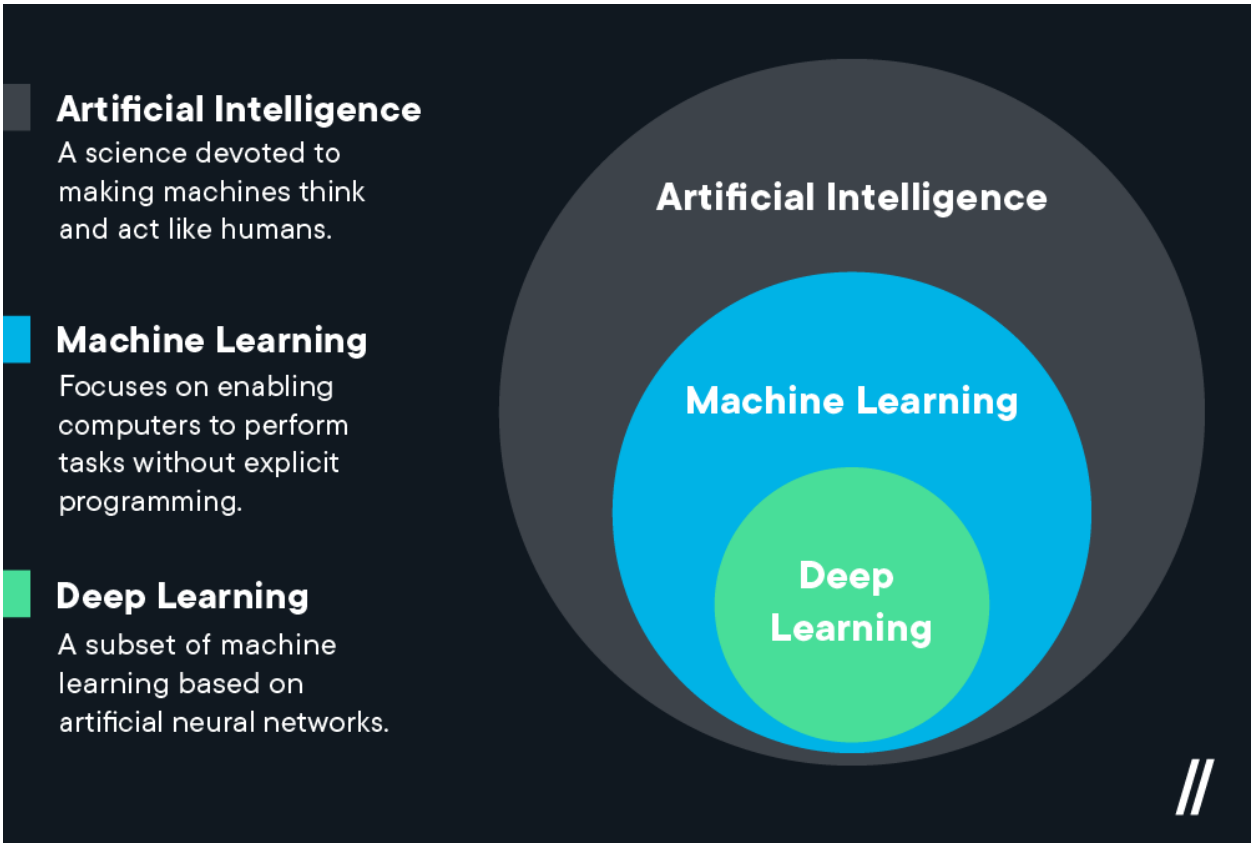
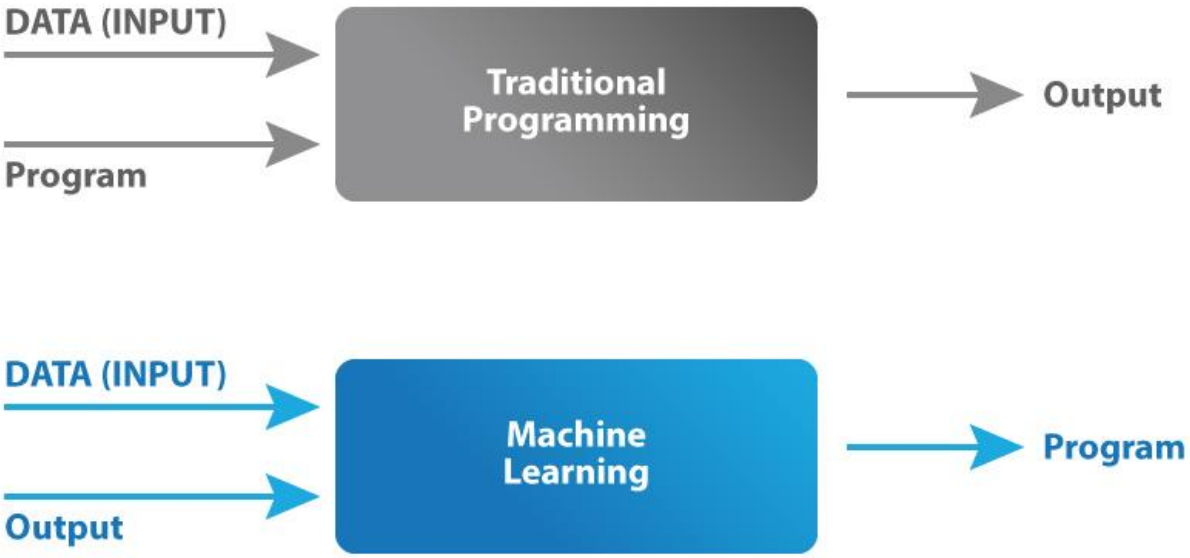
19ECT303- Artificial Intelligence and Machine Learning

