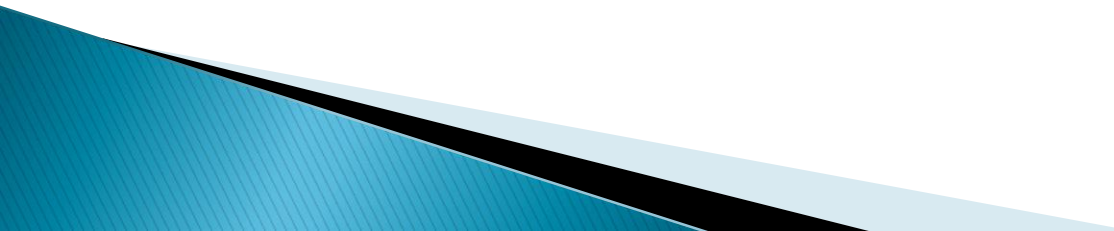


Elan Hartmann
Data Mining 3220
10.22.2012

Supervised Learning

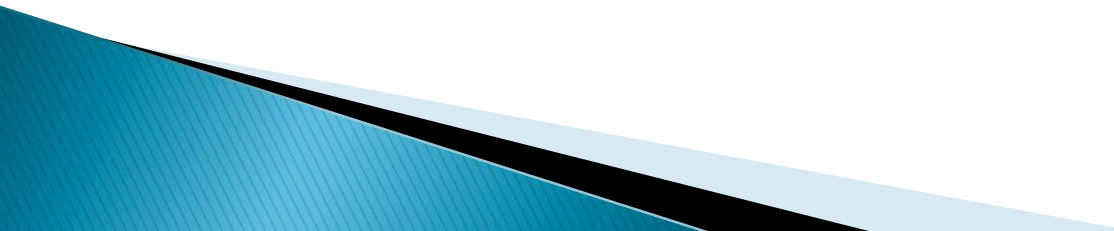
Outline:

- ▶ An overview of supervised learning
 - ▶ The Tasks for which it is used
 - ▶ As compared to unsupervised learning
 - ▶ A detailed look at the process
 - ▶ A list of the algorithms that are examples of this type of learning
 - ▶ An example using a decision tree
- 

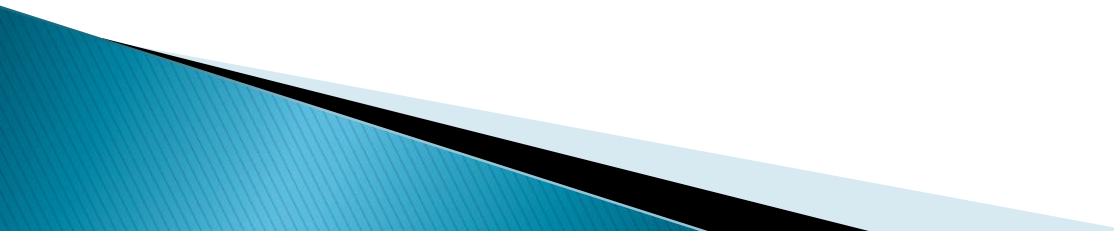
Supervised Learning: An Overview

A type of machine learning in which an algorithm infers a function from *training data* (sometimes called *training examples*). Following the completion of training, *test data* is input into the function to test how well the machine has learned. Both the training and test data have known classes or output values.

Tasks

- ▶ In *regression*, the output for a given input datum is in the form of a real number value.
 - ▶ In *classification*, the output for a given input datum is a class.
- 

Compared to Unsupervised Learning

- ▶ In *supervised learning* the classes or output values of the input must be established.
 - ▶ *Unsupervised learning* seeks to find hidden structure in an unlabeled data set.
- 

