Unit - 1

1. Explain machine learning in brief. Discuss application & limitations of machine learning.

Ans: The field of study known as machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.

The field of study that gives computers the ability to learn without being explicitly programmed . However, there is no universally accepted definition for machine learning.

Machine learning is programming computers to optimize a performance criterion using example data or past experience . We have a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience. The model may be predictive to make predictions in the future, or descriptive to gain knowledge from data.

**Applications:**

* Predictions while CommutingVideos Surveillance
* Email Spam and Malware Filtering
* Search Engine Result Refining
* Product Recommendations

**Limitations:**

* ML needs enough time to let the algorithms learn and develop enough to fulfill their purpose with a considerable amount of accuracy and relevancy. It also needs massive resources to function. This can mean additional requirements of computer power for you.
* Machine Learning requires massive data sets to train on, and these should be inclusive/unbiased, and of good quality. There can also be times where they must wait for new data to be generated.
* Another major challenge is the ability to accurately interpret results generated by the algorithms. You must also carefully choose the algorithms for your purpose.

2. Compare Low dimensional & High dimensional Data in reference to machine learning

Ans:

**High dimensional data** refers to a dataset in which the number of features *p* is larger than the number of [observations](https://www.statology.org/observation-in-statistics/) *N*, often written as *p* >> *N*.For example, a dataset that has *p* = 6 features and only *N* = 3 observations would be considered high dimensional data because the number of features is larger than the number of observations.One common mistake people make is assuming that “high dimensional data” simply means a dataset that has a lot of features. However, that’s incorrect. A dataset could have 10,000 features, but if it has 100,000 observations then it’s not high dimensional.

**Low-dimensional** representation refers to the outcome of a dimension reduction process on high-dimensional data. The low-dimensional representation of the data is expected to retain as much information as possible from the high-dimensional data. Usually, there is a tradeoff between how low the dimension can be reduced and how much information can be retained.

3. Differentiate among AI, ML & DL.

Ans:

**Artificial Intelligence:** Artificial Intelligence is basically the mechanism to incorporate human intelligence into machines through a set of rules(algorithm). AI is a combination of two words: “Artificial” meaning something made by humans or non-natural things and “Intelligence” meaning the ability to understand or think accordingly. Another definition could be that **“AI is basically the study of training your machine(computers) to mimic a human brain and it’s thinking capabilities”**. AI focuses on 3 major aspects(skills): **learning, reasoning and self-correction** to obtain maximum efficiency possible.

**Machine Learning:** Machine Learning is basically the study/process which provides the system(computer) to learn automatically on its own through experiences it had and improve accordingly without being explicitly programmed. **ML is an application or subset of AI.** ML focuses on the development of programs so that it can access data to use it for themselves. The entire process makes observations on data to identify the possible patterns being formed and make better future decisions as per the examples provided to them. **The major aim of ML is to allow the systems to learn by themselves through the experience without any kind of human intervention or assistance.**

**Deep Learning:** Deep Learning is basically a sub-part of the broader family of Machine Learning which makes use of **Neural Networks**(similar to the neurons working in our brain) to mimic human brain-like behavior. DL algorithms focus on **information processing patterns** mechanism to possibly identify the patterns just like our human brain does and classifies the information accordingly. DL works on larger sets of data when compared to ML and **prediction mechanism is self-administered by machines**.

4. Discuss various types of machine learning algorithm.

Ans: There are **four** types of machine learning algorithms.

**Supervised Learning:**

“*The outcome or output for the given input is known* *before itself”* and the machine must be able to map or assign the given input to the output. Multiple images of a cat, dog, orange, apple etc here the images are labelled. It is fed into the machine for training and the machine must identify the same. Just like a human child is shown a cat and told so, when it sees a completely different cat among others still identifies it as a cat, the same method is employed here.

* ***Regression and classification*** problems are mainly solved here.
* ***Labelled data*** is used for training here.
* ***Popular Algorithms:*** Linear Regression, Support Vector Machines (SVM), Neural Networks, Decision Trees, Naive Bayes, Nearest Neighbor.
* It is mainly used in ***Predicting Modelling***.

**Unsupervised Learning:**

*“The outcome or output for the given inputs is unknown”,* here input data is given and the model is run on it. The image or the input given are grouped together here and insights on the inputs can be found here(which is the most of the real world data available). The main algorithms include *Clustering algorithms( ) and* learning algorithms.

* It is used for ***Clustering problems(***grouping***), Anomaly Detection (in banks for unusual transactions)*** where there is a need for finding relationships among the data given.
* ***Unlabeled data*** is used in unsupervised learning.
* ***Popular Algorithms:*** *k-means clustering,* Association rule.
* It is mainly used in ***Descriptive Modelling.***

**Semi-supervised Learning:**

It is in-between that of ***Supervised and Unsupervised Learning***. Where the combination is used to produce the desired results and it is the most important in real-world scenarios where all the data available are a combination of ***labelled and unlabeled data***.

**Reinforced Learning:**

The machine is exposed to an ***environment where it gets trained by trial and error method***, here it is trained to make a much specific decision. The machine learns from past experience and tries to capture the best possible knowledge to make ***accurate decisions*** based on the feedback received.

* Basic reinforcement is modelled as ***Markov Decision Process***
* The most popular algorithms used here is ***Q-Learning***, ***Deep Adversarial Networks.***
* Its practical applications include computer playing board games such as ***chess*** and ***GO***, **Self-driving cars** also use this learning.

5. Explain, why the knowledge of linear algebra, statistic & probability theory in beneficial

in machine learning development.

Ans:

**Use of Linear Algebra**

There is no denying the fact that calculus trumps linear algebra when it comes to advanced mathematics. Integral and differential calculus help you a lot more than just with integration, differentiation, and limits, they also serve as fundamental knowledge required for applications, such as tensors and vectors.Learning these things will help you have a better understanding of linear equations and linear functions amongst other areas. You will also know about advanced concepts, such as the Simplex method and spatial vectors. If you need help with linear programming, you can use the Simplex method.

