

## **1. What is the difference between supervised and unsupervised machine learning?**

### **Supervised Machine learning:**

Supervised machine learning requires training labelled data. Let's discuss it in bit detail, when we have

### **Unsupervised Machine learning:**

Unsupervised machine learning doesn't required labelled data.

## **2. What is bias, variance trade off ?**

### **Bias:**

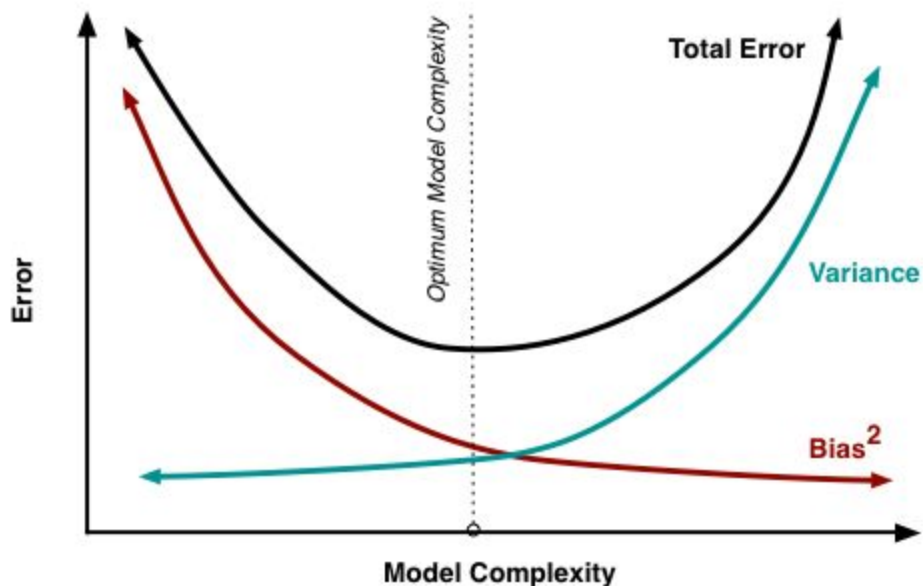
"Bias is error introduced in your model due to over simplification of machine learning algorithm." It can lead to under fitting. When you train your model at that time model makes simplified assumptions to make the target function easier to understand.

Low bias machine learning algorithms — Decision Trees, k-NN and SVM  
High bias machine learning algorithms — Linear Regression, Logistic Regression

### **Variance:**

"Variance is error introduced in your model due to complex machine learning algorithm, your model learns noise also from the training data set and performs bad on test data set." It can lead high sensitivity and over fitting.

Normally, as you increase the complexity of your model, you will see a reduction in error due to lower bias in the model. However, this only happens till a particular point. As you continue to make your model more complex, you end up over-fitting your model and hence your model will start suffering from high variance.



### **Bias, Variance trade off:**

The goal of any supervised machine learning algorithm is to have low bias and low variance to achieve good prediction performance.

1. The k-nearest neighbours algorithm has low bias and high variance, but the trade-off can be changed by increasing the value of k which increases the number of neighbours that contribute to the prediction and in turn increases the bias of the model.
2. The support vector machine algorithm has low bias and high variance, but the trade-off can be changed by increasing the C parameter that influences the number of violations of the margin allowed in the training data which increases the bias but decreases the variance.

There is no escaping the relationship between bias and variance in machine learning. Increasing the bias will decrease the variance. Increasing the variance will decrease the bias.

### 3. What is exploding gradients ?

#### Gradient:

Gradient is the **direction and magnitude** calculated during training of a neural network that is used to update the network weights in the right direction and by the right amount.

“Exploding gradients are a problem where **large error gradients** accumulate and result in very large updates to neural network model weights during training.” At an extreme, the values of weights can become so large as to overflow and result in NaN values.

This has the effect of your model being unstable and unable to learn from your training data. Now let's understand what is the gradient.

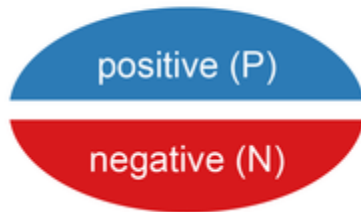
#### 4. What is a confusion matrix ?

The confusion matrix is a 2X2 table that contains 4 outputs provided by the **binary classifier**. Various measures, such as error-rate, accuracy, specificity, sensitivity, precision and recall are derived from it. *Confusion Matrix*

		Predicted class	
		$P$	$N$
Actual Class	$P$	True Positives (TP)	False Negatives (FN)
	$N$	False Positives (FP)	True Negatives (TN)

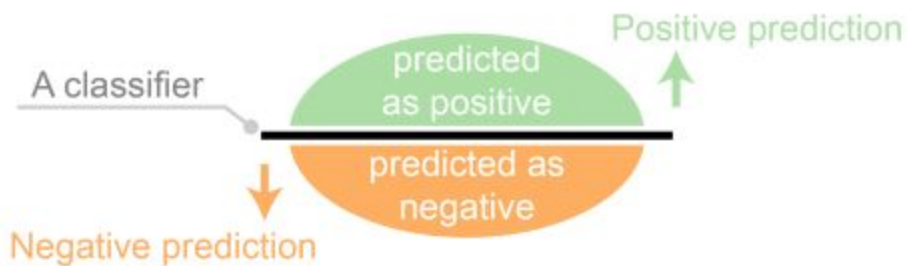
A data set used for performance evaluation is called test data set. It should contain the correct labels and predicted labels.

### Two actual classes or observed labels

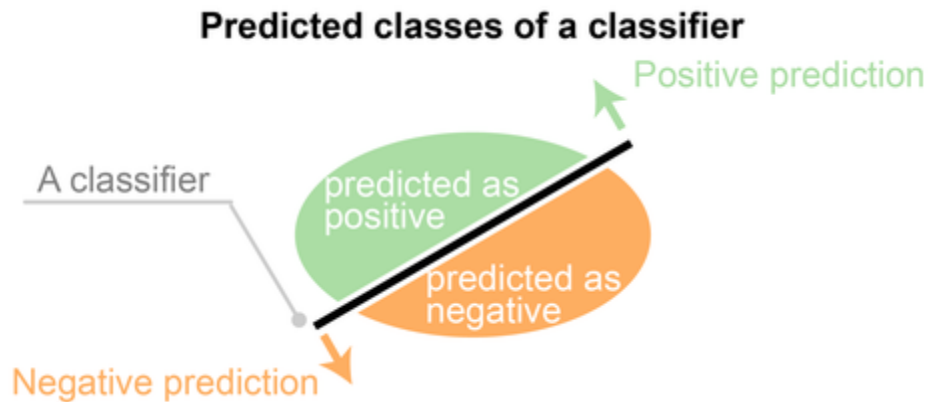


The predicted labels will exactly the same if the performance of a binary classifier is perfect.

### Predicted classes of a perfect classifier

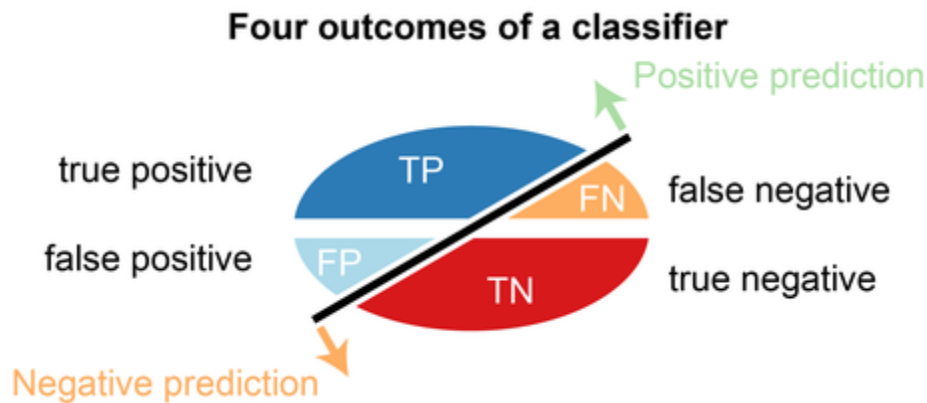


The predicted labels usually match with part of the observed labels in real world scenarios.



