Classification:

-Form of Data analysis that extracts models describing important data classes
-Data analysis task where a model classifier is construct is constructed to predict class

Prediction:

-Data analysis model
-To predict or fill missing data or values
-Prediction model predict function value than class

Training Data (Train Model)

Applications of Classification

-Credit approval
-Target marketing
-Medical Diagnosis
-Fraud detection

Classification:

Supervised Classification:

-Class label provided
-New data is classified based on training set

Unsupervised Learning:

-Classes labels are unknown in learning phase
-Clustering example

Supervised Learning examples

-SVM
-Logistic Regression
-Naïve Bayes

Unsupervised learning examples

-SVD (Single value decomposition)
-PCA (Principal component Analysis)

Classification steps:

*Model Construction:

-Set of predetermined classes

Model Usage:

-For classifying future or unknown objects

-Each tuple/sample is assumed to belong to a predefined class.

-The set of tuples used for model construction is training set .

-Model is represented as classification rules, decision trees, or mathematical formula.

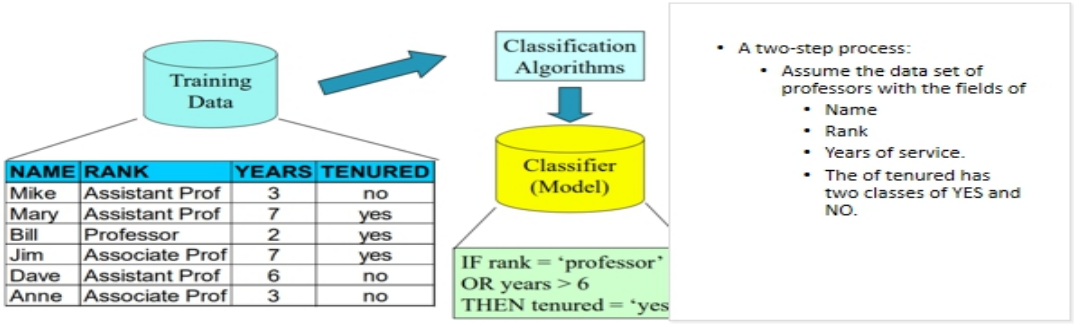
Accuracy Measure

-Known label of test sample is compared with the result from model.

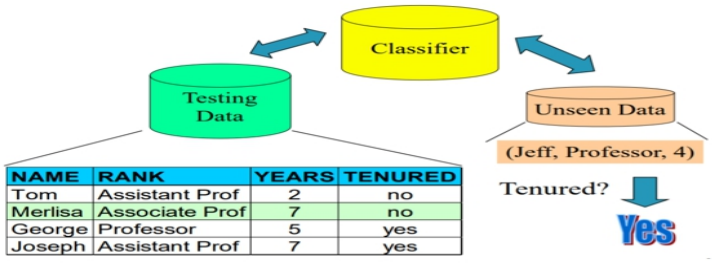
-Test set independent to training set .

-If accuracy is acceptable then use model for unknown class.

Classification: Steps



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

Data preparation:

-Data cleaning.
-Preprocess data in order to reduce noise and handle missing values
-Smoothing .
-Data Filling.

Classification: Issues

Relevance Analysis

-Presence of irrelevant data should be removed
-Correlation analysis to find relevance in data such as Chi Square etc.

Data Transformation

-Normalization:
-Same scale,

Generalization

-Uses of concept hierarchies

Data Reduction

-Binning
-Histogram analysis
-Clustering

Data Mining.

Comparing Classification Methods

How to compare classification/prediction methods?

