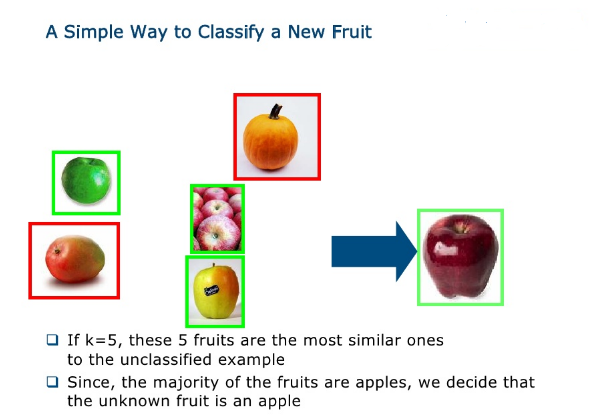
## Subject: DATA MINING code:BCACsT5.11

**Chapter-2**

**Classification**

* **Classification is a predictive Data mining task which maps the data into predefined groups or classes.**
* **Classification derives a model to determine the class of an object based on its attributes.**
* **Classification comes under a supervised learning .The first step of the classification process can also be viewed as the learning of a mapping or function, Y = f (X), that can predict the associated class label y of a given tuple X.**

**Classification Example**

****

**Classification derives a model to determine the class of an object based on its attributes**. A collection of records will be available, each record with a set of attributes. Goal of classification task is assigning a class attribute to new set of records as accurately as possible.

**Classification can be used in direct marketing that** is to reduce marketing costs by targeting a set of customers who are likely to buy a new product.Using the available data, it is possible to know which customers purchased similar products and who did not purchase in the past.

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 customers (loan applicant) are 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.

**Applications of classification data mining task.**

1. Internet traffic interception - certain governments (possibly from the middle east) would like to restrict certain categories of web pages. For example, due to religious restrictions, certain movie pages may be restricted/censored. This is a clear application of classification.
2. Video classification - as and when you upload a video on youtube, the video has to be classified into appropriate categories and meta-data added to it .
3. Voice and image classification in for google assistance and other applications.
4. Weather Predictions
5. Resume sorting
6. Gene expression analysis

**Classification Process**

**Classification is a derivation of a model or function which determines the class of an object based on attributes.**

**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).

In the first step, a classifier is built describing a predetermined set of data classes or concepts. This is the **learning step** (or training phase), where a classification algorithm builds the classifier by analyzing or “learning from” a **training set** made up of database tuples and their associated class labels.

In the context of classification, data tuples can be referred to as *samples, examples, instances, data points*, or *objects.* ( Bank, College, Student etc……. Database sample).

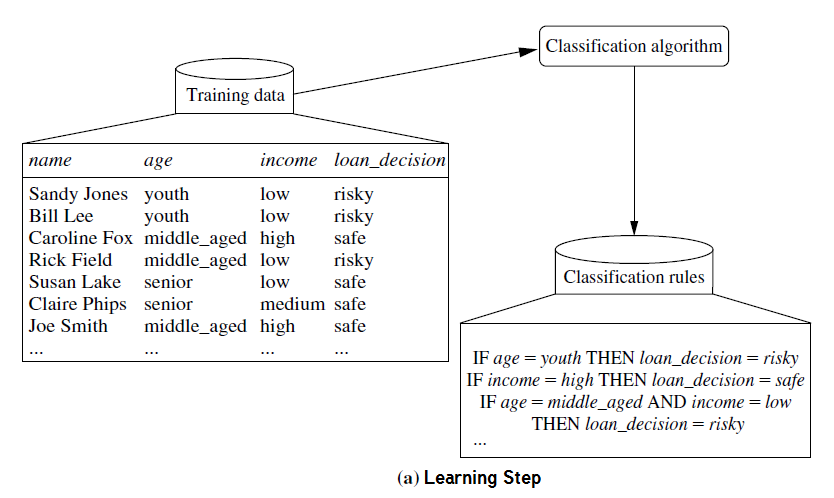
A tuple, ***X***, is represented by an *n*-dimensional **attribute vector**, ***X*** =( .*x*1, *x*2, : : : , *xn*), depicting *n* measurements made on the tuple from *n* database attributes, respectively, *A*1, *A*2, : : : , *An*.1 .

Ex: if **A** is Income attribute then it is represented by **X** attribute vector (low, medium ,high)

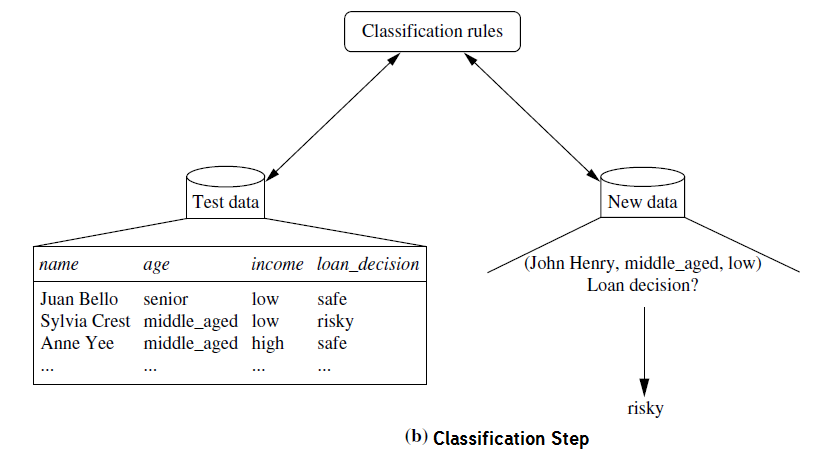
***X***, is **class label attribute**, predefined class which is determined by another database attribute. The class label attribute is discrete-valued and unordered. It is *categorical*(or nominal) in that each value serves as a category or class.

The individual tuples making up the training set are referred to as **training tuples** and are randomly sampled from the database under analysis

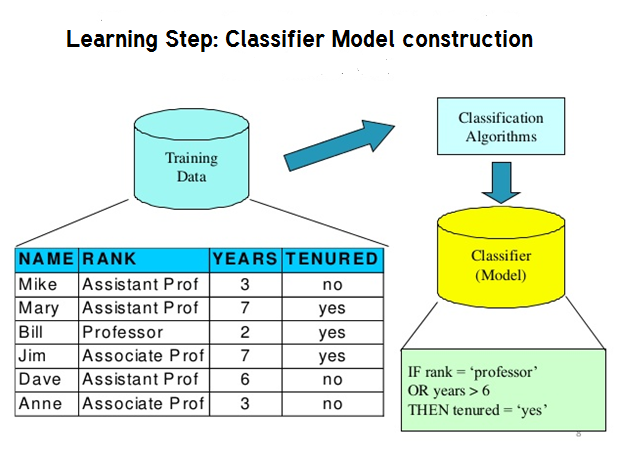
**(a) *Learning***: Training data are analyzed by a classification algorithm. Here, the class label attribute is *loan decision*, and the learned model or classifier is represented in the form of classification rules.

