GL BAJAJ
Group of Institutions
Mathura
FIND YOUR SPARK
Approved by AICTE, CoA & Affiliated to AKTU

Data Analytics

3rd year

Notes for unit -2

TOPICS COVERED IN UNIT 2

Data Analysis:

- Regression modeling,
- multivariate analysis,
- Bayesian modeling, inference
- and Bayesian networks,
- support vector and kernel methods,
- analysis of time series: linear
- systems analysis;
- nonlinear dynamics,
- rule induction,

Neural Networks:

- learning and generalisation,
- competitive learning,
- principal component analysis
- neural networks,

Fuzzy Logic:

- extracting fuzzy models from data,
- fuzzy decision trees,
- stochastic search methods.

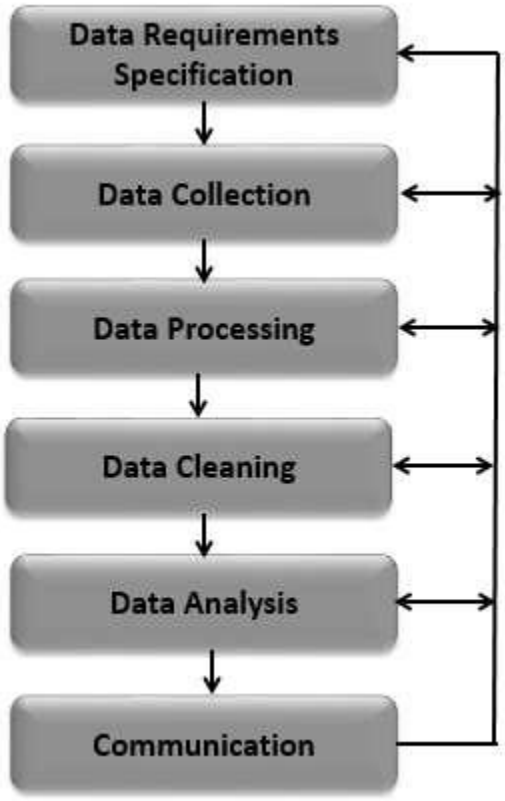
Data Analysis:

Data Analysis is the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data.

Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making

A simple example of Data analysis is **whenever we take any decision in our day-to-day life** is by thinking about what happened last time or what will happen by choosing that particular decision. This is nothing but analyzing our past or future and making decisions based on it.

Data Analysis Process consists of the following phases that are iterative in nature −

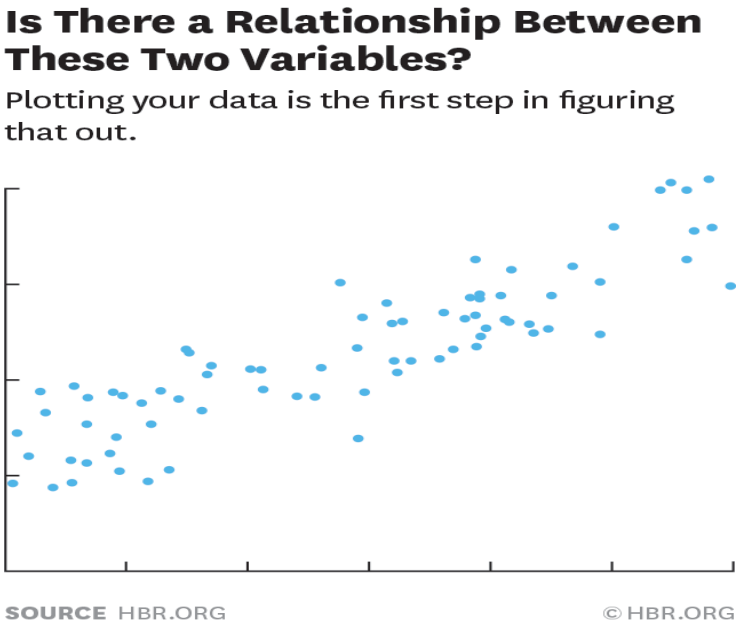


Regression modeling

Regression analysis is a way of mathematically sorting out which of those variables does indeed have an impact. It answers the questions: Which factors matter most? Which can we ignore? How do those factors interact with each other? And, perhaps most importantly, how certain are we about all of these factors?

In regression analysis, those factors are called variables. You have your **dependent variable** — the main factor that you're trying to understand or predict. In Redman's example above, the dependent variable is monthly sales. And then you have your **independent variables** — the factors you suspect have an impact on your dependent variable.

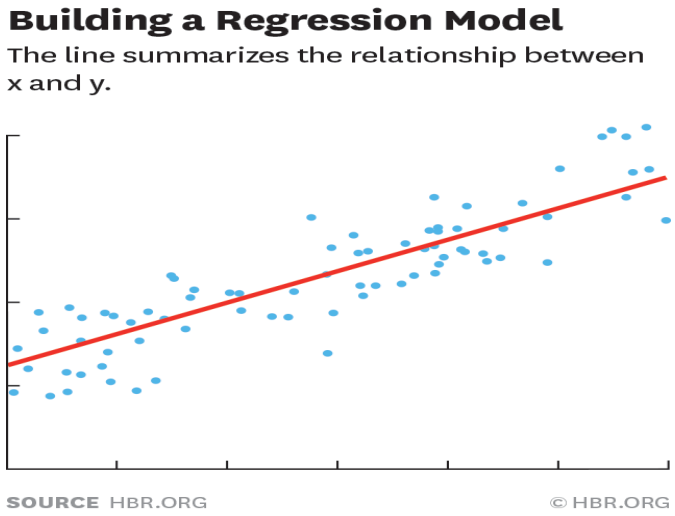
In order to conduct a regression analysis, you gather the data on the variables in question. (Reminder: you likely don't have to do this yourself, but it's helpful for you to understand the process your data analyst colleague uses.) You take all of your monthly sales numbers for, say, the past three years and any data on the independent variables you're interested in. So, in this case, let's say you find out the average monthly rainfall for the past three years as well. Then you plot all of that information on a chart that looks like this:



The y-axis is the amount of sales (the dependent variable, the thing you're interested in, is always on the y-axis) and the x-axis is the total rainfall. Each blue dot represents one month's data—how much it rained that month and how many sales you made that same month.

