MACHINE LEARNING ALGORITHMS

Abstract

Now-a-days, everyone is familiar with the term “data” and it is everywhere. But, this is huge in size and may be generated by people or devices. The problem with data is that, it could be in different forms like text, audio, video, and image etc., Due to this the data can be categorized as structured or unstructured. Analyzing and producing results out of this unstructured data is a time-consuming process. However, it would be easy to derive output from unbalanced data if it could be converted into balanced data. Here comes the role of *Machine Learning*, which is a subset of *Artificial Intelligence* (AI) that enables machines or other systems to learn on their own without any kind of explicit programming. These systems are designed in such a way that, they use knowledge to extract information from the unbalanced data. To deal with these data problems, various techniques have been supported by machine learning. For instance, to develop decision–making insights, many data-intensive problems require implementation of *regression* or *classification* techniques. This falls within the area of machine learning. Machine learning algorithms can be categorized as supervised, unsupervised and reinforcement learning strategies based on the desired outcome of the algorithm. Examples of various Machine learning algorithms include Linear Regression, Logistic regression, k-nearest neighbors, k-means, Naïve Bayes, Support Vector Machine (SVM), Random forest, Decision tree, Dimensionality reduction, Gradient boosting and Ada Boosting algorithm etc., could be applied on data for future predictions.

Keywords: Artificial Intelligence, Machine learning, Regression, Classification, Support Vector Machine.

Authors

Rajani Rajalingam
Computer Science and Engineering
Ravindra College of Engineering for Women, Kurnool, Andhra Pradesh
India
rajaninec@gmail.com

Dr. Madhusudhana Reddy Barusu
Electronics and Communication Engineering
Ravindra College of Engineering for Women, Kurnool, Andhra Pradesh
India
Madhu.barusuau@gmail.com

G. Prathibha Priyadarshini
Computer Science and Engineering
Ravindra College of Engineering for Women, Kurnool, Andhra Pradesh
India

Pulagouni Priyanka
Computer Science and Engineering
Ravindra College of Engineering for Women, Kurnool, Andhra Pradesh
India

I. INTRODUCTION

Machine learning is primarily used for **"prediction".** It uses its knowledge to predict the future output based on its previous history. As per human psychology, we are very anxious to know future but we can't, right? This is the reason, why we are depending on machines and that is the essence of machine learning. The efficiency of prediction depends on two factors: Accuracy and Speed.

There are different machine learning strategies and the following figure depicts it:

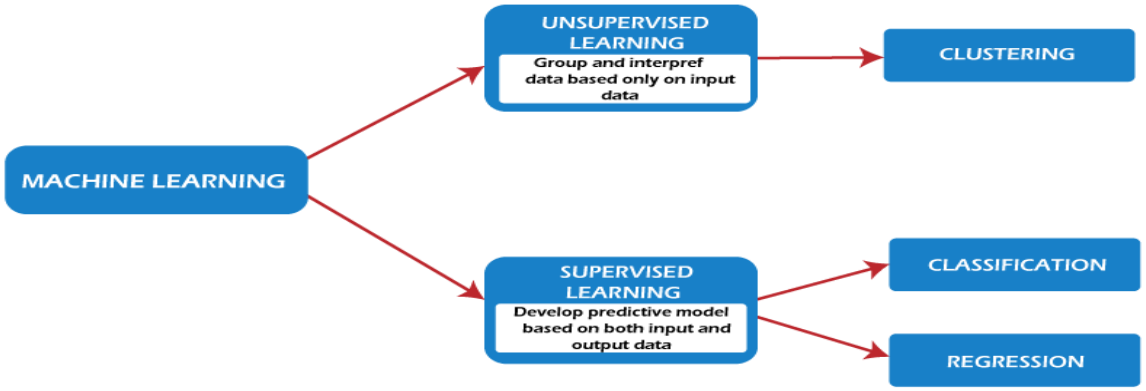

Figure 1: Machine Learning Strategies

II. MACHINE LEARNING STRATEGIES

Mainly, machine learning follows 2 strategies, i.e., Supervised and Unsupervised Learning.

1. **Supervised learning:** It's a type of learning mechanism. In this method, there will be a trainer who trains the model with known data (labeled data). The machine learns things from trained data and this knowledge is applied once the test data is given as input to it. Finally, it tries to produce the desired output.

   This learning strategy uses techniques like classification and regression to develop machine learning models.

