Classification: Basic Concept

- **What is Classification?**

➢ There are two forms of data analysis that can be used for extracting models describing important classes or to predict future data trends. These two forms are as follows −

- Classification
- Predictions

Classification models predict categorical class labels; and prediction models predict continuous valued functions.

Classification

➢ It is a Data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts.
Classification is the problem of identifying to which of a set of categories (subpopulations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known.

➢ Following are the examples of cases where the data analysis task is Classification −

- A bank loan officer wants to analyze the data in order to know which customer (loan applicant) is risky or which are safe.
- A marketing manager at a company needs to analyze a customer with a given profile, who will buy a new computer.

➢ In both of the above examples, a model or classifier is constructed to predict the categorical labels.

- These labels are risky or safe for loan application data and yes or no for marketing data.

Prediction

- Prediction deals with some variables or fields, which are available in the data set to predict unknown values regarding other variables of interest.

- Numeric prediction is the type of predicting continuous or ordered values for given input.

- **For example:** The company may wish to predict the potential sales of a new product given with its price. In this example we are bothered to predict a numeric value. In this case, a model or a predictor will be constructed that predicts a continuous-valued-function or ordered value.

- The most widely used approach for numeric prediction is regression.

- **Difference between classification and prediction**

Parameters	Classification	Prediction
Definition	Classification is the process of identifying to which category, a new observation belongs to on the basis of a training data set containing observations whose category membership is known.	Predication is the process of identifying the missing or unavailable numerical data for a new observation.
Accuracy	In classification, the accuracy depends on finding the class label correctly.	In predication, the accuracy depends on how well a given predicator can guess the value of a predicated attribute for a new data.
Model	A model or the classifier is constructed to find the categorical labels.	A model or a predictor will be constructed that predicts a continuous-valued function or ordered value.
Synonyms for the model	In classification, the model can be known as the classifier.	In predication, the model can be known as the predictor.

- **General Approach to Classification**

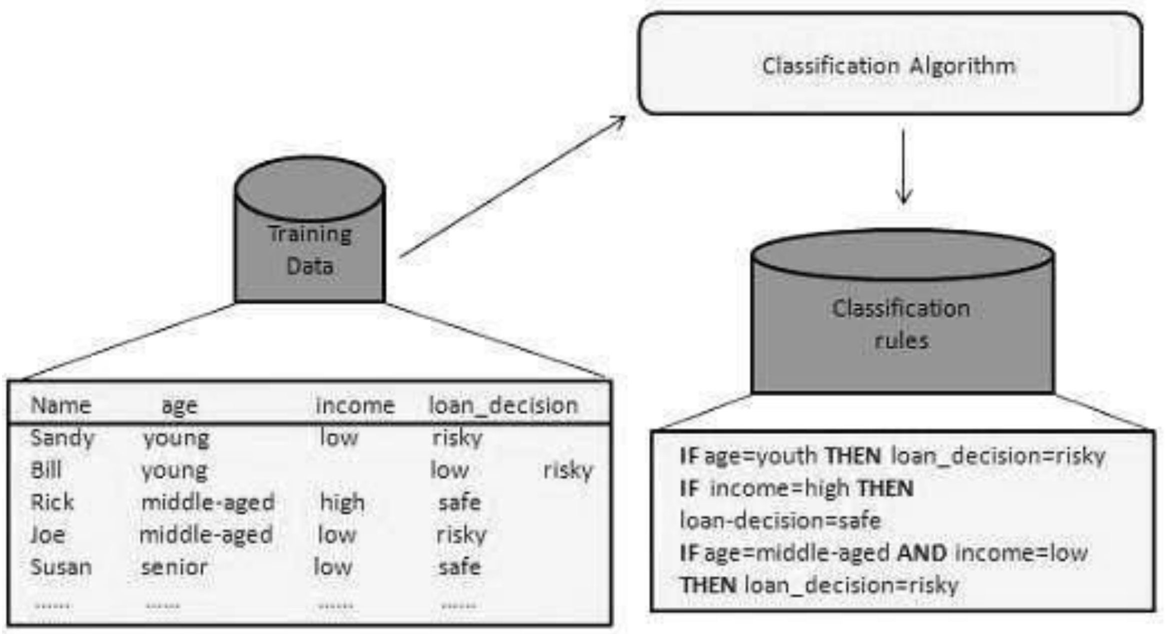
➢ Data classification is a two-step process, consisting of a learning step (where a classification model is constructed) and a classification step (where the model is used to predict class labels for given data).