UNIT I - FUNDAMENTALS OF MACHINE LEARNING

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. Machine learning is actively being used today, perhaps in many more places than one would expect. Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so.

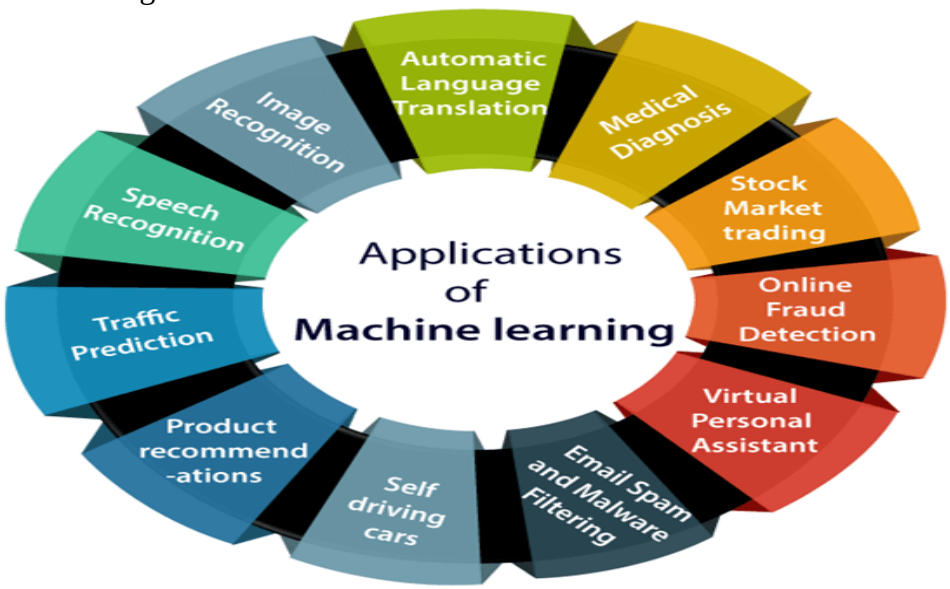




Goals and application of Machine Learning

The primary goal of machine learning research is to develop general purpose algorithms of practical value. Such algorithms should be efficient. First, the results of using machine learning are often more accurate than what can be created through direct programming.

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:



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.

It is based on the Facebook project named "Deep Face," which is responsible for face recognition and person identification in the picture.

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 Perceptron, 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.

These virtual assistants use machine learning algorithms as an important part.

These assistant record our voice instructions, send it over the server on a cloud, and decode it using ML algorithms and act accordingly.

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.