-Accuracy
-Speed
-Robustness
-Scalability
-Interpretability

Accuracy

-Ability of classifier to predict class label correctly
-How well a given predictor can guess the value of predicted for a new data
-Take algorithm with best accuracy

Speed

-Robustness (have ability to work with noisy data).
-It refers to the ability of classifier or predictor to make correct predictions from given noisy data

Scalability

-Ability to construct the classifier or predictor efficiently; given large amount of data

Interpretability

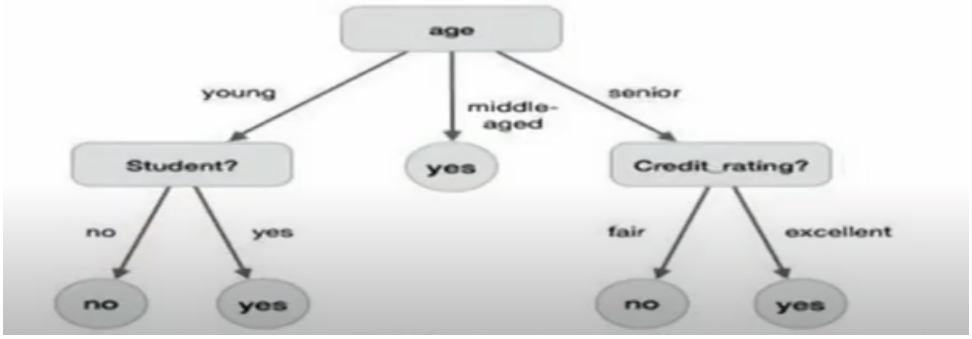
-It refers to what extent the classifier or predictor understands

---

Decision Tree induction:

Decision Tree

-A decision tree is a structure that includes a
-Root node
-Branches
-Leaf Node
-Each internal node denotes a test on an attribute
-Each branch denotes outcome of a test
-And each leaf node holds class label
-The topmost node is root node



Example of Decision

Decision Tree Induction

Decision Tree Induction

-Learning of decision trees from class labeled training tuples

Tree structure:

Each branch represents outcome of the test and the leaf node holds a class label.

Decision Tree Induction

Decision Tree Induction Algorithm

-Construct tree in a top-down recursive divide and conquer manner
-Training examples are at the root at first
-Partition examples recursively based on selected attributes
-Select attributes on the basis of heuristic or statistical measure

Stopping condition

-All samples for a given node belong to the same class
-No remaining attributes for further partitioning
-No samples left

DTI Algorithm

DTI Algorithm

-Generating a decision tree from training tuples of data partitioning D
-Algorithm : Generate_decision_tree (Name of Algo)

-Input:
-Data partition, D, Which is a set of training tuples and their associated class labels
-Atrribute_list, the set of candidate attributes.
-Attribute selection method, procedure to determine the splitting criterion that best partitions that the data tuples into individual classes

Tree pruning

-In machine learning and data mining, pruning is a technique associated with decision trees. Pruning reduces the size of decision trees by removing parts of the tree that do not provide power to classify instances.

Overfitting and Tree pruning

-Overfitting occurs when many branches reflect anomalies due to noise or outliers

-Pruning methods address this problem of overfitting the data.

Approaches

-Prepruning: Halt tree construction early

-Do not split a node if this would result in the goodness measure falling below a threshold

-Difficult to choose an appropriate threshold

Postpruning

Remove branches from a “fully grown” tree

Entropy:

-In information theory Entropy is measure of uncertainty associated with random variable

-Average amount of information needed to identify the class label of a tuple

-Definition:

Lets assume a discrete random variable Y with m number of values

Y = {y1, y2, y3 . . . . ym}

$$H(Y) = -\sum_{i=1}^{m} p_i \log(p_i) \text{ , where } p_i = P(Y = y_i)$$

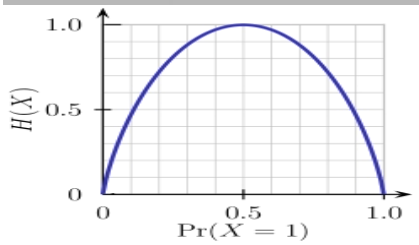
Where H is the entropy of Y.

