MACHINE LEARNING

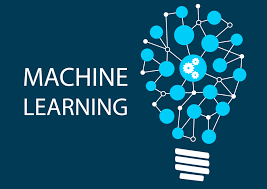
*Introduction-*

Machine Learning is a subfield of Artificial Intelligence (AI). The goal of Machine Learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people. Although ML is a field within computer science, it differs from traditional computational approaches. ML algorithms allow computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range.

Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

Any technology user today has benefitted from machine learning. Facial recognition technology allows social media platforms to help users tag and share photos of friends. Optical character recognition (OCR) technology converts images of text into movable type. Recommendation engines, powered by machine learning, suggest what movies or television shows to watch next based on user preferences. Self-driving cars that rely on machine learning to navigate may soon be available to consumers.

Machine learning is a continuously developing field. Because of this, there are some considerations to keep in mind as you work with machine learning methodologies, or analyze the impact of machine learning processes.

[](https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.blumeglobal.com%2Flearning%2Fmachine-learning%2F&psig=AOvVaw0BNrgsgmOZTEqFhYEYxZ-K&ust=1590574079188000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCKiu5LSk0ekCFQAAAAAdAAAAABAD)

*Machine Learning Methods-*

In machine learning, tasks are generally classified into broad categories. These categories are based on how learning is received or how feedback on the learning is given to the system developed.

Two of the most widely adopted machine learning methods are supervised learning which trains algorithms based on example input and output data that is labeled by humans, and unsupervised learning which provides the algorithm with no labeled data in order to allow it to find structure within its input data. The third method is Reinforcement Learning which is like a hit and trail method of learning where the machine gets a reward or penalty point for each action. Let’s explore these methods in more detail.

In any Machine Learning Method, the dataset used is first and foremost split into two. The training and testing dataset. Generally, the entire dataset is split in the ratio of 4:1 or in 80:20 percentage, where 80% of data is randomly chosen for training and the remaining data is for testing the model.

**Machine Learning**

**Reinforcement**

**Unsupervised**

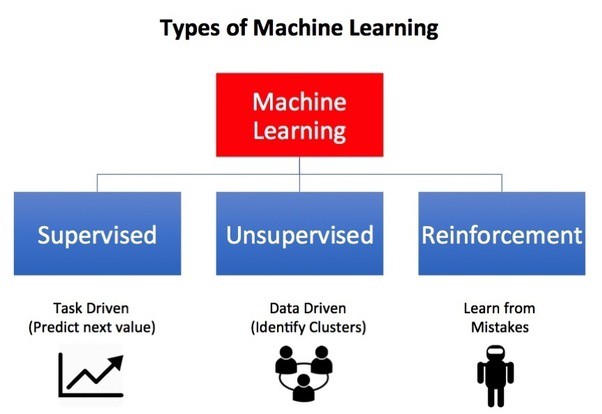
**Supervised**

Clustering

Association

Regression

Classification

[](https://www.google.com/url?sa=i&url=https%3A%2F%2Ftowardsdatascience.com%2Fwhat-are-the-types-of-machine-learning-e2b9e5d1756f&psig=AOvVaw0b_zwmKHZelH_TC5Hga6yq&ust=1590574141970000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCKC5r86k0ekCFQAAAAAdAAAAABAD)

*Supervised Learning –*

It consists of a target / outcome variable which is to be predicted from given set of predictors. Using these set of variables, we generate a function that map inputs to desired outputs. Training process continues until model achieves desired level of accuracy on training data. E.g.- Regression, Decision Tree, Support Vector Machine (SVM), KNN, Random Forest, etc.

In such learning machine is taught using labeled data. These algorithms are mainly used for predicting or forecasting outcomes. When data is fed into supervised learning models, data is labelled and sent. For instance, the machine should determine whether a picture shown is dog or cat. For this example, we first classify various dog and cat pictures (i.e. Labeled data) and input to the machine. The machine then analyses the data provided to it and comes to conclusion that if so and so features are present it might be a dog. The machine can never be sure. It always predicts the probability of an occurrence. In other words, from the above example when a new picture(testing dataset) is input apart from the training dataset, the machine gives an output saying, the picture is a probably a dog(with 80% ), meaning the machine sure for 80% but there is always a left percentage of its surety. Real time applications for such learning techniques are Risk evaluation, Sales forecast, etc, Churn prediction, etc.

*Unsupervised Learning –*

In unsupervised learning we do not have any target/outcome variable to predict or estimate. It is used for clustering population in different groups, which are widely used for segmenting customers in different groups for specific intervention.

Classification- used to predict outcome of given sample when output variable is in form of categories. E.g. sick/healthy, yes /no, sunny / rainy, etc.

