Unit 5

Data Mining Techniques

Data Mining Definition

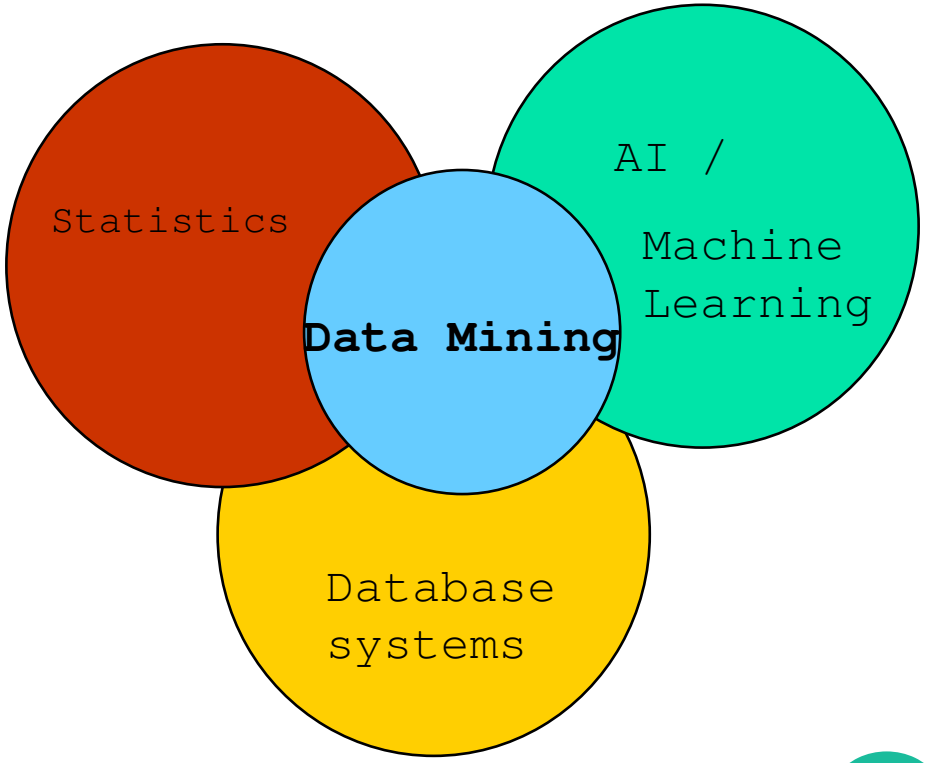
Refer Chapter 1

Origins of Data Mining

Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems

Must address:

- **Greatness of data**
- **High dimensionality of data**
- **Heterogeneous, distributed nature of data**



Why Data Preprocessing?

- Data in the real world is dirty
  - incomplete: lacking *attribute values*, lacking certain *attributes of interest*, or containing only aggregate data
  - noisy: containing errors or outliers
  - inconsistent: containing discrepancies in codes or names
- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
  - Data warehouse needs consistent integration of quality data
  - Required for both OLAP and Data Mining!

Why can Data be Incomplete?

- Attributes of interest are not available (e.g., customer information for sales transaction data)
- Data were not considered important at the time of transactions, so they were not recorded!
- Data not recorded because of misunderstanding or malfunctions
- Data may have been recorded and later deleted!
- Missing/unknown values for some data

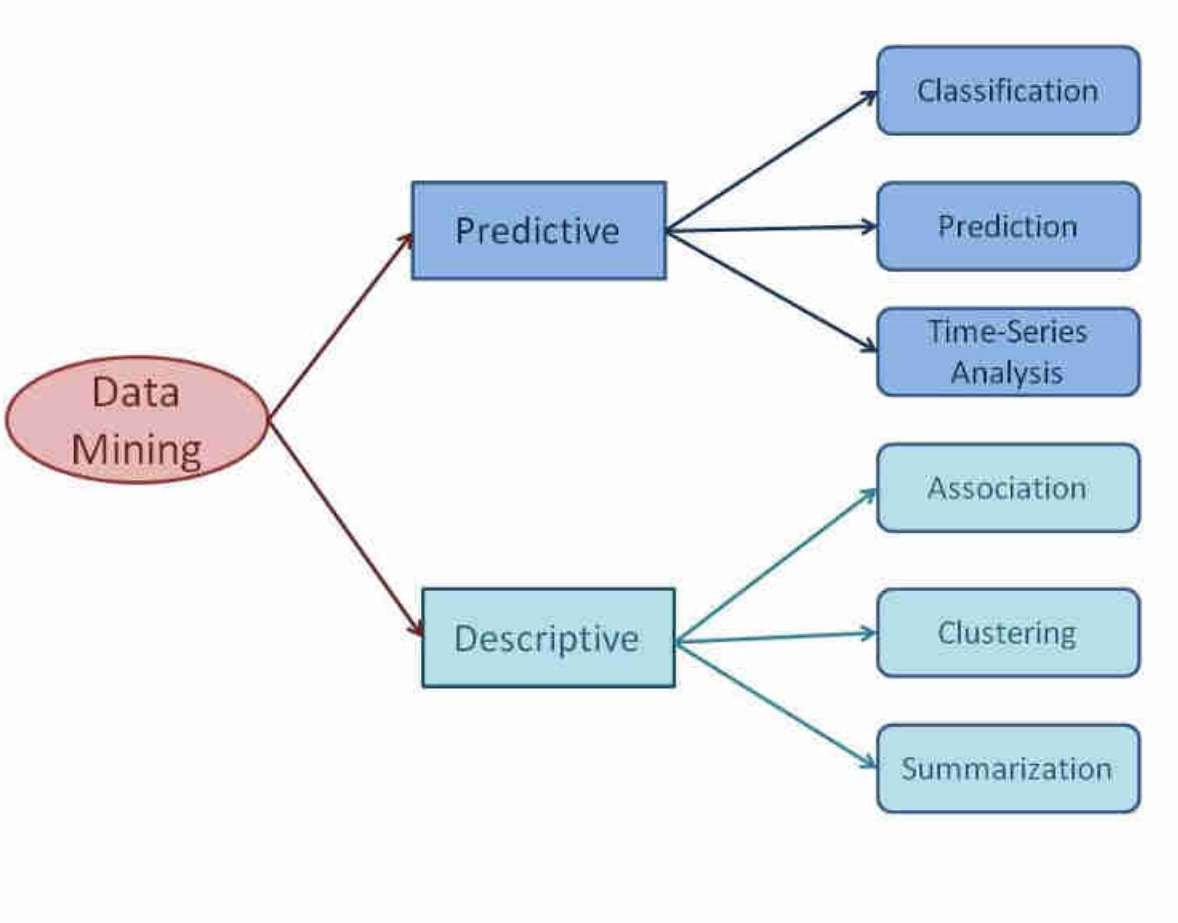
Data Cleaning

- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data

Data Mining Tasks

For a given dataset D, language of facts L, interestingness function $I_{D,L}$ and threshold c, find the expression E such that $I_{D,L}(E) > c$ efficiently.

