Certificate Course in Machine Learning using Python [6 Weeks]

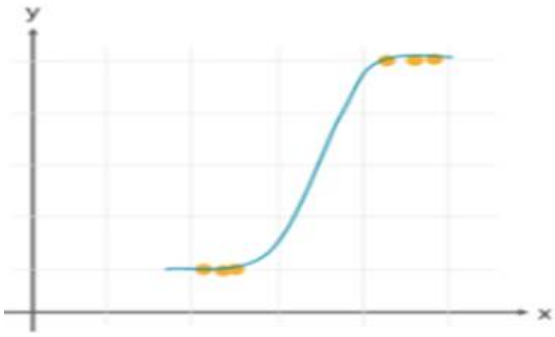
Mathematics behind Regression Algorithms

Attempt: 1

Machine Learning Algorithm

1. Logistic Regression

It is a classification algorithm in machine learning that uses one or more independent variables to determine an outcome. The outcome is measured with a dichotomous variable meaning it will have only two possible outcomes.



The goal of logistic regression is to find a best-fitting relationship between the dependent variable and a set of independent variables. It is better than other binary classification algorithms like nearest neighbor since it quantitatively explains the factors leading to classification.

Advantages and Disadvantages

- Logistic regression is specifically meant for classification.
- It is useful in understanding how a set of independent variables affect the outcome of the dependent variable.
- The main disadvantage of the logistic regression algorithm is that it only works when the predicted variable is binary.
- It assumes that the data is free of missing values and assumes that the predictors are independent of each other.

Use Cases

- Identifying risk factors for diseases
- Word classification
- Weather Prediction

- Voting Applications

2. Naive Bayes Classifier

It is a classification algorithm based on Bayes's theorem which gives an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Even if the features depend on each other, all of these properties contribute to the probability independently. Naive Bayes model is easy to make and is particularly useful for comparatively large data sets. Even with a simplistic approach, Naive Bayes is known to outperform most of the classification methods in machine learning. Following is the Bayes theorem to implement the Naive Bayes Theorem.

$$P(C_i|\, x_1, x_2 \ldots, x_n) = \frac{P(x_1, x_2 \ldots, x_n|C_i).P(C_i)}{P(x_1, x_2 \ldots, x_n)} \; for\; 1 < i < k$$

Advantages and Disadvantages

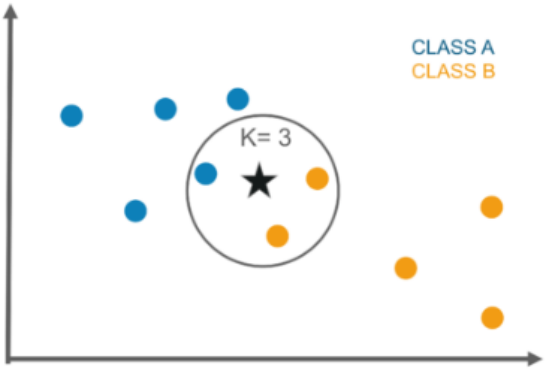
- The Naive Bayes classifier requires a small amount of training data to estimate the necessary parameters to get the results. They are extremely fast in nature compared to other classifiers.
- The only disadvantage is that they are known to be a bad estimator.

Use Cases

- Disease Predictions
- Document Classification
- Spam Filters
- Sentiment Analysis

K-Nearest Neighbor

It is a lazy learning algorithm that **stores all instances corresponding to training data in n-dimensional space**. It is a **lazy learning algorithm** as it does not focus on constructing a general internal model; instead, it works on storing instances of training data.



Classification is computed from a simple majority vote of the k nearest neighbors of each point. It is supervised and takes a bunch of labeled points and uses them to label other points. To label a new point, it looks at the labeled points closest to that new point also known as its nearest neighbors. It has those neighbors vote, so whichever label the most of the neighbors have is the label for the new point. The "k" is the number of neighbors it checks.

