SUPERVISED LEARNING
(SESSION-2018-19)

(Regression & Classification)

OVERVIEW OF MACHINE LEARNING

Machine Learning is the process of creating models that can perform a certain task without the need for a human explicitly programming it to do something.

WHAT IS SUPERVISED LEARNING?

Relationship → "Dog"

Relationship → "Cat"

FEATURES

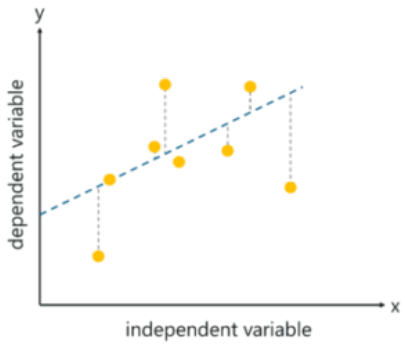
- Dogs and cats both have 4 legs and a tail.
- Dogs come in small to large sizes. Cats, on the other hand, are always small.
- Dogs have a long mouth while cats have smaller mouths.
- Dogs bark while cats meow.
- Different dogs have different ears while cats have almost the same kind of ears.

WHY IS IT IMPORTANT?

- Supervised learning gives the algorithm experience which can be used to output the predictions for new unseen data
- Experience also helps in optimizing the performance of the algorithm

TYPES OF SUPERVISED LEARNING (REGRESSION & CLASSIFICATION)

- Regression: Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent and independent variable.



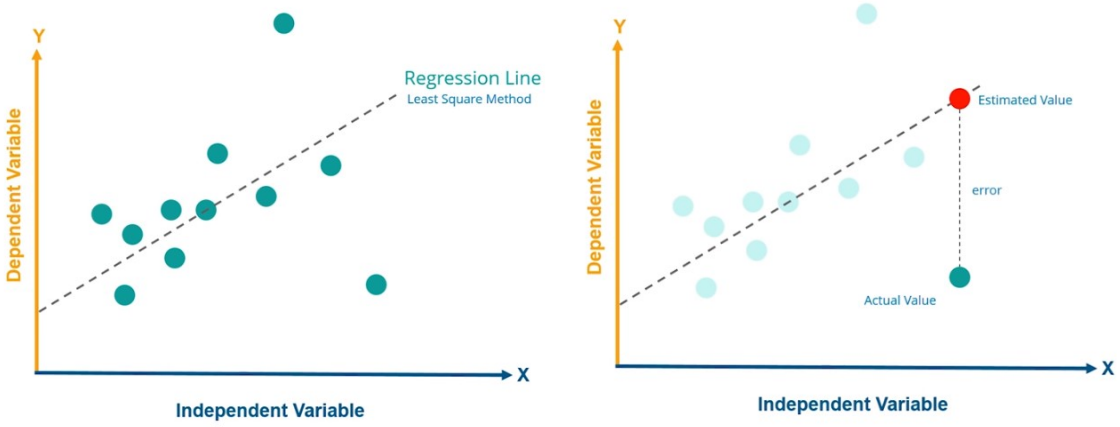
USES OF REGRESSION

- Determining the strength of predictors (strength of the effect that the independent variable have on the dependent variable)
- Forecasting an effect
- Trend forecasting

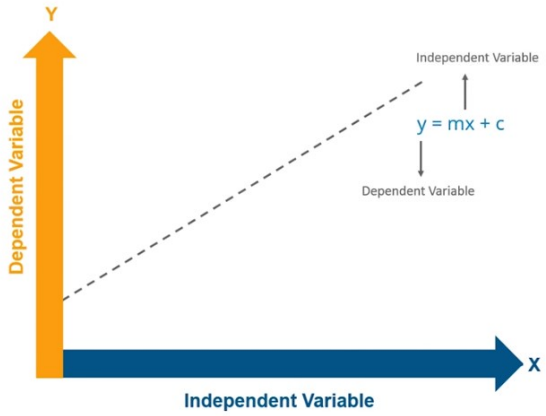
LINEAR VS LOGISTIC REGRESSION

Basis	Linear Regression	Logistic Regression
Core Concept	The data is modelled using a straight line	The probability of some obtained event is represented as a linear function of a combination of predictor variables.
Used with	Continuous Variable	Categorical Variable
Output/Prediction	Value of the variable	Probability of occurrence of event
Accuracy and Goodness of fit	measured by loss, R squared, Adjusted R squared etc.	Accuracy, Precision, Recall, F1 score, ROC curve, Confusion Matrix, etc

