# Unit 2
# Data Types, Input and Output of Data Mining Algorithms

# Introduction

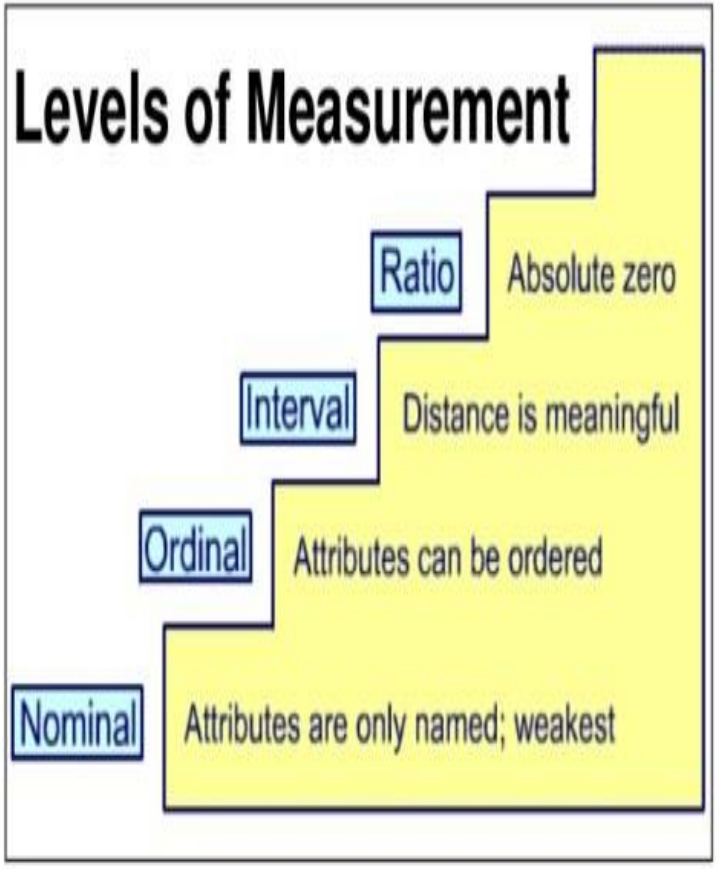
- Data types and characteristics of dataset has to be analysed
- Data is classified as structured, unstructured and semi-structured data
- Business databases mostly are structured
- Scientific data may be of all three classes
- Unstructured data may be like multimedia recordings
- Structured data is also called traditional data
- Other are non-traditional data
- Most of the DM methods are applied to traditional data

# Instances and Features

- Potential measurement is called as Features
- Features are measured uniformly over many cases
- Representation of structured data is in tabular form or in the form of single relation
- Each row is the instances or it is sample

# Types of Features

- Nominal variables – Features which help to identify unique entities
- Ordinal variables – categories that can be rationally listed in some order
- Interval variables – ordinal variables in which the distance between the ordered categories can be measured
- Ratio variables – interval variables which can have zero starting point



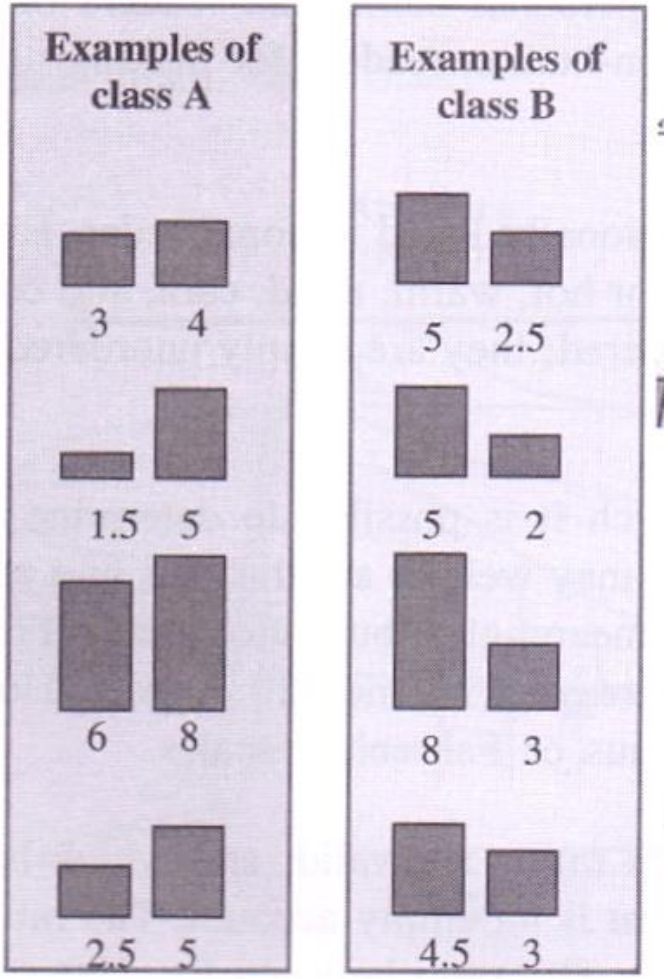
- Data Mining algorithms,
  - Do not distinguish nominal and ordinal type
  - Do not distinguish ratio and interval variables
- They are concerned whether the variables are unordered labels or ordered labels

