Unit-II

Statistical Thinking in the Age of Big Data

Big Data phrase means three things. First, it is a bundle of technologies. Second, it is a potential revolution in measurement. Third, it is a point of view, or philosophy, about how decisions will be and perhaps should be made in the future.

When we're developing your skill set as a data scientist, - data preparation and munging, modelling, coding, visualization, and communication and will begin by getting grounded in statistical inference.

Statistical Inference

The world we live in is Complex, Random and Uncertain, but it's big data generating machine. Data represents the traces of the real-world processes, and exactly which traces we gather are decided by our data collection or sampling method. After separating the process from the data collection, we can see clearly that there are two sources of *Randomness and Uncertainty*. Namely, the *Randomness and Uncertainty* underlying the process itself, and the uncertainty associated with your underlying data collection methods.

This overall process of going from the world to the data, and then from the data back to the world, is the field of *statistical inference*. Data could be represented as mathematical models or functions of the data, known as statistical estimators.

Statistical Inference is the discipline that concerns itself with the development of procedures, methods, and theorems that allow us to extract meaning and information from data that has been generated by stochastic (random) processes.

Population and Samples

In statistical inference, *Population* isn't used to simply describe only people. It could be any set of objects or units, such as tweets or photographs or stars.

If we could measure the characteristics or extract characteristics of all those objects, we'd have a complete set of *Observations*, and the convention is to use to represent the total number of observations in the population.

When we take a *Sample*, we take a subset of the units of size to examine theobservations to draw conclusions and make inferences about the population.

Sampling mechanism can introduce *biases* into the data, and distort it, so that the subset [sample] is not exact representation of the population. Once that happens, any conclusion/inference you draw will simply be wrong and distorted.

For example, Population refers Entire data set i.e., collection of records. Observation refers a single record. Sample refers set of observations from Population without bias and represent the given population.

In the age of Big Data, where we can record all users' actions all the time, don't we observe everything? Is there really still this notion of population and sample?

There are multiple aspects of this that need to be addressed.

- Sampling solves some engineering challenges
- Bias
- Sampling distribution
- New kinds of data

Statistical Modeling

Statistical model is a mathematical representation of observed data. The mathematical expressions will be general enough that they must include parameters, but the values of these parameters are not yet known. Statistical model can be thought of as a statistical assumption with a certain property: that the assumption allows us to calculate the probability of any event.

In mathematical terms, a statistical model is usually thought of as a pair (S,P), where S is the set of possible observations, i.e. the sample space, and P is a set of probability distributions on S.

In mathematical expressions, the convention is to use Greek letters for parameters and Latin letters for data.

$$y = \beta_0 + \beta_1 * x$$

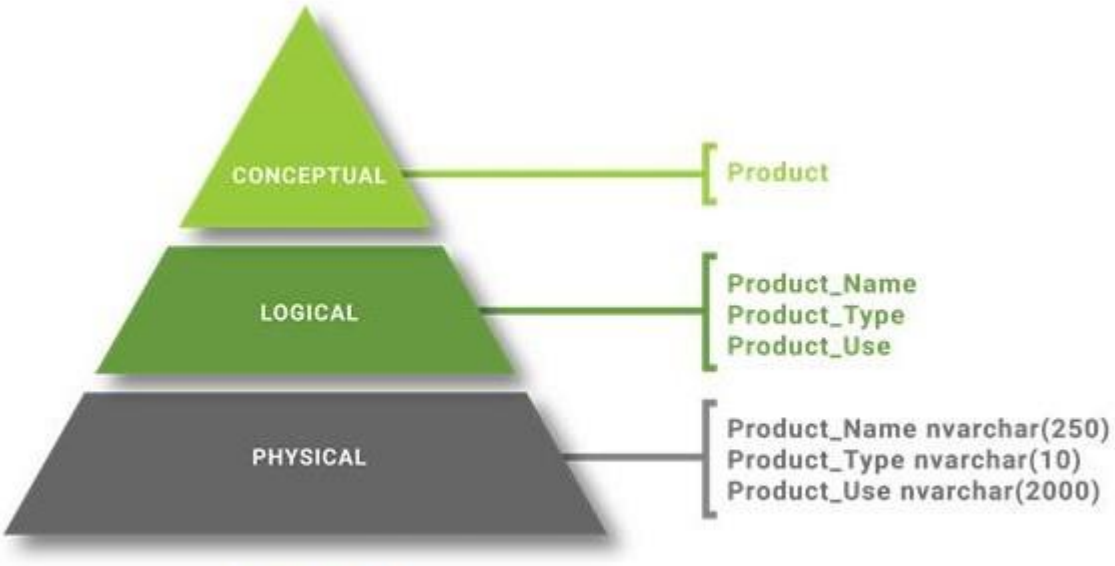
β_0 and β_1 are parameters, x and y are input and output variables.

A Model is an informative representation of an object, person, or system. Data model is the representation to store one's data, which is the realm of database managers.

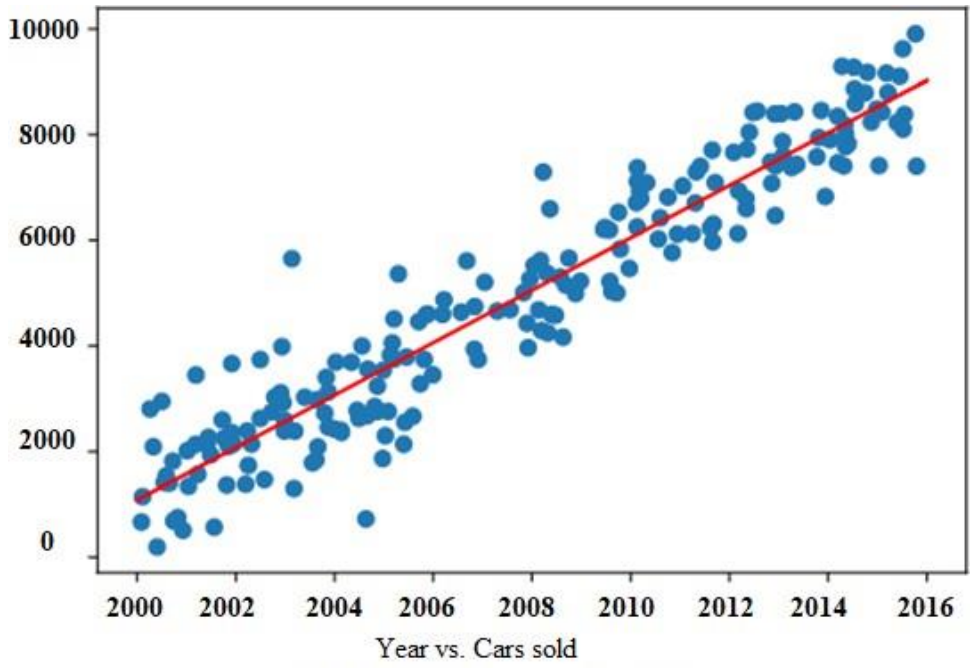
- Conceptual model [SQL Queries]
- Logical model [Tables]
- Physical model [Store]

Statistical model is a mathematical representation of observed data.

- Mean
- Mode
- Median



Data Model



Statistical / Mathematical Model

A Model is our attempt to understand and represent the nature of reality through a particular lens, be it architectural, biological, or mathematical. A model is an artificial construction where all extraneous detail has been removed or abstracted. Statisticians and Data scientists capture the *uncertainty and randomness* of data-generating processes with *mathematical functions* that express the *shape and structure* of the data itself.

