Literacy Review

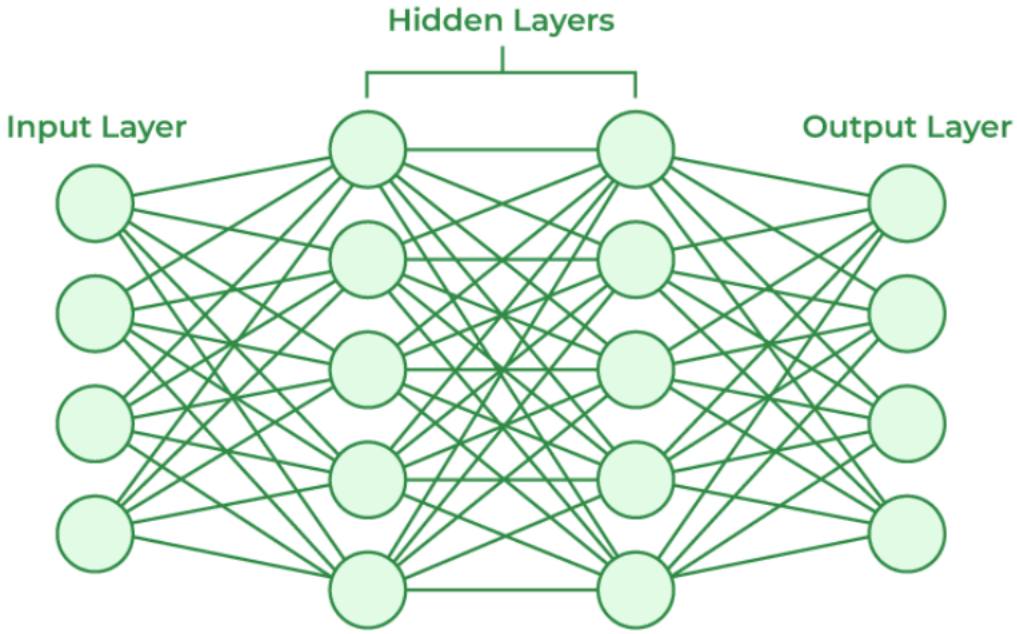
A. Supervised Machine Learning Algorithms

Supervised machine learning algorithms learns from a set of pre-labelled data, with the possible outputs for the corresponding spends have been already been given. This algorithm learns gradually using the labelled data provided and eventually builds up its own probabilistic mapping system to use for new inputs. This technique has two different subtypes called, Regression and classification. This technique is mostly used to generate the outputs which are in categorical nature. In this context; Spam and Ham as the two categories.

SUPERVISED LEARNING BASED MACHINE LEARNING ALGORITHMS

Artificial Neutral Network (ANN)

ANN is built using artificial neutrons, Hence the name come from. The number of artificial neutrons that are been used in the system can be varied and depend on the requirements of the system. These neurons are connected to different layers such as Input layer, Hidden layers and output layers. ANN systems 'learns' through a process named, 'Back-Propagation'. The produced new output of the network is compared and matched with the ideal match that should have been produced. The variation is taken into account and adjust the weights between the neutron connections with many iterations .



Naïve Based Machine Learning Algorithm

This is one of the commonly used supervised machine learning algorithm. This has been developed using the Bayes' rule which tries to derive the probability of an event occurrence based on even related prior knowledge and conditions . This approach is highly scalable, fast and easy to implement into a system. Na¨ ıve Based algorithm treats the features as independent from each other.

Decision tree (DT)

Decision tree machine learning algorithm is another algorithm that have been used more commonly in the reviewed supervised learning approach studies. The reasons to use this more often are this is an algorithm that can be used easily, easier explanations and visualizations. This can be used with both large and small data sets.

Unsupervised Machine Learning Algorithms

As the name describes, in this technique there are no labelled data or explicit instructions to pre trained the designed model. Therefore, these systems are not provided with a training. In this algorithm the analysis is carried out based on the dataset and feature out the common characteristics, structures and features in a group. Then rearrange the output data in different based structure or the pattern [10]. The output data can be organized in different types such as clustering, anomaly detection, association.

UNSUPERVISED LEARNING BASED MACHINE LEARNING ALGORITHMS

K-nearest Neighbour machine learning algorithm (KNN)

This algorithm is effective to use when there is noise in the input dataset. This can be used to generate both classifications and regression outputs for the developed system. The main drawback of this algorithm is it is highly sensitive for the outliers in the data set. Apart from that, computational cost for this algorithm is comparatively higher with regard to other machine learning algorithms.

K- means Clustering machine learning algorithm

This algorithm has straightforward implementation mechanism and the computational cost is comparatively lower than KNN ML algorithm. These are the reasons for this algorithm to be one of the commonly used unsupervised machine learning algorithm in spam classification field. In the K means clustering the data mining process initiates with the first group which is selected randomly. There is a randomly selected centroid for each cluster to begin the process. Repetitive calculations are carried out starting from that centroid to generate the optimized position.