A binary classifier predicts all data instances of a test dataset as either positive or negative. This produces four outcomes-

1. True positive(TP) – Correct positive prediction
2. False positive(FP) – Incorrect positive prediction
3. True negative(TN) – Correct negative prediction
4. False negative(FN) – Incorrect negative prediction

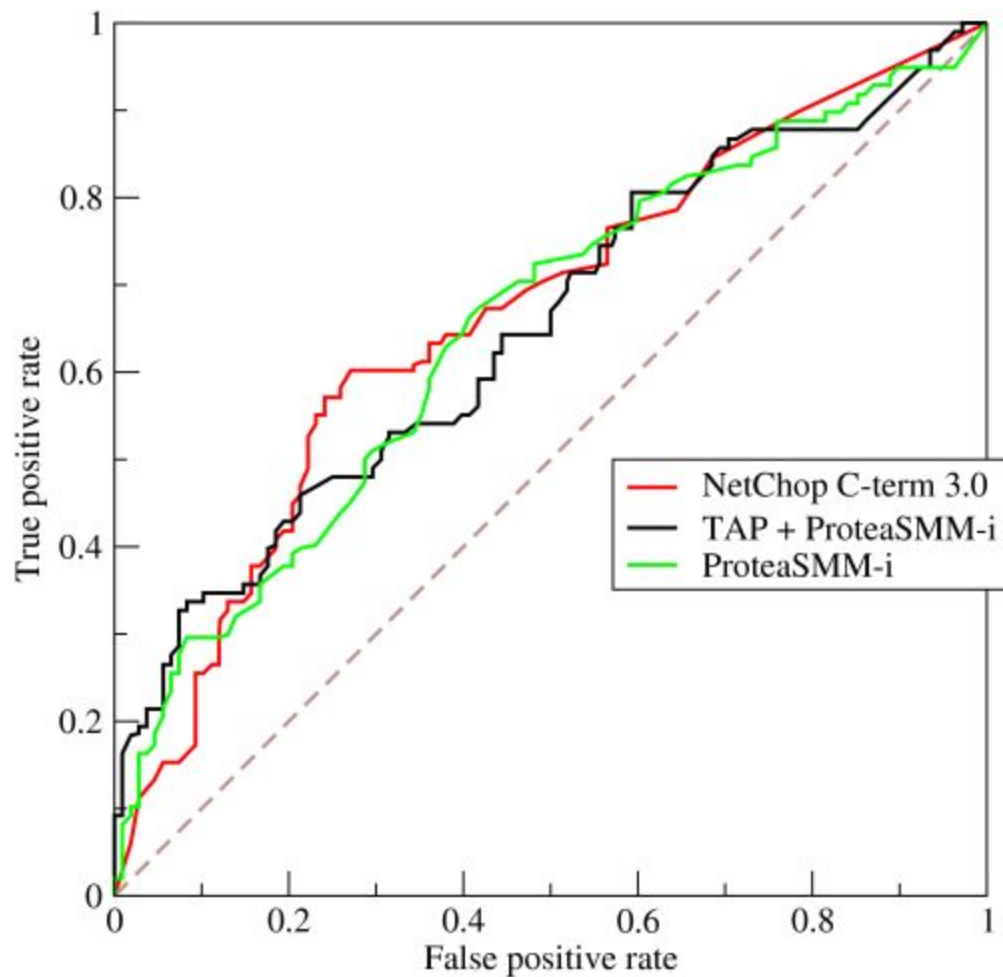


### Basic measures derived from the confusion matrix

1. Error Rate =  $(FP+FN)/(P+N)$
2. Accuracy =  $(TP+TN)/(P+N)$
3. Sensitivity(Recall or True positive rate) =  $TP/P$
4. Specificity(True negative rate) =  $TN/N$
5. Precision(Positive predicted value) =  $TP/(TP+FP)$
6. F-Score(Harmonic mean of precision and recall) =  $(1+b)(PREC.REC)/(b^2PREC+REC)$  where  $b$  is commonly 0.5, 1, 2.

### 6. Explain how a ROC curve works ?

The **ROC** curve is a graphical representation of the contrast between true positive rates and false positive rates at various thresholds. It is often used as a proxy for the trade-off between the sensitivity(true positive rate) and false positive rate.



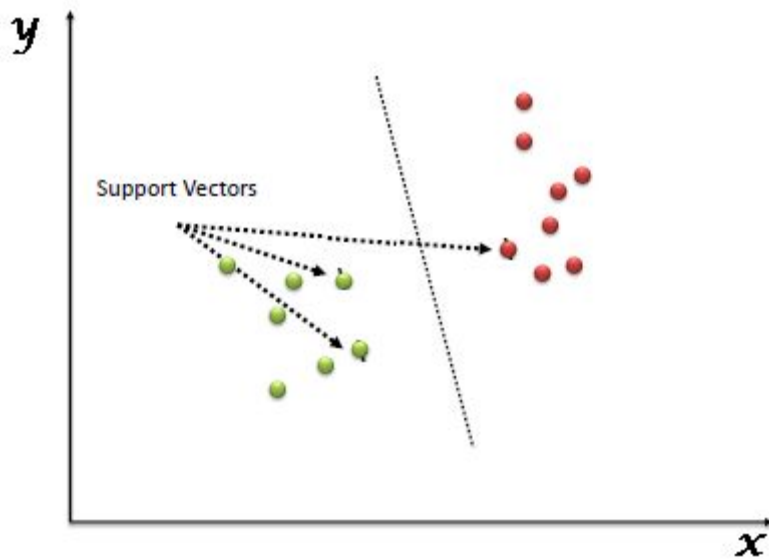
## 7. What is selection Bias ?

Selection bias occurs when sample obtained is not representative of the population intended to be analysed.

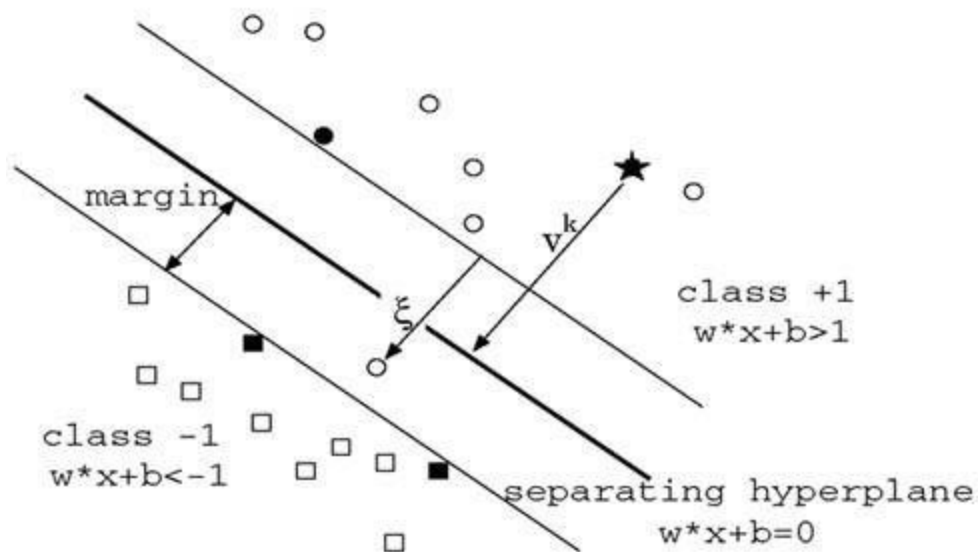
## 8. Explain SVM machine learning algorithm in detail.



SVM stands for support vector machine, it is a supervised machine learning algorithm which can be used for both **Regression and Classification**. If you have  $n$  features in your training data set, SVM tries to plot it in  $n$ -dimensional space with the value of each feature being the value of a particular coordinate. SVM uses hyper planes to separate out different classes based on the provided kernel function.



**9. What are support vectors in SVM.**



In the above diagram we see that the thinner lines mark the distance from the classifier to the closest data points called the support vectors (darkened data points). The distance between the two thin lines is called the margin.

## 10. What are the different kernels functions in SVM ?

There are four types of kernels in SVM.

1. Linear Kernel
2. Polynomial kernel
3. Radial basis kernel
4. Sigmoid kernel

## 11. Explain Decision Tree algorithm in detail.

Decision tree is a supervised machine learning algorithm mainly used for the **Regression and Classification**. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. Decision tree can handle both categorical and numerical data.

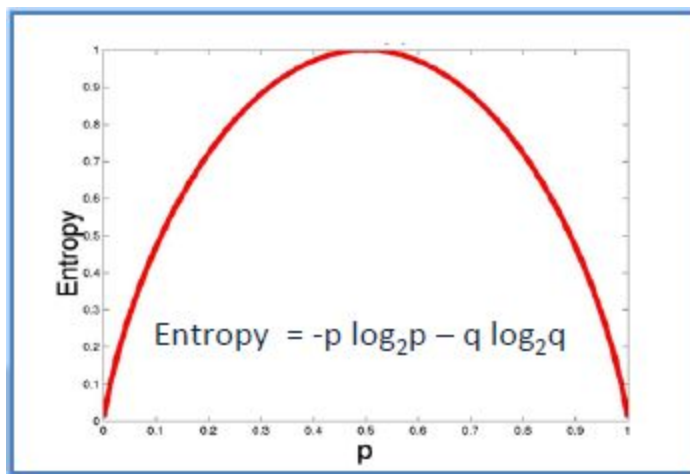


## 12. What is Entropy and Information gain in Decision tree algorithm ?

The core algorithm for building decision tree is called **ID3**. **ID3** uses **Entropy** and **Information Gain** to construct a decision tree.

### Entropy

A decision tree is built top-down from a root node and involve partitioning of data into homogenous subsets. **ID3** uses enteropy to check the homogeneity of a sample. If the sample is completely homogenous then entropy is zero and if the sample is an equally divided it has entropy of one.



$$\text{Entropy} = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$$

## Information Gain

The **Information Gain** is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attributes that returns the highest information gain.