Entropy

-The result of entropy can be interpreted as
-High Entropy => High Uncertainty
-Lower Entropy => Lower Uncertainty

-Conditional Entropy

$$H(Y|X) = \sum_x p(x) H(Y|X = x)$$



Attribute selection measure

Attribute selection measure

-A heuristic for selecting and splitting criterion that “best” separates a given partition, D, of class labeled training tuples into individual classes.
-Information Gain
-Gain Ratio
-Gini Index

Information Gain:

Information with the highest gain is selected

Let pi be the probability that an arbitrary tuple in D belongs to class Ci, Estimated by |Ci, d|/|D|

Expected information (entropy) needed to classify a tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

PTO

Information needed (after using A to split D into v Partitions) to classify D:
Here A is one column and D is total

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

With respect to one column

Information gained by branching on attribute A

$$Gain(A) = Info(D) - Info_A(D)$$

- Class P: buys_computer = "yes"
- Class N: buys_computer = "no"

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

$$Info(D) = I(9,5) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.940$$

age	p_i	n_i	$I(p_i, n_i)$
<=30	2	3	0.971
31…40	4	0	0
>40	3	2	0.971

$$Info_{age}(D) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) = 0.694$$

$$Gain(income) = 0.029$$
$$Gain(student) = 0.151$$
$$Gain(credit\_rating) = 0.048$$

Gain Ratio:

C4.5 (a successor of ID3) overcomes the problems (normalization to informationgain)

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2(\frac{|D_j|}{|D|})$$

Gain Ratio(A) = Gain(A)/SplitInfo(A)

$$SplitInfo_{income}(D) = -\frac{4}{14} \times \log_2\left(\frac{4}{14}\right) - \frac{6}{14} \times \log_2\left(\frac{6}{14}\right) - \frac{4}{14} \times \log_2\left(\frac{4}{14}\right) = 1.557$$

gain_ratio(income) = 0.029/1.557 = 0.019

The attribute with the maximum gain ratio is selected as the splitting attribute

---

Gini Index

If a data set D contains examples from n classes, gini index , gini(D) is defined as

$$gini(D) = 1 - \sum_{j=1}^{n} p_j^2$$

Where pj is the relative frequency of class j in D. If a data set D is split on A into two subsets D1 and D2, the gini index gini(D) is defined

$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2)$$

Reduction in Impurity

$$\Delta gini(A) = gini(D) - gini_A(D)$$

Total gini index – gini index of column

The attribute provides the smallest gini split (D) (or the largest reduction in impurity) is chosen to split the node

D has 9 tuples in buys_computer = “yes ” and 5 in “no”

$$gini(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459$$

Suppose the attribute income partitions D into 10 in D1: {low, medium} and 4 in D2

$$gini_{income \in \{low, medium\}}(D) = \left(\frac{10}{14}\right) Gini(D_1) + \left(\frac{4}{14}\right) Gini(D_2)$$

$$= \frac{10}{14}\left(1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2\right) + \frac{4}{14}\left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right)$$
$$= 0.443$$
$$= Gini_{income \in \{high\}}(D).$$

$Gini_{\{low,high\}}$ is 0.458; $Gini_{\{medium,high\}}$ is 0.450.

ASM Comparison:

Attribute selection measure comparison

*Information Gain

-Biased towards multivalued attributes
-Gender and ages
-It will prefer multivalued attribute age

Gain Ratio

-Tend to prefer unbalanced splits in which one partition is much smaller than the other

Gini Index

-Biased to Multivalued attributes
-Has difficulty when number of classes is large
-Works best when result is equal sized partitions