LINEAR REGRESSION



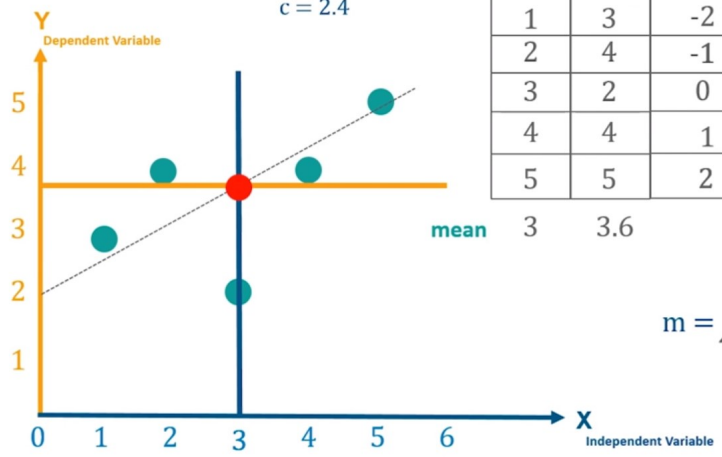
LINEAR REGRESSION



LINEAR REGRESSION

$y = mx+c$

$c = 2.4$



x	y	$x-\dot{x}$	$y-\dot{y}$	$(x-\dot{x})^2$	$(x-\dot{x})(y-\dot{y})$
1	3	-2	-0.6	4	1.2
2	4	-1	0.4	1	-0.4
3	2	0	-1.6	0	0
4	4	1	0.4	1	0.4
5	5	2	1.4	4	2.8
mean	3	3.6		$\Sigma = 10$	$\Sigma = 4$

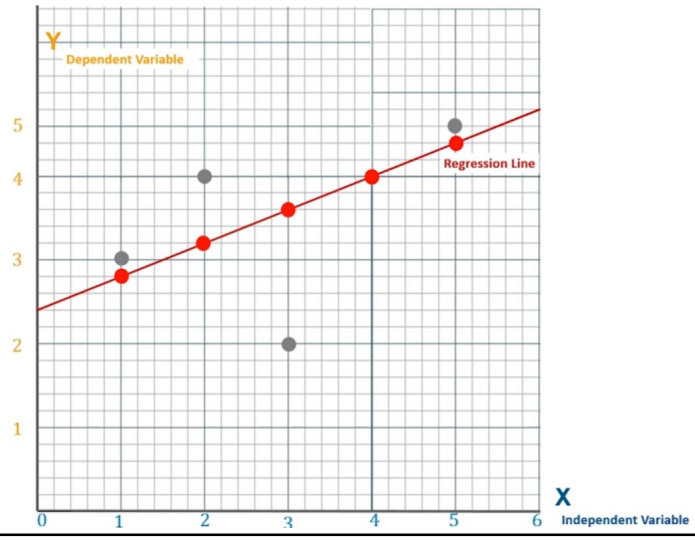
$$m = \sum \frac{(x-\dot{x})(y-\dot{y})}{(x-\dot{x})^2} = \frac{4}{10}$$

$m = 0.4$

$c = 2.4$

$y = 0.4x + 2.4$

LINEAR REGRESSION



$m = 0.4$
$c = 2.4$
$y = 0.4x + 2.4$

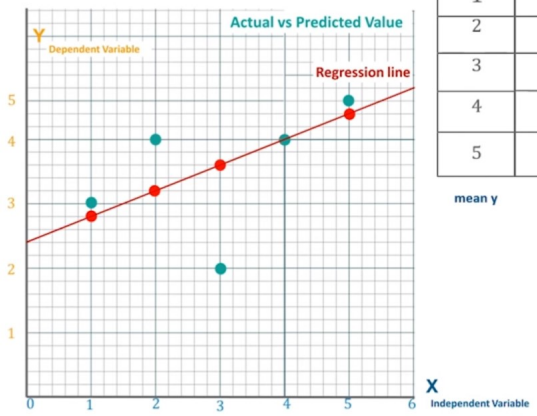
For given $m = 0.4$ & $c = 2.4$, lets predict values for y for $x = \{1,2,3,4,5\}$

