International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 4, Issue 7, July-2018

Prediction of Promotion and Position of Employee in Company and Analysis of Machine Learning Algorithms

Neelkumar N. Patel[1], Professor Vatsal H. Shah[2]

[1,2]*Information Technology Department & Birla Vishvakarma Mahavidhyalaya*

Abstract— ***This research paper demonstrates the application of Machine Learning in real life by predicting the promotion and position of the employee in the company based on the passed data. It also explains the types of Machine Learning and their algorithm. Difference between Data Mining, Artificial Intelligence, Machine Learning and Deep Learning. Three Machine Learning Algorithm is analysed in the paper. They are Decision Tree Machine Learning Algorithm, Naïve Bayes Classifier Machine Learning Algorithm, K-Means Clustering Machine Learning Algorithm.***

Keywords— Prediction, Machine Learning, Artificial Intelligence, Decision Tree ML Algorithm, Naïve Bayes ML Algorithm, K-means Clustering ML Algorithm.

INTRODUCTION

With increasing usage of internet and computer in the 21st century, there has been great expansion of the data available in the raw format. To effectively utilize the data and produce effective results and outcomes of the data several computer science algorithms have been developed so far. For effective analysis of data several algorithms are developed which can analyse the large publicly available data as well as draw some output or predict certain outcomes on the several inputs. From buying commodities from internet to application of university is taking place over the internet. Hence there is large cluster of raw data over the internet. To effectively visualize and analyse the data Machine Learning Algorithms are used. What is Machine Learning?

1.1 Machine Learning

"Machine Learning is developing and optimizing an algorithm which can analyse past data and predict the outcome for forthcoming input by learning and training its algorithm and experience."

DATA is the centre of the machine learning algorithm and main part of it. This data fetched to the machine learning algorithm has some kind of hidden pattern in it. Machine learning algorithms find this pattern in the raw data and analyse it and improves the performance of the algorithm. Statistics play an important role in machine learning algorithm. Also, many of the machine learning algorithm make use of statistics in algorithm. Basic concept of function of the machine learning algorithm is statistical method applied while processing the dataset. There are two methods further in the statistical method as shown below:

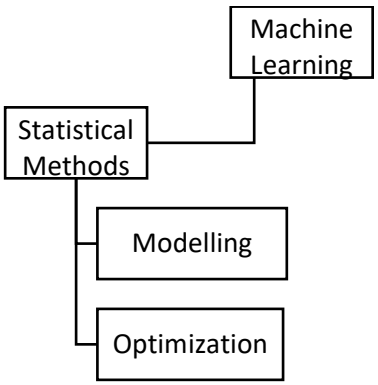


Figure 1.1 Statistical Methods

Statistical Modelling – Statistical Modelling is a process of manipulating the raw data mathematically and analysing the obtained output and to predict the outcomes. Generally, the prediction is the optional part in statistical modelling.

Machine learning algorithms are statistical modelling methods which analyse the data mathematically and predict the outcomes of the input dataset.

Optimization – Optimization means extracting meaningful information from the data and optimizing the data to predict the outcomes. Using this optimized data and knowledge to predict the similar type of data. It provides techniques for modelling the data.

Machine learning algorithm extract information from data without human guidance. Study, design, and development of algorithm giving computers capability to learn without being explicitly programmed. There are several applications in which this type of Machine Learning Algorithms is applied. Artificial Intelligence and Machine Learning are already being use in Image Recognition, Prediction, Extraction of Data, Neural Networks, Data Mining and Data Warehousing etc.