Advantages of DTI

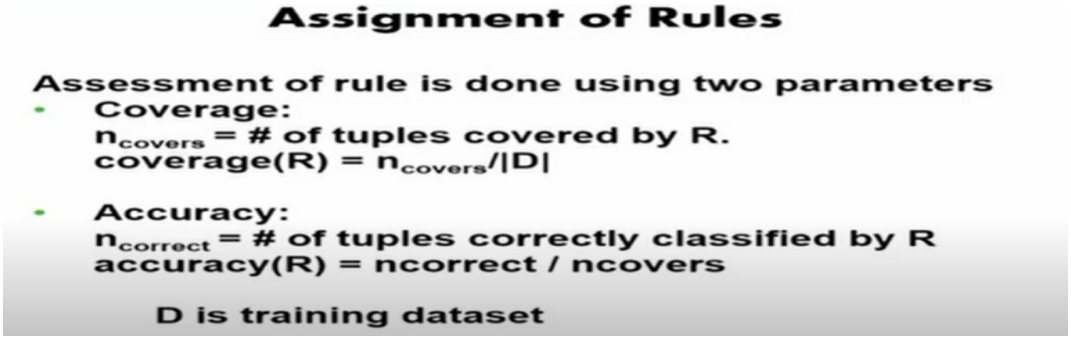
-Relatively faster learning
-Convertible simple and easy to understand classification rules
-Can use SQL queries to access databases
-Comparable classifications accuracy with other methods

Rule based Classification

-Rules are a good way of representing information or bits or knowledge
-If else rules

IF THEN Rules for classification

-Rule based classifiers are classifiers where the learned model is represented as a set of IF-THEN rules.



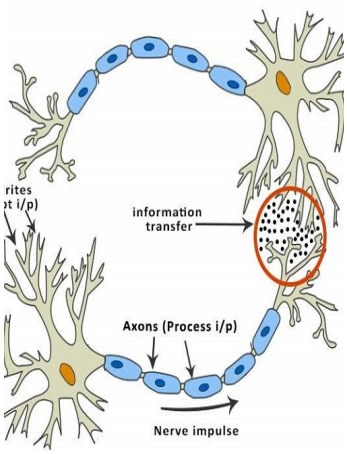
----------Chapeter end----------Prepared by : MEHRAN KHAN----------

CHP:LAST (Artificial Neural Networks)

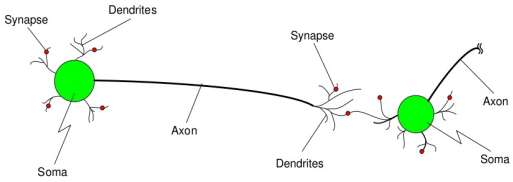
Basic Structure of ANNs :

-The idea of ANNs is based on the belief that working of human brain by making the right connections,
-can be imitated using silicon and wires as living neurons and dendrites
-The human brain is composed of 86 billion nerve cells called neurons
-hey are connected to other thousand cells by Axons
-Stimuli from external environment or inputs from sensory organs are accepted by dendrites.
-These inputs create electric impulses, which quickly travel through the neural network

A neuron can then send the message to other neuron to handle the issue or does not send it forward.



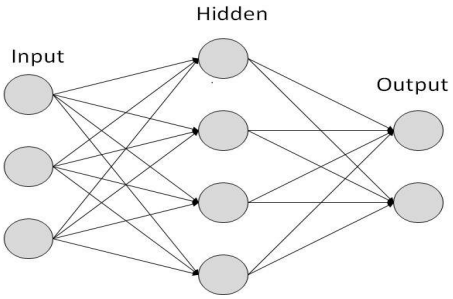
Biological Neural Networks

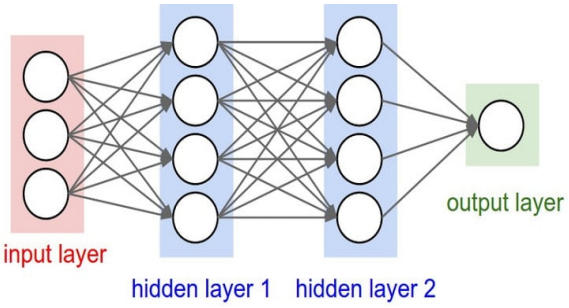


- Two interconnected brain cells (neurons)

ANNS :

-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





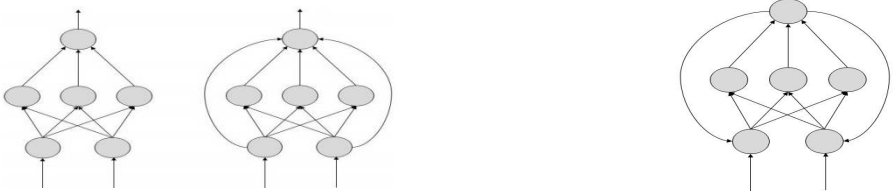
Types of Artificial Neural Networks

There are two Artificial Neural Network topologies

– FeedForward and Feedback

FeedForward ANN:

-In this ANN, the information flow is unidirectional
-A unit sends information to other unit from which it does not receive any information
-There are no feedback loops
-hey are used in pattern generation/recognition/classification
-They have fixed inputs and outputs.



