

Machine Learning: Basic fundamental and Algorithm

# Introductions: Machine learning

## Definition:

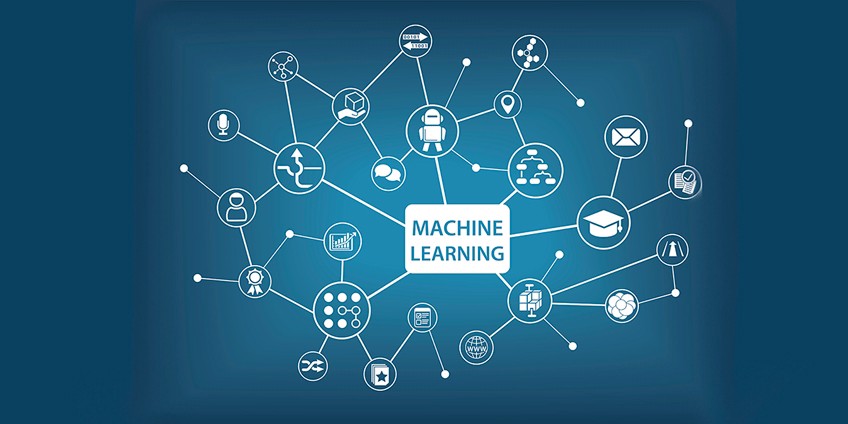


Figure 1: Machine Learning

Machine learning (ML) is the process of using mathematical models of data to help a computer learn without direct instruction. It’s considered a subset of artificial intelligence (AI). Machine learning uses algorithms to identify patterns within data, and those patterns are then used to create a data model that can make predictions. With increased data and experience, the results of machine learning are more accurate—much like how humans improve with more practice.

The adaptability of machine learning makes it a great choice in scenarios where the data is always changing, the nature of the request or task are always shifting, or coding a solution would be effectively impossible.

## Why Machine learning is important?

Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, uncovering key insights within data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics. As big data continues to expand and grow, the market demand for data scientists will increase, requiring them to assist in the identification of the most relevant business questions and subsequently the data to answer them.

## How machine learning relates to AI?

Machine learning is considered a subset of AI. An “intelligent” computer thinks like a human and carries out tasks on its own. One way to train a computer to mimic human reasoning is to use a neural network, which is a series of algorithms that are modeled after the human brain.

## How machine learning relates to predictive analytics?

Although machine learning is a type of predictive analytics, a notable nuance is that machine learning is significantly easier to implement with real-time updating as it gains more data. Predictive analytics usually works with a static dataset and must be refreshed for updates.

## How machine learning relates to deep learning?

Deep learning is a specialized form of machine learning, using neural networks (NN) to deliver answers. Able to determine accuracy on its own, deep learning classifies information like a human brain—and powers some of the most human-like AI.

# Type of Machine learning

1. Supervised Machine Learning
2. Unsupervised Machine Learning
3. Reinforcement Machine Learning
4. Semi-Supervised Machine Learning

## Structure of the Machine learning

Diagram

Description automatically generated

## Supervised Machine Learning

Supervised learning algorithms are used when the output is classified or labeled. These algorithms learn from the past data that is inputted, called training data, runs its analysis and uses this analysis to predict future events of any new data within the known classifications. The accurate prediction of test data requires large data to have a sufficient understanding of the patterns. The algorithm can be trained further by comparing the training outputs to actual ones and using the errors for modification of the algorithms.

Real-Life Example-

Image Classification – The algorithm is drawn from feeding with labeled image data. An algorithm is trained, and it is expected that the algorithm classifies it correctly in the case of the new image.

Market Prediction – It is also called Regression. Historical business market data is fed to the computer. Then, with analysis and regression algorithm, the new price for the future is predicted depending on variables.