For each genuine transaction, the output is converted into some hash values, and these values become the input for the next round. For each genuine transaction, there is a specific pattern which gets change for the fraud transaction hence, it detects it and makes our online transactions more secure.

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.

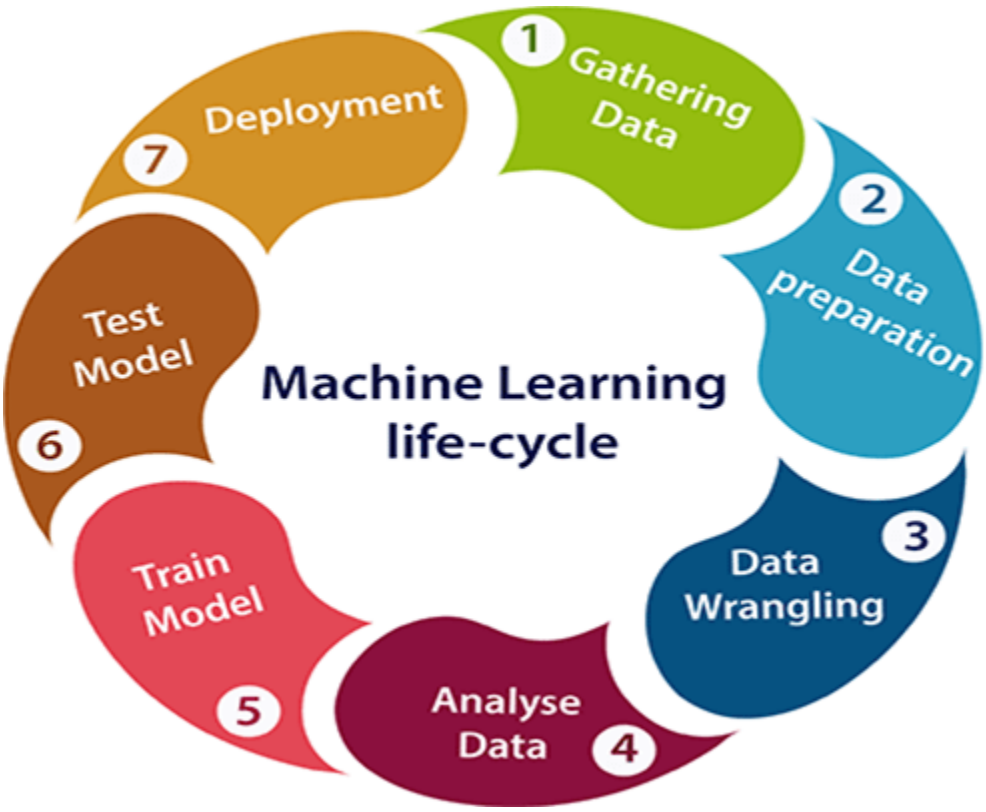
Machine learning Life cycle

Machine learning has given the computer systems the abilities to automatically learn without being explicitly programmed. But how does a machine learning system work? So, it can be described using the life cycle of machine learning. 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:

- Gathering Data
- Data preparation
- Data Wrangling
- Analyze Data

- Train the model
- Test the model
- Deployment



Machine learning Life cycle

The most important thing in the complete process is to understand the problem and to know the purpose of the problem. Therefore, before starting the life cycle, we need to understand the problem because the good result depends on the better understanding of the problem.

In the complete life cycle process, to solve a problem, we create a machine learning system called "model", and this model is created by providing "training". But to train a model, we need data, hence, life cycle starts by collecting data.

1. Gathering Data:

Data Gathering is the first step of the machine learning life cycle. 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. It is one of the most important steps of the life cycle. 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
By performing the above task, we get a coherent set of data, also called as a dataset. It will be used in further steps.

2. 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 have to 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 preprocessing of data for its analysis.
3. 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. It is one of the most important steps of the complete process. Cleaning of data is required to address the quality issues.

It is not necessary that data we have collected is always of our use as some of the data may not be useful. 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.

4. 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.

Hence, in this step, we take the data and use machine learning algorithms to build the model.

5. Train Model

Now the next step is to train the model, in this step 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.

6. Test Model

Once our machine learning model has been trained on a given dataset, then we test the 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.