1. Learning Step (Training Phase):

➢ Construction of Classification Model different algorithms is used to build a classifier by making the model learn using the training set available. The model has to be trained for the prediction of accurate results.

➢ **Building the Classifier or Model**

- In this step the classification algorithms build the classifier.
- The classifier is built from the training set made up of database tuples and their associated class labels.
- Each tuple that constitutes the training set is referred to as a category or class. These tuples can also be referred to as sample, object or data points.



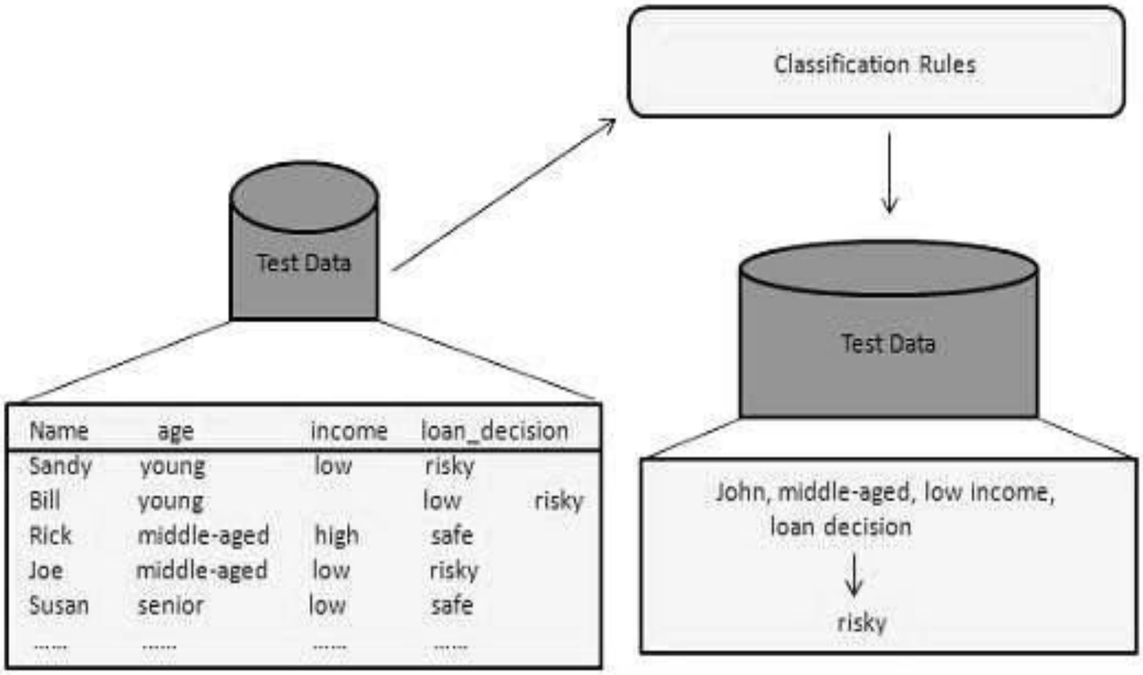
Name	age	income	loan_decision	
Sandy	young	low	risky	
Bill	young		low	risky
Rick	middle-aged	high	safe	
Joe	middle-aged	low	risky	
Susan	senior	low	safe	
......	......	......	......	

IF age=youth **THEN** loan_decision=risky
IF income=high **THEN**
loan-decision=safe
IF age=middle-aged **AND** income=low
THEN loan_decision=risky

2. Classification Step:

- Model used to predict class labels and testing the constructed model on test data and hence estimate the accuracy of the classification rules.

- In this step, the classifier is used for classification. Here the test data is used to estimate the accuracy of classification rules. The classification rules can be applied to the new data tuples if the accuracy is considered acceptable.