Probability Distributions

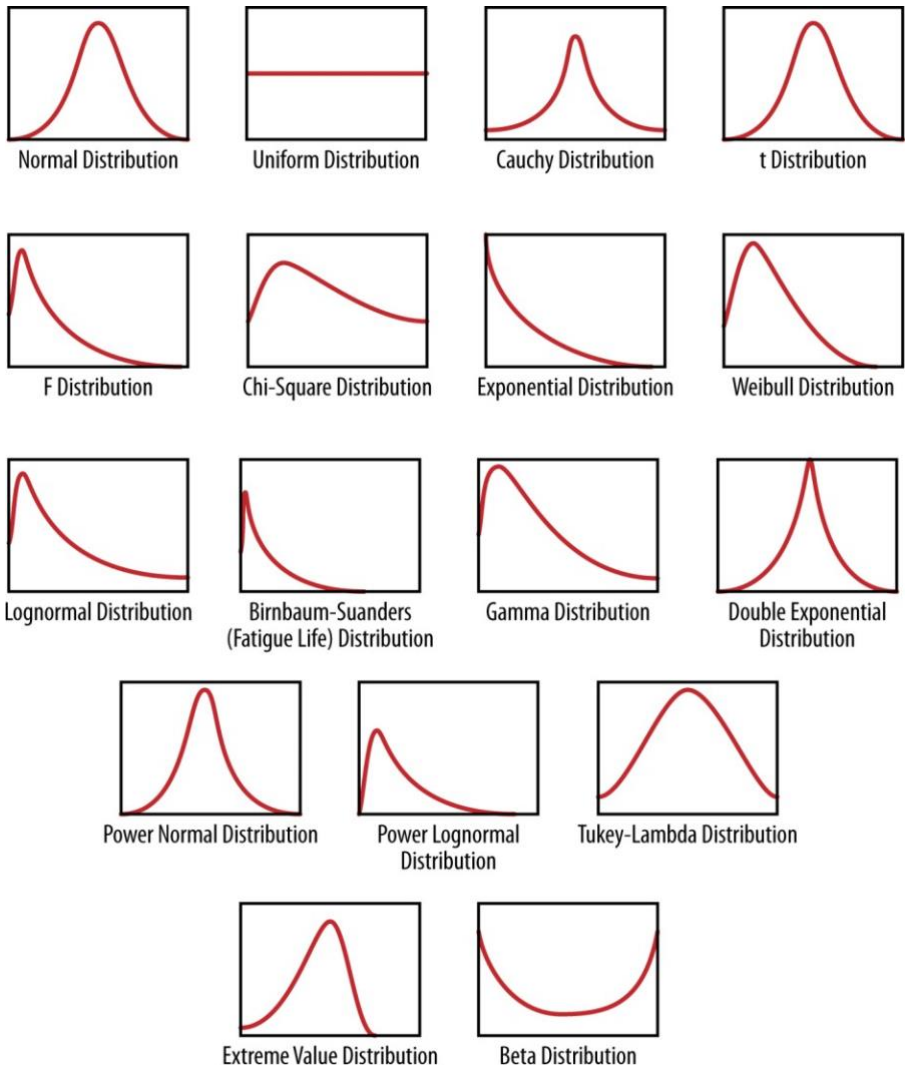
Probability distributions are the foundation of statistical models. Probability Distribution is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment. They are to be interpreted as assigning a probability to a subset of possible outcomes and have corresponding functions.

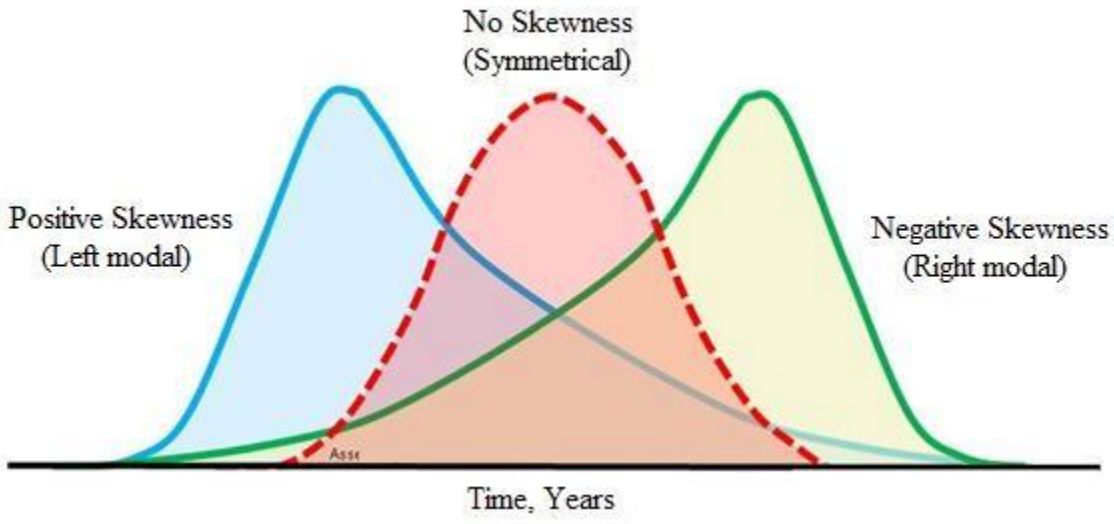
Example:

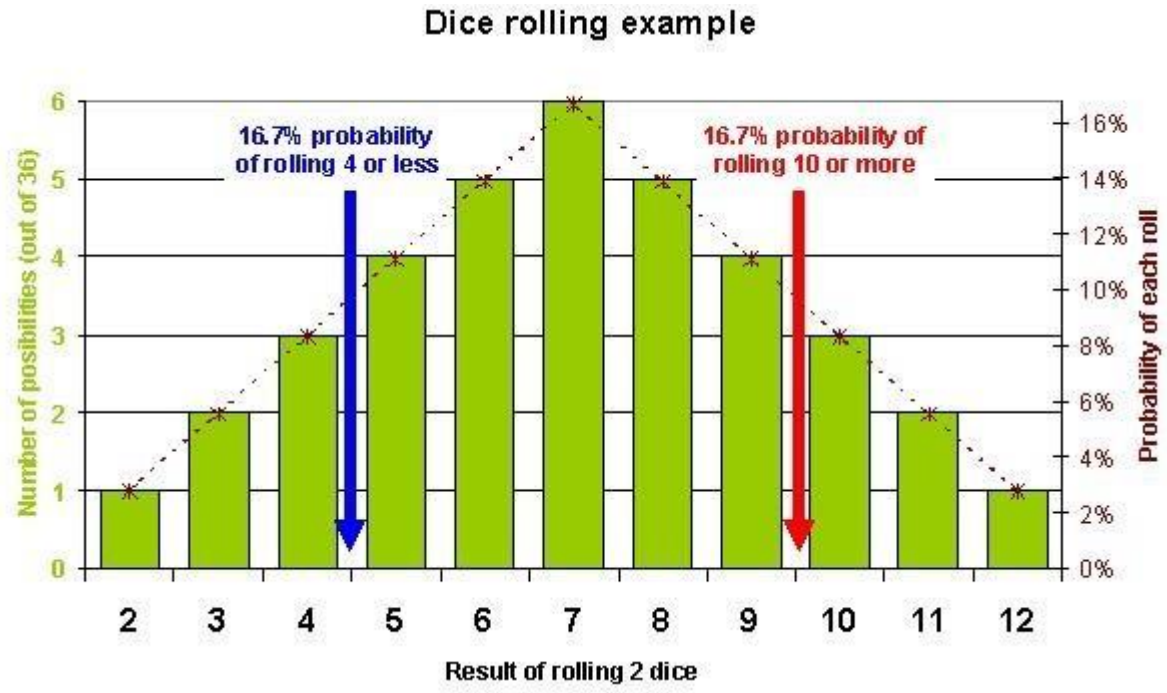
Normal distribution [Bell curve] is written as,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The parameter μ is the mean and median and controls where the distribution is centered (because this is a symmetric distribution), and the parameter σ controls how spread out the distribution is. This is the general functional form, but for specific real-world phenomenon, these parameters have actual numbers as values, which we can estimate from the data.







