BHARATHIDASAN UNIVERSITY
Tiruchirappalli- 620024
Tamil Nadu, India.

Programme: M.Sc. Statistics

Course Title: Introduction to Machine learning

Course Code: 20ST01VAC

Unit-I

Introduction to Machine Learning

Dr. T. Jai Sankar
Associate Professor and Head
Department of Statistics

Ms. E. Devi
Guest Faculty
Department of Statistics

UNIT-I

Introduction to Machine Learning

1.1 Introduction to Machine Learning:

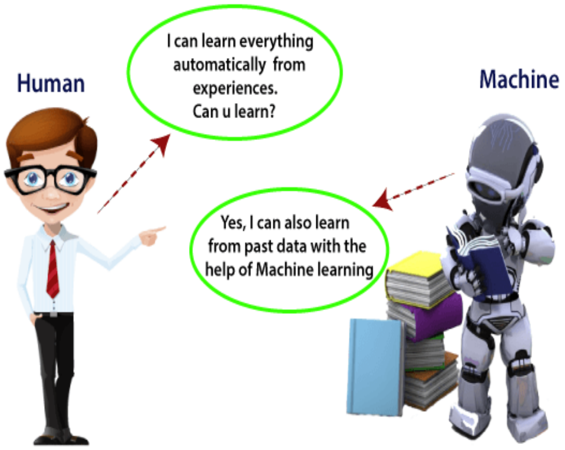
Machine Learning tutorial provides basic and advanced concepts of machine learning. Our machine learning tutorial is designed for students and working professionals.

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.

This machine learning tutorial gives you an introduction to machine learning along with the wide range of machine learning techniques such as **Supervised**, **Unsupervised**, and **Reinforcement** learning. You will learn about regression and classification models, clustering methods, hidden Markov models, and various sequential models.

Machine Learning

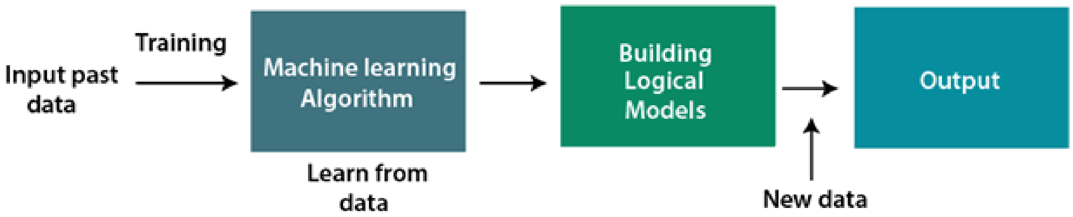
In the real world, we are surrounded by humans who can learn everything from their experiences with their learning capability, and we have computers or machines which work on our instructions. But can a machine also learn from experiences or past data like a human does? So here comes the role of Machine Learning.


Human
I can learn everything automatically from experiences. Can u learn?
Machine
Yes, I can also learn from past data with the help of Machine learning

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**.

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. Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the output. Machine learning has changed our way of thinking about the problem. The below block diagram explains the working of Machine Learning algorithm:



Features of Machine Learning

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

Need for Machine Learning

The need for machine learning is increasing day by day. The reason behind the need for machine learning is that it is capable of doing 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.

The importance of machine learning can be easily understood by its uses cases, 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 build machine learning models that are using a vast amount of data to analyze the user interest and recommend product accordingly.

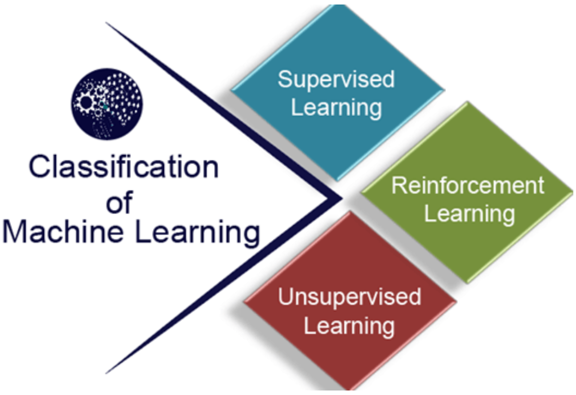
Following are some key points which show the importance of Machine Learning:

- Rapid increment in the production of data
- Solving complex problems, which are difficult for a human
- Decision making in various sector including finance
- Finding hidden patterns and extracting useful information from data.

Classification of Machine Learning

At a broad level, machine learning can be classified into three types:

1. **Supervised learning**
2. **Unsupervised learning**
3. **Reinforcement learning**



1) Supervised Learning

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

The system creates a model using labelled data to understand the datasets and learn about each data, once the training and processing are done then we test the model by providing a 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. The supervised learning is based on supervision, and it is the same as when a student learns things in the supervision of the teacher. The example of supervised learning is **spam filtering**.