Name	age	income	loan_decision	
Sandy	young	low	risky	
Bill	young		low	risky
Rick	middle-aged	high	safe	
Joe	middle-aged	low	risky	
Susan	senior	low	safe	
......	......	......	......	

- **Decision Tree Induction**

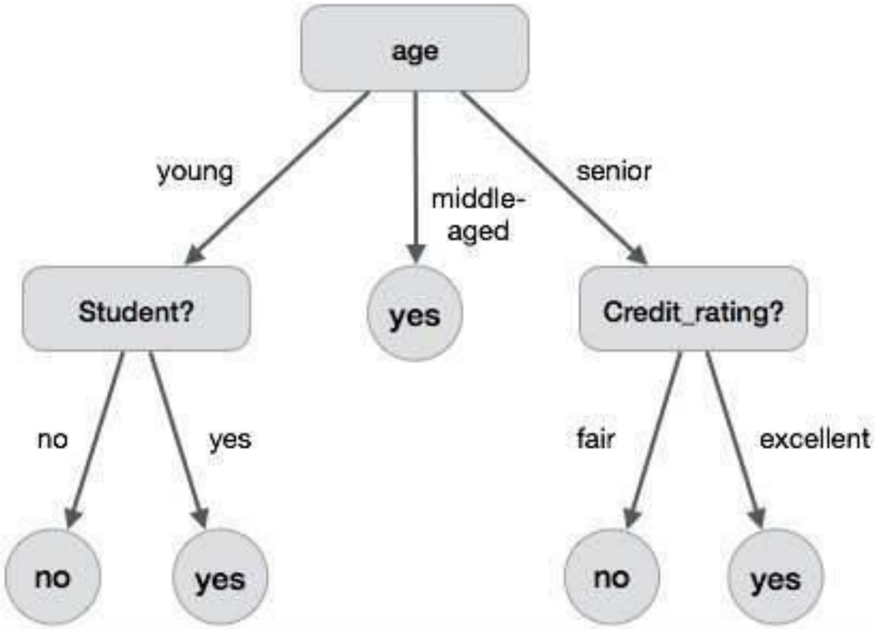
- Decision tree induction is the method of learning the decision trees from the training set. The training set consists of attributes and class labels. Applications of decision tree induction include astronomy, financial analysis, medical diagnosis, manufacturing, and production.

- A decision tree is a structure that includes a **root node, branches, and leaf nodes.**

- **Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.**

- **Example:**

  The following decision tree is for the concept buy_computer that indicates whether a customer at a company is likely to buy a computer or not. Each internal node represents a test on an attribute. Each leaf node represents a class.



The benefits of having a decision tree are as follows −

- It does not require any domain knowledge.
- It is easy to comprehend.
- The learning and classification steps of a decision tree are simple and fast.

Attribute Selection Measures

- **Attribute selection measures are also called splitting rules to decide how the tuples are going to split**.

- The splitting criteria are used to best partition the dataset. These measures provide a ranking to the attributes for partitioning the training tuples.

- **The most popular methods of selecting the attribute are information gain, Gini index.**

1. Information Gain

- This method is the main method that is used to build decision trees. It reduces the information that is required to classify the tuples. It reduces the number of tests that are needed to classify the given tuple. The attribute with the highest information gain is selected.

- The original information needed for classification of a tuple in dataset D is given by:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

- Where p is the probability that the tuple belongs to class C. The information is encoded in bits, therefore, log to the base 2 is used. E(s) represents the average amount of information required to find out the class label of dataset D. This information gain is also called **Entropy.**

- **Entropy - A** decision tree is built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values (homogeneous).

- The information required for exact classification after portioning is given by the formula:

$$E(T,X)=\sum_{c\in X}P(c)E(c)$$

- Where P (c) is the weight of partition. This information represents the information needed to classify the dataset D on portioning by X.

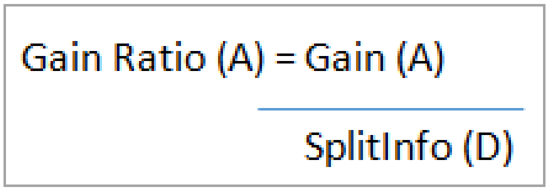
- Information gain is the difference between the original and expected information that is required to classify the tuples of dataset D.

$$Gain(T,X)=Entropy(T)-Entropy(T,X)$$

- Gain is the reduction of information that is required by knowing the value of X. The attribute with the highest information gain is chosen as “best”.

