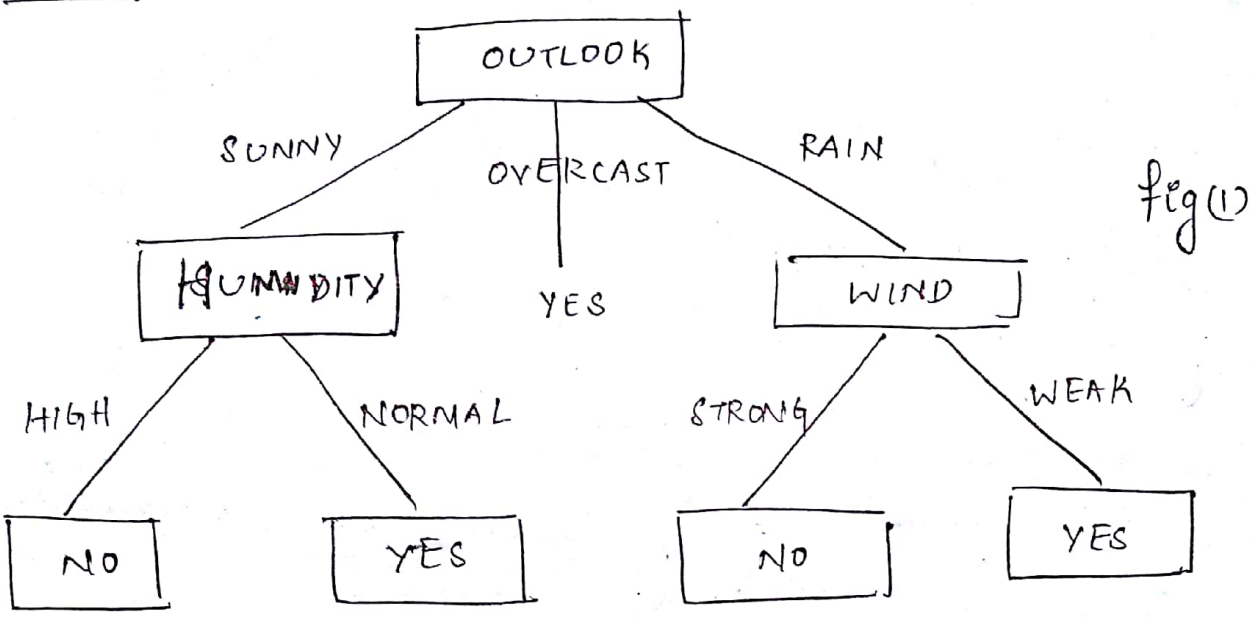
DT learning is a ML task which learns or infers discrete valued target function which is represented as a decision tree.

Decision tree representation:-

Decision trees classify instances by sorting them down the tree from the root to leaf node which provides the classification of the instance.

* Each node in the tree specifies a test of some attribute
* Each branch descending from the node is one of the possible values for this attribute.
* An instance is classified by starting from the root node down to the leaf checking the tests on the nodes.

Example:- Decision Tree for Concept: PLAY TENNIS.



fig(1)

The above tree represents a learned decision tree that can be represented as a

disjunction of conjunction of constraints on the attribute values of the instances.

(ie) (Outlook = Sunny ∧ Humidity = Normal)

∨ (Outlook = Overcast)

∨ (Outlook = Rain ∧ Wind = Weak)

How does Decision tree learning approach Concept?

Core Algorithm

Employs: 1) A top down

2) Greedy

search through the space of possible Decision Trees.

Variations of Decision Tree learning Algorithms

(1) ID3 (2) C4.5.

ID3 is the basic decision tree algorithm. All other decision tree algorithms are variations of ID3.

Decision tree learning using ID3.

* ID3 learns decision trees in a top down fashion.
* It begins by selecting the best attribute to be tested at the root level.
* Each instance attribute goes through a statistical test which determines how well each one classifies the training examples.
* The best one is selected for testing at the root node level.
* A descendant is created for the root node for each possible value of the root node.
* The entire process is repeated for all the other attributes using the training samples.
* This is a greedy construction of the DT where there is no backtracking.

Attribute Selection

* How to select the best attribute for testing at each level?
* How to measure the worth of an attribute?
* ID3 uses a quantitative measure called Information Gain that finds how well an attribute classifies the training examples.