**Use of Statistics**

Statistics is generally considered a prerequisite to the field of applied machine learning.We need statistics to help transform observations into information and to answer questions about samples of observations.Statistics is a collection of tools developed over hundreds of years for summarizing data and quantifying properties of a domain given a sample of observations.

**Use of Probability**

### Sampling – Dealing with non-deterministic processes

### Pattern recognition

### Training – use in Maximum likelihood estimation

### Developing specific algorithms

### Hyperparameter optimization

### Model evaluation

Unit - 2

1. Explain what bias and variance and their effect on accuracy of ML model developed.

Also discuss bias-variance trade-off in brief.

Ans -: Bias is a phenomenon that skews the result of an algorithm in favor or against an idea.

Bias is considered a systematic error that occurs in the machine learning model itself due to incorrect assumptions in the ML process.Technically, we can define bias as the error between average model prediction and the ground truth. Moreover, it describes how well the model matches the training data set.

* A model with a higher bias would not match the data set closely.
* A low bias model will closely match the training data set.

Variance refers to the changes in the model when using different portions of the training data set.Simply stated, variance is the variability in the model prediction—how much the ML function can adjust depending on the given data set. Variance comes from highly complex models with a large number of features

* Models with high bias will have low variance.
* Models with high variance will have a low bias.

Bias and variance are inversely connected. It is impossible to have an ML model with a low bias and a low variance.When a [data engineer](https://www.bmc.com/blogs/data-engineer-vs-data-scientist/) modifies the ML algorithm to better fit a given data set, it will lead to low bias—but it will increase variance. This way, the model will fit with the data set while increasing the chances of inaccurate predictions.The same applies when creating a low variance model with a higher bias. While it will reduce the risk of inaccurate predictions, the model will not properly match the data set. It’s a delicate balance between these bias and variance. Importantly, however, having a higher variance does not indicate a bad ML algorithm. Machine learning algorithms should be able to handle some variance.

2. Differentiate between Overfitting &amp; Underfitting in machine learning.

Ans Overfitting occurs when our [machine learning](https://www.javatpoint.com/machine-learning) model tries to cover all the data points or more than the required data points present in the given dataset. Because of this, the model starts caching noise and inaccurate values present in the dataset, and all these factors reduce the efficiency and accuracy of the model. The overfitted model has **low bias** and **high variance.**

The chances of occurrence of overfitting increase as much we provide training to our model. It means the more we train our model, the more chances of occurring the overfitted model.

Overfitting is the main problem that occurs in [supervised learning](https://www.javatpoint.com/supervised-machine-learning).

### How to avoid the Overfitting in Model

Both overfitting and underfitting cause the degraded performance of the machine learning model. But the main cause is overfitting, so there are some ways by which we can reduce the occurrence of overfitting in our model.

* **Cross-Validation**
* **Training with more data**
* **Removing features**
* **Early stopping the training**
* **Regularization**
* **Ensembling**

Underfitting occurs when our machine learning model is not able to capture the underlying trend of the data. To avoid the overfitting in the model, the fed of training data can be stopped at an early stage, due to which the model may not learn enough from the training data. As a result, it may fail to find the best fit of the dominant trend in the data.

In the case of underfitting, the model is not able to learn enough from the training data, and hence it reduces the accuracy and produces unreliable predictions.

An underfitted model has high bias and low variance.

### How to avoid underfitting:

* By increasing the training time of the model.
* By increasing the number of features.

3. Explain cross validation (CV) in brief using diagram and discuss different-different

CV techniques used in machine learning.

Ans Cross-validation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set.

The three steps involved in cross-validation are as follows :

1. Reserve some portion of sample data-set.
2. Using the rest data-set train the model.
3. Test the model using the reserve portion of the data-set.

**Validation**

In this method, we perform training on the 50% of the given data-set and rest 50% is used for the testing purpose. The major drawback of this method is that we perform training on the 50% of the dataset, it may possible that the remaining 50% of the data contains some important information which we are leaving while training our model i.e higher bias.

**LOOCV (Leave One Out Cross Validation)**

In this method, we perform training on the whole data-set but leaves only one data-point of the available data-set and then iterates for each data-point. It has some advantages as well as disadvantages also.

An advantage of using this method is that we make use of all data points and hence it is low bias.

The major drawback of this method is that it leads to higher variation in the testing model as we are testing against one data point. If the data point is an outlier it can lead to higher variation. Another drawback is it takes a lot of execution time as it iterates over ‘the number of data points’ times.

**K-Fold Cross Validation**

In this method, we split the data-set into k number of subsets(known as folds) then we perform training on the all the subsets but leave one(k-1) subset for the evaluation of the trained model. In this method, we iterate k times with a different subset reserved for testing purpose each time

4. What do you mean by performance metrics/KPI? List different-2 KPI used to evaluate ML

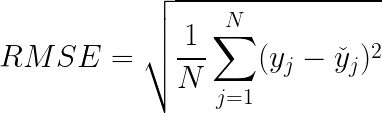
model performance.

Ans [Performance metrics](https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce) are a part of every machine learning pipeline. They tell you if you’re making progress, and put a number on it. All machine learning models, whether it’s linear regression, or a SOTA technique like [BERT](https://neptune.ai/blog/bert-and-the-transformer-architecture-reshaping-the-ai-landscape), need a metric to judge performance.

Every machine learning task can be broken down to either *Regression* or *Classification*, just like the performance metrics. There are dozens of metrics for both problems, but we’re gonna discuss popular ones along with what information they provide about model performance. It’s important to know how your model sees your data!

### Root Mean Squared Error (RMSE)

Root Mean Squared Error corresponds to the square root of the average of the squared difference between the target value and the value predicted by the regression model. Basically, sqrt(MSE). Mathematically it can be represented as:

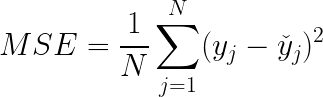


It addresses a few downsides in MSE.

Few key points related to RMSE:

* It retains the differentiable property of MSE.
* It handles the penalization of smaller errors done by MSE by square rooting it.
* Error interpretation can be done smoothly, since the scale is now the same as the random variable.
* Since scale factors are essentially normalized, it’s less prone to struggle in the case of outliers.