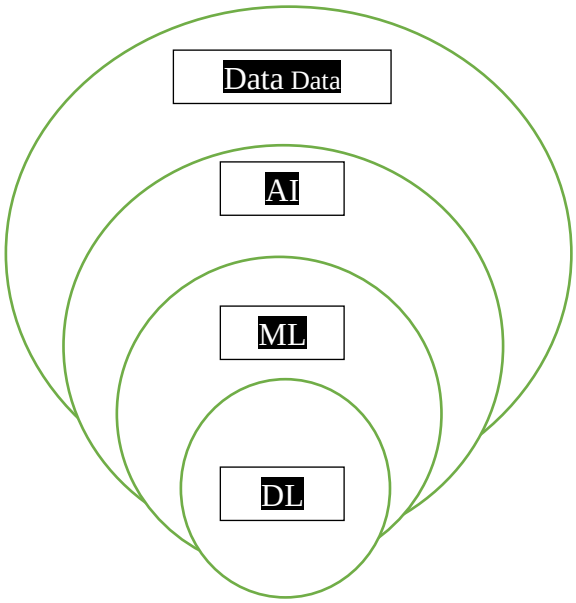


Figure 1.2 Chart of Data Mining, AI, ML, Deep Learning

As shown in the figure that the Machine Learning is a part of Artificial Intelligence and shows the comparison between the Data Mining, Artificial Intelligence, Machine Learning and Deep Learning. Data Mining is general field in which Artificial Intelligence, Machine Learning Algorithms and Deep Learning is applied to manipulate the data and analyse the data. Artificial is broad term which make computer to think like human brain. To do so, Artificial Intelligence make use of Machine Learning Algorithms and Artificial Neural Networks in Deep Learning. Machine Learning make use of the Algorithms to make decision or to predict the outcomes on the other hand Deep Learning uses Artificial Neural Network to make human brain like decision. Artificial Neural Network are human brain like neuron network which works to take decision without human guidance.

There are limitations to the Machine Learning. Like when more large and Complex dataset is provided for the analysis of the data, the errors increases in the outcomes provided by the machine learning algorithm. Such large datasets are solved easily by the Deep Learning using Artificial Neural Network.

1.2 Types of Machine Learning Methods

There are three types of Machine Learning Methods

1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning

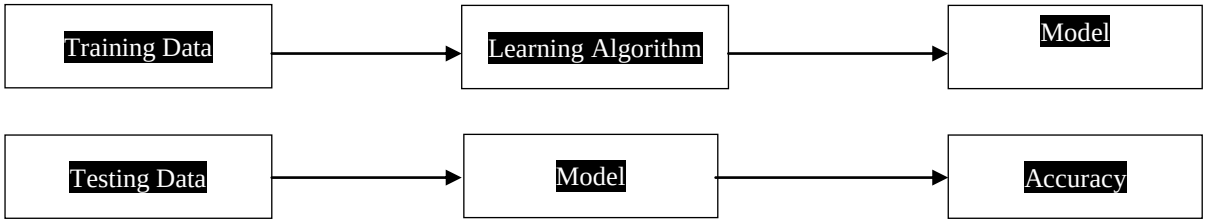
1.2.1 Supervised Learning

In Supervised learning the dataset along with the specific output of the input is provided to the machine learning algorithm. According to the dataset provided to the algorithm, it predicts the outcomes of similar type of the data by analysed data. Function in the supervised learning is trained with the input variables say X (input variable) and the output variable say Y (output variable). The algorithm is trained with the training data and a model is prepared from that data.

After that, the model is used to predict the outcomes of the similar type of the data. The input data is divided into two sets of data:

1. Training Data
2. Testing Data

Training data is feed to the machine learning algorithm and a model is prepared. After the model is tested with the testing data and accuracy of the model is calculated.



Example of Supervised Learning:

Spam filtering is type of the supervised learning. Some emails labelled as spam are feed to the algorithm and the algorithm is trained with the data according to the dataset. Incoming unlabelled email are then analysed by the algorithm and labelled as spam mail or unspam mail. Thus, the algorithm maps the input into several classification of the data.

1.2.2 Unsupervised Learning

In unsupervised learning the dataset only contains the input parameters and output of the dataset is not present. Thus, dataset is neither classified nor labelled. In this type of learning the algorithm owns itself for analysing the dataset a predicting the outcomes. The unsupervised algorithm predicts the outcomes by finding the hidden patterns in the dataset and analysing it. Dataset contains repetitive patterns in it. Unsupervised learning algorithm finds such repetitive hiding patterns in the dataset and forms several classes or cluster according to the pattern of the data. This type of approach is also known as density estimation approach. Because, the cluster are formed according to the statistical property of the data.