$y = 0.4 \times 1 + 2.4 = 2.8$
$y = 0.4 \times 2 + 2.4 = 3.2$
$y = 0.4 \times 3 + 2.4 = 3.6$
$y = 0.4 \times 4 + 2.4 = 4.0$
$y = 0.4 \times 5 + 2.4 = 4.4$

R-SQUARED VALUE

- R-squared value is a statistical measure of how close the data are to the fitted regression line.
- It is also known as coefficient of determination, or the coefficient of multiple determination

GOODNESS OF FIT

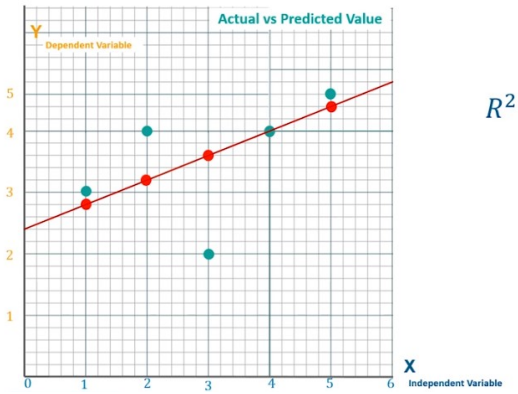


x	y	$y-y$	$(y-y)^2$	y_p	(y_p-y)	$(y_p-y)^2$
1	3	– 0.6	0.36	2.8	-0.8	0.64
2	4	0.4	0.16	3.2	-0.4	0.16
3	2	−1.6	2.56	3.6	0	0
4	4	0.4	0.16	4.0	0.4	0.16
5	5	1.4	1.96	4.4	0.8	0.64
mean y	3.6		$\sum$ 5.2			$\sum$ 1.6

$$R^2 = \frac{1.6}{5.2} = \frac{\sum \left(y_p - \bar{y}\right)^2}{\sum (y-\bar{y})^2}$$

CONT....

- When the value of R square is equal to 1 then the actual values lies on the regression line.



$$R^2 \approx 0.3$$

MEAN SQUARED ERROR

$$ms\varepsilon = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (mx_i + b)\right)^2$$

Cost Function

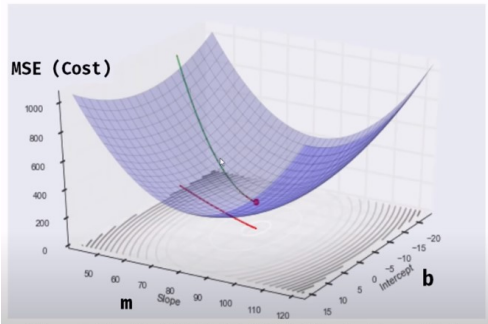
GRADIENT DESCENT

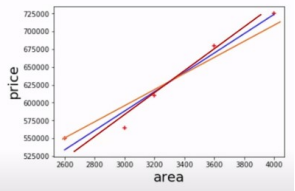
Gradient descent is an algorithm that finds best fit line for given training data set

EXAMPLE

area = [2600,3000,3200,3600,4000]

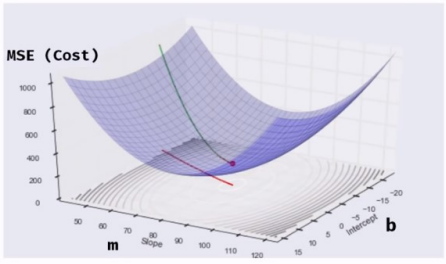
price = [550k,565k,610k,680k,725k]

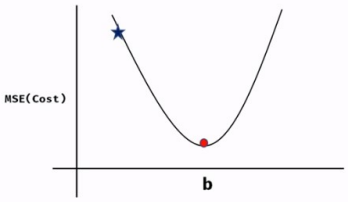


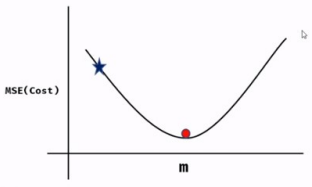


Reference: https://am207.github.io/2017/wiki/gradientdescent.html

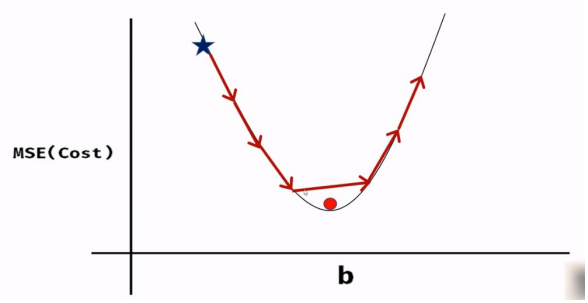
CONT....