Mean squared error is perhaps the most popular metric used for regression problems. It essentially finds the average of the squared difference between the target value and the value predicted by the regression model.



Where:

* y\_j: ground-truth value
* y\_hat: predicted value from the regression model
* N: number of datums

5. What do you understand by fine tuning a ML model? Discuss about Grid Search &amp;

Randomized search method in brief.

Ans *Fine-tuning* is a way of applying or utilizing transfer learning. Specifically, fine-tuning is a process that takes a model that has already been trained for one given task and then tunes or tweaks the model to make it perform a second similar task.Assuming the original task is similar to the new task, using an artificial neural network that has already been designed and trained allows us to take advantage of what the model has already learned without having to develop it from scratch.

When building a model from scratch, we usually must try many approaches through trial-and-error.

### Grid Search

Grid search is a technique which tends to find the right set of hyperparameters for the particular model. Hyperparameters are not the model parameters and it is not possible to find the best set from the training data. Model parameters are learned during training when we optimise a loss function using something like a gradient descent. In this tuning technique, we simply build a model for every combination of various hyperparameters and evaluate each model. The model which gives the highest accuracy wins.

### Random Search

Random search is a technique where random combinations of the hyperparameters are used to find the best solution for the built model. It is similar to grid search, and yet it has proven to yield better results comparatively. The drawback of random search is that it yields high variance during computing. Since the selection of parameters is completely random; and since no intelligence is used to sample these combinations, luck plays its part.

6. What do you understand by ensemble learning? Also discuss concept of bagging &amp;

boosting in detail.

Ans Ensemble methods combine several decision trees classifiers to produce better predictive performance than a single decision tree classifier. The main principle behind the ensemble model is that a group of weak learners come together to form a strong learner, thus increasing the accuracy of the model.

Bagging is used when the goal is to reduce the variance of a decision tree classifier. Here the objective is to create several subsets of data from training sample chosen randomly with replacement. Each collection of subset data is used to train their decision trees. As a result, we get an ensemble of different models. Average of all the predictions from different trees are used which is more robust than a single decision tree classifier.

Advantages:

* Reduces over-fitting of the model.
* Handles higher dimensionality data very well.
* Maintains accuracy for missing data.

Boosting is used to create a collection of predictors. In this technique, learners are learned sequentially with early learners fitting simple models to the data and then analysing data for errors. Consecutive trees (random sample) are fit and at every step, the goal is to improve the accuracy from the prior tree. When an input is misclassified by a hypothesis, its weight is increased so that next hypothesis is more likely to classify it correctly. This process converts weak learners into better performing model.

Advantages:

* Supports different loss function (we have used ‘binary:logistic’ for this example).
* Works well with interactions.

Unit - 3

1. Explain regression & classification problem using suitable example.

**Ans-1**

Difference between Regression and Classification

| **Regression Algorithm** | **Classification Algorithm** |
| --- | --- |
| In Regression, the output variable must be of continuous nature or real value. | In Classification, the output variable must be a discrete value. |
| The task of the regression algorithm is to map the input value (x) with the continuous output variable(y). | The task of the classification algorithm is to map the input value(x) with the discrete output variable(y). |
| Regression Algorithms are used with continuous data. | Classification Algorithms are used with discrete data. |
| In Regression, we try to find the best fit line, which can predict the output more accurately. | In Classification, we try to find the decision boundary, which can divide the dataset into different classes. |
| Regression algorithms can be used to solve the regression problems such as Weather Prediction, House price prediction, etc. | Classification Algorithms can be used to solve classification problems such as Identification of spam emails, Speech Recognition, Identification of cancer cells, etc. |
| The regression Algorithm can be further divided into Linear and Non-linear Regression. | The Classification algorithms can be divided into Binary Classifier and Multi-class Classifier. |
| **Example:** The best example to understand the Classification problem is Email Spam Detection. The model is trained on the basis of millions of emails on different parameters, and whenever it receives a new email, it identifies whether the email is spam or not. If the email is spam, then it is moved to the Spam folder. | **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. |

2. Explain Linear regression using suitable diagram.

**Ans-2**

**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 getting used.

**Hypothesis function for Linear Regression :**

While training the model we are given :

**x:** input training data (univariate – one input variable(parameter))

**y:** labels to data (supervised learning)

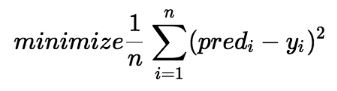
When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and θ2 values.

**θ1:** intercept

**θ2:** coefficient of x

Once we find the best θ1 and θ2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

**Cost Function (J):**

By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ1 and θ2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).



Cost function(J) of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value (pred) and true y value (y).

3. Discuss Gradient Descent Algorithm in detail.

**Ans-3**

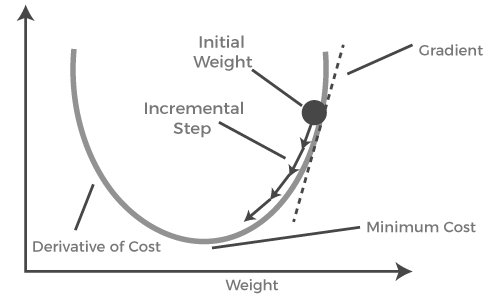
Gradient Descent is known as one of the most commonly used optimization algorithms to train machine learning models by means of minimizing errors between actual and expected results. Further, gradient descent is also used to train Neural Networks.

Gradient descent was initially discovered by **"Augustin-Louis Cauchy"** in mid of 18th century. ***Gradient Descent is defined as one of the most commonly used iterative optimization algorithms of machine learning to train the machine learning and deep learning models. It helps in finding the local minimum of a function.***

2.3K

Machine Learning -The Nature of Probability and Statistics - Data Collection and Sampling Techniques

The best way to define the local minimum or local maximum of a function using gradient descent is as follows:

* If we move towards a negative gradient or away from the gradient of the function at the current point, it will give the **local minimum** of that function.
* Whenever we move towards a positive gradient or towards the gradient of the function at the current point, we will get the **local maximum** of that function.