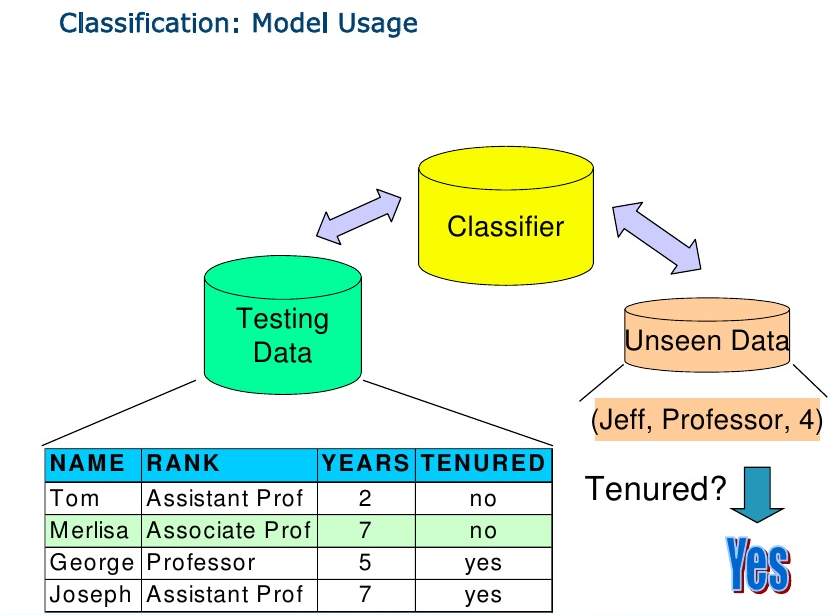


. **(b) *Classification*:** Test data are used to estimate the accuracy of the classification rules. If the accuracy is considered acceptable, the rules can be applied to the classification of new data tuples.



Ex2: Classifying faculties of institution to give tenure( Permanent Posting)





**Prediction**

**Prediction task predicts the possible values of missing or future data. It models continuous valued functions.**

**Prediction involves developing a model based on the available data and this model is used in predicting future values of a new data set of interest.**

For example, a model can predict the income of an employee based on education, experience and other demographic factors like place of stay, gender etc. Also prediction analysis is used in different areas including medical diagnosis, fraud detection etc.

Prediction derives the relationship between a thing you know and a thing you need to predict for future reference.

For example, prediction models in data mining are used by a marketing manager who predict that how much amount a particular customer will spend during a sale, so that upcoming sale amount can be planned accordingly. The prediction in [data mining](https://www.cogneesol.com/data-mining-services) is known as Numeric Prediction. Generally regression analysis is used for prediction.

Typical applications

* Credit approval
* Target marketing
* Medical diagnosis
* Fraud detection

**Comparison b/w Classification and Prediction model**

Accuracy :

classifier accuracy: predicting class label

predictor accuracy: guessing value of predicted attributes

Speed :

computational costs: time to construct the model (training time)

computational costs: time to use the model(classification/prediction time)

Robustness :

handling noise and missing values

Scalability

efficiency in large amounts of data

Interpretability :

understanding and insight provided by the model

Other measures, e.g., goodness of rules, such as decision tree size or

compactness of classification rules

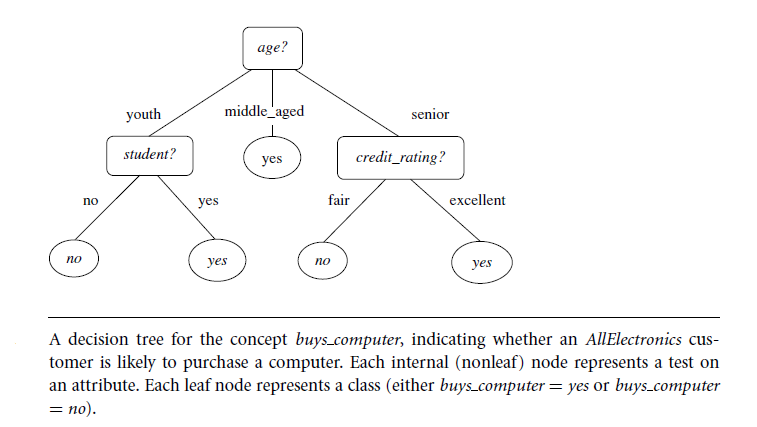
**Decision tree**

**Decision Tree Induction**

**Decision tree induction** is the learning of decision trees from class-labeled training tuples.

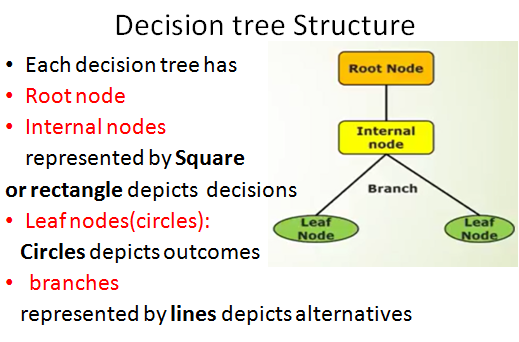
A **decision tree** is a flowchart-like tree structure, where each **internal node** (nonleaf node) denotes a test on an attribute, each **branch** represents an outcome of the test, and each **leaf node** (or *terminal node*) holds a class label. The topmost node in a tree is the **root** node.

A typical decision tree for buying computer is shown in Figure:



It represents the concept *buys computer*, that is, it predicts whether a customer at *AllElectronics* is

likely to purchase a computer. Each leaf node represents a class either buys\_computer = yes or buys\_computer=no.



Some decision tree algorithms produce only *binary* trees (whereeach internal node branches to exactly two other nodes), whereas others can produce nonbinary trees.

*“How are decision trees used for classification?”* Given a tuple, ***X***, for which the associated

class label is unknown, the attribute values of the tuple are tested against the

decision tree. A path is traced from the root to a leaf node, which holds the class

prediction for that tuple. Decision trees can easily be converted to classification rules.

**Steps for drawing Decision Trees**

**1. Write the main decision.**

Begin the decision tree by drawing a box (the root node) on 1 edge of your paper.

Write the main decision on the box.

**2. Draw the lines**

Draw line leading out from the box for each possible solution or action. Make at least 2,

but better no more than 4 lines. Keep the lines as far apart as you can to enlarge the tree later.