7. Deployment

The last step of machine learning life cycle is deployment, where we deploy the model in the real-world system.

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 similar to making the final report for a project.

Types of Machine learning

Supervised Learning

Supervised learning is one of the most basic types of machine learning. In this type, the machine learning algorithm is trained on labeled data. Even though the data needs to be labeled accurately for this method to work, supervised learning is extremely powerful when used in the right circumstances.

In supervised learning, the ML algorithm is given a small training dataset to work with. This training dataset is a smaller part of the bigger dataset and serves to give the algorithm a basic idea of the problem, solution, and data points to be dealt with. The training dataset is also very similar to the final dataset in its characteristics and provides the algorithm with the labeled parameters required for the problem.

The algorithm then finds relationships between the parameters given, essentially establishing a cause and effect relationship between the variables in the dataset. At the end of the training, the algorithm has an idea of how the data works and the relationship between the input and the output.

This solution is then deployed for use with the final dataset, which it learns from in the same way as the training dataset. This means that supervised machine learning algorithms will continue to improve even after being deployed, discovering new patterns and relationships as it trains itself on new data.

Unsupervised Learning

Unsupervised machine learning holds the advantage of being able to work with unlabeled data. This means that human labor is not required to make the dataset machine-readable, allowing much larger datasets to be worked on by the program.

In supervised learning, the labels allow the algorithm to find the exact nature of the relationship between any two data points. However, unsupervised learning does not have labels to work off of, resulting in the creation of hidden structures. Relationships between data points are perceived by the algorithm in an abstract manner, with no input required from human beings.

The creation of these hidden structures is what makes unsupervised learning algorithms versatile. Instead of a defined and set problem statement, unsupervised learning algorithms can adapt to the data by dynamically changing hidden structures. This offers more post-deployment development than supervised learning algorithms.

