Machine Learning

Regression, Classification and Clustering

Regression:

Regression is a process of finding the correlations between dependent and independent variables. It helps in predicting the continuous variables such as prediction of **Market Trends**, prediction of House prices, etc.

The task of the Regression algorithm is to find the mapping function to map the input variable(x) to the continuous output variable(y).

Example: Suppose we want to do weather forecasting, so for this, we will use the Regression algorithm. In weather prediction, the model is trained on the past data, and once the training is completed, it can easily predict the weather for future days.

What is Classification?

We use the training dataset to get better boundary conditions which could be used to determine each target class. Once the boundary conditions are determined, the next task is to predict the target class. The whole process is known as classification.

Target class examples:

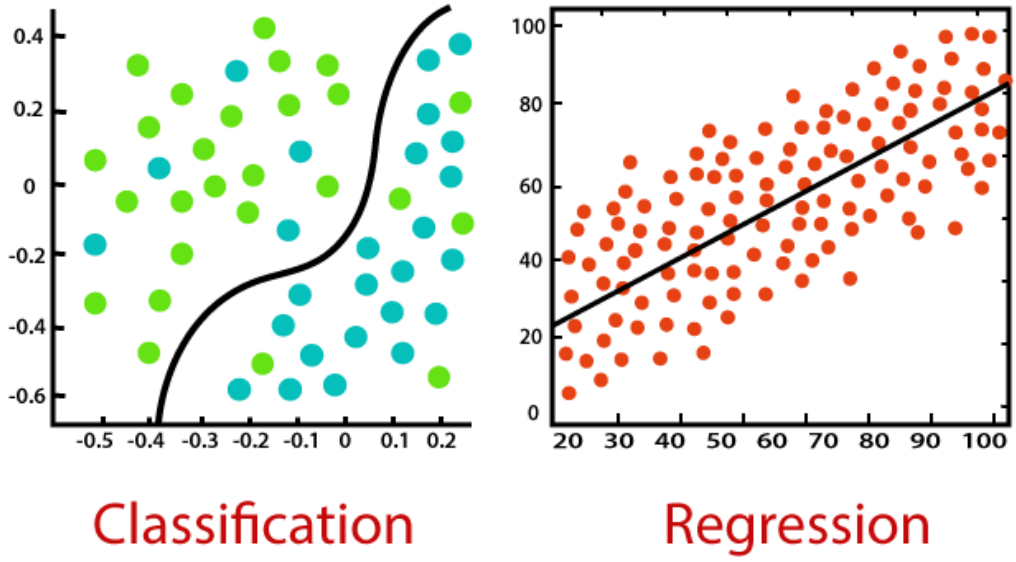
- Analysis of the customer data to predict whether he will buy computer accessories **(Target class: Yes or No)**
- Classifying fruits from features like color, taste, size, weight **(Target classes: Apple, Orange, Cherry, Banana)**
- Gender classification from hair length **(Target classes: Male or Female)**

Let's understand the concept of classification algorithms with gender classification using hair length (by no means am I trying to stereotype by gender, this is only an example). To classify gender **(target class)** using hair length as feature parameter we could train a model using any classification algorithms to come up with some set of boundary conditions which can be used to differentiate the male and female

genders using hair length as the training feature. In gender classification case the boundary condition could the proper hair length value. Suppose the **differentiated boundary** hair length value is 15.0 cm then we can say that if hair length is **less than 15.0 cm** then gender could be male or else female.

Classification Algorithms vs Regression

The main difference between Regression and Classification algorithms that Regression algorithms are used to predict the continuous values such as price, salary, age, etc. and Classification algorithms are used to predict/Classify the discrete values such as Male or Female, True or False, Spam or Not Spam, etc.



Classification Algorithms vs Clustering Algorithms

In clustering, the idea is not to predict the target class as in classification, it's more ever trying to group the similar kind of things by considering the most satisfied condition, **all the items in the same group should be similar and no two different group items should not be similar.**

Group items Examples:

- While grouping similar language type documents **(Same language documents are one group.)**
- While categorizing the news articles **(Same news category(Sport) articles are one group)**

Let's understand the concept with clustering genders based on hair length example. To determine gender, different similarity measure could be used to categorize male and female genders. This could be done by finding the similarity between two hair lengths and keep them in the same group if the similarity is less **(Difference of hair length is less)**. The same process could continue until all the hair length properly grouped into two categories.

