Machine Learning

BY DATAFLAIR TEAM · UPDATED · JANUARY 9, 2020

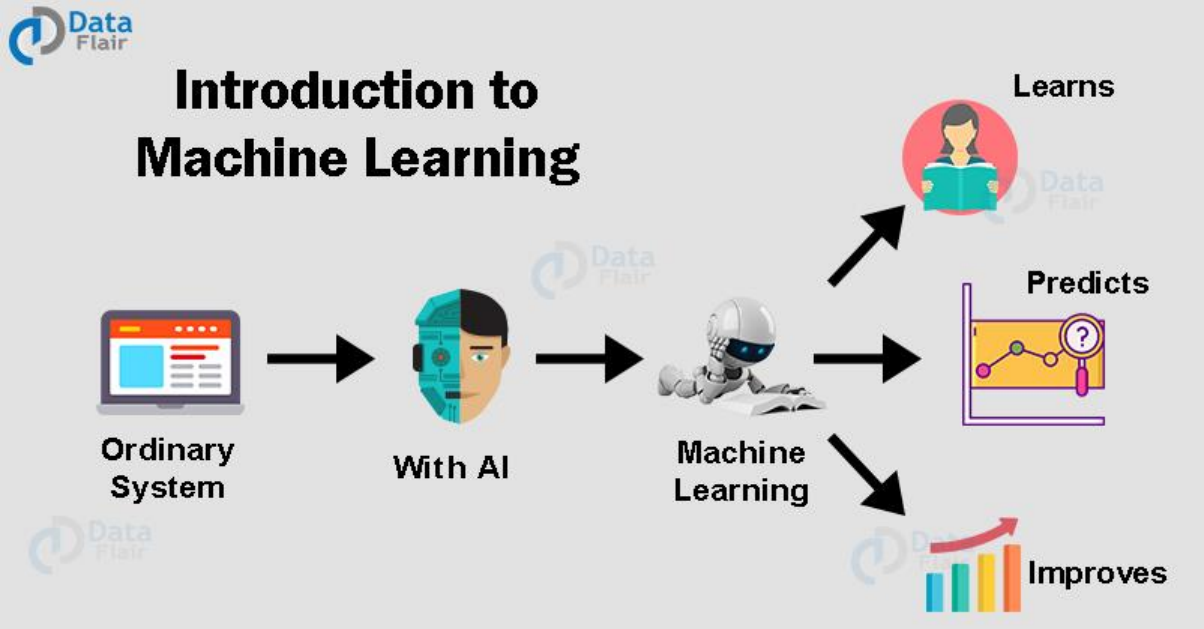
Machine Learning tutorial for beginners. As it is a beginners tutorial, I will try to make it as simple as it could be.

Have you ever went for grocery shopping? What do you do before going to the market? I always prepare a list of ingredients beforehand. Also, I make the decision according to the previous purchasing experience.

Then, I go and purchase the items. But, with the rising inflation, it's not too easy to work in the budget. I have observed that my budget gets deviated a lot of times. This happens because the shopkeeper changes the quantity and price of a product very often. Due to such factors, I have to modify my shopping list. It takes a lot of effort, research and time to update the list for every change. This is where ***Machine Learning*** can come to your rescue.
Confused MUCH?
Don't worry! Read this DataFlair's latest Machine learning tutorial to get deep insight and understand why machine learning is trending.

What is Machine Learning?

Machine Learning is the most popular technique of predicting the future or classifying information to help people in making necessary decisions. Machine Learning algorithms are trained over instances or examples through which they learn from past experiences and also analyze the historical data. Therefore, as it trains over the examples, again and again, it is able to identify patterns in order to make predictions about the future.



Machine Learning Tutorial: Introduction to Machine Learning

After knowing what machine learning is, let's take a quick introduction to machine learning and start the tutorial.

With the help of Machine Learning, we can develop intelligent systems that are capable of taking decisions on an autonomous basis. These algorithms learn from the past instances of data through statistical analysis and pattern matching. Then, based on the learned data, it provides us with the predicted results.

Data is the core backbone of machine learning algorithms. With the help of the historical data, we are able to create more data by training these machine learning algorithms. For example, ***Generative Adversarial Networks*** are an advanced concept of Machine Learning that learns from the historical images through which they are capable of generating more images. This is also applied towards speech and text synthesis. Therefore, Machine Learning has opened up a vast potential for data science applications.
Machine Learning combines computer science, mathematics, and statistics. Statistics is essential for drawing inferences from the data. Mathematics is

useful for developing ***machine learning models*** and finally, computer science is used for implementing algorithms.
However, simply building models is not enough. You must also optimize and tune the model appropriately so that it provides you with accurate results. Optimization techniques involve tuning the hyperparameters to reach an optimum result.

Machine Learning is used in every domain. It is being used to impart intelligence to static systems. With the knowledge acquired from the data, it is used to build intelligent products.

Why Machine Learning?

The world today is evolving and so are the needs and requirements of people. Furthermore, we are witnessing a fourth industrial revolution of data. In order to derive meaningful insights from this data and learn from the way in which people and the system interface with the data, we need computational algorithms that can churn the data and provide us with results that would benefit us in various ways. Machine Learning has revolutionized industries like medicine, healthcare, manufacturing, banking, and several other industries. Therefore, Machine Learning has become an essential part of modern industry.

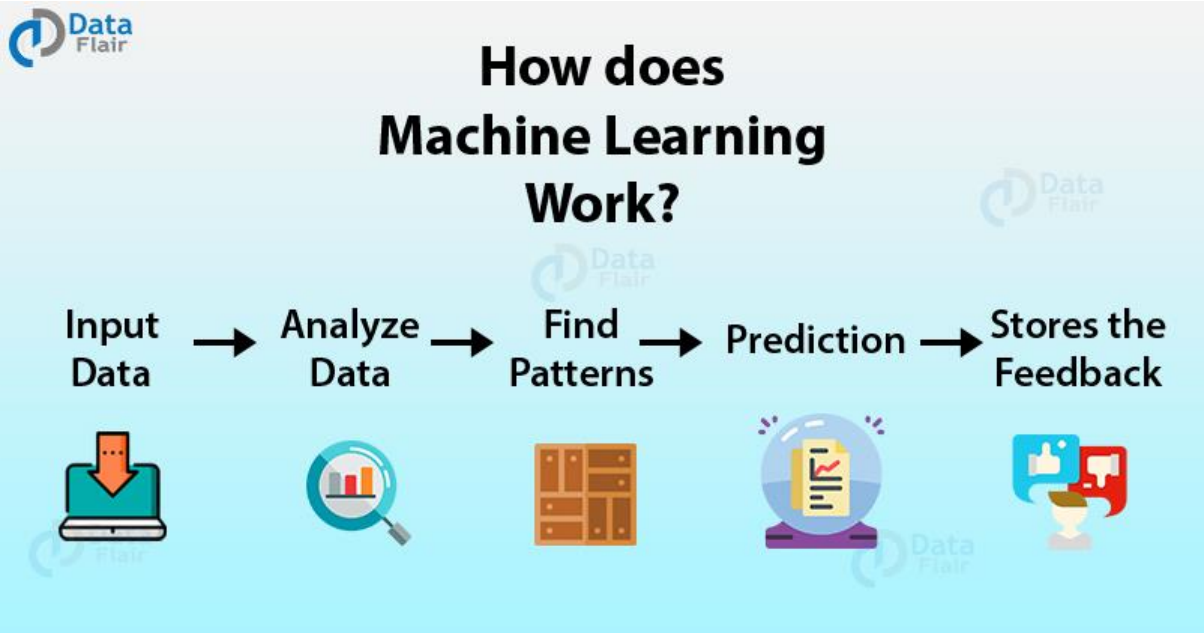
Data is expanding exponentially and in order to harness the power of this data, added by the massive increase in computation power, Machine Learning has added another dimension to the way we perceive information. Machine Learning is being utilized everywhere. The electronic devices you use, the applications that are part of your everyday life are powered by ***powerful machine learning algorithms***.
Machine Learning example – Google is able to provide you with appropriate search results based on browsing habits. Similarly, Netflix is capable of recommending the films or shows that you would want to watch based on the machine learning algorithms that perform predictions based on your watch history.