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1
Gain = 0.029			

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1
Gain = 0.152			

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3
Gain = 0.048			

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

$$G(\text{PlayGolf}, \text{Outlook}) = E(\text{PlayGolf}) - E(\text{PlayGolf}, \text{Outlook})$$

$$= 0.940 - 0.693 = 0.247$$

13. What is pruning in Decision Tree ?

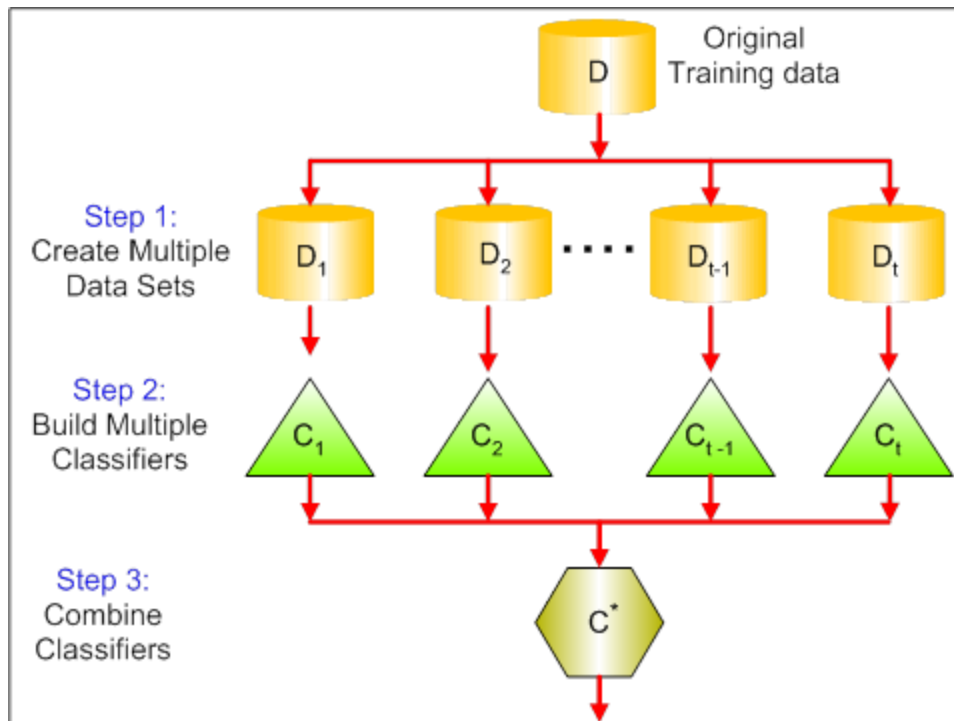
When we remove sub-nodes of a decision node, this process is called pruning or opposite process of splitting.

#### **14. What is Ensemble Learning ?**

Ensemble is the art of combining diverse set of learners(Individual models) together to improvise on the stability and predictive power of the model. Ensemble learning has many types but two more popular ensemble learning techniques are mentioned below.

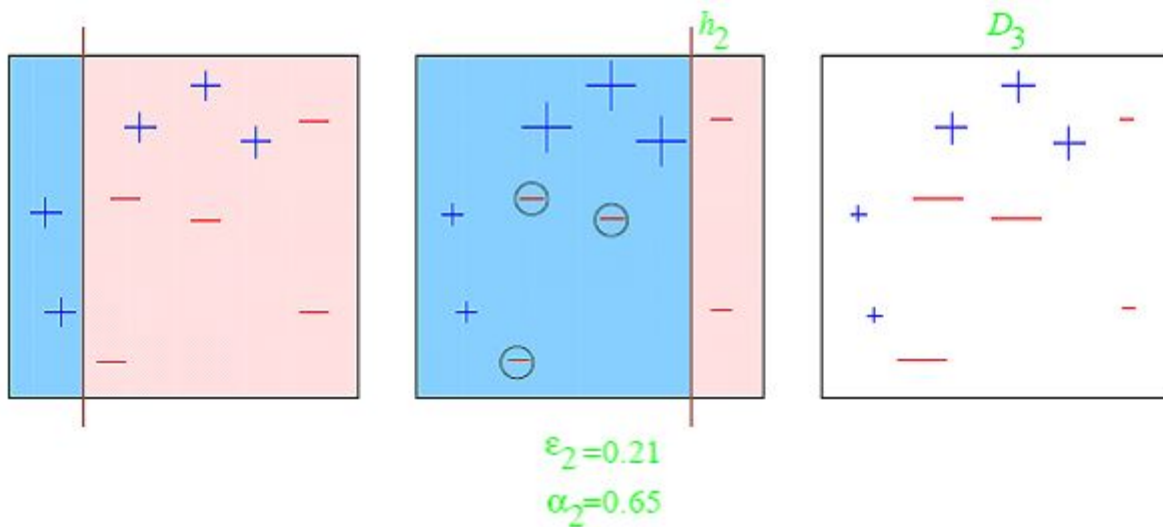
##### **Bagging**

Bagging tries to implement similar learners on small sample populations and then takes a mean of all the predictions. In generalised bagging, you can use different learners on different population. As you expect this helps us to reduce the variance error.



## Boosting

Boosting is an iterative technique which adjust the weight of an observation based on the last classification. If an observation was classified incorrectly, it tries to increase the weight of this observation and vice versa. Boosting in general decreases the bias error and builds strong predictive models. However, they may over fit on the training data.



### 15. What is Random Forest? How does it work ?

Random forest is a versatile machine learning method capable of performing both regression and classification tasks. It is also used for dimensionality reduction, treats missing values, outlier values. It is a type of ensemble learning method, where a group of weak models combine to form a powerful model.

In Random Forest, we grow multiple trees as opposed to a single tree. To classify a new object based on attributes, each tree gives a classification. The forest chooses the classification having the **most votes** (Over all the trees in the forest) and in case of regression, it takes the **average** of outputs by different trees.

### 16. What cross-validation technique would you use on a time series data set.



Instead of using k-fold cross-validation, you should be aware to the fact that a time series is not randomly distributed data — It is inherently ordered by chronological order.

In case of time series data, you should use techniques like forward chaining — Where you will be model on past data then look at forward-facing data.

fold 1: training[1], test[2]

fold 1: training[1 2], test[3]

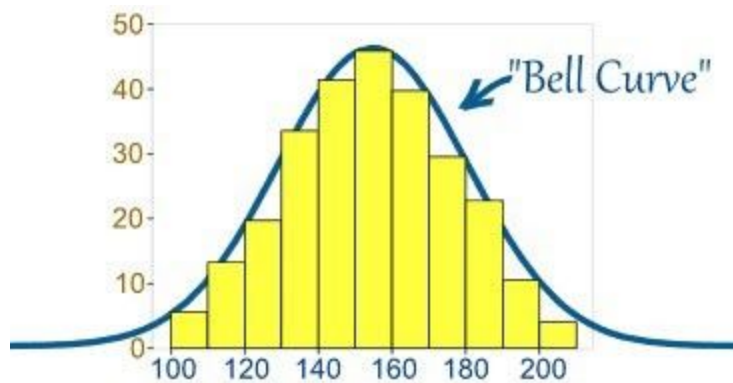
fold 1: training[1 2 3], test[4]

fold 1: training[1 2 3 4], test[5]

**17. What is logistic regression? Or State an example when you have used logistic regression recently.**

Logistic Regression often referred as logit model is a technique to predict the binary outcome from a linear combination of predictor variables. For example, if you want to predict whether a particular political leader will win the election or not. In this case, the outcome of prediction is binary i.e. 0 or 1 (Win/Lose). The predictor variables here would be the amount of money spent for election campaigning of a particular candidate, the amount of time spent in campaigning, etc.

