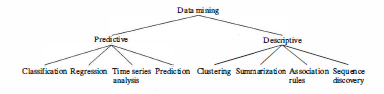
**1.1 Data Mining Tasks:**

Data Mining involves several algorithms to accomplish a particular data

The basic outline of the tasks can be viewed as the following

****

**Predictive Model:** A predictive model makes a prediction using known results found from different data.

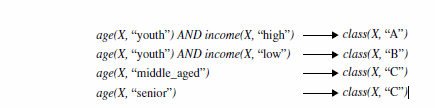
1. **Classification**

**Classification** is the process of finding a **model.** The model is derived based on the analysis of a set of **training data** (i.e., data objects for which the class labels are known). The model is used to predict the class label of objects for which the class label is unknown.

**Supervised learning** is basically a synonym for classification.

The derived model may be represented in various forms, such as classification rules (i.e., IF THEN rules), decision tree and neural networks

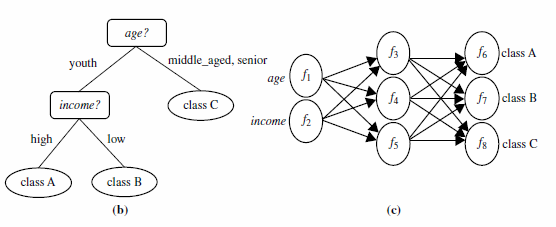
*IF-THEN rule*



**Decision Tree and Neural Network**

A **decision tree** is a flowchart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions.

A **neural network**, is typically a collection of neuron-like processing units with weighted connections between the units.



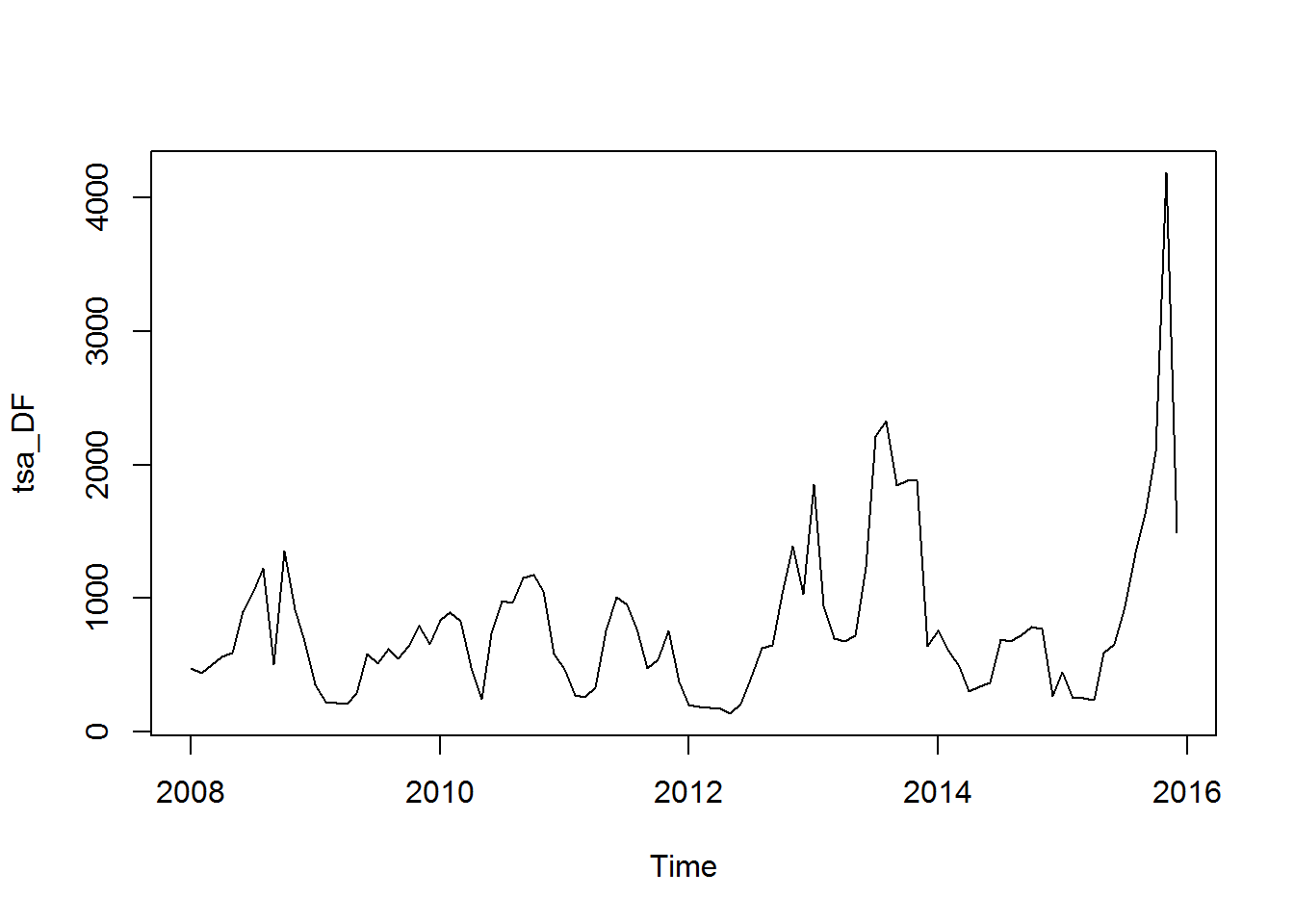
Other methods for classification:

* Naïve Bayesian classification
* Support vector machines,
* and *k*-nearest-neighbor classification

1. **Regression:**  Regression is used to predict missing or unavailable *numerical data values* rather than (discrete) class labels.
2. **Time series analysis:**

Time series analysis is a [statistical technique](http://www.statisticssolutions.com/directory-of-statistical-analyses/) that deals with time series data, or trend analysis.  Time series data means that data is in a series of particular time periods or intervals. A time series plot is used to visualize the time series.

Dengue time series project



1. **Prediction:** Prediction can be viewed as a type of classification. Prediction is predicting a future state rather than a current state.

**Example: Predicting flooding**

One approach is placing the sensors at various points in the river. These sensors collect the data water level, rain amount, time, humidity and so on. Then by comparing the water levels with respect to time can predict the flooding.

**Descriptive model:** Descriptive model identifies patterns and relationships in the data

1. **Clustering:** The method of identifying similar groups of **data** in a **data** set is called **clustering**. **Clustering** analyzes data objects without consulting class labels. **Unsupervised learning** is essentially a synonym for clustering.

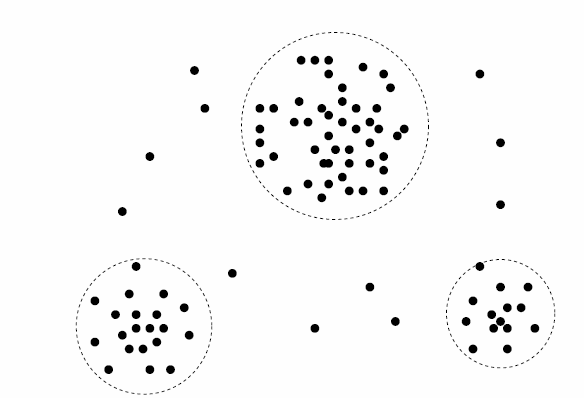
Objects within a cluster have high similarity in comparison.

**Examples for clustering**

**Example 1**



**Example 2**



2D Plot of customer data with respect to customer location

**Segmentation:** A special type of clustering is called segmentation. With segmentation a database is partitioned into disjointed groupings of similar tuples called segments.

1. **Association Rule:** An association rule is a model that identifies specific types or patterns of data associations. These associations are often used in the retail sales community to identify items that are frequently purchased together.
2. **Sequential analysis** is used to determine sequential patterns in data. These patterns are based on a time sequence of actions. These patterns are similar to associations in that data (or events) are found to be related, but the relationship is based on time.

For example, most people who purchase CD players may be found to purchase CDs within one week.

1. **Summarization** It extracts or derives representative information about the database.

Alternatively, summary type information can be derived from the data.

**1.2 Data Mining Issues:**

There are many important implementation issues associated with data mining:

1. Human interaction: Technical experts are used to formulate the queries and assist in interpreting the results. Users are needed to identify training data and desired results.

2. Over fitting: When a model is generated that is associated with a given database state it is desirable that the model also fit future database states. Over fitting occurs when the model does not fit future states.

3. Outliers: The data entries that do not fit nicely into the derived model are called Outlier. This becomes even more of an issue with very large databases. If a model is developed that includes these outliers, then the model may not behave well for data that are not outliers .

4. Interpretation of results: Currently, data mining output may require experts to correctly interpret the results.

5. Visualization of results: To easily view and understand the output of data mining algorithms, visualization of the results is helpful.

6. Large datasets: When a model (Designed for the small data sets) is applied to the large data sets, they may produce inefficient results. Sampling and parallelization are effective tools to attack this scalability.

7. High dimensionality: If a data base contains too many attributes then it is called a high Dimensional data set. The use of other attributes may simply increase the overall complexity and decrease the efficiency of an algorithm.