Entropy and Information Gain

Entropy is a measure of disorder in the dataset. (ie) an indicator of how messy a dataset is. Entropy is used in Decision Trees to seperate data and group samples together into the classes they belong to. The aim is to maximise the purity of the groups (ie) homogenity of the group.

It characterises the impurity of a collection of samples. Given a collection S containing positive and negative examples of some target concept, the entropy of S relative to this boolean classification is

$$\text{Entropy}(S) \equiv -P_{\oplus} \log_2 P_{\oplus} - P_{\ominus} \log_2 P_{\ominus}$$

PTO

where

$P_{\oplus}$ is the proportion of +ve samples in S

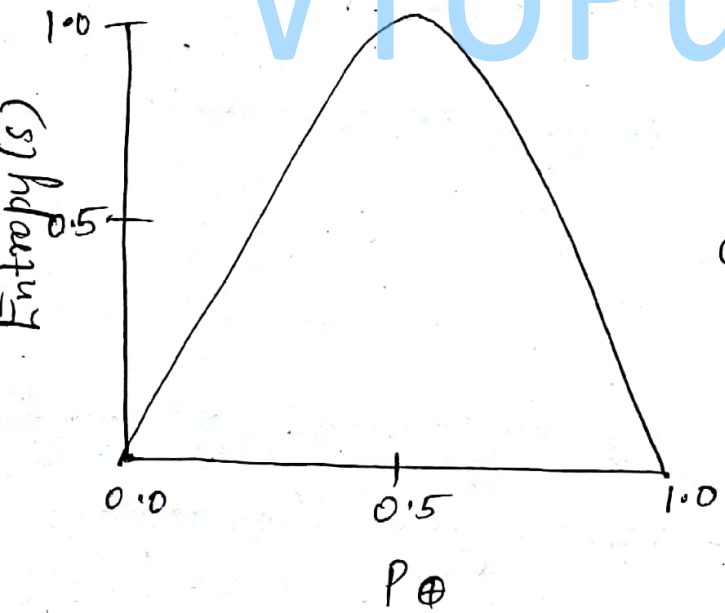
$P_{\ominus}$ is the proportion of -ve samples in S.

Entropy lies between 0 and 1.

Entropy is 0 if all members of S belong to the same class.

Entropy is 1 if the collection S has equal no of +ve and -ve training examples.

Entropy is between 0 and 1 if the collection has unequal no of +ve and -ve examples.



If the target attribute has c different values, entropy of S for c-wise classification is

$$\text{Entropy}(S) \equiv \sum_{i=1}^{c} -P_i \log_2 P_i$$

where P_i is the proportion of i belonging to S.

The logarithm is base 2, since it is a measure of the expected encoding length measured in bits.

Information Gain

Given entropy as a measure of impurity in a collection of training samples, a measure used for the effectiveness of an attribute in classifying the training data is called Information Gain.

Info Gain is the expected reduction in entropy caused by partitioning data according to an attribute.

Information Gain of an attribute A relative to S.

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

where Values(A) is the set of all possible values for the attribute A, S_v is the subset of S for which the attribute A has the value v.

In other words, Gain(S, A) is the information provided about the target function value, given the value of some other attribute A.

Capabilities and limitations of ID3.

The hypothesis space searched by ID3 is the set of all possible decision trees. It uses the information gain measure to evaluate the best attribute at each level that should be tested to construct the decision tree.

(1) ID3's hypothesis space of all decision trees is a complete space of finite discrete valued functions.

Searches complete hypothesis space -

Adv Since every hypothesis is a discrete valued function, there is no risk of searching for a hypothesis space that does not contain the target concept.

(2) ID3 maintains only a single current hypothesis as it searches through the hypothesis space, unlike CE which maintains the subset of all hypothesis consistent with the training examples.

Disadv Hence it does not have the ability to consider alternate decision trees.

(3) ID3 does not allow backtracking. Hence there is a possibility that it may converge to a locally optimal solution that is not globally optimal. The solution found may be less desirable than other trees along a different path of search.

Disadv

(4) ID3 uses all training examples at each step to make or construct the decision tree so that it refines the current hypothesis. In contrast CE or find-S incrementally constructs the hypothesis as and when a training sample is encountered.

Adv

ID3 uses statistical properties of all training examples and hence is less prone to errors.