## Unsupervised Machine Learning

Unsupervised learning algorithms are used when we are unaware of the final outputs, and the classification or labeled outputs are not at our disposal. These algorithms study and generate a function to describe completely hidden and unlabeled patterns. Hence, there is no correct output, but it studies the data to give out unknown structures in unlabeled data.

Real-Life Example-

Clustering – Data with similar traits are asked to group together by the algorithm; this grouping is called clusters. These prove helpful in the study of these groups, which can be applied to the entire data within a cluster.

High Dimension Data – High dimension data is normally not easy to work with. With the help of unsupervised learning, visualization of high dimension data becomes possible.

Generative Models – Once your algorithm analyses and comes up with the probability distribution of the input, it can be used to generate new data. This proves to be very helpful in cases of missing data.

## Reinforcement Machine Learning

This type of machine learning algorithm uses the trial-and-error method to churn out output based on the highest efficiency of the function. The output is compared to find out errors and feedback fed back to the system to improve or maximize its performance. The model is provided with rewards which are basically feedback and punishments in its operations while performing a particular goal.

Example: Applications in self-driving cars

## Semi-Supervised Machine Learning

These algorithms normally undertake labeled and unlabeled data, where the unlabeled data amount is large as compared to labeled data. As it works with both and in between supervised and unsupervised learning algorithms, therefore is called semi-supervised machine learning. Systems using these models are seen to have improved learning accuracy.

Example – An image archive can contain only some of its data labeled, e.g. Dog, cat, mouse, and a large chunk of images remain unlabeled.

# Machine Algorithm

## Linear Regression

Linear Regression analysis is used to predict the value of a variable (dependent) based on the value of another variable (independent). This analysis estimates the coefficients of the linear equation that involves one or more independent variables that best predict the value of dependent variable. The linear regression algorithm uses data points to find the best fit line to model the data. A line can be represented by the equation, y=m\*x+c where y & x are dependent variable and independent variable respectively. Basic calculus theories are then applied to find the values of m & c using the given data set. This analysis fits a straight line/surface, which minimizes the inconsistencies between the predicted and actual output values. The linear regression method can be employed in a variety of environments and programs like Sklearn linear regression, R linear regression, Linear Regression in Python, etc. Linear Regression is perhaps the simplest machine learning algorithm, and as ML practitioner Jason Brownlee emphasized in his blog, Linear Regression belongs to both the “statistics and machine learning arena.” While it is essentially a powerful statistical model, it is widely applied in machine learning for predictive modeling. Use Case: Common use cases include forecasting sales based on trends, risk assessment in financial services, house price predictions in real estate, and understanding customer behavior. Learning Resource: Want to get started with Linear Regression analysis? Click here Learn More: Top 10 Python Libraries for Machine Learning

## Logistic Regression

Logistic Regression analysis, also known as logit model, is used in predictive analytics and modeling and extends to machine learning applications. In this approach, dependent variable is finite or categorical.

Binary regression – either A or B

Multinomial regression – A range of finite options A, B, C or D

Machine Learning utilizes statistical concepts to enable computers to learn without explicit programming. The logistic approach fits best when the task which the machine is learning is based on binary classification (based on 2 values). Your computer can use this analysis to predict the likelihood of a choice being made or an event happening. Case in point, if you want to predict the likelihood of a visitor choosing an offer on your website or not. The machine leverages logistic regression analysis to make determinations about promoting your offer and take actions itself. And, with more data provided, the self-learning algorithm evolves over time. Again, this is another statistical technique applied in machine learning for a range of classification tasks. Data Analysts also use Excel Analysis ToolPak to apply logistic regression models and make predictions. Use Case: Primarily used for classification, some of the use cases include spam detection, credit card fraud detection, product categorization, and medical image classification. Learning Resource: Here’s a great primer for getting started with Logistic Regression. Learn More: How Does AI Learn Through Machine Learning Algorithms