Fitting a Model

Fitting a model means that you estimate the parameters of the model using the observed data. You are using your data as evidence to help approximate the real-world mathematical process that generated the data. Fitting the model often involves optimization methods and algorithms, such as maximum likelihood estimation, to help get the parameters. In fact, when you estimate the parameters, they are estimators, meaning they themselves are functions of the data.

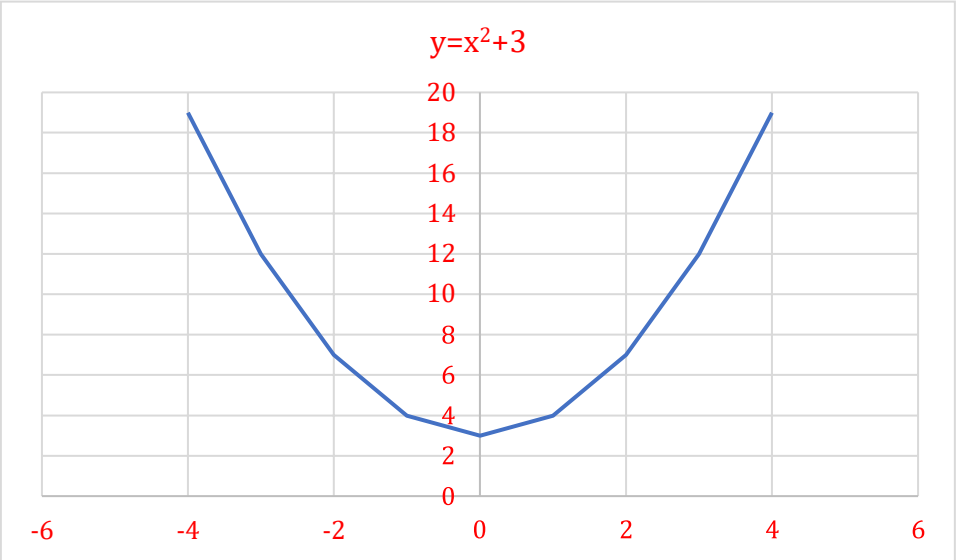
The equation or functional form expresses the relationship between your two variables, based on your assumption that the data followed a linear pattern. Fitting the model is when you start coding: your code will read in the data, and you'll specify the functional form that you

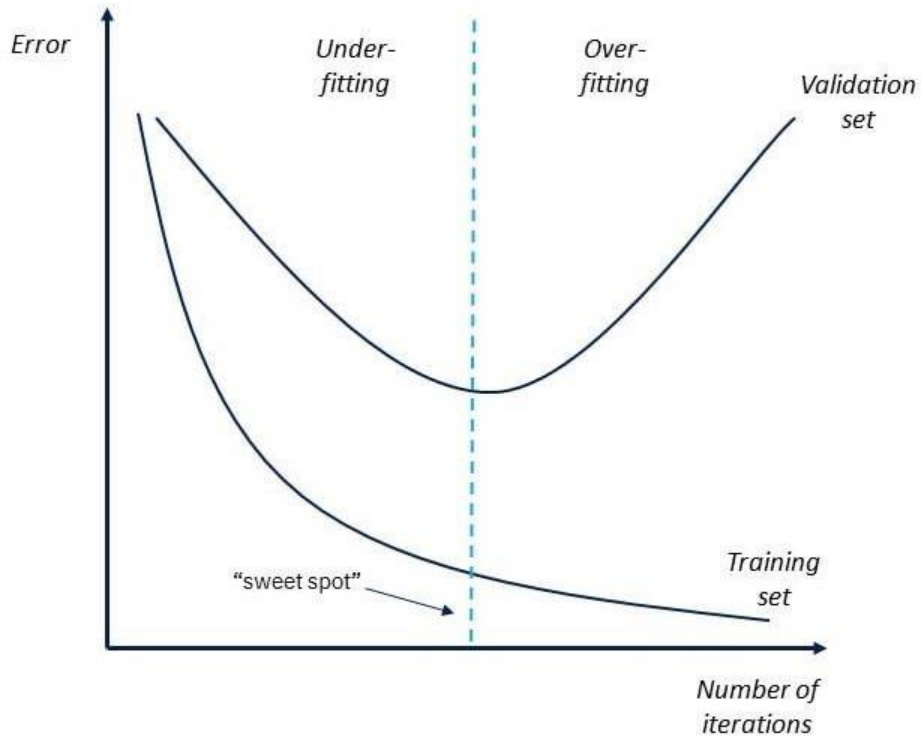
wrote down on the piece of paper. Then R or Python will use built-in optimization methods to give you the most likely values of the parameters given the data.

Example:

x = {-4,-3,-2,-1,0,1,2,3,4}

y = {19,12,7,4,3,4,7,12,19}





Sampling

Data Sampling is a statistical analysis technique used to select, manipulate, and analyze a representative subset of data points to identify patterns and trends in the larger data set being examined.

Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population and can be used for Predictive Analysis. The methodology used to sample from a larger population depends on the type of analysis being performed, but it may include simple random sampling or systematic sampling.

Sampling can be based on **probability**, an approach that uses random numbers that correspond to points in the data set to ensure that there is no correlation between points chosen for the sample.

Sampling can also be based on **nonprobability**, an approach in which a data sample is determined and extracted based on the judgment of the analyst. As inclusion is determined by the analyst, it can be more difficult to extrapolate whether the sample accurately represents the larger population than when probability sampling is used.

Probability Sampling types:

- Simple random sampling:
  Software is used to randomly select subjects from the whole population.
- Stratified sampling:
  Subsets of the data sets or population are created based on a common factor, and samples are randomly collected from each subgroup.
- Cluster sampling:
  The larger data set is divided into subsets (clusters) based on a defined factor, then a random sampling of clusters is analyzed.
- Multistage sampling:
  It involves dividing the larger population into a number of clusters. Second-stage clusters are then broken out based on a secondary factor, and those clusters are then sampled and analyzed. This staging could continue as multiple subsets are identified, clustered, and analyzed.
- Systematic sampling:
  A sample is created by setting an interval at which to extract data from the larger population. for example, selecting every 10th row in a spreadsheet of 200 items to create a sample size of 20 rows to analyze.

