Machine Learning Algorithms

ALGORITHM BEGINNER MACHINE LEARNING

This article was published as a part of the Data Science Blogathon.

Table of Contents

Introduction

We all know how Artificial Intelligence is leading nowadays. Machine Learning is a part of it. Artificial Intelligence is achieved by both Machine Learning and Deep Learning. There are three steps in the workflow of an AI project. They are Data collection, model training, and Deploying it. We use machine learning for models.

Types of Machine Learning Algorithms

First of all, we will start by learning types of Machine Learning Algorithms.

They are,

1. Supervised Learning

2. Unsupervised Learning

3. Reinforcement Learning

1. Supervised Learning: The data which is used in supervised learning is labeled data. Labeling is something known as categorizing. Using this labeled data machine learning model is trained and then with that model, we will predict the outcome of
untrained datasets.

2. Unsupervised Learning: The data which is used in unsupervised learning is unlabeled data. Unlabeled data is given to the machine learning model and is trained. Here model will form clusters according to similar features and characteristics and then clusters are formed. Now when untrained data is sent, the model will recognize it and predicts it to the corresponding clusters.

3. Reinforcement Learning: Here in reinforcement learning machine learning model is not provided with any of the data either it is labeled or unlabeled. Instead, here Machine tries with different actions and whenever the machine has done the correct
model then the reward signal is given. And in this way model is trained and predicts the outcome in the future with past experiences.

Supervised Learning is classified into Regression and Classification algorithms.

Regression algorithms are used whenever prediction is needed for continuous target variables. Like predicting salary, predicting age, stock market prediction, etc…For example linear regression, Multilinear regression, polynomial regression.

Whereas Classification algorithms are used for the prediction of discrete variables.Like for predicting either True or False, predicting Yes or No, predicting 0 or 1, predicting pass or fail, etc…. for example, logistic regression, Decision Tree, SVM, KNN.

Unsupervised Learning is completely based on clustering. The model will analyze a similar pattern among the input variables and forms cluster. For example, K means clustering, Hierarchical clustering.

Some common basic machine learning algorithms which are used:

1. Simple Linear Regression

2. Multilinear Regression

3. Logistic Regression

4. Decision Tree

5. SVM

6. KNN

7. K Means Clustering

Simple Linear Regression

Linear regression is a supervised learning model which is used to analyze continuous data. It is a data plot that graphs the linear relationship between independent and dependent variables.

Features are called independent variables and outcome or label is known as dependent variables which are dependent on features. The equation for linear regression is

$$y=a+bx+e$$

where,

y=dependent variable(outcome)

x=independent variable(feature)

a=intercept

b=slope

e=model error

Training model means finding slope and intercept. With that slope and intercepts model will predict y with a change in x.

linear regression is imported from sklearn.

Let's have look at the code.

from sklearn.linear_model import LinearRegression

We can see in the Image that 1st step is creating a model. Here Linear Regression model is created and then the model is trained by using the fit method. Prediction is done by using predict method. we can find the slope and intercept by using coef_ and intercept_ methods respectively.

Multi-Linear Regression

Multilinear regression is almost similar to simple linear regression except, here model takes multiple feature variables to predict the target variable. All the syntax and code are the same as simple linear regression. In simple terms, the model will predict one dependent variable with two or more than two independent variables.

The equation for a multilinear regression is,

$$y=b_0+b_1x_1+\ldots+b_nx_n+ e$$

Where,

y=Dependent variable

b_0=y-intercept

b_1x_1=Regression coefficient(b_1) of independent variable x_1

b_nx_n=Regression coefficient(b_n) of independent variable x_n

e=Model Error

Logistic Regression

The logistic regression model is a supervised learning model which is a generalization of a linear regression model, which is mainly used for categorical data. By the name regression in it, many used to think of it as a Regression algorithm but it is a classification algorithm.

Let's understand it in detail. Think of an example about grading either it can be pass or fail. And you will be provided with marks for each subject. And with those marks model finds a pattern to decide pass or fail. In our terms, let's say the pass percentage is 35%. Then all the candidates who get more than 35% will pass and the remaining will fail. Similarly, the model predicts pass or fail.

Logistic Regression is imported from sklearn.

from sklearn.linear_model import LogisticRegression

Let's have look at the code.

We can see in the Image that 1st step is creating a model. Here LogisticRegreesion model is created and then the model is trained by using the fit method. Prediction is done by using predict method. Finally predicted results are viewed.

Decision Tree

A decision tree is a supervised Machine learning technique. A decision tree algorithm is used for both regression and classification type problems.

Decision Node: When sub-node divides into sub-nodes, then it is called decision node.

Leaf/Terminal Node: Node with no children.

Pruning: The process of reducing the size of the decision tree by removing nodes.

Entropy: Entropy is the measure of the randomness of elements. It is the measure of uncertainty in the given set.

If entropy = 0, then the sample is completely homogeneous.

Information gain: Information gain measures how much “Information” a feature variable gives us about the class. Decision tree algorithms always try to maximize Information gain.

An attribute with the highest information gain will be split first.

Information gain = Entropy(parent) – (average weight) * Entropy(children)

For decision tree out of all features, which will be the root node, which will be the next decision node???

This will be decided by entropy. Attributes with the highest information gain will be split first.

Let’s have a look at the code.

We can see in the Image that 1st step is creating a model. Here DecisionTreeClassifier model is created and then the model is trained by using the fit method. Prediction is done by using predict method. And finally predicted results are viewed.