Regression- used to predict outcome of given sample when output variables is in real values. E.g. amount of rainfall, height of a person.

Ensembling- combining predictions of multiple machine learning models that are individually weak to produce a more accurate prediction on new sample. E.g. bagging with random forest, boasting with XGBoost.

Association- used to discover the probability of co-occurrence of variables in a collection. It is extensively used in market-based analysis. E.g. used to discover that if customer purchases bread, she/he is likely to also purchase eggs.

Clustering- use to group samples such that objects within same cluster are more similar to each other than to objects from another cluster.

Dimensionality Reduction- used to reduce number of variables of dataset while ensuring important information is still conveyed. It can be done using Feature Extraction and Feature Selection methods. It performs data transformation from high-dimensional space to low dimensional space. Selected from a subset of original variables. E.g. PCA

*Algorithms grouped based on Similarity:*

1. Regression Algorithm

* Linear Regression
* Logistic Regression
* Logistic Regression
* Ordinary Least Square Regression (OLSR)
* Stepwise Regression
* Multivariate Adaptive Regression Splines (MARS)
* Locally Estimated Scatterplot Smoothing (LOESS)

It is concerned with modelling relationship between variables that is iteratively refined using a measure of error in prediction made by model

1. Clustering Algorithm

* K-means
* K-median
* Expectation Maximization (EM)
* Hierarchical Clustering

Typically organised by modelling approaches such as centroid based or hierarchal

1. Decision Tree Algorithm

* Classification and Regression Tree (CART)
* Iterative Dichotomiser (ID3)
* Chi-squared Automatic Interaction Detection (CHAID)
* Decision Stump
* M5
* Conditional Decision Tree

It constructs a model of decisions made based on actual values of attributes in data. Decision Tree are often fast and accurate in machine learning.

1. Instance based Algorithm

* K-nearest neighbour (kNN)
* Learning vector quantization (LVQ)
* Self-Organizing Map (SOM)
* Locally Weighted learning (LWL)
* Support Vector Machine (SVM)

It is a decision problem with instances / examples of training data that are deemed important or required to model.

1. Regularization Algorithm

* Ridge Regression
* Least Absolute Shrinkage and Select Operator (LASSO)
* Elastic Net
* Least Angle Regression

An extension to another method that penalizes models based on their complexity favouring simple models that are also better at generalizing. They are popular , powerful and generally simple modifications made to othe r methods.

1. Bayesian Algorithm

* Naïve Bayes
* Gaussian Naïve Bayes
* Multinomial Naïve Bayes
* Averaged One – Dependence Estimators
* Bayesian Belief Network

These methods explicitly apply Bayes Theorum.

1. Dimensionality Reduction Algorithm

* Principal Component Analysis (PCA)
* Principal Component Regression
* Partial Least Squared Regression (PLSR)
* Sammon Mapping
* Multidimensional Scaling (MDS)
* Project Pursuit
* Linear Discriminant Analysis (LDA)
* Mixture Discriminant Analysis (MDA)
* Quadratic Discriminant Analysis (QDA)
* Flexible Discriminant Analysis (FDA)

Dimensionality Reduction seeks and exploit inherent structure in data, but in this case in a unsupervised manner or order to summarize or describe data using less information. It can be useful visualize data or simplify data which can then be used in supervised learning method.

1. Ensemble Algorithm

* Boosting
* Bootstrapped aggregation (Boosting)
* Ada boost
* Weighted Average
* Stacked Generalization (Stacking)
* Gradient Boosting Machine (GBM)
* Gradient Boosted Regression Tree (GBRT)
* Random Forest

These models are comprised of multiple weaker models that are independently trained whose predictions are combined in same way to make overall prediction.

1. Associate Rule Learning Algorithm

* Apriori algorithm
* Eclat algorithm

It extracts rules that explain observed relationships between variables in data. These rules can discover important and commercially useful associations in large multi-dimensional datasets that can be exploited by an organization.

1. Deep Learning Algorithm

* Convolutional Neural Network (CNN)
* Recurrent Neural Network (RNN)
* Long Short-Term Memory (LSTM)
* Stacked Auto Encoders
* Deep Boltzmann Machine (DBM)
* Deep Belief Network

They are connected with building much larger and more complex neural network and many methods are concerned with large datasets of labelled analog data like image, text, audio and video.

1. Artificial Neural Network Algorithm

* Perceptron
* Multi-layer Perceptrons (MLP)
* Back propagation
* Stochastic Gradient Descent
* Hopfield Network
* Radial Basis Function Network (RBFN)

They are a class of pattern matching that are commonly used for regression and classification problems but are really enormous subfield comprised of hundreds of algorithms and variations.