# Concept learning and Concept Description

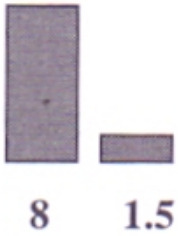
- 4 types of DM methodologies
  - Classification learning
  - Association learning
  - Clustering
  - Regression

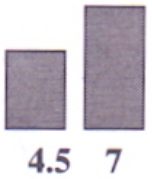
# Classification learning

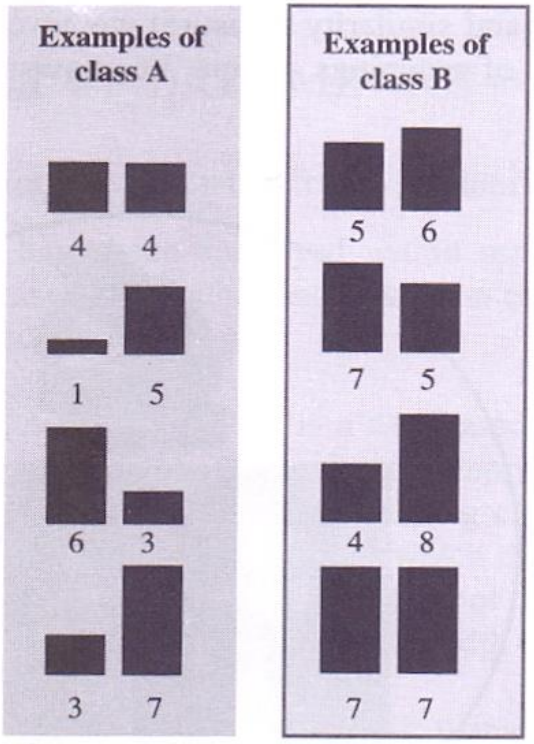
- Here, you have set of classified examples from which you are expected to learn a way of classifying unseen examples
- The required set of rules to partition the data into exclusive groups is called classification rules. The data used for deriving such rules is called training set.
- The rules can then be used to discover as to which group a new customer belongs to.


Examples of class A
3 4
1.5 5
6 8
2.5 5


Examples of class B
5 2.5
5 2
8 3
4.5 3


8 1.5


4.5 7


Examples of class A
4 4
1 5
6 3
3 7
Examples of class B
5 6
7 5
4 8
7 7


6 6

# Association learning

- Association between the features is derived.
- The output will be association rules.

| TID | Produce |
|---|---|
| 1 | MILK, BREAD, EGGS |
| 2 | BREAD, SUGAR |
| 3 | BREAD, CEREAL |
| 4 | MILK, BREAD, SUGAR |
| 5 | MILK, CEREAL |
| 6 | BREAD, CEREAL |
| 7 | MILK, CEREAL |
| 8 | MILK, BREAD, CEREAL, EGGS |
| 9 | MILK, BREAD, CEREAL |

# Clustering

- Groups of examples that belong together are sought
- Grouping is based on the similarity of features
- Clustering can be applied to
  - Marketing : finding group of customers with similar behaviour
  - Biology: grouping of plants and animals based on their features
  - Libraries: book ordering
  - Insurance: identifying groups of policy holders, identifying frauds
  - Earthquakes: Identifying group of places which receives earthquakes
  - www: Document classification, clustering web log data

| Subject | A | B |
|---|---|---|
| 1 | 1.0 | 1.0 |
| 2 | 1.5 | 2.0 |
| 3 | 3.0 | 4.0 |
| 4 | 5.0 | 7.0 |
| 5 | 3.5 | 5.0 |
| 6 | 4.5 | 5.0 |
| 7 | 3.5 | 4.5 |