Glancing at this data, you probably notice that sales are higher on days when it rains a lot. That's interesting to know, but by how much? If it rains 3 inches, do you know how much you'll sell? What about if it rains 4 inches?

Now imagine drawing a line through the chart above, one that runs roughly through the middle of all the data points.



Regression analysis includes several variations, such as linear, multiple linear, and nonlinear. The most common models are simple linear and multiple linear. Nonlinear regression analysis is commonly used for more complicated data sets in which the dependent and independent variables show a nonlinear relationship.

Regression analysis offers numerous applications in various disciplines, including finance.

Regression Analysis – Linear Model Assumptions

Linear regression analysis is based on six fundamental assumptions:

1. The dependent and independent variables show a linear relationship between the slope and the intercept.
2. The independent variable is not random.
3. The value of the residual (error) is zero.
4. The value of the residual (error) is constant across all observations.
5. The value of the residual (error) is not correlated across all observations.
6. The residual (error) values follow the normal distribution.

Regression Analysis – Simple Linear Regression

Simple linear regression is a model that assesses the relationship between a dependent variable and an independent variable. The simple linear model is expressed using the following equation:

$$\mathbf{Y = a + bX + \epsilon}$$

Where:

- **Y** – Dependent variable
- **X** – Independent (explanatory) variable
- **a** – Intercept
- **b** – Slope
- **ϵ** – Residual (error)

Regression Analysis – Multiple Linear Regression

Multiple linear regression analysis is essentially similar to the simple linear model, with the exception that multiple independent variables are used in the model. The mathematical representation of multiple linear regression is:

$$\mathbf{Y = a + bX_1 + cX_2 + dX_3 + \epsilon}$$

Where:

- **Y** – Dependent variable
- $\mathbf{X_1, X_2, X_3}$ – Independent (explanatory) variables
- **a** – Intercept
- **b, c, d** – Slopes
- **ϵ** – Residual (error)

Multiple linear regression follows the same conditions as the simple linear model. However, since there are several independent variables in multiple linear analysis, there is another mandatory condition for the model:

- **Non-collinearity:** Independent variables should show a minimum correlation with each other. If the independent variables are highly correlated with each other, it will be difficult to assess the true relationships between the dependent and independent variables.

Regression Analysis in Finance

Regression analysis comes with several applications in finance. For example, the statistical method is fundamental to the Capital Asset Pricing Model (CAPM). Essentially, the CAPM equation is a model that determines the relationship between the expected return of an asset and the market risk premium.

The analysis is also used to forecast the returns of securities, based on different factors, or to forecast the performance of a business. Learn more forecasting methods in CFI's Budgeting and Forecasting Course!

1. Beta and CAPM

In finance, regression analysis is used to calculate the Beta (volatility of returns relative to the overall market) for a stock. It can be done in Excel using the Slope function.

Regression Tools

Excel remains a popular tool to conduct basic regression analysis in finance, however, there are many more advanced statistical tools that can be used.

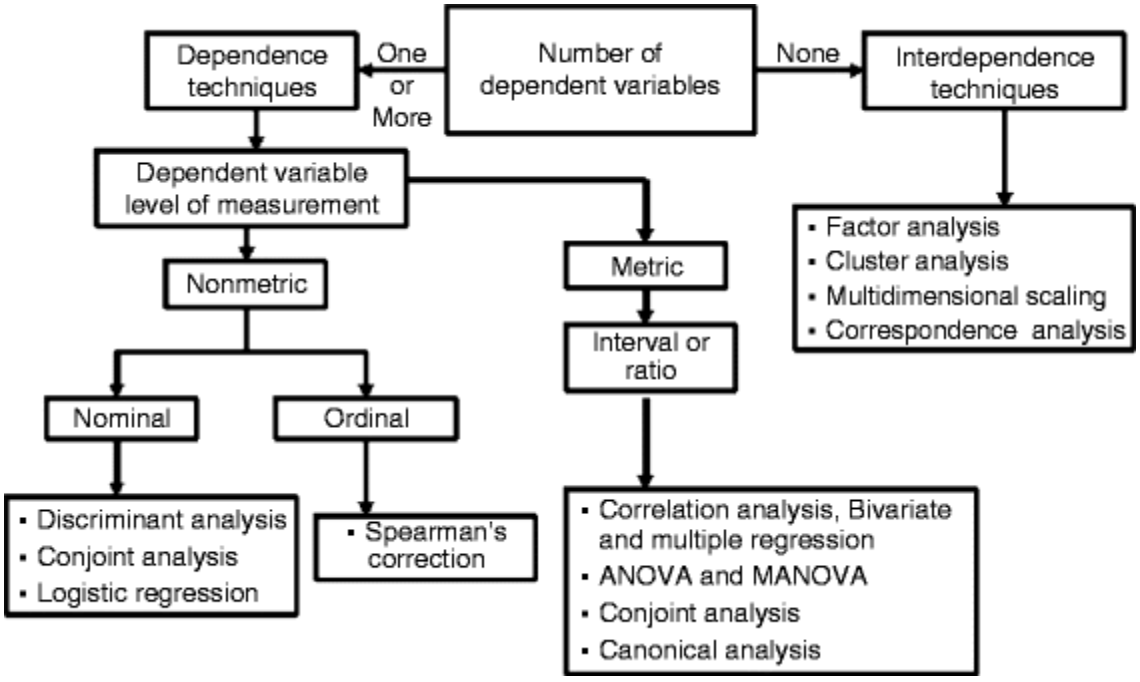
Python and R are both powerful coding languages that have become popular for all types of financial modeling, including regression. These techniques form a core part of data science and machine learning where models are trained to detect these relationships in data.