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

- **Classification**
- **Regression**

2) 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. It can be further classifieds into two categories of algorithms:

- **Clustering**
- **Association**

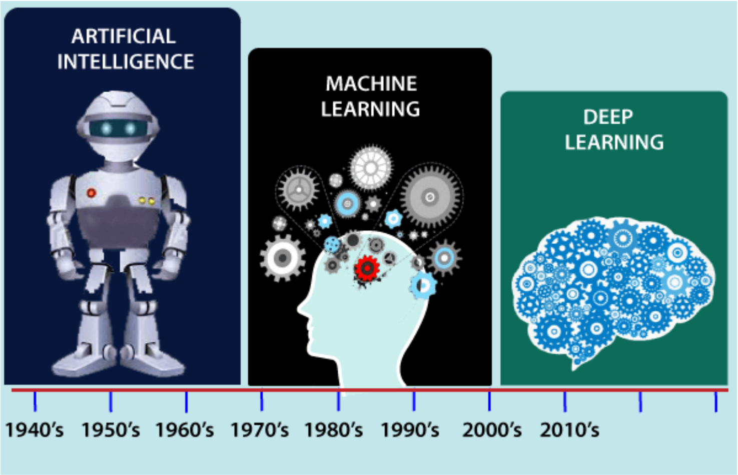
3) Reinforcement Learning

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. In reinforcement 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.

History of Machine Learning

Before some years (about 40-50 years), machine learning was science fiction, but today it is the part of our daily life. Machine learning is making our day to day life easy from **self-driving cars** to **Amazon virtual assistant "Alexa"**. However, the idea behind machine learning is so old and has a long history. Below some milestones are given which have occurred in the history of machine learning:



Machine Learning at 21st century

- **2006:** In the year 2006, computer scientist Geoffrey Hinton has given a new name to neural net research as "**deep learning**," and nowadays, it has become one of the most trending technologies.
- **2012:** In 2012, Google created a deep neural network which learned to recognize the image of humans and cats in YouTube videos.
- **2014:** In 2014, the Chabot "**Eugen Goostman**" cleared the Turing Test. It was the first Chabot who convinced the 33% of human judges that it was not a machine.
- **2014: Deep Face** was a deep neural network created by Face book, and they claimed that it could recognize a person with the same precision as a human can do.
- **2016: Alpha Go** beat the world's number second player **Lee sedol** at **Go game**. In 2017 it beat the number one player of this game **KeJie**.
- **2017:** In 2017, the Alphabet's Jigsaw team built an intelligent system that was able to learn the **online trolling**. It used to read millions of comments of different websites to learn to stop online trolling

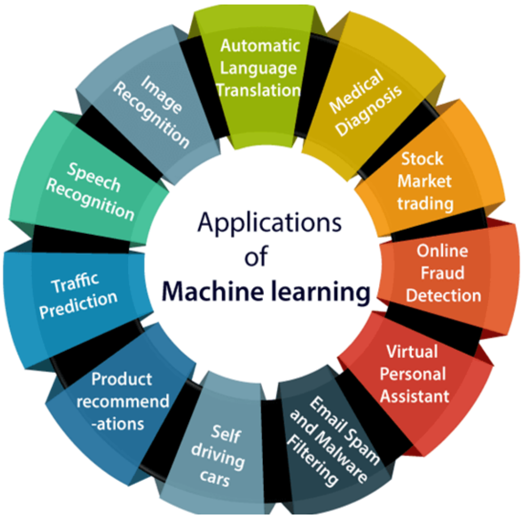
Machine Learning at present

Now machine learning has got a great advancement in its research, and it is present everywhere around us, such as **self-driving cars**, **Amazon Alexa**, **Catboats**, **recommender system**, and many more. It includes **Supervised**, **unsupervised**, and **reinforcement learning with clustering**, **classification**, **decision tree**, **SVM algorithms**, etc.

Modern machine learning models can be used for making various predictions, including **weather prediction**, **disease prediction**, **stock market analysis**, etc.

Applications of Machine learning

Machine learning is a buzzword for today's technology, and it is growing very rapidly day by day. We are using machine learning in our daily life even without knowing it such as Google Maps, Google assistant, Alexa, etc. Below are some most trending real-world applications of Machine Learning:


Applications of Machine learning
Automatic Language Translation
Medical Diagnosis
Stock Market trading
Online Fraud Detection
Virtual Personal Assistant
Email Spam and Malware Filtering
Self driving cars
Product recommend -ations
Traffic Prediction
Speech Recognition
Image Recognition

1. Image Recognition:

Image recognition is one of the most common applications of machine learning. It is used to identify objects, persons, places, digital images, etc. The popular use case of image recognition and face detection is, **Automatic friend tagging suggestion.**

Facebook provides us a feature of auto friend tagging suggestion. Whenever we upload a photo with our Facebook friends, then we automatically get a tagging suggestion with name, and the technology behind this is machine learning's **face detection** and **recognition algorithm**.

2. Speech Recognition

While using Google, we get an option of "**Search by voice**," it comes under speech recognition, and it's a popular application of machine learning.