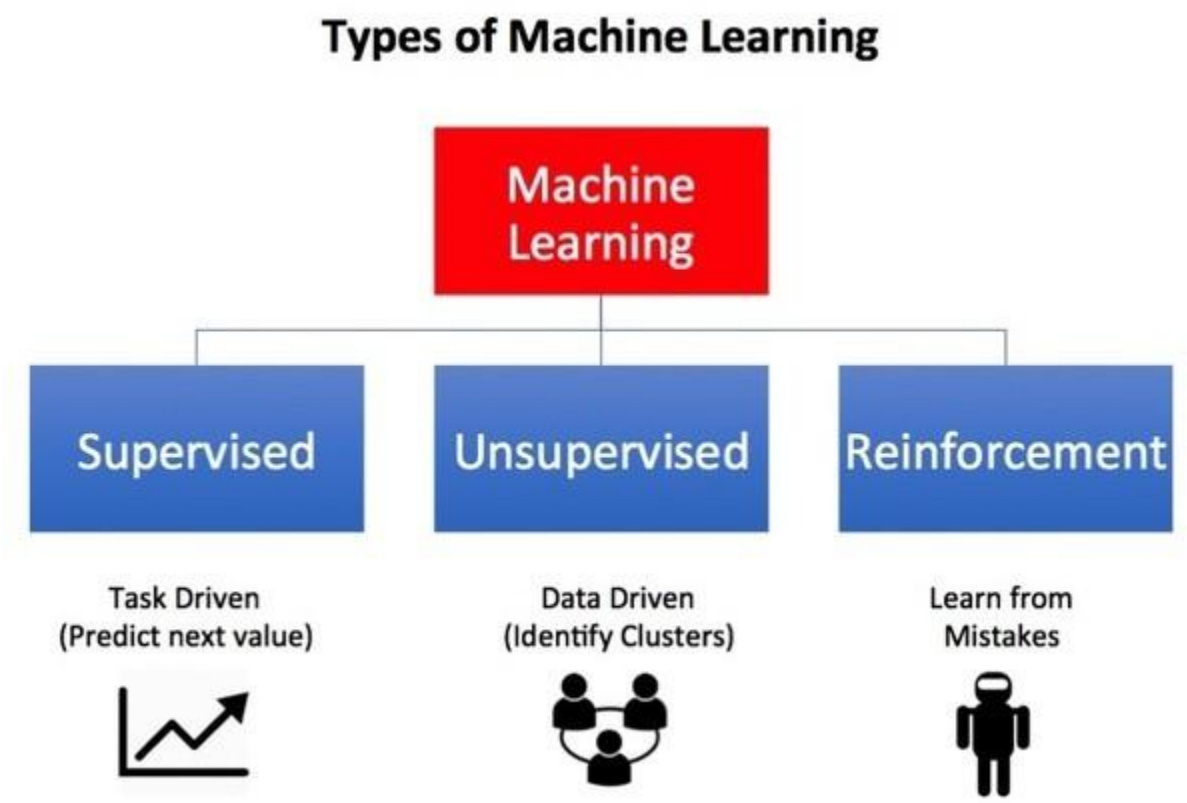
Reinforcement Learning

Reinforcement learning directly takes inspiration from how human beings learn from data in their lives. It features an algorithm that improves upon itself and learns from new situations using a trial-and-error method. Favorable outputs are encouraged or 'reinforced', and non-favorable outputs are discouraged or 'punished'.

Based on the psychological concept of conditioning, reinforcement learning works by putting the algorithm in a work environment with an interpreter and a reward system. In every iteration of the algorithm, the output result is given to the interpreter, which decides whether the outcome is favorable or not.

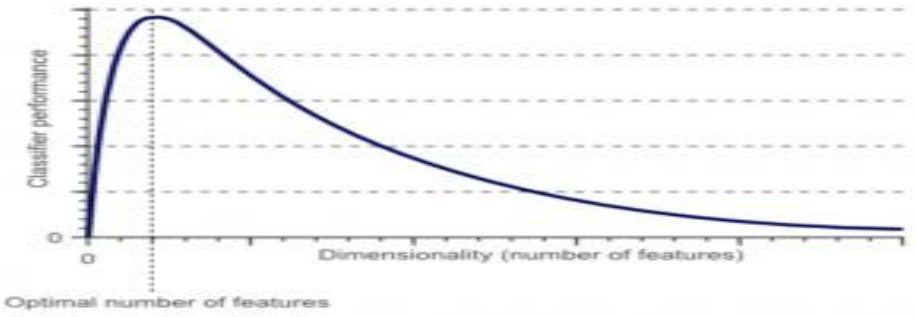
In case of the program finding the correct solution, the interpreter reinforces the solution by providing a reward to the algorithm. If the outcome is not favorable, the algorithm is forced to reiterate until it finds a better result. In most cases, the reward system is directly tied to the effectiveness of the result.

In typical reinforcement learning use-cases, such as finding the shortest route between two points on a map, the solution is not an absolute value. Instead, it takes on a score of effectiveness, expressed in a percentage value. The higher this percentage value is, the more reward is given to the algorithm. Thus, the program is trained to give the best possible solution for the best possible reward.



Curse of dimensionality in machine learning

The curse of dimensionality basically means that the error increases with the increase in the number of features. It refers to the fact that algorithms are harder to design in high dimensions and often have a running time exponential in the dimensions.



Curse of Dimensionality refers to a set of problems that arise when working with high-dimensional data. The dimension of a dataset corresponds to the number of attributes/features that exist in a dataset. A dataset with a large number of attributes, generally of the order of a hundred or more, is referred to as high dimensional data. Some of the difficulties that come with high dimensional data manifest during analyzing or visualizing the data to identify patterns, and some manifest while training machine learning models. The difficulties related to training machine learning models due to high dimensional data are referred to as 'Curse of Dimensionality'. The popular aspects of the curse of dimensionality; 'data sparsity' and 'distance concentration' are discussed in the following sections.

Data Sparsity

The supervised machine learning models are trained to predict the outcome for a given input data sample accurately. While training a model, the available data is used such that part of the data is used for training the model, and a part of the data is used to evaluate how the model performs on unseen data. This evaluation step helps us establish whether the model is generalized or not. Model generalization refers to the models' ability to predict the outcome for an unseen input data accurately. It is important to note that the unseen input data has to come from the same distribution as the one used to train the model. A generalized model's prediction accuracy on the unseen data should be very close to its accuracy on the training data. An effective way to build a generalized model is to capture different possible combinations of the values of predictor variables and the corresponding targets.