Learn more about regression analysis, Python, and Machine Learning in CFI's Business Intelligence & Data Analysis certification.

Multivariate analysis (**MVA**) is a Statistical procedure for analysis of data involving more than one type of measurement or observation. It may also mean solving problems where more than one dependent variable is analyzed simultaneously with other variables.

Classification Chart of Multivariate Techniques

Selection of the appropriate multivariate technique depends upon-



Most of multivariate analysis deals with **estimation, confidence sets, and hypothesis testing** for means, variances, covariances, correlation coefficients, and related, more complex population characteristics.

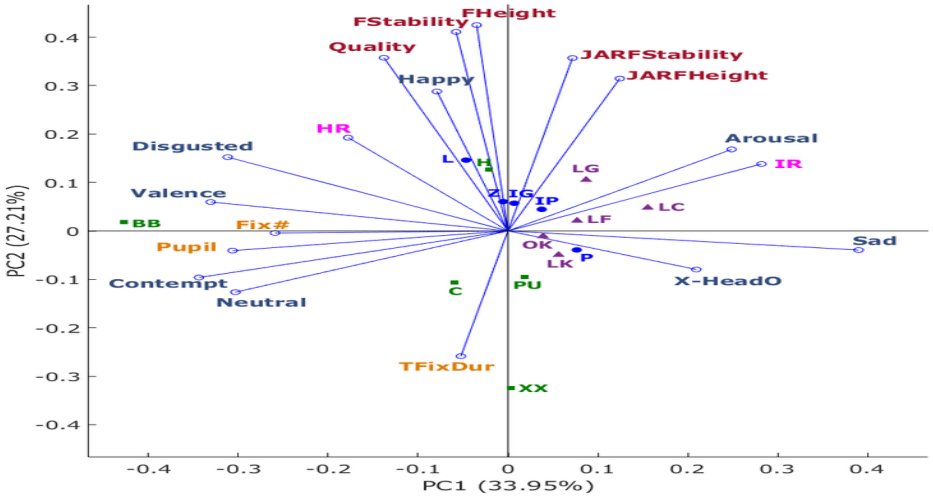
Multivariate Data Analysis Techniques

There are many techniques of Multivariate Analysis starting with quality of the data to structural equation modelling, each one of the techniques has its own purpose, and are used depending on the data and the type of outcome realized by the data analyst. These techniques provide statistical data given a specific data set but requires caution when interpreting and putting them to use remember as I always say people do the most important part than what technology does for us.

The techniques are as follows:

- Multiple Regression Analysis
- Discriminant Analysis
- Multivariate Analysis of Variance (MANOVA)
- Factor Analysis
- Cluster Analysis
- Canonical Correlation
- Classification Analysis
- Principal Component Analysis

There are two main factor analysis methods: **common factor analysis**, which extracts factors based on the variance shared by the factors, and principal component analysis, which extracts factors based on the total variance of the factors. A example of multivariate analysis for a school data.

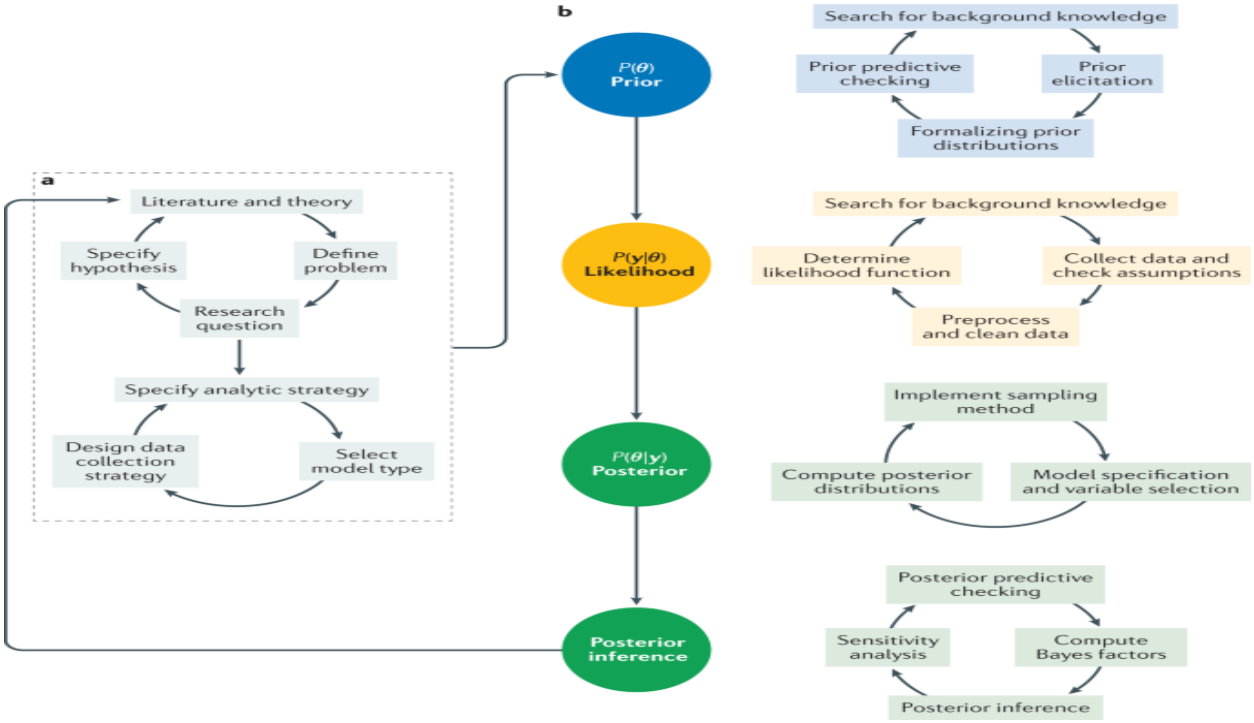


Bayesian model

A Bayesian model is **a statistical model where you use probability to represent all uncertainty within the model**, both the uncertainty regarding the output but also the uncertainty regarding the input (aka parameters) to the model.