# Regression

- Regression helps to perform numeric prediction
- Here the outcome to be predicted is not a discrete class but a numeric quantity

| State | District | Year | Jan | Feb | Mar | Apr |
|---|---|---|---|---|---|---|
| Karnataka | Uttar Kannada | 2006 | 0.0 | 0.0 | 8.3 | 0.6 |
| Karnataka | Uttar Kannada | 2007 | 0.0 | 0.1 | 1.9 | 27.8 |
| Karnataka | Uttar Kannada | 2008 | 0.0 | 5.3 | 95.3 | 11.9 |
| Karnataka | Uttar Kannada | 2009 | 0.0 | 0.0 | 27.6 | 16.4 |
| Karnataka | Uttar Kannada | 2010 | 8.9 | 0.0 | 1.7 | 48.9 |

| cycle time (ns) | memory min (kB) | memory max (kB) | cache (kB) | chan min | chan max | perfor-mance |
|---|---|---|---|---|---|---|
| 125 | 256 | 6000 | 256 | 16 | 128 | 198 |
| 29 | 8000 | 32000 | 32 | 8 | 32 | 269 |
| 29 | 8000 | 32000 | 32 | 8 | 32 | 220 |
| 29 | 8000 | 32000 | 32 | 8 | 32 | 172 |
| 29 | 8000 | 16000 | 32 | 8 | 16 | 132 |
| ... | | | | | | |
| 125 | 2000 | 8000 | 0 | 2 | 14 | 52 |
| 480 | 512 | 4000 | 32 | 0 | 0 | 67 |
| 480 | 1000 | 4000 | 0 | 0 | 0 | 45 |

# Output of Decision Trees

| Tid | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

Training Data


Model: Decision Tree

# Output of Data Mining

- Knowledge output from classified learners
  - Decision trees
  - Neural Networks
- Decision tree is a tree based knowledge representation methodology used to represent classification rules.
- The leaf node represents the class labels while other node represents the attributes associated with the objects being classified.
- The branches of the tree represent each possible value of the attribute node from which they originate

# Output of Association Rules

| TID | Produce |
|---|---|
| 1 | MILK, BREAD, EGGS |
| 2 | BREAD, SUGAR |
| 3 | BREAD, CEREAL |
| 4 | MILK, BREAD, SUGAR |
| 5 | MILK, CEREAL |
| 6 | BREAD, CEREAL |
| 7 | MILK, CEREAL |
| 8 | MILK, BREAD, CEREAL, EGGS |
| 9 | MILK, BREAD, CEREAL |

**{Cereal, Milk} → Bread**
**[sup=22%, conf=50%]**

Association rule:

22% chance that customers buy all the three products together

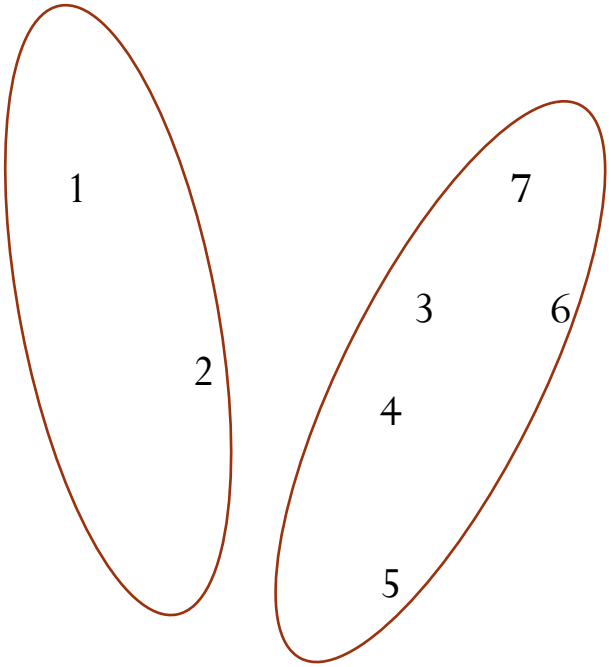
50% chance that customers who buy cereal and milk will buy bread