Data Mining Tasks



Data Mining Tasks

1. **Classification:**
   learning a function that maps an item into one of a set of predefined classes

2. **Regression/Prediction:**
   learning a function that maps an item to a real value

3. **Time – Series Analysis**
   sequence of events where the next event is determined by one or more of the preceding events

Data Mining Tasks

4. **Clustering:**
   - identify a set of groups of similar items

5. **Dependencies and associations:**
   - identify significant dependencies between data attributes

6. **Summarization:**
   - find a compact description of the dataset or a subset of the dataset

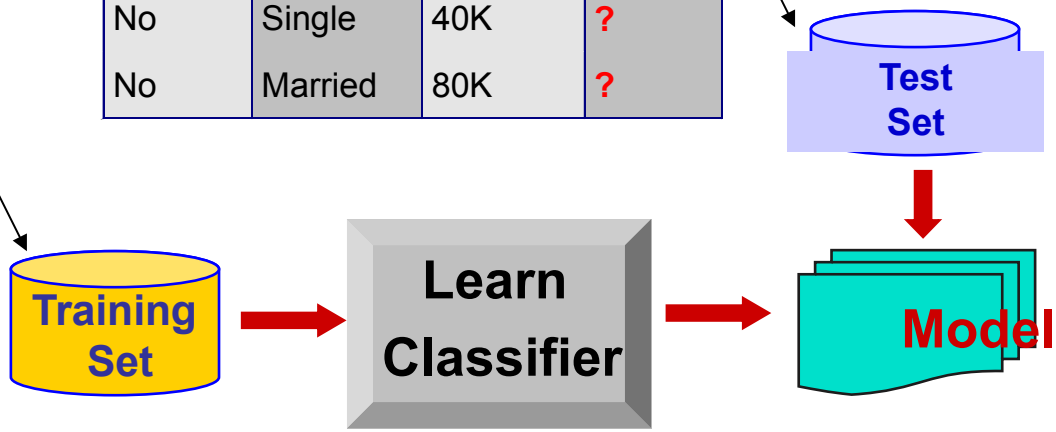
Classification: Definition

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification Example

Tid	Home Owner (categorical)	Marital Status (categorical)	Taxable Income (continuous)	Default (class)
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

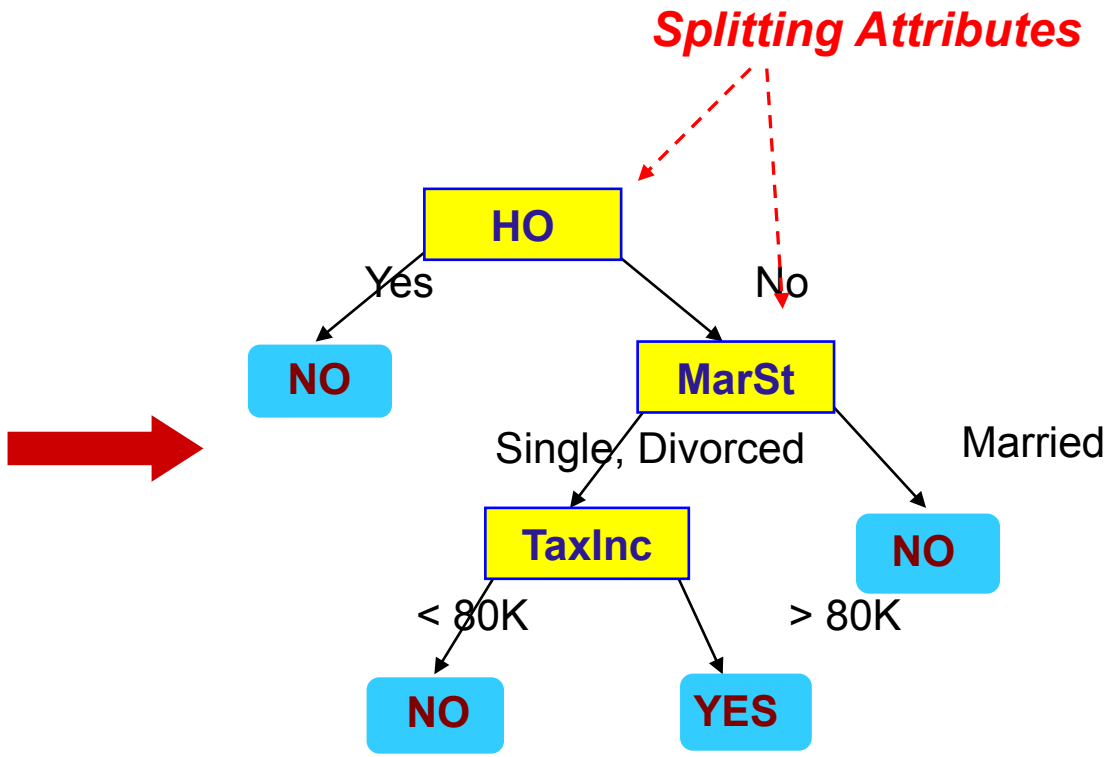
Home Owner	Marital Status	Taxable Income	Default
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification Example

Tid	Home Owner (categorical)	Marital Status (categorical)	Taxable Income (continuous)	Default (class)
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