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}$$

A posterior distribution comprises a prior distribution about a parameter and a likelihood model providing information about the parameter based on observed data. Depending on the chosen prior distribution and likelihood model, the posterior distribution is either available analytically or approximated by, for example, one of the Markov chain Monte Carlo (MCMC) methods.



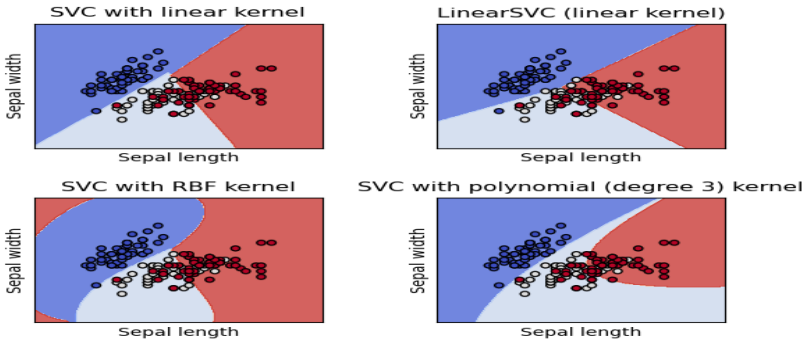
Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection.

The advantages of support vector machines are:

- Effective in high dimensional spaces.
- Still effective in cases where number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

The disadvantages of support vector machines include:

- If the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernel functions and regularization term is crucial.
- SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation (see Scores and probabilities, below).



Support vector machines (SVMs) are **a set of supervised learning methods used for classification, regression and outliers detection**. The advantages of support vector machines are: Effective in high dimensional spaces. Still effective in cases where number of dimensions is greater than the number of samples.

Kernels Methods are employed in **SVM** (Support Vector Machines) which are used in classification and regression problems. The SVM uses what is called a “Kernel Trick” where the data is transformed and an optimal boundary is found for the possible outputs.

In machine learning, **kernel machines** are a class of algorithms for pattern analysis, whose best known member is the support-vector machine (SVM). The general task of pattern analysis is to find and study general types of relations (for example clusters, rankings, principal components, correlations, classifications) in datasets. For many algorithms that solve these tasks, the data in raw representation have to be explicitly transformed into feature vector representations via a user-specified *feature map*: in contrast, kernel methods require only a user-specified *kernel*, i.e., a similarity function over pairs of data points in raw representation.

Kernel methods owe their name to the use of kernel functions, which enable them to operate in a high-dimensional, *implicit* feature space without ever computing the coordinates of the data in that space, but rather by simply computing the inner products between the images of all pairs of data in the feature space. This operation is often computationally cheaper than the explicit computation of the coordinates. This approach is called the "**kernel trick**. Kernel functions have been introduced for sequence data, graphs, text, images, as well as vectors.

Applications

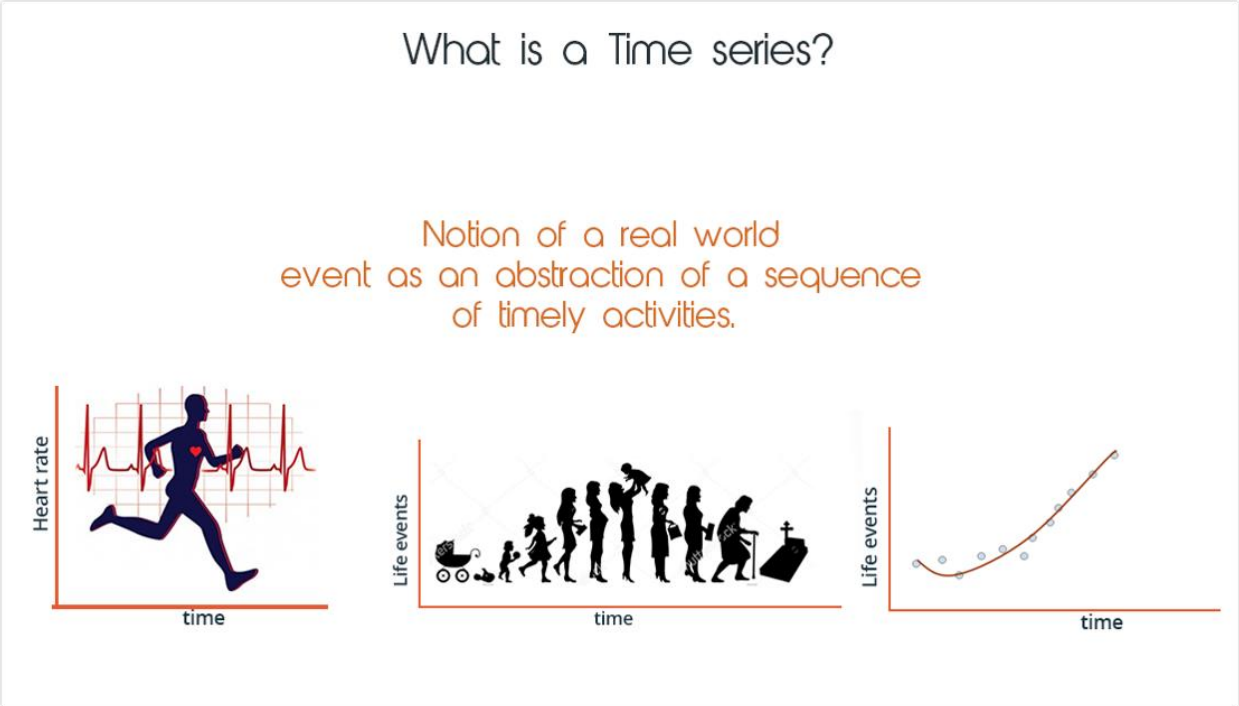
Application areas of kernel methods are diverse and include geostatistics kriging, inverse distance weighting, 3D reconstruction, bioinformatics, chemoinformatics, information extraction and handwriting recognition.