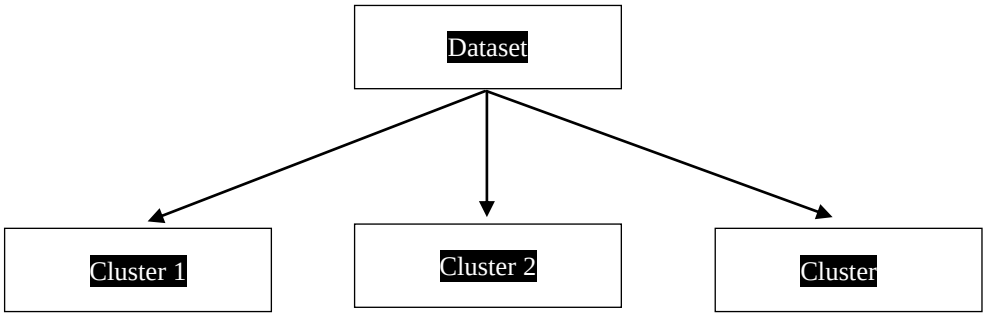


Figure 1.4 Unsupervised Learning

Example of Unsupervised Learning:

Classification of the spoon in the kitchen is unsupervised learning. The cluster of the spoon is given as an input to the algorithm. The algorithm classifies the input according to the feature of the spoon in the dataset and produces class or cluster of similar type of spoon.

1.2.3 Reinforcement Learning

Like unsupervised learning dataset in reinforcement learning only contains input parameters and output of the dataset in not present. But, unlike unsupervised learning, the data generated as output and predicted there is not assured that it is correct or incorrect. The goal is not set in reinforcement learning like supervised and unsupervised learning. The machine learning algorithm continuously learn and improve the efficiency by analysing its own generated outcomes and data. It produces effective outcomes at each iteration of applying algorithm.

Example of Reinforcement Learning:

Self-Driving cars are the example of the reinforcement learning. The navigation system and algorithm continuously learn and improves the effectiveness of the cars.

PREDICTION AND ANALYSIS OF MACHINE LEARNING ALGORITHM

There are three ways in which the learning methods can be applied

1. Supervised Learning Methods
2. Unsupervised Learning Methods
3. Reinforcement Learning Methods

In the same the machine learning algorithms are divided according to this type of methods

1. Supervised Learning Algorithm
2. Unsupervised Learning Algorithm
3. Semi-Supervised Learning Algorithm

Here we are going to analysis following type of machine learning algorithms which are divided as supervised and unsupervised learning algorithm

1. Decision Tree Machine Learning Algorithm
2. Naïve Bayes Classifier Machine Learning Algorithm
3. K – Means Clustering Machine Learning Algorithm

1. Decision Tree Machine Learning Algorithm

It is a type of supervised learning algorithm. The main object of the decision tree machine learning algorithm is the decision trees. Such trees are known as classification tree or regression trees. As the name suggest that the decision tree is used to visually represent the decision making. It is one of the most important algorithms for prediction of the data and outcomes. Tree is a type of data structure. Algorithm uses recursive partitioning approach to produce a decision tree in the decision tree machine learning algorithm. It produces the tree which partition the logical connections and partition the data according to the tree and predicts the outcomes according to the obtained decision tree. It is graphical based machine learning algorithm. The decision is made by computing the several mathematical instructions on the data. Each branch in the decision tree represents the outcome of the data in partial form and final outcome is obtained from the bottom most branch.

A decision tree begins with the root node at the top of the tree and further the tree splits into the internal node based on the gain. To calculate the gain, the entropy of class and the entropy of attribute is to be calculated.

Entropy of class = $\left|\frac{-P}{P+N}\log_2\left(\frac{P}{P+N}\right) - \frac{N}{P+N}\log_2\left(\frac{N}{P+N}\right)\right.$

where, P = Number of positive value

N = Number of negative value

For each of the attribute calculate the entropy and then calculate the gain. The attribute with the highest gain is the root node of the decision tree and further based on the gain value of the attributes the node is added to the tree.