**3. Illustrate the outcomes of the solution at the end of each line.**

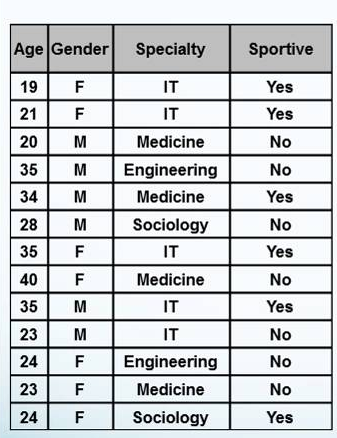
**4. Continue adding boxes or cicles and lines .**

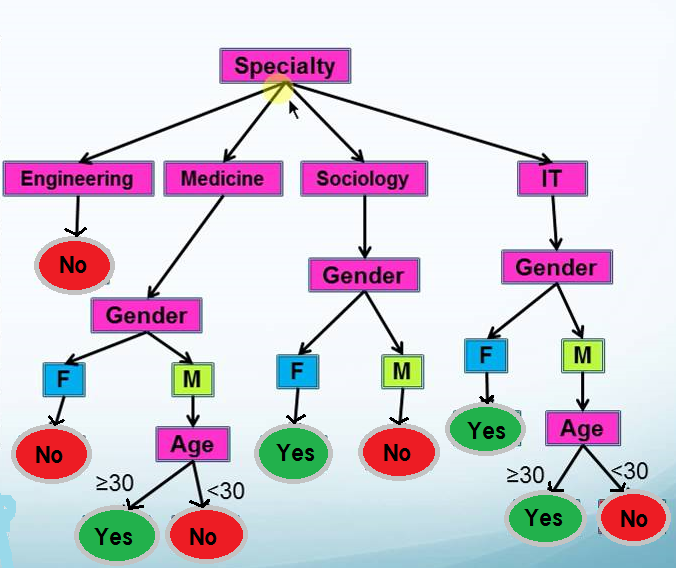
Continue until there are no more problems, and all lines have either uncertain

outcome or blank ending.

**5. Finish the tree with final outcome.**

Construct a decision tree for the training data set given below.





*“Why are decision tree classifiers so popular?”*

The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery.

Decision trees can handle multidimensionaldata.

Their representation of acquired knowledge in tree form is intuitive and

generally easy to assimilate by humans.

The learning and classification steps of decision tree induction are simple and fast.

**Advantages and disadvantages of Decision tree**

**Advantages:**

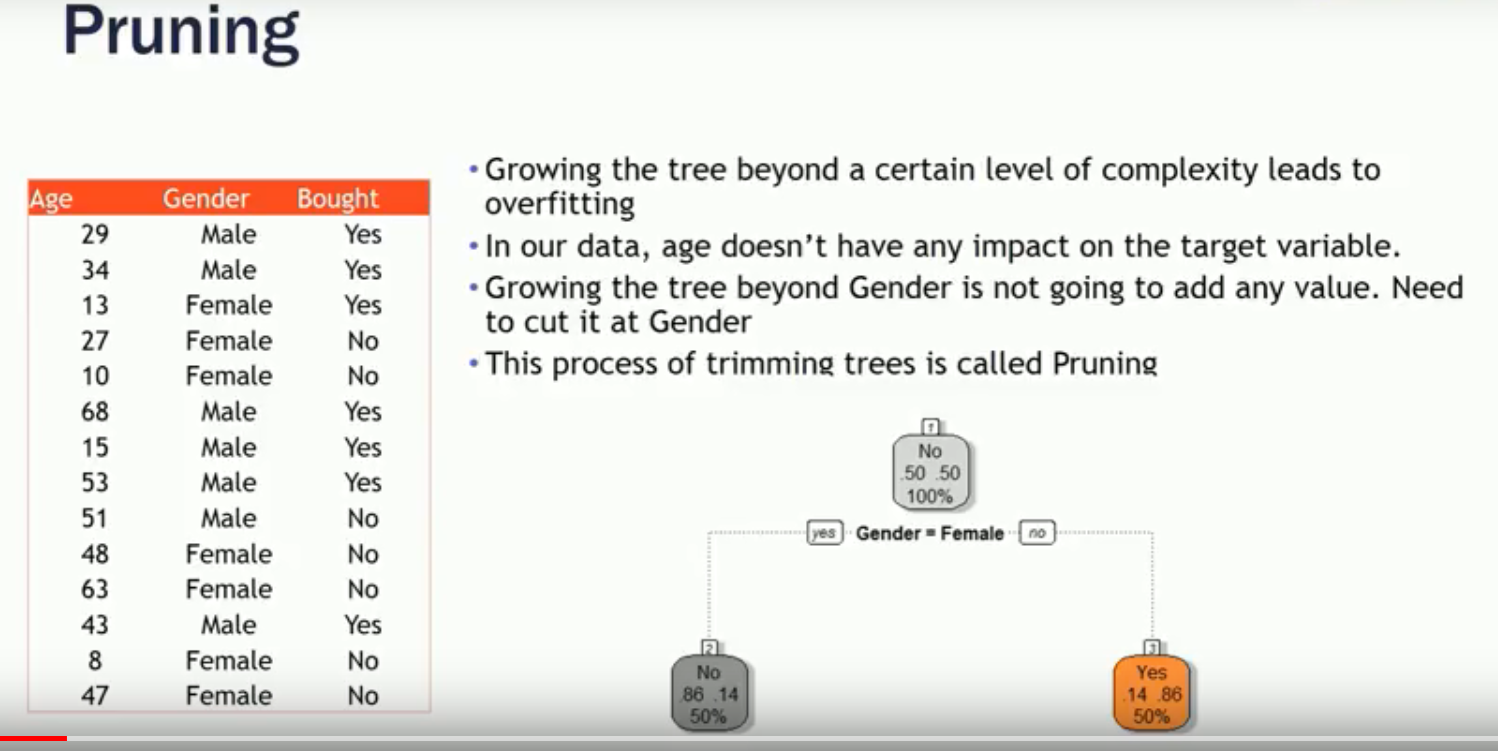
* Simple to understand and to interrupt,trees can be visualized.
* Requires little data preparation
* Uses a white box model

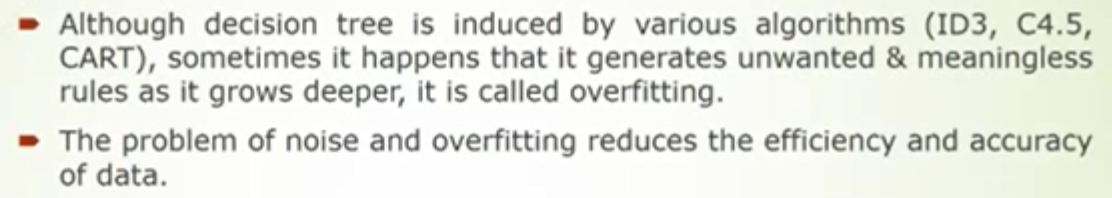
**Disadvantages:**

* Over fitting due to noise and outliers
* Large decision trees are over-complex and difficult to understand
* Inefficiency and inaccuracy of the decision tree due to over fitting
* Memory dependency and swapping of training data in and out of memories
* Inefficiency due to predefined algorithm design

**Over fitting and Pruning of Decision Tree**

**Overfitting: An induced or constructed tree may overfit the training data.ie too many branches may reflect in anamolies due to noise or outliers which result in poor accuracy of unseen samples.**