2. Gain Ratio

- Information gain might sometimes result in portioning useless for classification. However, the Gain ratio splits the training data set into partitions and considers the number of tuples of the outcome with respect to the total tuples. The attribute with the max gain ratio is used as a splitting attribute.



$$\text{Gain Ratio (A)} = \frac{\text{Gain (A)}}{\text{SplitInfo (D)}}$$

Gini Index

- Gini index says, if we select two items from a population at random then they must be of same class and probability for this is 1 if population is pure.

1. It works with categorical target variable “Success” or “Failure”.
2. It performs only Binary splits
3. Higher the value of Gini higher the homogeneity.
4. CART (Classification and Regression Tree) uses Gini method to create binary splits.

Tree Pruning

Tree pruning is performed in order to remove anomalies in the training data due to noise or outliers. The pruned trees are smaller and less complex.

Tree Pruning Approaches

There are two approaches to prune a tree −

1. **Pre-pruning** −The tree is pruned by halting its construction early. (e.g.,by deciding not to further split or partition the subset of training tuples at a given node).Upon halting, the node becomes a leaf. The leaf may hold the most frequent class among the subset tuples or the probability distribution of those tuples.

2. **Post-pruning** - This approach removes a sub-tree from a fully grown tree. A subtree at a given node is pruned by removing its branches and replacing it with a leaf. The leaf is labeled with the most frequent class among the subtree being replaced

- **What Is Cluster Analysis?**

- Cluster analysis or simply clustering is the process of partitioning a set of data objects (or observations) into subsets.

- Each subset is a cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters. **The set of clusters resulting from a cluster analysis can be referred to as a clustering.**

- Clustering is also called **data segmentation** because clustering partitions large data sets into groups according to their similarity. Clustering can also be used for outlier detection,

- **<u>Requirements of Clustering in Data Mining</u>**

- The following points throw light on why clustering is required in data mining −

- **Scalability**

- We need highly scalable clustering algorithms to deal with large databases otherwise clustering may lead to biased results.

- **Ability to deal with different kinds of attributes**

- Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.

- **Discovery of clusters with attribute shape**

- The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.

- **High dimensionality**

- The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.

- **Ability to deal with noisy data**

- Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.

- **Interpretability**

- The clustering results should be interpretable, comprehensible, and usable.

- **Overview of Basic Clustering Methods**

- Clustering methods can be classified into the following categories

1. Partitioning Method
2. Hierarchical Method
3. Density-based Method
4. Grid-Based Method
5. Model-Based Method
6. Constraint-based Method

1. Partitioning Methods

- Suppose we are given a database of ‘n’ objects and the partitioning method constructs ‘k’ partition of data. Each partition will represent a cluster and k ≤ n. It means that it will classify the data into k groups, which satisfy the following requirements −
  - Each group contains at least one object.
  - Each object must belong to exactly one group.
- **Points to remember**
  - For a given number of partitions (say k), the partitioning method will create an initial partitioning.
  - Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other.

2. Hierarchical Methods

- This method creates a hierarchical decomposition of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed. There are two approaches here −

- **Agglomerative Approach**
- **Divisive Approach**

- **Agglomerative Approach**

➢ **This approach is also known as the bottom-up approach**. In this, we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keeps on doing so until all of the groups are merged into one or until the termination condition holds.

- **Divisive Approach**

➢ **This approach is also known as the top-down approach**. In this, we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid, i.e., once a merging or splitting is done, it can never be undone.

3. <u>Density-based Method</u>

➢ **This method is based on the notion of density.** The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

4. <u>Grid-based Method</u>

➢ In this, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure.

Advantages

- The major advantage of this method is fast processing time.
- It is dependent only on the number of cells in each dimension in the quantized space.

5. Model-based methods

- In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points.
- This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods.

6. Constraint-based Method

- In this method, the clustering is performed by the incorporation of user or application-oriented constraints.
- A constraint refers to the user expectation or the properties of desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. Constraints can be specified by the user or the application requirement.