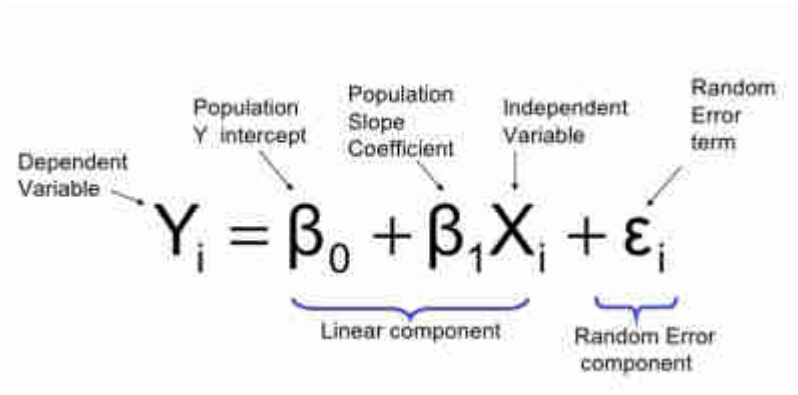
Training Data

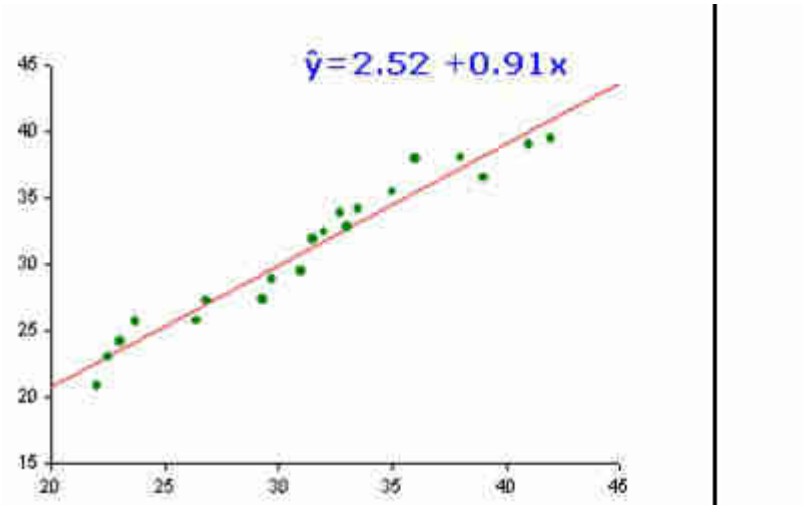


Model: Decision Tree

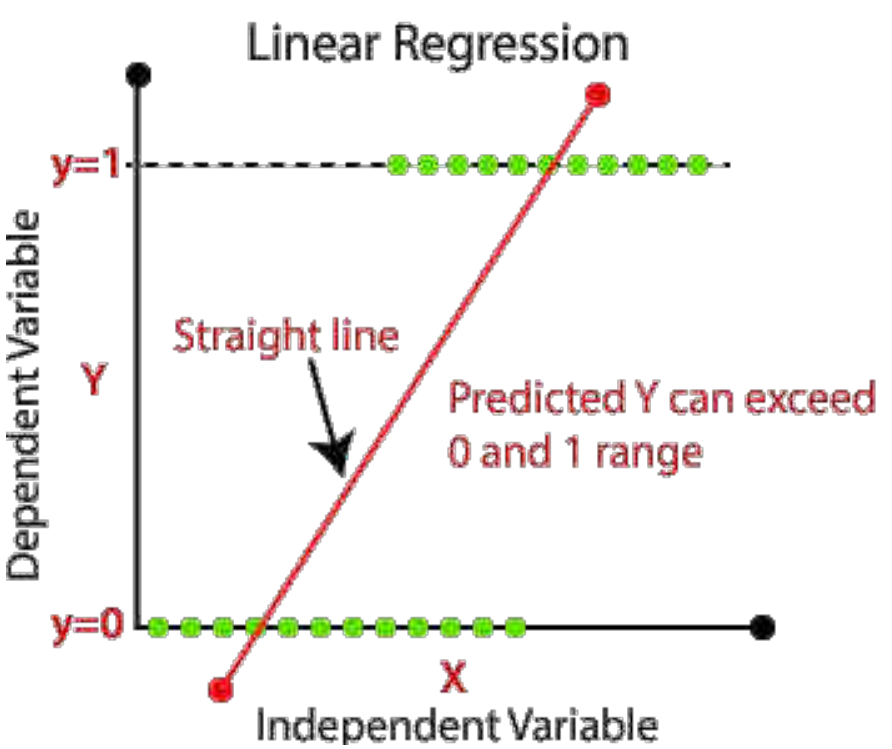
Regression

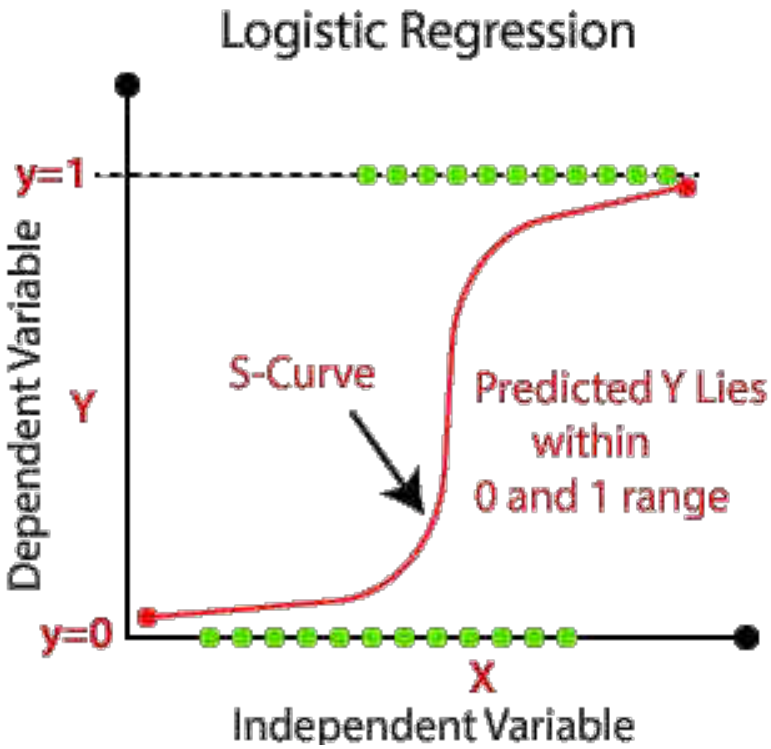
- Regression can be defined as a data mining technique that is generally used for the purpose of predicting a range of continuous values (which can also be called "numeric values") in a specific data set.
  - For example, Regression can predict sales, profits, temperature, distance and so on





Regression → Types





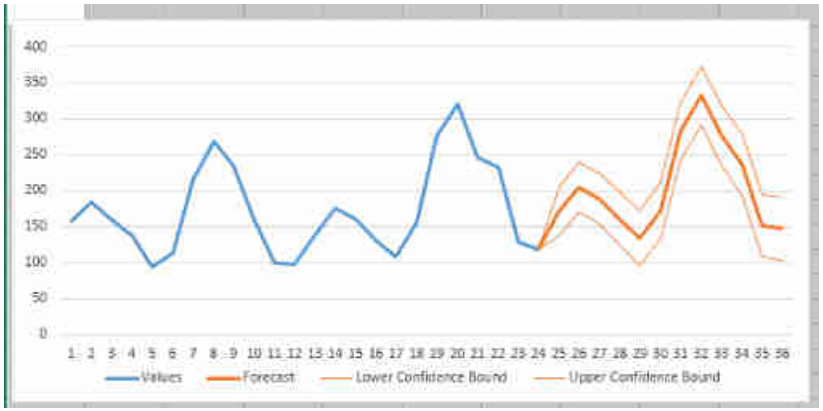
Regression VS Classification

Classification	Regression
Classification is mainly used for allotting the given data into discrete categories.	Regression is particularly used for the aim of predicting numeric or continuous values.
The nature of the predicted data is unordered.	The nature of the predicted data is ordered.
In the method of classification, the calculations are usually done by measuring the accuracy.	In the method of Regression, the calculations are usually done by using the root mean square error.
Classification can be further classified into the binary classifier and multi- class classifier.	The algorithm can be further classified into linear Regression and non- linear Regression.
Examples of classification are decision tree, logistic Regression and many others.	The examples of Regression can be Random forest or Regression tree, linear regression, to name a few.

Time Series Analysis

Time series is a sequence of events where the next event is determined by one or more of the preceding events.

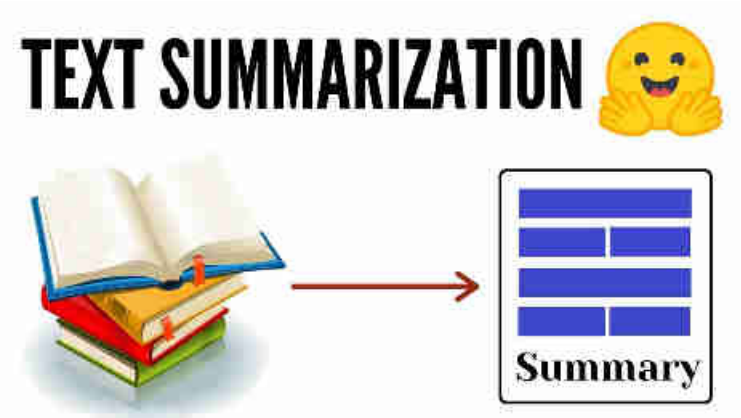
Time series reflects the process being measured and there are certain components that affect the behavior of a process. Time series analysis includes methods to analyze time-series data in order to extract useful patterns, trends, rules and statistics. Stock market prediction is an important application of time- series analysis.