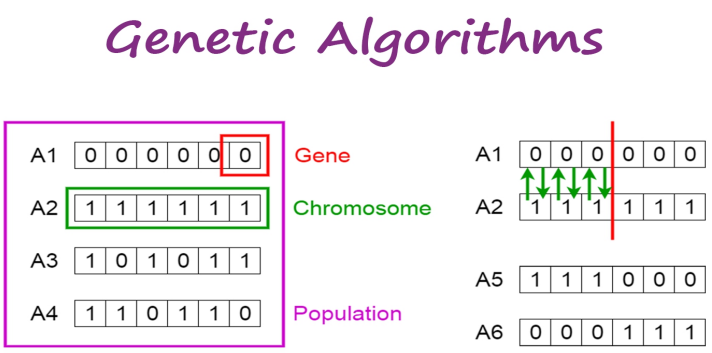
FeedBack ANN Here, feedback loops are allowed . # 3 PIC

Working of ANNs

In the topology diagrams shown, each arrow represents a connection between two neurons and indicates the pathway for the flow of information
Each connection has a weight, an integer number that controls the signal between the two neurons
If the network generates a “good or desired” output, there is no need to adjust the weights
However, if the network generates a “poor or undesired” output or an error, then the system alters the weights in order to improve subsequent results.

Genetic Algorithms

- A genetic algorithm is a search heuristic that is inspired by Charles Darwin's theory of natural evolution
-This algorithm reflects the process of natural selection where the fittest Individuals are selected for reproduction in order to produce offspring of the next generation

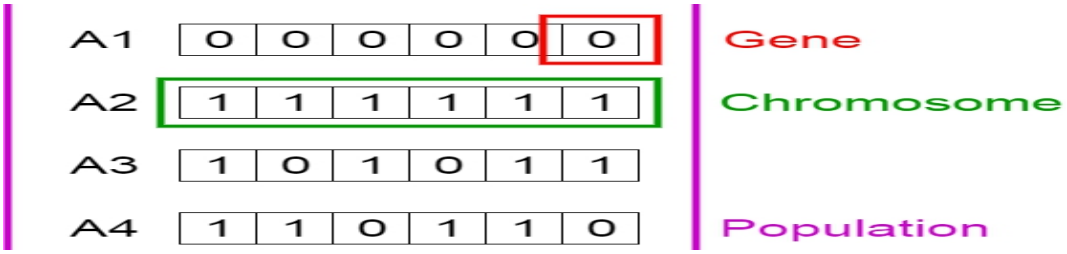


Notion of Natural Selection

-The process of natural selection starts with the selection of fittest individuals from a population
-They produce offspring which inherit the characteristics of the parents and will be added to the next generation
-If parents have better fitness, their offspring will be better than parents and have a better chance at surviving
-This process keeps on iterating and at the end, a generation with the fittest individuals will be found.
-This notion can be applied for a search problem
-We consider a set of solutions for a problem and select the set of best ones out of them.
-Five phases are considered in a genetic algorithm.
-Initial population
-Fitness function
-Selection
-Crossover
-Mutation

Initial Population

-The process begins with a set of individuals which is called a Population.
-Each individual is a solution to the problem you want to solve.
-An individual is characterized by a set of parameters (variables) known as Genes
-Genes are joined into a string to form a Chromosome (solution).
-In a genetic algorithm, the set of genes of an individual is represented using a string, in terms of an alphabet
-Usually, binary values are used (string of 1s and 0s)
-We say that we encode the genes in a chromosome.