Furthermore, machine learning has facilitated the automation of redundant tasks that have taken away the need for manual labor. All of this is possible due to the massive amount of data that you generate on a daily basis. Machine Learning facilitates several methodologies to make sense of this data and provide you with steadfast and accurate results.

Wait! You need to check how Netflix and Machine Learning work together

How does Machine Learning Work?

With an exponential increase in data, there is a need for having a system that can handle this massive load of data. Machine Learning models like ***Deep Learning*** allow the vast majority of data to be handled with an accurate generation of predictions. Machine Learning has revolutionized the way we perceive information and the various insights we can gain out of it.



These machine learning algorithms use the patterns contained in the training data to perform classification and future predictions. Whenever any new input is introduced to the ML model, it applies its learned patterns over the new data to make future predictions. Based on the final accuracy, one can optimize their models using various standardized approaches. In

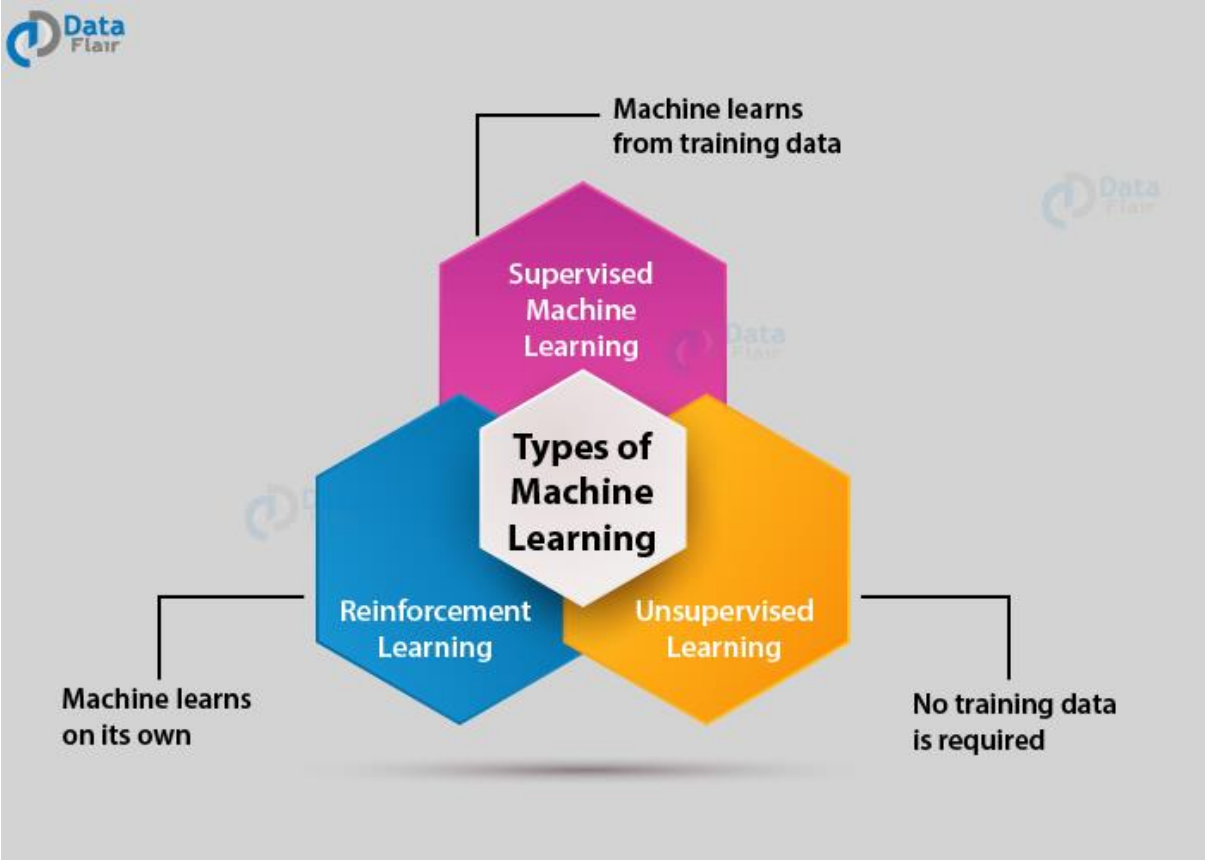
this way, Machine Learning model learns to adapt to new examples and produce better results.

Next in Machine Learning tutorial is its types. Have a look –

Types of Machine Learning

Machine Learning Algorithms can be classified into 3 types as follows –

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning



Supervised learning

In Supervised Learning, the dataset on which we train our model is labeled. There is a clear and distinct mapping of input and output. Based on the example inputs, the model is able to get trained in the instances. An example of supervised learning is spam filtering. Based on the labeled data, the model is able to determine if the data is spam or ham. This is an easier form of training. Spam filtering is an example of this type of machine learning algorithm.