**18. What do you understand by the term Normal Distribution?**

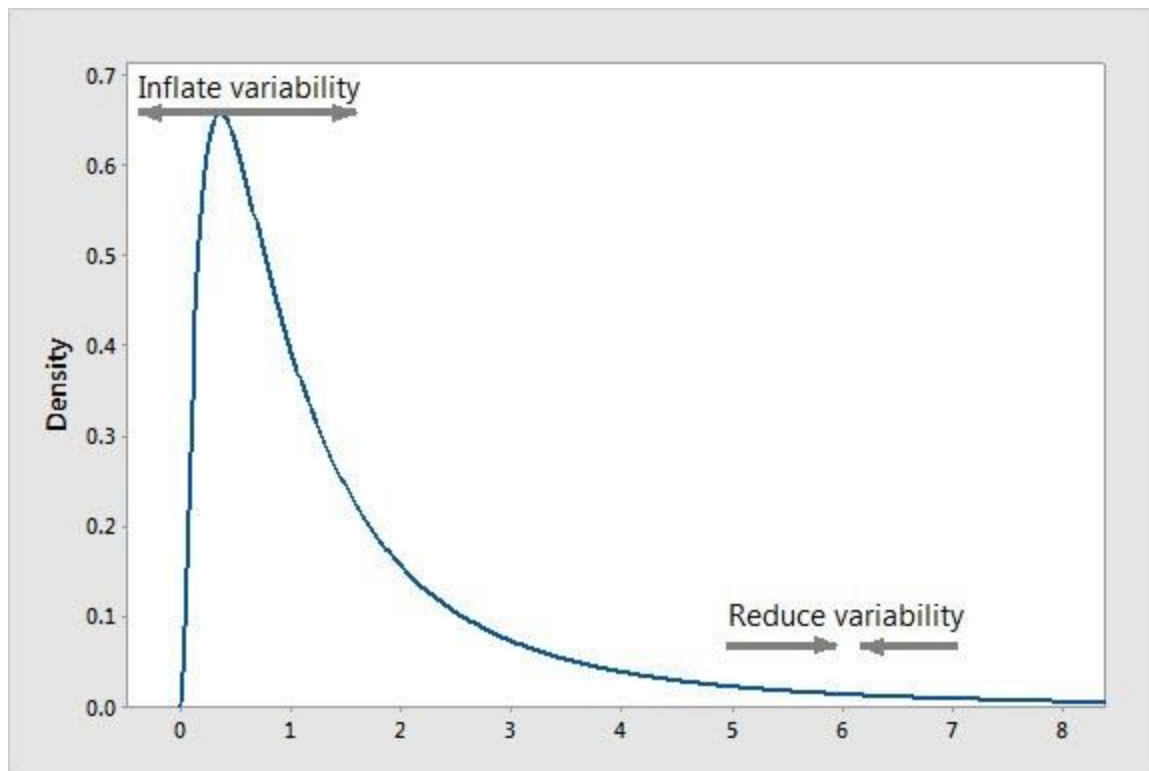


Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up. However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell shaped curve. The random variables are distributed in the form of an symmetrical bell shaped curve.

### 19. What is a Box Cox Transformation?

Dependent variable for a regression analysis might not satisfy one or more assumptions of an ordinary least squares regression. The residuals could either curve as the prediction increases or follow skewed distribution. In such scenarios, it is necessary to transform the response variable so that the data meets the required assumptions. A Box cox transformation is a statistical technique to transform non-normal dependent variables into a normal shape. If the given data is not normal then most of the statistical

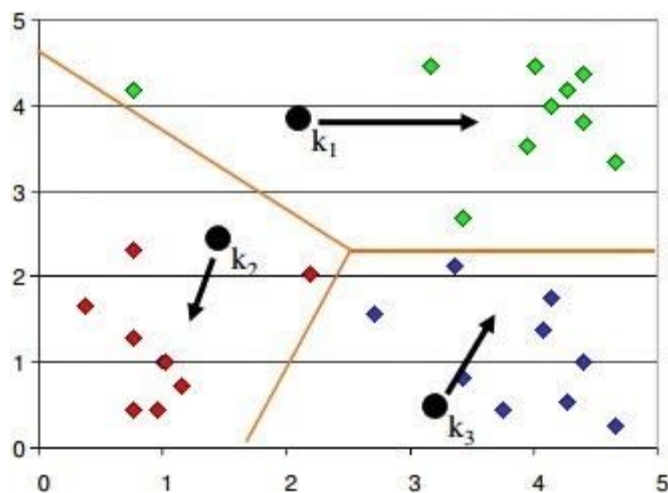
techniques assume normality. Applying a box cox transformation means that you can run a broader number of tests.



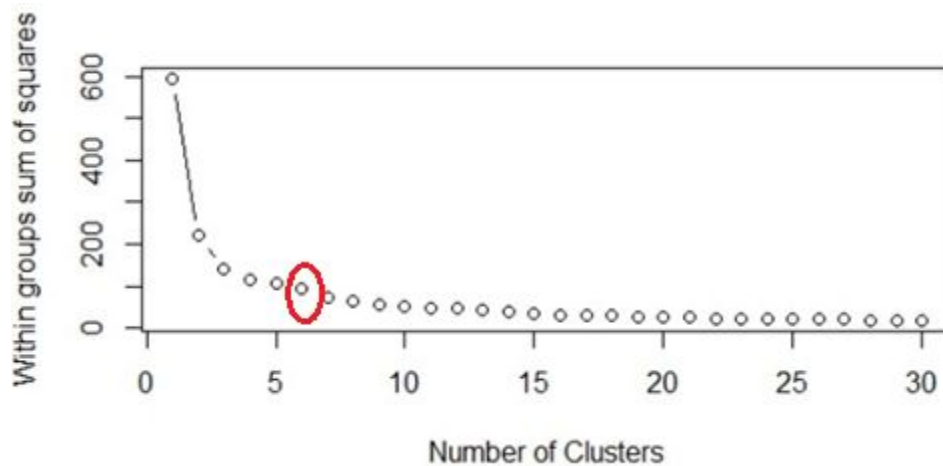
A Box Cox transformation is a way to transform non-normal dependent variables into a normal shape. Normality is an important assumption for many statistical techniques, if your data isn't normal, applying a Box-Cox means that you are able to run a broader number of tests. The Box Cox transformation is named after statisticians **George Box** and **Sir David Roxbee Cox** who collaborated on a 1964 paper and developed the technique.

## 20. How will you define the number of clusters in a clustering algorithm?

Though the Clustering Algorithm is not specified, this question will mostly be asked in reference to K-Means clustering where “K” defines the number of clusters. For example, the following image shows three different groups.



Within Sum of squares is generally used to explain the homogeneity within a cluster. If you plot WSS for a range of number of clusters, you will get the plot shown below. The Graph is generally known as Elbow Curve.

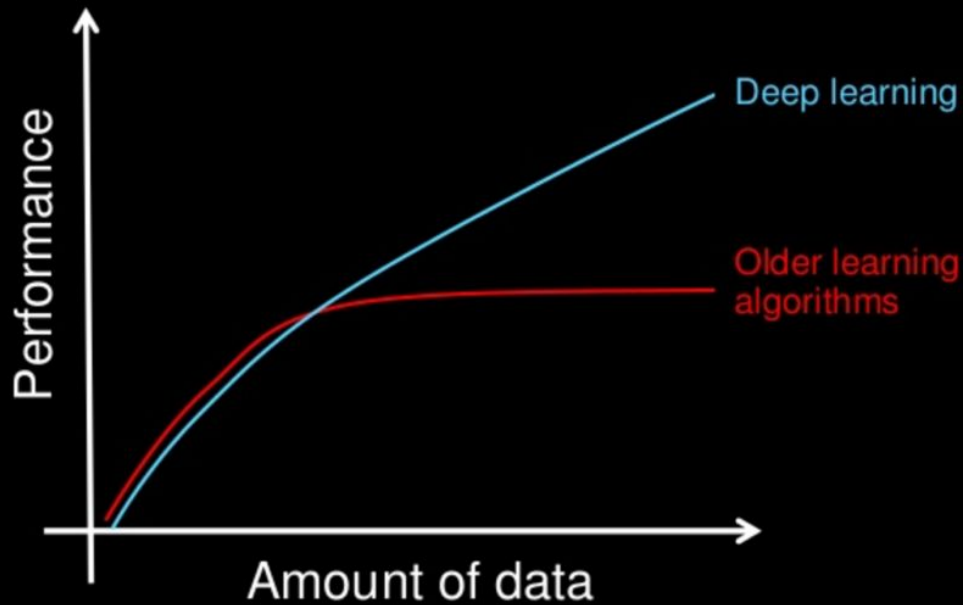


Red circled point in above graph i.e. Number of Cluster =6 is the point after which you don't see any decrement in WSS. This point is known as bending point and taken as K in K – Means. This is the widely used approach but few data scientists also use Hierarchical clustering first to create dendograms and identify the distinct groups from there.

## 21. What is deep learning?

Deep learning is sub field of machine learning inspired by structure and function of brain called artificial neural network. We have a lot numbers of algorithms under machine learning like Linear regression, SVM, Neural network etc and deep learning is just an extension of Neural networks. In neural nets we consider small number of hidden layers but when it comes to deep learning algorithms we consider a huge number of hidden layers to better understand the input output relationship.

