Chapter 12

Discovering New Knowledge – Data Mining

Chapter Objectives

- Introduce the student to the concept of **Data Mining** (DM), also known as **Knowledge Discovery in Databases** (KDD).
  - How it is different from knowledge elicitation from experts
  - How it is different from extracting existing knowledge from databases.
- The objectives of data mining
  - Explanation of past events (descriptive DM)
  - Prediction of future events (predictive DM)
- (continued)

Chapter Objectives (cont.)

- Introduce the student to the different classes of statistical methods available for DM
  - Classical statistics (e.g., regression, curve fitting, …)
  - Induction of symbolic rules
  - Neural networks (a.k.a. “connectionist” models)
- Introduce the student to the details of some of the methods described in the chapter.

Historical Perspective

- DM, a.k.a. KDD, arose at the intersection of three independently evolved research directions:
  - Classical statistics and statistical pattern recognition
  - Machine learning (from symbolic AI)
  - Neural networks

Objectives of Data Mining

- **Descriptive DM** seeks patterns in past actions or activities to affect these actions or activities
  - eg, seek patterns indicative of fraud in past records
- **Predictive DM** looks at past history to predict future behavior
  - **Classification** classifies a new instance into one of a set of discrete predefined categories
  - **Clustering** groups items in the data set into different categories
  - **Affinity** or **association** finds items closely associated in the data set

Classical statistics & statistical pattern recognition

- Provide a detailed description of the most important statistical methods for data mining
  - Curve fitting with least squares method
  - Multi-variate correlation
  - K-Means clustering
  - Market Basket analysis
  - Discriminant analysis
  - Logistic regression

Figure 12.14 – 2-D input data plotted on a graph

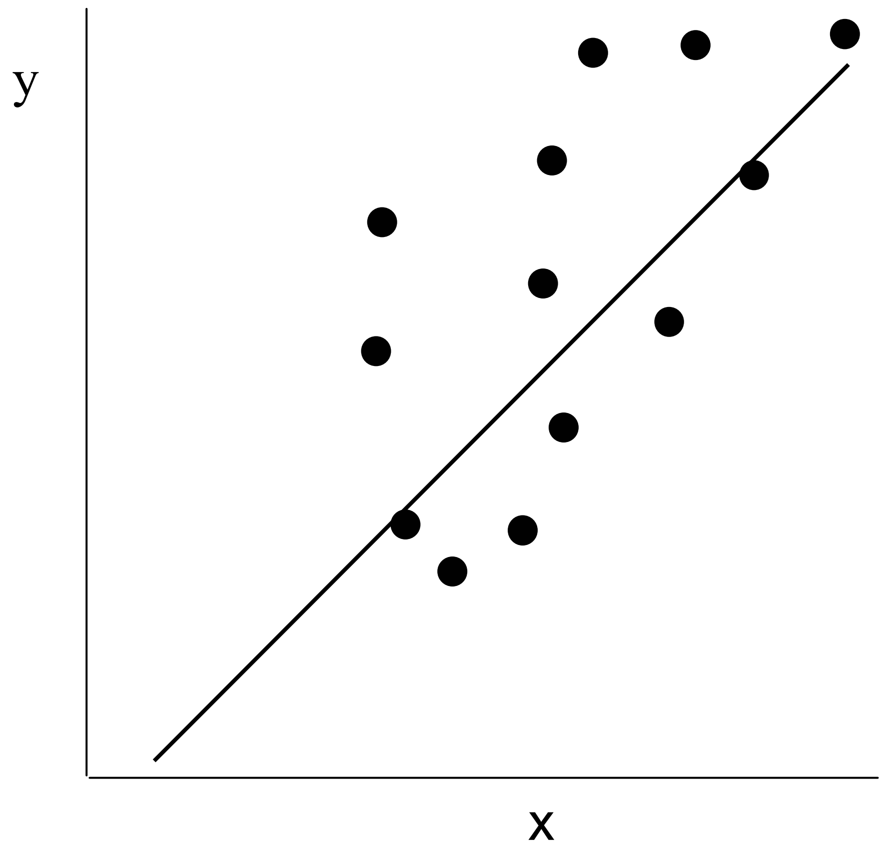
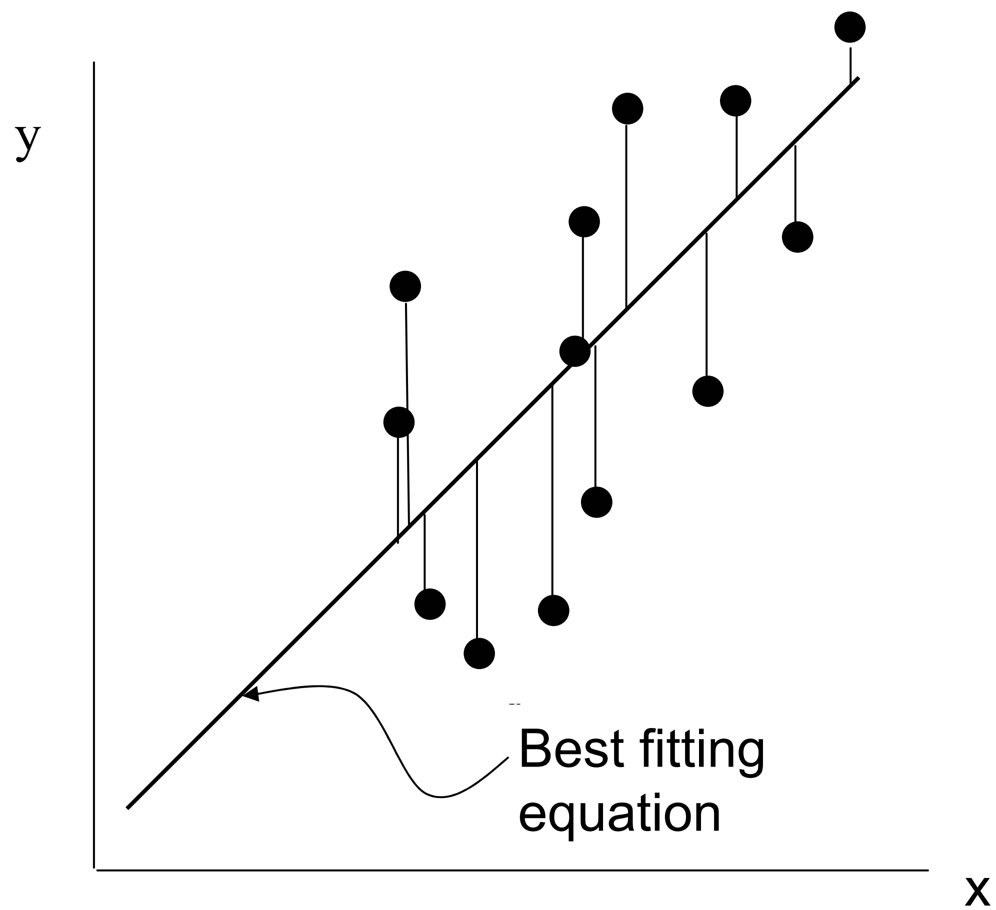