▪ k-Means: A Centroid-Based Technique

- K-means algorithm is an iterative algorithm that tries to partition the dataset into *K* pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group.

- K-Means clustering intends to partition n objects into k clusters in which each object belongs to the cluster with the nearest mean.

- This method produces exactly k different clusters of greatest possible distinction. The best number of clusters k leading to the greatest separation (distance) is not known as a priori and must be computed from the data. The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function:

$$\text{objective function} \leftarrow J = \overset{\text{number of clusters}}{\sum_{j=1}^{k}} \overset{\text{number of cases}}{\sum_{i=1}^{n}} \underbrace{\left\| \overset{\text{case } i}{x_i^{(j)}} - \overset{\text{centroid for cluster } j}{c_j} \right\|^2}_{\text{Distance function}}$$

- **The way k-means algorithm works is as follows:**

1. Clusters the data into k groups where k is predefined.
2. Select k points at random as cluster centers.
3. Assign objects to their closest cluster center Calculate the centroid or mean of all objects in each cluster.
4. Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.

- **Data Mining Applications**

Here is the list of areas where data mining is widely used −

- Financial Data Analysis
- Retail Industry
- Telecommunication Industry
- Biological Data Analysis
- Other Scientific Applications
- Intrusion Detection

Financial Data Analysis

- The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining. Some of the typical cases are as follows −
  - Design and construction of data warehouses for multidimensional data analysis and data mining.
  - Loan payment prediction and customer credit policy analysis.
  - Classification and clustering of customers for targeted marketing.
  - Detection of money laundering and other financial crimes.

Retail Industry

- Data Mining has its great application in Retail Industry because it collects large amount of data from on sales, customer purchasing history, goods transportation, consumption and services.
- It is natural that the quantity of data collected will continue to expand rapidly because of the increasing ease, availability and popularity of the web.

- Data mining in retail industry helps in identifying customer buying patterns and trends that lead to improved quality of customer service and good customer retention and satisfaction. Here is the list of examples of data mining in the retail industry
  - Design and Construction of data warehouses based on the benefits of data mining.
  - Multidimensional analysis of sales, customers, products, time and region.
  - Analysis of effectiveness of sales campaigns.
  - Customer Retention.
  - Product recommendation and cross-referencing of items.

Telecommunication Industry

- Today the telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, internet messenger, images, e-mail, web data transmission, etc.
- Data mining in telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Here is the list of examples for which data mining improves telecommunication services −
  - Multidimensional Analysis of Telecommunication data.
  - Fraudulent pattern analysis.
  - Identification of unusual patterns.
  - Multidimensional association and sequential patterns analysis.
  - Mobile Telecommunication services.
  - Use of visualization tools in telecommunication data analysis.

Biological Data Analysis

- Biological data mining is a very important part of Bioinformatics. Following are the aspects in which data mining contributes for biological data analysis −
  - Semantic integration of heterogeneous, distributed genomic and proteomic databases.
  - Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
  - Discovery of structural patterns and analysis of genetic networks and protein pathways.
  - Association and path analysis.
  - Visualization tools in genetic data analysis.

Other Scientific Applications

- Huge amount of data have been collected from scientific domains such as geosciences, astronomy, etc. A large amount of data sets is being generated because of the fast numerical simulations in various fields such as climate and ecosystem modeling, chemical engineering, fluid dynamics, etc.
- Following are the applications of data mining in the field of Scientific Applications −
  - Data Warehouses and data preprocessing.
  - Graph-based mining.
  - Visualization and domain specific knowledge.

Intrusion Detection

- Intrusion refers to any kind of action that threatens integrity, confidentiality, or the availability of network resources. In this world of connectivity, security has become the major issue.
- With increased usage of internet and availability of the tools and tricks for intruding and attacking network prompted

intrusion detection to become a critical component of network administration. Here is the list of areas in which data mining technology may be applied for intrusion detection −

- Development of data mining algorithm for intrusion detection.
- Association and correlation analysis, aggregation to help select and build discriminating attributes.
- Analysis of Stream data.
- Distributed data mining.
- Visualization and query tools.