Unsupervised Learning

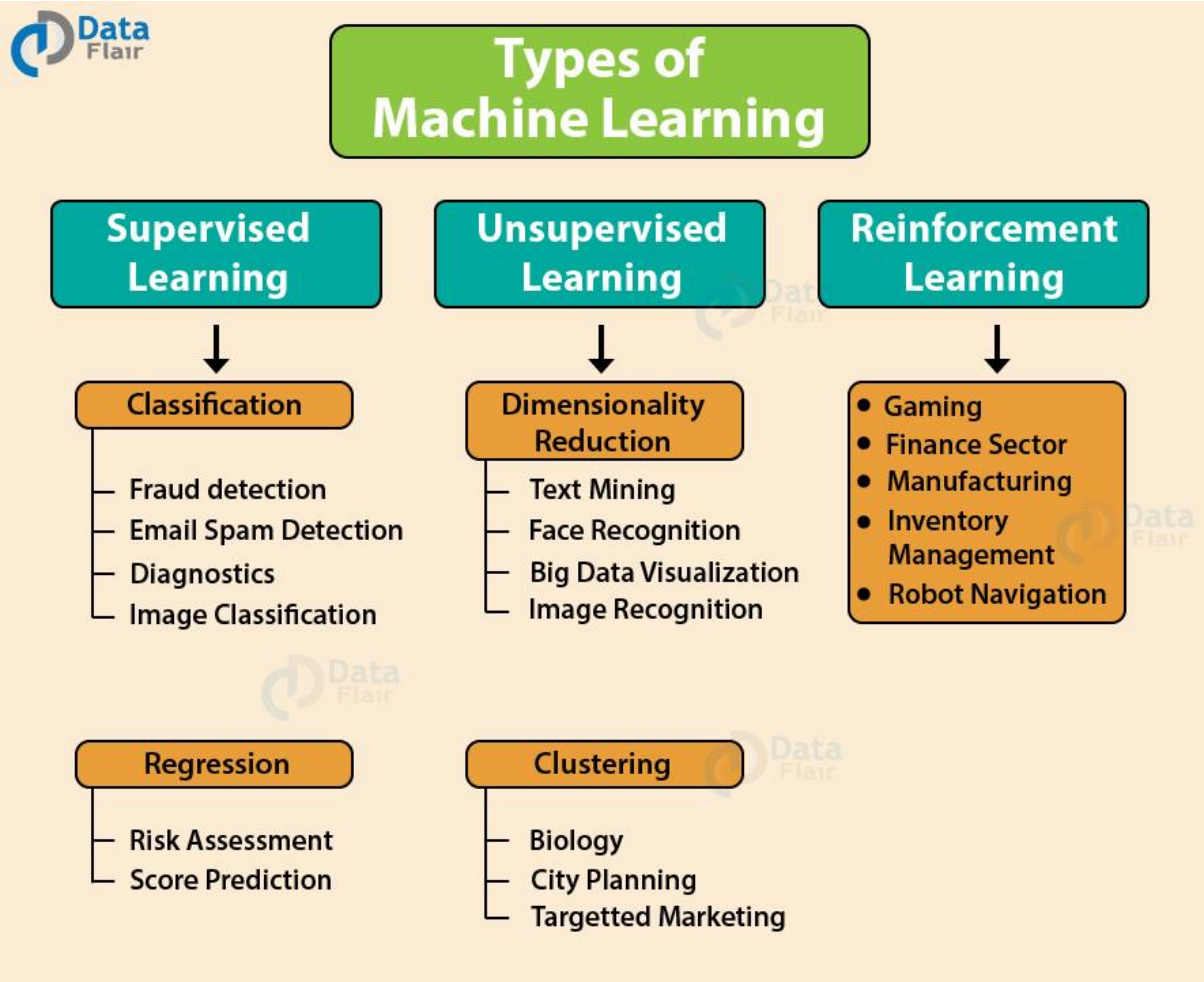
In Unsupervised Learning, there is no labeled data. The algorithm identifies the patterns within the dataset and learns them. The algorithm groups the data into various clusters based on their density. Using it, one can perform visualization on high dimensional data. One example of this type of Machine learning algorithm is the Principle Component Analysis. Furthermore, ***K-Means Clustering*** is another type of Unsupervised Learning where the data is clustered in groups of a similar order.
The learning process in Unsupervised Learning is solely on the basis of finding patterns in the data. After learning the patterns, the model then makes conclusions.

Reinforcement Learning

Reinforcement Learning is an emerging and most popular type of Machine Learning Algorithm. It is used in various autonomous systems like cars and industrial robotics. The aim of this algorithm is to reach a goal in a dynamic environment. It can reach this goal based on several rewards that are provided to it by the system.

It is most heavily used in programming robots to perform autonomous actions. It is also used in making intelligent self-driving cars. Let us consider the case of robotic navigation. Furthermore, the efficiency can be improved with further experimentation with the agent in its environment.

This the main principle behind reinforcement learning. There are similar sequences of action in a reinforcement learning model.



Machine Learning Algorithms

Let us see some most common machine learning approaches:

1. Regression

Regression models are used extensively to predict values based on the variables that are dependent on several factors. The most common example of regression is Linear Regression where there is a linear relationship or correlation between the predictor variable and the response variable. There are also other types of regression such as ARIMA regression that makes use an auto-correlation regression model to forecast continuous values

provided by the time-series data. They are used in forecasting the stock prices and other values that are based on time.

2. Decision Tree Learning

Decision Trees are a supervised type of machine learning algorithms. These trees are mainly used for predictive modeling. We create a decision tree that is able to take decisions based on user input. Decision Trees can be used for both regressions as well as classification. These trees are used to provide graphical outputs to the user based on several independent variables.

3. Support Vector Machines

Support Vector Machines or SVMs are machine learning algorithms that are used to classify data into two categories or classes. It is a type of supervised learning algorithms that makes use of several types of kernels to classify the data. Based on the prediction performed, it can categorize whether it falls into one class or any other class.
With the help of SVMs, one can perform both linear as well as non-linear classification. An SVM classifier divides the data into two classes using a hyperplane.

4. Association Rule Learning

Association Rule Mining is used for finding relationships between several variables that are present in the database. It is a type of ***data mining technique*** through which you can discover association between several items. It applied in sale industries mostly to predict if the customer will buy item Y if he has purchased the item X.

5. Artificial Neural Networks (ANN)

An Artificial Neural Network is an advanced form of machine learning technique. These neural networks are modeled after the human nervous system and are therefore called neural networks. There is a connection of several neurons which compute the information. These neurons capture the statistical structure and are therefore able to create a joint probability distribution over the input variables.

These neural networks are apt at finding patterns over large datasets. Neural Networks can perform classification as well as regression tasks with high accuracy. Furthermore, they eliminate the requirement for doing heavy statistical tasks in pre-processing as they are quite adequate in realizing patterns on their own.

6. Inductive Logic Programming

In this, logic programming forms the core part to produce a rule-like learning model. Inductive Logic Programming or ILP presents the input information, hypothesis as well as the background contextual knowledge in the form of several rules that have to be followed with logic. It makes use of functional programs to carry out inductive programming to process hypothesis in part rules.
Training models are quite often used for developing this model which is then used to forge relationships between several variables.