Figure 12.15 – data and deviations



Induction of symbolic rules

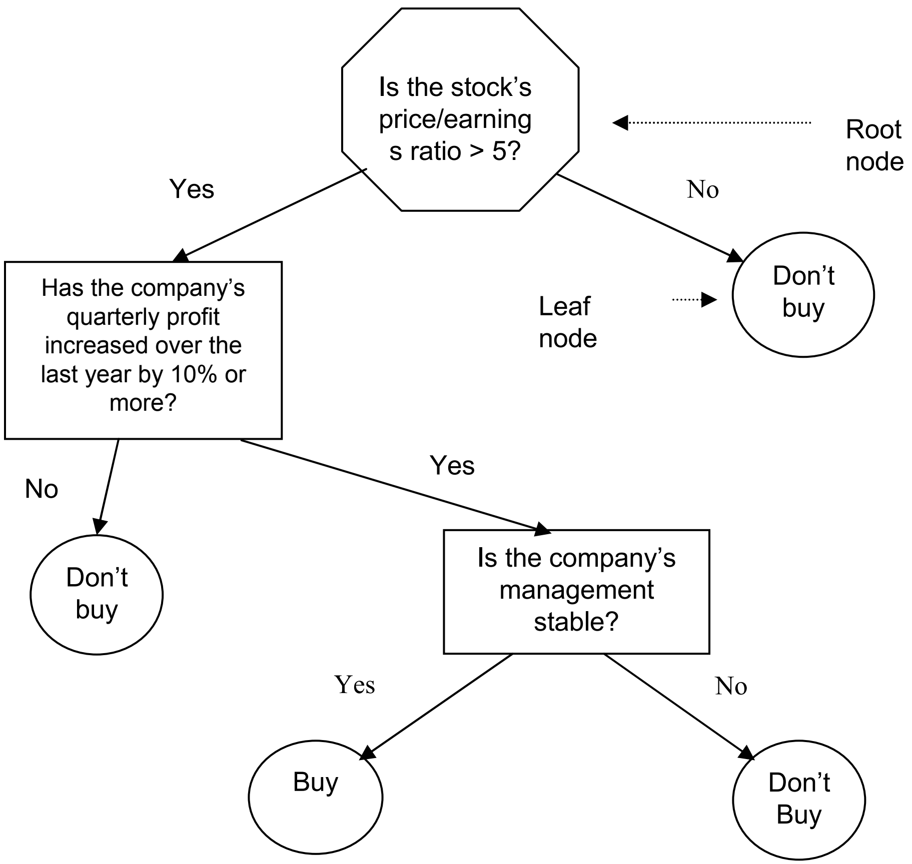
- Present a detailed description of the symbolic approach to data mining - rule induction by learning decision trees
- Present the main algorithm for rule induction
  - C5.0 and its ancestors, ID3 and CLS (from machine learning)
  - CART (Classification And Regression Trees) and CHAID, very similar algorithms for rule induction (independently developed in statistics)
- Present several example applications of rule induction

Table 12.1 – decision tables (if ordered, then **decision lists**)

Name	Outlook	Temperature	Humidity	Class
Data sample1	Sunny	Mild	Dry	Enjoyable
Data sample2	Cloudy	Cold	Humid	Not Enjoyable
Data sample3	Rainy	Mild	Humid	Not Enjoyable
Data sample4	Sunny	Hot	Humid	Not Enjoyable

Note: DS = Data Sample

Figure 12.1 – decision trees (a.k.a. classification trees)



Induction trees

- An **induction tree** is a decision tree holding the data samples (of the **training set**)
- Built progressively by gradually segregating the data samples

Figure 12.2 – simple induction tree (step 1)

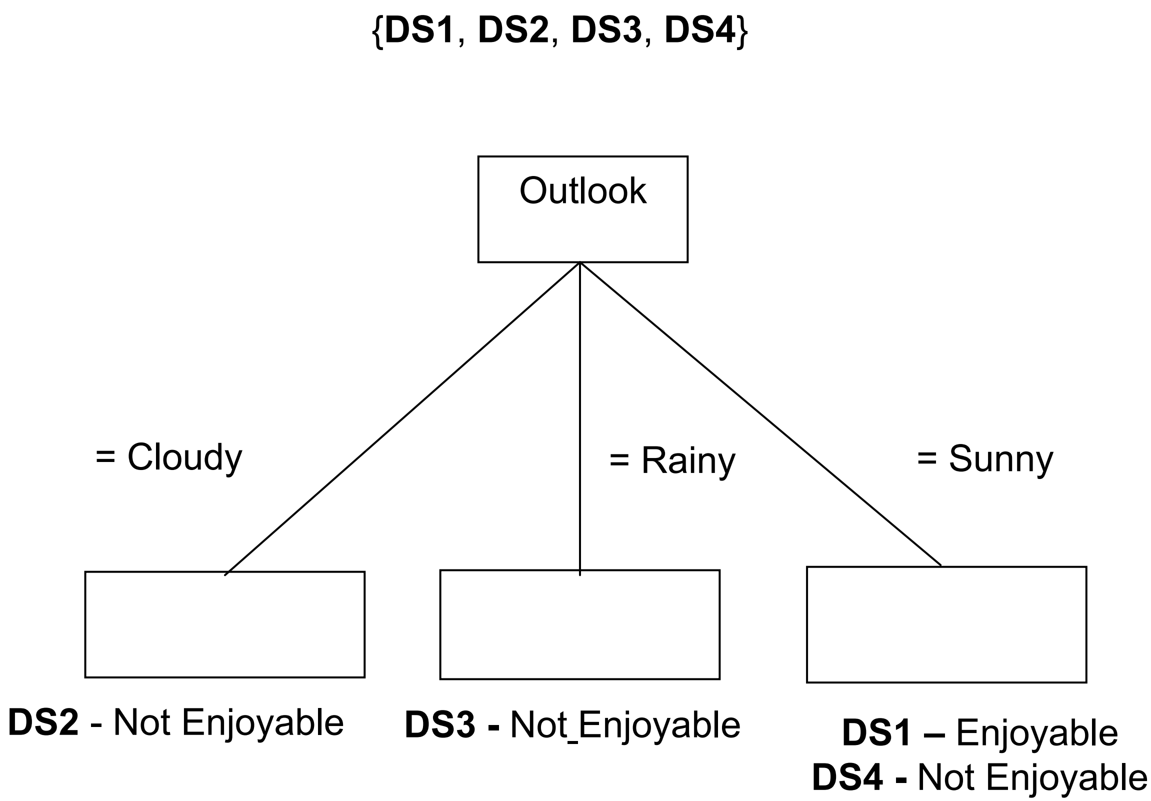
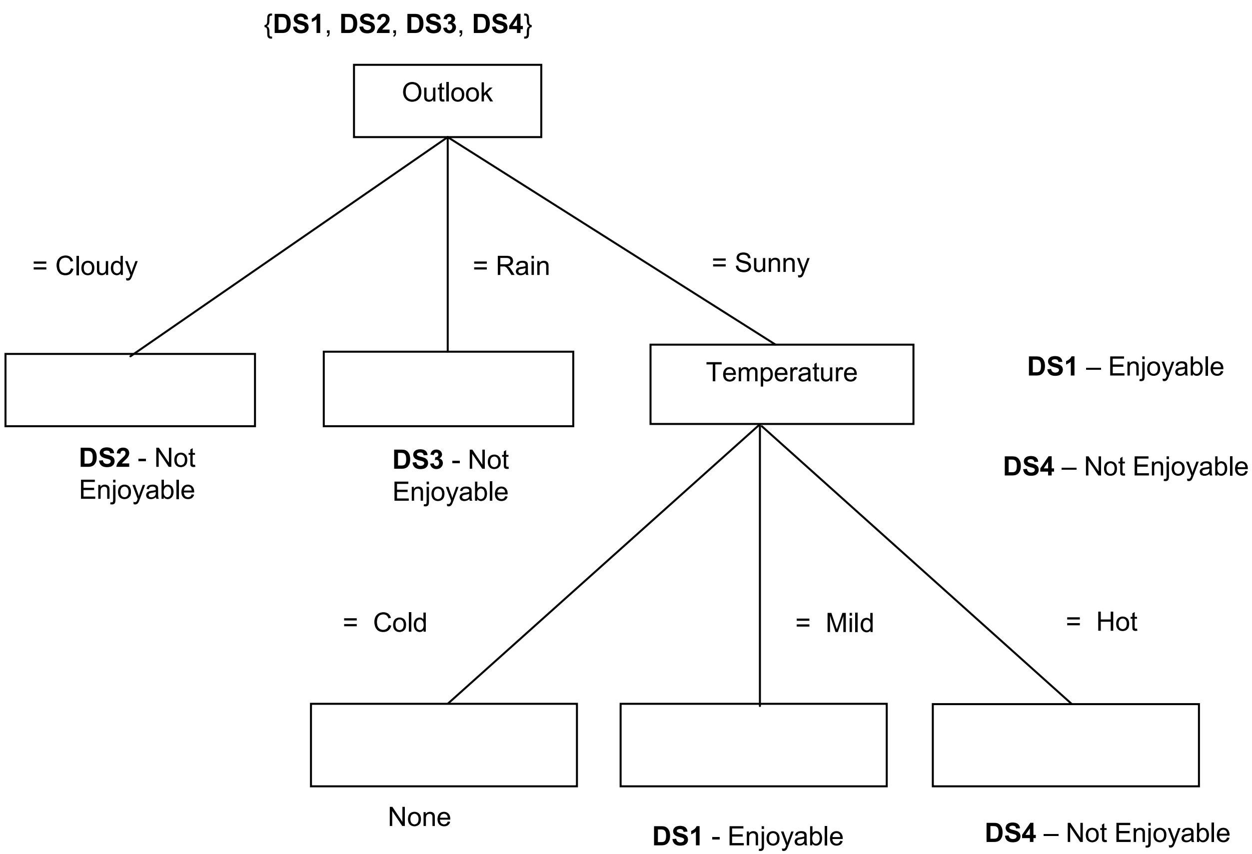