Non-Probability Sampling types:

- Convenience sampling:

  Data is collected from an easily accessible and available group.
- Consecutive sampling:

  Data is collected from every subject that meets the criteria until the predetermined sample size is met.
- Purposive or judgmental sampling:

  The researcher selects the data to sample based on predefined criteria.
- Quota sampling:

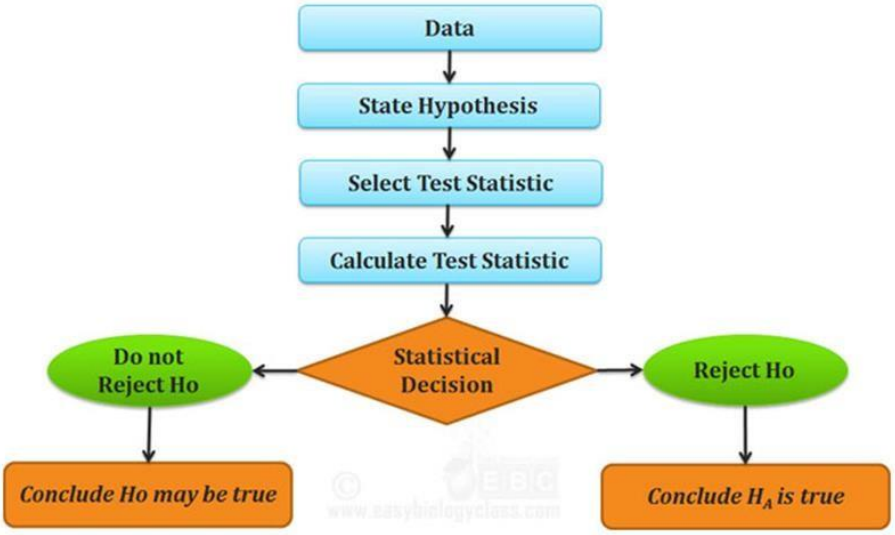
  The researcher ensures equal representation within the sample for all subgroups in the data set or population.

Sampling Tests:

Hypothesis is a proposition [true/false] made as a basis for reasoning, without any assumption of its truth. Hypothesis testing in statistics is a way to test the results of a survey or experiment to see if we have meaningful results. Statistical hypothesis is an assumption about a population parameter (mean, standard deviation, etc.…) and this assumption may or may not be true. A hypothesis test uses sample data to determine whether to reject the null hypothesis.

- The **Null hypothesis (H_0)** of a test always predicts no effect or no relationship between variables
- The **Alternative hypothesis (H_A)** is what you might believe to be true or hope to prove true. {opposite of Null hypothesis}

Steps in Hypothesis testing:



intro to R

R is an open-source programming language that is widely used as a statistical software and data analysis tool.

R generally comes with the Command-line interface.

R is available across widely used platforms like Windows, Linux, and macOS.

the R programming language is the latest cutting-edge(highly advanced/ innovative) tool.

R was initially written by **Ross Ihaka** and **Robert Gentleman** at the Department of Statistics of the University of Auckland in Auckland, New Zealand. R made its first appearance in 1993.

A large group of individuals has contributed to R by sending code and bug reports.

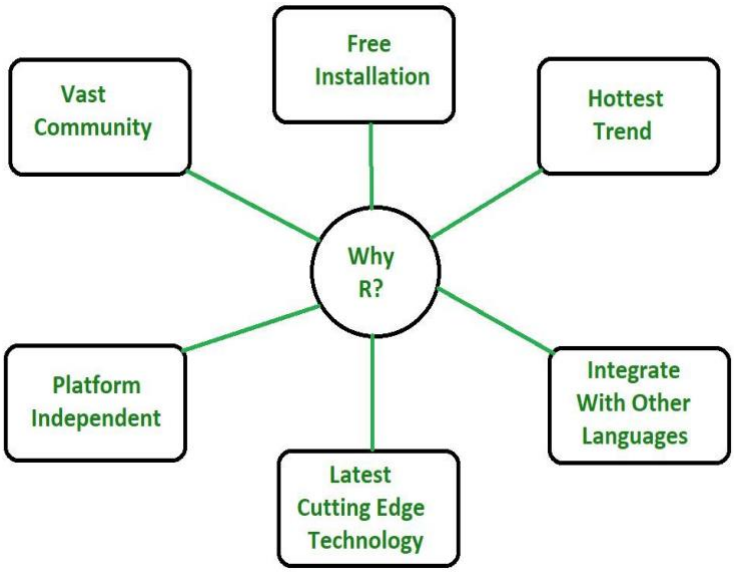
Since mid-1997 there has been a core group (the "R Core Team") who can modify the R source code archive.

R is freely available under the GNU General Public License

This programming language was named **R**, based on the first letter of first name of the two R authors (Robert Gentleman and Ross Ihaka)

Partly a play on the name of the Bell Labs Language **S**.

Why R



features of R

The following are the important features of R

R is a well-developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.

R has an effective data handling and storage facility,

R provides a suite of operators for calculations on arrays, lists, vectors and matrices.

R provides a large, coherent and integrated collection of tools for data analysis.

R provides graphical facilities for data analysis and display either directly at the computer or printing at the papers.

R Packages

One of the major features of R is it has a wide availability of libraries.

R has CRAN(Comprehensive R Archive Network), which is a repository holding more than 100000 packages.

Programming in R

Since R is much similar to other widely used languages syntactically, it is easier to code and learn in R.

Programs can be written in R in any of the widely used IDE like R Studio, Rattle, Tinn-R, etc.

After writing the program save the file with the extension .r

To run the program use the following command on the command line:

R file_name.r
Example: R

```
> # We can use the print() function
> print("Hello World!")
[1] "Hello World!"

> # Quotes can be suppressed in the output
> print("Hello World!", quote = FALSE)
[1] Hello World!

> # If there are more than 1 item, we can concatenate using paste()
> print(paste("How","are","you?"))
[1] "How are you?"
```

In this program, we have used the built-in function print() to print the string Hello World!

The quotes are printed by default. To avoid this we can pass the argument quote = FALSE

If there are more than one item, we can use the paste() or cat() function to concatenate the strings together.

R Program to Sample from a Population:

In this example, you will learn to take sample from a population using sample() function.

To understand this example, you should have the knowledge of following R programming topics:

R Variables and Constants
R Functions
We generally require to sample data from a large population.

R has a function called sample() to do the same. We need to provide the population and the size we wish to sample.

Additionally, we can specify if we want to do sampling with replacement. By default it is done without replacement.