Summerization

Summarization is the generalization of data. A set of relevant data is summarized which result in a smaller set that gives aggregated information of the data.

For example, the shopping done by a customer can be summarized into total products, total spending, offers used, etc. Such high level summarized information can be useful for sales or customer relationship team for detailed customer and purchase behavior analysis.



Clustering

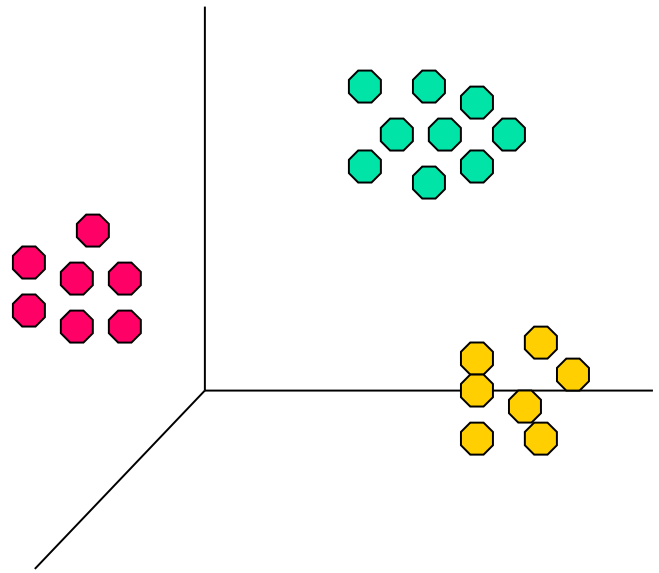
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Clustering → Illustration

Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



What is the difference between KDD and Data mining?

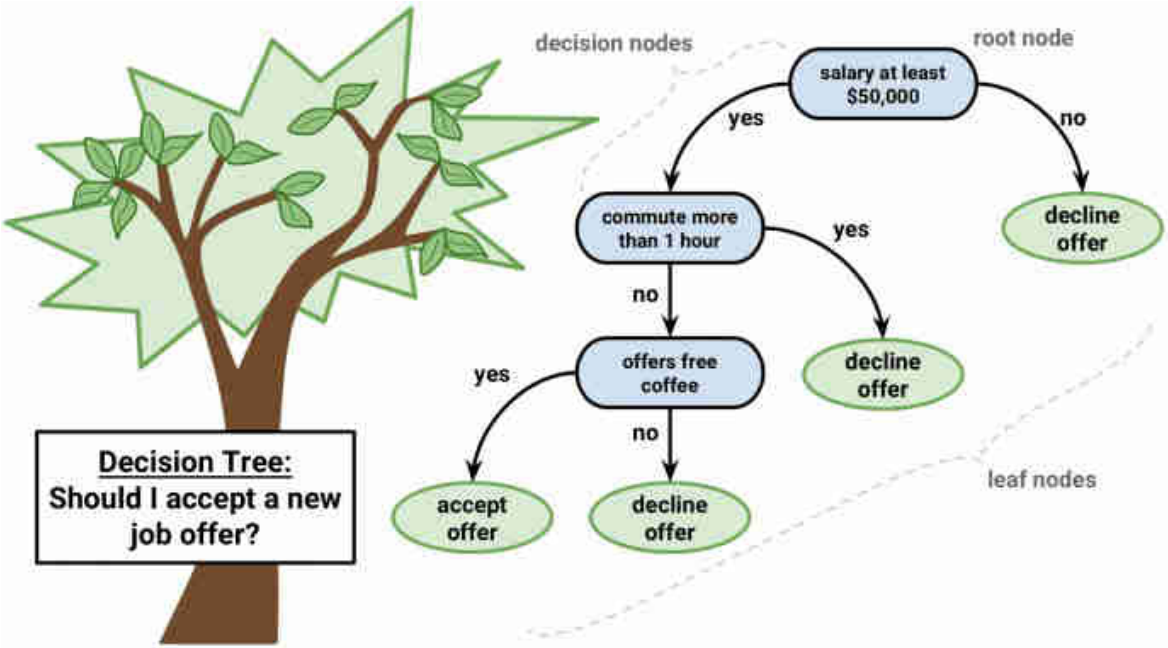
- Although, the two terms KDD and Data Mining are heavily used interchangeably, they refer to two related yet slightly different concepts.
- KDD is the overall process of extracting knowledge from data while Data Mining is a step inside the KDD process, which deals with identifying patterns in data.
- In other words, Data Mining is only the application of a specific algorithm based on the overall goal of the KDD process.

Data Mining Methods

1. Decision Tree Classifiers:
   - Used for modeling, classification
2. Association Rules:
   - Used to find associations between sets of attributes
3. Sequential patterns:
   - Used to find temporal associations in time series
4. Hierarchical clustering:
   - used to group customers, web users, etc

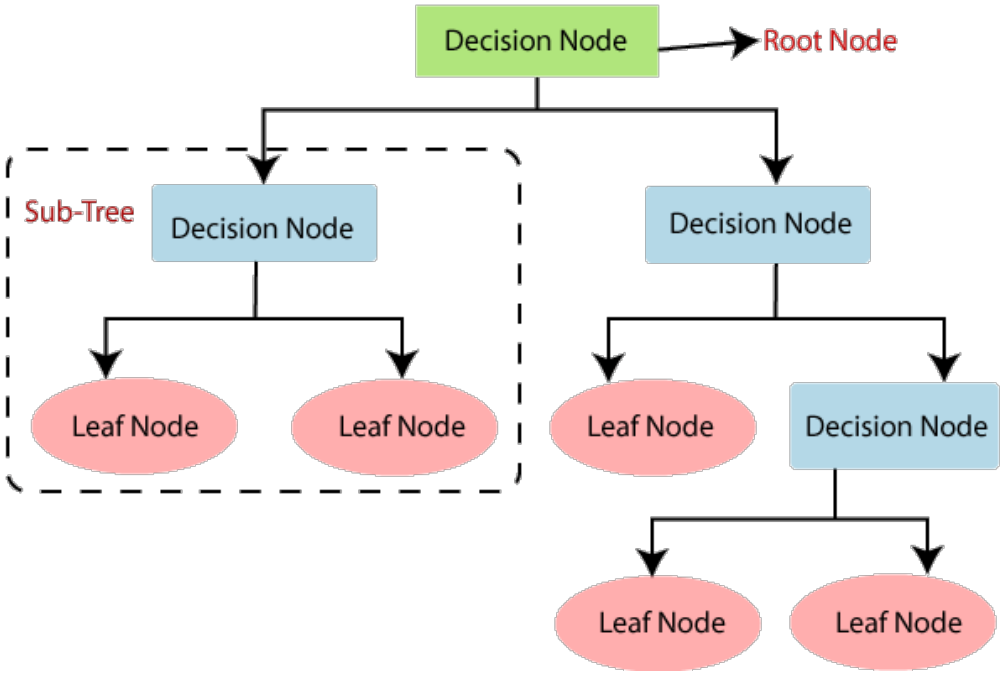
Decision Tree Classifiers

- Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving classification problems.