Time series

A time series is a sequence of data points that occur in successive order over some period of time. This can be contrasted with cross-sectional data, which captures a point-in-time.

Suppose you wanted to analyze a time series of daily closing stock prices for a given stock over a period of one year. You would obtain a list of all the closing prices for the stock from each day for the past year and list them in chronological order. This would be a one-year daily closing price time series for the stock.

Delving a bit deeper, you might analyze time series data with technical analysis tools to know whether the stock's time series shows any seasonality. This will help to determine if the stock goes through peaks and troughs at regular times each year. Analysis in this area would require taking the observed prices and correlating them to a chosen season. This can include traditional calendar seasons, such as summer and winter, or retail seasons, such as holiday seasons.

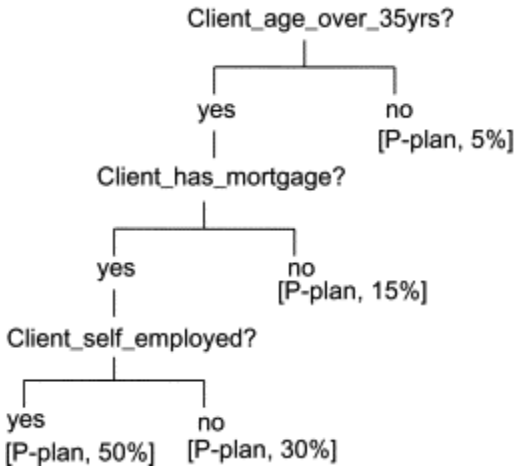


Time series data can be classified into two types:

- Measurements gathered at regular time intervals (metrics)
- Measurements gathered at irregular time intervals (events)

Rule induction

Rule induction is **an area of machine learning in which formal rules are extracted from a set of observations**. The rules extracted may represent a full scientific model of the data, or merely represent local patterns in the data.



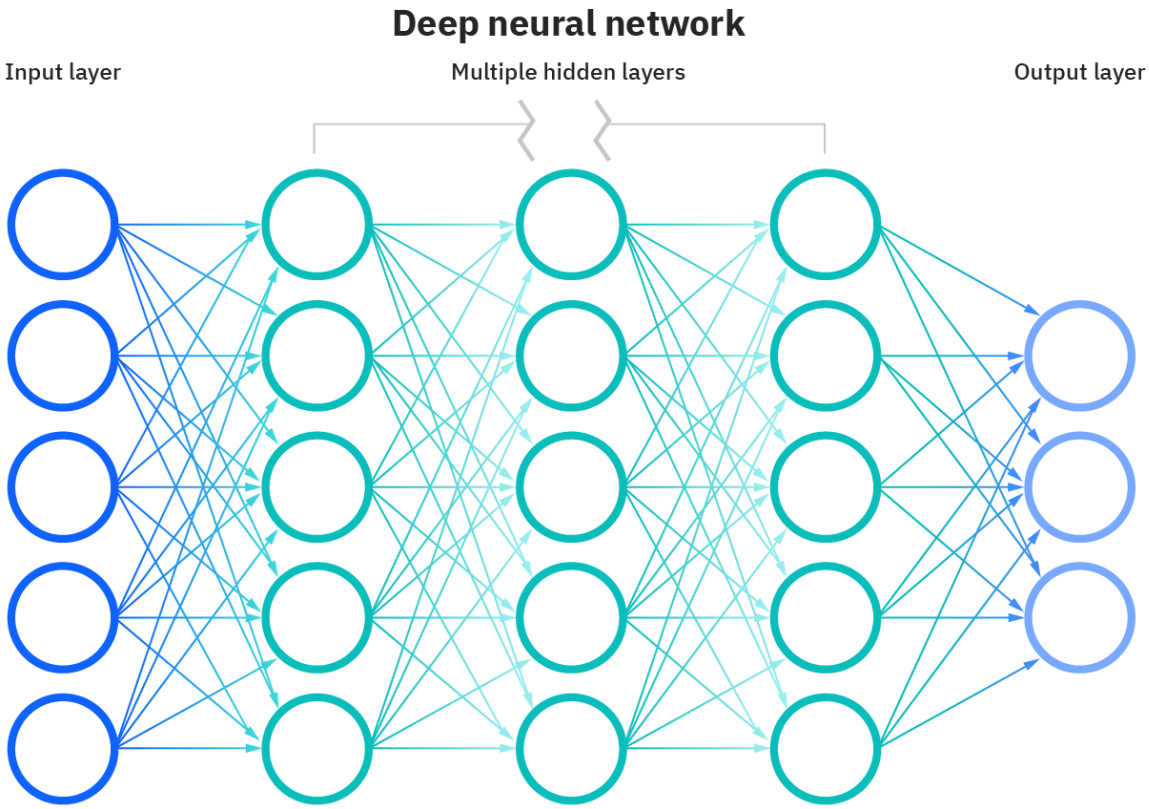
illustrates a tree induction applied to a set of input variables with respect to an output variable. The input variables are client_age_over_35yrs, client_has_mortgage, and client_self_employed, and the output variable is P-plan (pension plan). The percentage after P-plan indicates the percentage of clients at the given leaf node who actually have a pension plan. As observed earlier, the variable highest in the tree is the most general, which in this case is client_age_over_35yrs; that is, the client's age is a key aspect that determines whether they contract a pension plan or not. The variable lowest in the tree is client_self_employed; that is, the type of employment is a more specific criterion related to contracting a pension plan. The other two input variables were given to the tree induction technique, but they were pruned from the tree because they didn't reach the minimum information support level. The information support level can be considered as a sort of relevance measure or correlation with the business objective (output label); hence the algorithm uses this threshold to decide whether or not to include a variable in the tree data model.