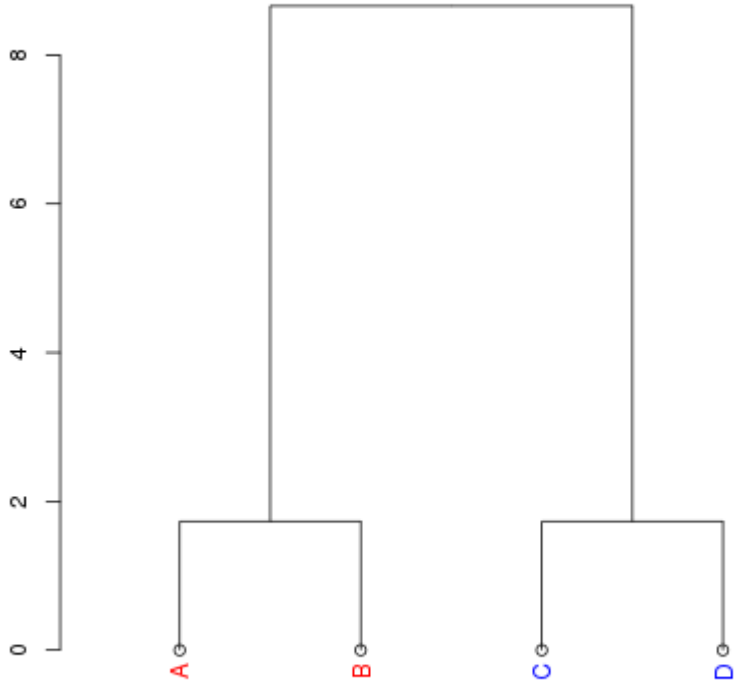
# Clustering

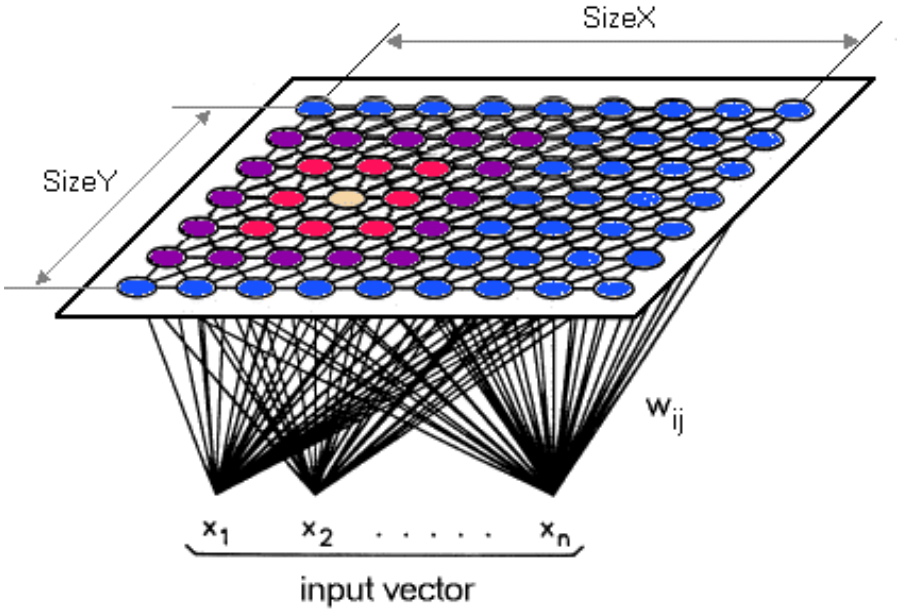
| Subject | A | B |
|---|---|---|
| 1 | 1.0 | 1.0 |
| 2 | 1.5 | 2.0 |
| 3 | 3.0 | 4.0 |
| 4 | 5.0 | 7.0 |
| 5 | 3.5 | 5.0 |
| 6 | 4.5 | 5.0 |
| 7 | 3.5 | 4.5 |

Cluster1: 1,2
Cluster2: 3,4,5,6,7

- The cluster output can be visualised as
  - Table form
  - Venn diagram
  - Dendrograms
  - Self-organizing maps