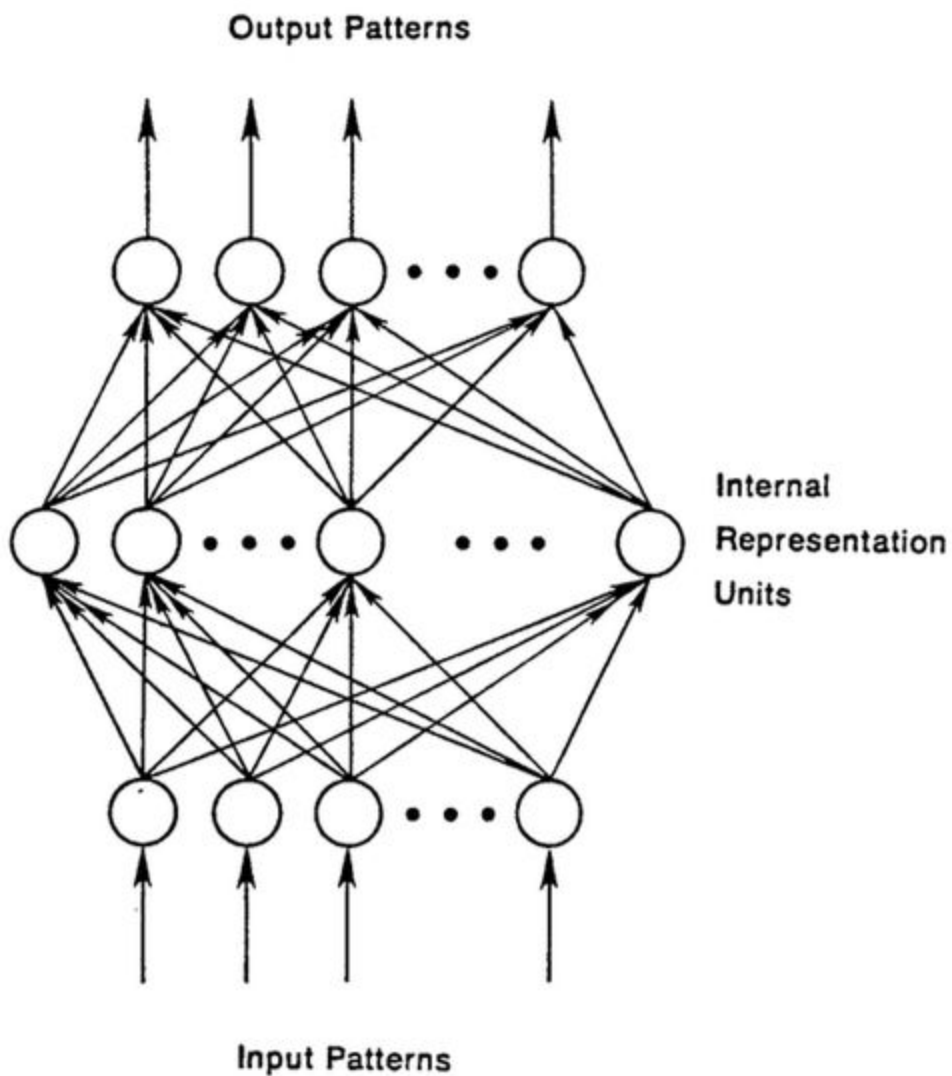
# Why deep learning



How do data science techniques scale with amount of data?

## 22. What are Recurrent Neural Networks(RNNs) ?

Recurrent nets are type of artificial neural networks designed to recognise pattern from the sequence of data such as Time series, stock market and government agencies etc. To understand recurrent nets, first you have to understand the basics of feed forward nets. Both these networks RNN and feed forward named after the way they channel information through a series of mathematical orations performed at the nodes of the network. One feeds information through straight(never touching same node twice), while the other cycles it through loop, and the latter are called recurrent.



Recurrent networks on the other hand, take as their input not just the current input example they see, but also the what they have perceived previously in time. The BTSXPE at the bottom of the drawing represents the input example in the current moment, and CONTEXT UNIT represents the output of the previous moment. The decision a recurrent

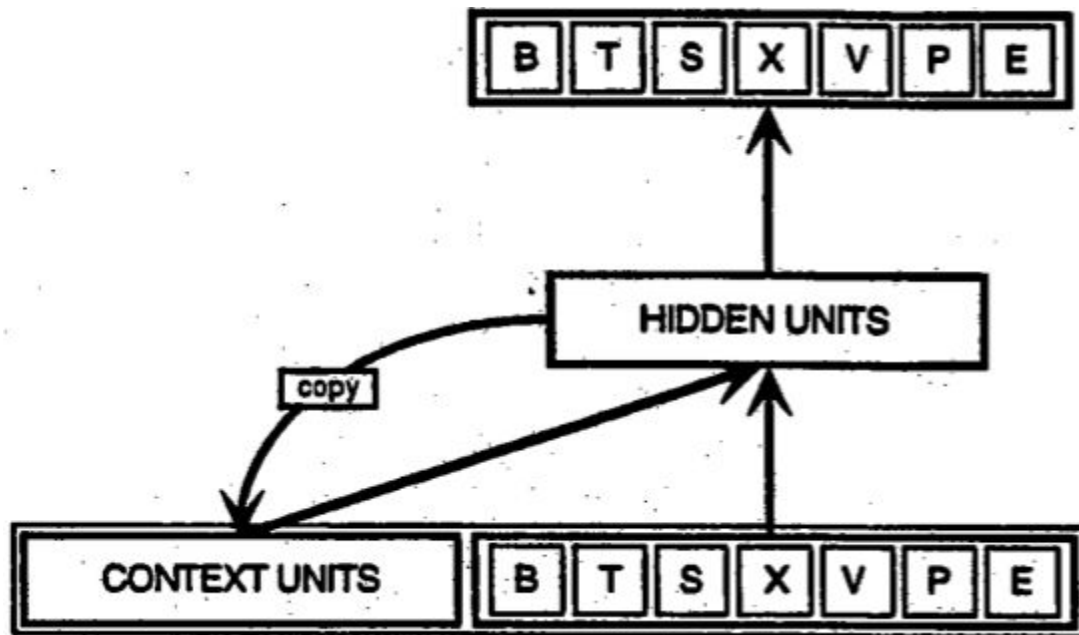
neural network reached at time  $t-1$  affects the decision that it will reach one moment later at time  $t$ . So recurrent networks have two sources of input, the present and the recent past, which combine to determine how they respond to new data, much as we do in life.

The error they generate will return via back propagation and be used to adjust their weights until error can't go any lower. Remember, the purpose of recurrent nets is to accurately classify sequential input. We rely on the back propagation of error and gradient descent to do so.

Back propagation in feed forward networks moves backward from the final error through the outputs, weights and inputs of each hidden layer, assigning those weights responsibility for a portion of the error by calculating their partial derivatives —  $\partial E / \partial w$ , or the relationship between their rates of change. Those derivatives are then used by our learning rule, gradient descent, to adjust the weights up or down, whichever direction decreases error.

Recurrent networks rely on an extension of back propagation called back propagation through time, or BPTT. Time, in this case, is simply expressed by a well-defined, ordered series of calculations linking one time step to the next, which is all back propagation needs to work.





**23. What is the difference between machine learning and deep learning?**

**Machine learning:**

Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed. Machine learning can be categorised in following three categories.

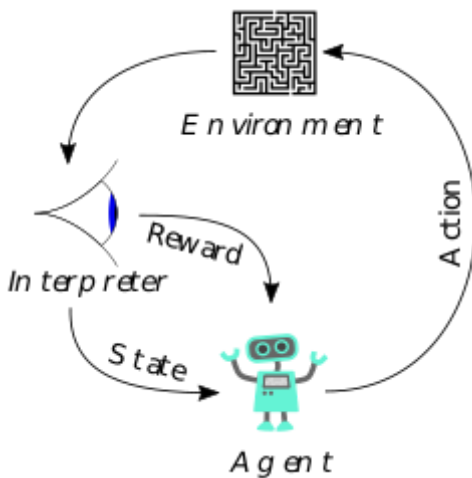
1. Supervised machine learning,
2. Unsupervised machine learning,
3. Reinforcement learning

**Deep learning:**

Deep Learning is a sub field of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks.

## 24. What is reinforcement learning ?

### Reinforcement learning



Reinforcement Learning is learning what to do and how to map situations to actions.

The end result is to maximise the numerical reward signal. The learner is not told which action to take, but instead must discover which action will yield the maximum reward. Reinforcement learning is inspired by the learning of human beings, it is based on the reward/punishment mechanism.

## **25. What is selection bias ?**

Selection bias is the bias introduced by the selection of individuals, groups or data for analysis in such a way that proper randomisation is not achieved, thereby ensuring that the sample obtained is not representative of the population intended to be analysed. It is sometimes referred to as the selection effect. The phrase “selection bias” most often refers to the distortion of a statistical analysis, resulting from the method of collecting samples. If the selection bias is not taken into account, then some conclusions of the study may not be accurate.

## **26. Explain what regularisation is and why it is useful.**

Regularisation is the process of adding tuning parameter to a model to induce smoothness in order to prevent overfitting. This is most often done by adding a constant multiple to an existing weight vector. This constant is often the L1(Lasso) or L2(ridge). The model predictions should then minimize the loss function calculated on the regularized training set.

## **27. What is TF/IDF vectorization ?**

tf-idf is short for term frequency-inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining. The tf-idf value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

## **28. What are Recommender Systems?**

A subclass of information filtering systems that are meant to predict the preferences or ratings that a user would give to a product. Recommender systems are widely used in movies, news, research articles, products, social tags, music, etc.

## **29. What is the difference between Regression and classification ML techniques.**

Both Regression and classification machine learning techniques come under **Supervised machine learning algorithms**. In Supervised machine learning algorithm, we have to train the model using labelled data set, While training we have to explicitly provide the correct labels and algorithm tries to learn the pattern from input to output. If our labels are discrete values then it will be a classification problem, e.g A,B etc. but if our labels are continuous values then it will be a regression problem, e.g 1.23, 1.333 etc.