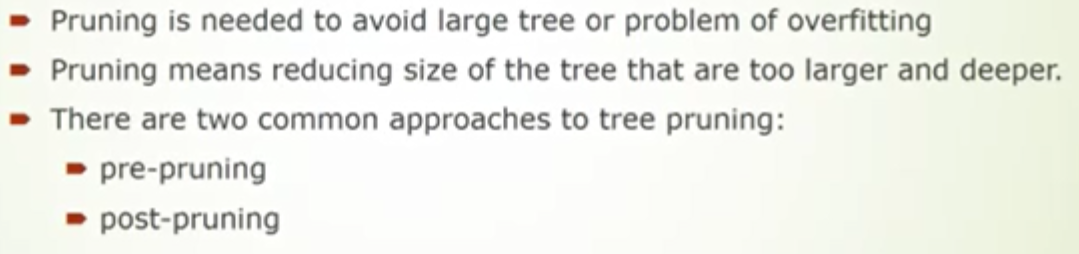
****

****

**Pruning: The process of trimming the decision tree.ie, removing the branches which are not needed for the classification.**

**Pruning of decision tree is done by replacing a whole subtree by a leaf node.**

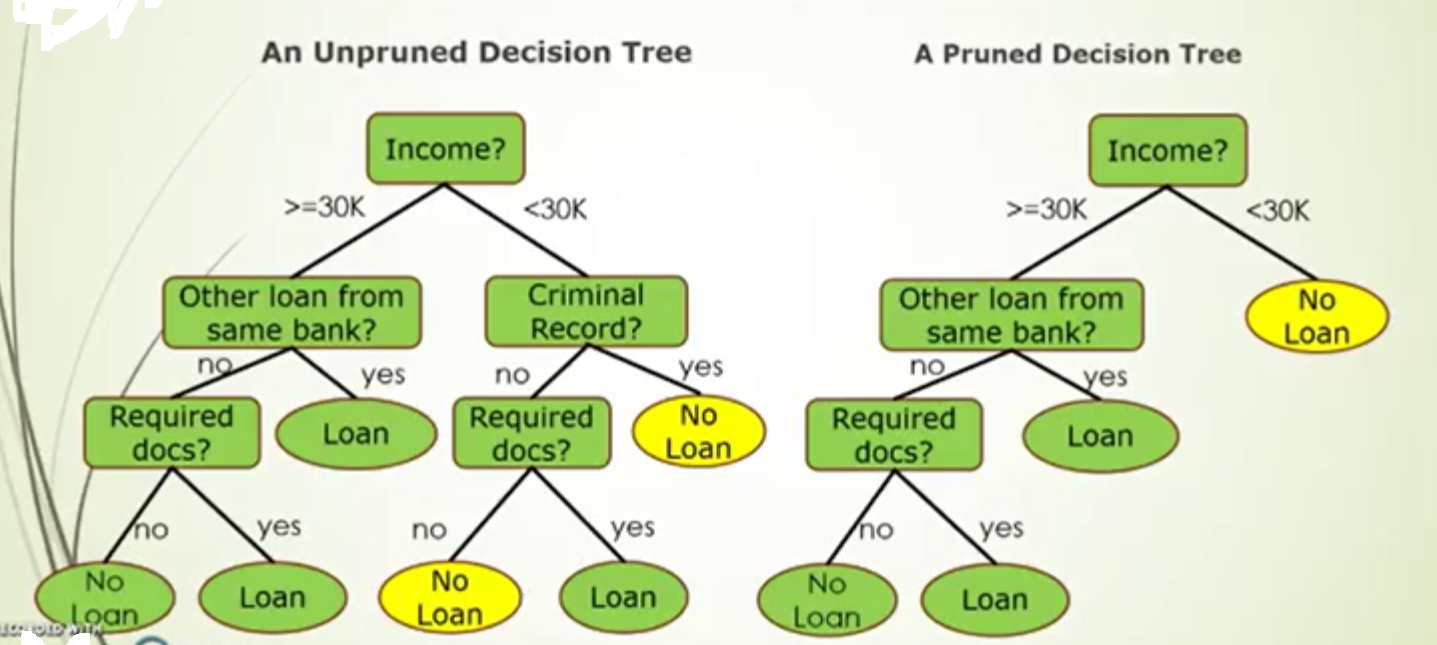
**The replacement takes place if a decision rule establishes that the expected error rate in the subtree is greater than in the single leaf.**

****

**Pruning approach to avoid overfitting**

****

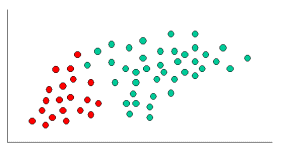
**Pruned Tree**

****

**Naive Bayes Classifie****r**

**Naive Bayes Classifier Introductory Overview**

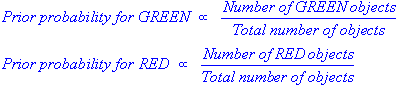
It is a classification technique based on [Bayes’ Theorem](https://en.wikipedia.org/wiki/Bayes%27_theorem) with an assumption of independence(naïve) among predictors. This theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods.



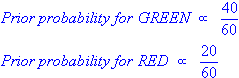
consider an example displayed which consists of 60 objects can be classified as either GREEN or RED. Our task is to classify new cases as they arrive, i.e., decide to which class label they belong, based on the currently exiting objects.

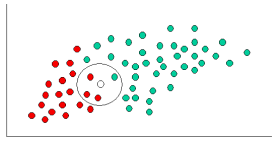
Since there are twice as many GREEN objects as RED, it is reasonable to believe that a new case (which hasn't been observed yet) is twice as likely to have membership GREEN rather than RED. In the Bayesian analysis, this belief is known as the prior probability. Prior probabilities are based on previous experience, in this case the percentage of GREEN and RED objects, and often used to predict outcomes before they actually happen.

Thus, we can write:

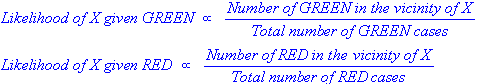


Since there is a total of 60 objects, 40 of which are GREEN and 20 RED, our prior probabilities for class membership are:





Having formulated our prior probability, we are now ready to classify a new object (WHITE circle). Since the objects are well clustered, it is reasonable to assume that the more GREEN (or RED) objects in the vicinity of X.To measure this likelihood, we draw a circle around X which gives a count of number of points chosen irrespective of their class labels. Then we calculate the number of points in the circle belonging to each class label. From this we calculate the likelihood:

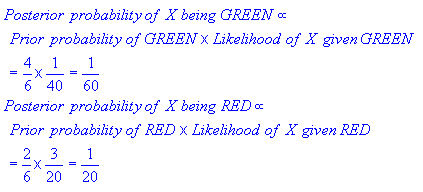


From the illustration above, it is clear that Likelihood of X given GREEN is smaller than Likelihood of X given RED, since the circle encompasses 1 GREEN object and 3 RED ones. Thus:

http://www.statsoft.com/textbook/NBEquation.gif

http://www.statsoft.com/textbook/NaiveBayesIntro6.gif

Although the prior probabilities indicate that X may belong to GREEN (given that there are twice as many GREEN compared to RED) the likelihood indicates otherwise; that the class membership of X is RED (given that there are more RED objects in the vicinity of X than GREEN). In the Bayesian analysis, the final classification is produced by combining both sources of information, i.e., the prior and the likelihood, to form a posterior probability using the so-called Bayes' rule (named after Rev. Thomas Bayes 1702-1761).