Distance Concentration

Another facet of curse of dimensionality is 'Distance Concentration'. Distance concentration refers to the problem of all the pairwise distances between different samples/points in the space converging to the same value as the dimensionality of the data increases. Several machine learning models such as clustering or nearest neighbours' methods use distance-based metrics to identify similar or proximity of the samples. Due to distance concentration, the concept of

proximity or similarity of the samples may not be qualitatively relevant in higher dimensions. Figure 3 shows this aspect graphically . A fixed number of random points are generated from a uniform distribution on a ‘d’ dimensional torus. The ‘d’ here corresponds to the number of dimensions considered at a time.

Mitigating Curse of Dimensionality

To mitigate the problems associated with high dimensional data a suite of techniques generally referred to ‘Dimensionality reduction techniques’ are used. Dimensionality reduction techniques fall into one of the two categories- ‘Feature selection’ or ‘Feature extraction’.

Feature selection Techniques

In feature selection techniques, the attributes are tested for their worthiness and then selected or eliminated. Some of the commonly used Feature selection techniques are discussed below.

Low Variance filter: In this technique, the variance in the distribution of all the attributes in a dataset are compared and attributes with very low variance are eliminated. Attributes that do not have such much variance will assume an almost constant value and do not contribute to the predictability of the model.

High Correlation filter: In this technique, the pair wise correlation between attributes are determined. One of the attributes in the pairs that show very high correlation are eliminated and the other retained. The variability in the eliminated attribute is captured through the retained attribute.

Multi-collinearity: In some cases, high correlation may not be found for pairs of attributes but if each attribute is regressed as a function of others, we may see that variability of some of the attributes are completely captured by the others. This aspect is referred to as multicollinearity and variance Inflation Factor (VIF) is a popular technique used to detect multicollinearity. Attributes with high VIF values, generally greater than 10, are eliminated.

Feature Ranking: Decision Tree models such as CART can rank the attributes based on their importance or contribution to the predictability of the model. In high dimensional data, some of the lower ranked variables could be eliminated to reduce the dimensions.

Forward selection: In building Multi-linear regression models with high dimensional data, a process can be followed in which, at the beginning, only one attribute is selected to build the regression model. later the remaining attributes are added one by one and tested for their worthiness using ‘Adjusted-R2’ values. If the Adjusted-R2 shows a noticeable improvement then the variable is retained else it is discarded.

Feature Extraction Techniques

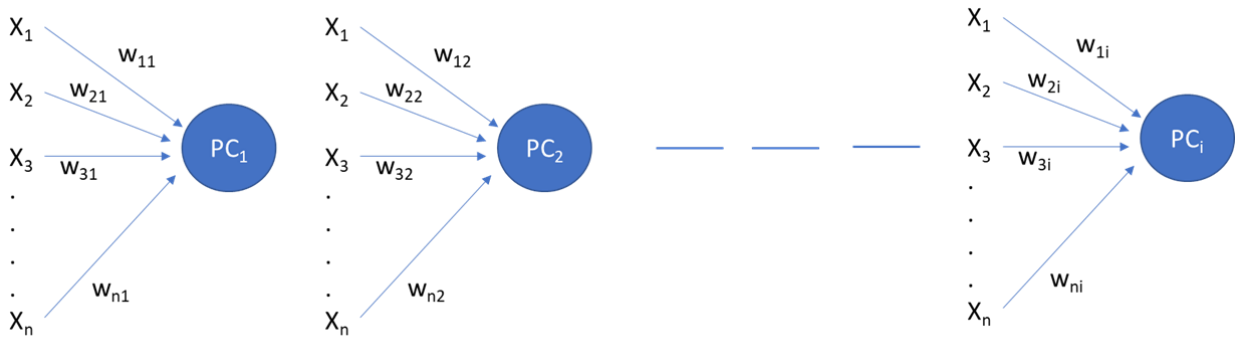
In feature extraction techniques, the high dimensional attributes are combined in low dimensional components (PCA or ICA) or factored into low dimensional factors (FA).