Figure 12.3 – simple induction tree (step 2)



Writing the induced tree as rules

- **Rule 1.** If the *Outlook* is *cloudy*, then the *Weather* is *not enjoyable*.
- **Rule 2.** If the *Outlook* is *rainy*, then the *Weather* is *not enjoyable*.
- **Rule 3.** If the *Outlook* is *sunny* and *Temperature* is *mild*, then the *Weather* is *enjoyable*.
- **Rule 4.** If the *Outlook* is *sunny* and *Temperature* is *cold*, then the *Weather* is *not enjoyable*.

Learning decision trees for classification into multiple classes

- In the previous example, we were learning a function to predict a boolean (*enjoyable* = true/false) output.
- The same approach can be generalized to learn a function that predicts a class (when there are multiple predefined classes/categories).
- For example, suppose we are attempting to select a KBS shell for some application:
  - with the following as our options:
    - ThoughtGen, Offsite, Genie, SilverWorks, XS, MilliExpert
  - using the following attributes and range of values:
    - Development language: { java, C++, lisp }
    - Reasoning method: { forward, backward }
    - External interfaces: { dBase, spreadsheetXL, ASCII file, devices }
    - Cost: any positive number
    - Memory: any positive number

Table 12.2 –
collection of data samples **(training set)** described as vectors of attributes **(feature vectors)**

Language	Reasoning method	Interface Method	Cost	Memory	Classification
Java	Backward	SpreadsheetXL	250	128MB	MilliExpert
Java	Backward	ASCII	250	128MB	MilliExpert
Java	Backward	dBase	195	256MB	ThoughtGen
Java	*	Devices	985	512MB	OffSite
C++	Forward	*	6500	640MB	Genie
LISP	Forward	*	15000	5GB	Silverworks
C++	Backward	*	395	256MB	XS
LISP	Backward	*	395	256MB	XS

Figure 12.4 – decision tree resulting from selection of the language attribute

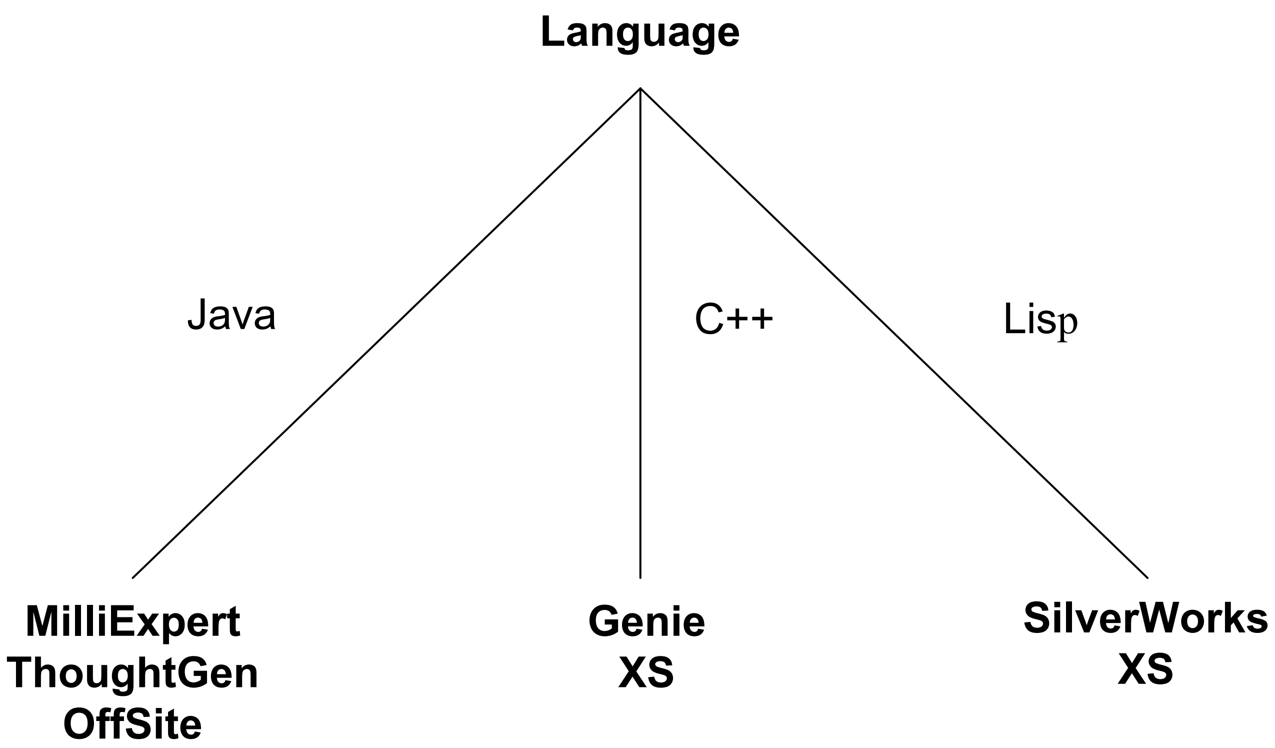


Figure 12.5 – decision tree resulting from addition of the reasoning method attribute