Speech recognition is a process of converting voice instructions into text, and it is also known as "**Speech to text**", or "**Computer speech recognition**." At present, machine learning algorithms are widely used by various applications of speech recognition. **Google assistant**, **Siri**, **Cortana**, and **Alexa** are using speech recognition technology to follow the voice instructions.

3. Traffic prediction:

If we want to visit a new place, we take help of Google Maps, which shows us the correct path with the shortest route and predicts the traffic conditions.

It predicts the traffic conditions such as whether traffic is cleared, slow-moving, or heavily congested with the help of two ways:

- **Real Time location** of the vehicle form Google Map app and sensors
- **Average time has taken** on past days at the same time.

Everyone who is using Google Map is helping this app to make it better. It takes information from the user and sends back to its database to improve the performance.

4. Product recommendations:

Machine learning is widely used by various e-commerce and entertainment companies such as **Amazon**, **Netflix**, etc., for product recommendation to the user. Whenever we search for some product on Amazon, then we started getting an advertisement for the same product while internet surfing on the same browser and this is because of machine learning.

Google understands the user interest using various machine learning algorithms and suggests the product as per customer interest.

As similar, when we use Netflix, we find some recommendations for entertainment series, movies, etc., and this is also done with the help of machine learning.

5. Self-driving cars:

One of the most exciting applications of machine learning is self-driving cars. Machine learning plays a significant role in self-driving cars. Tesla, the most popular car manufacturing company is working on self-driving car. It is using unsupervised learning method to train the car models to detect people and objects while driving.

6. Email Spam and Malware Filtering:

Whenever we receive a new email, it is filtered automatically as important, normal, and spam. We always receive an important mail in our inbox with the important symbol and spam emails in our spam box, and the technology behind this is Machine learning. Below are some spam filters used by Gmail:

- Content Filter
- Header filter
- General blacklists filter
- Rules-based filters
- Permission filters

Some machine learning algorithms such as **Multi-Layer Preceptor**, **Decision tree**, and **Naïve Bayes classifier** are used for email spam filtering and malware detection.

7. Virtual Personal Assistant:

We have various virtual personal assistants such as **Google assistant**, **Alexa**, **Cortana**, **Siri**. As the name suggests, they help us in finding the information using our voice instruction. These assistants can help us in various ways just by our voice instructions such as Play music, call someone, Open an email, Scheduling an appointment, etc.

8. Online Fraud Detection:

Machine learning is making our online transaction safe and secure by detecting fraud transaction. Whenever we perform some online transaction, there may be various ways that a fraudulent transaction can take place such as **fake accounts**, **fake ids**, and **steal money** in the middle of a transaction. So to detect this, **Feed Forward Neural network** helps us by checking whether it is a genuine transaction or a fraud transaction.

9. Stock Market trading:

Machine learning is widely used in stock market trading. In the stock market, there is always a risk of up and downs in shares, so for this machine learning's **long short term memory neural network** is used for the prediction of stock market trends.

10. Medical Diagnosis:

In medical science, machine learning is used for diseases diagnoses. With this, medical technology is growing very fast and able to build 3D models that can predict the exact position of lesions in the brain.

It helps in finding brain tumors and other brain-related diseases easily.

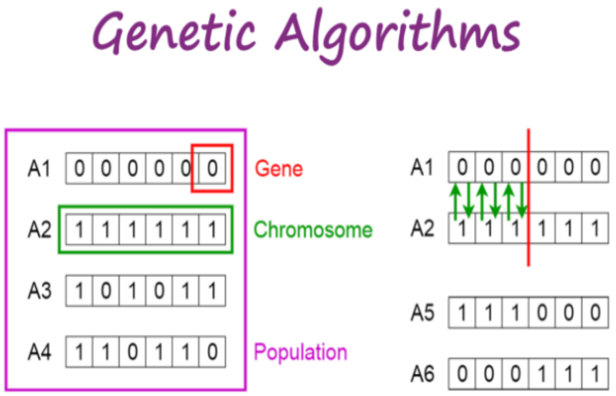
11. Automatic Language Translation:

Nowadays, if we visit a new place and we are not aware of the language then it is not a problem at all, as for this also machine learning helps us by converting the text into our known languages. Google's GNMT (Google Neural Machine Translation) provide this feature, which is a Neural Machine Learning that translates the text into our familiar language, and it called as automatic translation.

The technology behind the automatic translation is a sequence to sequence learning algorithm, which is used with image recognition and translates the text from one language to another language.

1.2 Introduction to Genetic Algorithms — Including Example Code

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.

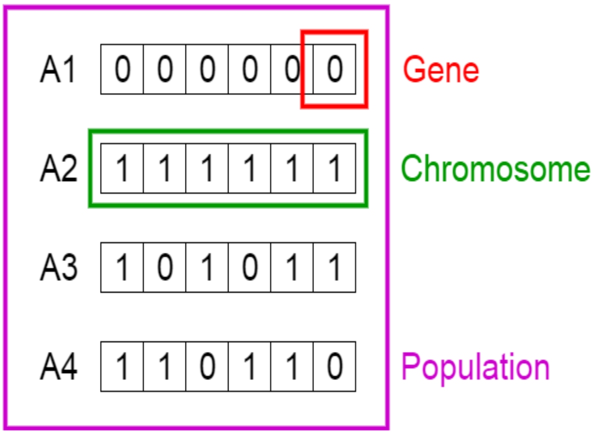
1. Initial population
2. Fitness function
3. Selection
4. Crossover
5. Mutation

1. 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.