Sample 2 items from x

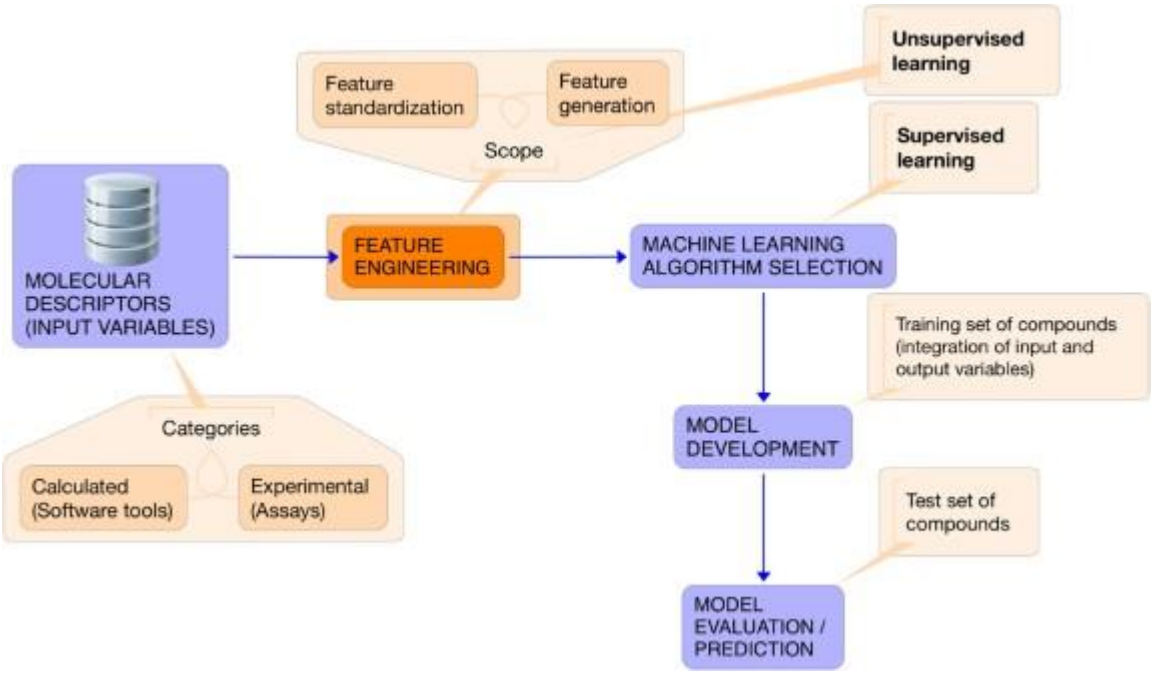
```
> x
[1]  1  3  5  7  9 11 13 15 17

> # sample 2 items from x
> sample(x, 2)
[1] 13  9
```

If we don't provide the size to sample, it defaults to the length of the population. This can be used to scramble x.

Feature generation:

Feature generation is the process of creating new features from one or multiple existing features, potentially for using in statistical analysis. This process adds new information to be accessible during the model construction and therefore hopefully result in more accurate model.

Feature Engineering, also known as feature creation or generation, is the process of constructing new features from existing data to train a machine learning model.



The most challenging part of the ML workflow is, arguably, knowing what data to feed into the models. The main question to be answered is “How can we find the most relevant data to train the models to solve the specific problem?”.

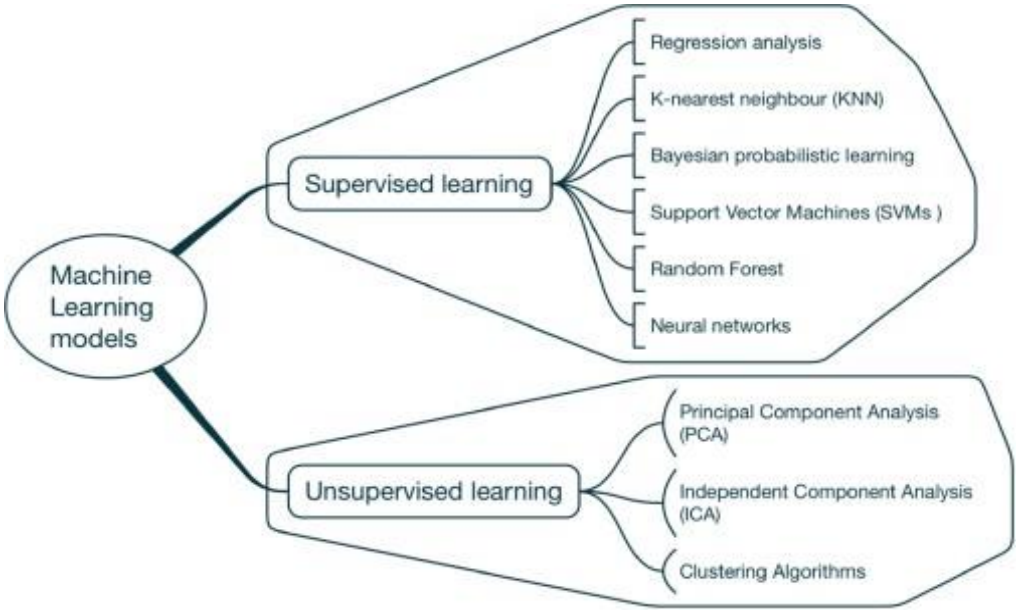
Three key considerations to identify the most relevant data are:

1. the right data attributes from which to infer realistic patterns and rules,
2. enough samples to train the model, and
3. engineering the data, the right way (e.g., aggregate, integrate, transform, etc., the data) before feeding into the models.

Feature engineering comprises feature generation and feature standardization. Feature generation entails looking at the attributes of the available data and, second, integrating and

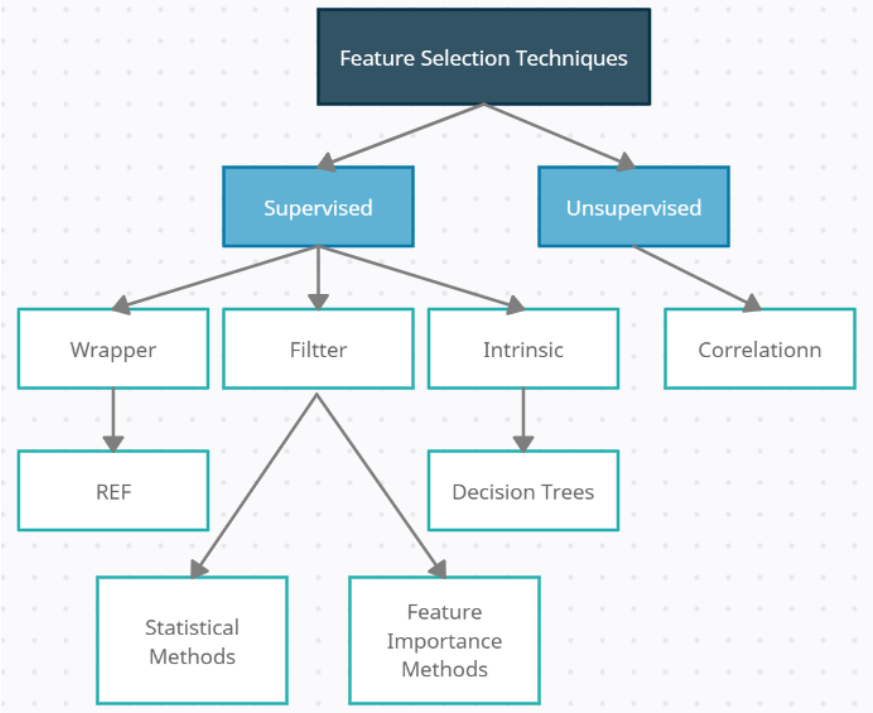
testing a new data source. Given the large number of possible molecular descriptors that represent potential input variables, it is a dataset without structure. Unsupervised learning is often used to develop a structure from a large dataset. Feature standardization refers to tools that make data “more appealing” to the model (e.g., imputation, normalization, and scaling).

A well-trained and validated ML model based on a large library of polyphenols could provide a predictive tool for evaluating the potential of newer polyphenols for effective antioxidant action in biological systems. ML models that are applied in quantitative structure–activity relationship (QSAR) studies and potentially useful for developing AI-based antioxidant assays are classified into (a) supervised learning, and (b) unsupervised learning.