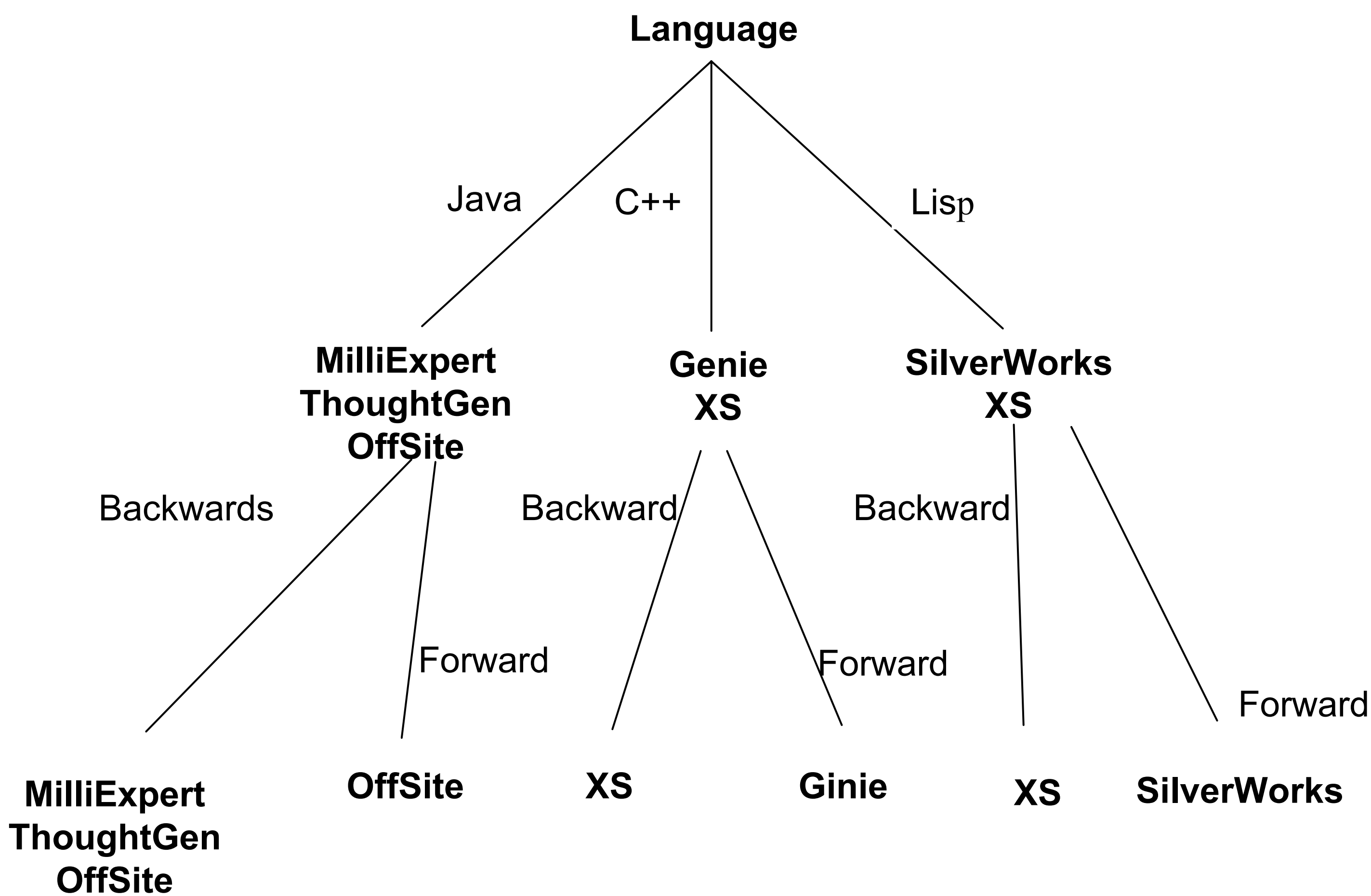
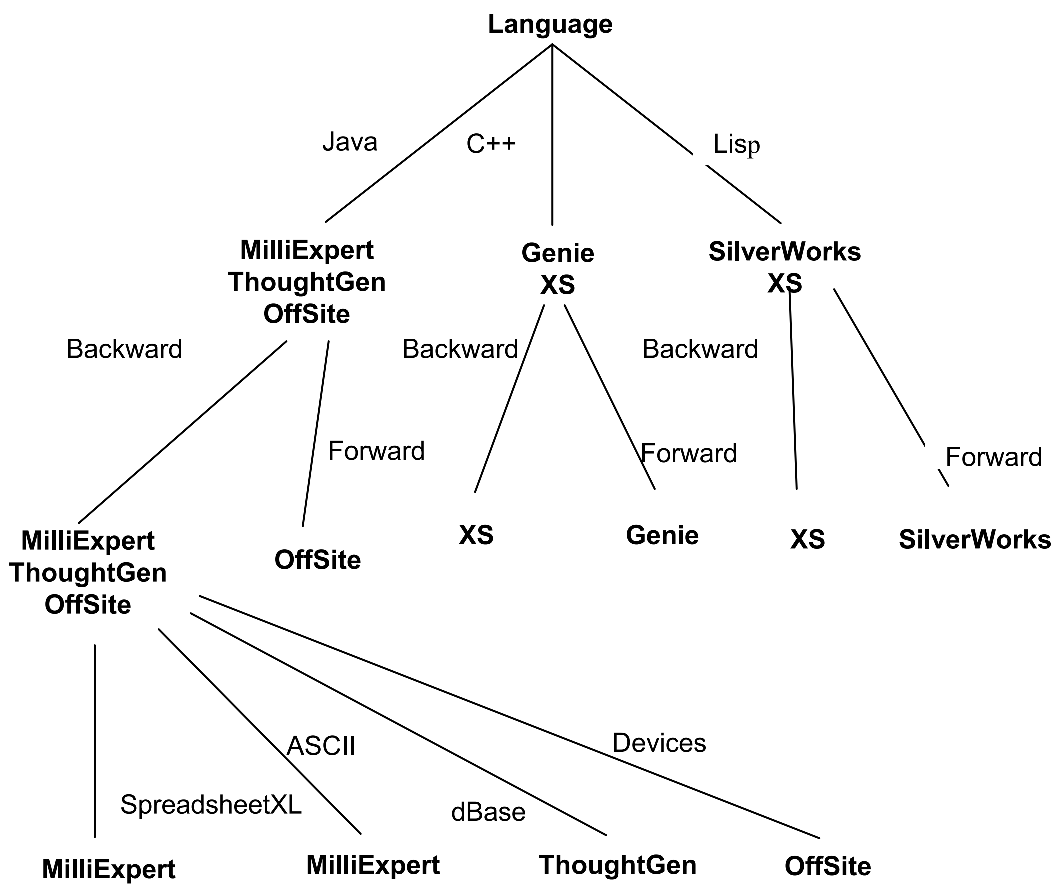


