**Machine Learning 101**

Machine Learning has become the baseline for majority of the intelligent systems today. Whether the system is in the domain of Artificial Intelligence or Data Science, it is considered the building block for embryonic systems. Despite its prevalence in present-day systems, the basics of Machine Learning are often confused with those of AI or Data Science. This blog would make an effort to clarify what Machine Learning is about, and what it entails.

Machine Learning is generally defined as enabling computers to learn on the basis of data being fed, without being explicitly programmed. The fact that there are almost 3.9 billion internet users all over the world indicates the data being generated is humungous. To deal with massive data, Machine Learning models are highly convenient given they have optimum resources.

**Machine Learning Terminologies:**

1. Dataset

The dataset is defined as the collection of data. It can be comma separated values (csv), database or a matrix. Each dataset is divided into three subsets:

* 1. Training set:

This subset is used for training the machine learning model. The parameters of the model are gauged using the training data.

* 1. Test set:

This subset is used for assessing the machine learning model. It is independent of the training data but follows the same distribution as the training data.

* 1. Cross Validation set:

This portion of dataset is used to compare the performances of the machine learning models that were created based on the training set. This enables us to tune the parameters of the model for better performance. The algorithm with the best performance is then chosen.

The most common approach towards separating the dataset into the above-mentioned subsets is dividing the dataset randomly into 80/20 ratio. The 80% training data is then again split into 80/20 ratio, with 20% allocated to cross validation set.

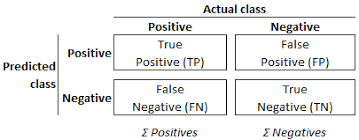
1. Cost Function

The cost function associated with every algorithm is defined as the ‘cost’ or error it incurs. It measures how wrong an algorithm is while predicting the output labels.

1. Confusion Matrix

A confusion matrix is a visual representation of the performance of the machine learning model. It classifies all predictions into four categories and calculates various metrics e.g. precision, recall and others.

* 1. True Positives: When the predictions and actual label, both are true.
  2. False Positives: When the predictions are true, but the actual label are not.
  3. True Negatives: When the predictions and actual label, both are false.
  4. False Negatives: When the predictions are false, but the actual label are not.



1. Outliers:

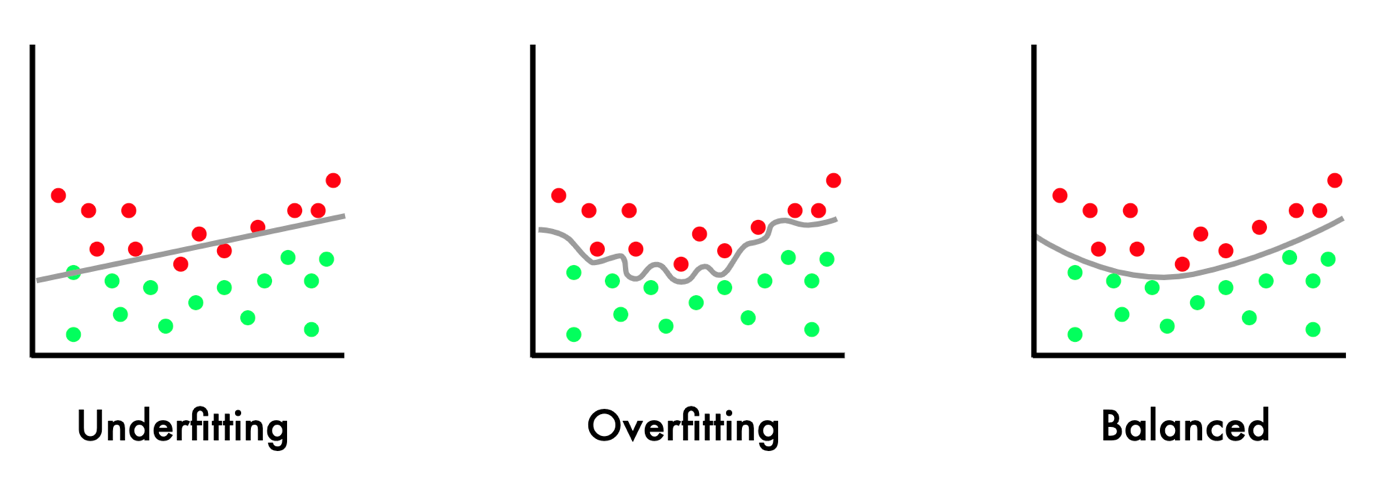
Outliers or anomalies are the data that does not fit to the underlying pattern of the data.

1. Underfitting:

Underfitting occurs when the model is unable to capture the trend in underlying data. The training set error and test set error is both high, and the model is unable to generalize to new data.

1. Overfitting

Overfitting occurs when the model performs well on training set but is unable to fit well to test data. The training set error is significantly lower than that of test set error.



**Machine Learning Types:**

The machine learning algorithms are generally divided into three types:

1. Supervised Learning

This approach involves using a labeled dataset i.e. the output of each training example in the dataset is provided. Some examples that rely on this type of learning are Linear Regression and Decision Trees.

1. Unsupervised Learning

This approach involves using an unlabeled dataset i.e. the output of each training example in the dataset is not provided. Some examples that rely on this type of learning are K-Mean Clustering.

