1. Write short notes on:
   1. Ad boost

**Boosting** is an ensemble modeling technique that attempts to build a strong classifier from the number of weak classifiers. It is done by building a model by using weak models in series. Firstly, a model is built from the training data. Then the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models are added.

**AdaBoost** was the first really successful boosting algorithm developed for the purpose of binary classification. *AdaBoost* is short for *Adaptive Boosting* and is a very popular boosting technique that combines multiple “weak classifiers” into a single “strong classifier”.

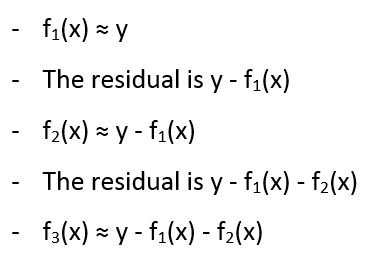
**Algorithm:** 

1. *Initialise the dataset and assign equal weight to each of the data point.*
2. *Provide this as input to the model and identify the wrongly classified data points.*
3. *Increase the weight of the wrongly classified data points.*
4. *if (got required results)   
     Goto step 5   
   else   
     Goto step 2*
5. *End*
   1. Gradient Tree Boosting

Boosting means combining a learning algorithm in series to achieve a strong learner from many sequentially connected weak learners. In case of gradient boosted decision trees algorithm, the weak learners are decision trees.

Each tree attempts to minimize the errors of previous tree. Trees in boosting are weak learners but adding many trees in series and each focusing on the errors from previous one make boosting a highly efficient and accurate model. Unlike bagging, boosting does not involve bootstrap sampling. Everytime a new tree is added, it fits on a modified version of initial dataset

Gradient boosting algorithm sequentially combines weak learners in way that each new learner fits to the residuals from the previous step so that the model improves. The final model aggregates the results from each step and a strong learner is achieved. **Gradient boosted decision trees** algorithm uses decision trees as week learners. A loss function is used to detect the residuals. For instance, mean squared error (MSE) can be used for a regression task and logarithmic loss (log loss) can be used for classification tasks. It is worth noting that existing trees in the model do not change when a new tree is added. The added decision tree fits the residuals from the current model. The steps are as follows:



* 1. Voting Classifier

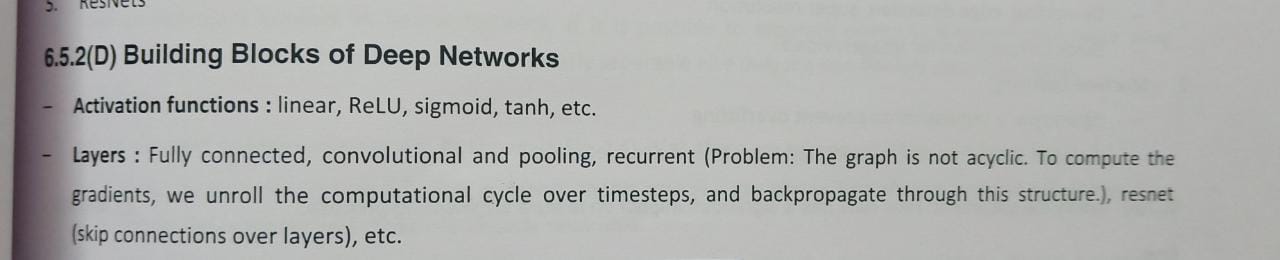
A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output.  
It simply aggregates the findings of each classifier passed into Voting Classifier and predicts the output class based on the highest majority of voting. The idea is instead of creating separate dedicated models and finding the accuracy for each them, we create a single model which trains by these models and predicts output based on their combined majority of voting for each output class.

* Hard Voting: Voting is calculated on the predicted output class.
* Soft Voting: Voting is calculated on the predicted probability of the output class.

1. With reference to Hierarchical Clustering, explain the issue of connectivity constraints.

In agglomerative clustering, we can restrict which clusters to join by adding connectivity constraints. These constraints specify which examples are considered connected and only clusters with connected examples, from one cluster to the other, can be joined into larger clusters. This helps solve some problems like Figure 19.5 illustrates. The left panel shows the result of agglomerative clustering without connectivity constraints. Since the linkage method used (Ward) takes into account only distances between the points, in order to minimize the SSE, the clusters include examples across the gap separating different stretches of the “ribbon” in which the data is structured. A connectivity constraint that restricts the connection of each example only to the 10 nearest neighbours creates a graph of connections that respects the structure of the data and prevents these inadequate clusters from forming.

1. What are building blocks of deep networks, elaborate.



**An Activation Function** decides whether a neuron should be activated or not. This means that it will decide whether the neuron’s input to the network is important or not in the process of prediction using simpler mathematical operations.