Support Vector Machine(SVM)

Support Vector Machine is a supervised Machine Learning algorithm. Support Vector Machine algorithm can be used for both Regression and Classification problems. But mostly SVM is used for classification problems. Here in SVM, we plot all the data points in a three-dimensional space. And then we have to find a hyperplane between categories to differentiate all the categories well.

Let’s Understand the Support Vector Machine algorithm in detail.

The main task of the SVM algorithm is to find the Right hyperplane between groups. For the given groups there will be many possible hyperplanes in between them. But which is right among all? Let’s find it.

Suppose Take 2 groups as stars and circles.

Here There are 3 hyperplanes namely A, B, and C. What do you think?? Which is the right hyperplane. Now you have to keep one main point in your mind. Hyperplane should segregate the groups very well. Here clearly B hyperplane Separates them in the best way.

Let's take Another Example.

Now have a look at this graph. What do you think is the best hyperplane?? Here all 3 hyperplanes segregate them well. In this case, we have to see a margin. Margin is the distance between the hyperplane to the nearest data point. B has maximum margin when compared to A and C. Hyperplane with the highest margin is the best hyperplane. Because the chances of getting the wrong classification will be less if the margin is more.

Let's have another example.

Now, which hyperplane will you decide on??? Hyperplane A has the highest margin and Hyperplane B segregates them well. So here Classification is our main motto. Hyperplane A has a classification error. whereas, B classifies well. So Hyperplane B is correct.

In this way, Hyperplane is decided.

Let's have look at the code.

We can see in the Image that 1st step is creating a model. Here SVC model is created using the SVM library and then the model is trained by using the fit method. Prediction is done by using predict method. We can find the accuracy of the model by using the accuracy_score method. Here we got 72% accuracy.

K Nearest neighbours (KNN)

K Nearest Neighbors(KNN) is a supervised Machine Learning algorithm that can be used for regression and classification type problems. KNN algorithm is used to predict data based on similarity measures from past data. One of the Industrial use cases of the KNN algorithm is recommendations in websites like amazon.

k= number of nearest neighbours.

Here we have to learn about something called Euclidean Distance. It is the distance between two data points which are Query and Trained data points. Here Query data point is a dependent variable which we have to find. The formula for Euclidean distance is,

$$d = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]}$$

Here,

(x_1,y_1) = Query data point

(x_2,y_2) = Trained data point

Let's take an example and understand it in deep. Here is the data which shows students Mathematics and Chemistry Marks and also label for this data is given which is either pass or fail. Now we have to find out if a student with Mathematics marks 5 and chemistry marks 6 will fail or pass.

Mathematics	Chemistry	label(pass/fail)
5	5	fail
7	8	pass
4	5	fail
3	5	fail
9	7	pass

Well, Here the query point$_{(x_1,y_1)}$ is (5,6).Find Euclidean distances to all the points.

d1 = $\sqrt{(5-5)^2+(5-6)^2} = 1$

d2 = $\sqrt{(7-5)^2+(8-6)^2} = 2.828$

d3 = $\sqrt{(4-5)^2+(5-6)^2} = 1.414$

d4 = $\sqrt{(3-5)^2+(5-6)^2} = 2.236$

d5 = $\sqrt{(9-5)^2+(7-6)^2} = 4.123$

Mathematics	Chemistry	label(pass/fail)	Euclidean distance	Ranking
5	5	fail	1	1
6	8	pass	2.828	4
4	5	fail	1.414	2
3	5	fail	2.236	3
9	7	pass	4.123	5

For n=5, k is 3 so we have to find 3 nearest neighbours which rank 1,2, and 3 . Results of these 3 neighbours are 3 fails. There are 3 fails and 0 passes. So with the majority failing, the student with mathematics marks 5 and chemistry marks 6 will fail.

This is how the KNN algorithm works.

Let’s have a look at the code.

We can see in the Image that 1st step is creating a model. Here KNeighborsClassifier model is created and then the model is trained by using the fit method. Prediction is done by using predict method. We can find the accuracy of the model by using the accuracy_score method. Here we got 98% accuracy.

K Means Clustering

K Means clustering is an unsupervised machine learning algorithm. Here there will be no labeled data. Data will be categorized into clusters. This algorithm is a centroid-based algorithm. Each group has a centroid. The motto of this algorithm is to minimize the distance between centroid and data points.

In the K Means algorithm, we find the best centroids by alternatively assigning random centroids to a dataset. And from the resulting clusters mean data points are selected to form new centroids. This process continues iteratively until the model is optimized.

For unsupervised learning datasets, there are no labels, only features are present. with the absence of labels, we have to identify which data points in the dataset are similar. A cluster is formed by a group of similar data points.

Here for all the random data points, two clusters are formed and random centroids were assigned. For the first iteration, clusters were like this with centroids.

After the second iteration, centroids were reassigned and clusters will be like this.

After the third iteration, again centroids were reassigned and finally, the model has optimized. Even if we iterate again, centroid points were not changing. And the final clusters will be like this.

In this way, the K Means clustering algorithm works.

Let’s have a look at the code.

We can see in the Image that 1st step is creating a model. Here KMeans model is created and then the model is trained by using the fit method. Prediction is done by using predict method. And then coming to visualization we can see all the data points are divided into 5 clusters with centroids.

Conclusion

As I said this article is for beginners and also those who need revision. Some basic Machine learning Algorithms are explained in this article in detail. Hope you guys have gained some knowledge through my article on Machine Learning Algorithms.

Read more blogs on machine learning algorithms on our website.

Connect with me on Linkedin: https://www.linkedin.com/in/amrutha-k-6335231a6vl/

Article Url - https://www.analyticsvidhya.com/blog/2022/01/machine-learning-algorithms/

Amrutha K