Machine Learning

Tom M. Mitchell, McGraw-Hill, 1997

# Chapter 3: Decision Tree Learning

Decision Tree learning is one of the most practical and widely used methods for inductive learning. It is a method for approximating discrete-valued functions that is robust to noisy data and capable of learning disjunctive expressions. This chapter describes a family of decision tree learning algorithms that widely used algorithms such as ID3, ASSSITANT, and C4.5. These decision tree learning methods search a completely exhaustive hypothesis space. Their inductive bias is a preference for small trees over large trees.

## Introduction

Decision Trees approximate discrete valued functions and represent the target function as a decision. This is equivalent to a disjunction of conjunctions each representing a set of constraints on particular attribute values.

## Appropriate Problems for Decision Tree Learning

* Instances are represented by attribute-value pairs
* The target function has discrete output values
* Disjunctive descriptions may be required
* Training data may contain errors
* Training data may contain missing attributes

## The Decision Tree Algorithm

The central choice in the construction of the decision tree is which attribute to test next. Decision trees use a statistical property call *information gain* that measures how well a given attribute separates the training instances.

Entropy measures the homogeneity of examples.

***Entropy(S)*** = g(pi)

Information gain is the expected reduction in entropy caused by partitioning the examples according to a particular attribute.

***Gain(S,A)*** = Entropy(S) -

## Hypothesis Space Search in Decision Trees

ID3 searches a complete space of hypotheses consisting of all decision trees for the a given set of attributes and their values. ID3 incompletely searches this space terminating as soon as tree is found that satisfies the training examples.

Because of this is susceptible to standard hill-climbing issues of local minima and maxima.

In contrast, the Candidate-Elimination algorithm completely searches an incomplete space of hypotheses.

* ID3’s space of hypotheses is a complete space of all discreet-valued functions relative to the available attributes.
* ID3’s maintains a single current hypothesis
* ID3 in it’s pure form performs no backtracking and is susceptible to the perils of hill-climbing.
* ID3 uses all training examples at each step in the search. This can help enable noise-tolerance.

## Inductive Bias in Decision Tree Learning

Inductive bias is the set of assumptions that with the training data deductively justify the classification of future training instances. There are potentially many decision that could accommodate the training data.

The inductive bias of decision trees is to on the one hand prefer smaller trees and on the hand prefer attributes with higher gain at higher level nodes. Note: This is in agreement with Occam ’s razor.

A restriction bias is a bias on the hypotheses represented in the search space.

A preference or search bias is a bias on the search strategy.

Typically a preference bias is preferable.

## Issues in Decision Tree Learning

### Overfitting

A hypothesis is said to overfit the training data if there exists another hypothesis that performs better over all the data, beyond the training instances.

Two approaches to Overfitting:

1. Stop growing the tree before it perfectly classifies the training data
2. Post-pruning the tree

The second approach has been found to be more useful in practice given the difficulty in estimating when to stop growing the tree.

Methods for determining correct tree-size in both approaches

1. Use a spate set of examples to evaluate the utility of post-pruning
2. Use a statistical test to see whether expanding/pruning will to produce an improvement beyond the training data.
3. Use an explicit measure of the complexity of encoding the training examples. Minimum Description Length principle.

### Reduced Error Pruning

Try replacing each node and subtree rooted at that node with a leaf node and consider the effect on the additional training data. If the tree performs no worse then, prune that node.

### Rule Post-Pruning

1. Build Decision Tree
2. Convert Tree to Rules
3. Prune/Generalize rules by removing terms that improve overall performance
4. Sort rules by estimated accuracy

### Continuous-valued Attributes

Map continuous-valued attribute to a discrete set of intervals. Target attribute however must remain discrete-valued.

### Alternative Measures for Selecting Attributes

### Missing Attribute Values

### Attributes with Differing Costs

## Chapter 4: Artificial Neural Networks (ANN)

Artificial Neural Networks provide a general practical method for learning real-valued, discrete-values and vector-valued functions from examples. Algorithms such as BACKPROPAGATION use gradient descent ti tune network parameters to best fit a training set of input-output pairs. ANN learning is robust to errors in the training data and has been successfully applied to problems such as interpreting visual scenes, speech recognition, and learning robot control strategies.

## Biological Motivation

Human brain is estimated to have about 1011 neurons each connected to 104 other neurons.

## Neural Network Representations

## Appropriate Problems for Neural Network Learning

## Perceptrons

A perceptron takes a vector of real-valued inputs, computes a linear combination of these inputs and then outputs a 1 if the result is greater and a -1 otherwise.

### Representation Power

A perceptron represents a *hyperplane decision surface* in an n-dimension space instances that can be used to separate *linear separable* instances.

A single perceptron can represent many Boolean functions such as AND, OR and NOT, but not XOR because it is not linearly separable.

### Perceptron Training Rule

### Gradient descent and the Delta Rule