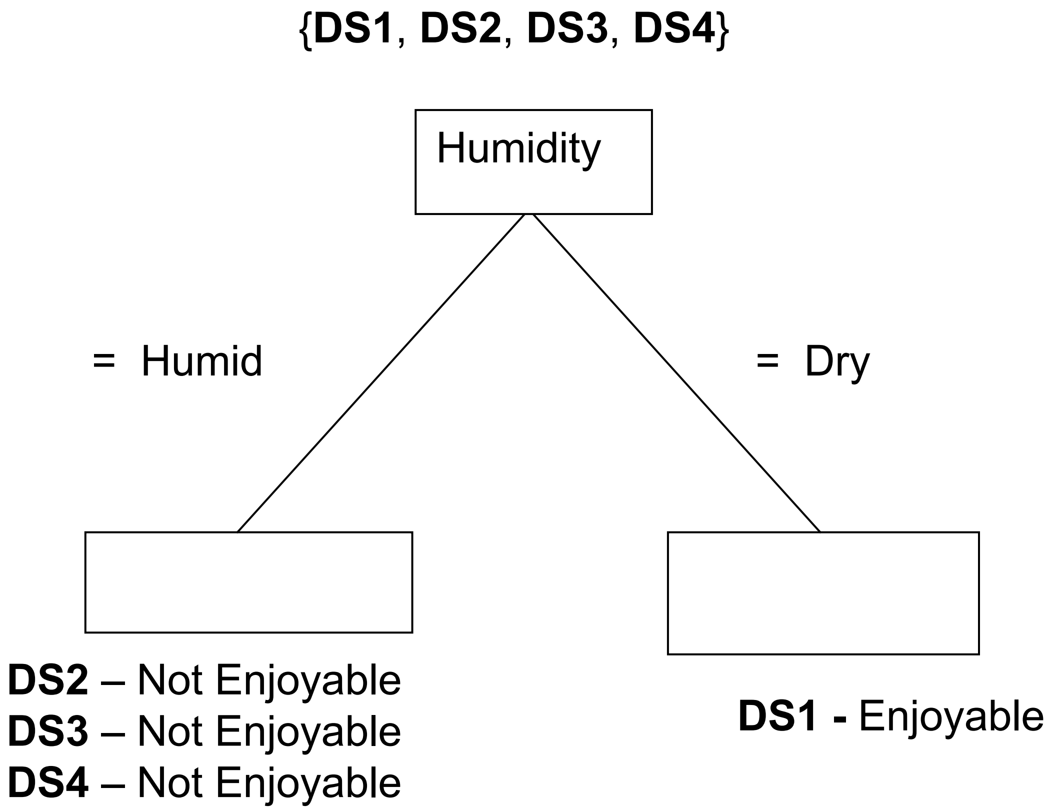
Figure 12.6 – final decision tree



Order of choosing attributes

- Note that the decision tree that is built depends greatly on which attributes you choose first

Figure 12.7



{**DS1**, **DS2**, **DS3**, **DS4**}
Humidity
= Humid
= Dry
DS2 – Not Enjoyable
DS3 – Not Enjoyable
DS4 – Not Enjoyable
DS1 - Enjoyable

Order of choosing attributes (cont)

- One sensible objective is to seek the **minimal tree**, ie, the smallest tree required to classify all training set samples correctly
  - **Occam's Razor** principle: the simplest explanation is the best
- What order should you choose attributes in, so as to obtain the minimal tree?
  - Often too complex to be feasible
  - Heuristics used
  - **Information gain**, computed using information theoretic quantities, is the best way in practice

Artificial Neural Networks

- Provide a detailed description of the connectionist approach to data mining - neural networks
- Present the basic neural network architecture - the multi-layer feed forward neural network
- Present the main supervised learning algorithm - backpropagation
- Present the main unsupervised neural network architecture - the Kohonen network