Rule induction is an area of machine learning in which formal **rule**s are extracted from a set of observations. The **rule**s extracted may represent a full scientific model of the data, or merely represent local patterns in the data.

Neural networks

Neural networks, also known as artificial neural networks (ANNs) or simulated neural networks (SNNs), are a subset of machine learning and are at the heart of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another.

Artificial neural networks (ANNs) are comprised of a node layers, containing an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network.



What is learning in neural?

Learning rule or Learning process is **a method or a mathematical logic**. It improves the Artificial Neural Network's performance and applies this rule over the network. Thus learning rules updates the weights and bias levels of a network when a network simulates in a specific data environment.

An artificial neural network's learning rule or learning process is a method, mathematical logic or algorithm which improves the network's performance and/or training time.

Depending upon the process to develop the network there are three main models of machine learning:

Unsupervised learning.

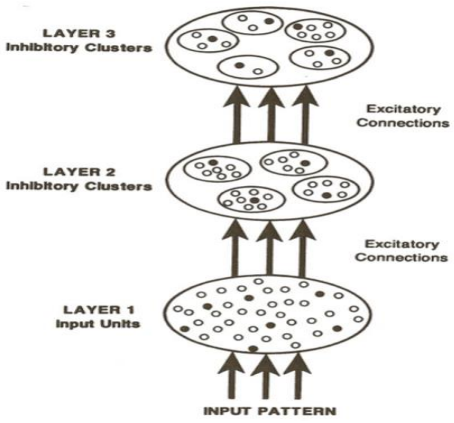
Supervised learning.

- **Supervised Learning**
  The learning algorithm would fall under this category if the desired output for the network is also provided with the input while training the network. By providing the neural network with both an input and output pair it is possible to calculate an error based on it's target output and actual output. It can then use that error to make corrections to the network by updating it's weights.

- **Unsupervised Learning**
  In this paradigm the neural network is only given a set of inputs and it's the neural network's responsibility to find some kind of pattern within the inputs provided without any external aid. This type of learning paradigm is often used in data mining and is also used by many recommendation algorithms due to their ability to predict a user's preferences based on the preferences of other similar users it has grouped together.

- **Competitive learning**

  Competitive learning is **a type of unsupervised learning model used in machine learning and artificial intelligence systems**. Some of the interesting new formats of machine learning projects are partially based on competitive learning include self-organizing component neural networks

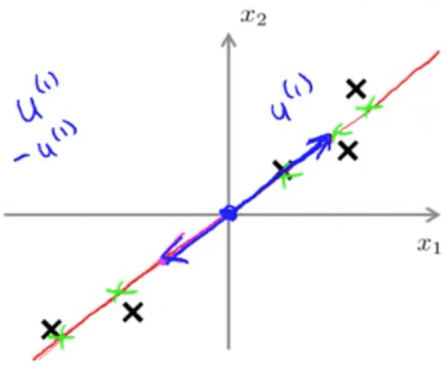


Principal Component Analysis?

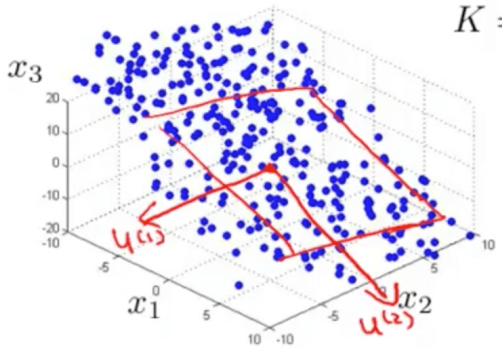
Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process.

Principal Component Analysis (PCA) problem formulation



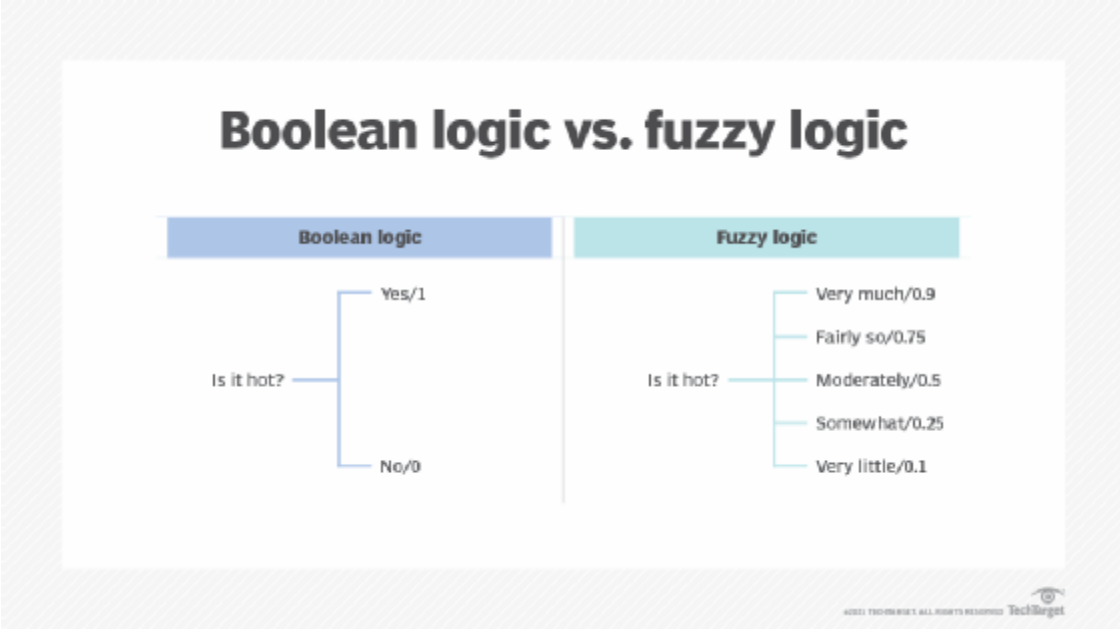
$3D \to 2D$
$K = 2$



Reduce from 2-dimension to 1-dimension: Find a direction (a vector $u^{(1)} \in \mathbb{R}^n$) onto which to project the data so as to minimize the projection error.