## Decision Trees

Decision Trees algorithms are decision support tools that use tree-like models or graphs of decisions or possible consequences, such as outcome based on chance-events. Decision Trees enable you to tackle the issue at hand in a structured and systematic way to logically deduce the outcome. To work on complex, high dimensional data machine learning based decision trees come up with the cascade of questions automatically by looking at the tagged data. Earlier versions such as CART trees were used for simple data; however, for the larger dataset the commonly employed advanced algorithms used nowadays are Random Forests and Boosting Trees. Random Forests build different classifiers on a random subset of attributes and combine them for output. Boosting trees train a cascade of trees one on top of others, by automatically correcting the mistakes of ones below them. This supervised method can be used to classify data points. Use Case: Decision Tree clustering algorithms are applied for both classification and regression problems. Some of the use cases for Decision Tree include loan approvals, business decision-making. Learning Resource: Check out Artificial Intelligence: A Modern Approach by Peter Norvig and Stuart J. Russell

## K-Means

K-means algorithm is a type of unsupervised learning method that is iterative and non-deterministic. It is used when you have unlabeled data i.e. data without any defined groups/categories. The algorithm is used to find groups in the data, with the number of groups represented by the variable K. The various data points are clustered based on the feature similarity. The K-means clustering algorithm results in:

The centroids of the K clusters, that may be used to label new data

Labels for the training data where each data point is assigned to a single cluster.

Instead of letting you define groups before looking at the data, clustering enables you to find and analyze groups that have been formed organically. Use Case: Google uses K-means to cluster pages by similarities and discover the relevance of search results. It’s a machine learning tool with a moderate training time and good accuracy. As it doesn’t consist of many parameters, it’s easy to arrive at the best possible combination. The clustering tool is also applied for customer segmentation, identifying traffic patterns, and fraud detection. Learning Resource: Here’s a great guide by AI heavyweight Andrew Ng and Adam Coates on K-means and its many advantages. Learn More: 15 Best Machine Learning Books for 2020

## K-Nearest neighbors

K Nearest Neighbor (KNN) is the most popular machine learning classification and regression algorithm that’s simple, yet extremely effective. It finds applications in data mining, pattern recognition, and intrusion detection. It uses some prior data/training data that classifies coordinates into groups identified by an attribute. In other words, the algorithm stores all available cases and classifies any new cases by taking a majority vote of its k neighbors. Then the case is allotted to the class with which it matches the most. The measurement procedure is performed by a distance function. Though it takes less time to train, its accuracy may be degraded by high dimension data as there isn’t much of a difference between the nearest and farthest neighbor. It’s computationally expensive and requires the variables to be normalized. Use Case: Widely applied in retail for credit card usage detection, facial recognition and various search applications. Learning Resource: Get started on this supervised machine learning algorithm with this tutorial.

## Naïve Bayes

This is a type of classification machine learning algorithm that’s based on Bayes Theorem in probability. It works on the assumption that each feature is independent of another feature. Even if any relation exists it considers each individually, when predicting the probability of an outcome. It is the most popular learning algorithm that groups similarities together. It’s used for document classification, run checks on text expressing positive or negative emotions, spam detection, facial recognition software and predict diseases. The model has moderate training time and makes use of linearity. The Naïve Bayes algorithm is not only easy to use but it uses massive data sets efficiently. It works quite well when you have a medium to large data set to train your model. This classification algorithm is applied in a wide number of domains. Use Case: Document classification is one of the most popular use cases. Other applications include weather forecasting and fraud analysis. Learning Resource: This Google Tech Developer Guide provides resources on Naive Bayes algorithm.

## Support Vector Machine

Support Vector Machine (SVM) are supervised machine learning classification type algorithms. SVMs are used for regression as well as classification problems. They’re often used in situations where the training datasets teach the algorithm about specific classes to classify the newly included data. The raw data is a plot in n-dimensional space (where n is the number of features). Then the value of each feature is matched to a particular coordinate, making it easier to classify data. Lines known as classifiers are used to split the data and plot them on a graph. These algorithms are moderate in their accuracy and training times as it assumes linear approximation. To get the work done, these algorithms need an average number of parameters. SVM’s find application in image-based gender detection, display advertising, and image classification with large feature sets. Use Case: Some real-life applications of SVMs include image classification, handwriting recognition, and text-classification. Learn Resource: Here’s a good starting point on Support Vector Machines for text-classification.