Entropy of attribute = $\frac{\sum_{i=0}^{n} Pi+Ni}{P+N}$ (I(Pi,Ni))

Gain is calculated as follows:

Gain = Entropy class – entropy attribute

A Decision Tree Machine Learning Algorithm to the predict the promotion of an employee in a certain company using past data. Dataset used in this algorithm is as follows:

Here type is the type of work employee does at work and points gained while working.

Age	Type	Points > 50	Promotion
Old	Software	Yes	No
Old	Hardware	No	No
Mid	Software	No	Yes
Mid	Hardware	Yes	Yes
New	Software	Yes	Yes
New	Hardware	No	No

Table 2.1 Decision Tree Training Data

First of all, Entropy of the Class is found out by the formula as shown above.

Entropy of class = 1

For each attribute the gain is calculated by calculating the entropy gain for each attribute.

We get gain of each attribute as follows:

For Age, gain = 0.6667

For Type, gain = 0.2803

For Points, gain = 0.5407

Thus, gain of age attribute is highest. So, the root node is age.

Again, for the internal node the entropy is calculated and internal node is formed by repeating the same procedure. At last we get a decision tree as follows:

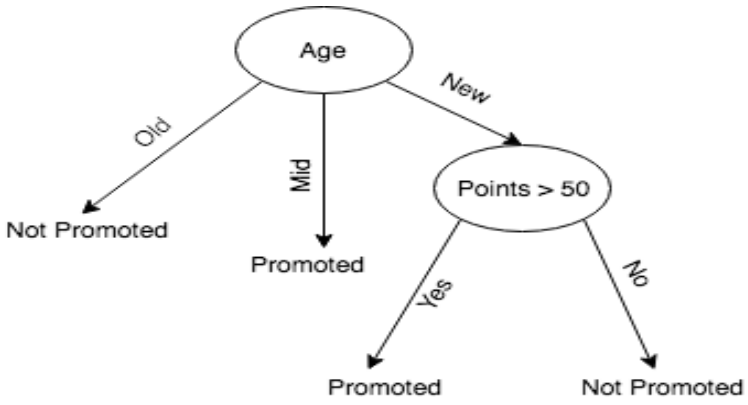


Figure 2.1 Decision Tree

Now, whenever new data comes like this [New, Software, Yes] then the algorithm from tree will predict that the employee is to be promoted.

2. Naïve Bayes Classifier Machine Learning Algorithm

Naïve Bayes is a classifies machine learning algorithm. The algorithm classifies the input dataset into several classes based on the function of the machine learning algorithm. The algorithm generally acts as a classifier. Naïve Bayes Classifier Learning Algorithm works on the basic principle of Bayes theorem of probability. It uses probability to predict the classes of the input dataset. For each different feature in the input dataset it generates a new class. According to the similarity of the features in the input data, the data is separated in the classes and outcomes is predicted for the upcoming input data. This algorithm can work on large dataset at once.

Naïve Bayes Classifier Machine Learning Algorithm is effective when output is categories of similar input data. That is when the problem is of classification of data type it outperforms the other type of Machine Learning Algorithms.

The Bayes theorem of probability can be expressed as follows:

$$P\left(x/y\right) = \frac{P\left(y/x\right) * P(x)}{P(y)}$$

Here, P(y/x) is known as likelihood which is probability of a given class. P(x/y) is probability of x with given probability of y.

Let us apply Naïve Bayes Classifier Machine Learning Algorithm to find the position of an employee in the industry or company.