Basic Terminology in Classification Algorithms

- **Classifier:** An algorithm that maps the input data to a specific category.
- **Classification model:** A classification model tries to draw some conclusion from the input values given for training. It will predict the class labels/categories for the new data.
- **Feature:** A feature is an individual measurable property of a phenomenon being observed.
- **Binary Classification:** Classification task with two possible outcomes. **Eg: Gender classification (Male / Female)**

Applications of Classification Algorithms

- Email spam classification
- Bank customers loan pay willingness prediction.
- Cancer tumor cells identification.
- Sentiment analysis
- Drugs classification
- Facial key points detection
- Pedestrians detection in an automotive car driving.

Types of Classification Algorithms

Classification Algorithms could be broadly classified as the following:

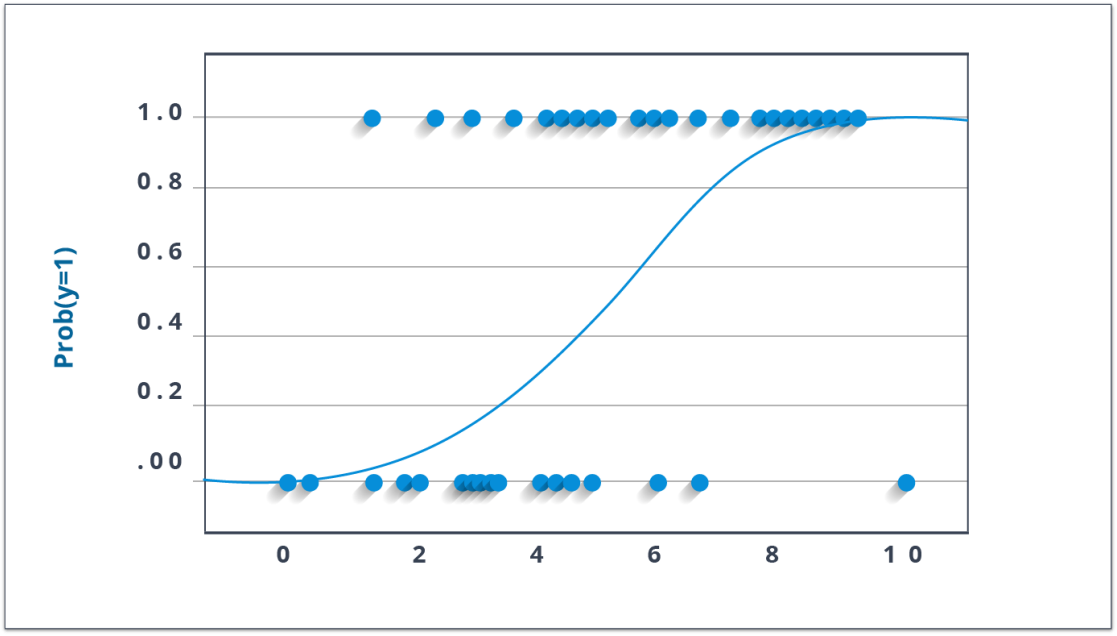
- ***Linear Classifiers***
    - Logistic regression
    - Naive Bayes classifier
- ***Support vector machines***
- ***Kernel estimation***
    - k-nearest neighbor
- ***Decision trees***
    - Random forests
- ***Neural networks***

Examples of a few popular Classification Algorithms are given below.

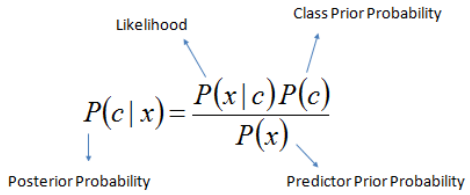
Logistic Regression

As confusing as the name might be, you can rest assured. Logistic Regression is a classification and not a regression algorithm. It estimates discrete values **(Binary values like 0/1, yes/no, true/false)** based on a given set of independent variable(s).