The role of the Activation Function is to derive output from a set of input values fed to a node (or a layer).

A convolution layer **transforms the input image in order to extract features from it** into convolution matrix or convolution mask.

Pooling layers **provide an approach to down sampling feature maps by summarizing the presence of features in patches of the feature map**.

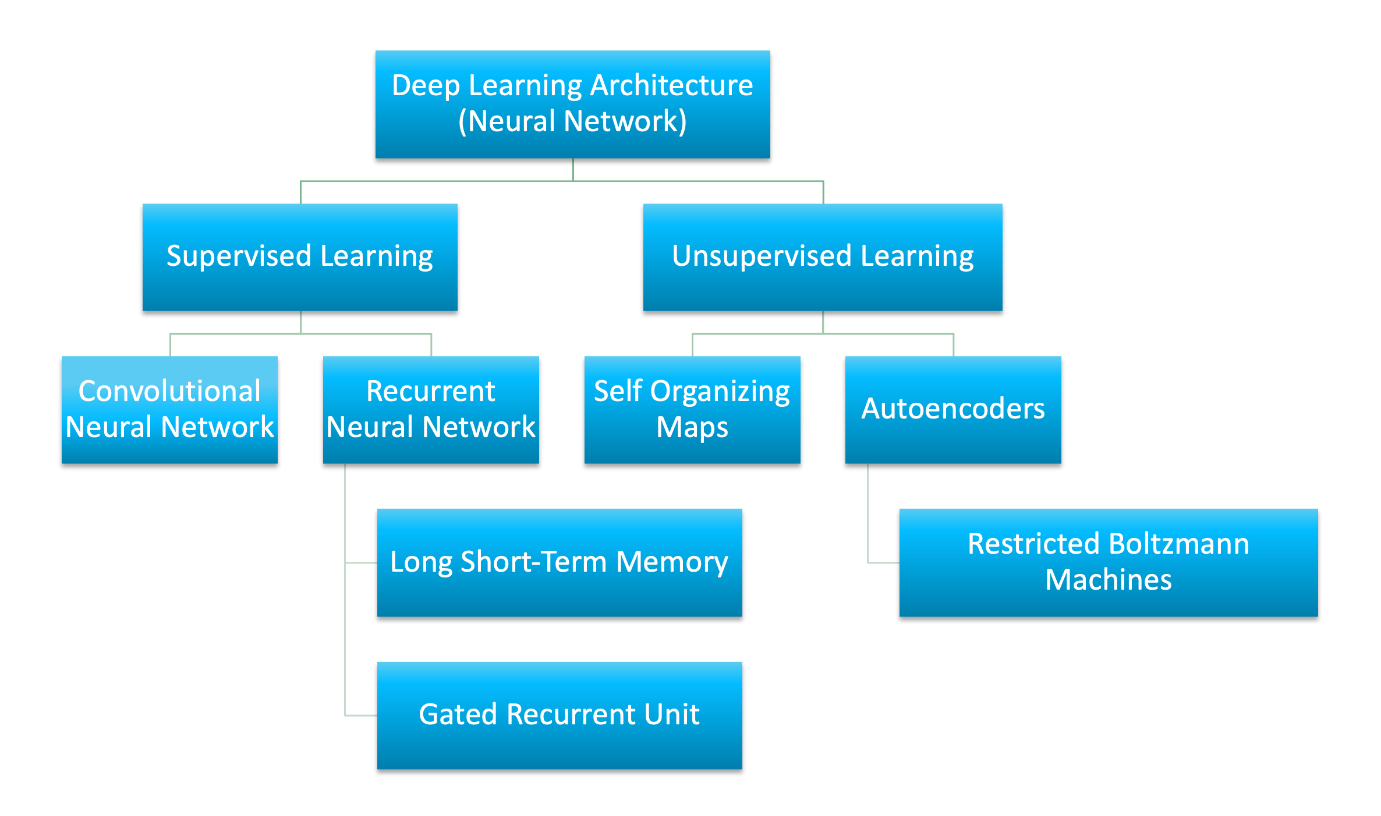
The input to the fully connected layer is the output from the final Pooling or Convolutional Layer, which is flattened and then fed into the fully connected layer.

Recurrent neural networks recognize data's sequential characteristics and use patterns to predict the next likely scenario.