This entire procedure is known as Gradient Ascent, which is also known as steepest descent. ***The main objective of using a gradient descent algorithm is to minimize the cost function using iteration.*** To achieve this goal, it performs two steps iteratively:

* Calculates the first-order derivative of the function to compute the gradient or slope of that function.
* Move away from the direction of the gradient, which means slope increased from the current point by alpha times, where Alpha is defined as Learning Rate. It is a tuning parameter in the optimization process which helps to decide the length of the steps.

4. Discuss Multiple linear regression (MLR) using example.

**Ans-4**

In simple Linear Regression, where a single Independent/Predictor(X) variable is used to model the response variable (Y). But there may be various cases in which the response variable is affected by more than one predictor variable; for such cases, the Multiple Linear Regression algorithm is used.

Moreover, Multiple Linear Regression is an extension of Simple Linear regression as it takes more than one predictor variable to predict the response variable. We can define it as:

*“Multiple Linear Regression is one of the important regression algorithms which models the linear relationship between a single dependent continuous variable and more than one independent variable.”*

***Example:***

*Prediction of CO2 emission based on engine size and number of cylinders in a car.*

***Some key points about MLR:***

* *For MLR, the dependent or target variable(Y) must be the continuous/real, but the predictor or independent variable may be of continuous or categorical form.*
* *Each feature variable must model the linear relationship with the dependent variable.*
* *MLR tries to fit a regression line through a multidimensional space of data-points.*

5. Explain Polynomial Regression in brief and discuss how it is different from multiple

linear regression.

**Ans-5**

* Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial. The Polynomial Regression equation is given below:

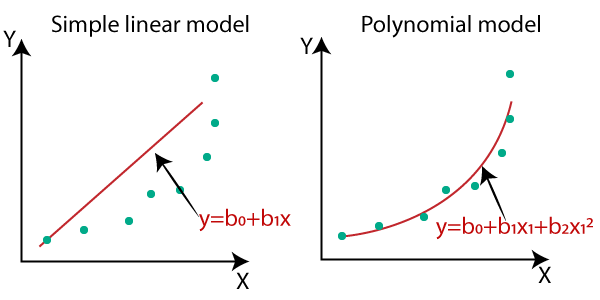
y= b0+b1x1+ b2x12+ b2x13+...... bnx1n

* It is also called the special case of Multiple Linear Regression in ML. Because we add some polynomial terms to the Multiple Linear regression equation to convert it into Polynomial Regression.
* It is a linear model with some modification in order to increase the accuracy.
* The dataset used in Polynomial regression for training is of non-linear nature.
* It makes use of a linear regression model to fit the complicated and non-linear functions and datasets.
* **Hence, *"In Polynomial regression, the original features are converted into Polynomial features of required degree (2,3,..,n) and then modeled using a linear model."***

## *Need for Polynomial Regression:*

The need of Polynomial Regression in ML can be understood in the below points:

* If we apply a linear model on a linear dataset, then it provides us a good result as we have seen in Simple Linear Regression, but if we apply the same model without any modification on a non-linear dataset, then it will produce a drastic output. Due to which loss function will increase, the error rate will be high, and accuracy will be decreased.
* So for such cases, where data points are arranged in a non-linear fashion, we need the Polynomial Regression model. We can understand it in a better way using the below comparison diagram of the linear dataset and non-linear dataset.

**

* *In the above image, we have taken a dataset which is arranged non-linearly. So if we try to cover it with a linear model, then we can clearly see that it hardly covers any data point. On the other hand, a curve is suitable to cover most of the data points, which is of the Polynomial model.*
* *Hence, if the datasets are arranged in a non-linear fashion, then we should use the Polynomial Regression model instead of Simple Linear Regression.*

Unit - 4

1. Why visualization is used in machine learning? Discuss its importance in data exploration

phase of ML.

2. What do you understand by “Curse of Dimensionality”? Does proper Feature Selection

help in avoiding this? Please discuss in brief.

3. Explain the term Dimensionality Reduction in brief. Also discuss the working of

Principal

4. Component Analysis (PCA) algorithm.

5. Explain K Nearest Neighbors (KNN) and its process in detail using suitable diagra

### Unit - 5

#### **1. What do you understand by clustering in ML? Discuss circumstances, when it is applicable to use for ML model development.**

Clustering is an unsupervised machine learning method of identifying and grouping similar data points in larger datasets without concern for the specific outcome. Clustering (sometimes called cluster analysis) is usually used to classify data into structures that are more easily understood and manipulated.

Clustering can not only give you a great starting point but shed light on important features of your data that can be enhanced with deeper analytics. These are just some of the applications of clustering algorithms:

**When you’re starting from a large, unstructured data set**

Clustering can take large datasets and, without instruction, quickly organize them into something more usable. The best part is that if you’re not looking to perform a massive analysis, clustering can give you fast answers about your data.

**When you don’t know how many or which classes your data is divided into**

Clustering is a great first step in your data prep because it starts to answer key questions about your dataset. For instance, you may discover that what you thought were two main subsets are actually four, or what categories you weren’t aware of were their own classes.

**When manually dividing and annotating your data is too resource-intensive**

For smaller datasets, manual annotation and organization is feasible, if not ideal. However, as your data begins to scale, annotation, classification, and categorization become exponentially harder. Clustering — depending on the algorithm you’re using — can cut down your annotation and classification time because it’s less interested in specific outcomes and more concerned with the categorization itself.

**When you’re looking for anomalies in your data**

One of the more valuable uses of clustering is that due to many algorithms’ sensitivity to outlier data points, they can serve as identifiers for data anomalies.Understanding your anomalous data can help you optimize your existing data collection tools, and lead to more accurate results in the long term.