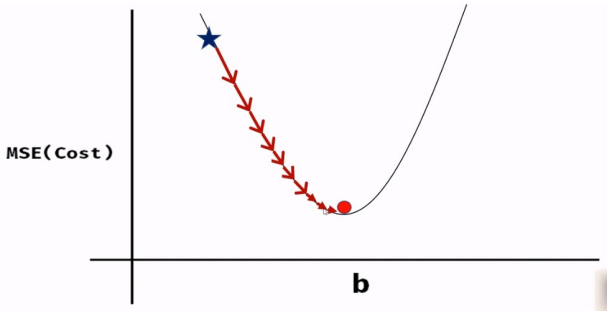






CONT....





CONT.....

$$mse = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (mx_i + b)\right)^2$$

$$\partial/\partial m = \frac{2}{n}\sum_{i=1}^{n} -x_i\left(y_i - (mx_i + b)\right)$$

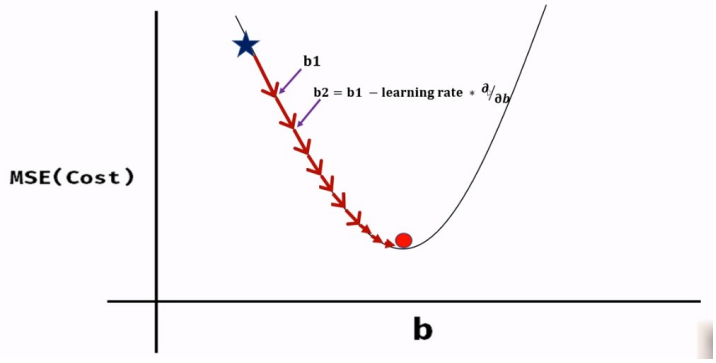
$$\partial/\partial b = \frac{2}{n}\sum_{i=1}^{n} -\left(y_i - (mx_i + b)\right)$$

For Slope

CONT....

$$m = m - \text{learning rate} * \partial/\partial m$$

$$b = b - \text{learning rate} * \partial/\partial b$$



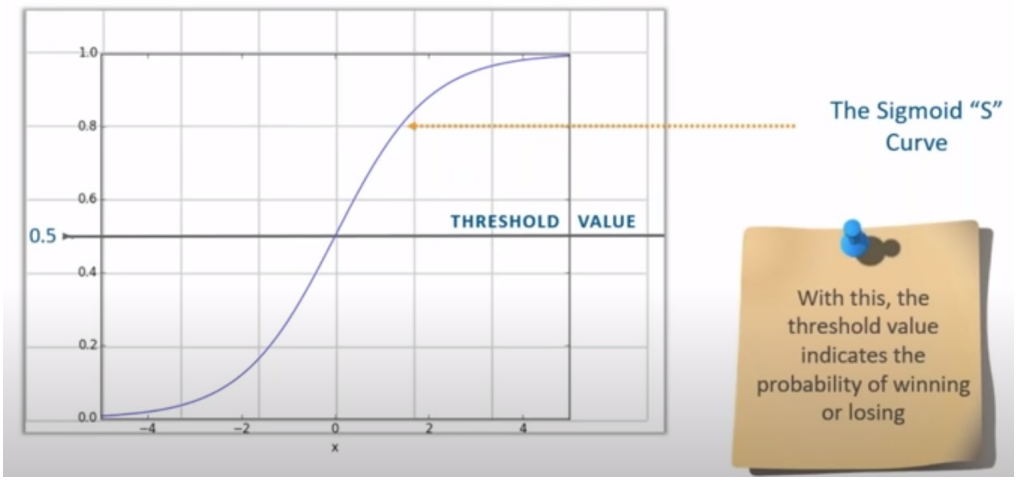
CODE

```
import numpy as np

def gradient_descent(x,y):
    m_curr = b_curr = 0
    iterations = 10000
    n = len(x)
    learning_rate = 0.08

    for i in range(iterations):
        y_predicted = m_curr * x + b_curr
        cost = (1/n) * sum([val**2 for val in (y-y_predicted)])
        md = -(2/n)*sum(x*(y-y_predicted))
        bd = -(2/n)*sum(y-y_predicted)
        m_curr = m_curr - learning_rate * md
        b_curr = b_curr - learning_rate * bd
        print ("m {}, b {}, cost {} iteration {}".format(m_curr,b_curr,cost, i))

x = np.array([1,2,3,4,5])
y = np.array([5,7,9,11,13])

gradient_descent(x,y)
```