Figure 12.8 – simple model of a neuron

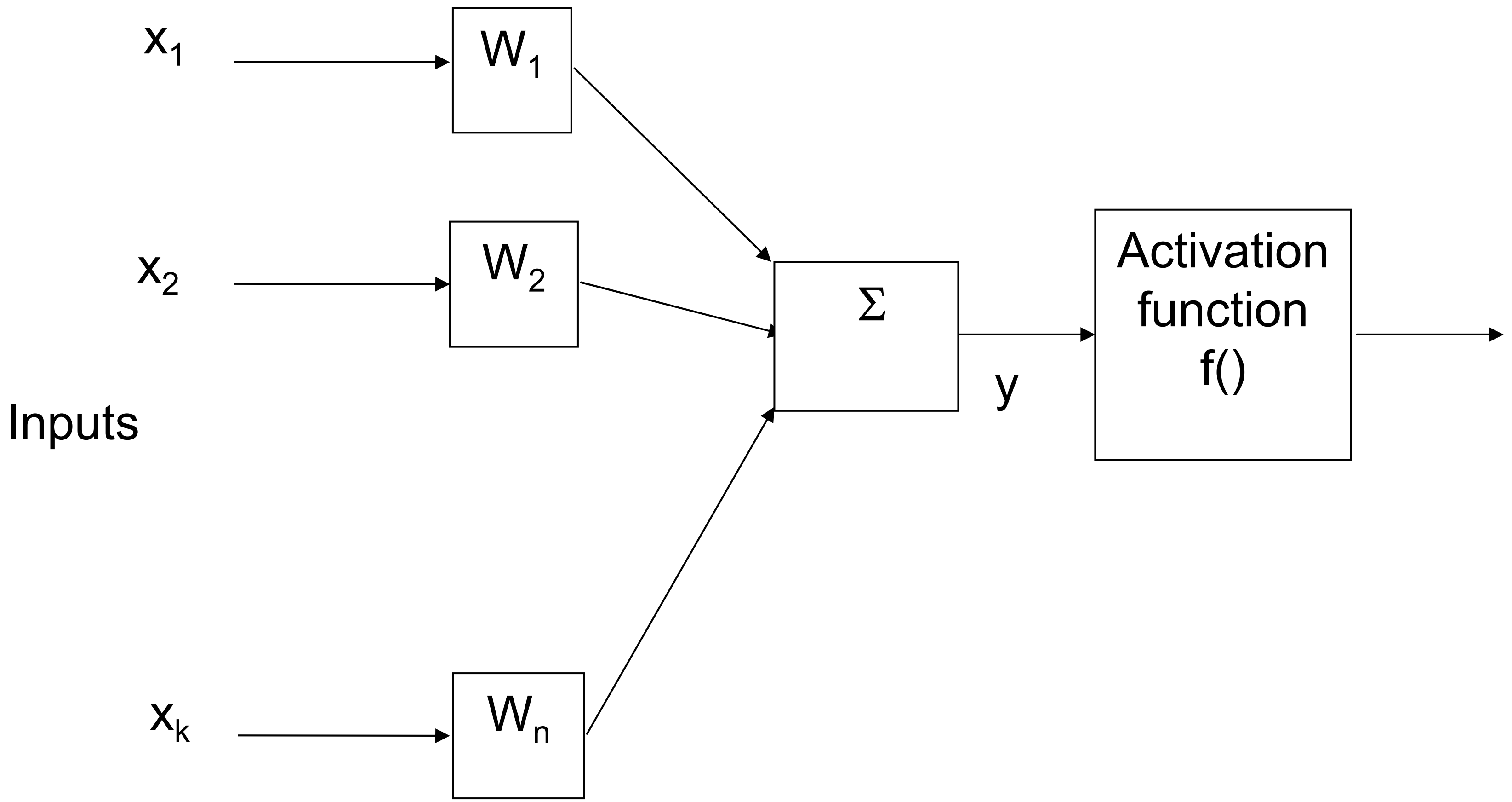


Figure 12.9 – three common activation functions

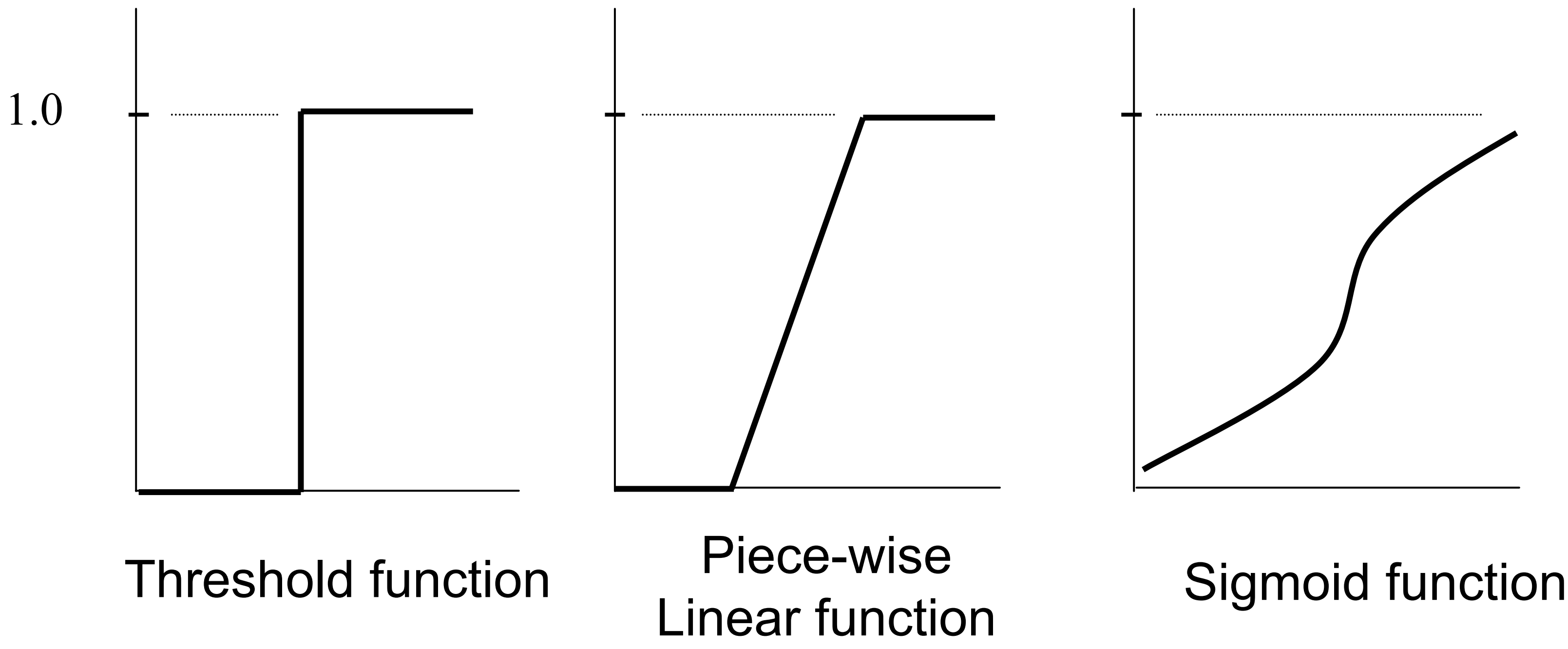


Figure 12.10 – simple single-layer neural network

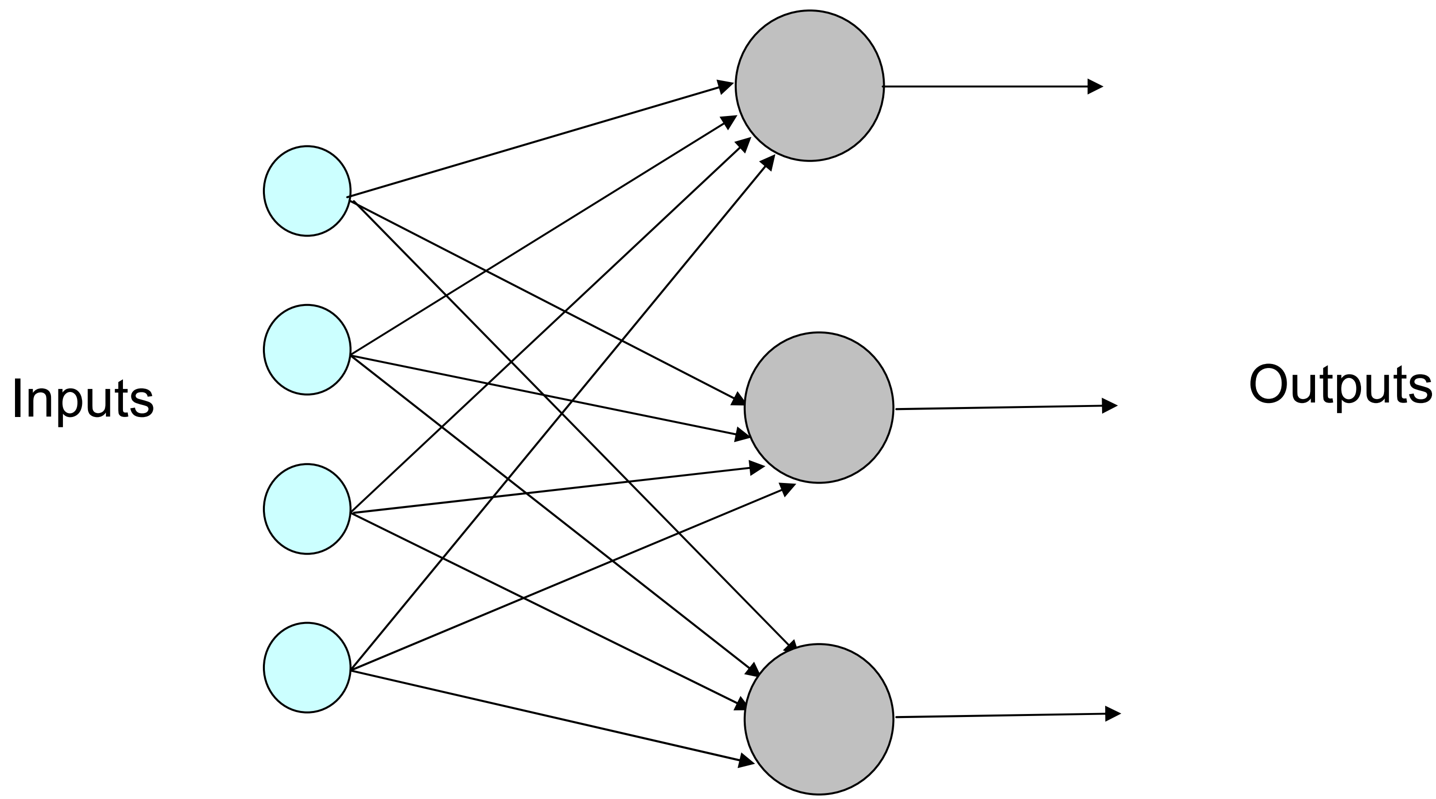
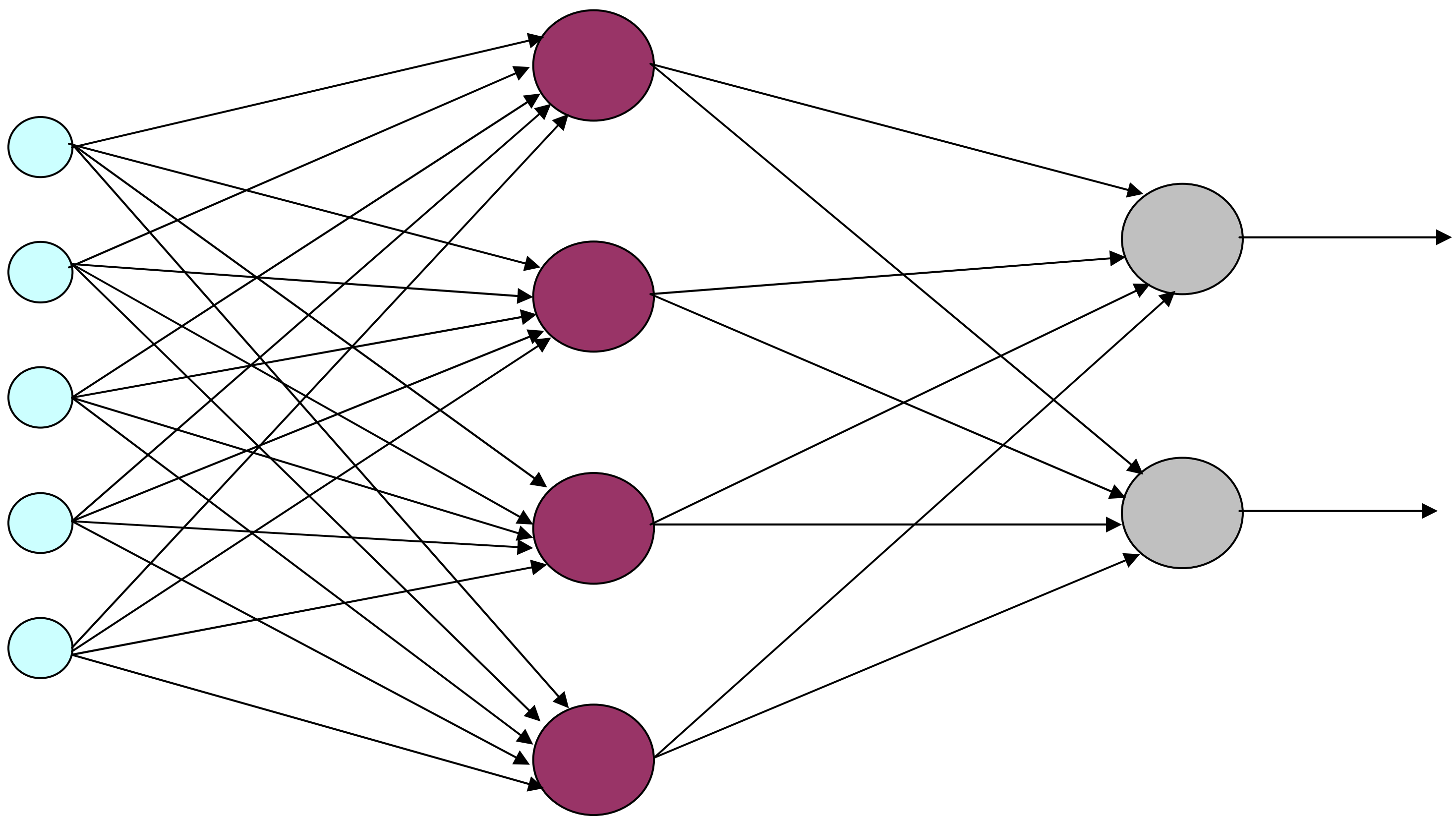