Reduce from n-dimension to k-dimension: Find k vectors $u^{(1)}, u^{(2)}, \ldots, u^{(k)}$ onto which to project the data, so as to minimize the projection error.

Fuzzy logic



Fuzzy logic is **an approach to computing based on "degrees of truth"** rather than the usual "true or false" (1 or 0) Boolean logic on which the modern computer is based. ... It may help to see fuzzy logic as the way reasoning really works and binary, or Boolean, logic is simply a special case of it.

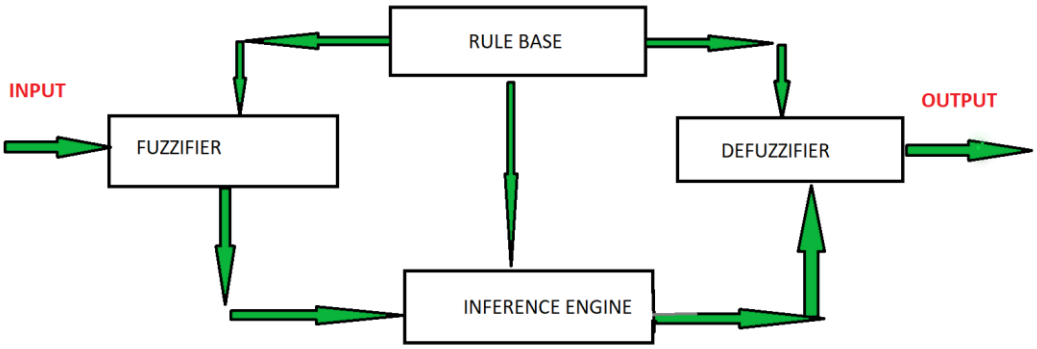
What is fuzzy logic used for?

Fuzzy logic has been used in numerous applications such as **facial pattern recognition, air conditioners, washing machines, vacuum cleaners**, antiskid braking systems, transmission systems, control of subway systems and unmanned helicopters, knowledge-based systems for multiobjective optimization of power systems,

Its Architecture contains four parts :

- RULE BASE: It contains the set of rules and the IF-THEN conditions provided by the experts to govern the decision-making system, on the basis of linguistic information. Recent developments in fuzzy theory offer several effective methods for the design and tuning of fuzzy controllers. Most of these developments reduce the number of fuzzy rules.
- FUZZIFICATION: It is used to convert inputs i.e. crisp numbers into fuzzy sets. Crisp inputs are basically the exact inputs measured by sensors and passed into the control system for processing, such as temperature, pressure, rpm's, etc.
- INFERENCE ENGINE: It determines the matching degree of the current fuzzy input with respect to each rule and decides which rules are to be fired according to the input field. Next, the fired rules are combined to form the control actions.

- DEFUZZIFICATION: It is used to convert the fuzzy sets obtained by the inference engine into a crisp value. There are several defuzzification methods available and the best-suited one is used with a specific expert system to reduce the error.



FUZZY LOGIC ARCHITECTURE

Membership function

Definition: A graph that defines how each point in the input space is mapped to membership value between 0 and 1. Input space is often referred to as the universe of discourse or universal set (u), which contains all the possible elements of concern in each particular application.

There are largely three types of fuzzifiers:

- Singleton fuzzifier
- Gaussian fuzzifier
- Trapezoidal or triangular fuzzifier

extracting fuzzy models from data,

Fuzzy-Logic theory has introduced a framework whereby human knowledge can be formalized and used by machines in a wide variety of applications, ranging from cameras to trains. The basic ideas that we discussed in the earlier posts were concerned with only this aspect with regards to the use of Fuzzy Logic-based systems; that is the application of human experience into machine-driven applications. While there are numerous instances where such techniques are relevant; there are also applications where it is challenging for a human user to articulate the knowledge that they hold. Such applications include driving a car or recognizing images. Machine learning techniques provide an excellent platform in such circumstances, where sets of inputs and corresponding

outputs are available, building a model that provides the transformation from the input data to the outputs using the available data.

Step 1 — Divide the input and output spaces into fuzzy regions.

Step 2 — Generate Fuzzy Rules from data.

Step 3 — Assign a degree to each rule

Step 4 — Create a Combined Fuzzy Rule Base

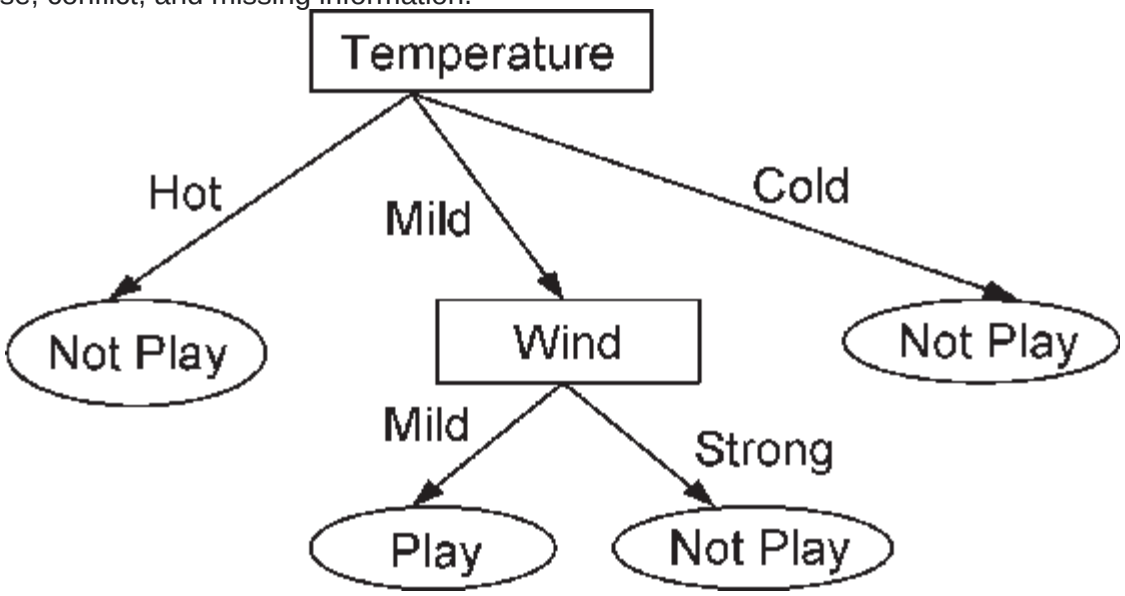
Step 5 — Determine a mapping based on the Combined Fuzzy Rule Base.