LOGISTIC REGRESSION

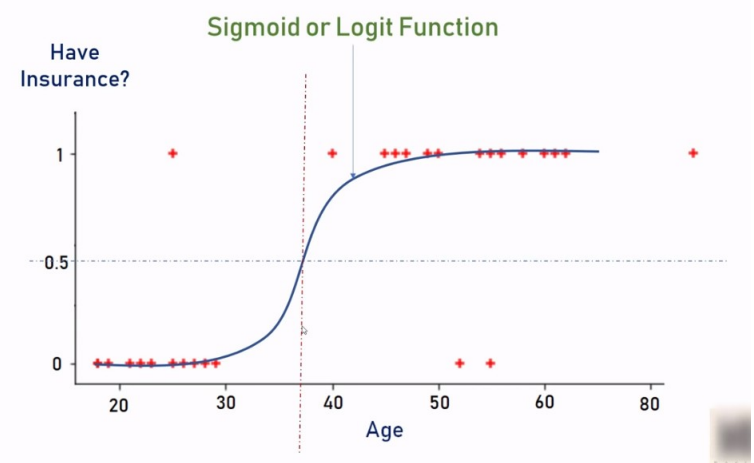
Logistic Regression produces result in a binary format which is used to predict the outcome of a categorical dependent variable. So the outcome should be discrete/categorical such as:

LOGISTIC REGRESSION CURVE



EXAMPLE

age	have_insurance
22	0
25	0
47	1
52	0
46	1
56	1
55	0
60	1
62	1
61	1
18	0
28	0
27	0
29	0
49	1



CONT....

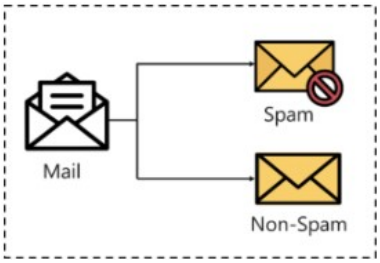
$$sigmoid(z) = \frac{1}{1 + e^{-z}}$$

e = Euler's number ~ 2.71828

Sigmoid function converts input into range 0 to 1

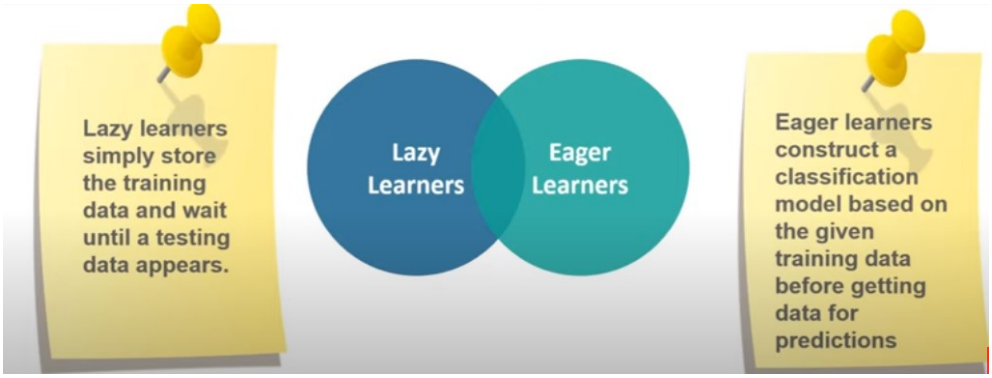
CLASSIFICATION

- Classification is a process of categorizing a given set of data into classes, It can be performed on both structured or unstructured data.
- The process starts with predicting the class of given data points. The classes are often referred to as target, label or categories.