The Employee dataset is as follows:

Employee id	Employee Name	Gender	Salary	Position
1	Ramesh	Male	80,000	Senior
2	Suresh	Male	40,000	Junior
3	Jack	Male	55,000	Senior
4	Jill	Female	60,000	Senior
5	Hitesh	Male	40,000	Junior
6	Yagnesh	Male	90,000	Senior
7	Rita	Female	65,000	Senior
8	Seema	Female	30,000	Junior

Table 2.2 Naïve Bayes Classifier Machine Learning Training Data

Probability of Position:

P(Senior) = 5/8

P(Junior) = 3/8

Range of Salaries: [0 - 30000]

[30,001 – 40,000]

[40,001 - 50,000]

[50,001 – 60,000]

[60,001 – 70,000]

[70,001 – 80,000]

[80,000 – 90,000]

Now find the probability of each attribute. Here, Gender and Salary range

Attribute	Value		Probability	
Gender	Senior	Junior	Senior	Junior
Male	3	2	3/5	2/3
Female	2	1	2/5	1/3
Salary				
0-30000	0	1	0	1/3
30,001 – 40,000	0	2	0	2/3
40,001 - 50,000	0	0	0	0
50,001 – 60,000	2	0	2/5	0
60,001 – 70,000	1	0	1/5	0
70,001 – 80,000	1	0	1/5	0
80,000 – 90,000	1	0	1/5	0

Table 2.3 Naïve Bayes Probability Table

We input a new data like a= [9, Shashikant, Male, 75,000]

P(a/senior) = 3/25

P(a/junior) = 0

Likelihood for senior = P(a/senior) * P(senior) = 3/40

Likelihood for junior = P(a/junior) * P(junior) = 0

Applying Bayes theorem of probability

P(junior/a) = 0

P(senior/a) = 1

Thus, Shashikant belongs to Senior position class.

3. K – Means Clustering Machine Learning Algorithm

As the names suggest that this machine learning algorithm is used for the cluster analysis. Also, this machine learning algorithm is unsupervised learning algorithm. The dataset operated by this machine learning algorithm is a cluster which is divided into k number of cluster. The value of k is the number of cluster to be formed from input cluster of data. Clustering is formation of cluster or classes. Statistical mean is used in this machine learning algorithm. The outcome is predicted using the clusters and classification of the structure of data. It is non-deterministic and iterative machine learning algorithm. The outcome of the k-means clustering machine learning algorithm is k number of cluster and classes. The clusters are made by feature of the data. Data having similar features are grouped into same clusters.

There are two types of cluster

1. Central
2. Hierarchy

Central cluster is a type of centroid like cluster. The classes are formed based on categories and cluster are formed as centroids. In hierarchy cluster the cluster is further divided into hierarchy.

Example of Clustering – E-Commerce website like amazon.com uses k-means clustering machine learning algorithm to classify it products into categories. The cluster or classes are formed on the basis of structure of products like electronics, clothes, hardware etc. The incoming product is classified and automatically added to its cluster and categories.

Steps for performing k-means clustering machine learning algorithm:

1. For given dataset of input choose k number of random number and assign them as mean. There will be k number of means.
2. Find the nearest number to the mean and again form k number of clusters.
3. Repeat step 1 and 2 until we get same mean for previous step and current step.

Let the input to the algorithm be data as follows:

{11,12,13,16,17,21,26,27,29,30} and k = 2

m1 = 14 and M2 = 22

k1 = {11,12,13,16,17} and k2 = {21,26,27,29,30}

m1 = 14 and m2 = 27

k1 = {11,12,13,16,17} and k2 = {21,26,27,29,30}

m1 = 14 and m2 = 27

The cluster formed are {11,12,13,16,17} and {21,26,27,29,30}

CONCLUSION

Thus, Machine Learning Algorithms are useful in predicting future result with different type of dataset. However, there are some errors in predicting the outcome. Further, for several dataset the algorithm to be used is different i.e. for supervised learning we can use Decision tree machine learning algorithm and for unsupervised learning we can use K-means clustering machine learning algorithm. The aim of this thesis is to get familiar with the machine learning algorithms for supervised and unsupervised learning. Also, to determine the type of algorithm to be used for particular dataset.

REFERENCES

[1] Introduction to Machine Learning (https://www.digitalocean.com/community/tutorials/an-introduction-to-machine-learning

[2] Introduction to Machine Learning, Alex Smola and S. V. N. Vishwanathan, Cambridge University Press.

What is Machine Learning? (https://www.expertsystem.com/machine-learning-definition/)

Machine Learning, Tom Mitchell, McGraw Hill Publication

Introduction to Machine Learning, Ethem Alpaydin, The MIT Press