7. Reinforcement Learning

The aim of ***Reinforcement Learning*** is to direct the agent towards maximizing rewards and reach its goal. This takes place in a dynamic environment where the agent has to chart its way to the goal through a series of trials and errors. Each time it takes a correct route, its profit is maximized and when it encounters a wrong approach, its profit is minimized. Reinforcement Learning is widely used in self-driving cars and autonomous robotics that require self-decision making capability. Reinforcement Learnings are experimental in nature and through a series of trials are able to reach their goals with maximum accuracy (or rewards).

8. Clustering

In ***clustering***, the observations are divided into groups or clusters. These clusters are formed based on similar data and have similar criteria. These criteria can be density or similar structure of the data. There are several clustering techniques that make use of different criteria to cluster the data. For instance, the distance between the data, the density of the data and graph connectivity are some of the criteria that define techniques for clustering in machine learning. Since there are no labeled data or input-

output mapping, this type of technique is an unsupervised machine learning procedure.

9. Similarity and Metric Learning

Similarity determination is one of the key functions of machine learning. In this form of learning, the ML model is provided a mix of similar as well as dissimilar data objects. The machine learning model learns to map similar objects together and learns a similarity function that allows it to group similar objects together in the future.

10. Bayesian Networks

A ***Bayesian Network*** is an acyclic directed graphical model. This model is also called DAG which represents the probability of several independent conditioned variables. One can illustrate the relationship between disease and symptoms. It can be used to compute the probabilities of various diseases. They can be used to find the diagnosis of several diseases through a calculated approach of listing probabilities of various factors that could have contributed towards it. More advanced forms of Bayesian Networks are Deep Bayesian Networks.
The basic principle behind the Bayesian Network is the Bayes theorem which is the most important part of the probability theory. With the help of Bayes Theorem, we determine the conditional probability of an event. This conditional probability is of a known event. The conditional probability itself is the hypothesis. And, we calculate this probability based on the previous evidence.

$$P(A/B) = P(B/A)*P(A)/P(B)$$

Using a well-defined network of a connected graph, a user can make a DAG to model conditional dependencies

11. Representation Learning

In order to represent the data in a more structured format, we make use of representation learning. This formats the data efficiently so that the model can train better to provide accurate results. The representation of data is one of the key factors that can affect the performance of the machine learning method. This allows the algorithm to learn better from the data.

Using representation learning, algorithms are able to preserve the input data and essential information. Therefore, a model is able to capture most of the information during pre-processing.

Furthermore, the inputs present in pre-processing are able to gather data generating a defined distribution.

12. Sparse Dictionary Learning

In the method of Sparse Dictionary, a linear combination of basis functions as well as sparse coefficients are assumed. The elements of a sparse dictionary are called atoms. These atoms altogether compose a dictionary. It is an extension of representation learning. It is used most widely in compressed sensing and signal recovery. In this method, we represent a datum as a linear combination of basis functions and then assume the coefficients to be sparse.

So, this was all in the latest Machine learning tutorial for beginners. Many of you might find the umbrella terms Machine learning, Deep learning, and AI confusing. So, here is some additional help; below is the difference between machine learning, deep learning, and AI in simple terms.

Machine Learning vs Deep Learning vs AI

Machine Learning

Machine Learning is a part of Artificial Intelligence that involves implementing algorithms that are able to learn from the data or previous instances and are able to perform tasks without explicit instructions. The procedure for learning from the data involves statistical recognition of patterns and fitting the model so as to evaluate the data more accurately and provide us with precise results.

Deep Learning

Deep Learning is a part of Machine Learning that involves the usage of artificial neural networks. Deep Learning machine learning algorithms are the most popular choice in many industries due to the ability of neural networks to learn from large data more accurately and provide steadfast results to the user.

Artificial Intelligence

AI is the greater pool that contains an amalgamation of all the above-discussed technologies. ***Artificial Intelligence*** is still under research and involves imparting sentient intelligence to the machines. However, Artificial General Intelligence is still far fetched and will require years of research before we can have even a basic version of it.

Summary

In this machine learning tutorial, we went through the basics of machine learning and how computing power has evolved over time to accommodate advanced machine learning algorithms. Computers are gaining intelligence owing to the data that is generated in a vast amount. We went through the different types of machine learning algorithms and further took a brief look at some of the popular ML algorithms. We hope that you are now well acquainted with machine learning. Now, what next? Here is our another article – ***Applications of Machine Learning***

Top Machine Learning Algorithms used by Data Scientists

BY DATAFLAIR TEAM · UPDATED · OCTOBER 11, 2019

If you are learning machine learning for getting a ***high profile data science job*** then you can't miss out learning these 11 best machine learning algorithms.

Here, we will first go through supervised learning algorithms and then discuss about the unsupervised learning ones. While there are many more algorithms that are present in the arsenal of machine learning, our focus will be on the most popular machine learning algorithms.

These ML algorithms are quite essential for developing predictive modeling and for carrying out classification and prediction. These ML algorithms are the most useful for carrying out prediction and classification in both supervised as well as unsupervised scenarios.

Top Machine Learning Algorithms

Below are some of the best machine learning algorithms –

- Linear Regression
- Logistic Regression
- Decision Trees
- Naive Bayes
- Artificial Neural Networks
- K-means Clustering
- Anomaly Detection
- Gaussian Mixture Model
- Principal Component Analysis
- KNN
- Support Vector Machines

1. Linear Regression

The methodology for measuring the relationship between the two continuous variables is known as Linear regression. It comprises of two variables –

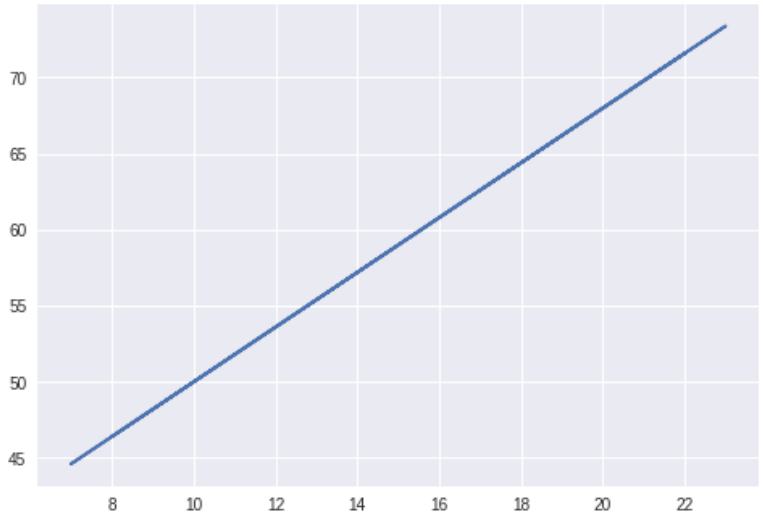
- Independent Variable – "x"
- Dependent Variable – "y"

In a simple linear regression, the predictor value is an independent value that does not have any underlying dependency on any variable. The relationship between x and y is described as follows –

$$\mathbf{y = mx + c}$$

Here, m is the slope and c is the intercept.