Feature Selection algorithms

Feature selection is the process of selecting the subset of the relevant features and leaving out the irrelevant features present in a dataset to build a model of high accuracy. In other words, it is a way of selecting the optimal features from the input dataset.



Filters:

In this method, the dataset is filtered, and a subset that contains only the relevant features is taken.

Some common techniques of filters method are:

- Correlation
- Chi-Square Test
- ANOVA
- Information Gain, etc.

Wrappers:

The wrapper method has the same goal as the filter method, but it takes a machine learning model for its evaluation. In this method, some features are fed to the ML model, and evaluate the performance. The performance decides whether to add those features or remove to increase the accuracy of the model. This method is more accurate than the filtering method but complex to work.

Some common techniques of wrapper methods are:

- Forward Selection
- Backward Selection
- Bi-directional Elimination

Information gain:

Information gain is the reduction in entropy by transforming a dataset and is often used in training decision trees. Information gain is calculated by comparing the entropy of the dataset before and after a transformation.

Entropy is a measure of the uncertainty associated with a random variable. Entropy is the measurement of impurities or randomness in the data points.

Information gain is used for determining the best features/attributes that render maximum information about a class. It follows the concept of entropy while aiming at decreasing the level of entropy, beginning from the root node to the leaf nodes. Information gain computes the difference between entropy before and after split and specifies the impurity in class elements.

Information Gain = Entropy before splitting - Entropy after splitting

Given a probability distribution such that

$$P = (p_1, p_2, \dots, p_n)$$

Where, p_i is the probability of a data point in the subset of D_i of a dataset D,

$$Entropy(P) = -\sum_{i=1}^{n} p_i * log_2 (p_i)$$

Information gain is non-negative. Information Gain is symmetric such that switching of the split variable and target variable, the same amount of information gain is obtained. Information gain determines the reduction of the uncertainty after splitting the dataset on a particular feature such that **if the value of information gain increases, that feature is most useful for classification**. The feature having the highest value of information gain is accounted for as the best feature to be chosen for split.

Gini Index:

The gini index, or gini coefficient, or gini impurity computes the degree of probability of a specific variable that is wrongly being classified when chosen randomly and a variation of gini coefficient. It works on categorical variables; provides outcomes either be "successful" or "failure" and hence conducts binary splitting only.

The degree of gini index varies from 0 to 1, Where 0 depicts that all the elements be allied to a certain class, or only one class exists there. The gini index of value as 1 signifies that all the elements are randomly distributed across various classes, and A value of 0.5 denotes the elements are uniformly distributed into some classes.

Gini index is an impurity measure for decision tree learning and given as,

$$Gini(P) = 1 - \sum_{i=1}^{n} p_i^2$$

An attribute or feature with **least gini index is preferred as root node** while creating a decision tree.

Gini index favours larger partitions (distributions) and is very easy to implement whereas **information gain** supports smaller partitions (distributions) with various distinct values, i.e., there is a need to perform an experiment with data and splitting criterion.

While working on categorical data variables, gini index gives results either in "success" or "failure" and performs binary splitting only, in contrast to this, information gain measures the entropy differences before and after splitting and depicts the impurity in class variables.

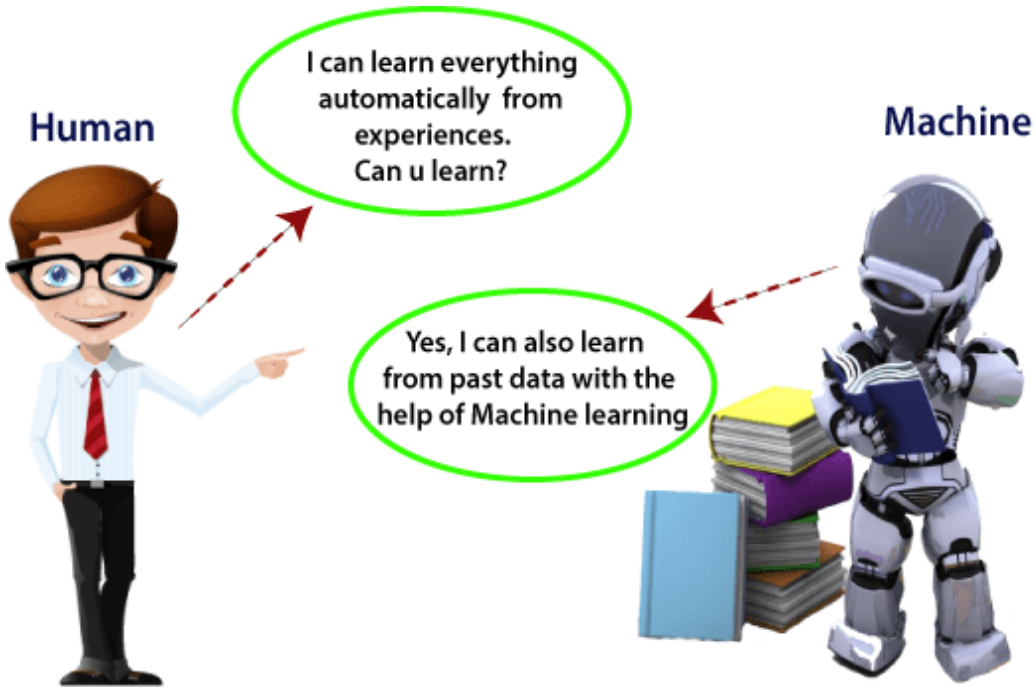
Machine Learning

Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for **building mathematical models and making predictions using historical data or information**. Currently, it is being used for various tasks such as image recognition, speech recognition, email filtering, Facebook auto-tagging, recommender system, and many more.

Machine Learning is said as a subset of artificial intelligence that is mainly concerned with the development of algorithms which allow a computer to learn from the data and past experiences on their own.

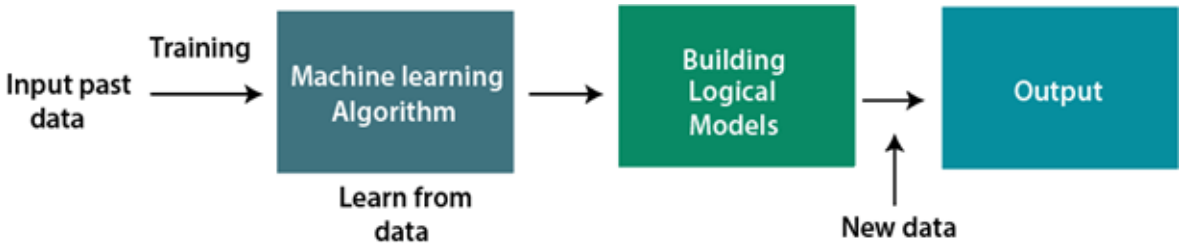
The term machine learning was first introduced by Arthur Samuel in 1959. We can define it in a summarized way as: "*Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed*."

With the help of sample historical data, which is known as **training data**, machine learning algorithms build a **mathematical model** that helps in making predictions or decisions without being explicitly programmed. Machine learning brings computer science and statistics together for creating predictive models. Machine learning constructs or uses the algorithms that learn from historical data. The more we will provide the information, the higher will be the performance. A machine can learn if it can improve its performance by gaining more data.