ResNets(Residual networks) can easily gain accuracy from greatly increased depth, producing results which are better than previous networks

1. With reference to Deep Learning, Explain the concept of Deep Architectures.

Deep architectures are composed of multiple levels of non-linear operations, such as in neural nets with many hidden layers or in complicated propositional formulae re-using many sub-formulae. Searching the parameter space of deep architectures is a difficult task, but learning algorithms such as those for Deep Belief Networks have recently been proposed to tackle this problem with notable success, beating the state-of-the-art in certain areas



1. Justify with elaboration the following statement: The k-means algorithm is based on the strong initial condition to decide the Number of clusters through the assignment of ‘k’ initial centroids or means

**poor initialization can cause the algorithm to get stuck into an inferior local minimum**.

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

In traditional k-Means algorithm starting initial points are selected randomly and result of traditional kMeans is highly depend upon selection on initial centroids so In case of traditional k-Means one of the major weakness is different initial k objects may produce different clustering results.

1. What problems are faced by SVM when used with real datasets?

* SVM algorithm is not suitable for large data sets.

the complexity of the algorithm’s training is highly dependent on the size of the dataset. In other words, training time grows with the dataset to a point where it becomes infeasible to train and use due to compute constraints.

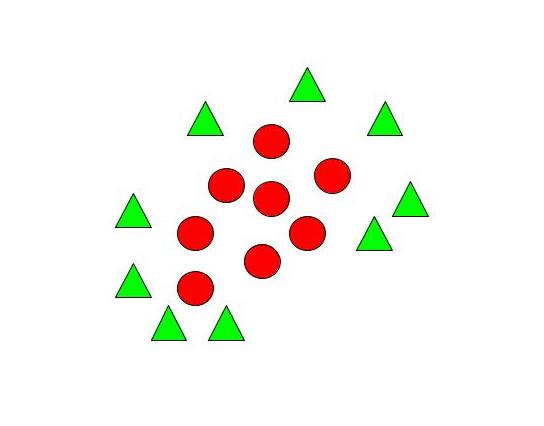
* SVM does not perform very well when the data set has more noise i.e. target classes are overlapping.
* In cases where the number of features for each data point exceeds the number of training data samples, the SVM will underperform.
* As the support vector classifier works by putting data points, above and below the classifying hyperplane there is no probabilistic explanation for the classification.

1. Explain the non-linear SVM with example

<https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm>

When we can easily separate data with hyperplane by drawing a straight line is Linear SVM. When we cannot separate data with a straight line we use **Non – Linear SVM**. In this, we have **Kernel functions.** They transform non-linear spaces into linear spaces. It transforms data into another dimension so that the data can be classified.

It transforms two variables x and y into three variables along with z. Therefore, the data have plotted from 2-D space to 3-D space. Now we can easily classify the data by drawing the best hyperplane between them



1. Write shorts notes on:

**Naive Bayes** is a supervised classification algorithm of Machine Learning based on Bayes theorem which gives the likelihood of occurrence of the event. Naive Bayes classifier is a probabilistic classifier which means that given an input, it predicts the probability of the input being classified for all the classes. It is also called conditional probability.

**P(A|B) = P(A) \* P(B|A)/P(B)**

Where we are calculating the probability of class A when predictor B is already provided.

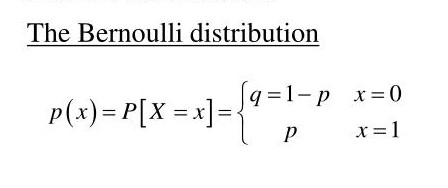
P(B) = prior probability of B

P(A) = prior probability of class A

P(B|A) = occurrence of predictor B given class A probability

* 1. Bernoulli naive Bayes.

This is used for discrete data and it works on Bernoulli distribution. The main feature of Bernoulli Naive Bayes is that it accepts features only as binary values like true or false, yes or no, success or failure, 0 or 1 and so on. So when the feature values are binary we know that we have to use Bernoulli Naive Bayes classifier.



* 1. multinomial naive Bayes.

Multimodal naive bayes is a specialized version of naive bayes designed to handle text documents using word counts as it's underlying method of calculating probability.

It's a simple but yet elegant model to handle classification that involve simple clsses that do not involve sentiment analysis (complex expressions of emotions such as sarcasm).  
Based on an arbitrary document that is within your set of classes, for example, you have a set of classes that represent subjects in school (mathematics, art, music, and etc..). You are given a document with words closely related to art because they used a lot of keywords but there could be some mathematics, history, and etc too but after going through all the classes and determining their probability it turns out it's most likely art.  
Even with low amounts of test data you can still produce an accurate or almost accurate classification model.

All of your classification results should be within your set of classe

* 1. Gaussian naive Bayes.

Gaussian Naive Bayes supports continuous valued features and models each as conforming to a Gaussian (normal) distribution.

An approach to create a simple model is to assume that the data is described by a Gaussian distribution with no co-variance (independent dimensions) between dimensions. This model can be fit by simply finding the mean and standard deviation of the points within each label, which is all what is needed to define such a distribution.