## **30. If you are having 4GB RAM in your machine and you want to train your model on 10GB data set. How would you go about this problem. Have you ever faced this kind of problem in your machine learning/data science experience so far ?**

First of all you have to ask which ML model you want to train.

**For Neural networks:** Batch size with Numpy array will work.

**Steps:**

1. Load the whole data in Numpy array. Numpy array has property to create mapping of complete data set, it doesn't load complete data set in memory.
2. You can pass index to Numpy array to get required data.
3. Use this data to pass to Neural network.
4. Have small batch size.

**For SVM:** Partial fit will work

**Steps:**

1. Divide one big data set in small size data sets.
2. Use partial fit method of SVM, it requires subset of complete data set.
3. Repeat step 2 for other subsets.

### **31. What is p-value?**

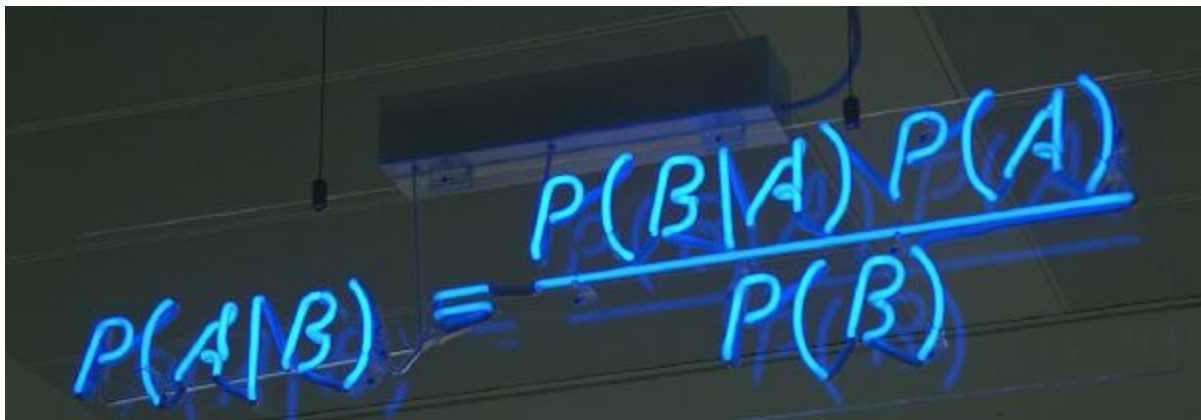
When you perform a hypothesis test in statistics, a p-value can help you determine the strength of your results. p-value is a number between 0 and 1. Based on the value it will denote the strength of the results. The claim which is on trial is called Null Hypothesis.

Low p-value ( $\leq 0.05$ ) indicates strength against the null hypothesis which means we can reject the null Hypothesis. High p-value ( $\geq 0.05$ ) indicates strength for the null hypothesis which means we can accept the null Hypothesis p-value of 0.05 indicates the Hypothesis could go either way. To put it in another way,

High P values: your data are likely with a true null. Low P values: your data are unlikely with a true null.

### 32. What is 'Naive' in a Naive Bayes ?

The Naive Bayes Algorithm is based on the Bayes Theorem. Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

A photograph of a chalkboard with the Bayes' Theorem formula written in blue chalk. The formula is  $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ . The chalkboard has a dark surface, and the lighting is somewhat dim, with the chalk being the primary source of light in the image.
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

### What is Naive ?

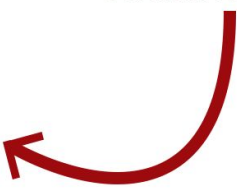
The Algorithm is 'naive' because it makes assumptions that may or may not turn out to be correct.

### 33. Why we generally use Softmax non-linearity function as last operation in network ?

It is because it takes in a vector of real numbers and returns a probability distribution. Its definition is as follows. Let  $x$  be a vector of real numbers (positive, negative, whatever, there are no constraints). Then the  $i$ 'th component of  $\text{Softmax}(x)$  is –

$$P(y=j \mid \theta^{(i)}) = \frac{e^{\theta_j^{(i)}}}{\sum_{k=0}^k e^{\theta_k^{(i)}} \quad \text{Softmax function}$$

where  $\theta = w_0x_0 + w_1x_1 + \dots + w_kx_k = \sum_{i=0}^k w_ix_i = w^T x$



It should be clear that the output is a probability distribution: each element is non-negative and the sum over all components is 1.

### 34. What are different ranking algorithms?

Traditional ML algorithms solve a prediction problem (classification or regression) on a single instance at a time. E.g. if you are doing spam detection on email, you will look at all the features associated with that email and classify it as spam or not. The aim of

traditional ML is to come up with a class (spam or no-spam) or a single numerical score for that instance.

Ranking algorithms like LTR solves a ranking problem on a list of items. The aim of LTR is to come up with optimal ordering of those items. As such, LTR doesn't care much about the exact score that each item gets, but cares more about the relative ordering among all the items. **RankNet**, **LambdaRank** and **LambdaMART** are all LTR algorithms developed by Chris Burges and his colleagues at Microsoft Research.

1. **RankNet** — The cost function for RankNet aims to minimize the number of inversions in ranking. RankNet optimizes the cost function using Stochastic Gradient Descent.
2. **LambdaRank** — Burgess et. al. found that during RankNet training procedure, you don't need the costs, only need the gradients ( $\lambda$ ) of the cost with respect to the model score. You can think of these gradients as little arrows attached to each document in the ranked list, indicating the direction we'd like those documents to move. Further they found that scaling the gradients by the change in NDCG found by swapping each pair of documents gave good results. The core idea of LambdaRank is to use this new cost function for training a RankNet. On experimental datasets, this shows both speed and accuracy improvements over the original RankNet.
3. **LambdaMart** — LambdaMART combines LambdaRank and MART (Multiple Additive Regression Trees). While MART uses gradient boosted decision trees for prediction tasks, LambdaMART uses gradient boosted decision trees using a cost function derived from LambdaRank for solving a ranking task. On experimental datasets, LambdaMART has shown better results than LambdaRank and the original RankNet.



### **35. What is the Difference between Ridge and Lasso Regularisation ?**

Ridge and Lasso regression uses two different penalty functions. Ridge uses  $l_2$  where as lasso go with  $l_1$ . In ridge regression, the penalty is the sum of the squares of the coefficients and for the Lasso, it's the sum of the absolute values of the coefficients. It's a shrinkage towards zero using an absolute value ( $l_1$  penalty) rather than a sum of squares ( $l_2$  penalty).

As we know that ridge regression can't zero coefficients. Here, you either select all the coefficients or none of them whereas LASSO does both parameter shrinkage and variable selection automatically because it zero out the co-efficients of collinear variables. Here it helps to select the variable(s) out of given  $n$  variables while performing lasso regression.

### **36. why logodds is important in logistic Regression ?**

### **37. what is the need of Label Encoding Explain ?**

### **38. What is cost function of SVM ?**

### **39. Gradient descent vs. best fit line ?**

### **40. What is a reinforcement algorithm ?**

### **41 How can you generate a random number between 1 – 7 with only a die?**

- Any die has six sides from 1-6. There is no way to get seven equal outcomes from a single rolling of a die. If we roll the die twice and consider the event of two rolls, we now have 36 different outcomes.
- To get our 7 equal outcomes we have to reduce this 36 to a number divisible by 7. We can thus consider only 35 outcomes and exclude the other one.

- A simple scenario can be to exclude the combination (6,6), i.e., to roll the die again if 6 appears twice.
- All the remaining combinations from (1,1) till (6,5) can be divided into 7 parts of 5 each. This way all the seven sets of outcomes are equally likely.

**42. A certain couple tells you that they have two children, at least one of which is a girl. What is the probability that they have two girls?**

In the case of two children, there are 4 equally likely possibilities

BB, BG, GB and GG;

where B = Boy and G = Girl and the first letter denotes the first child.

From the question, we can exclude the first case of BB. Thus from the remaining 3 possibilities of BG, GB & BB, we have to find the probability of the case with two girls.

Thus,  $P(\text{Having two girls given one girl}) = 1 / 3$

**43 . A jar has 1000 coins, of which 999 are fair and 1 is double headed. Pick a coin at random, and toss it 10 times. Given that you see 10 heads, what is the probability that the next toss of that coin is also a head?**

There are two ways of choosing the coin. One is to pick a fair coin and the other is to pick the one with two heads.

Probability of selecting fair coin =  $999/1000 = 0.999$

Probability of selecting unfair coin =  $1/1000 = 0.001$

Selecting 10 heads in a row = Selecting fair coin \* Getting 10 heads + Selecting an unfair coin

$$P(A) = 0.999 * (1/2)^5 = 0.999 * (1/1024) = 0.000976$$

$$P(B) = 0.001 * 1 = 0.001$$

$$P(A / A + B) = 0.000976 / (0.000976 + 0.001) = 0.4939$$

$$P(B / A + B) = 0.001 / 0.001976 = 0.5061$$

$$\text{Probability of selecting another head} = P(A/A+B) * 0.5 + P(B/A+B) * 1 = 0.4939 * 0.5 + 0.5061 = 0.7531$$

Python or R – Which one would you prefer for text analytics?

