MODULE -1

Data mining is defined as the process of discovering patterns in data either automatically or semi-automatically.

**Structural patterns:** The result of data mining is a structural pattern that provides useful and meaningful information. Typical popular structures are rules, decision trees, and clusters. Part 3 of this module provides details on the output structures.

What is not data mining and what is not?

* Looking up phone numbers in phone directory is NOT data mining; finding names that are popular in different locations
* Querying a web search engine for information about say “Big data” is NOT data mining; grouping returned documents based on similar occurring terms (called data classification)

**Instance:**

* This is the most common form of input to data mining tasks
* This is often referred to as sample data in statistics
* Data is expressed as a table of instances
* It could be expressed in different formats

Relation (e.g., sister-of relation)

**Attribute:**

* Each instance consists of set of features or attributes
* There could be fixed or variable number of attributes in an instance (continuous, integer, real)
* An attribute could be numeric or categorical (categorical, discrete, nominal, ordinal, enumerated, boolean)
* Examples: temperature=35 (numeric),  outlook = {sunny, overcast, rainy} (categorical, enumerated) play = {yes, no} (categorical, boolean)
* It is optional for the categorical values of an attribute to be ordered sunny>overcast>rainy) or no

**Concept:**

* Result of learning from data mining applications
* Examples
  + Classification learning
    - model to classify a given instance into a class
  + Association learning
    - associations among attributes/features
  + Clustering---grouping instances
  + Numeric prediction
    - Regression
* Tables
  + Example, weather data with exhaustive values
  + Linear models---regression equation---a linear equation with LHS being the predicted value and RHS being a linear equation involving the input attributes
  + May be used in binary classification with decision boundary specified for the output---points above the boundary are one class and the ones below are the second class.
* Trees
  + Nodes where an attribute is tested
  + Branches go out from a node
  + The leaf nodes indicate the final decision
  + When comparing enumerated (or nominal) values, the number of branches is usually the number of values it takes
  + When numeric attributes are tested, we could test based on <, =, and >
  + **Missing values are to be handled in a special way**
  + **For predicting numerical values, the leaf nodes are regression equations instead of classes, and called model trees**
  + **When the leaf node is an average value along that branch, it is called a regression tree**
* Rules
  + Popular alternatives to decision trees
  + **The LHS of a rule is the antecedent or precondition is a series of tests ANDed together**
  + Ex: if x=1 and y=0 then class =a
  + **The RHS is the consequent or conclusion**
  + In the above, class is the outcome or result attribute
* Types of rules
  + Classification rules: consequent is the class or outcome attribute
    - Ex: outlook=sunny and temperature=hot and humidity=high then play=no
  + Association rules: Similar to classification rules except that consequent may be any attribute and not necessarily the class or outcome attribute
    - If temperature = cool then humidity=normal
  + Rules with exceptions---an extension to classification rules—more like if then else in programming languages
    - Ex: If petal-length ≥ 2.45 and petal-length < 4.45 then iris-versicolor EXCEPT if petal-width < 1.0 then Iris-setosa
* Instance-based representation:
  + The output is a table of instances with a class
  + An unknown instance’s class is predicted by comparing it with those closest to it in the table
    - Ex: nearest-neighbor, k-nearest neighbor classification
* Cluster---instances are grouped into clusters
  + Each cluster is associated with several instances
  + In some clustering methods, clusters might overlap in which case an instance could belong to multiple clusters