1
2
7
3
6
4
5

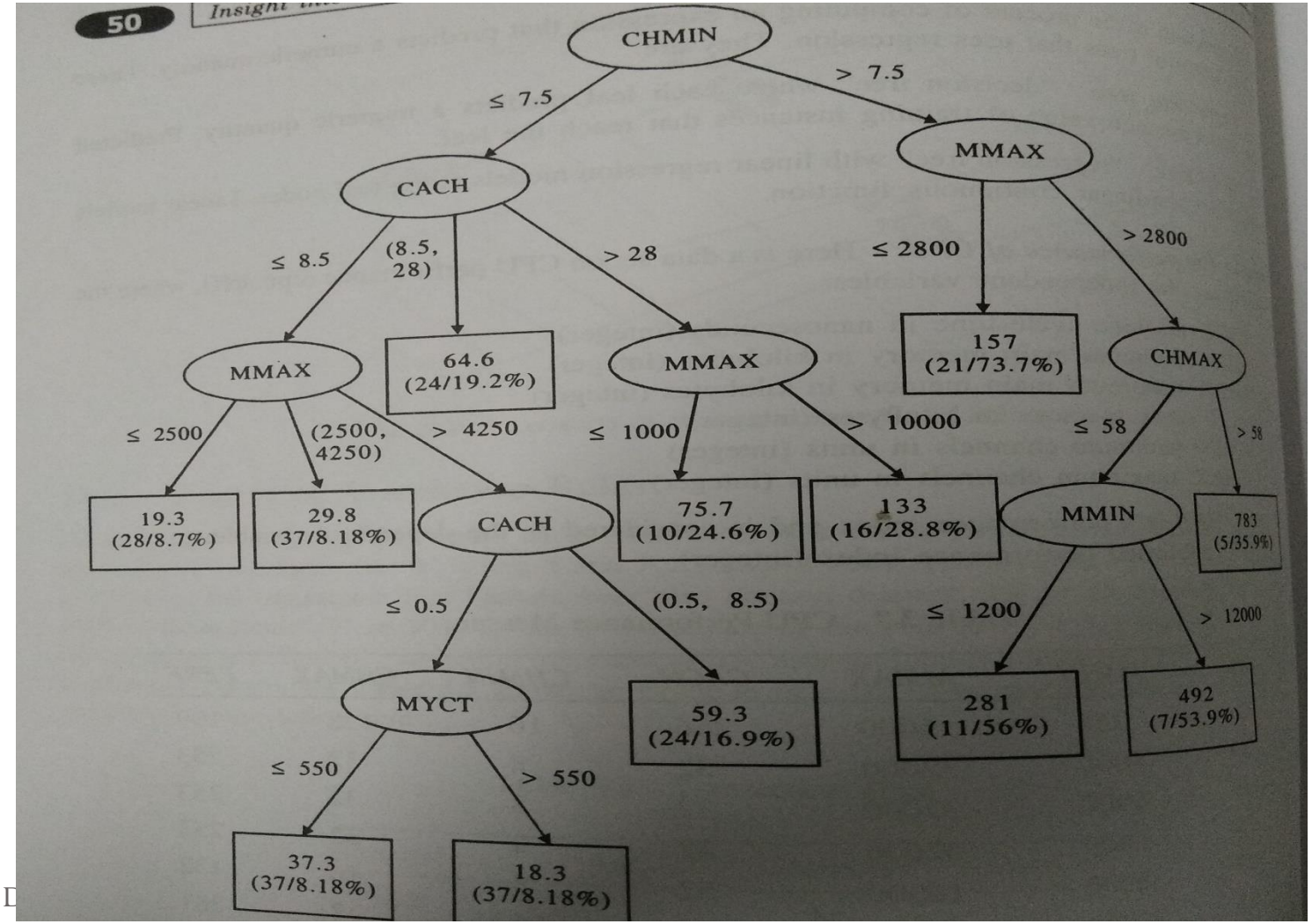

0
2
4
6
8
A
B
C
D


SizeX
SizeY
$w_{ij}$
$x_1$ $x_2$ . . . . . $x_n$
input vector

# Regression

- Generally trees for numeric prediction
  - Regression tree – leaf predicts a numeric qty. Predicted value is the average value of training instances that reach the leaf
  - Model tree – linear regression models at the leaf nodes.