Advantages and Disadvantages

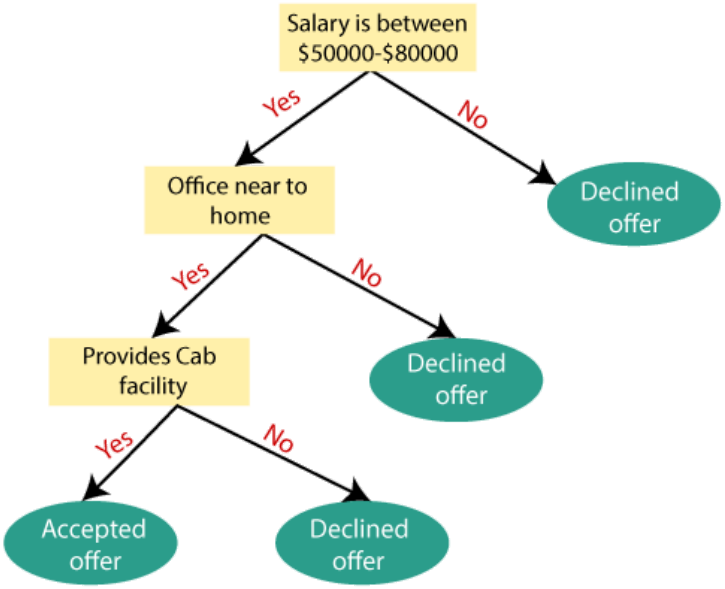
- This algorithm is quite simple in its implementation and is robust to noisy training data.
- Even if the training data is large, it is quite efficient.
- The only disadvantage with the KNN algorithm is that there is no need to determine the value of K and computation cost is pretty high compared to other algorithms.

Use Cases

- Industrial applications to look for similar tasks in comparison to others
- Handwriting detection applications
- Image recognition
- Video recognition
- Stock analysis

Decision Tree

The decision tree algorithm builds the classification model in the form of a **tree structure**. It utilizes the if-then rules which are equally exhaustive and mutually exclusive in classification. The process goes on with breaking down the data into smaller structures and eventually associating it with an incremental decision tree. The final structure looks like a tree with nodes and leaves. The **rules are learned sequentially** using the training data one at a time. Each time a rule is learned, the tuples covering the rules are removed. The process continues on the training set until the termination point is met.



- The tree is constructed in a top-down recursive divide and conquers approach.
- A decision node will have two or more branches and a leaf represents a classification or decision.
- The topmost node in the decision tree that corresponds to the best predictor is called the root node,
- A decision tree can handle both categorical and numerical data.

Advantages and Disadvantages

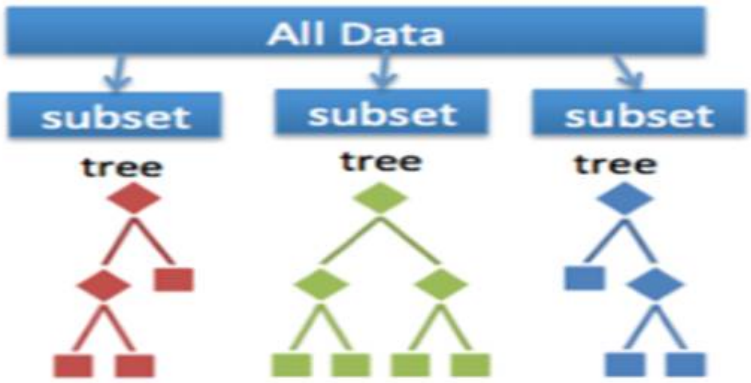
- A decision tree gives an advantage of simplicity to understand and visualize, it requires very little data preparation as well.
- The disadvantage that follows with the decision tree is that it can create complex trees that may not categorize efficiently.
- They can be quite unstable because even a simplistic change in the data can hinder the whole structure of the decision tree.

Use Cases

- Data exploration
- Pattern Recognition
- Option pricing in finances
- Identifying disease and risk threats

Random Forest

Random decision trees or random forest are an **ensemble learning method** for classification, regression, etc. It operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction (regression) of the individual trees.