Finally, we classify X as RED since its class membership achieves the largest posterior probability.

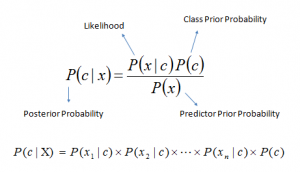
## Naive Bayes algorithm

It is a classification technique based on [Bayes’ Theorem](https://en.wikipedia.org/wiki/Bayes%27_theorem) with an assumption of independence(naïve) among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c).

the equation is given as:

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/09/Bayes_rule-300x172.png)where,

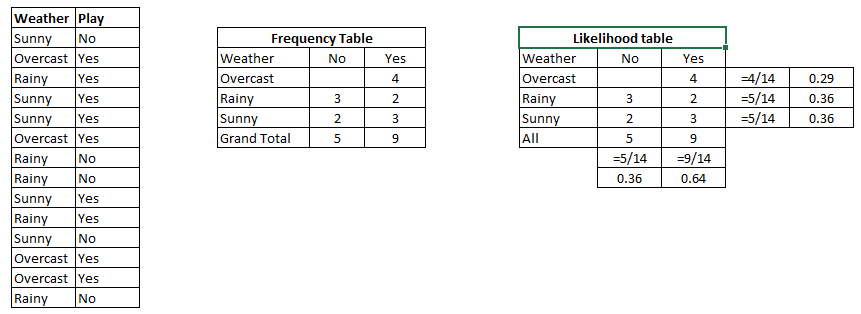
* *P*(*c|x*) is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
* *P*(*c*) is the prior probability of *class*.
* *P*(*x|c*) is the likelihood which is the probability of *predictor* given *class*.
* *P*(*x*) is the prior probability of *predictor*.

## How Naive Bayes algorithm works?

Let’s understand it using an example. Below we have a training data set of weather and corresponding target variable ‘Play’ (suggesting possibilities of playing). Now, we need to classify whether players will play or not based on weather condition. Let’s follow the below steps to perform it.

**Step 1: Convert the data set into a frequency table**

**Step 2: Create Likelihood table by finding the probabilities like Overcast probability = 0.29 and probability of playing is 0.64.**

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Bayes_41.png)

**Step 3: Now, use Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.**

**Problem:**Players will play if weather is sunny. Is this statement is correct?

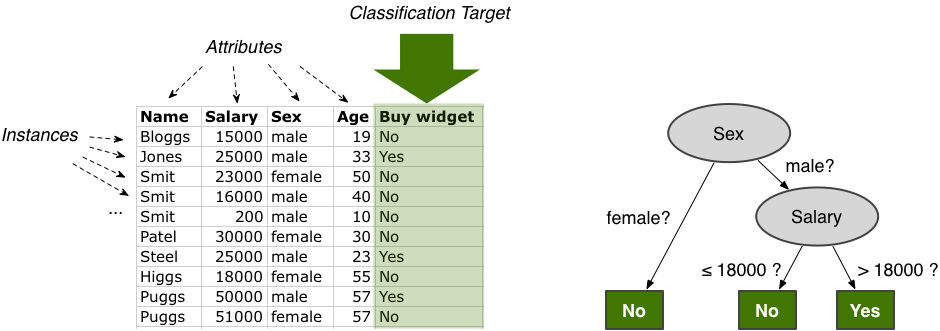
We can solve it using above discussed method of posterior probability.

P(Yes | Sunny) = P( Sunny | Yes) \* P(Yes) / P (Sunny)

Here we have P (Sunny |Yes) = 3/9 = 0.33, P(Sunny) = 5/14 = 0.36, P( Yes)= 9/14 = 0.64

Now, P (Yes | Sunny) = 0.33 \* 0.64 / 0.36 = 0.60, which has higher probability.

Naive Bayes uses a similar method to predict the probability of different class based on various attributes. This algorithm is mostly used in text classification and with problems having multiple classes.

2.Problem-2

i) .male with salary >18000 buys widget. Is the statement correct

ii) Female irrespective of age will not buy the widget. Evaluate the statement

i)

P(y)=3/10

P(n)=7/10

P(male)=6/10

P(>18000)=6/10

|  |  |  |
| --- | --- | --- |
| **Gender** | **Y** | **N** |
| **Male** | **3** | **3** |
| **Female** | **0** | **4** |

**Likelyhood(m/y) L(M/N)=3/3. L(M/N)=3/7**

|  |  |  |
| --- | --- | --- |
| **Salary** | **Y** | **N** |
| **>18000** | **3** | **3** |
| **<=18000** | **0** | **4** |

**Likelyhood. L(3/3). L(<=18000)=(3/4)**

**P(male/yes)=3/3**

**P(male/no)=3/7**

**P(>18000/yes)=3/3**

**P(>18000/no)=3/7**

**P(yes/male>18000)=p(male/yes)\*P(>18000/yes)\*p(y) / (male)\*P(>18000)**

**=(3/3\*3/3\*3/10)/(6/10\*6/10)**

**=o.8333**

**ii) Female irrespective of age buys widget**

**P(no/female)=p(female/no)\*p(n)/p(female)**

**=(4/7\*7/10)/(4/10)**

**= 1**

**Female certainly not going to buy the widget irrespective of age**

**Advantages and disadvantages of Naive Bayes**

***Advantages:***

* It is easy and fast to predict class of test data set. It also perform well in multi class prediction
* When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.
* It perform well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption).

***Disadvantages:***

* If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as “Zero Frequency”. To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.
* On the other side naive Bayes is also known as a bad estimator, so the probability outputs from predict\_proba are not to be taken too seriously.
* Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.

## Applications of Naive Bayes Algorithms

* **Real time Prediction:**Naive Bayes is an eager learning classifier and it is sure fast. Thus, it could be used for making predictions in real time.
* **Multi class Prediction:**This algorithm is also well known for multi class prediction feature. Here we can predict the probability of multiple classes of target variable.
* **Text classification/ Spam Filtering/ Sentiment Analysis:** Naive Bayes classifiers mostly used in text classification (due to better result in multi class problems and independence rule) have higher success rate as compared to other algorithms. As a result, it is widely used in Spam filtering (identify spam e-mail) and Sentiment Analysis (in social media analysis, to identify positive and negative customer sentiments)
* **Recommendation System:**Naive Bayes Classifier and [Collaborative Filtering](https://en.wikipedia.org/wiki/Collaborative_filtering) together builds a Recommendation System that uses machine learning and data mining techniques to filter unseen information and predict whether a user would like a given resource or not

**Classification by Back propagation.**

**It is a classification method which uses a neural network learning algorithm for classifying and predicting the outcomes. It iteratively process a set of training tuples and compare networks prediction with the actual target value.**

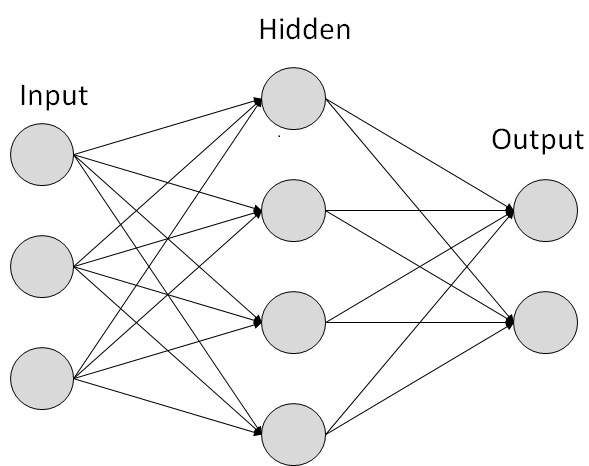
**A neural network is a set of connected inputs/output units where each connection has a weight associated with it.** Neural networks represent a brain metaphor for information processing. These models are biologically inspired rather than an exact replica of how the brain actually functions.