#### **2. Explain working of K-Means Clustering using appropriate diagram.**

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of predefined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

* Determines the best value for K center points or centroids by an iterative process.
* Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

The working of the K-Means algorithm is explained in the below steps:

**Step-1:** Select the number K to decide the number of clusters.

**Step-2:** Select random K points or centroids. (It can be other from the input dataset).

**Step-3:** Assign each data point to their closest centroid, which will form the predefined K clusters.

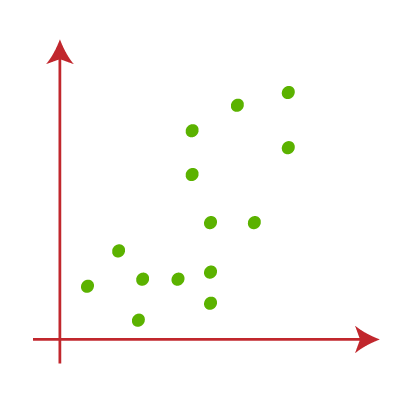
**Step-4:** Calculate the variance and place a new centroid of each cluster.

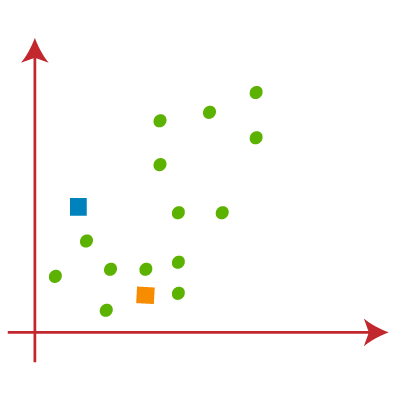
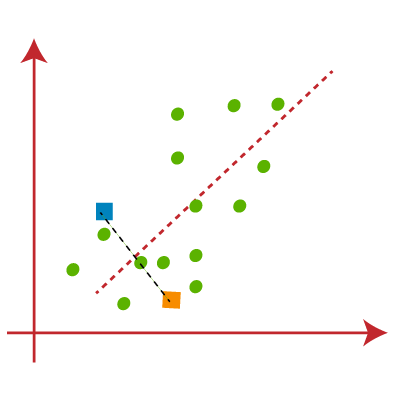
**Step-5:** Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

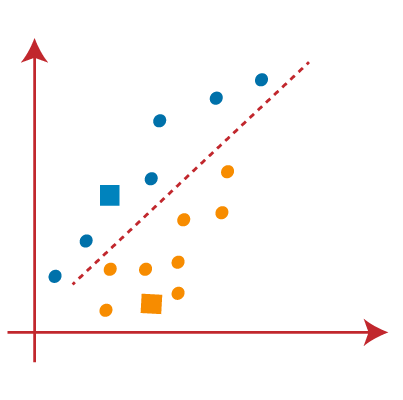
**Step-6:** If any reassignment occurs, then go to step-4 else go to FINISH.

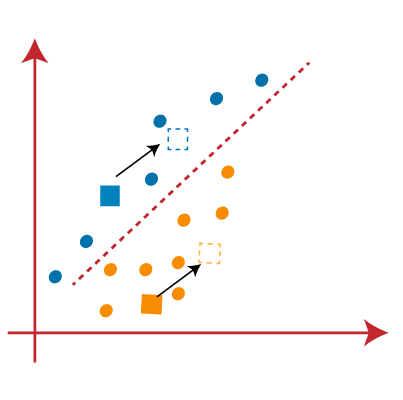
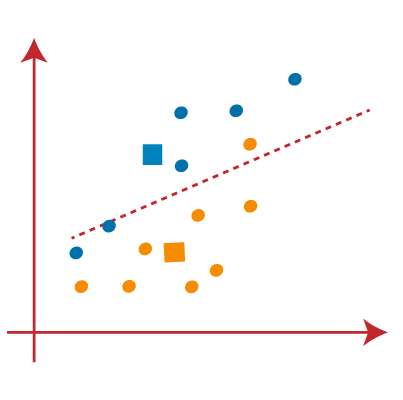
**Step-7**: The model is ready.

Suppose we have two variables M1 and M2. The x-y axis scatter plot of these two variables is given below:

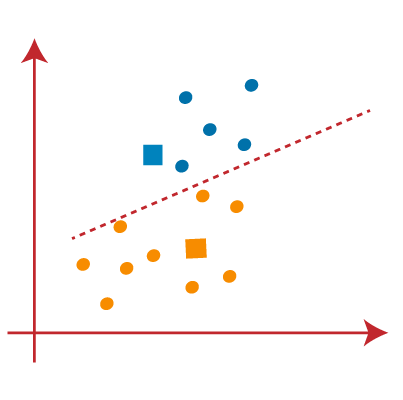


* Let's take number k of clusters, i.e., K=2, to identify the dataset and to put them into different clusters. It means here we will try to group these datasets into two different clusters.
* We need to choose some random k points or centroid to form the cluster. These points can be either the points from the dataset or any other point. So, here we are selecting the below two points as k points, which are not the part of our dataset. 
* Now we will assign each data point of the scatter plot to its closest K-point or centroid. 

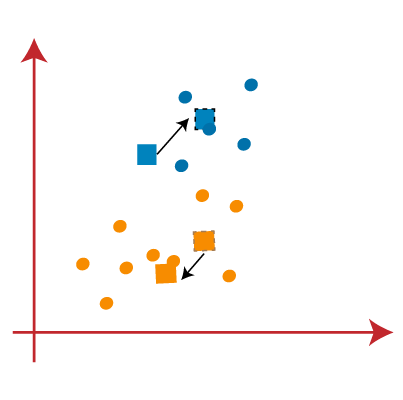
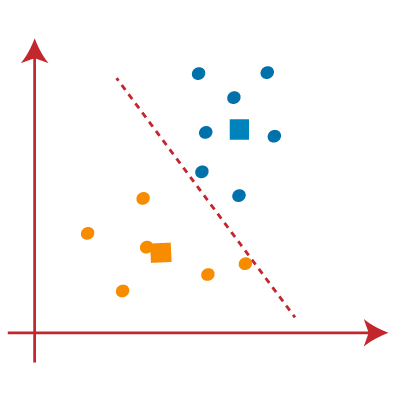
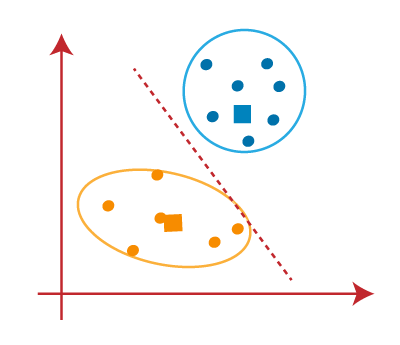
From the above image, it is clear that points left side of the line is near to the K1 or blue centroid, and points to the right of the line are close to the yellow centroid. 