This problem is sometimes referred to as the dimensionality curse, meaning that there are many attributes (dimensions) involved and it is difficult to determine which ones should be used.

One solution to this high dimensionality problem is to reduce the number of attributes, which is known as dimensionality reduction.

8. Multimedia data: Most previous data mining algorithms are targeted to traditional data types (numeric, character, text, etc.).

9. Missing data: During the preprocessing phase of KDD, missing data may be replaced with estimates. This and other approaches to handling missing data can lead to invalid results in the data mining step.

10. Irrelevant data: Some attributes in the database might not be of interest to the data mining task being developed.

11. Noisy data: Some attribute values might be invalid or incorrect. These values are often corrected before running data mining applications.

12. Changing data: Databases cannot be assumed to be static. However, most data mining algorithms do assume a static database. This requires that the algorithm be completely rerun anytime the database changes.

13. Integration: The KDD process is not currently integrated into normal data processing activities. KDD requests may be treated as special, unusual, or one-time needs. This makes them inefficient, ineffective, and not general enough to be used on an ongoing basis. Integration of data mining functions into traditional DBMS systems is certainly a desirable goal.

14. Application: Determining the intended use for the information obtained from the data mining function is a challenge. Indeed, how business executives can effectively use the output is sometimes considered the more difficult part, not the running of the algorithms themselves. Because the data are of a type that has not previously been known, business practices may have to be modified to determine how to effectively use the information uncovered.

These issues should be addressed by data mining algorithms and products

**Data Mining Metrics:**

**Metrics** are parameters or measures of quantitative assessment used for measurement, comparison, or to track performance.

**Example:**

Return On Investment (ROI) is used in business oriented industries.

Time and space complexity is used in estimating the performance of algorithms

Accuracy can be used for the classification.

**Metrics for Supervised Learning**

**A confusion matrix** is a table that is often used to **describe the performance of a classification model** (or "classifier") on a set of test data.

* **True positives (TP):** These are cases in which we predicted yes (they have the disease), and they do have the disease.
* **True negatives (TN):** We predicted no, and they don't have the disease.
* **False positives (FP):** We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")
* **False negatives (FN):** We predicted no, but they actually do have the disease. (Also known as a "Type II error.")

**Accuracy:** Overall, how often is the classifier correct?

* + (TP+TN)/total

**Error Rate: Misclassification Rate:** Overall, how often is it wrong?

* + (FP+FN)/total
  + equivalent to 1 minus Accuracy

**Recall: or True Positive Rate or Sensitivity**

When it's actually yes, how often does it predict yes?

**Recall** =TP/actual yes

**Precision: When** it predicts yes, how often is it correct?

* + Precision=TP/predicted yes

**Specificity: (True Negative Rate)** When it's actually no, how often does it predict no?

* + Specificity=TN/actual no

**ROC Curve:** This is a commonly used graph that summarizes the performance of a classifier over all possible thresholds. It is generated by plotting the True Positive Rate (y-axis) against the False Positive Rate (x-axis)

**Example:**



* **Accuracy:** Overall, how often is the classifier correct?
  + (TP+TN)/total = (100+50)/165 = 0.91
* **Error Rate: Misclassification Rate:** Overall, how often is it wrong?
  + (FP+FN)/total = (10+5)/165 = 0.09
  + equivalent to 1 minus Accuracy
  + also known as "Error Rate"
* **Recall: True Positive Rate or sensitivity** When it's actually yes, how often does it predict yes?
  + TP/actual yes = 100/105 = 0.95
* **Specificity: True Negative Rate:** When it's actually no, how often does it predict no?
  + TN/actual no = 50/60 = 0.83
  + equivalent to 1 minus False Positive Rate
* **Precision:** When it predicts yes, how often is it correct?
  + TP/predicted yes = 100/110 = 0.91

**Metrics for Unsupervised Learning:**

**Squared Sum Error: SSE is typically used in clustering algorithms to measure the quality of clusters obtained.**

**If SSE is small when the points are close to their cluster center, indicating a good clustering. If it is high it indicates a poor clustering.**

**Data Mining From a Database Perspective:**

The study of data mining from a database perspective involves looking at all types of data mining applications and techniques. We are concerned about the following implementation issues:

• **Scalability:** Algorithms that do not scale up to perform well with massive real world datasets are of limited application. Related to this is the fact that techniques should work regardless of the amount of available main memory.

• **Real-world data:** Real-world data are noisy and have many missing attribute values. Algorithms should be able to work even in the presence of these problems.

• **Update:** Many data mining algorithms work with static datasets. This is not a realistic assumption.

• **Ease of use**: Although some algorithms may work well, they may not be well received by users if they are difficult to use or understand.