Summary

Adv. (1) ID3 searches complete hypothesis space

Disadv (2) Maintains only a single current hypothesis.

Disadv (3) No backtracking in its pure form

Adv (4) Considers all training examples and its statistical properties in each step of construction.

Inductive Bias in Decision Tree Learning

Every ML algorithm with ability to generalise beyond the training data that it sees, has some type of inductive bias. This are the assumptions made by the model to learn the target function and generalise beyond the training data.

Inductive bias of a learner is the set of assumptions a learner uses to predict results, given inputs it has not yet encountered.

Given a collection of training examples, there are typically many decision trees consistent with the training examples. How does ID3 choose one among these trees?

ID3's search strategy chooses a tree based on the following:

Inductive Bias of ID3

(a) selects in favour of shorter trees over longer ones

(b) selects trees that place attributes with highest information gain closer to the root.

The above 2 constitute the inductive bias of ID3.

Preference Bias Vs Restriction Bias

Preference Bias / Search bias :- in learning has a preference for certain hypothesis over others. with no hard restriction on the hypothesis space.

Ex:- shorter hypothesis over longer ones.

* Preference bias is preferred over restriction bias.
* It allows the learner to learn from a complete hypothesis space that is assured to contain the unknown target function.

Ex:- ID3 has a preference bias in its learning since it searches incompletely through a complete hypothesis space.

* It ends up with a hypothesis that follows from its search strategy
* Searches from simple to complex hypotheses until it finds one consistent with the training data.
* The hypothesis space does not introduce any additional bias in ID3.
* The bias is due to the search strategy.

Restriction bias / Language Bias in learning has Restriction on the type of hypothesis to be learned (ie) limits the set of hypothesis to be learned/expressed.

* The restriction bias strictly limits the set of potential hypotheses considered and hence has the possibility of excluding the target function altogether.

Example:- Candidate Elimination Algorithm exhibits restriction bias or language bias

* CE algorithm searches an incomplete hypothesis space completely finding every hypothesis consistent with the ~~features~~ training data.

* The bias introduced is due to the hypothesis representation and not due to the search strategy.

Why does ID3 prefer shorter hypotheses?

The justification of the inductive bias of ID3 is based on the famous theory in Philosophy

Occam's Razor: Prefer the simplest hypothesis that fits the data.

Issues in Decision Tree Learning.

1. Avoid overfitting data.
   * Reduced Error Pruning
   * Rule Post Pruning.
2. Incorporating continuoused valued attributes.
3. Alternative measures for selecting attributes.
4. Handling Training data with missing values.
5. Handling attributes with differing costs.

~ ~

1. Avoid overfitting data

Decision tree learning grows each branch of the tree just deeply enough to perfectly classify the training samples.