1. Reinforcement Learning

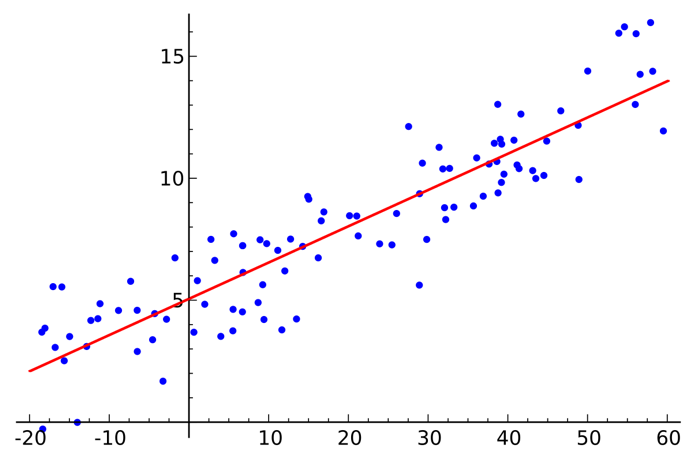
This approach is used when the purpose is to increase efficiency. The model interacts with the environment and then calculates the cost or reward. One of the major characteristic of this approach is training itself continuously via iterative trial and error and choosing the one with maximum performance.

**Machine Learning Algorithms:**

While the approaches to solving a problem statement differ, one of the most vital part is to choose the right algorithm to decipher it. Below are the few algorithms that are most commonly used:

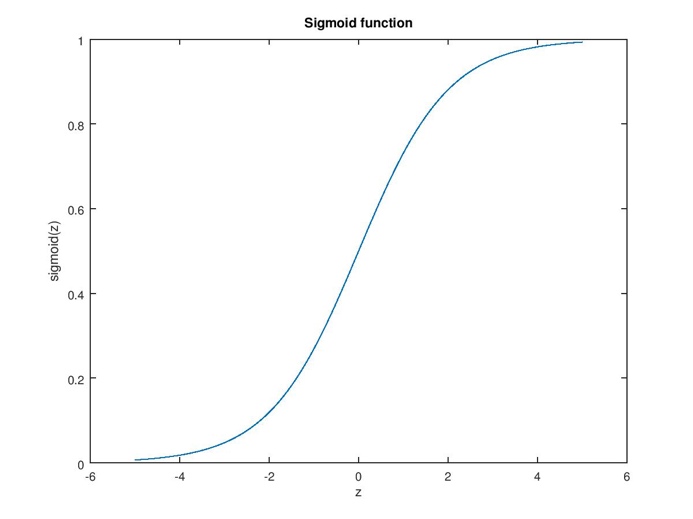
1. Linear Regression

Linear Regression is used to predict continuous output values. It is represented by an equation that uses ‘line of best fit’ to fit to the underlying data. It is one of the oldest, simplest and fastest technique, and a good algorithm to begin with if you have just started with Machine Learning Algorithms.



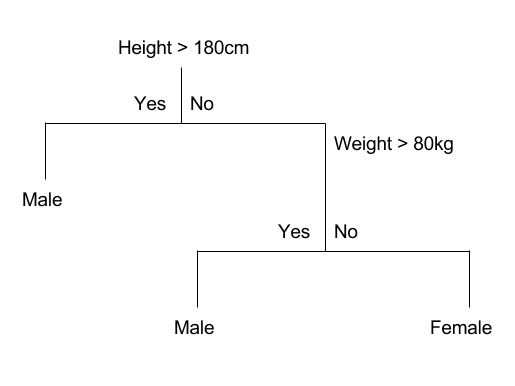
1. Logistic Regression

Logistic Regression is used to predict discrete values. To fit the values between 0 and 1, a sigmoid function is used. You can either use the output directly as probability or define a threshold for outputs e.g. if output is greater than 0.5, assign 1 to the label. It is also an efficient and simple algorithm.



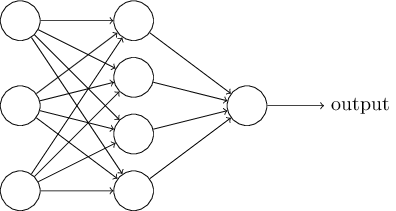
1. Decision Trees

Decision Trees solve regression and classification problems by using tree-like model for decision making. The dataset is split into smaller subsets, and the resulting tree is made up of decision nodes and leaf nodes. The leaf nodes are the output labels.



1. Neural Networks

Neural Networks are developed on the working of human brain. It is made up of input layer, output layer and sometimes, hidden layers i.e. the layers between input and output layer. Each layer is made up of nodes, which can be called a function per se. Each node has a ‘weight’, which numerates the preference of the node.



1. Naïve Bayes

Naïve Bayes is based on the assumption that each feature is independent of the other feature. It calculates conditional probability of each class based on the features, using it to determine the output labels. It is a simple algorithm, which is easy to train, and can be more appropriate at times than complex algorithms.

In this blog, I have tried to give an overview of Machine Learning for beginners. There are many complex algorithms, which were not mentioned due to the limited scope of this blog.