Based on this equation, we can calculate the output that will be through the relationship exhibited between the dependent and the independent variable.

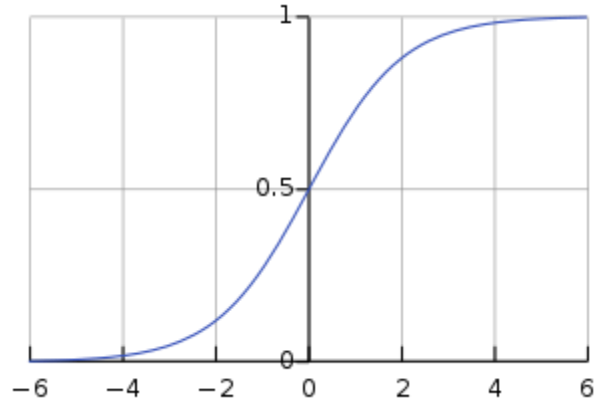


Learn linear regression in detail with DataFlair

2. Logistic Regression

This is the most popular ML algorithm for binary classification of the data-points. With the help of **logistic regression**, we obtain a categorical classification that results in the output belonging to one of the two classes. For example, predicting whether the price of oil would increase or not based on several predictor variables is an example of logistic regression.
Logistic Regression has two components – **Hypothesis and Sigmoid Curve**. Based on this hypothesis, one can derive the resultant likelihood of the event.

Data obtained from the hypothesis is then fit into the log function that forms the S-shaped curve called 'sigmoid'. Through this log function, one can determine the category to which the output data belongs to.
The sigmoid S-shaped curve is visualized as follows –



The above-generated graph is a result of this logistic equation –

1 / (1 + e^-x)

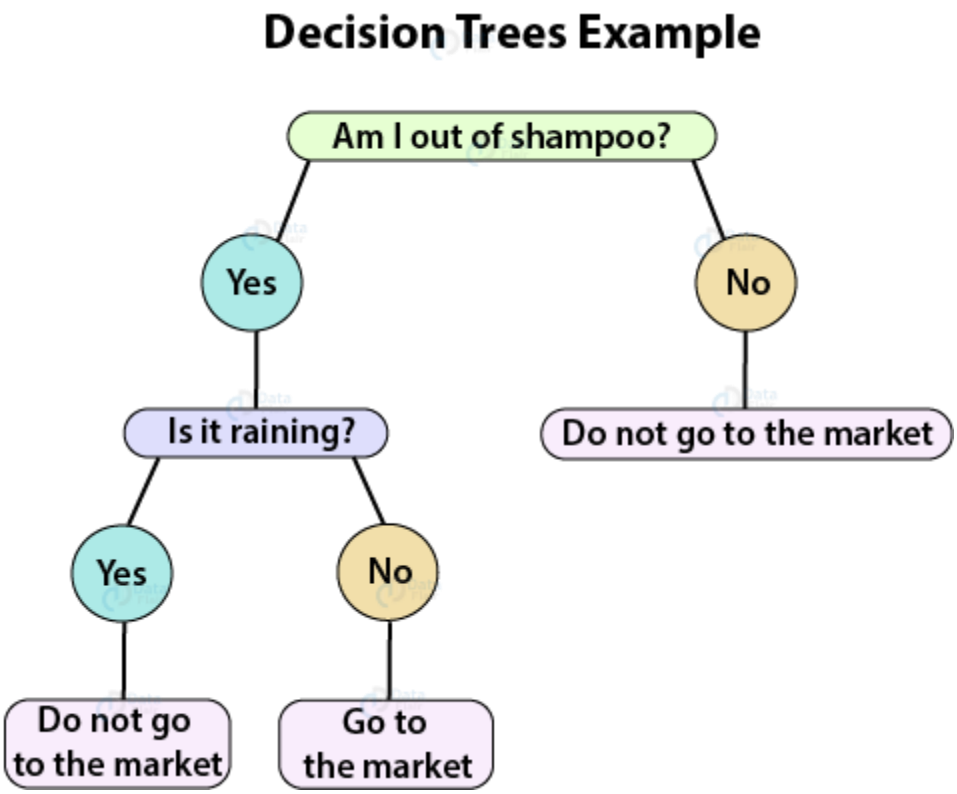
In the above equation, e is the base of the natural log and the S-shaped curve that we obtain is between 0 and 1. We write the equation for logistic regression as follows –

y = e^(b0 + b1*x) / (1 + e^(b0 + b1*x))

b0 and b1 are the two coefficients of the input x. We estimate these coefficients using the maximum likelihood function.

3. Decision Trees

Decision Trees facilitate prediction as well as classification. Using the decision trees, one can make decisions with a given set of input. Let us understand decision trees with the following example –
Let us assume that you want to go to the market to purchase a shampoo. First, you will analyze if you really do require shampoo. If you run out of it, then you will have to buy it from the market. Furthermore, you will look outside and assess the weather. That is, if it is raining, then you will not go and if it is not, you will. We can visualize this scenario intuitively with the following visualization.

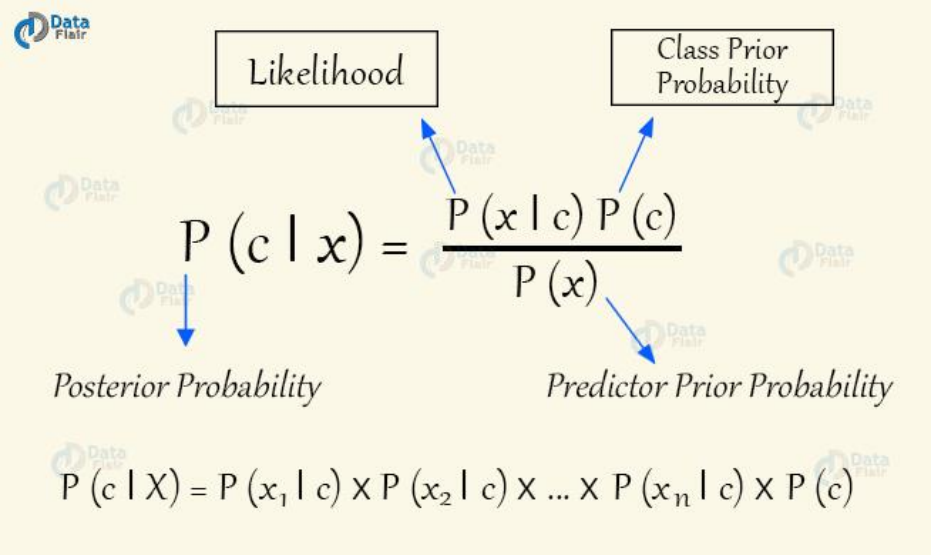


With the same principle, we can construct a hierarchical tree to obtain our output through several decisions. There are two procedures towards building a decision tree – Induction and Pruning. In Induction, we build the decision tree and in pruning, we simplify the tree by removing several complexities.