2. 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.

3. Selection

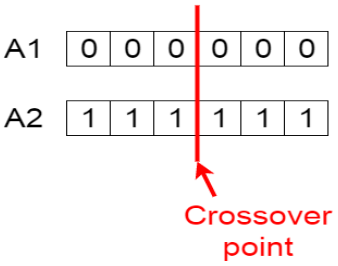
The idea of **selection** phase is to select the fittest individuals and let them pass their genes to the next generation.

Two pairs of individuals (**parents**) are selected based on their fitness scores. Individuals with high fitness have more chance to be selected for reproduction.

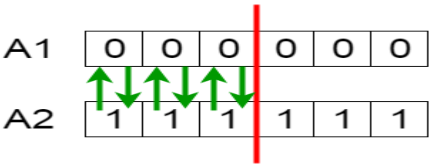
4. Crossover

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.

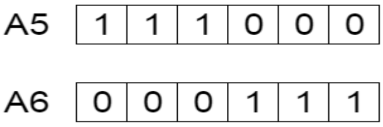
For example, consider the crossover point to be 3 as shown below.



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.



New offspring

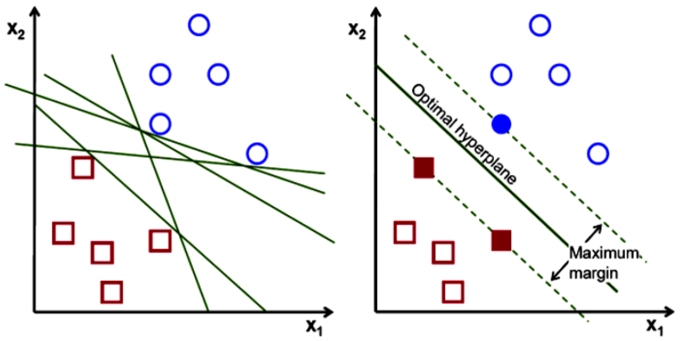
Termination

The algorithm terminates if the population has converged (does not produce offspring which are significantly different from the previous generation). Then it is said that the genetic algorithm has provided a set of solutions to our problem.

1.3 Support Vector Machine — Introduction to Machine Learning Algorithms:

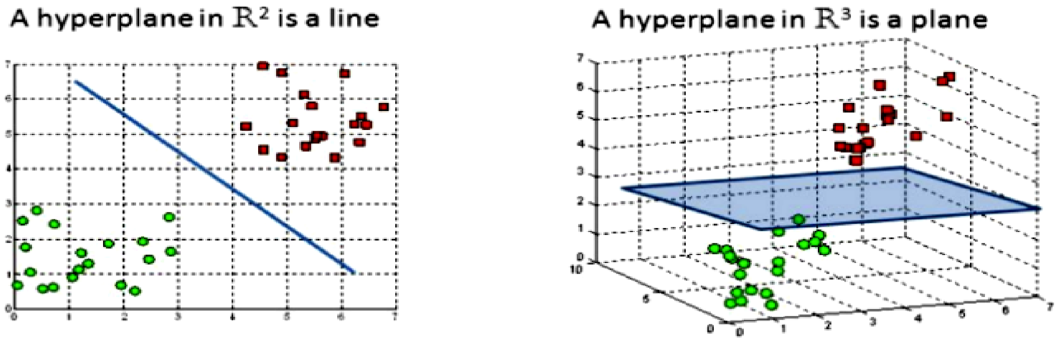
Support Vector Machine

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.



Possible hyperplanes

Hyperplanes and Support Vectors:


A hyperplane in $\mathbb{R}^2$ is a line

A hyperplane in $\mathbb{R}^3$ is a plane

Hyperplanes in 2D and 3D feature space

Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. If the number of input features is 2, then the hyperplane is just a line. If the number of input features is 3, then the hyper plane becomes a two-dimensional plane. It becomes difficult to imagine when the number of features exceeds 3.

Support vector machines:

Definition: Support vector machine or SVM is one of the most popular supervised learning algorithm, which is used for classification and regression problems.

It is used for classification problems in machine learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

Large Margin Intuition : In logistic regression, we take the output of the linear function and squash the value within the range of [0,1] using the sigmoid function. If the squashed value is greater than a threshold value(0.5) we assign it a label 1, else we assign it a label 0. In SVM, we take the output of the linear function and if that output is greater than 1, we identify it with one class and if the output is -1, we identify is with another class. Since the threshold values are changed to 1 and -1 in SVM, we obtain this reinforcement range of values ([-1,1]) which acts as margin.