We now define the centre of a fuzzy region as the point that has the smallest absolute value among all points at which the membership function for this region is equal to 1 as illustrated below;

- The fuzzy system is generated from the test data directly.
- The sets were created using the recommendation in the original paper, that is evenly spaced. It is, however, interesting to see the effects of changing this method. One idea is to have sets created around the dataset mean with a spread relatable to the standard deviation — this might be investigated in a future post.
- The system created does not cater for categorical data implicitly, and this is a future improvement that can affect the performance of the system considerably in real-life scenarios.

Fuzzy logic decision tree

A fuzzy decision tree induction method, which is based on the reduction of classification ambiguity with fuzzy evidence, is developed. Fuzzy decision trees represent **classification knowledge more naturally to the way of human thinking** and are more robust in tolerating imprecise, conflict, and missing information.



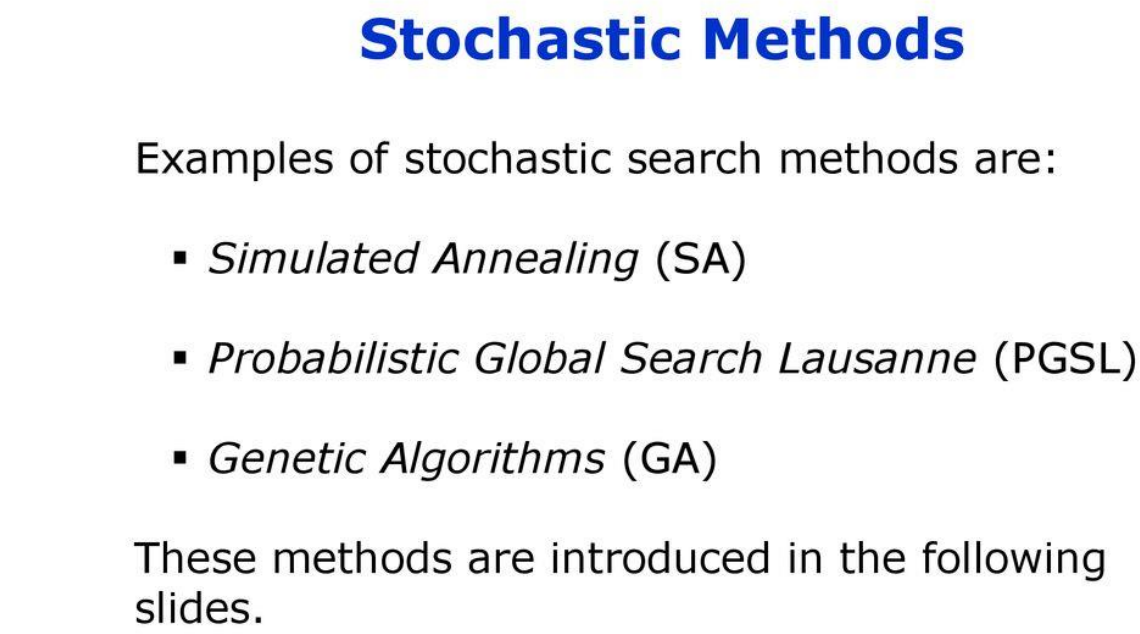
A decision tree is a very specific type of probability tree that enables you to make a decision about some kind of process. For example, you might want to **choose between manufacturing item A or item B, or investing in choice 1, choice 2, or choice 3**.

Difference between Neural Network And Fuzzy Logic

Neural Network	Fuzzy Logic
This system can not easily modified.	This system can easily modified.
It trains itself by learning from data set	Everything must be defined explicitly.
It is complex than fuzzy logic.	It is simpler than neural network.
It helps to perform predictions.	It helps to perform pattern recognition.
Difficult to extract knowledge.	Knowledge can easily extracted.
It based on learning.	It doesn’t base on learning.

Stochastic search method:

The behavior and performance of many machine learning algorithms are referred to as stochastic. Stochastic refers to **a variable process where the outcome involves some randomness and has some uncertainty**. ... A variable or process is stochastic if there is uncertainty or randomness involved in the outcomes.



Sophisticated search techniques form the backbone of modern machine learning and data analysis. Computer systems that are able to extract information from huge data sets (data mining), to recognize patterns, to do classification, or to suggest diagnoses, in short, systems that are adaptive and — to some extent — able to learn, fundamentally rely on effective and

efficient search techniques. The ability of organisms to learn and adapt to signals from their environment is one of the core features of life. Technically, any adaptive system needs some kind of search operator in order to explore a feature space which describes all possible configurations of the system. Usually, one is interested in “optimal” or at least close to “optimal” configurations defined with respect to a specific application domain: the weight settings of a neural network for correct classification of some data, parameters that describe the body shape of an airplane with minimum drag, a sequence of jobs assigned to a flexible production line in a factory resulting in minimum idle time for the machine park, the configuration for a stable bridge with minimum weight or minimum cost to build and maintain, or a set of computer programs that implement a robot control task with a minimum number of commands.