## Apriori

Apriori is an unsupervised go-to algorithm for association rule mining. By sorting information it enhances the data management process, as the data users are appraised on the new information helping them figure out the data they are working with. It works on two basic principles, first that in case an itemset occurs frequently then all subset of itemset occur frequently and the other is that if the itemset occurs infrequently then all supersets will have infrequent occurrences. From the given data set this algorithm generates associated rules. A bottom-up approach is followed where frequently used subsets are extended one at a time and the algorithm terminates when no further extension is possible. The data can be entered into an artificial neural network or some other form of AI (artificial intelligence). However, it should be present with guiding information like dates or timestamps. The guiding information helps machine learning algorithm process categories to find patterns. A list of numbers can only be sorted based on the frequency or general amount. It often works with huge data sets including thousands of entries of qualitative or quantitative data. This algorithm is easy to implement and can be used in parallel. It is generally expensive, slow and inefficient. Use Case: One of the historical uses of Apriori algorithm includes Market Basket analysis – boosting customer sales by predicting their shopping basket. It also can be applied in healthcare to detect drug reactions. Google auto-complete is another popular application of Apriori algorithm. Learning Resource: Here’s a good primer on Apriori algorithm. Learn More: Top 10 Machine Learning Projects of 2020

## 3.9. Principal Component Analysis (PCA)

PCA is an unsupervised, non-parametric statistical technique commonly employed to reduce dimensionality in machine learning. It utilizes orthogonal transformation to alter a set of observations of possibly correlated variables into a set of values to principal components (linearly uncorrelated variable). Here are the steps involved in principal component analysis:

The first step in PCA is to standardize the data (with mean =0 and variance = 1).

Then the Covariance matrix of dimensions has to be computed.

Post this obtain the Eigenvectors and Eigenvalues from the Covariance matrix.

The Eigenvalues have to be sorted in descending order and the top k eigenvectors have to be chosen that correspond to the k largest eigenvalues (where k is the no. of dimensions of the new feature subspace k≤d, and d is the no. of original dimensions)

Construct the projection matrix W from selected k Eigenvectors

Transform X (original data set) via W to get new k-dimensional feature subspace Y.

It makes the large data simple and easy to explore and visualize. It also reduces the computational complexity of the model that makes machine learning algorithms run faster. It is not suitable for noisy data (as all components of principal component analysis have high variance). Use Case: This unsupervised algorithm finds use in classification problems and is also used for facial recognition. Learning Resource: Here’s how to implement PCA in Azure ML Studio.

## Gradient Boosting & AdaBoost

Boosting is an ensemble modeling technique that creates more powerful accurate algorithms from weak algorithms. It involves building a model by using weak models in a series. First of all, a model is built from the training data. Second model is then built which tries to rectify errors found in the first model. Gradient boosting algorithms can handle massive amounts of data accurately and speedily. There are multiple gradient boosting algorithms such as XGBoost (uses linear & tree algorithm) and LightGBM (uses tree-based algorithms). AdaBoost is the first successful boosting algorithm that was developed for binary classification. Also known as Adaptive Boosting it combines multiple weak classifiers into a single strong classifier. The main difference between AdaBoost and Gradient Boosting is that in Adaboost,“shortcomings” are identified by high-weight data points, whereas in Gradient Boosting shortcomings are identified by gradients. Use Case: Dubbed as the algorithms for champions by practitioners (used in Kaggle competitions), some real-world applications include predicting customer churn, and ad click-through rates. Learning Resource: This guide is useful for getting started on boosting algorithms. Also check out the paper on XGBoost by its creator Tianqi Chen.