4. Naive Bayes

Naive Bayes are a class of conditional probability classifiers that are based on the Bayes Theorem. They assume independence of assumptions between the features.

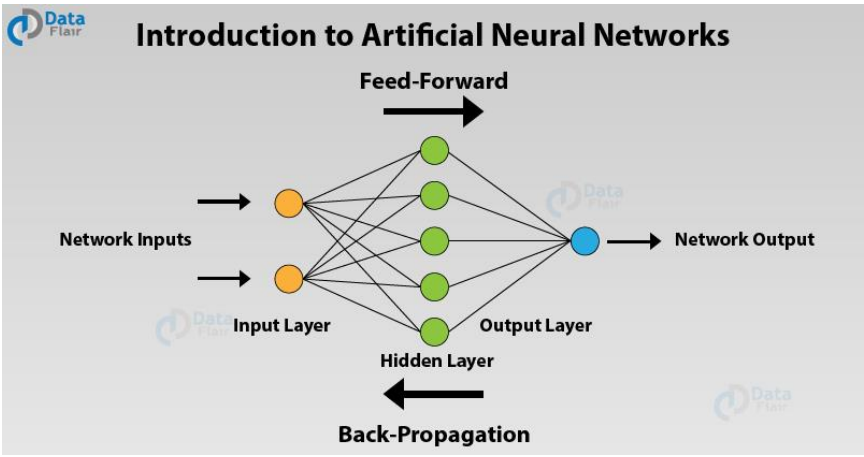
Bayes Theorem lays down a standard methodology for the calculation of posterior probability P(c|x), from P(c), P(x), and P(x|c). In a Naive Bayes classifier, there is an assumption that the effect of the values of the predictor on a given class(c) is independent of other predictor values.



Bayes Theorem has many advantages. They can be easily implemented. Furthermore, Naive Bayes requires a small amount of training data and the results are generally accurate.

5. Artificial Neural Networks

Artificial Neural Networks share the same basic principle as the neurons in our nervous system. It comprises of neurons that act as units stacked in layers that propagate information from input layer to the final output layer. These Neural Networks have an input layer, hidden layer and a final output layer. There can be a single layered Neural Network (Perceptron) or a multi-layered neural network.



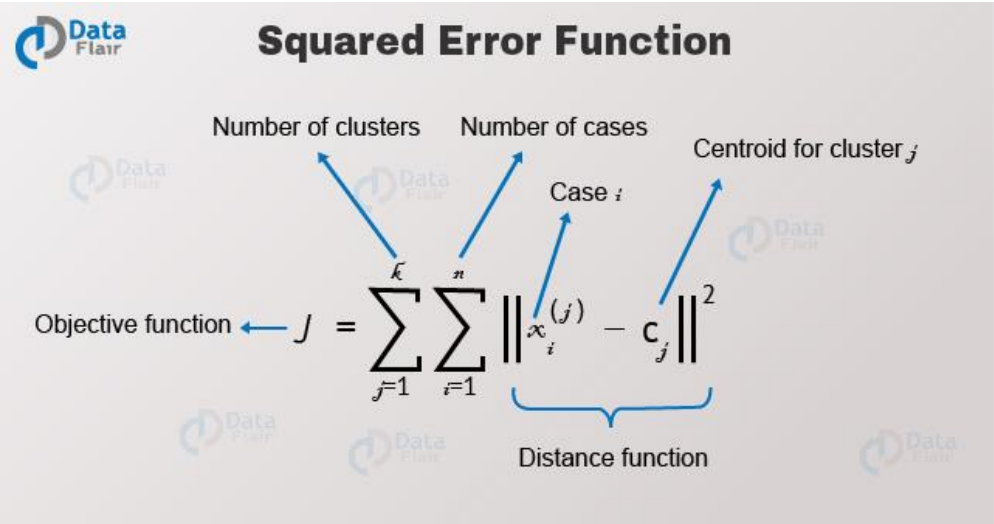
In this diagram, there is a single input layer that takes the input which is in the form of an output. Afterwards, the input is passed to the hidden layer that performs several mathematical functions to perform computation to get the desired output. For example, given an image of cats and dogs, the hidden layers compute maximum probability of the category to which our image

belongs. This is an example of binary classification in which the cat or dog is assigned an appropriate place.

6. K-Means Clustering

K-means clustering is an iterative machine learning algorithm that performs partitioning of the data consisting of n values into subsequent k subgroups. Each of the n values with the nearest mean belongs to the k cluster.
Given a group of objects, we perform partitioning of the group into several sub-groups. The sub-groups have a similar basis where the distance of each data point in the sub-group has a meaning related to their centroids. It is the most popular form of unsupervised machine learning algorithm as it is quite easy to comprehend and implement.

The main objective of a K-means clustering algorithm is to reduce the Euclidean Distance to its minimum. This distance is the intra-cluster variance which we minimize using the following squared error function –



Here, J is the objective function of the centroid of the required cluster. There are K clusters and n are the number of cases in it. There are C centroids and j are the number of clusters.We determine the Euclidean Distance from the X data-point. Let us now look at some of the important algorithms for K-means clustering –

- In the first step, we initialize and select the k-points. These k-points denote the means.

- Using the Euclidean Distance, we find the data points that lie closest to the center of the cluster.
- We then proceed to calculate the mean of all the points which will help us to find the centroid.
- We perform iterative repeat of steps 1,2 and 3 until we have all the points assigned to the right cluster.