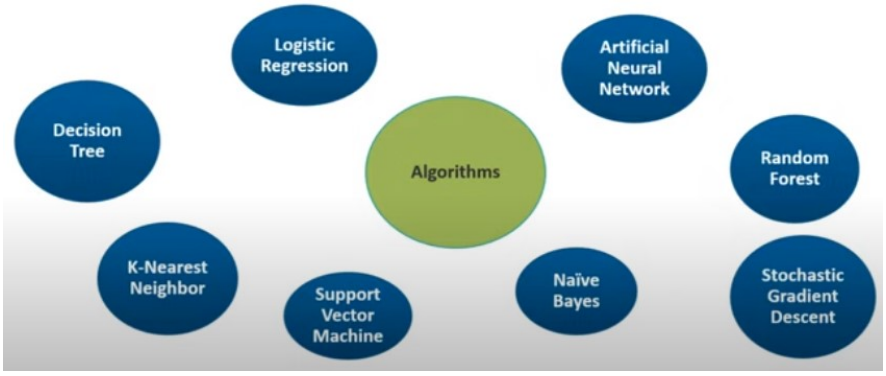
CLASSIFICATION TERMINOLOGIES

TYPES OF LEARNERS IN CLASSIFICATION



CLASSIFICATION ALGORITHMS

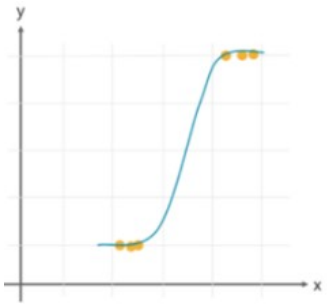
In machine learning, classification is a supervised learning concept which basically categorizes a set of data into classes.



LOGISTIC REGRESSION

It is a classification algorithm in machine learning that uses one or more independent variables to determine an outcome.

It will have only two possible outcomes.



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.

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

STOCHASTIC GRADIENT DESCENT

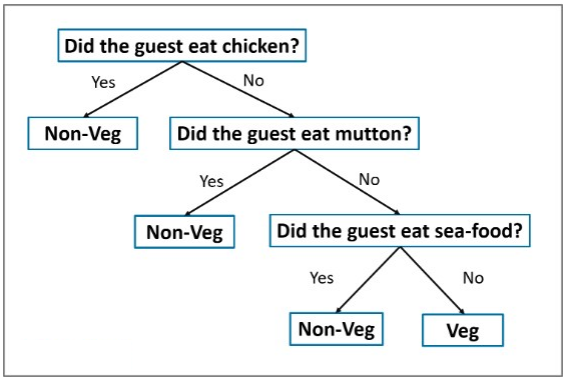
- It is a very effective and simple approach to fit linear models.
- Stochastic Gradient Descent is particularly useful when the sample data is in a large number.

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.

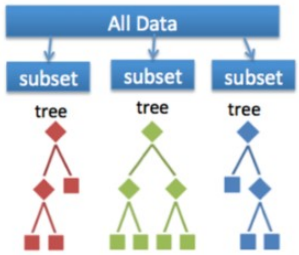
DECISION TREE

The decision tree algorithm builds the classification model in the form of a tree structure.



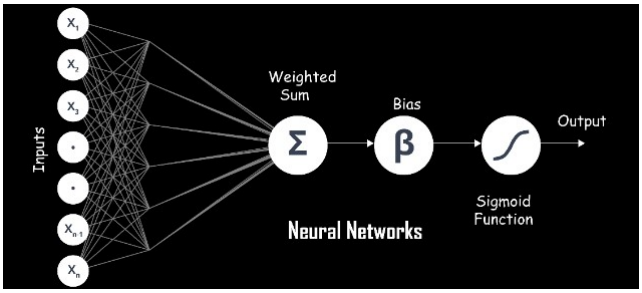
RANDOM FOREST

- 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.

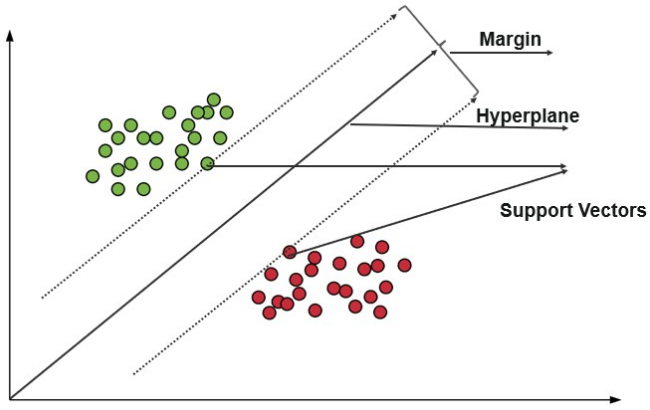


ARTIFICIAL NEURAL NETWORKS

A neural network consists of neurons that are arranged in layers, they take some input vector and convert it into an output. The process involves each neuron taking input and applying a function which is often a non-linear function to it and then passes the output to the next layer.