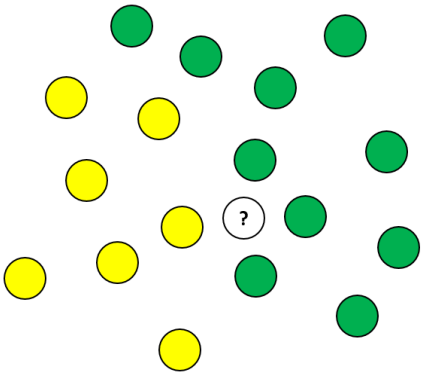
Let us try and understand this through an example.



Naive Bayes classifier

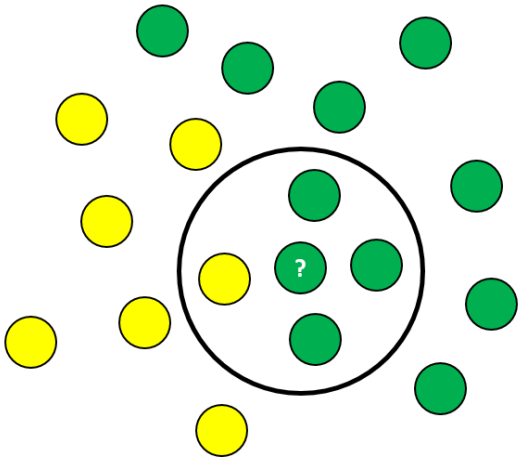


$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$



P(yellow) = $\frac{7}{17}$ P(green) = $\frac{10}{17}$

We have a dataset with two labels
(green and yellow class).
We can calculate the probability of these classes



P(yellow) = $\frac{7}{17}$

P(green) = $\frac{10}{17}$

P'(? | green) = $\frac{3}{10}$

P'(? | yellow) = $\frac{1}{7}$

prior probabilities
number of samples in a given class
divided by the total number of samples

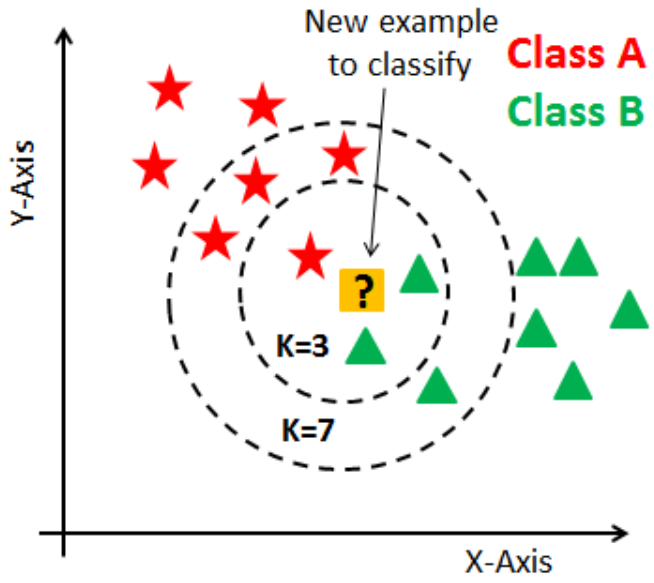
we consider just the vicinity
of the new sample we wan to classify

posterior probability P''(? is green) = P(green) * P'(? | green) = $\frac{10}{17} * \frac{3}{10} = \frac{30}{170}$

posterior probability P''(? is yellow) = P(yellow) * P'(? | yellow) = $\frac{7}{17} * \frac{1}{7} = \frac{7}{119}$

KNN (k- Nearest Neighbors)

K nearest neighbors is a simple algorithm used for both classification and regression problems. It basically stores all available cases to classify the new cases by a majority vote of its k neighbors. The case assigned to the class is most common amongst its K nearest neighbors measured by a distance function (Euclidean, Manhattan, Minkowski, and Hamming).