Figure 12.11 – two-layer neural network



Supervised Learning: Back Propagation

- An iterative learning algorithm with three phases:
  1. Presentation of the examples (input patterns with outputs) and feed forward execution of the network
  2. Calculation of the associated errors when the output of the previous step is compared with the expected output and back propagation of this error
  3. Adjustment of the weights

Unsupervised Learning: Kohonen Networks

- Clustering by an iterative competitive algorithm
- Note relation to CBR

Figure 12.12 – clusters of related data in 2-D space

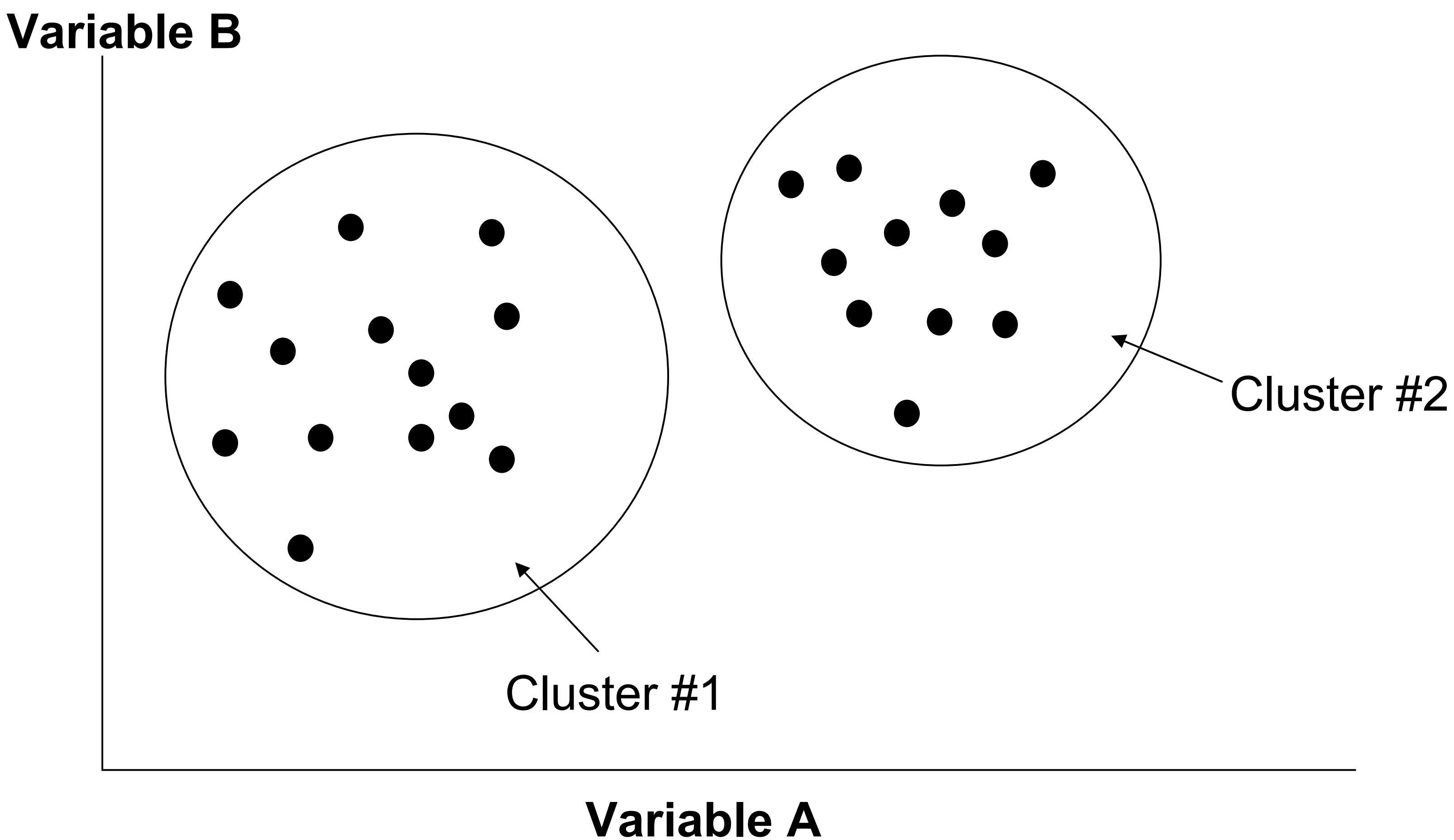
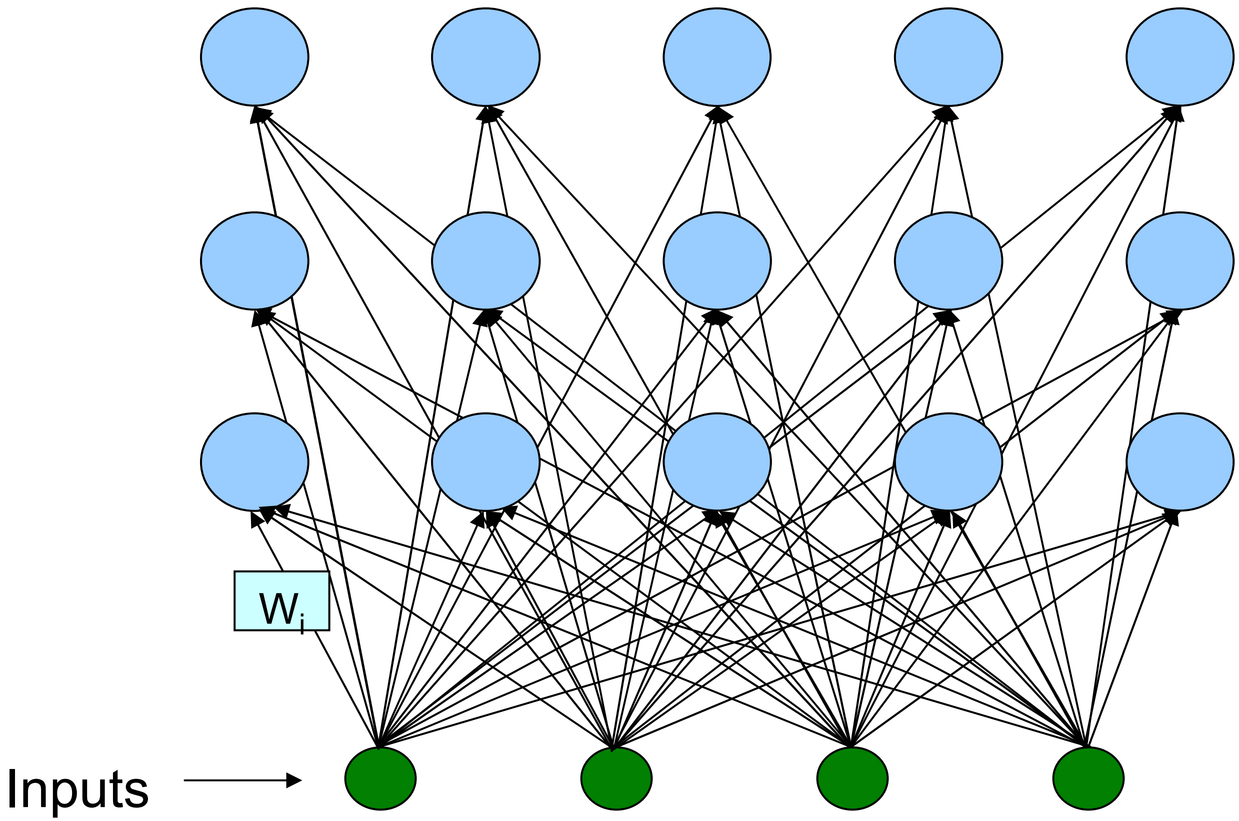


