Machine Learning Algorithms

|  |  |  |  |
| --- | --- | --- | --- |
| Rajani Rajalingam,Department of Computer Science and EngineeringRavindra College of Engineering for Women, Kurnool, Andhra Pradesh, Indiarajaninec@gmail.com | Madhusudhana Reddy Barusu, Department of Electronics and Communication Engineering,Ravindra College of Engineering for Women, Kurnool, Andhra Pradesh, India | Pulagouni PriyankaDepartment of Computer Science and Engineering ( Artificial Intelligence)Ravindra College of Engineering for Women, Kurnool, Andhra Pradesh, India | G. Prathibha Priyadarshini,Department of Computer Science and EngineeringRavindra College of Engineering for Women, Kurnool, Andhra Pradesh, India |

**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*, a subset of *Artificial Intelligence* (AI) which 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 work on this kind of 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 machine learning realm. 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.

**I. INTRODUCTION**

Machine learning primarily used for **“prediction”.** It uses its knowledge to predict the future based on the past history. As per human psychology, we are very anxious to know future but we can’t, right? This is the reason, where 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:



Fig 1: Machine Learning Strategies

**II. MACHINE LEARNING STRATEGIES**

Mainly machine learning follows two strategies, i.e. supervised learning and unsupervised learning.

1. **Supervised Learning**

It’s a type of learning mechanism in which there will be a supervisor who trains the model with training data and once the test data is given as input, the model predicts the exact output. Supervised learning uses classification and regression techniques to develop machine learning models.

### How Supervised Machine Learning Works

Fig 2: Supervised Machine Learning

### Unsupervised Learning

In this mechanism, the model predicts the output without being given any training. There is no trained data whereas there exists only test data. The model doesn’t depend on any supervisor whereas through observation itself it tries to predict the desired outcome.

Clustering is a common unsupervised learning technique. Applications for cluster analysis include gene sequence analysis, market research, and commodity identification.

**III. MACHINE LEARNING ALGORITHMS**

To implement machine learning, the various algorithms are utilized such linear regression, k-nearest algorithm, 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 find the correlation among variables. A regression problem is when the output variable is a value of type real or continuous.
* 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 ***linear relationship*** between the independent and the dependent variables, and hence it can be named as ***linear regression****.*
* In turn, we have 2 types of linear regression such a*s* ***simple*** and ***multiple linear regression****.*
* *In* ***simple linear regression,*** *there will be a single input variable (x), And*
* *In* ***multiple linear regression*,** *there will be more than one input variable.*
* But, in both kinds of regression, there will be only one dependent variable.
1. **Linear Regression**

This model produces a sloped straight line that describes the relationship within the variables.

Let’s say we want to estimate the salary of an employee based on years of experience. Here, we can assume that, ***year of experience*** is an ***independent variable***, and the ***salary*** of an employee is a ***dependent variable***, as salary of an employee is dependent on the experience of an employee. Using this insight, we can predict the future salary of the employee based on current & past information.



 Fig 3: Predictions based on Linear Regression

1. **What is the k-nearest neighbors’ algorithm?**
* It is a type of [***supervised machine learning***](https://learn.g2.com/supervised-learning) algorithm, which is used to solve classification and regression problems. Particularly, used for classification problems.
* This method is used for estimating the likelihood that a data point will become a member of one group or another based on the nearest group it belongs to.
* It simply stores the trained data and do not perform any computations. Even, it doesn’t construct a model unless until 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 and it is recommended to choose an ***odd value*** for best ***accuracy***.



Fig 4: Finding nearest neighbor

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 test 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.

**Selecting the right value of K**

Here, for different K values, we will try to test the accuracy of the model. The value of K that delivers the best accuracy for both training and testing data is selected, yet, there is no specific technique for determining the value of K.

1. **k-Means Algorithm**

 key features of k-means clustering are as follows:

* It comes under the category of unsupervised leaning. Because, it works well on unlabeled, numerical data.
* It supports Hierarchical clustering technique in which, it operates at a 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 at faster rate.
* One good thing with this technique is, it is robust and uncomplicated to understand.

Fig 5: Finding Clusters

This algorithm mainly performs two tasks:

* Through an iterative process, it determines the best value for K center points or centroids.
* Assigns each data point to its closest k-center.
* These data points which are near to the particular k-center will create a cluster.

The areas where we can apply this algorithm include are: *Market segmentation, Document Clustering, Image segmentation, Image compression, Customer segmentation* and *Analyzing the trend on dynamic data* etc.,

### 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)

These methods explicitly apply Bayes’ theorem for problems of type classification and regression.

**Naive Bayes Classification**

* It’s a probabilistic machine learning algorithm that can be used in a wide variety of classification tasks.
* For a given event B, it produces the conditional probability of an event A.
* Applications include filtering spam, classifying documents and sentiment prediction etc.

For example, to predict whether a person will purchase a product on a particular combination of day, discount, and free delivery using a Naive Bayes classifier.

****

**Fig 6:**

Fig.6. Prediction on purchase of a product

**It’s called ‘Naive’ - because**

This algorithm is comprised of two words Naïve and Bayes, which can be described as:

* **Naïve**: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.
* **Bayes**: As it depends on Bayes’ theorem, so it’s called Bayes.
1. **Support Vector Machine Algorithm**

It is one of the most popular Supervised Learning algorithms, which is used for classification as well as Regression problems. However, it is used for Classification problems, particularly.



Fig 7: Accurate classification

**Example:**  Consider the diagram,suppose we saw a cat with strange features because which also has some features of dogs. So, if we want to predict whether it is cat or dog, we must choose a model which can perform correct prediction. We can build such a model using SVM. We will first train the model with lots of images of cats and dogs with different features and then we test it with this strange cat. SVM generates a decision boundary between these two objects (cat and dog) and choose extreme cases of support vectors, it will observe the extreme case of cat and dog. Now, it will classify it as a cat on the basis of the support vectors.

SVMs are used in applications like handwriting recognition, intrusion detection, face detection, email classification, gene classification, and in web pages etc.

**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

Ravi, Sanjay Churiwala, 2019.

1. A Brief Introduction to Machine Learning for Engineers Now Publishers Osvaldo Simeone,2018
2. E. Alpaydin “Introduction to Machine Learning”, third Edition, MIT Press, 2014
3. An Introduction to Machine Learning Springer International Publishing Miroslav Kubat (auth.), 2017
4. An introduction to machine learning Interpretability, O’Reilly, Patrick Hall and Navadeep Gill,2018
5. A brief introduction to machine learning for engineers, kings college London, Osvaldo Simeone,2018
6. 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>