1.4 Distance Measures for Machine Learning

- Distance measures play an important role in machine learning.
- They provide the foundation for many popular and effective machine learning algorithms like k-nearest neighbour for supervised learning and k-means clustering for unsupervised learning.
- Different distance measures must be chosen and used depending on the types of the data. As such, it is important to know how to implement and calculate a range of different popular distance measures and the intuitions for the resulting scores.
- In this tutorial, you will discover distance measures in machine learning.
- After completing this tutorial, you will know:
- The role and importance of distance measures in machine learning algorithms.
- How to implement and calculate Hamming, Euclidean, and Manhattan distance measures.
- How to implement and calculate the Minkowski distance that generalizes the Euclidean and Manhattan distance measures.

Importance of Distance Metrics in Machine Learning Modelling

A number of Machine Learning Algorithms - Supervised or Unsupervised, use Distance Metrics to know the input data pattern in order to make any Data Based decision. A good distance metric helps in improving the performance of Classification, Clustering and Information Retrieval process significantly. In this article, we will discuss about different Distance Metrics and how do they help in Machine Learning Modelling.

What is similarity?

Similarity is essentially a vast umbrella term that encompasses a wide range of scores and measures for assessing the differences among several kinds of data. Honestly, similarity refers to far more than you could really cover in one place.

Similarity and Dissimilarity

Distance or similarity measures are essential in solving many pattern recognition problems such as classification and clustering. Various distance/similarity measures are available in the literature to compare two data distributions. As the names suggest, a similarity measures how close two distributions are. For multivariate data complex summary methods are developed to answer this question. Similarity Measure Numerical measure of how alike two data objects often fall between 0 (no similarity) and 1 (complete similarity). Dissimilarity Measure Numerical measure of how different two data objects are range from 0 (objects are alike) to ∞ (objects are different)

Similarity/Dissimilarity for Simple Attributes

Here, p and q are the attribute values for two data objects.

Attribute Type	Similarity	Dissimilarity
Nominal	s={1 if p=q0 if p≠q	d={0 if p=q1 if p≠q
Ordinal	s=1−‖p−q‖n−1 (values mapped to integer 0 to n-1, where n is the number of values)	d=‖p−q‖n−1
Interval or Ratio	s=1−‖p−q‖,s=11+‖p−q‖	d=‖p−q‖

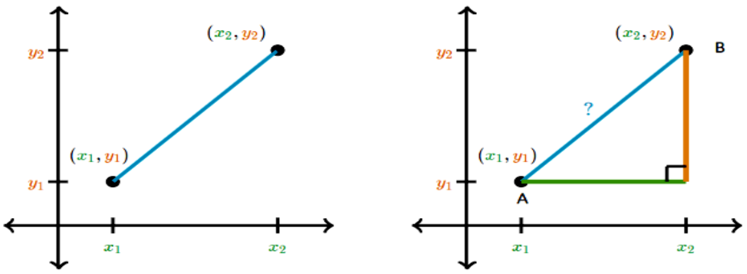
Distance, such as the Euclidean distance, is a dissimilarity measure and has some well-known properties: Common Properties of Dissimilarity Measures

1. d(p, q) ≥ 0 for all p and q, and d(p, q) = 0 if and only if p = q,
2. d(p, q) = d(q,p) for all p and q,
3. d(p, r) ≤ d(p, q) + d(q, r) for all p, q, and r, where d(p, q) is the distance (dissimilarity) between points (data objects), p and q.

A distance that satisfies these properties is called a **metric**. Following is a list of several common distance measures to compare multivariate data. We will assume that the attributes are all continuous.

Distance Function:

Do you remember studying Pythagorean theorem? If you do, then you might remember calculating distance between two data points using the theorem.



In order to calculate the distance between data points A and B Pythagorean theorem considers the length of x and y axis.

$$? = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

A distance function is nothing but a mathematical formula used by distance metrics. The distance function can differ across different distance metrics. Let’s talk about different distance metrics and understand their role in machine learning modelling.

Distance Metrics

There are a number of distance metrics, but to keep this article concise, we will only be discussing a few widely used distance metrics. We will first try to understand the mathematics behind these metrics and then we will identify the machine learning algorithms where we use these distance metrics. Below are the commonly used distance metrics -

Minkowski distance is the generalized distance metric. Here generalized means that we can manipulate the above formula to calculate the distance between two data points in different ways. As mentioned above, we can manipulate the value of p and calculate the distance in three different ways.

What are the types of Distance similarity?

p = 1, Manhattan Distance

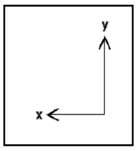
p = 2, Euclidean Distance

p = ∞, Chebychev Distance

We will discuss these distance metrics below in detail.

1. Manhattan Distance: We use Manhattan Distance if we need to calculate the distance between two data points in a grid like path. As mentioned above, we use **Minkowski distance** formula to find Manhattan distance by setting **p’s** value as **1**.