1. Define Bayes Theorem. Elaborate Naive Bayes Classifier working with example.

Bayes' theorem is also known as **Bayes' Rule** or **Bayes' law**, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.

The formula for Bayes' theorem is given as:

Naïve Bayes Classifier Algorithm

**P(A|B) is Posterior probability**: Probability of hypothesis A on the observed event B.

**P(B|A) is Likelihood probability**: Probability of the evidence given that the probability of a hypothesis is true.

**P(A) is Prior Probability**: Probability of hypothesis before observing the evidence.

**P(B) is Marginal Probability**: Probability of Evidence.

1. Convert the given dataset into frequency tables.
2. Generate Likelihood table by finding the probabilities of given features.
3. Now, use Bayes theorem to calculate the posterior probability.

<https://www.javatpoint.com/machine-learning-naive-bayes-classifier>

1. What are Linear support vector machines? Explain with example.

<https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm>

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.

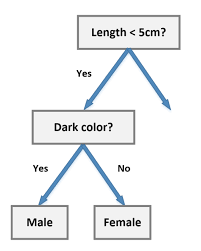
1. Explain with example the variant of SVM, the Support vector regression.

Support Vector Regression is a supervised learning algorithm that is used to predict discrete values. Support Vector Regression uses the same principle as the SVMs. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points.

Unlike other Regression models that try to minimize the error between the real and predicted value, the SVR tries to fit the best line within a threshold value. The threshold value is the distance between the hyperplane and boundary line. The fit time complexity of SVR is more than quadratic with the number of samples which makes it hard to scale to datasets with more than a couple of 10000 samples.

1. Explain the structure of binary decision tree for a sequential decision process.

A binary decision tree is a structure based on a sequential decision process. Starting from the root, a feature is evaluated and one of the two branches is selected. This procedure is repeated until a final leaf is reached, which normally represents the classification target we're looking for. Considering other algorithms, decision trees seem to be simpler in their dynamics; however, if the dataset is splittable while keeping an internal balance, the overall process is intuitive and rather fast in its predictions. Moreover, decision trees can work efficiently with unnormalized datasets because their internal structure is not influenced by the values assumed by each feature



1. With reference to Clustering, explain the issue of “Optimization of clusters”
2. Explain Evaluation methods for clustering algorithms.

Clustering is an unsupervised machine learning algorithm. It helps in clustering data points to groups. Validating the clustering algorithm is bit tricky compared to supervised machine learning algorithm as clustering process does not contain ground truth labels. If one want to do clustering with ground truth labels being present, validation methods and metrics of supervised machine learning algorithms can be used.

**How Clustering can be evaluated?**

Three important factors by which clustering can be evaluated are

(a) Clustering tendency - the data does not contain clustering tendency, then clusters identified by any state of the art clustering algorithms may be irrelevant. Non-uniform distribution of points in data set becomes important in clustering.

(b) Number of clusters, **k -** If k is too high, each point will broadly start representing a cluster and if k is too low, then data points are incorrectly clustered. Finding the optimal number of clusters leads to granularity in clustering

There is no definitive answer for finding right number of cluster as it depends upon (a) Distribution shape (b) scale in the data set (c) clustering resolution There are two major approaches to find optimal number of clusters:  
(1) Domain knowledge  
(2) Data driven approach

(c) Clustering quality- how well the clustering has performed can be quantified by a number of metrics. Extrinsic Measures Intrinsic Measures

1. With reference to Meta Classifiers, explain the concepts of Weak and eager learner.

**Weak Classifier**: A weak classifier is a model for binary classification that performs slightly better than random guessing. Formally, a classifier that achieves slightly better than 50 percent accuracy.

It is based on formal computational learning theory that proposes a class of learning methods that possess weakly learnability, meaning that they perform better than random guessing. Weak learnability is proposed as a simplification of the more desirable strong learnability, where a learnable achieved arbitrary good classification accuracy.

weak learning models:

* **k-Nearest Neighbors**, with k=1 operating on one or a subset of input variables.
* **Multi-Layer Perceptron**, with a single node operating on one or a subset of input variables.
* **Naive Bayes**, operating on a single input variable.
* **Strong Classifier**: A strong classifier is a model for binary classification that performs with arbitrary performance, much better than random guessing. Formally, a classifier that achieves arbitrarily good accuracy.
* Logistic Regression.
* Support Vector Machine.
* k-Nearest Neighbors.
* **Weak Learner**: Easy to prepare, but not desirable due to their low skill.
* **Strong Learner**: Hard to prepare, but desirable because of their high skill.