Neural computing refers to a pattern recognition methodology for machine learning.

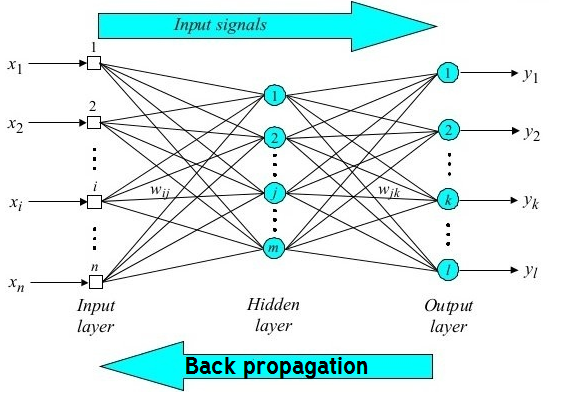
**Artificial Neural Network**

ANNs are composed of multiple **nodes**, which imitate biological **neurons** of human brain. The neurons are connected by links and they interact with each other. The nodes can take input data and perform simple operations on the data. The result of these operations is passed to other neurons. The output at each node is called its **activation** or **node value.**

Each link is associated with **weight.** ANNs are capable of learning, which takes place by altering weight values. The following illustration shows a simple ANN −



**Back propagation as a process**

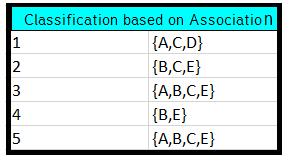
****

**During learning phase, the network learns by adjusting weights so as to able to predict the correct class label of the input tuples. During classification** desired outputs are compared to achieve system outputs, and then the systems are tuned by adjusting connection ie,  **for the expected class label the weights are updated with the connections backwards from output to input, hence the name back propagation.**

Neural networks have been used in many business applications for pattern recognition (radar systems, face identification, object recognition and more), forecasting, prediction, and classification.

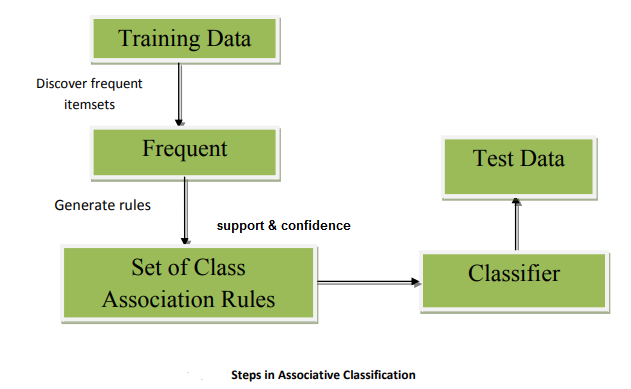
**Associative Classification (AC)**

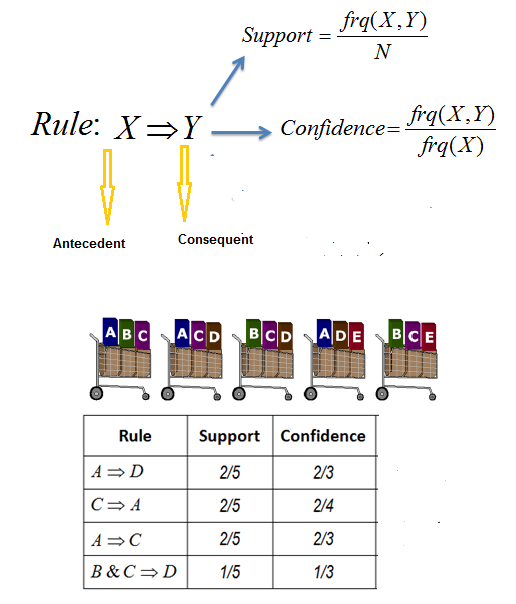
Associative classification mining is a promising approach in data mining that utilizes the association rule discovery techniques to construct classification systems, also known as associative classifiers.



AC integrates two known data mining tasks, association rule discovery and classification, to build a model (classifier) for the purpose of prediction. Classification and association rule discovery are similar tasks in data mining, with the exception that the main aim of classification is the prediction of class labels, while association rule discovery describes correlations between items in a transactional database.

Associative classification Association rules are generated and analyzed for use in classification Search for strong associations between frequent patterns (conjunctions of attribute-value pairs) and class labels Classification: Based on evaluating a set of rules in the form of P1 ^ p 2 … ^ pl “Aclass = C” (conf, sup).





An **association rule** is a rule of type X → Y, where X and Y are itemsets

– If transaction contains itemset X, it (probably) also contains

itemset Y

The **support** of itemset X in database D is the number

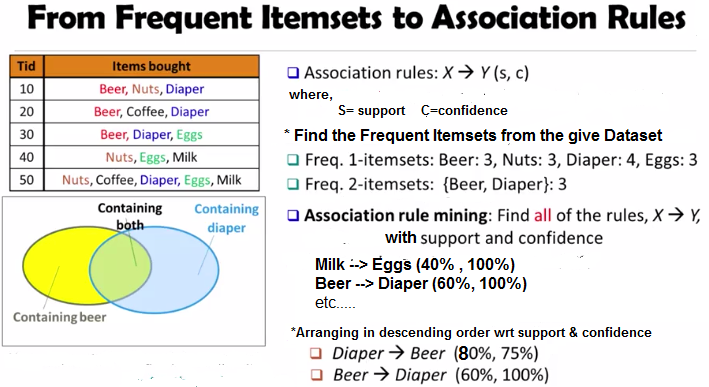
of transactions in D that contain it:

supp(X, D) = |{t ∈ D : t contains X}

The **confidence** of rule X → Y in data D is c(X → Y, D)

– The confidence is the empirical conditional probability that

transaction contains Y given that it contains X



Advantage of Associative classification.

It explores highly confident associations among multiple attributes and may overcome some constraints introduced by decision-tree induction, which considers only one attribute at a time In many studies, associative classification has been found to be more accurate than some traditional classification methods, such as C4.5, CBA,CART

Applications

CRMs Customer relationship management.

In Sales and marketing to recommend products , offers to the customers.