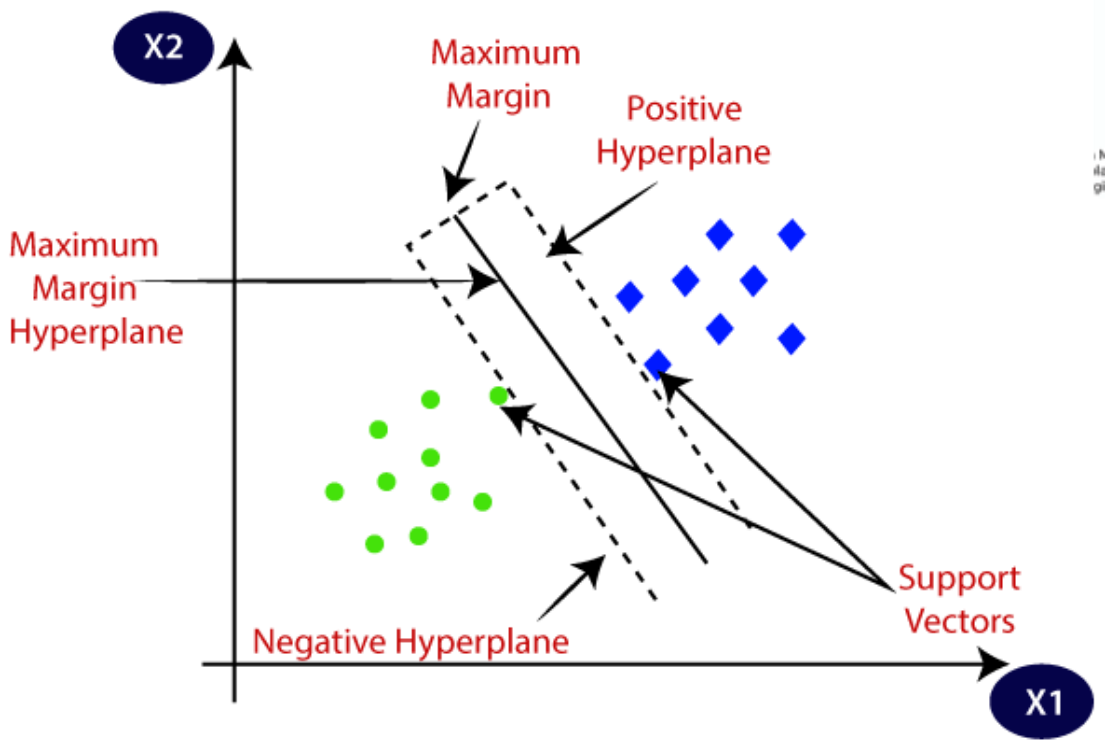
While the three former distance functions are used for continuous variables, Hamming distance function is used for categorical variables. If K = 1, then the case is simply assigned to the class of its nearest neighbor. At times, choosing K turns out to be a challenge while performing kNN modeling.

SVM(Support Vector Machine)

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

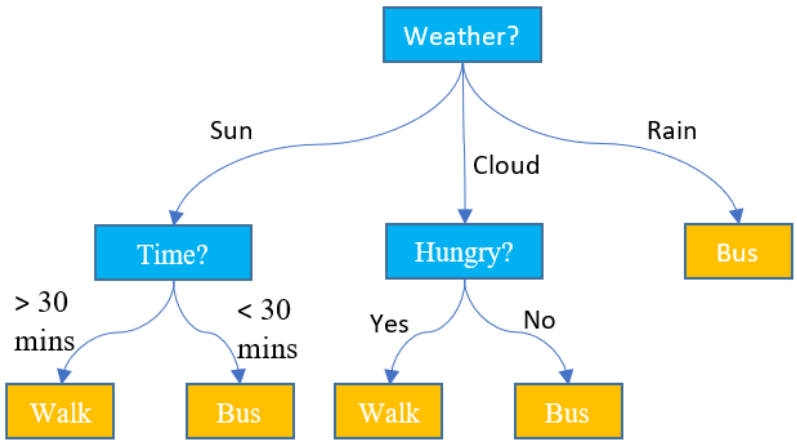
The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