* As we need to find the closest cluster, so we will repeat the process by choosing **a new centroid**. To choose the new centroids, we will compute the center of gravity of these centroids, and will find new centroids as below:  
   
* Next, we will reassign each datapoint to the new centroid. For this, we will repeat the same process of finding a median line.  
   

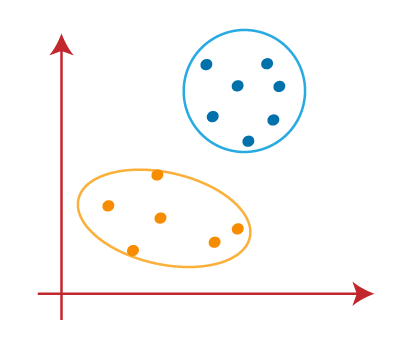
From the above image, we can see, one yellow point is on the left side of the line, and two blue points are right to the line. So, these three points will be assigned to new centroids.



As reassignment has taken place, so we will again go to the step-4, which is finding new centroids or K-points.

* We will repeat the process by finding the center of gravity of centroids, so the new centroids will be as shown in the below image:  
   
* As we got the new centroids so again will draw the median line and reassign the data points. So, the image will be:  
   
* We can see in the above image; there are no dissimilar data points on either side of the line, which means our model is formed. Consider the below image:  
   

As our model is ready, so we can now remove the assumed centroids, and the two final clusters will be as shown in the below image:



#### **3. Explain working of Decision Tree based machine learning algorithm using suitable example.**

* Decision Tree is a **Supervised learning technique** that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where **internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.**
* In a Decision tree, there are two nodes, which are the **Decision Node** and **Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
* The decisions or the test are performed on the basis of features of the given dataset.
* ***It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.***
* It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

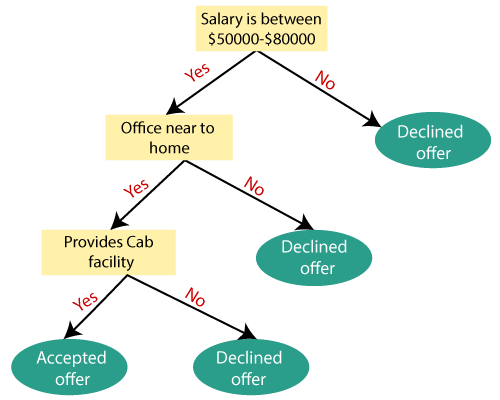


In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.

For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below algorithm:

* **Step-1:** Begin the tree with the root node, says S, which contains the complete dataset.
* **Step-2:** Find the best attribute in the dataset using **Attribute Selection Measure (ASM).**
* **Step-3:** Divide the S into subsets that contains possible values for the best attributes.
* **Step-4:** Generate the decision tree node, which contains the best attribute.
* **Step-5:** Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

**Example:** Suppose there is a candidate who has a job offer and wants to decide whether he should accept the offer or Not. So, to solve this problem, the decision tree starts with the root node (Salary attribute by ASM). The root node splits further into the next decision node (distance from the office) and one leaf node based on the corresponding labels. The next decision node further gets split into one decision node (Cab facility) and one leaf node. Finally, the decision node splits into two leaf nodes (Accepted offers and Declined offer). Consider the below diagram:

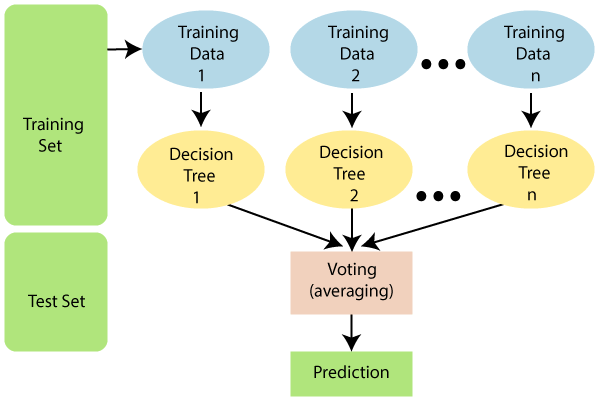


#### **4. Discuss process of Random Forest algorithm in detail.**

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.

**The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.**



Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier:

* There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
* The predictions from each tree must have very low correlations.

**Advantages of Random Forest :**

* It takes less training time as compared to other algorithms.
* It predicts output with high accuracy, even for the large dataset it runs efficiently.
* It can also maintain accuracy when a large proportion of data is missing.

**Working:**

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

**Step-1:** Select random K data points from the training set.

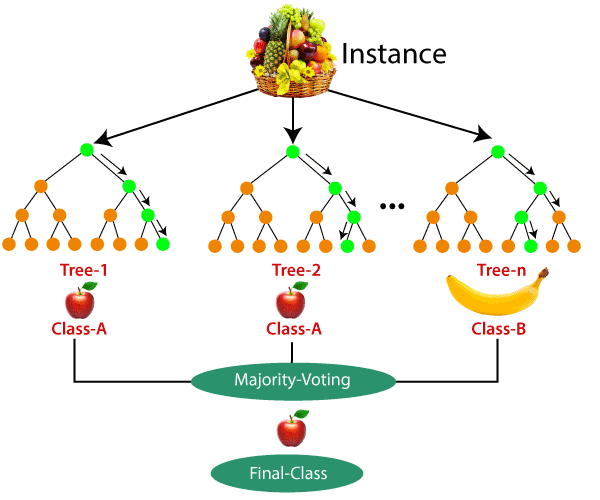
**Step-2:** Build the decision trees associated with the selected data points (Subsets).