# Regression tree



# Model tree

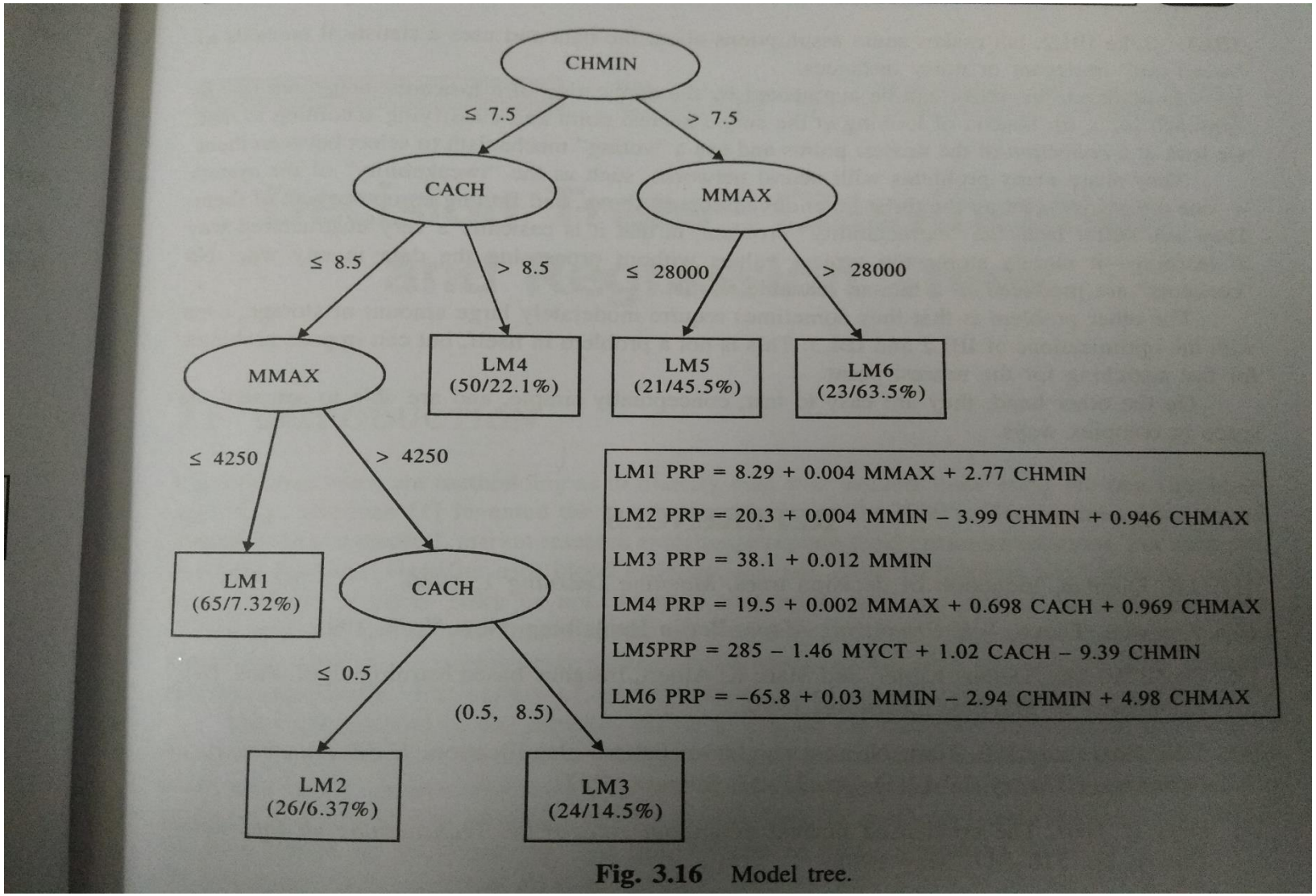



**Fig. 3.16** Model tree.

# Unit 2– Part II
# Preprocessing and Postprocessing

# Introduction

- Essential - Preparation and transformation of the initial dataset.
- If data comes from data warehouse, some transformation might be already done.
- It may not be the fullest extent.
- Some transformations may be done during data mining process only.
- Many transformations may be needed to produce features more useful for selected data mining methods like prediction or classification

# Steps in Preprocessing

- **Choosing the object representation**
  - Choose the appropriate attributes
  - Distinguish the attributes as categorical or numerical
  - Whether Structured ?
- **Mapping and collecting data**
  - Represent the data uniformly Like date 08/02/2018 08-02-2018
- **Scaling large datasets**
  - See that the complete dataset could be placed in memory
  - If not use windowing/batch incremental mode

# Steps in Preprocessing

- **Handling noise and errors**
  - Handle External errors and internal errors separately
- **Processing unknown attribute values**
  - Unknown values / missing values
  - Caused due to data lost, NA, doesnot exist, irrelevant to context
- **Discretization/Fuzzification of numerical attributes**
  - Discretize the numerical value
  - Extension of discretization is fuzzification
  - Offline or online processing can be done

# Normalization/Fuzzification

V={8,10,12,20} actual min = 8 actual max=20

New min = 0

New max=1

$$V' = \frac{(V\text{-}min)}{max\text{-}min} \times (newmax\text{-}newmin) + newmin$$

V'[8] = 8-8/12 * 1 = 0/12=0

V'[10] = 10-8/12*1=2/12=0.16

# Steps in Preprocessing

- **Processing of continuous classes**
  - We may have discretized symbolic classes or continuous numeric classes
- **Grouping of values of symbolic attributes**
  - Grouping of symbolic values if classes are more
  - It helps in making decision/classification rules easier
- **Attribute selection and ordering**
  - Select attributes and order according to the target concept
- **Attribute construction and transformation**
  - Reform and Transform the attribute if required such that the learning becomes easy
- **Consistency checking**
  - Check for the consistency of the data. First do it offline and then online