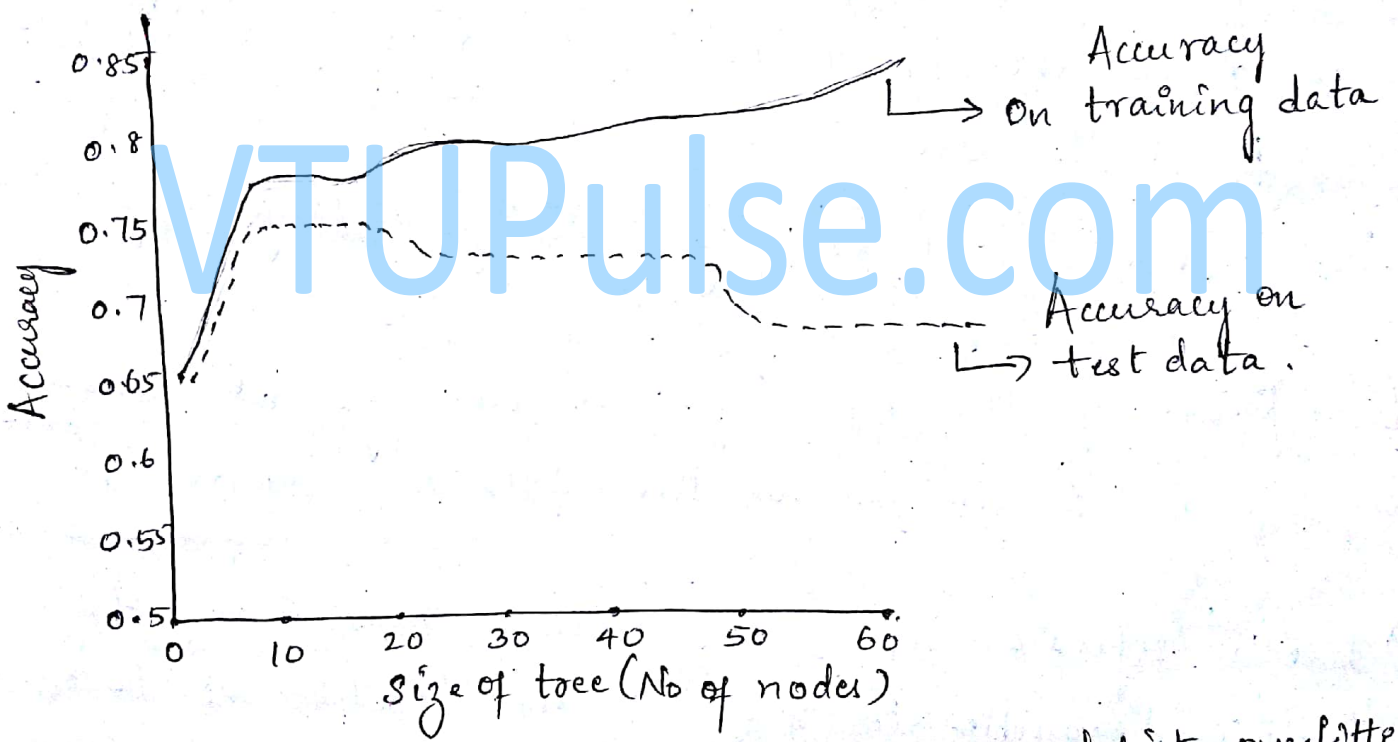
* But such a learning technique can result in overfitting of data when
  * There is noise in the data
  * The training samples is too small to be representative of the target function. (Chumber)

What is overfitting of data?

Definition:- Given a hypothesis space H, a hypothesis $h \in H$, is said to overfit the training data if there exists some alternate hypothesis $h' \in H$, such that h has smaller error than h' over the training data, but h' has a smaller error than h over the entire distribution of instances

↓

includes instances beyond training data.



The above graph is an example that depicts overfitting of a diabetic dataset. The solid line shows monotonic increase of accuracy as no of nodes are increased in the decision tree when the data used is only the given training data. (ie) Accuracy of Prediction over the training samples

The dotted line shows the accuracy of Prediction over the test samples. Although accuracy increases as far as 25 nodes, it decreases the accuracy when no of nodes are further increased.

So how to avoid overfitting of data?

Two different Approaches

(1) stop growing the tree earlier, before it reaches a point where it perfectly classifies the training data.

(2) Allow the tree to overfit the data, and then post prune the tree.

(ie) Find correct tree size ← by stopping early
                          ← Post Pruning

Irrespective of the approaches for pruning, the correct tree size is determined by the following approaches:-

(1) Use a seperate set of examples distinct from the training examples to evaluate the utility of post pruning nodes from the tree. ⇒ Training & Validation

(2) Use statistical test like Chi-square test to estimate whether expanding or pruning a particular node will produce an improvement beyond the training data.

Rule Post-Pruning :-

is a technique of pruning the decision tree learnt for improving accuracy over unseen instances.

It involves the following steps :-

(1) Infer the decision tree from the training set, growing the tree until the training data is fit and allow overfitting to occur.

(2) Convert the learned tree into an equivalent set of rules by creating one rule for each path from the root node to the leaf node.

(3) Generalise each rule by removing preconditions - this will improve accuracy.

(4) Sort the pruned rules by their estimated accuracy and use them in the sorted sequence for classifying subsequent instances.

Ex:- In rule post-pruning, one rule is generated for each leaf node in the tree.

(ie) For the tree in fig (1) [Decision tree for Play Tennis]

antecedants.

IF (Outlook = Sunny) ∧ (Humidity = High)

THEN PlayTennis = No.

is a rule.

* This rule is pruned by removing any antecedant (pre-condition), ~~that~~ whose removal does not worsen accuracy
* The accuracy of rules is determined by testing them with a validation set of examples different from the training set.

Advantages of converting decision tree to Rules before Pruning