Fitness Function

-The fitness function determines how fit an individual is (the ability of an individual to compete with other individuals)
-It gives a fitness score to each individual
-The probability that an individual will be selected for reproduction is based on its fitness score

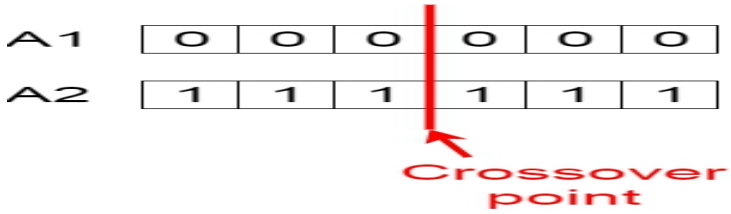
---

Selection

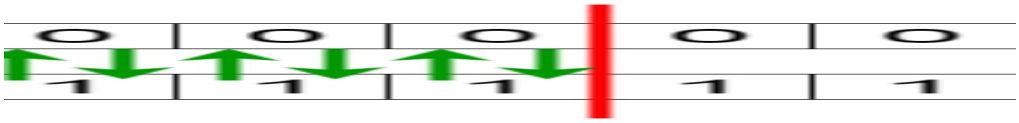
-The idea of selection phase is to select the fittest individuals and let them pass their genes to the next generation
-Two pairs of individuals (parents) are selected based on their fitness scores
-Individuals with high fitness have more chance to be selected for reproduction

Crossover:

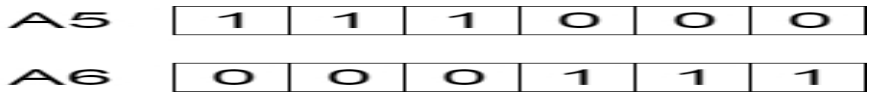
-Crossover is the most significant phase in a genetic algorithm.
-For each pair of parents to be mated, a crossover point is chosen at random from within the genes
-For example, consider the crossover point to be 3 as shown below.



Offspring are created by exchanging the genes of parents among themselves until the crossover point is reached

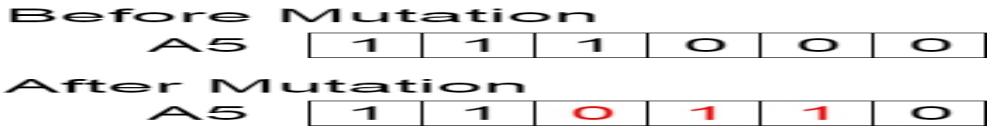


The new offspring are added to the population



Mutation

-In certain new offspring formed, some of their genes can be subjected to a mutation with a low random probability
-This implies that some of the bits in the bit string can be flipped



Association rules

-Association Rule is one of the very important concepts of machine learning being used in market basket analysis
-Market Basket Analysis is the study of customer transaction databases to determine dependencies between the various items they purchase at different times
-Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases.
-It identifies frequent if-then associations called association rules which consists of an antecedent (if) and a consequent (then).
-For example: “If tea and milk, then sugar” (“If tea and milk are purchased, then sugar would also be bought by the customer”)
- Antecedent: Tea and Milk
– Consequent: Sugar.
-There are three common metrics to measure association
-Support
-Confidence
-Lift

Support: is an indication of how frequently the items appear in the data. Mathematically, support is the fraction of the total number of transactions in which the item set occurs.

$$Support(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Total\ number\ of\ transactions}$$

Confidence : indicates the number of times the if-then statements are found true. Confidence is the conditional probability of occurrence of consequent given the antecedent.

$$Confidence(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Transactions\ containing\ X}$$

Lift :can be used to compare confidence with expected confidence. This says how likely item Y is purchased when item X is purchased, while controlling for how popular item Y is. Mathematically,

$$Lift(\{X\} \rightarrow \{Y\}) = \frac{(Transactions\ containing\ both\ X\ and\ Y)/(Transactions\ containing\ X)}{Fraction\ of\ transactions\ containing\ Y}$$