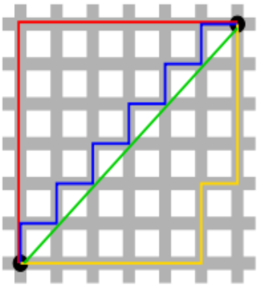
Let’s say, we want to calculate the distance, **d**, between two data points- **x** and **y**.



Manhattan

Distance **d** will be calculated using an **absolute sum of difference** between its Cartesian co-ordinates as below :

$$d= \sum_{i=1}^{n} |x_i - y_i|$$

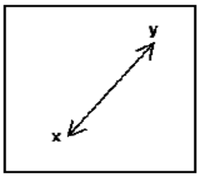


Manhattan distance is also known as Taxicab Geometry, City Block Distance etc.

2. Euclidean Distance: Euclidean distance is one of the most used distance metric. It is calculated using Minkowski Distance formula by setting **p's** value to **2**. This will update the distance **'d'** formula as below :

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Euclidean distance formula can be used to calculate the distance between two data points in a plane.



Euclidean

3. Cosine Distance:

Mostly Cosine distance metric is used to find similarities between different documents. In cosine metric we measure the degree of angle between two documents/vectors (the term frequencies in different documents collected as metrics). This particular metric is used when the magnitude between vectors does not matter but the orientation.

Cosine similarity formula can be derived from the equation of dot products :-

$$\vec{a} \cdot \vec{b} = \|\vec{a}\| \|\vec{b}\| \cos\theta$$

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \|\vec{b}\|}$$

4. Mahalanobis Distance:

Mahalanobis Distance is used for calculating the distance between two data points in a multivariate space. The benefit of using mahalanobis distance is, it takes covariance in account which helps in measuring the strength/similarity between two different data objects. The distance between an observation and the mean can be calculated as below -

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}.$$

5. Jaccard distance

The Jaccard distance measures the similarity of the two data set items as the intersection of those items divided by the union of the data items.

6. Minkowski distance

The Minkowski distance is the generalized form of the Euclidean and Manhattan Distance Measure. Minkowski distance is a metric in Normed vector space. What is Normed vector space? A Normed vector space is a vector space on which a norm is defined. Suppose X is a vector space then a norm on X is a real valued function ||x||which satisfies below conditions.

Zero Vector- Zero vector will have zero length.

Scalar Factor- The direction of vector doesn't change when you multiply it with a positive number though its length will be changed.

Triangle Inequality- If distance is a norm then the calculated distance between two points will always be a straight line.

The distance can be calculated using below formula -

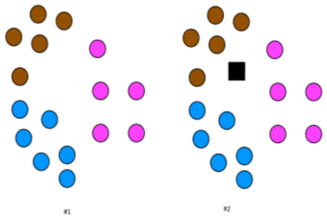
$$\left(\sum_{i=1}^{n}|x_i-y_i|^p\right)^{1/p}$$

1.5. Machine Learning Modelling and distance metrics

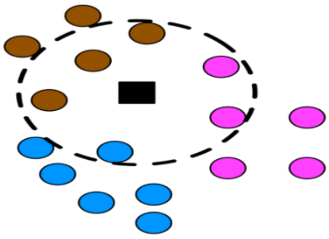
In this section, we will be working on some basic **classification and clustering** use cases. This will help us in understanding the usage of distance metrics in machine learning modelling. We will start with quick introduction of supervised and unsupervised algorithms and slowly will move on to the examples.

1. Classification

K-Nearest Neighbours (KNN) : KNN is a non-probabilistic supervised learning algorithm i.e. it doesn't produce the probability of membership of any data point rather KNN classifies the data on hard assignment, e.g the data point will either belong to 0 or 1. Now, you must be thinking how does KNN work if there is no probability equation involved. KNN uses distance metrics in order to find similarities or dissimilarities. Let's take iris dataset which has three classes and see how KNN will identify the classes for test data.

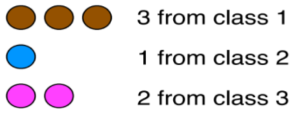


In the #2 image above the black square is a test data point. Now, we need to find which class this test data point belong to, with the help of KNN algorithm. We will now prepare the dataset to create machine learning model to predict the class for our test data. In KNN classification algorithm, we define the constant “K”. K is the number of nearest neighbours of a test data point. These K data points then will be used to decide the class for test data point.(Note this is in a training data set)



Are you wondering that how would we find the nearest neighbours. Well that’s where the distance metric comes into pictures. First, we calculate the distance between each train and test data point and then select the top nearest according to the value of k.