How does Machine Learning work

A Machine Learning system **learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it**. The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.



Features of Machine Learning include:

- Machine learning uses data to detect various patterns in each dataset.
- It can learn from past data and improve automatically.
- It is a data-driven technology.
- Machine learning is much like data mining as it also deals with the huge amount of the data.

The reason behind the need for machine learning is that it can do tasks that are too complex for a person to implement directly. As a human, we have some limitations as we cannot access the huge amount of data manually, so for this, we need some computer systems and here comes the machine learning to make things easy for us.

We can train machine learning algorithms by providing them the huge amount of data and let them explore the data, construct the models, and predict the required output automatically. The performance of the machine learning algorithm depends on the amount of data, and it can be determined by the cost function. With the help of machine learning, we can save both time and money.

Currently, machine learning is used in self-driving cars, cyber fraud detection, face recognition, and friend suggestion by Facebook, etc. Various top companies such as Netflix and Amazon have built machine learning models that are using a vast amount of data to analyze the user interest and recommend product accordingly.

The importance of Machine Learning is due to Rapid increment in the production of data, solving complex problems which are difficult for a human, Decision making in various

sector including finance, and Finding hidden patterns and extracting useful information from data.

Classification of Machine Learning

a. Supervised learning
b. Unsupervised learning
c. Reinforcement learning

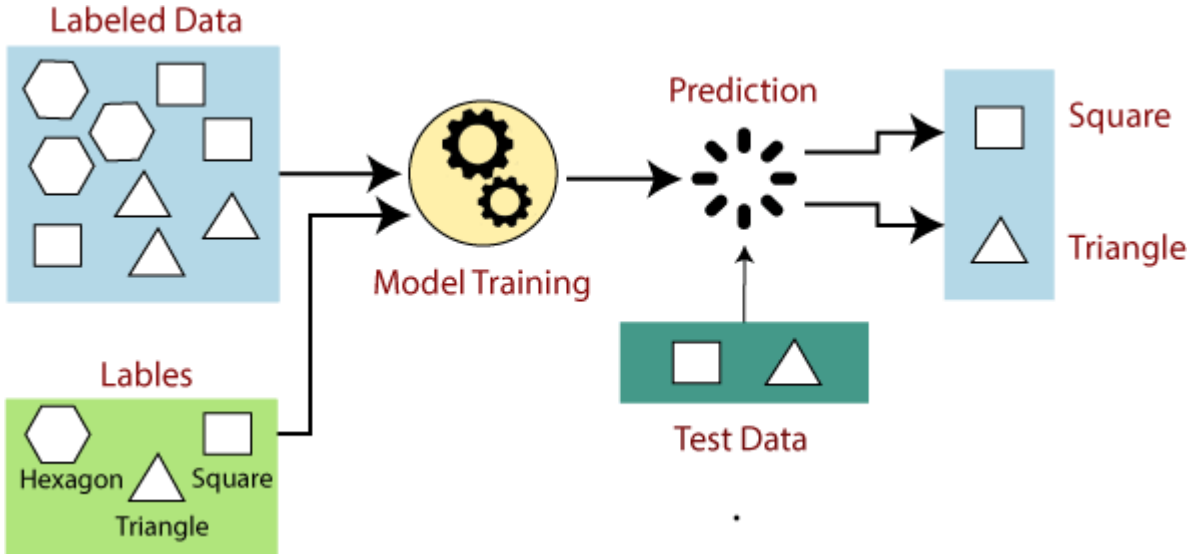
a. Supervised learning

Supervised learning is a type of machine learning method in which we provide **sample labeled data to the machine learning system to train it, and on that basis, it predicts the output.**

The system creates a model using labeled data to understand the datasets and learn about each data, once the training and processing are done then we test the model by providinga sample data to check whether it is predicting the exact output or not.

The goal of supervised learning is to map input data with the output data. Thesupervised learning is based on supervision, and it is the same as when a student learns thingsin the supervision of the teacher. The example of supervised learning is spam filtering.

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples.



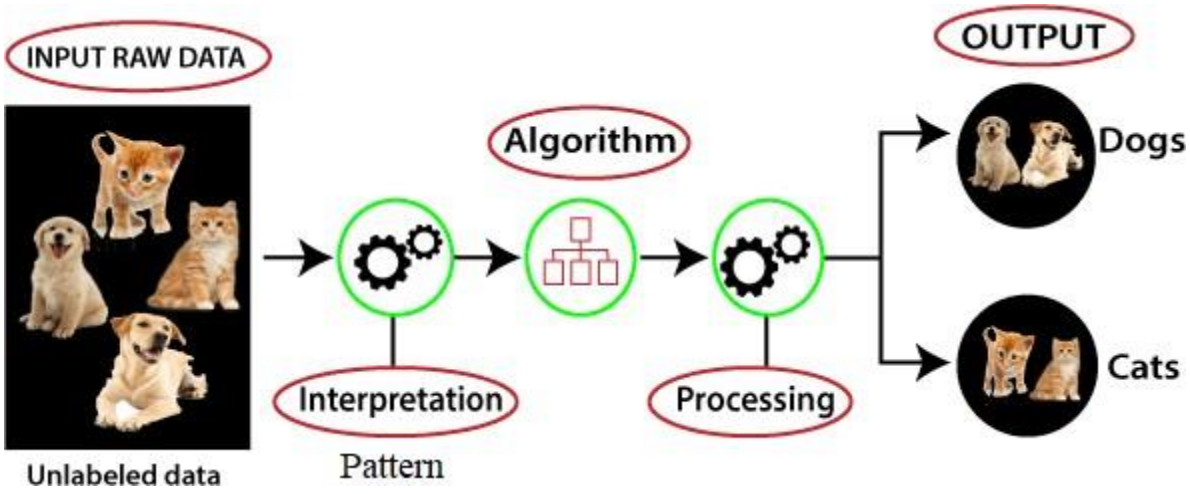
Supervised learning can be grouped further in two categories of algorithms:

- Classification
- Regression

b. Unsupervised Learning

Unsupervised learning is a learning method in which a machine learns without any supervision. The training is provided to the machine with the set of data that has not been labeled, classified, or categorized, and the algorithm needs to act on that data without any supervision. **The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns.**

In unsupervised learning, we don't have a predetermined result. The machine tries to find useful insights from the huge amount of data. Unsupervised learning refers to the use of artificial intelligence algorithms to identify patterns in data sets containing data points that are neither classified nor labeled. In other words, **unsupervised learning allows the system to identify patterns within data sets on its own.**



Unsupervised learning can be grouped further in two categories of algorithms:

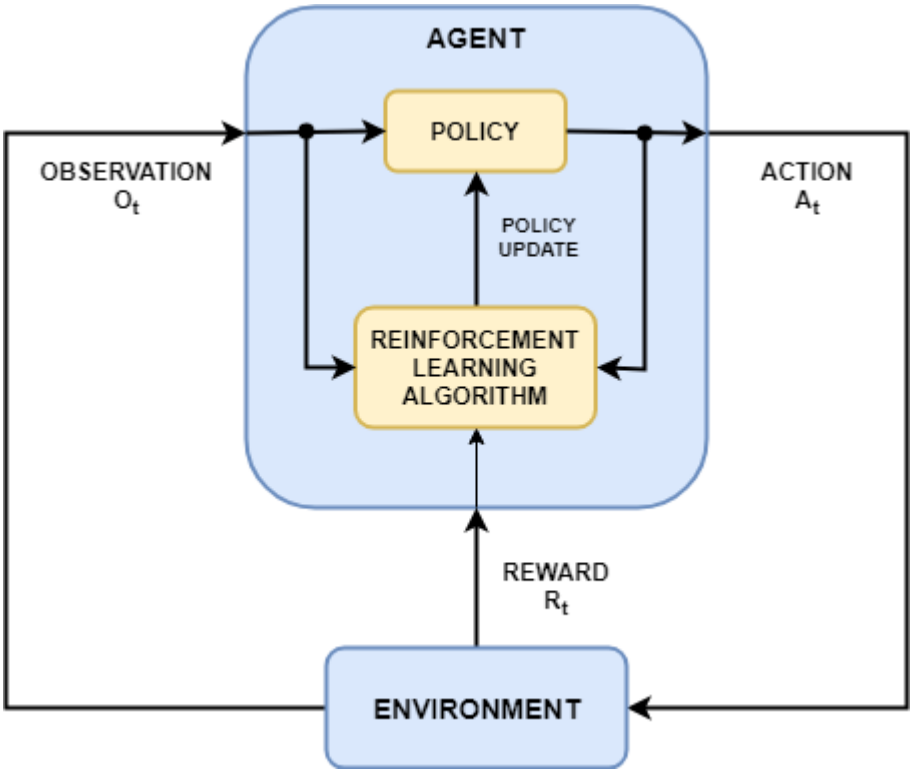
- Clustering
- Association

c. Reinforcement Learning

Reinforcement learning is a **feedback-based learning method, in which a learning agent gets a reward for each right action and gets a penalty for each wrong action**. The agent learns automatically with these feedbacks and improves its performance. Inreinforcement learning, the agent interacts with the environment and explores it. The goal of an agent is to get the most reward points, and hence, it improves its performance.

The robotic dog, which automatically learns the movement of his arms, is an example of Reinforcement learning.

Reinforcement learning is a machine learning training method based on rewarding desired behaviours and/or punishing undesired ones. In general, a reinforcement learning agent can perceive and interpret its environment, take actions, and learn through trial and error.

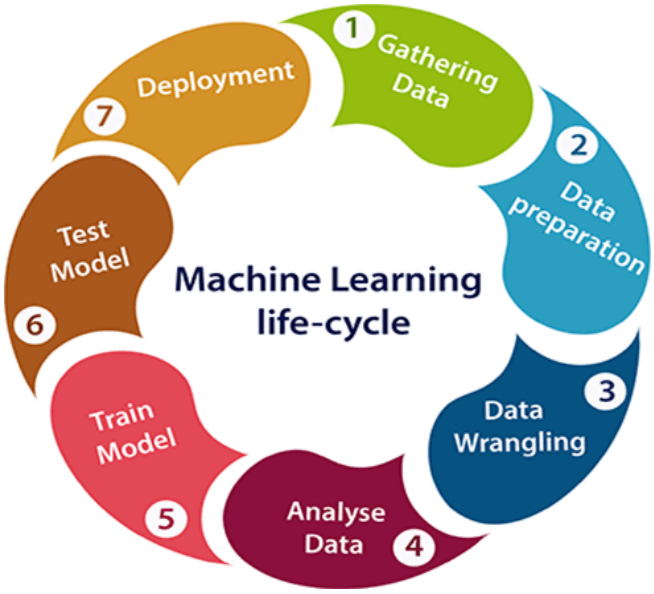


Machine learning Life cycle

Machine learning has given the computer systems the abilities to automatically learn without being explicitly programmed. But how does a machine learning system work. Machine learning life cycle is a cyclic process to build an efficient machine learning project. The main purpose of the life cycle is to find a solution to the problem or project.

Machine learning life cycle involves seven major steps, which are given below:

a. Gathering Data
b. Data preparation
c. Data Wrangling
d. Analyse Data
e. Train the model
f. Test the model
g. Deployment



a. Gathering Data:

Data Gathering is the first step of the machine learning life cycle in the form of data set. The goal of this step is to identify and obtain all data-related problems. In this step, we need to identify the different data sources, as data can be collected from various sources such as files, database, internet, or mobile devices.

The quantity and quality of the collected data will determine the efficiency of the output. The more will be the data, the more accurate will be the prediction.

This step includes the below tasks:

- Identify various data sources
- Collect data
- Integrate the data obtained from different sources

b. Data preparation:

After collecting the data, we need to prepare it for further steps. Data preparation is a step where we put our data into a suitable place and prepare it to use in our machine learning training. In this step, first, we put all data together, and then randomize the ordering of data. This step can be further divided into two processes:

- **Data exploration:**
    - It is used to understand the nature of data that we must work with. We need to understand the characteristics, format, and quality of data.
    - A better understanding of data leads to an effective outcome. In this, we find Correlations, general trends, and outliers.
- **Data pre-processing:**
    - Now the next step is pre-processing of data for its analysis.

c. Data Wrangling:

Data wrangling is the process of cleaning and converting raw data into a useable format. It is the process of cleaning the data, selecting the variable to use, and transforming the data in a proper format to make it more suitable for analysis in the next step. Cleaning of data is required to address the quality issues.

In real-world applications, collected data may have various issues, including:

- Missing Values
- Duplicate data
- Invalid data
- Noise

So, we use various filtering techniques to clean the data. It is mandatory to detect and remove the above issues because it can negatively affect the quality of the outcome.

d. Data Analysis:

Now the cleaned and prepared data is passed on to the analysis step. This step involves:

- Selection of analytical techniques
- Building models
- Review the result

The aim of this step is to build a machine learning model to analyze the data using various analytical techniques and review the outcome. It starts with the determination of the type of the

problems, where we select the machine learning techniques such as Classification, Regression, Cluster analysis, Association, etc. then build the model using prepared data, and evaluate the model.

e. Train Model:

Here, we train our model to improve its performance for better outcome of the problem. We use datasets to train the model using various machine learning algorithms. Training a model is required so that it can understand the various patterns, rules, and features.

f. Test Model

In this step, we check for the accuracy of our model by providing a test dataset to it. Testing the model determines the percentage accuracy of the model as per the requirement of project or problem.

g. Deployment:

If the above-prepared model is producing an accurate result as per our requirement with acceptable speed, then we deploy the model in the real system. But before deploying the project, we will check whether it is improving its performance using available data or not. The deployment phase is like making the final report for a project.

----------end of II Unit----------

Short Questions

1. Define Populations and samples.
2. Explain about Statistical modeling.
3. Define probability distribution.
4. Discuss the process of fitting a model.
5. Define Feature Generation.
6. How Feature Selection is done..
7. Define Object Oriented concepts of R
8. What is a Model and why we require model?
9. Define concept of Filters and Wrappers.
10. Define Information Gain and its usage.
11. Explain about Gini Index.

Essay Questions

1. What is Model and Explain about Statistical modeling. And process of Fitting a model.

2. Explain about Feature Generation and Feature Selection with example.

3. Explain about Filters and Wrappers in Feature Selection Algorithms with example.
4. Explain the Feature Selection Algorithms
5. Discuss about information Gain, Gini Index, and their impact on data.
6. Explain in detail about probability distribution in Fitting a model.