# Discretization

- The process of partitioning the continuous variables into categories is usually termed as discretization.
- Interval labels can then be used to replace actual data values
- Discretization can be done recursively on a attribute
- Discretization helps in reduce the data size
- It helps in transforming the quantitative data into qualitative data
- Unsupervised – Binning - equal width methods or equal frequency methods
- Supervised – entropy or information gain which measure strength of relationship to determine which group it belongs to
- Supervised – may lead to accurate calculation

- Manual approach
  - Manually define the cut points for the data.
  - They are defined subjectively
- Unsupervised method – Binning
- Binning does not use class information for discretization
  - Equi width binning
  - Equi frequency binning
- Supervised method – Entropy Selection

# Equi Width binning

- The algorithm divides the data into k intervals of equal size
- The width of interval is

$$W = (max-min)/k$$

The interval boundaries are

$$min+w, min+2w, \ldots . min+(k-1)w$$

# Equi frequency Binning

- The algorithm divides the data into k groups where each group contains approximately equal number of values
- For both methods, the best way of determining the k value is by seeing the histogram and trying different intervals or groups

# Entropy based discretization

- The goal of this algorithm is to find the split with the maximum information gain
- The boundary that minimizes the entropy over all possible boundaries is selected
- The process is recursively applied to partitions obtained until some stopping criterion is met
- Such a boundary may reduce data size and improve classification accuracy

# Calculating Entropy

- For m classes

$$Entropy(S) = -\sum_{i=1}^{m} p_i \log_2 p_i$$

- For 2 classes

$$Entropy(S) = -p_1 \log_2 p_1 - p_2 \log_2 p_2$$

- Calculated based on the class distribution of the samples in set S.
- pi is the probability of class i in S
- m is the number of classes (class values)

# Calculating entropy from Split

- Entropy of subsets S and S2 Entropy of subsets S1 and S2 are calculated .
- The calculations are weighted by their probability of being in set S and summed.
- In formula below,
  - S is the set
  - T is the value used to split S into S1 and S2

$$E(S,T) = \frac{|S_1|}{|S|} Entropy(S_1) + \frac{|S_2|}{|S|} Entropy(S_2)$$

# Calculation of information gain

- Information Gain = Difference in entropy between original set (S) and weighted split (S1 + S2)

$$Gain(S,T) = Entropy(S) - E(S,T)$$

- It is always desired to have maximum information gain for effective split.
- To find effective split take two split values and calculate information gain. Choose the split which has maximum information gain or less E(s1,s2)

- Calculate Entropy for your data.
- For each potential split in your data…
  - Calculate Entropy in each potential bin
  - Find the net entropy for your split
  - Calculate entropy gain
- Select the split with the highest entropy gain
- Recursively (or iteratively in some cases) perform the partition on each split until a termination criteria is met
  - Terminate once you reach a specified number of bins
  - Terminate once entropy gain falls below a certain threshold.

# Practice

| Hours Studied | A Grade on Test |
|---|---|
| 4 | N |
| 5 | Y |
| 8 | N |
| 12 | Y |
| 15 | Y |

| | A on Test | Lower than A |
|---|---|---|
| Overall | 3 | 2 |

$$Entropy(D) = -(\frac{3}{5}log_2(\frac{3}{5}) + \frac{2}{5}log_2(\frac{2}{5})) = .529 + .442 = .971$$

## Split 1: 4.5

Starting off, we split at 4.5 ((5+4)/2). Now we get two bins, as follows:

| | A on Test | Lower than A |
|---|---|---|
| <=4.5 | 0 | 1 |
| >4.5 | 3 | 1 |

$$Entropy(D_{<=4.5}) = -(\frac{1}{1}log_2(1) + 0log_2(0)) = 0 + 0 = 0$$

$$Entropy(D_{>4.5}) = -(\frac{3}{4}log_2(\frac{3}{4}) + \frac{1}{4}log_2(\frac{1}{4})) = .311 + .5 = .811$$

Now net entropy is:

$$Info_a(D_{new}) = \frac{1}{5}(0) + \frac{4}{5}(.811) = .6488$$

# Practice Problem on Binning

- Given are the recorded temperature in a day

  10,18,22,25,24,21,20,20,19,18,15.