**Step-3:** Choose the number N for decision trees that you want to build.

**Step-4:** Repeat Step 1 & 2.

**Step-5:** For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

**Example:** Suppose there is a dataset that contains multiple fruit images. So, this dataset is given to the Random forest classifier. The dataset is divided into subsets and given to each decision tree. During the training phase, each decision tree produces a prediction result, and when a new data point occurs, then based on the majority of results, the Random Forest classifier predicts the final decision. Consider the below image:



#### **5. Explain Support Vector Machine (SVM) algorithm with neat &amp; clean diagram.**

Support Vector Machine” (SVM) is a supervised machine learning algorithm that can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have)

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.

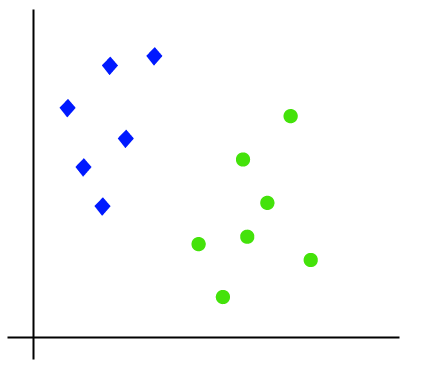
SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

**SVM can be of two types:**

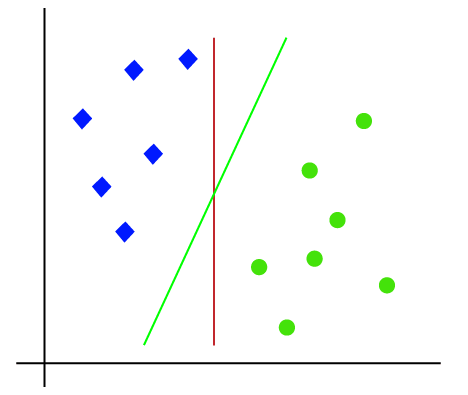
* **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
* **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

**Linear SVM:**

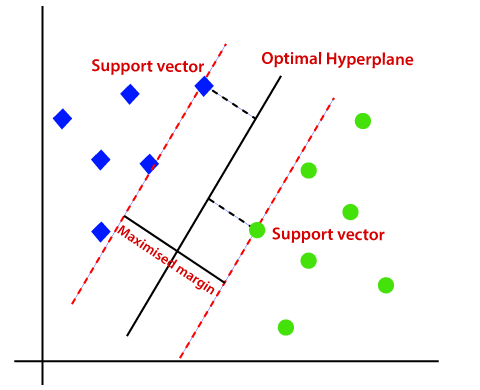
The working of the SVM algorithm can be understood by using an example. Suppose we have a dataset that has two tags (green and blue), and the dataset has two features x1 and x2. We want a classifier that can classify the pair(x1, x2) of coordinates in either green or blue. Consider the below image:



So as it is 2-d space so by just using a straight line, we can easily separate these two classes. But there can be multiple lines that can separate these classes. Consider the below image:

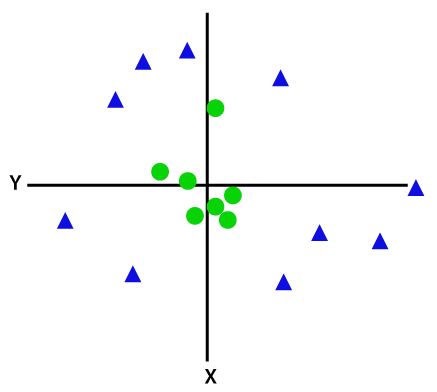


Hence, the SVM algorithm helps to find the best line or decision boundary; this best boundary or region is called as a **hyperplane**. SVM algorithm finds the closest point of the lines from both the classes. These points are called support vectors. The distance between the vectors and the hyperplane is called as **margin**. And the goal of SVM is to maximize this margin. The **hyperplane** with maximum margin is called the **optimal hyperplane**.



**Non-Linear SVM:**

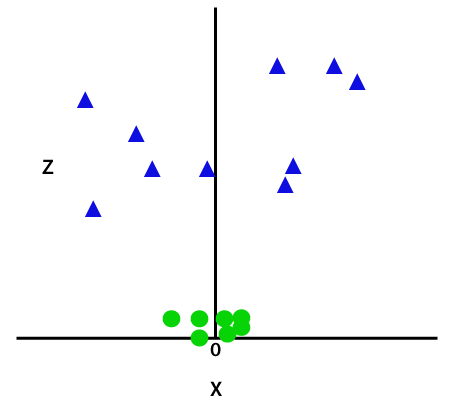
If data is linearly arranged, then we can separate it by using a straight line, but for non-linear data, we cannot draw a single straight line. Consider the below image:



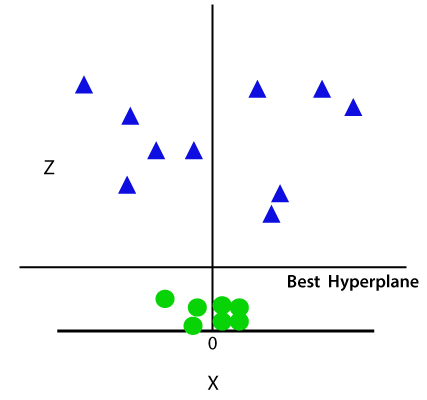
So to separate these data points, we need to add one more dimension. For linear data, we have used two dimensions x and y, so for non-linear data, we will add a third dimension z. It can be calculated as:

z=x2 +y2

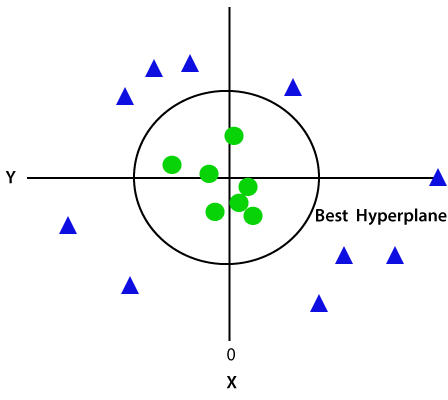
By adding the third dimension, the sample space will become as below image:



So now, SVM will divide the datasets into classes in the following way. Consider the below image:



Since we are in 3-d Space, hence it is looking like a plane parallel to the x-axis. If we convert it in 2d space with z=1, then it will become as:



Hence we get a circumference of radius 1 in case of non-linear data.