1. **Statistical Based Algorithms**

**Statistical classification mainly works on the principle of Logistic regression.**

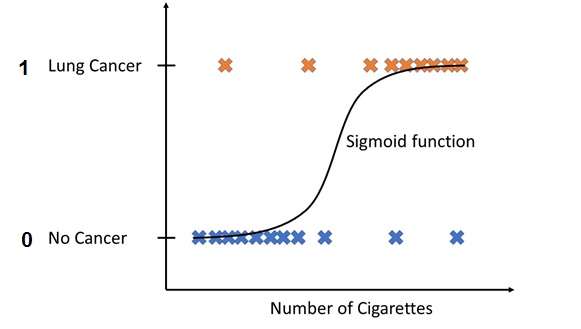
***Logistic Regression*** is basically a supervised classification algorithm. Logistic Regression is one of the most used Machine Learning algorithms for binary classification. Like many other machine learning techniques, it is borrowed from the field of statistics. Regression used for classification deals with estimation (prediction) of an output (class) value based on input values from the database. It gives you a discrete binary outcome between 0 and 1.

A simple example of a Logistic Regression problem would be an algorithm used for cancer detection that takes screening picture as an input and should tell if a patient has cancer (1) or not (0).

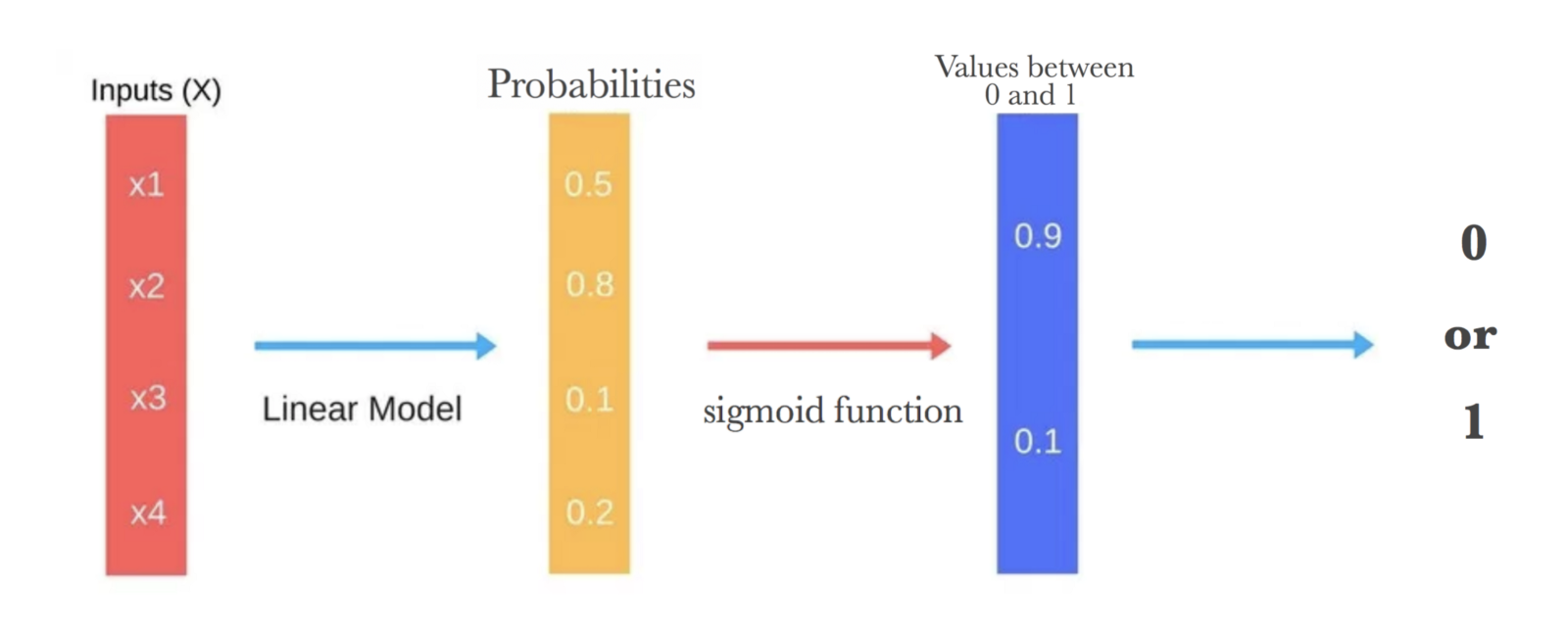
### How it works

Logistic Regression measures the relationship between the dependent variable (our label, what we want to predict) and the one or more independent variables (our features), by estimating probabilities using its underlying logistic function.

These probabilities must then be transformed into binary values in order to actually make a prediction. This is the task of the logistic function, also called the ***sigmoid function***. The Sigmoid-Function is an *S-shaped* curve that can take any real-valued number and map it into a value between the range of 0 and 1, but never exactly at those limits. These values between 0 and 1 will then be transformed into either 0 or 1 using a threshold classifier. EG:



The picture below illustrates the steps that logistic regression goes through to give you your desired output.



**Rule based Classification**

Rule-based classifier makes use of a set of IF-THEN rules for classification. e can express a rule in the following from −

IF condition THEN conclusion

Let us consider a rule R1,

R1: IF age = youth AND student = yes

THEN buy\_computer = yes

**Points to remember −**

* The IF part of the rule is called **rule antecedent** or **precondition**.
* The THEN part of the rule is called **rule consequent**.
* The antecedent part ,the condition consist of one or more attribute tests and these tests are logically ANDed.
* The consequent part consists of class prediction.

**Note** − We can also write rule R1 as follows −

R1: (age = youth) ^ (student = yes))(buys computer = yes)

If the condition holds true for a given tuple, then the antecedent is satisfied.