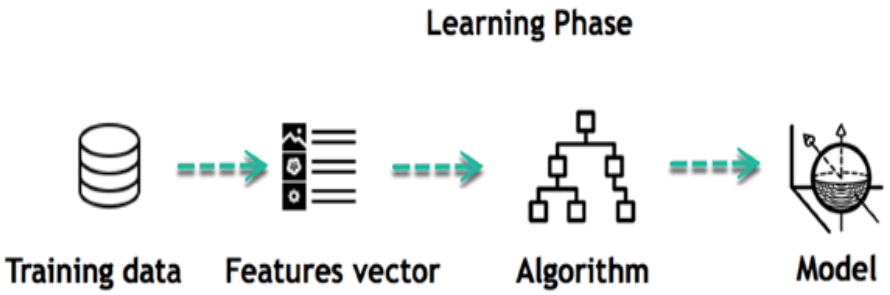

Figure 2: Supervised Machine Learning

2. **Unsupervised learning:** In this method, the model predicts the output without being given any training. There is no trained data and there exists only test data. The model

doesn't depend on any trainer. Through observation it learns on its own and tries to predict the expected outcome.

One of the common unsupervised learning techniques is clustering. Few applications of clustering technique include gene sequence analysis, market research, and commodity identification.

III. MACHINE LEARNING STRATEGIES

To implement machine learning, various algorithms are utilized such as linear regression, k-Nearest Neighbors, Bayesian algorithm etc. The basics of these algorithms are explained below.

Regression

- It falls under the category of supervised learning technique. This technique is used to uncover the correlation that exists among variables.
- In this method, the output that has to be produced must be a real or continuous variable value.
- It assumes independent and dependent variables as two different variables.
- The independent variable can be considered as an Input variable and is shown on X-axis.
- The dependent variable can be considered as an output variable and is shown on Y-axis.
- This technique shows the ***relationship*** between independent and dependent variables. This relationship is linear in nature and hence is named as ***linear regression***.
- In turn, linear regression may be classified as ***simple*** and ***multiple linear regression***.
  - *If there is a single input variable, then, it is* ***simple linear regression,*** *And*
  - *If there is more than one input variable, then, it is* ***multiple linear regression*****.**
  - Yet, in both kinds of regression, there will be only one dependent variable.

1. **Linear regression:** This model produces a slanted straight line that describes the relationship that exists between I/O variables.

Let's consider an example:

Say we want to estimate the salary of an employee based on years of experience. Here, we can assume that, ***years of experience*** as an ***input variable***, and the ***salary*** of an employee as an ***output variable***. Using this perception, we can predict the salary of an employee based on present & past data.

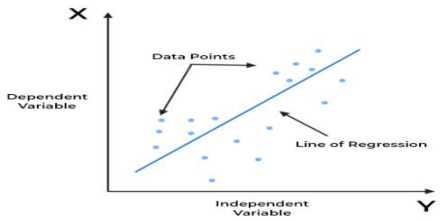

Figure 3: Predictions based on Linear Regression

2. k-Nearest Neighbors' algorithm

- It is a kind of ***supervised machine learning*** algorithm.
- It can be used to solve problems like classification, regression etc.
- Particularly, it concentrates on classification problems.
- This method is used for approximating the probability of a data point will become a member of one group or another group based on the nearest group as it belongs to.
- It simply stores the trained data and do not perform any computations.
- It doesn't construct a model unless the query is raised on the given dataset.
- Due to this reason, it's called as a *lazy learning* and *non-parametric* algorithm.
- In this method, K is a positive integer value.
- It is always recommended to choose an ***odd value for k*** for best ***accuracy***.

We will see an ***example:***

Consider**,** there are cats and dogs as two separate groups of objects and we trained the model with these objects during training time. During testing period, to the model the cat is supplied as test data and also the value of K as 5. Since there are **four cats** and just **one dog** in the vicinity of the 5 nearest neighbors, based on the vicinity of the 5 nearest neighbors in the red circle's boundaries, the algorithm would predict that the test data is a cat.

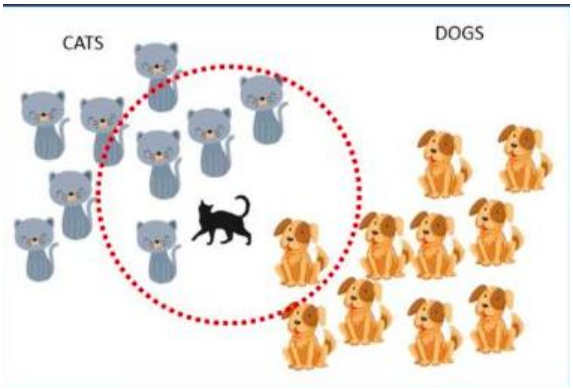

Figure 4: Finding Nearest Neighbor

Selecting the correct value of K : Here, we will try to test the accuracy of the model, for K different values. The value of K that produces the top accuracy for both training and testing data can be selected. But, there is no specific technique for determining the value of K.