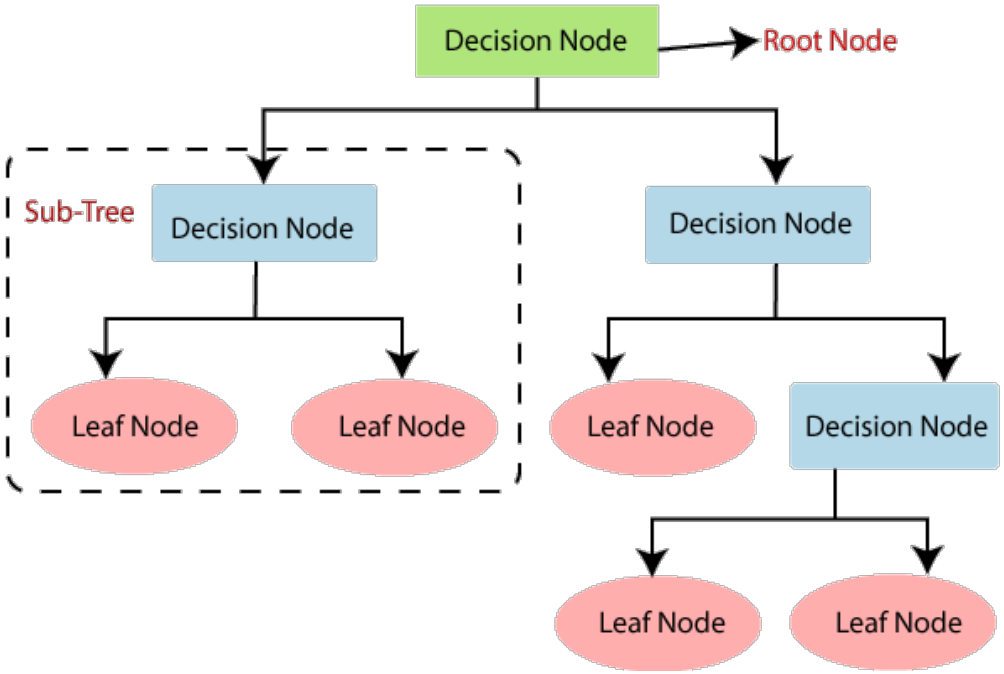
Decision Tree Classifiers

- Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems.
- It is a tree-structured classifier, where internal nodes represent the features of a data set, branches represent the decision rules and each leaf node represents the outcome.



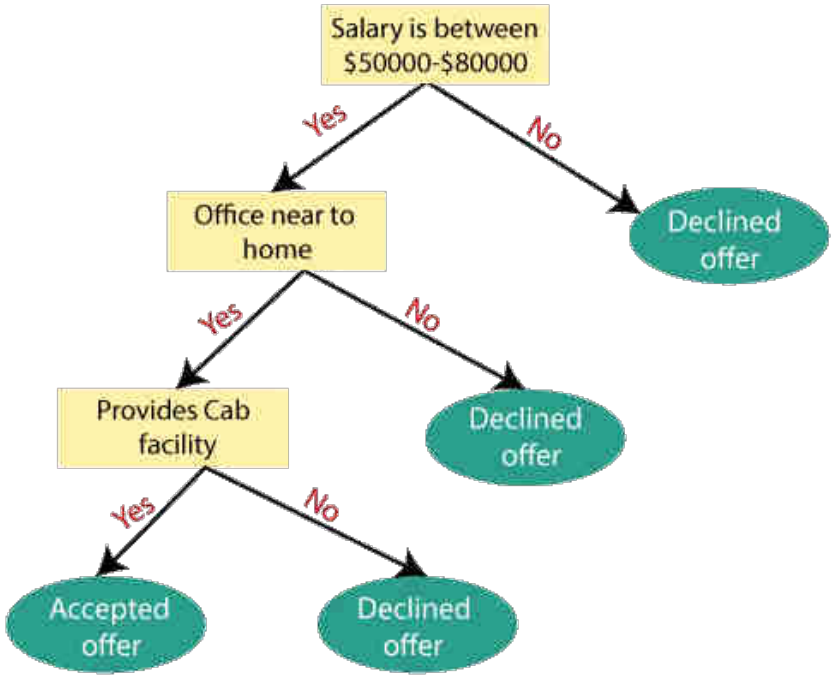
Decision Tree Classifiers

- It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.
- It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
- In order to build a tree, we use the **CART algorithm**, which stands for Classification and Regression Tree algorithm.



Decision Tree Classifiers → Example

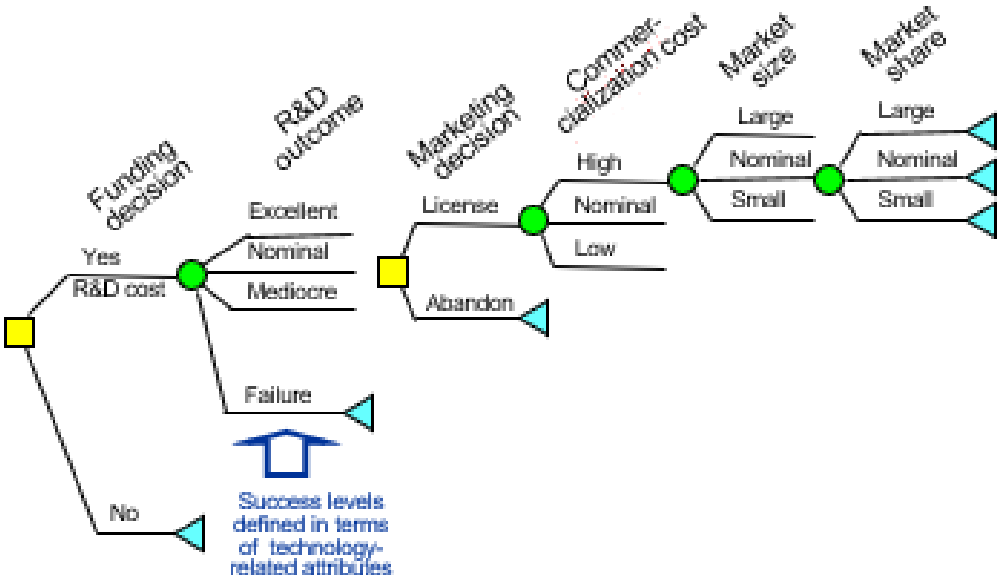
Suppose there is a candidate who has a job offer and wants to decide whether he should accept the offer or Not. So, to solve this problem, the decision tree starts with the root node (Salary attribute by ASM). The root node splits further into the next decision node (distance from the office) and one leaf node based on the corresponding labels. The next decision node further gets split into one decision node (Cab facility) and one leaf node. Finally, the decision node splits into two leaf nodes (Accepted offers and Declined offer). Consider the below diagram:



Decision Tree Classifiers → Application

Marketing:

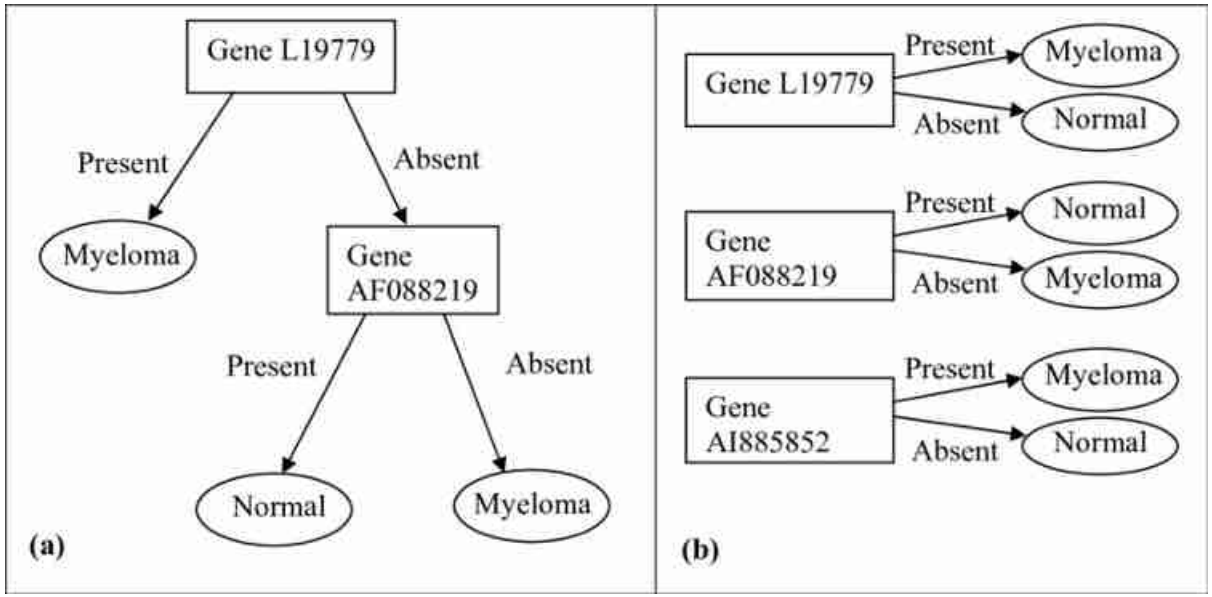
Businesses can use decision trees to enhance the accuracy of their promotional campaigns by observing the performance of their competitors products and services. Decision trees can help in audience segmentation and support businesses in producing better-targeted advertisements that have higher conversion rates.



Decision Tree Classifiers → Applications

Diagnosis of Diseases:

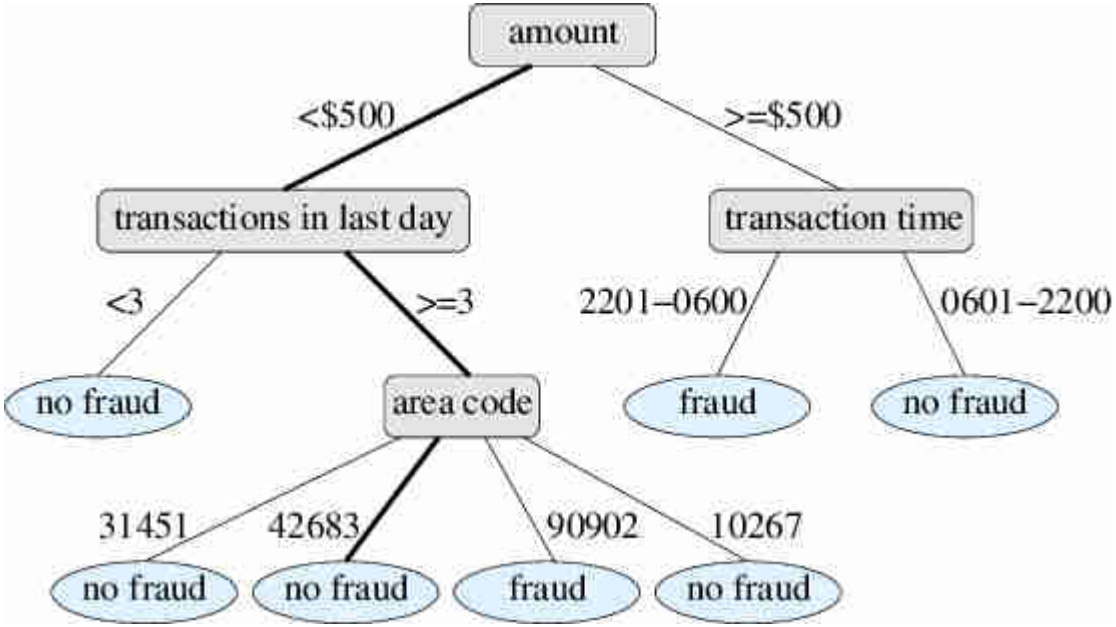
Decision trees can help physicians and medical professionals in identifying patients that are at a higher risk of developing serious (or preventable) conditions such as diabetes or dementia. The ability of decision trees to narrow down possibilities according to specific variables is quite helpful in such cases.



Decision Tree Classifiers → Application

Detection of Frauds:

Companies can prevent fraud by using decision trees to identify fraudulent behavior beforehand. It can save companies a lot of resources, including time and money.



Association Rule Learning

- Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable.
- It tries to find some interesting relations or associations among the variables of data set.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

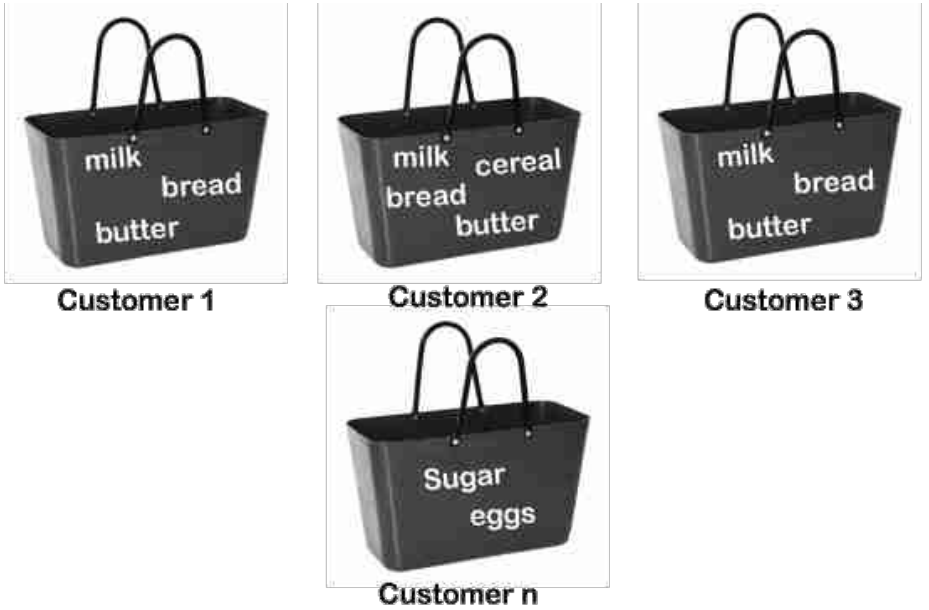
Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Association Rule Learning → Example

- For example, if a customer buys bread, he most likely can also buy butter, eggs, or milk, so these products are stored within a shelf or mostly nearby.