We will prefer Python because of the following reasons:

- *Python* would be the best option because it has Pandas library that provides easy to use data structures and high-performance data analysis tools.
- *R* is more suitable for machine learning than just text analysis.
- Python performs faster for all types of text analytics.

#### **45 . How does data cleaning plays a vital role in the analysis?**

Data cleaning can help in analysis because:

- Cleaning data from multiple sources helps to transform it into a format that data analysts or data scientists can work with.
- Data Cleaning helps to increase the accuracy of the model in machine learning.

- It is a cumbersome process because as the number of data sources increases, the time taken to clean the data increases exponentially due to the number of sources and the volume of data generated by these sources.
- It might take up to 80% of the time for just cleaning data making it a critical part of the analysis task.

#### **46. Differentiate between univariate, bivariate and multivariate analysis.**

*Univariate analyses* are descriptive statistical analysis techniques which can be differentiated based on the number of variables involved at a given point of time. For example, the pie charts of sales based on territory involve only one variable and can the analysis can be referred to as univariate analysis.

*The bivariate analysis* attempts to understand the difference between two variables at a time as in a scatterplot. For example, analyzing the volume of sale and spending can be considered as an example of bivariate analysis.

*Multivariate analysis* deals with the study of more than two variables to understand the effect of variables on the responses.

#### **47. Explain Star Schema.**

It is a traditional database schema with a central table. Satellite tables map IDs to physical names or descriptions and can be connected to the central fact table using the ID fields; these tables are known as lookup tables and are principally useful in real-time applications, as they save a lot of memory. Sometimes star schemas involve several layers of summarization to recover information faster.

#### **48 . What is Cluster Sampling?**

Cluster sampling is a technique used when it becomes difficult to study the target population spread across a wide area and simple random sampling cannot be applied. Cluster Sample is a probability sample where each sampling unit is a collection or cluster of elements.

For eg., A researcher wants to survey the academic performance of high school students in Japan. He can divide the entire population of Japan into different clusters (cities). Then the researcher selects a number of clusters depending on his research through simple or systematic random sampling.

Let's continue our Data Science Interview Questions blog with some more statistics questions.

#### **49. What is Systematic Sampling?**

*Systematic sampling* is a statistical technique where elements are selected from an ordered sampling frame. In systematic sampling, the list is progressed in a circular manner so once you reach the end of the list, it is progressed from the top again. The best example of systematic sampling is equal probability method.

#### **50. What are Eigenvectors and Eigenvalues?**

*Eigenvectors* are used for understanding linear transformations. In data analysis, we usually calculate the eigenvectors for a correlation or covariance matrix. Eigenvectors are the directions along which a particular linear transformation acts by flipping, compressing or stretching.

*Eigenvalue* can be referred to as the strength of the transformation in the direction of eigenvector or the factor by which the compression occurs.

**51. Can you cite some examples where a false positive is important than a false negative?**

Let us first understand what false positives and false negatives are.

- False Positives are the cases where you wrongly classified a non-event as an event a.k.a Type I error.
- False Negatives are the cases where you wrongly classify events as non-events, a.k.a Type II error.

*Example 1:* In the medical field, assume you have to give chemotherapy to patients. Assume a patient comes to that hospital and he is tested positive for cancer, based on the lab prediction but he actually doesn't have cancer. This is a case of false positive. Here it is of utmost danger to start chemotherapy on this patient when he actually does not have cancer. In the absence of cancerous cell, chemotherapy will do certain damage to his normal healthy cells and might lead to severe diseases, even cancer.

*Example 2:* Let's say an e-commerce company decided to give \$1000 Gift voucher to the customers whom they assume to purchase at least \$10,000 worth of items. They send free voucher mail directly to 100 customers without any minimum purchase condition because they assume to make at least 20% profit on sold items above \$10,000. Now the issue is if we send the \$1000 gift vouchers to customers who have not actually purchased anything but are marked as having made \$10,000 worth of purchase.

**52. Can you cite some examples where a false negative is important than a false positive?**

*Example 1:* Assume there is an airport 'A' which has received high-security threats and based on certain characteristics they identify whether a particular passenger can be a threat or not. Due to a shortage of staff, they decide to scan passengers being predicted as risk positives by their predictive model. What will happen if a true threat customer is being flagged as non-threat by airport model?

*Example 2:* What if Jury or judge decides to make a criminal go free?

*Example 3:* What if you rejected to marry a very good person based on your predictive model and you happen to meet him/her after a few years and realize that you had a false negative?

**53. Can you cite some examples where both false positive and false negatives are equally important?**

In the Banking industry giving loans is the primary source of making money but at the same time if your repayment rate is not good you will not make any profit, rather you will risk huge losses.

Banks don't want to lose good customers and at the same point in time, they don't want to acquire bad customers. In this scenario, both the false positives and false negatives become very important to measure.

**54. Can you explain the difference between a Validation Set and a Test Set?**

A *Validation set* can be considered as a part of the training set as it is used for parameter selection and to avoid overfitting of the model being built.

On the other hand, a *Test Set* is used for testing or evaluating the performance of a trained machine learning model.

In simple terms, the differences can be summarized as; training set is to fit the parameters i.e. weights and test set is to assess the performance of the model i.e. evaluating the predictive power and generalization.

### **55 . Explain cross-validation.**

Cross-validation is a model validation technique for evaluating how the outcomes of statistical analysis will generalize to an independent dataset. Mainly used in backgrounds where the objective is forecast and one wants to estimate how accurately a model will accomplish in practice.

The goal of cross-validation is to term a data set to test the model in the training phase (i.e. validation data set) in order to limit problems like overfitting and get an insight on how the model will generalize to an independent data set.

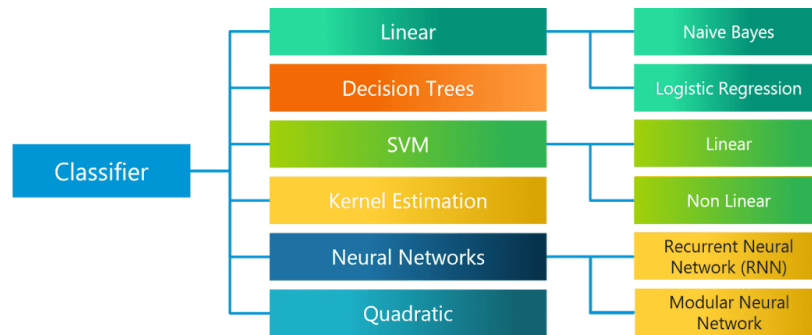
### **56. What is Machine Learning?**

*Machine Learning* explores the study and construction of algorithms that can learn from and make predictions on data. Closely related to computational statistics. Used to devise complex models and algorithms that lend themselves to a prediction which in commercial use is known as predictive analytics. Given below, is an image representing the various domains Machine Learning lends itself to.

### **57 . What are the various classification algorithms?**



The diagram lists the most important *classification algorithms*.



### 58 . What is 'Naive' in a Naive Bayes?

The *Naive Bayes Algorithm* is based on the Bayes Theorem. Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

The Algorithm is 'naive' because it makes assumptions that may or may not turn out to be correct.

### 59 . What are the different kernels in SVM?

There are four types of kernels in SVM.

1. Linear Kernel
2. Polynomial kernel
3. Radial basis kernel
4. Sigmoid kernel

### 60. What are Recommender Systems?

Recommender Systems are a subclass of information filtering systems that are meant to predict the preferences or ratings that a user would give to a product. Recommender

systems are widely used in movies, news, research articles, products, social tags, music, etc.

*Examples include movie recommenders in IMDB, Netflix & BookMyShow, product recommenders in e-commerce sites like Amazon, eBay & Flipkart, YouTube video recommendations and game recommendations in Xbox.*

## 61. What is Collaborative filtering?

The process of filtering used by most of the recommender systems to find patterns or information by collaborating viewpoints, various data sources and multiple agents.

Movie	Alice	Bob	Carol	Dave
Shutter Island	4	3	5	1
Fight Club	5	4	4	2
Dark Knight	5	3	4	?
21	4	3	?	5
Home Alone	4	4	5	5

**Figure:** Predicting the rating of Dave for Dark Knight and Carol for 21 using Collaborative Filtering

An example of collaborative filtering can be to predict the rating of a particular user based on his/her ratings for other movies and others' ratings for all movies. This concept is widely used in recommending movies in IMDB, Netflix & BookMyShow, product recommenders in e-commerce sites like Amazon, eBay & Flipkart, YouTube video recommendations and game recommendations in Xbox.

## 62. How can outlier values be treated?

Outlier values can be identified by using univariate or any other graphical analysis method. If the number of outlier values is few then they can be assessed individually but for a large number of outliers, the values can be substituted with either the 99th or the 1st percentile values.