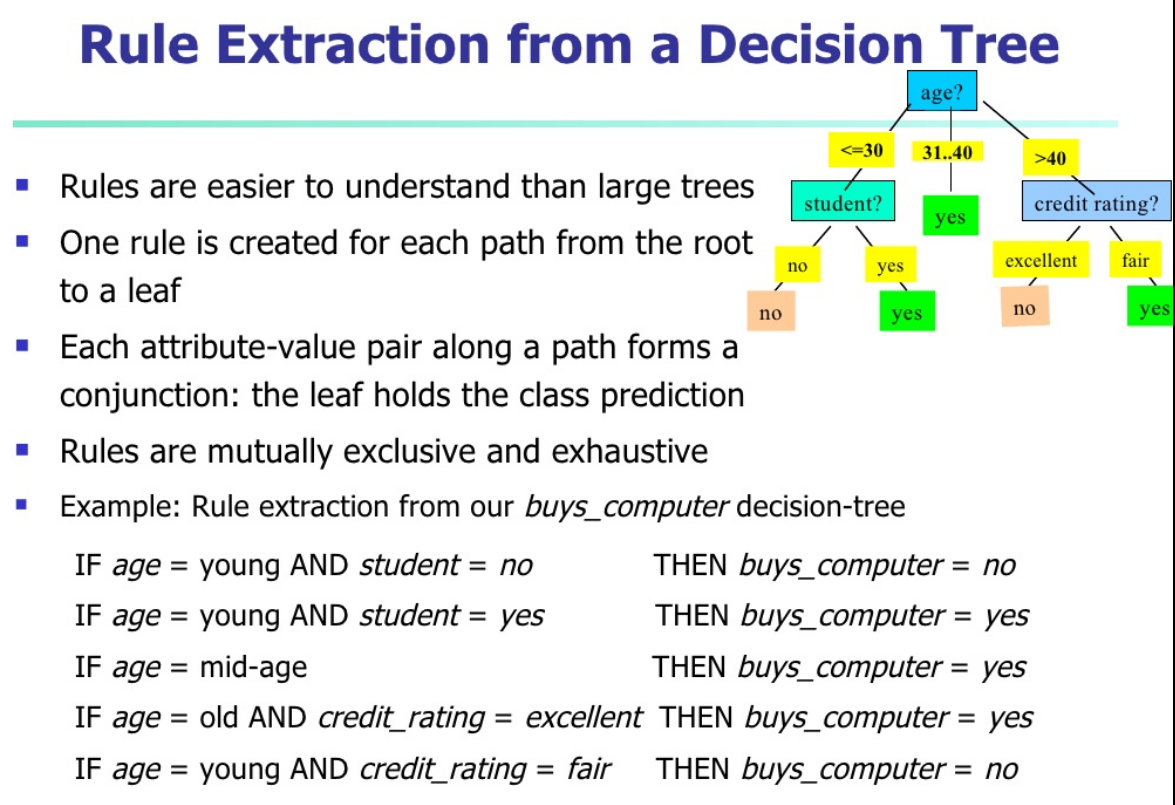
Rule Extraction

Here we will learn how to build a rule-based classifier by extracting IF-THEN rules from a decision tree.

**Points to remember −**

To extract a rule from a decision tree −

* One rule is created for each path from the root to the leaf node.
* To form a rule antecedent, each splitting criterion is logically ANDed.
* The leaf node holds the class prediction, forming the rule consequent.



**Accuracy and Error measures of Classification and prediction**

Measure predictor accuracy: measures how far off the predicted value is from the actual known value **Loss function**: measures the error betw. yi and the predicted value yi ’

Absolute error: | yi – yi’|

Squared error: (yi – yi’) 2

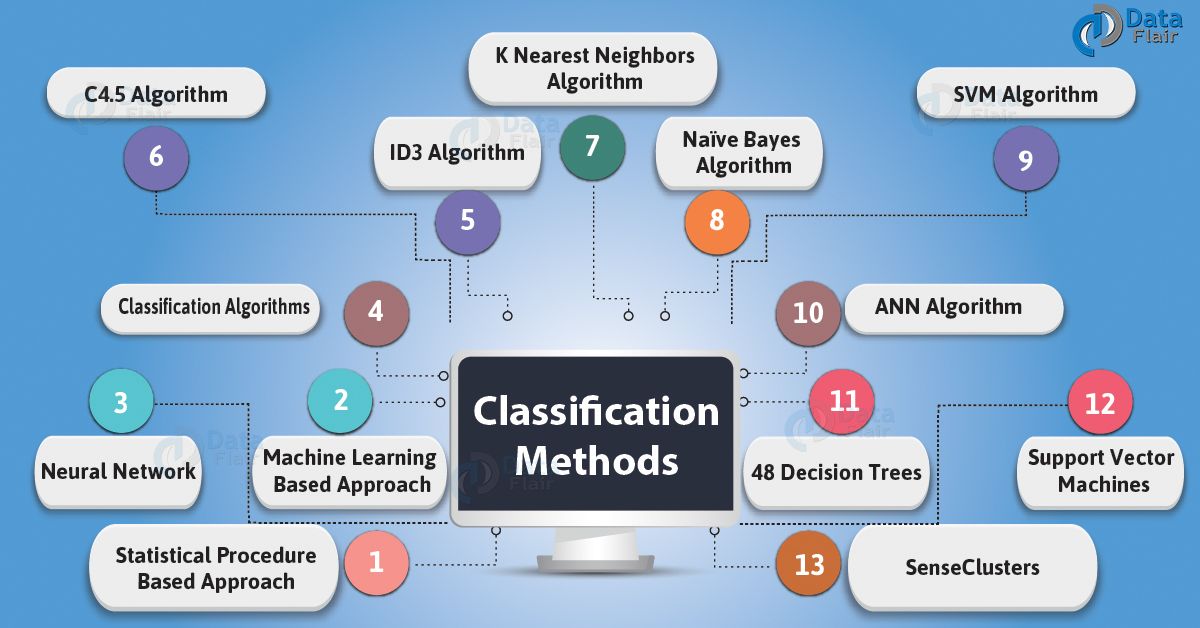
**Test error** (generalization error): the average loss over the test set Mean absolute error: Mean squared error: Relative absolute error: Relative squared error: The mean squared-error exaggerates the presence of outliers Popularly use (square) root mean-square error, similarly, root relative squared error

Bootstrap:

Works well with small data sets Samples the given training tuples uniformly with replacement i.e., each time a tuple is selected, it is equally likely to be selected again and re-added to the training set.

Several boostrap methods, and a common one is .632 boostrap

Other classification types:



* 1. **Genetic Algorithm**: it is based on an analogy to biological evolution An initial population is created consisting of randomly generated rules Each rule is
  2. represented by a string of bits E.g., if A1 and ¬A 2 then C 2 can be encoded as 100 If an attribute has k > 2 values, k bits can be used Based on the notion of survival of the fittest, a new population is formed to consist of the fittest rules and their offsprings The fitness of a rule is represented by its classification accuracy on a set of training examples
  3. **K-nearest neighbor algorithm**

All instances correspond to points in the n-D space The nearest neighbor are defined in terms of Euclidean distance, dist(X1, X2) Target function could be discrete- or real- valued For discrete-valued, k-NN returns the most common value among the k training examples nearest to xq

k-NN for real-valued prediction for a given unknown tuple Returns the mean values of the k nearest neighbors

**Classification Software**

**Weka**  
Weka is a Java based free and open source software licensed under the GNU GPL and available for use on Linux, Mac OS X and Windows. It comprises a collection of machine learning algorithms for data mining. It packages tools for data pre-processing, classification, regression, clustering, association rules and visualisation.

**RapidMiner**  
Rapid Miner is available in both FOSS and commercial editions and is a leading predictive analytic platform. Gartner, the US research and advisory firm, has recognised Rapid Miner and Knife as leaders in the magic quadrant for advanced analytic platforms in 2016. Rapid Miner is helping enterprises embed predictive analysis in their business processes with its user friendly, rich library of data science and machine learning algorithms through its all-in-one programming environments like Rapid Miner Studio.

**Orange**  
Python users playing around with data sciences might be familiar with Orange. It is a Python library that powers Python scripts with its rich compilation of mining and machine learning algorithms for data pre-processing, classification, modelling, regression, clustering and other miscellaneous functions.

* BayesiaLab, includes Bayesian classification algorithms for data segmentation and uses Bayesian networks to automatically cluster the variables.