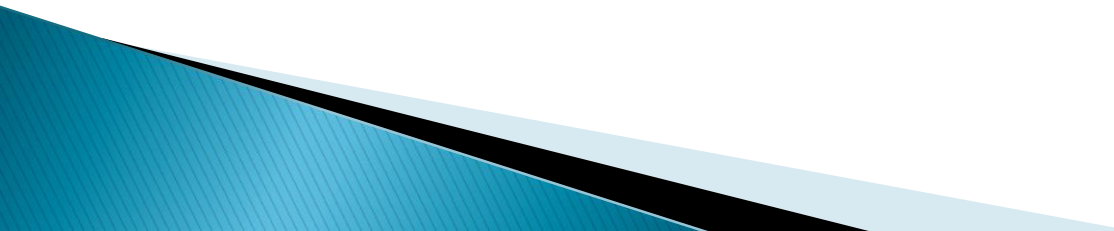
A Detailed look

- ▶ Start with a pool of ‘known’ data (which we’ll call x_p). Now split that into training data (which we’ll call x_j) and test data (which we’ll call x_t).
- ▶ Each instance has a known real number (regression) or class (classification) associated with it (which we’ll call y_p).

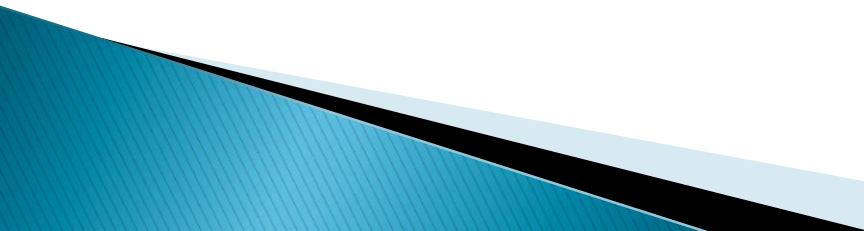
(Continued)

- ▶ Inputting x_i into the algorithm, a function, $f(x)$ is inferred in which for all x_i , $f(x_i) = y_i$.
 - ▶ x_t is then input into $f(x)$.
- ▶ The percentage accuracy of the function is:
 - $\{\#(f(x_t) = y_t) / \# x_t\} * 100\%$

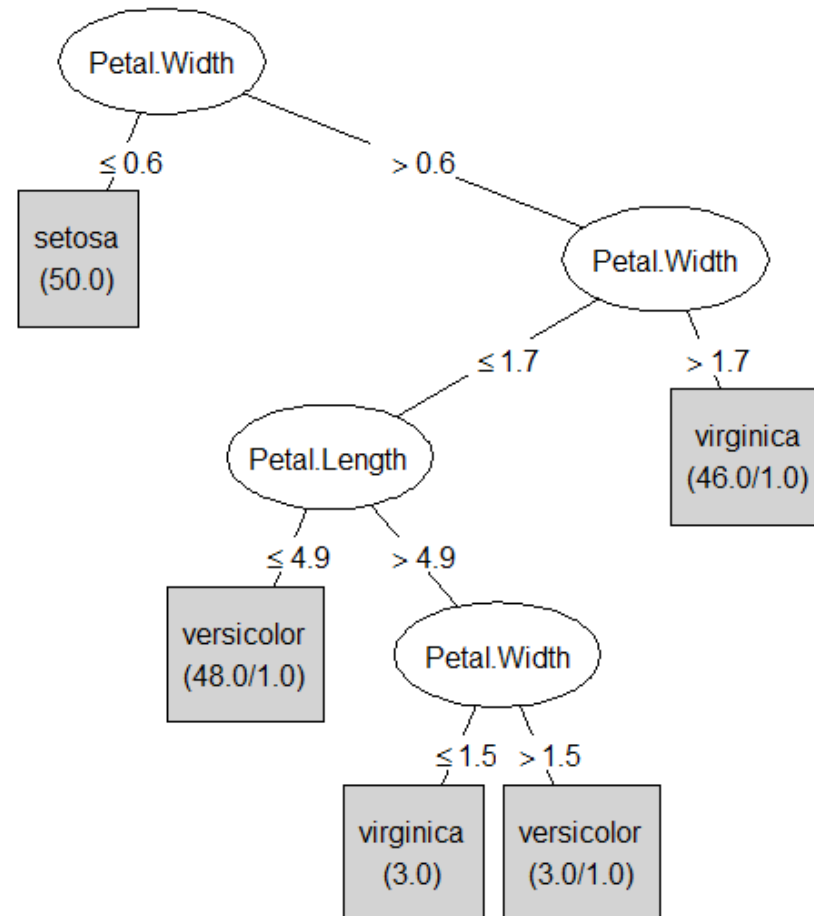
Algorithms

- ▶ “The most widely used learning algorithms are Support Vector Machines, linear regression, logistic regression, naive Bayes, linear discriminant analysis, decision trees, k-nearest neighbor algorithm, and Neural Networks (Multilayer perceptron).” –Wikipedia
- 