Association Rule Learning → Application

Market Basket

Association Rules are widely used to analyze retail basket or transaction data, and are intended to identify strong rules discovered in transaction data using measures of interestingness, based on the concept of strong rules.

$$Support = \frac{frq(X,Y)}{N}$$

$$Rule: X \Rightarrow Y \rightarrow Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

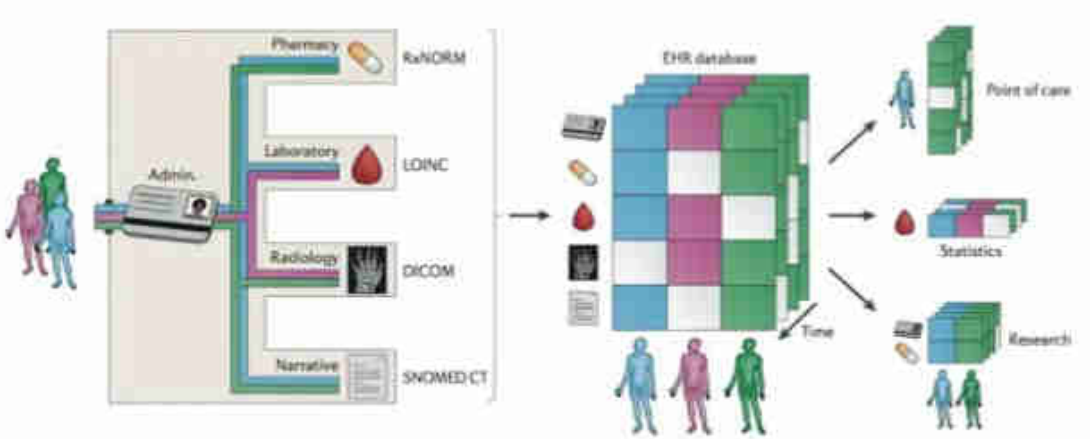
Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

Association Rule Learning → Application

Medical Diagnosis:

Using relational association rule mining, we can identify the probability of the occurrence of illness concerning various factors and symptoms.

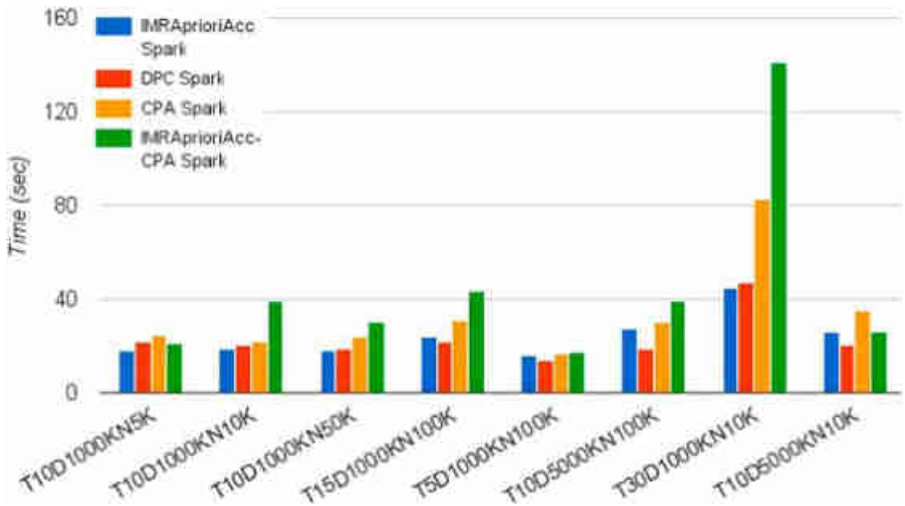
Further, using learning techniques, this interface can be extended by adding new symptoms and defining relationships between the new signs and the corresponding diseases.



Association Rule Learning → Application

Census Data:

Every government has tonnes of census data. This data can be used to plan efficient public services(education, health, transport) as well as help public businesses (for setting up new factories, shopping malls, and even marketing particular products). This application of association rule mining and data mining has immense potential in supporting sound public policy and bringing forth an efficient functioning of a democratic society.



Sequential Pattern Mining

Given a set of sequences and support threshold, find the complete set of frequent sub sequences

A *sequence* : <(ef)(ab)(df)cb>

A *sequence database*

SID	sequence
10	<a(abc)(ac)d(cf)>
20	<(ad)c(bc)(ae)>
30	<(ef)(ab)(df)cb>
40	<eg(af)cbc>

An element may contain a set of items.
Items within an element are unordered and we list them alphabetically.

<a(bc)dc> is a *subsequence* of <a(abc)(ac)d(cf)>

Given *support threshold* *min_sup* =2, <(ab)c> is a *sequential pattern*

Challenges on Sequential Pattern Mining

- A huge number of possible sequential patterns are hidden in databases
- A mining algorithm should
  - find the complete set of patterns, when possible, satisfying the minimum support threshold
  - be highly efficient, scalable, involving only a small number of database scans
  - be able to incorporate various kinds of user specific constraints

Hierarchical Clustering

- Hierarchical clustering refers to an unsupervised learning procedure that determines successive clusters based on previously defined clusters. It works via grouping data into a tree of clusters.

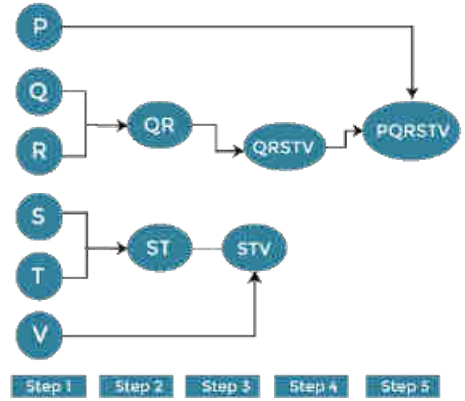
- The endpoint refers to a different set of clusters, where each cluster is different from the other cluster, and the objects within each cluster are the same as one another.

Hierarchical Clustering

There are two types of hierarchical clustering

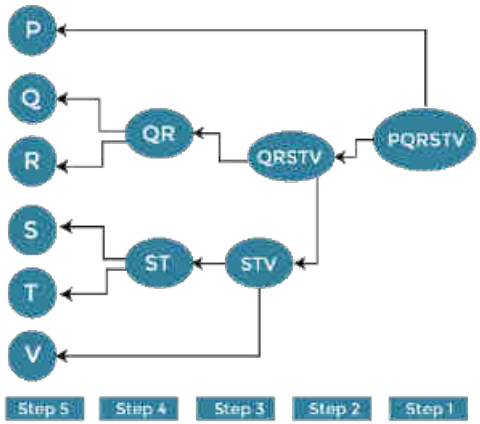
- **Agglomerative Hierarchical Clustering**

  Each data point act as an individual cluster and at each step, data objects are grouped in a bottom-up method. Initially, each data object is in its cluster.