A random forest is a meta-estimator that fits a number of trees on various subsamples of data sets and then uses an average to improve the accuracy in the model's predictive nature. The sub-sample size is always the same as that of the original input size but the samples are often drawn with replacements.

Advantages and Disadvantages

- The advantage of the random forest is that it is more accurate than the decision trees due to the reduction in the over-fitting.
- The only disadvantage with the random forest classifiers is that it is quite complex in implementation and gets pretty slow in real-time prediction.

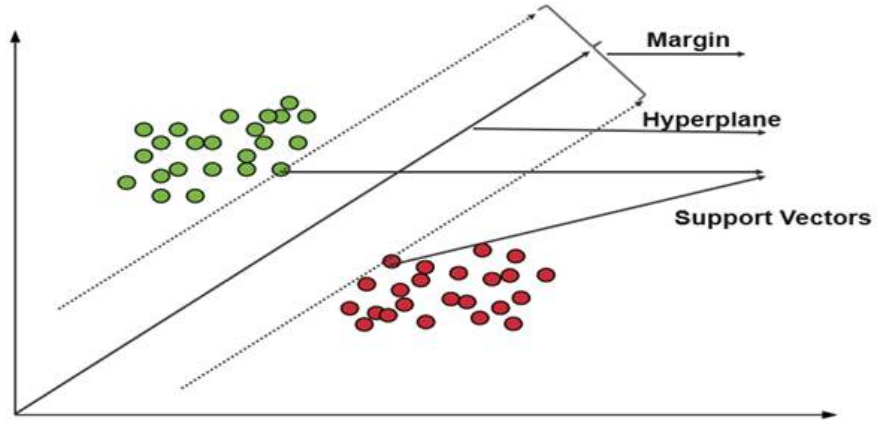
Use Cases

- Industrial applications such as finding if a loan applicant is high-risk or low-risk
- For Predicting the failure of mechanical parts in automobile engines

- Predicting social media share scores
- Performance scores

Support Vector Machine

The support vector machine is a classifier that represents the training data as points in space separated into categories by a gap as wide as possible. New points are then added to space by predicting which category they fall into and which space they will belong to.



Advantages and Disadvantages

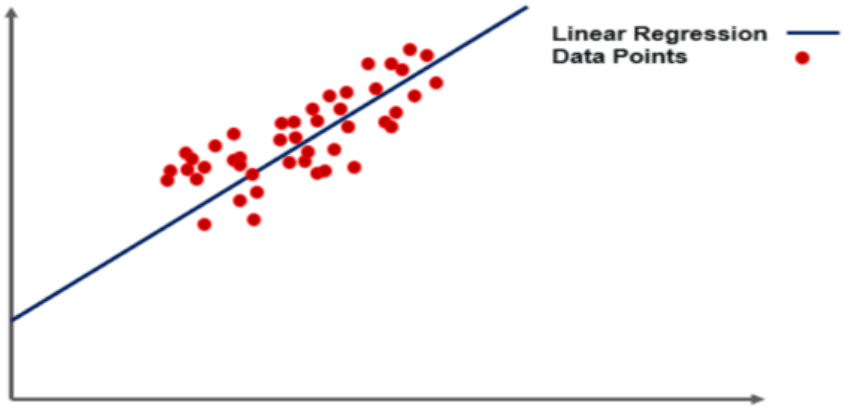
- It uses a subset of training points in the decision function which makes it memory efficient and is highly effective in high dimensional spaces.
- The support vector machine algorithm does not directly provide probability estimates.

Use cases

- Business applications for comparing the performance of a stock over a period of time
- Investment suggestions
- Classification of applications requiring accuracy and efficiency

Linear Regression

Linear Regression is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y (output). Hence, the name is Linear Regression.

PREVIOUS ACTIVITY

◄ Twitter Data Access: Python Code

Jump to...

NEXT ACTIVITY

Mathematics Behind Classification Algorithm ►

Stay in touch

Contact Us

http://nielit.gov.in/gorakhpur/

abhinav@nielit.gov.in or ajay.verma@nielit.gov.in