3. k-Means Algorithm: Significant features of k-means mechanism are as follows:

- It is a type of unsupervised leaning method.
- It works very well on unlabeled, numerical data.
- It supports Hierarchical clustering technique in which, it operates at faster rate even if the size of dataset is large.
- It is very smooth in terms of interpretation and resolution.
- Even a single instance can modify the cluster while re-determining the cluster Centre.
- It tries to improvise dense clusters.
- When datasets are well distinctive, it tries to yield best output with high speed.
- One good feature in this technique is, it is robust and uncomplicated to understand.

Figure 5: Finding Clusters

This algorithm mainly performs two tasks:

- It follows an iterative process, and determines the best value for K center points or centroids.
- Then assigns each data point to its closest k-center.
- The data points which are nearer to the particular k-center will then generate a cluster.
- We can apply this algorithm in areas like: *Market segmentation, Document Clustering, Image segmentation, Image compression, Customer segmentation* and *Analyzing the trend on dynamic data* etc.,

4. **Bayesian methods:** The most popular Bayesian methods are:

- Naive Bayes
- Gaussian Naive Bayes
- Multinomial Naive Bayes
- Averaged One-Dependence Estimators (AODE)
- Bayesian Belief Network (BBN)
- Bayesian Network (BN)

Explicitly, these methods depend on Bayes' theorem for problems of different categories such as classification and regression.

Naive bayes classification

- It's a probabilistic ML algorithm that can be used in a wide variety of classification tasks.
- It produces the conditional probability of an event B, for a given event A.

This specific algorithm can be applied in filtering spam messages, documents classification and prediction of sentiments etc.

For example, using a Naive Bayes classifier, it's possible to predict whether a person will purchase a product on attributes like date, discount, and at free delivery.

Figure 6: Prediction on purchase of a product

As this algorithm is a combination of two words Naïve and Bayes, which can be described as:

- It assumes that, the incidence of a definite feature is independent of the incidence of some other features.
- For Example, if the fruit is identified based on its color, shape, and taste, and then red, spherical, and sweet fruit is recognized as an apple. Hence each attribute individually contributes to find that it is an apple without depending on other attributes.
- As the algorithm depends on Bayes' theorem, it's called Bayes.

5. **Support vector machine mechanism:** One of the most popular Supervised Learning algorithms, which is used for classification as well as Regression problems is named as SVM. Yet, in specific, this can be widely used in classification problems, only.

Example:

Consider the diagram, suppose we saw a cat which resembles few features of a dog. As it looks to be strange, we would like to predict whether it is really a cat or dog. Hence, in this case, we must construct a model that can perform correct prediction. We have to build such a model using SVM.

Figure 7: SVM classification

First, we may have to train the model by providing input of cats and dogs images with different features and secondly, we test it with this strange cat.

SVM tries to generate a decision boundary between these two objects (cat and dog) and chooses extreme cases of support vectors; it will observe the extreme features of cat and dog. Now, it will classify it as a cat on the basis of the support vectors.

We can use SVMs in applications like recognition of handwriting, detecting intruders, detection of face, email messages classification, gene finding, etc.

TEXT BOOK REFERENCES

[1] A Concise Introduction to Machine Learning, Anitha C. Faul, CRC Press, 2020
[2] An Introduction to Machine Learning Springer International Publishing Gopinath Rebala, Ajay
[3] Ravi, Sanjay Churiwala, 2019.
[4] A Brief Introduction to Machine Learning for Engineers Now Publishers Osvaldo Simeone,2018
[5] E. Alpaydin "Introduction to Machine Learning", third Edition, MIT Press, 2014
[6] An Introduction to Machine Learning Springer International Publishing Miroslav Kubat (auth.), 2017
[7] An introduction to machine learning Interpretability, O'Reilly, Patrick Hall and Navadeep Gill,2018
[8] A brief introduction to machine learning for engineers, kings college London, Osvaldo Simeone,2018
[9] An introduction to machine learning, Springer, Kubat, Miroslav, 2015

ONLINE RESOURCES

[1] http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/091117.pdf
[2] https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/index.html
[3] https://alex.smola.org/drafts/thebook.pdf
[4] https://seat.massey.ac.nz/personal/s.r.marsland/MLBook.html