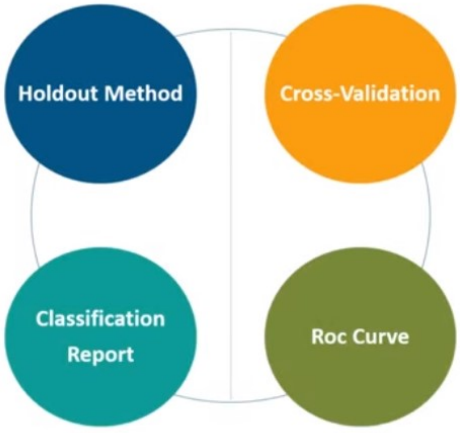


SUPPORT VECTOR MACHINE



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.

CLASSIFIER EVALUATION

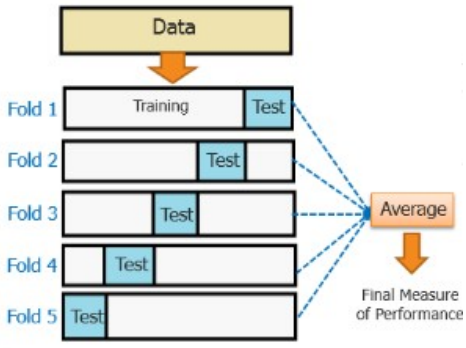


HOLDOUT METHOD

This is the most common method to evaluate a classifier. In this method, the given data set is divided into two parts as a test and train set 20% and 80% respectively.

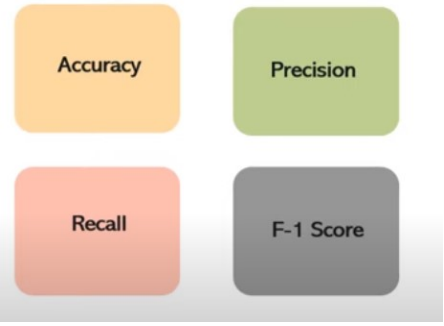
The train set is used to train the data and the unseen test set is used to test its predictive power.

CROSS-VALIDATION



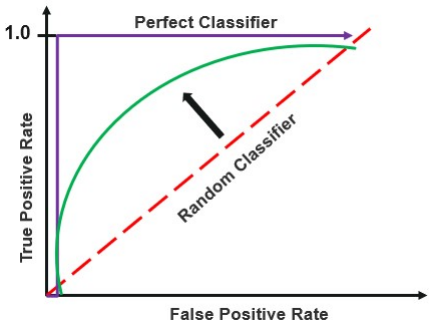
- Over-fitting is the most common problem prevalent in most of the machine learning models. K-fold cross-validation can be conducted to verify if the model is over-fitted at all.
- In this method, the data set is randomly partitioned into k mutually exclusive subsets, each of which is of the same size. Out of these, one is kept for testing and others are used to train the model. The same process takes place for all k folds.

CLASSIFICATION REPORT

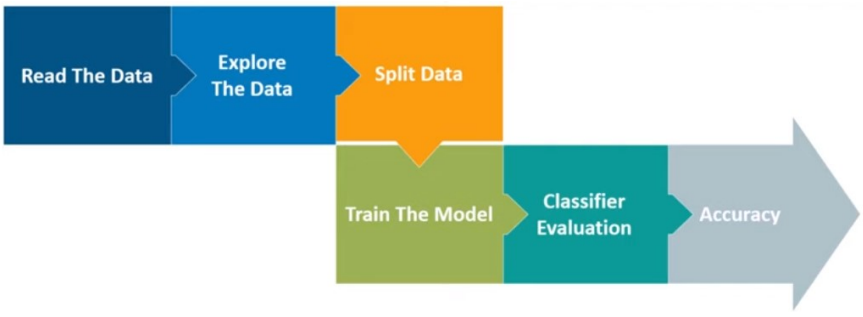


ROC CURVE

Receiver operating characteristics or ROC curve is used for visual comparison of classification models, which shows the relationship between the true positive rate and the false positive rate. The area under the ROC curve is the measure of the accuracy of the model.



ALGORITHM SELECTION



Thank You…