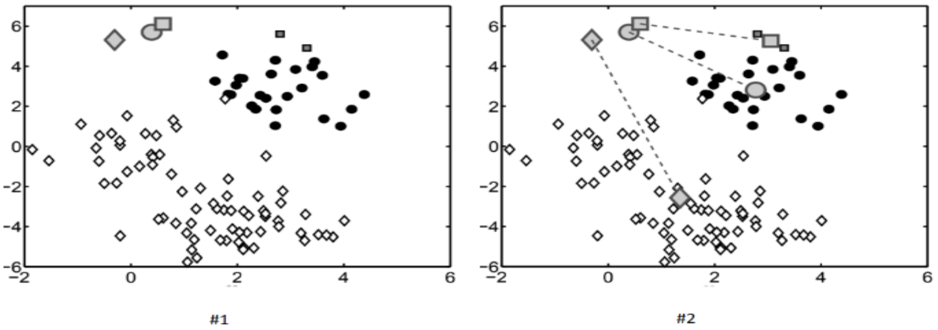
You can see in the above code we are using Murkowski distance metric with value of p as 2 i.e. KNN classifier is going to use Euclidean Distance Metric formula. As we move forward with machine learning modelling we can now train our model and start predicting the class for test data. Once the top nearest neighbours are selected, we check most voted class in neighbours -



2. Clustering

K-means : In classification algorithms, probabilistic or non-probabilistic we will be provided with labelled data so, it gets easier to predict the classes. Though in clustering algorithm we have no information on which data point belongs to which class. Distance metrics are important part of these kind of algorithm.

In K-means, we select number of centroids that define number of clusters. Each data point will then be assigned to its nearest centroid using distance metric (Euclidean). We will be using iris data to understand the underlying process of K-means.



In the above image #1 as you can see we randomly placed the centroids and in the image #2, using distance metric tried to find their closest cluster class.

3. Natural Language Processing

Information Retrieval : In information retrieval we work with unstructured data. The data can be an article, website, emails, text messages, a social media post etc. With the help of techniques used in NLP we can create vector data in a manner that can be used to retrieve information when queried. Once the unstructured data is transformed into vector form, we can use cosine similarity metric to filter out the irrelevant documents from the corpus.

Different Clustering Methods

Clustering Method	Description	Advantages	Disadvantages	Algorithms
Hierarchical Clustering	Based on top-to-bottom hierarchy of the data points to create clusters.	Easy to implement, the number of clusters need not be specified apriori, Dendrogram are easy to interpret.	Cluster assignment is strict and cannot be undone, high time complexity, cannot work for a larger dataset	DIANA, AGNES, halest etc.
Partitioning methods	Based on centroids and data points are assigned into a cluster based on its proximity to the cluster centred	Easy to implement, faster processing, can work on larger data, easy to interpret the outputs	We need to specify the number of cenrtroid sapriori, clusters that get created are of inconsistent sizes and densities, Effected by noise and outliers	k-means, k-medians, k-modes
Distribution-based Clustering	Based on the probability distribution of the data, clusters are derived from various metrics like mean, variance etc.	Number of clusters need not be specified apriority, works on real-time data, metrics are easy to understand and tune	Complex algorithm and slow, cannot be scaled to larger data	Gaussian Mixed Models, DBCLASD
Density-based Clustering (Model-based methods)	Based on density of the data points, also known as model based clustering	Can handle noise and outliers, need not specify number of clusters in the start, clusters that are created are highly homogenous, no restrictions on cluster shapes.	Complex algorithm and slow, cannot be scaled to larger data	DENCAST, DBSCAN
Fuzzy Clustering	Based on Partitioning Approach but data points can belong to more than one cluster	Can work on highly overlapped data, a higher rate of convergence	We need to specify the number of cancroids apriority, Effected by noise and outliers, Slow algorithm and cannot be scaled	Fuzzy C Means, Rough k means
Constraint Based (Supervised Clustering)	Clustering is directed and controlled by user constraints	Creates a perfect decision boundary, can automatically determine the outcome classes based on constraints, future data can be classified based on the training boundaries	Over fitting, high level of misclassification errors, cannot be trained on larger datasets	Decision Trees, Random Forest, Gradient Boosting

1.6. Deep learning:

Deep learning is a method in artificial intelligence (AI) that teaches computers to process data in a way that is inspired by the human brain. Deep learning models can recognize complex patterns in pictures, text, sounds, and other data to produce accurate insights and predictions.

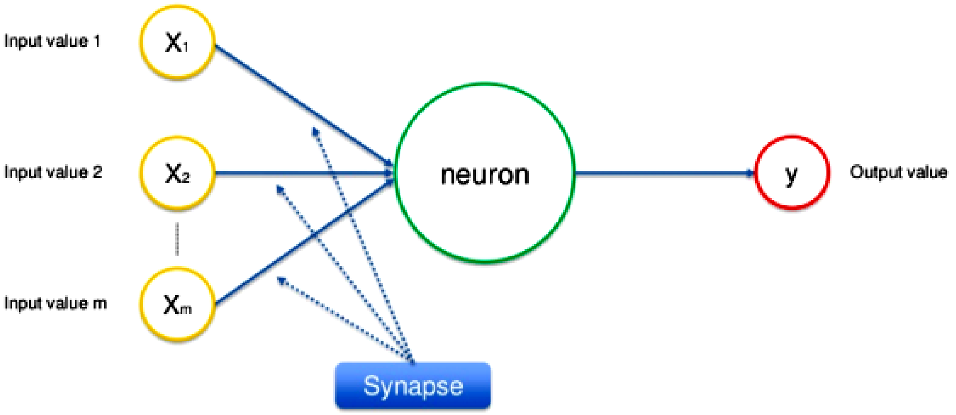
Deep learning used : Deep learning is used within businesses in a variety of industries for a wide range of use cases. Here are some examples of how deep learning is commonly used:

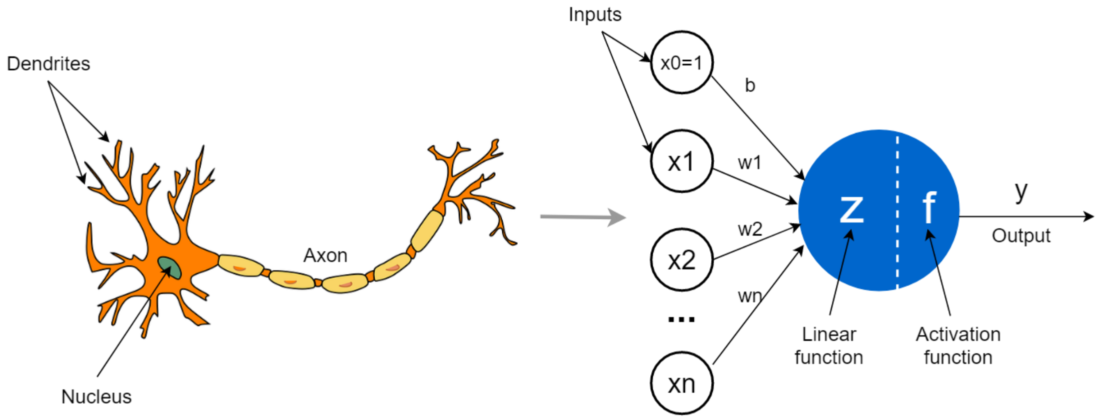
- **Image, speech, and emotion recognition :** Deep learning software is used to increase image, speech, and emotion recognition accuracy and to enable photo searches, personal digital assistants, driverless vehicles, public safety, digital security, and other intelligent technologies.
- **Tailored experiences :** Streaming services, e-commerce retailers, and other businesses use deep learning models to drive automated recommendations for products, movies, music, or other services and to perfect customer experiences based on purchase histories, past behavior, and other data.
- **Chat bots :** Savvy businesses use deep learning to power text- or voice-activated online chat bots for frequently asked questions, routine transactions, and especially for customer support. They replace teams of service agents and queues of waiting customers with automated, contextually appropriate, and useful responses.
- **Personal digital assistants :** Voice-activated personal digital assistants use deep learning to understand speech, respond appropriately to queries and commands in natural language, and even crack wise occasionally.
- **Driverless vehicles :** The unofficial representative for AI and deep learning, self-driving cars use deep learning algorithms to process multiple dynamic data feeds in split seconds, never have to ask for directions, and react to the unexpected—faster than a human driver.

Neural network

A neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain. It is a type of machine learning process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain.

It creates an adaptive system that computers use to learn from their mistakes and improve continuously. Thus, artificial neural networks attempt to solve complicated problems, like summarizing documents or recognizing faces, with greater accuracy.





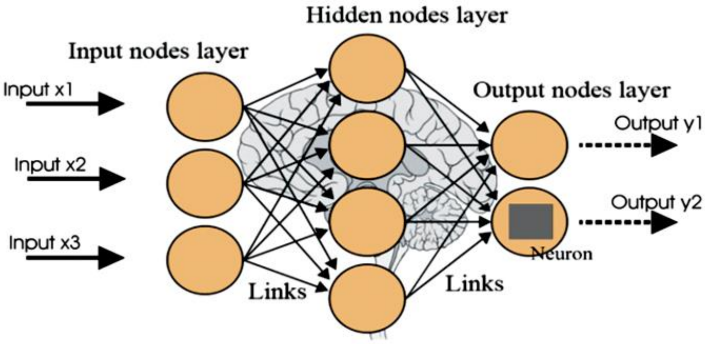
Working of a Neural Network

The rudimentary form of a neural network has three functional layers:

- The input layer
- The hidden layer
- The output layer

A neural network may contain the following 3 layers:

- Input layer – The activity of the input units represents the raw information that can feed into the network.
- Hidden layer – To determine the activity of each hidden unit. The activities of the input units and the weights on the connections between the input and the hidden units. There may be one or more hidden layers.
- Output layer – The behavior of the output units depends on the activity of the hidden units and the weights between the hidden and output units.



Neural networks uses

Neural networks have several use cases across many industries, such as the following:

- Medical diagnosis by medical image classification
- Targeted marketing by social network filtering and behavioural data analysis
- Financial predictions by processing historical data of financial instruments
- Electrical load and energy demand forecasting
- Process and quality control
- Chemical compound identification.