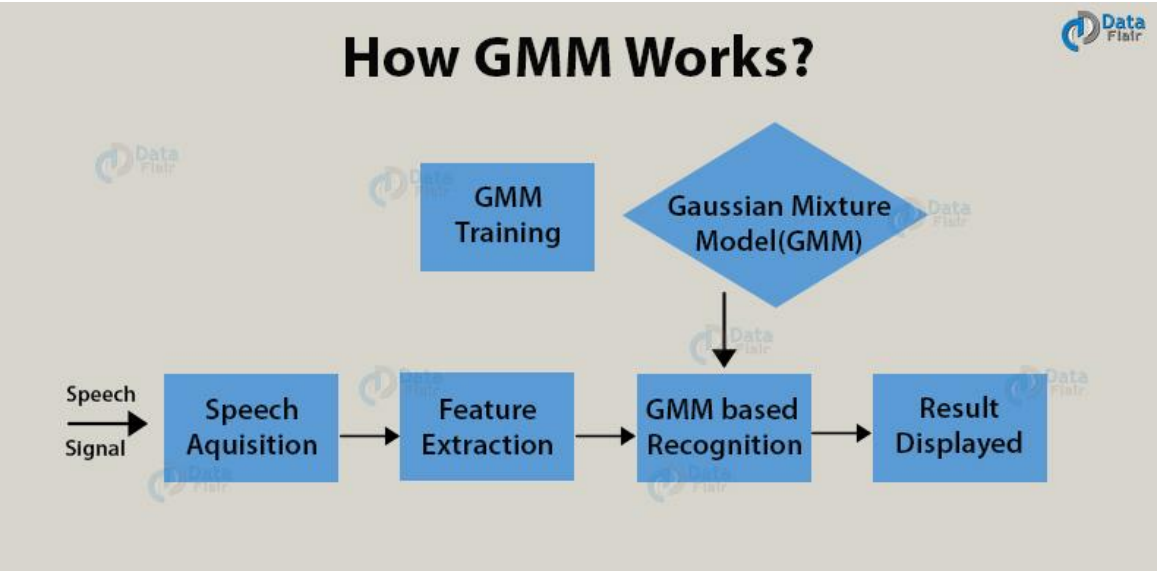
7. Anomaly Detection

In Anomaly Detection, we apply a technique to identify unusual patterns that are similar to the general pattern. These anomalous patterns or data points are known as outliers. The detection of these outliers is a crucial goal for many businesses that require intrusion detection, fraud detection, health system monitoring as well as fault detection in the operating environments.

Outlier is a rare occurring phenomena. It is an observation that is very different from the others. This could be due to some variability in measurement or simply the form of an error.

8. Gaussian Mixture Model

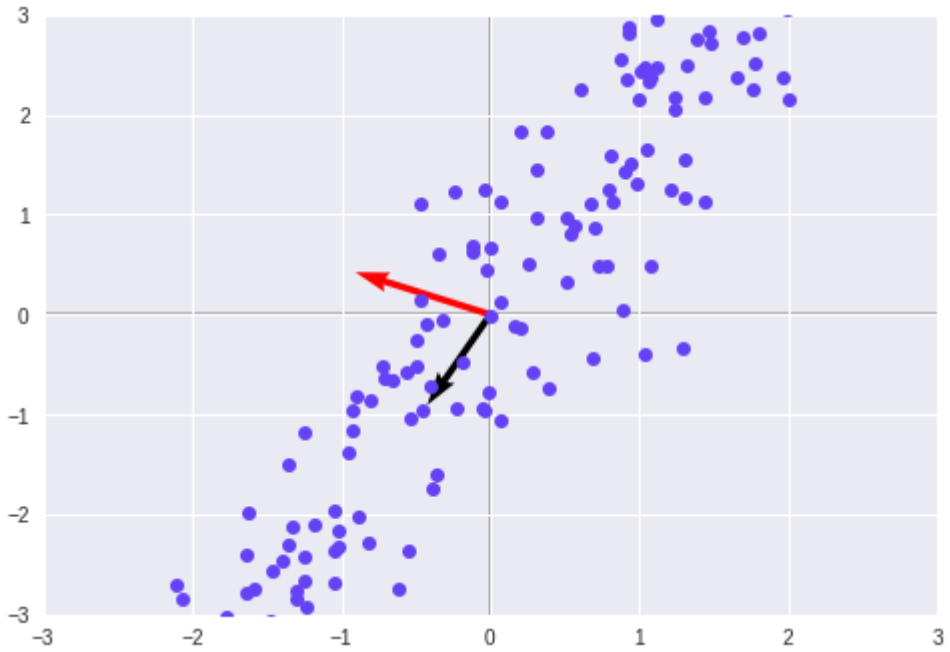
For representing a normally distributed subpopulation within an overall population, ***Gaussian Mixture Model*** is used. It does not require the data associated with the subpopulation. Therefore, the model is able to learn subpopulations automatically. As the assignment of the population is unclear, it comes under the category of unsupervised learning.
For example, assume that you have to create a model of the human height data. The mean height of males in male distribution is 5'8'' and for females, it is 5'4''. We are only aware of the height data and not the gender assignment. Distribution follows the sum of two scaled and two shifted normal distributions. We make this assumption with the help of the Gaussian Mixture Model or GMM. GMM can also have multiple components.
Using GMMs, we can extract important features from the speech data, we can also perform tracking of the objects in cases that have a number of mixture components and also the means that provide a prediction of the location of objects in a video sequence.



9. Principal Component Analysis

Dimensionality reduction is one of the most important concepts of Machine Learning. A data can have multiple dimensions. Let these dimensions be n. For instance, let there be a data scientist working on financial data which includes credit score, personal details, salary of the personnel and much more. For understanding significant labels contributing towards our model, we use dimensionality reduction. ***PCA*** is one of the most popular algorithms for reducing the dimensions.

Using PCA, one can reduce the number of dimensions while preserving the important features in our model. The PCAs are based on the number of dimensions and each PCA is perpendicular to the other. The dot product of all of the perpendicular PCAs is 0.



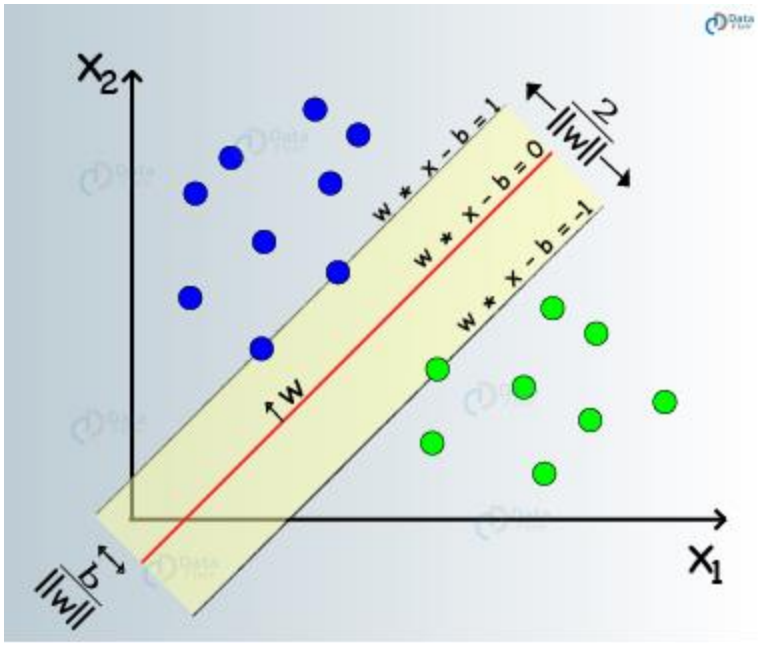
10. KNN

KNN is one of the many supervised machine learning algorithms that we use for data mining as well as machine learning. Based on the similar data, this classifier then learns the patterns present within. It is a non-parametric and a lazy learning algorithm. By non-parametric, we mean that the assumption for underlying data distribution does not hold valid. In lazy loading, there is no requirement for training data points for generating models.

The training data is utilized in testing phase causing the testing phase slower and costlier as compared with the training phase.