Principal Component Analysis (PCA)

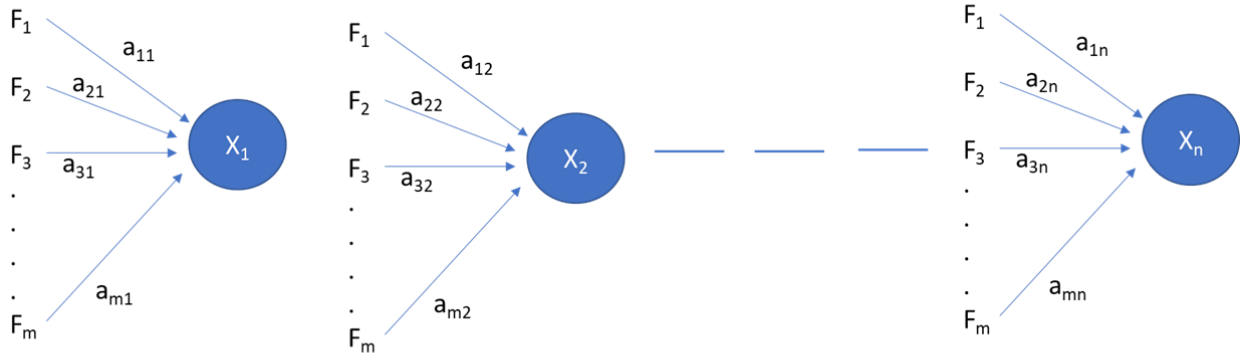
Principal Component Analysis, or PCA, is a dimensionality-reduction technique in which high dimensional correlated data is transformed to a lower dimensional set of uncorrelated components, referred to as principal components. The lower dimensional principle components capture most of the information in the high dimensional dataset. An ‘n’ dimensional data is transformed into ‘n’ principle components and a subset of these ‘n’ principle components is selected based on the percentage of variance in the data intended to be captured through the principle components. Figure 5 shows a simple example in which a 10-dimensional data is transformed to 10-principle components. To capture 90% of the variance in the data only 3 principle components are needed. Hence, we have reduced a 10-dimensional data to 3-dimensions.

Factor Analysis (FA)

Factor analysis is based on the assumption that all the observed attributes in a dataset can be represented as a weighted linear combination of latent factors. The intuition in this technique is that an ‘n’ dimensional data can be represented by ‘m’ factors ($m<n$). The main difference between PCA and FA is in the fact that While PCA synthesizes components from the base attributes, FA decomposes the attributes into latent factors .



Principle component analysis (i≤n)