Figure 12.13 – Kohonen self-organizing map



When to use what

- Provide useful guidelines for determining what technique to use for specific problems

Table 12.3

Goal	Input Variables (Predictors)	Output Variables (Outcomes)	Statistical Technique	Examples [SPSS, 2000]
Find linear combination of predictors that best separate the population	Continuous	Discrete	Discriminant Analysis	• Predict instances of fraud • Predict whether customers will remain or leave (churners or not) • Predict which customers will respond to a new product or offer •Predict outcomes of various medical procedures
Predict the probability of outcome being in a particular category	Continuous	Discrete	Logistic and Multinomial Regression	• Predicting insurance policy renewal • Predicting fraud • Predicting which product a customer will buy • Predicting that a product is likely to fail

Table 12.3 (cont.)

Goal	Input Variables (Predictors)	Output Variables (Outcomes)	Statistical Technique	Examples [SPSS, 2000]
Output is a linear combination of input variables	Continuous	Continuous	Linear Regression	• Predict expected revenue in dollars from a new customer • Predict sales revenue for a store • Predict waiting time on hold for callers to an 800 number. • Predict length of stay in a hospital based on patient characteristics and medical condition.
For experiments and repeated measures of the same sample	Most inputs must be Discrete	Continuous	Analysis of Variance (ANOVA)	• Predict which environmental factors are likely to cause cancer
To predict future events whose history has been collected at regular intervals	Continuous	Continuous	Time Series Analysis	• Predict future sales data from past sales records

Table 12.4

Goal	Input (Predictor) Variables	Output (Outcome) Variables	Statistical Technique	Examples [SPSS, 2000]
Predict outcome based on values of nearest neighbors	Continuous, Discrete, and Text	Continuous or Discrete	Memory-based Reasoning (MBR)	•Predicting medical outcomes
Predict by splitting data into subgroups (branches)	Continuous or Discrete (Different techniques used based on data characteristics)	Continuous or Discrete (Different techniques used based on data characteristics)	Decision Trees	•Predicting which customers will leave •Predicting instances of fraud
Predict outcome in complex non-linear environments	Continuous or Discrete	Continuous or Discrete	Neural Networks	•Predicting expected revenue •Predicting credit risk

Table 12.5

Goal	Input (Predictor) Variables	Output (Outcome) Variables	Statistical Technique	Examples [SPSS, 2000]
Predict by splitting data into more than two subgroups (branches)	Continuous, Discrete, or Ordinal	Discrete	Chi-square Automatic Interaction Detection (CHAID)	• Predict which demographic combinations of predictors yield the highest probability of a sale • Predict which factors are causing product defects in manufacturing
Predict by splitting data into more than two subgroups (branches)	Continuous	Discrete	C5.0	• Predict which loan customers are considered a "good" risk • Predict which factors are associated with a country's investment risk

Table 12.5 (cont.)

Goal	Input (Predictor) Variables	Output (Outcome) Variables	Statistical Technique	Examples [SPSS, 2000]
Predict by splitting data into binary subgroups (branches)	Continuous	Continuous	Classification and Regression Trees (CART)	• Predict which factors are associated with a country's competitiveness • Discover which variables are predictors of increased customer profitability
Predict by splitting data into binary subgroups (branches)	Continuous	Discrete	Quick, Unbiased, Efficient, Statistical Tree (QUEST)	•Predict who needs additional care after heart surgery

Table 12.6

Goal	Input Variables (Predictor)	Output Variables (Outcome)	Statistical Technique	Examples [SPSS, 2000]
Find large groups of cases in large data files that are similar on a small set of input characteristics,	Continuous or Discrete	No outcome variable	K-means Cluster Analysis	• Customer segments for marketing • Groups of similar insurance claims
To create large cluster memberships			Kohonen Neural Networks	• Cluster customers into segments based on demographics and buying patterns
Create small set associations and look for patterns between many categories	Logical	No outcome variable	Market Basket or Association Analysis with Apriori	• Identify which products are likely to be purchased together • Identify which courses students are likely to take together

Errors and their significance in DM

- Discuss the importance of errors in data mining studies
- Define the types of errors possible in data mining studies

Table 12.7

Heart Disease Diagnostic	*Predicted No Disease*	*Predicted Presence of Disease*
Actual No Disease	118 (72%)	46 (28%)
Actual Presence of Disease	43 (30.9%)	96 (69.1%)

Conclusions

- The student should be able to use:
  - Curve-fitting algorithms.
  - Statistical methods for clustering.
  - The C5.0 algorithm to capture rules from examples.
  - Basic feedforward neural networks with supervised learning.
  - Unsupervised learning, clustering techniques and the Kohonen networks.
  - Other statistical techniques.

Chapter 12

Discovering New Knowledge – Data Mining