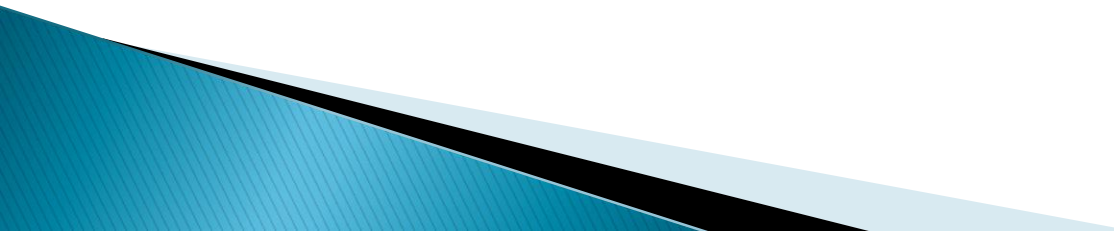
Decision Tree Example

- ▶ Suppose the Iris data set is used.
 - ▶ The task is classification.
 1. The Iris data is split into training and test data.
 2. The training data is input into the decision tree algorithm which produces a tree.
- 

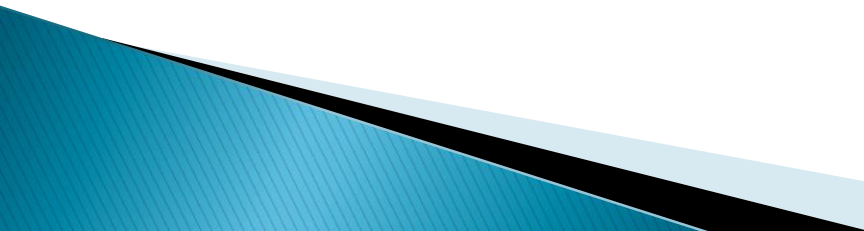
The decision tree *is* the inferred function in this case.



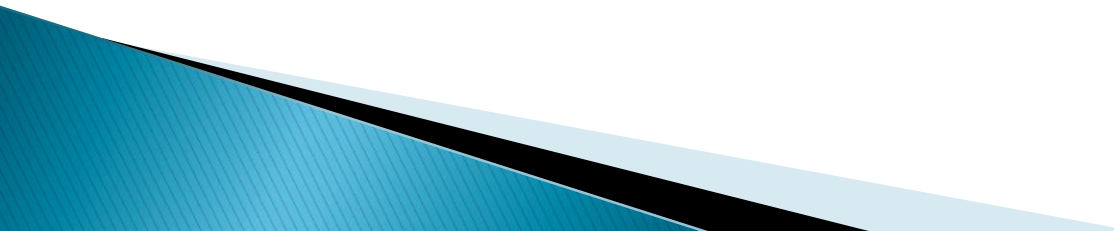
Testing

- ▶ The test data is then put through the tree.
 - ▶ The accuracy of the function is measured in terms of the relative frequency that it has accurately classified each test datum.
- 

Some Final Notes

- ▶ Supervised learning is more of a general approach to machine learning than a specific algorithm.
 - ▶ There sometimes exists partial but incomplete knowledge about the class or values of the output.
 - In this case, *semi-supervised learning* may be employed. This is any algorithm or technique that shares characteristics of both unsupervised and supervised learning.
- 

Review

- ▶ A pool of data is split into training data and test data.
 - ▶ An algorithm infers a function from input training data.
 - ▶ The test data is run through the function.
 - ▶ The accuracy of the output of the test data is the measure of the function's accuracy.
- 

Summary

- ▶ An overview of supervised learning
 - ▶ The Tasks for which it is used
 - ▶ As compared to unsupervised learning
 - ▶ A detailed look at the process
 - ▶ A list of the algorithms that are examples of this type of learning
 - ▶ An example using a decision tree
- 