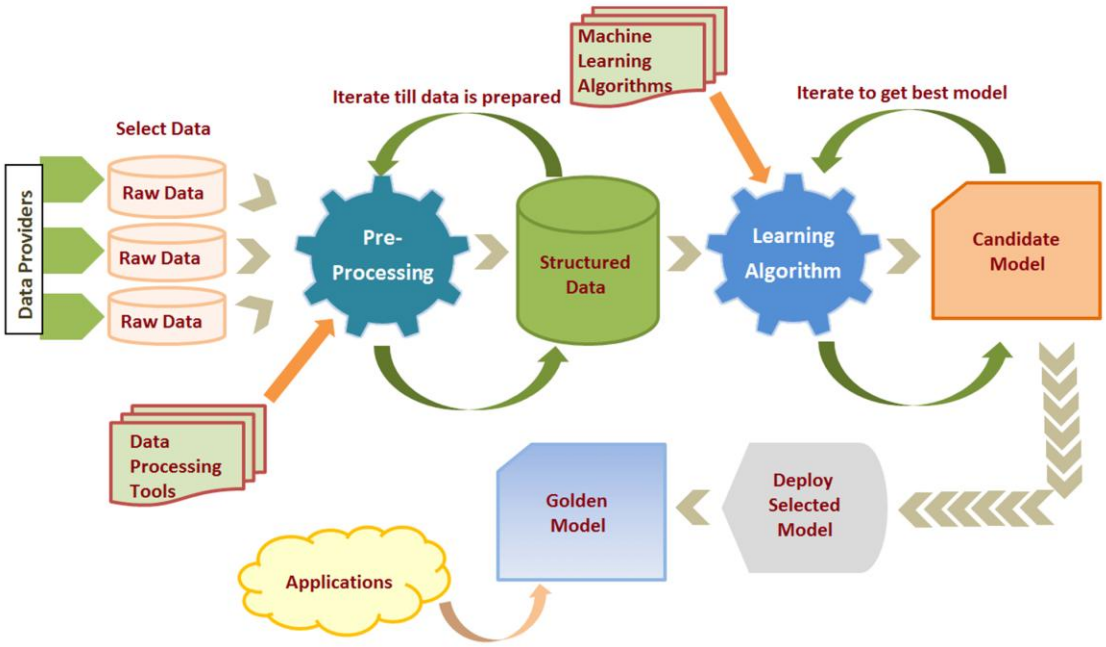


Factor analysis (m≤n)

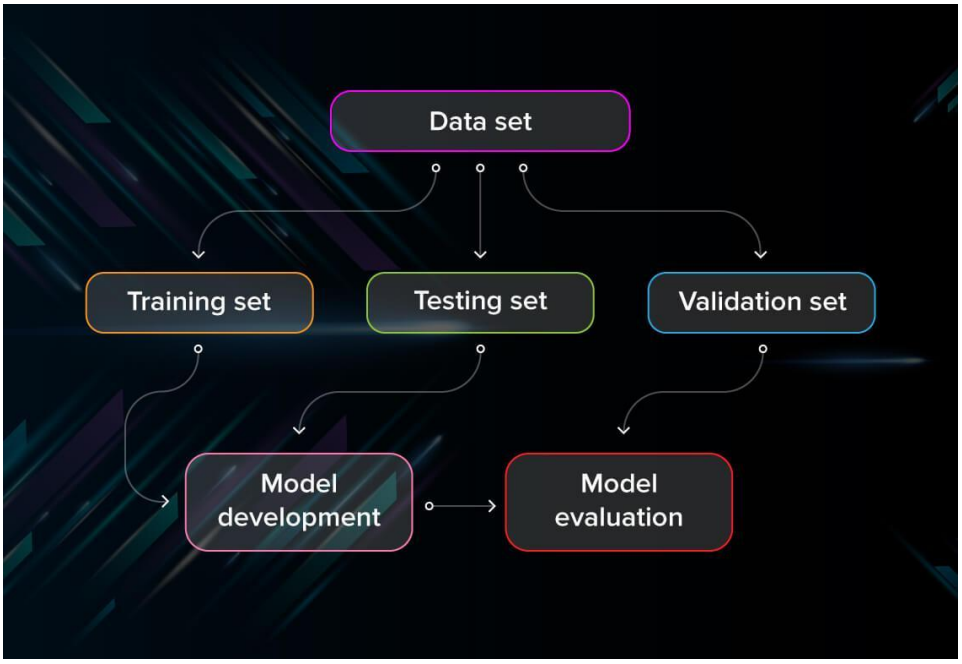
Independent Component Analysis (ICA)

ICA assumes that all the attributes are essentially a mixture of independent components and resolves the variables into a combination of these independent components. ICA is perceived to be more robust than PCA is generally used when PCA and FA fail.

Machine Learning Process



Testing Machine learning Algorithm



Performing ML tests is necessary if you care about the quality of the model. ML testing has a couple of peculiarities: it demands that you test the quality of data, not just the model, and go through a couple of iterations adjusting the hyperparameters to get the best results.

Validation set.

Having only a training set and a testing set is not enough if you do many rounds of hyperparameter-tuning (which is always). And that can result in overfitting. To avoid that, you can select a small validation data set to evaluate a model. Only after you get maximum accuracy on the validation set, you make the testing set come into the game.

Test set (or holdout set).

In order to assure that, you select samples for a testing set from your training set — examples that the machine hasn't seen before. It is important to remain unbiased during selection and draw samples at random. Also, you should not use the same set many times to avoid training on your test data. Your test set should be large enough to provide statistically meaningful results and be representative of the data set as a whole.

Cross-validation
Cross-validation is a model evaluation technique that can be performed even on a limited dataset. The training set is divided into small subsets, and the model is trained and validated on each of these samples.