Unit -4

Q1

Ans

Data visualization is the representation of data through use of common graphics, such as charts, plots, infographics, and even animations. These visual displays of information communicate complex data relationships and data-driven insights in a way that is easy to understand

*With the help of data visualization, we can see how the data looks like and what kind of correlation is held by the attributes of data. It is the fastest way to see if the features correspond to the output. With the help of following Python recipes, we can understand ML data with statistics.*

*Data exploration refers to the initial step in data analysis in which data analysts use* [*data visualization*](https://www.omnisci.com/learn/data-visualization) *and statistical techniques to describe dataset characterizations, such as size, quantity, and accuracy, in order to better understand the nature of the data.*

A Machine Learning project is as good as the foundation of data on which it is built. In order to perform well, machine learning data exploration models must ingest large quantities of data, and model accuracy will suffer if that data is not thoroughly explored first. Data exploration steps to follow before building a machine learning model include:

* Variable identification: define each variable and its role in the dataset
* Univariate analysis: for continuous variables, build box plots or histograms for each variable independently; for categorical variables, build bar charts to show the frequencies
* Bi-variable analysis - determine the interaction between variables by building visualization tools
* ~Continuous and Continuous: scatter plots
* ~Categorical and Categorical: stacked column chart
* ~Categorical and Continuous: boxplots combined with swarmplots
* Detect and treat missing values
* Detect and treat outliers

Q2

Ans

Curse of Dimensionality refers to a set of problems that arise when working with high-dimensional data. The dimension of a dataset corresponds to the number of attributes/features that exist in a dataset. A dataset with a large number of attributes, generally of the order of a hundred or more, is referred to as high dimensional data. Some of the difficulties that come with high dimensional data manifest during analyzing or visualizing the data to identify patterns, and some manifest while training machine learning models. The difficulties related to training machine learning models due to high dimensional data is referred to as ‘Curse of Dimensionality’.

Anomaly detection is used for finding unforeseen items or events in the dataset. In high-dimensional data anomalies often show a remarkable number of attributes which are irrelevant in nature; certain objects occur more frequently in neighbour lists than others.

The curse of dimensionality occurs because the sample density decreases exponentially with the increase of the dimensionality. When we keep adding features without increasing the number of training samples as well, the dimensionality of the feature space grows and becomes sparser and sparser. Due to this sparsity, it becomes much easier to find a “perfect” solution for the machine learning model which highly likely leads to overfitting.

Dimensionality reduction is the process of reducing the dimensionality of the feature space with consideration by obtaining a set of principal features. Dimensionality reduction can be further broken into feature selection and feature extraction.

Q3

Ans

In machine learning classification problems, there are often too many factors on the basis of which the final classification is done. These factors are basically variables called features. The higher the number of features, the harder it gets to visualize the training set and then work on it. Sometimes, most of these features are correlated, and hence redundant. This is where dimensionality reduction algorithms come into play. Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables. It can be divided into feature selection and feature extraction.

The various methods used for dimensionality reduction include:

* Principal Component Analysis (PCA)
* Linear Discriminant Analysis (LDA)
* Generalized Discriminant Analysis (GDA)

**Advantages of Dimensionality Reduction**

* It helps in data compression, and hence reduced storage space.
* It reduces computation time.
* It also helps remove redundant features, if any.

**Disadvantages of Dimensionality Reduction**

* It may lead to some amount of data loss.
* PCA tends to find linear correlations between variables, which is sometimes undesirable.
* PCA fails in cases where mean and covariance are not enough to define datasets.
* We may not know how many principal components to keep- in practice, some thumb rules are applied.

Q4

Ans

Principal Component Analysis is an unsupervised learning algorithm that is used for the dimensionality reduction in [machine learning](https://www.javatpoint.com/machine-learning). It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation. These new transformed features are called the Principal Components. It is one of the popular tools that is used for exploratory data analysis and predictive modeling. It is a technique to draw strong patterns from the given dataset by reducing the variances.

PCA generally tries to find the lower-dimensional surface to project the high-dimensional data.

PCA works by considering the variance of each attribute because the high attribute shows the good split between the classes, and hence it reduces the dimensionality. Some real-world applications of PCA are ***image processing, movie recommendation system, optimizing the power allocation in various communication channels.*** It is a feature extraction technique, so it contains the important variables and drops the least important variable.

The PCA algorithm is based on some mathematical concepts such as:

* Variance and Covariance
* Eigenvalues and Eigen factors

Some common terms used in PCA algorithm:

* **Dimensionality:** It is the number of features or variables present in the given dataset. More easily, it is the number of columns present in the dataset.
* **Correlation:** It signifies that how strongly two variables are related to each other. Such as if one changes, the other variable also gets changed. The correlation value ranges from -1 to +1. Here, -1 occurs if variables are inversely proportional to each other, and +1 indicates that variables are directly proportional to each other.
* **Orthogonal:** It defines that variables are not correlated to each other, and hence the correlation between the pair of variables is zero.
* **Eigenvectors:** If there is a square matrix M, and a non-zero vector v is given. Then v will be eigenvector if Av is the scalar multiple of v.
* **Covariance Matrix:** A matrix containing the covariance between the pair of variables is called the Covariance Matrix.

### Principal Components in PCA

As described above, the transformed new features or the output of PCA are the Principal Components. The number of these PCs are either equal to or less than the original features present in the dataset. Some properties of these principal components are given below:

* The principal component must be the linear combination of the original features.
* These components are orthogonal, i.e., the correlation between a pair of variables is zero.
* The importance of each component decreases when going to 1 to n, it means the 1 PC has the most importance, and n PC will have the least importance.

Q5

Ans

# K-Nearest Neighbor(KNN) Algorithm for Machine Learning

* K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
* K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
* K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
* K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
* K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
* It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
* KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

The K-NN working can be explained on the basis of the below algorithm:

* **Step-1:** Select the number K of the neighbors
* **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
* **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
* **Step-4:** Among these k neighbors, count the number of the data points in each category.
* **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
* **Step-6:** Our model is ready.