- **Divisive Clustering**

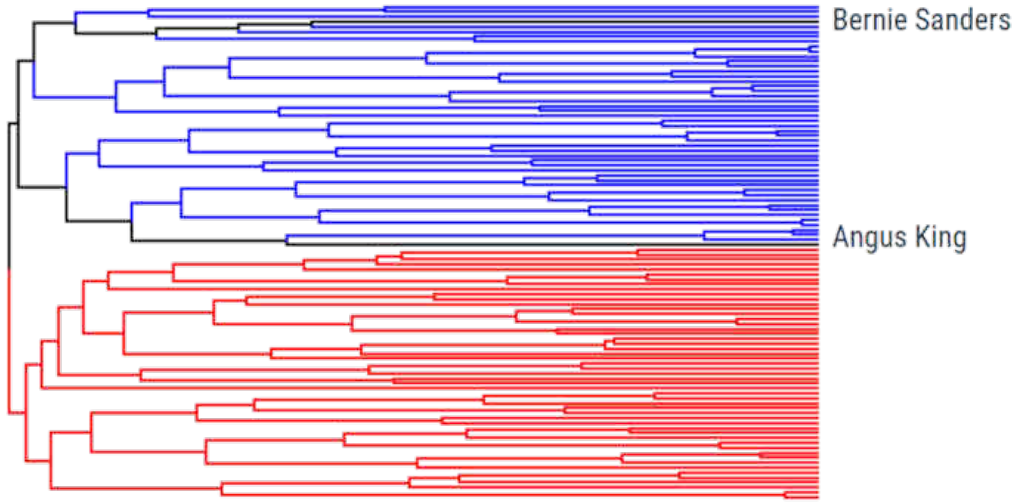
  All the data points are considered an individual cluster, and in every iteration, the data points that are not similar are separated from the cluster. The separated data points are treated as an individual cluster.



Hierarchical Clustering → Application

US Senator Clustering through Twitter

In this example, we use Twitter to cluster US senators into their respective parties. Our data is simple: we only look at which senators follow which senators. That defines a graph structure with senators as the nodes and follows as the edges.

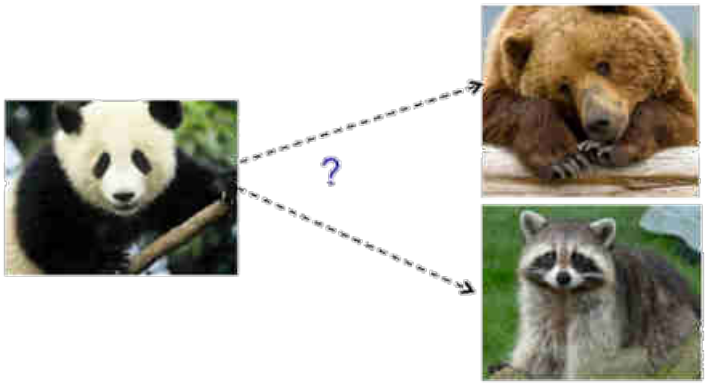


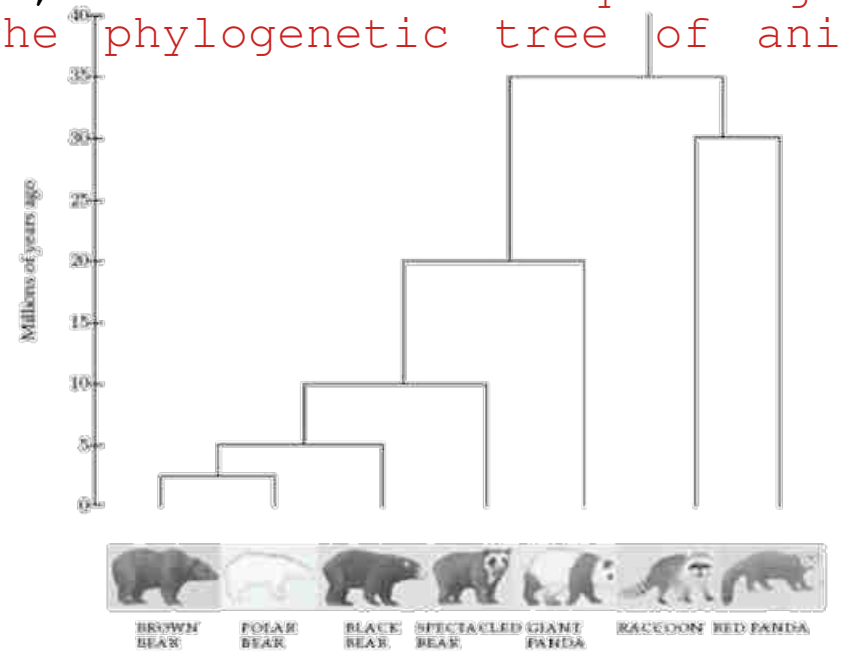
Reds are Republicans, Blues are Democrats, Blacks are independent

Hierarchical Clustering → Application

Charting Evolution through Phylogenetic Trees

In the decades before DNA sequencing was reliable, the scientists struggled to answer a seemingly simple question: Are giant pandas closer to bears or racoons? Nowadays, we can use DNA sequencing and hierarchical clustering to find the phylogenetic tree of animal evolution





Data Mining Tools

Data Mining tools have the objective of discovering patterns/ trends/groupings among large sets of data and transforming data into more refined info.

Data Mining Tools → Orange Data Mining

- It supports the visualization and is a software-based on components written in Python computing language and developed at the bioinformatics laboratory at the faculty of computer and information science, Ljubljana University, Slovenia.
- As it is a software-based on components, the components of Orange are called "widgets." These widgets range from preprocessing and data visualization to the assessment of algorithms and predictive modeling.

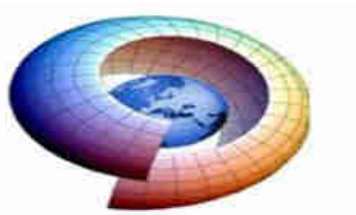
Data Mining Tools → SAS

- SAS stands for Statistical Analysis System. It is a product of the SAS Institute created for analytics and data management. SAS can mine data, change it, manage information from various sources, and analyze statistics. It offers a graphical UI for non-technical users.

- SAS data miner allows users to analyze big data and provide accurate insight for timely decision-making purposes.

Data Mining Tools → DataMelt Data Mining

- DataMelt is a computation and visualization environment which offers an interactive structure for data analysis and visualization. It is primarily designed for students, engineers, and scientists. It is also known as Dmelt.



- DMelt is a multi-platform utility written in JAVA. It can run on any operating system which is compatible with JVM (Java Virtual Machine).
- It consists of Science and mathematics libraries.

Data Mining Tools → Rattle

- Ratte is a data mining tool based on GUI. It uses the R stats programming language.
- Rattle exposes the statical power of R by offering significant data mining features.

Data Mining Tools → Rapid Miner

- Rapid Miner is one of the most popular predictive analysis systems created by the company with the same name as the Rapid Miner. It is written in JAVA programming language.
- It offers an integrated environment for text mining, deep learning, machine learning, and predictive analysis