Decision Trees

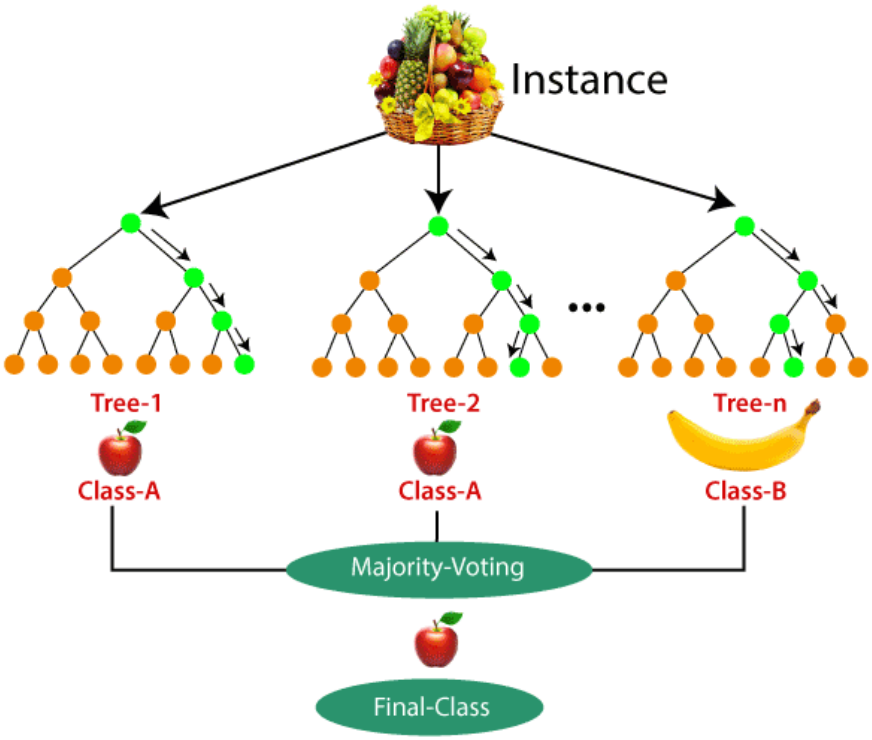
Now, the ***decision tree*** is by far, one of my favorite algorithms. With versatile features helping actualize both categorical and continuous dependent variables, it is a type of supervised learning algorithm mostly used for classification problems. What this algorithm does is, it splits the population into two or more homogeneous sets based on the most significant attributes making the groups as distinct as possible.



Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, **"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."** Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.


Instance
Tree-1
Tree-2
Tree-n
Class-A
Class-A
Class-B
Majority-Voting
Final-Class

Unsupervised learning

Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision.

Why use Unsupervised Learning?

Below are some main reasons which describe the importance of Unsupervised Learning:

Unsupervised learning is helpful for finding useful insights from the data.

Unsupervised learning is much similar as a human learns to think by their own experiences, which makes it closer to the real AI.

Unsupervised learning works on unlabeled and uncategorized data which make unsupervised learning more important.

In real-world, we do not always have input data with the corresponding output so to solve such cases, we need unsupervised learning.

Unsupervised Learning algorithms:

Below is the list of some popular unsupervised learning algorithms:

K-means clustering

KNN (k-nearest neighbors)

Hierarchal clustering

Anomaly detection

Neural Networks

Principle Component Analysis

Independent Component Analysis

Advantages of Unsupervised Learning

Unsupervised learning is used for more complex tasks as compared to supervised learning because, in unsupervised learning, we don't have labeled input data.

Unsupervised learning is preferable as it is easy to get unlabeled data in comparison to labeled data.

Disadvantages of Unsupervised Learning

Unsupervised learning is intrinsically more difficult than supervised learning as it does not have corresponding output.

The result of the unsupervised learning algorithm might be less accurate as input data is not labeled, and algorithms do not know the exact output in advance.

k-means Clustering

k-means clustering is one of the simplest algorithms which uses unsupervised learning method to solve known clustering issues. k-means clustering require following two inputs.

k = number of clusters

Training set(m) = {x1, x2, x3,……., xm}

Let’s say you have an unlabeled data set like the one shown below and you want to group this data into clusters.