Discretize using equi width binning and equi frequency binning with k=4

# Practice Problem on Entropy

- For the given data set calculate the entropy on the target variable
- For the given data set calculate the entropy of the predictor variable with a possible split of x<=17 and x>17
- Calculate the information gain on (21,S).

| Predictor variable x | Target variable y |
|---|---|
| 4 | M |
| 12 | M |
| 16 | M |
| 18 | F |
| 24 | F |
| 26 | F |

# Practice Problem on Quantile

- Find the quantile and IQR for the following dataset without applying correction factor for the recorded temperature in a day.

  10,18,22,25,24,21,20,20,19,18,15

# Feature Extraction, Selection and Construction

- Dimensionality reduction
  - Goal: represent instances with fewer variables
  - Try to preserve as much as structure in the data as possible
- Feature Extraction
  - Construct new set of dimensions ei=f(x1…xd)
  - (linear)combinations of original
- Can be done using Feature mapping algorithms
  - Feed forward networks – (Feature Extraction)
  - Principal component analysis ( Dimensionality reduction)

- Feed forward networks – preserve the features of the first hidden layer in neural network

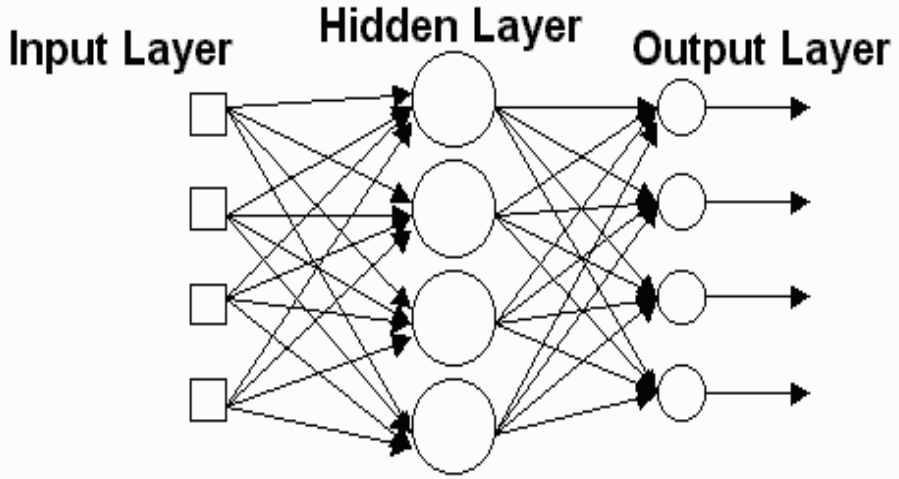



**Figure 2** The anatomy of a neural network.

- Principal component analysis is a method of extracting important variables (in form of components) from a large set of variables available in a data set.
- It extracts low dimensional set of features from a high dimensional data set with a motive to capture as much information as possible.
- With fewer variables, visualization also becomes much more meaningful. PCA is more useful when dealing with 3 or higher dimensional data.

- It is always performed on a symmetric correlation or covariance matrix. This means the matrix should be numeric and have standardized data.
- Feature Extraction by PCA

| Covariance matrix | | | | Eigen Values | Proportion |
|---|---|---|---|---|---|
| 1 | -0.1094 | 0.8718 | 0.818 | 2.9108 | 0.7277 |
| -0.1094 | 1 | -0.4205 | -0.3565 | 0.9212 | 0.2303 |
| 0.8718 | -0.4205 | 1 | 0.9628 | 0.1474 | 0.0368 |
| 0.818 | -0.3565 | 0.9628 | 1 | 0.0206 | 0.0052 |

# Missing data

- Missing data are the data we desired to collect but never got into our database for subsequent analysis
- Reasons for missing data
  - Respondents refusal to answer an item
  - Respondents does not know the answer
  - Data not applicable
- Consequence : inaccurate conclusion

- Three classes of missing data
  - Missing completely at random (MCAR) – the probability of missing data on Y is not dependent on Y or X
  - Missing at random (MAR) – the probability of missing data on Y depends on the value of X
  - Not missing at random(NMAR) – probability of missing data on Y is dependent on value of Y

- Handling missing values
  - Eliminate data objects
  - Estimate missing values
    - Missing data substitution (mean, median, specific value, regression substitution)
    - K-nearest neighbours
  - Ignore the missing values during analysis
  - Replace with all possible values(weighted by their probabilities)

# Post Processing

- Knowledge filtering
- Interpretation and explanation
- Evaluation
- Knowledge integration