11. Support Vector Machines (SVM)

Support Vector Machines are a type of supervised machine learning algorithms that facilitate modeling for data analysis through regression and classification. SVMs are used mostly for classification. In SVM, we plot our data in an n-dimensional space. The value of each feature in SVM is same as that of specific coordinate. Then, we proceed to find the ideal hyperplane differentiating between the two classes.
Support Vectors represent the coordinate representation of individual observation. Therefore, it is a frontier method that we utilize for segregating the two classes.



Conclusion

In this article, we went through a number of machine learning algorithms that are essential in the data science industry. We studied a mix of supervised as well as unsupervised learning algorithms that are quite essential for the implementation of machine learning models. So, now you are ready to apply these ML algorithms concepts in your next data science job.

Hope you liked this machine learning algorithms article. Do share your feedback in the comment section below.

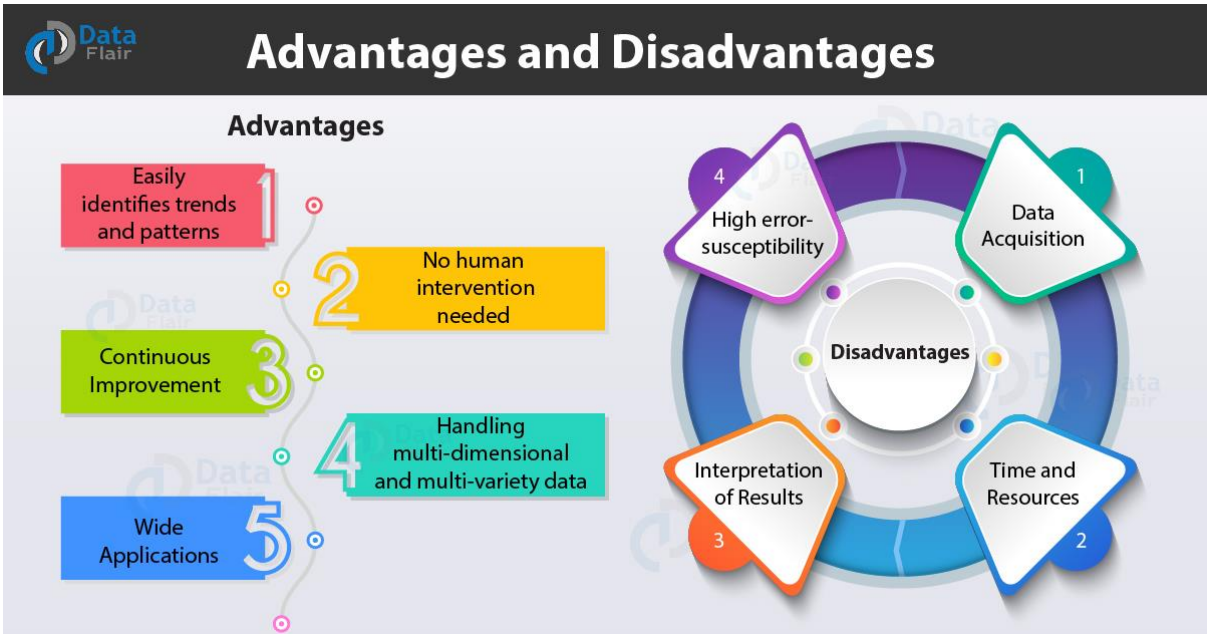
DON'T MISS – Top Machine Learning Projects with Source Code

Advantages and Disadvantages of Machine Learning Language

BY DATAFLAIR TEAM · UPDATED · MARCH 1, 2019

Amidst all the hype around **Big Data,** we keep hearing the term "Machine Learning". Not only does it offer a remunerative career, it promises to solve problems and also benefit companies by making predictions and helping them make better decisions. In this blog, we will learn the Advantages and Disadvantages of Machine Learning. As we will try to understand where to use it and where not to use Machine learning.

So, let's start the Advantages and Disadvantages of Machine Learning.



Advantages and Disadvantages of Machine Learning Language

Every coin has two faces, each face has its own property and features. It's time to uncover the faces of ML. A very powerful tool that holds the potential to revolutionize the way things work.

Advantages of Machine learning

1. Easily identifies trends and patterns

Machine Learning can review large volumes of data and discover specific trends and patterns that would not be apparent to humans. For instance, for an e-commerce website like Amazon, it serves to understand the browsing behaviors and purchase histories of its users to help cater to the right products, deals, and reminders relevant to them. It uses the results to reveal relevant advertisements to them.

[Do you know the Applications of Machine Learning?]

2. No human intervention needed (automation)

With ML, you don't need to babysit your project every step of the way. Since it means giving machines the ability to learn, it lets them make predictions

and also improve the algorithms on their own. A common example of this is anti-virus softwares; they learn to filter new threats as they are recognized. ML is also good at recognizing spam.

3. Continuous Improvement

As **ML algorithms** gain experience, they keep improving in accuracy and efficiency. This lets them make better decisions. Say you need to make a weather forecast model. As the amount of data you have keeps growing, your algorithms learn to make more accurate predictions faster.

4. Handling multi-dimensional and multi-variety data

Machine Learning algorithms are good at handling data that are multi-dimensional and multi-variety, and they can do this in dynamic or uncertain environments.

5. Wide Applications

You could be an e-tailer or a healthcare provider and make ML work for you. Where it does apply, it holds the capability to help deliver a much more personal experience to customers while also targeting the right customers.

Disadvantages of Machine Learning

With all those advantages to its powerfulness and popularity, Machine Learning isn't perfect. The following factors serve to limit it:

1. Data Acquisition

Machine Learning requires massive data sets to train on, and these should be inclusive/unbiased, and of good quality. There can also be times where they must wait for new data to be generated.

2. Time and Resources

ML needs enough time to let the algorithms learn and develop enough to fulfill their purpose with a considerable amount of accuracy and relevancy. It also needs massive resources to function. This can mean additional requirements of computer power for you.

Also, see the future of Machine Learning

3. Interpretation of Results

Another major challenge is the ability to accurately interpret results generated by the algorithms. You must also carefully choose the algorithms for your purpose.

4. High error-susceptibility

Machine Learning is autonomous but highly susceptible to errors. Suppose you train an algorithm with data sets small enough to not be inclusive. You end up with biased predictions coming from a biased training set. This leads to irrelevant advertisements being displayed to customers. In the case of ML, such blunders can set off a chain of errors that can go undetected for long periods of time. And when they do get noticed, it takes quite some time to recognize the source of the issue, and even longer to correct it.

Summary

As a result, we have studied Advantages and Disadvantages of Machine Learning. Also, this blog helps an individual to understand why one needs to choose machine learning. While Machine Learning can be incredibly powerful when used in the right ways and in the right places (where massive training data sets are available), it certainly isn't for everyone. You may also like to read

Tags: best machine learning algorithmsList of ml algorithms