All extreme values are not outlier values. The most common ways to treat outlier values

1. To change the value and bring it within a range.
2. To just remove the value.

**63. What are the various steps involved in an analytics project?**

**The following are the various steps involved in an analytics project:**

1. Understand the Business problem
2. Explore the data and become familiar with it.
3. Prepare the data for modelling by detecting outliers, treating missing values, transforming variables, etc.
4. After data preparation, start running the model, analyze the result and tweak the approach. This is an iterative step until the best possible outcome is achieved.
5. Validate the model using a new data set.
6. Start implementing the model and track the result to analyze the performance of the model over the period of time.

**64. During analysis, how do you treat missing values?**

The extent of the missing values is identified after identifying the variables with missing values. If any patterns are identified the analyst has to concentrate on them as it could lead to interesting and meaningful business insights.

If there are no patterns identified, then the missing values can be substituted with mean or median values (imputation) or they can simply be ignored. Assigning a default value which can be mean, minimum or maximum value. Getting into the data is important.

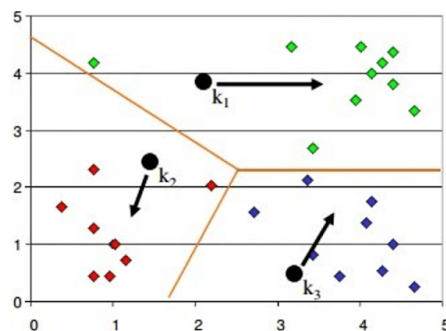
If it is a categorical variable, the default value is assigned. The missing value is assigned a default value. If you have a distribution of data coming, for normal distribution give the mean value.

If 80% of the values for a variable are missing then you can answer that you would be dropping the variable instead of treating the missing values.

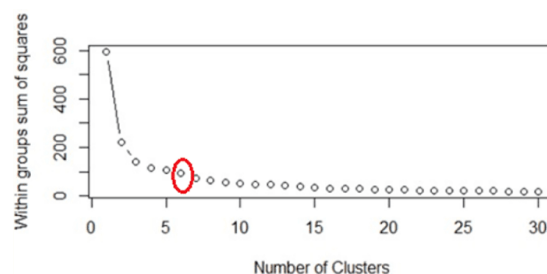
### 65. How will you define the number of clusters in a clustering algorithm?

Though the Clustering Algorithm is not specified, this question is mostly in reference to *K-Means clustering* where “K” defines the number of clusters. The objective of clustering is to group similar entities in a way that the entities within a group are similar to each other but the groups are different from each other.

For example, the following image shows three different groups.



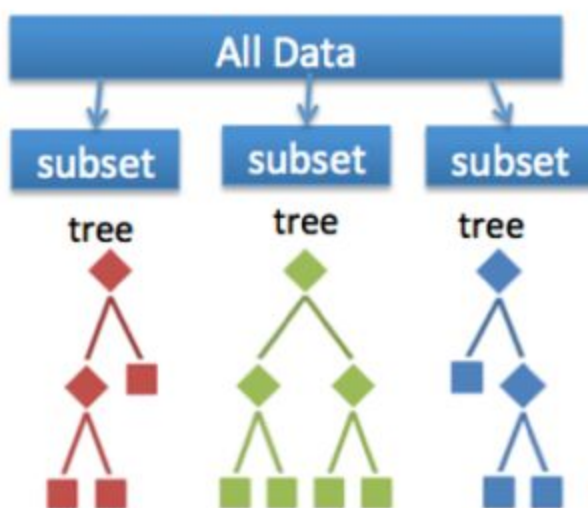
Within Sum of squares is generally used to explain the homogeneity within a cluster. If you plot WSS for a range of number of clusters, you will get the plot shown below.



- This is the widely used approach but few data scientists also use Hierarchical clustering first to create dendrograms and identify the distinct groups from there.

Ensemble Learning is basically combining a diverse set of learners(Individual models) together to improvise on the stability and predictive power of the model.

*Random forest* is a versatile machine learning method capable of performing both regression and classification tasks. It is also used for dimensionality reduction, treats missing values, outlier values. It is a type of ensemble learning method, where a group of weak models combine to form a powerful model.



In Random Forest, we grow multiple trees as opposed to a single tree. To classify a new object based on attributes, each tree gives a classification. The forest chooses the classification having the most votes(Overall the trees in the forest) and in case of regression, it takes the average of outputs by different trees.

### **68. How Do You Work Towards a Random Forest?**

The underlying principle of this technique is that several weak learners combined to provide a keen learner. The steps involved are

- Build several decision trees on bootstrapped training samples of data
- On each tree, each time a split is considered, a random sample of mm predictors is chosen as split candidates, out of all pp predictors
- Rule of thumb: At each split  $m = p \sqrt{m} = p$
- Predictions: At the majority rule

### **69. What cross-validation technique would you use on a time series data set?**

Instead of using k-fold cross-validation, you should be aware of the fact that a time series is not randomly distributed data — It is inherently ordered by chronological order.

In case of time series data, you should use techniques like forward=chaining — Where you will be model on past data then look at forward-facing data.

fold 1: training[1], test[2]

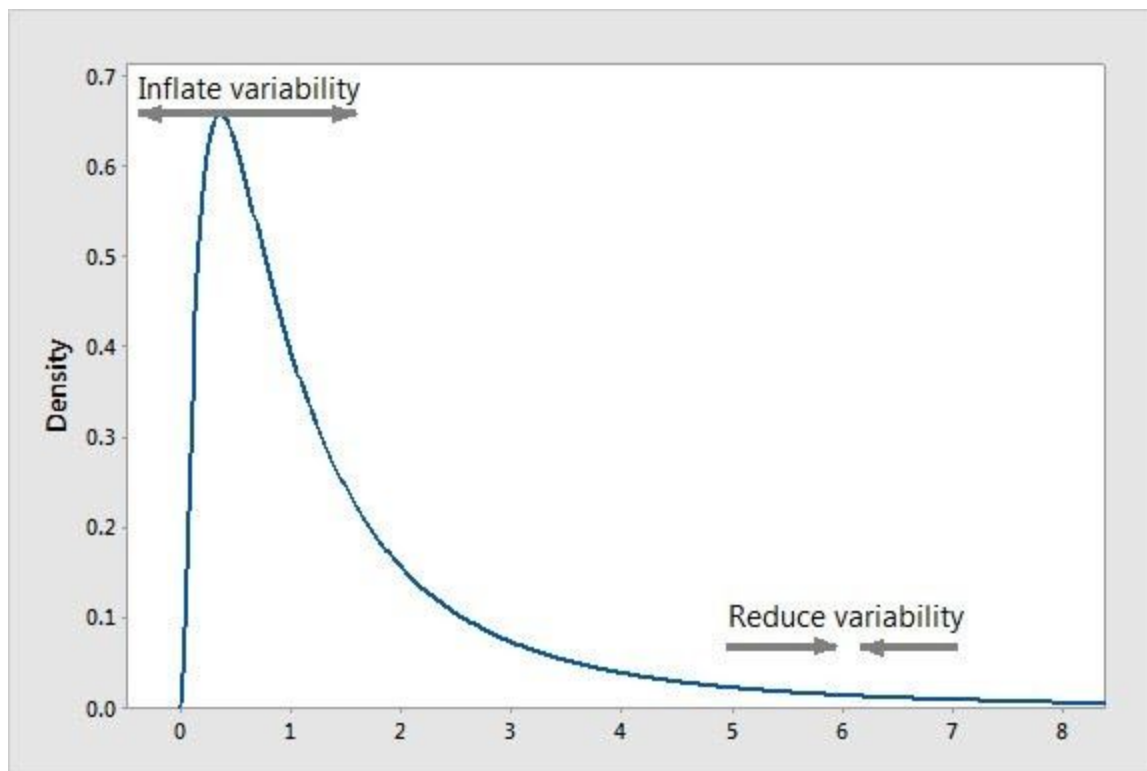
fold 1: training[1 2], test[3]

fold 1: training[1 2 3], test[4]

fold 1: training[1 2 3 4], test[5]

### **70. What is a Box-Cox Transformation?**

The dependent variable for a regression analysis might not satisfy one or more assumptions of an ordinary least squares regression. The residuals could either curve as the prediction increases or follow the skewed distribution. In such scenarios, it is necessary to transform the response variable so that the data meets the required assumptions. A Box cox transformation is a statistical technique to transform non-normal dependent variables into a normal shape. If the given data is not normal then most of the statistical techniques assume normality. Applying a box cox transformation means that you can run a broader number of tests.



A Box-Cox transformation is a way to transform non-normal dependent variables into a normal shape. Normality is an important assumption for many statistical techniques, if your data isn't normal, applying a Box-Cox means that you are able to run a broader number of tests. The Box-Cox transformation is named after statisticians *George Box*

and *Sir David Roxbee Cox* who collaborated on a 1964 paper and developed the technique.

## **71. What is Data Analyst ,Data scientist,Data Engineer ?**

### **Data Analyst**

Data Analysts deliver value to their companies by taking data, using it to answer questions, and communicating the results to help make business decisions. Common tasks done by data analysts include data cleaning, performing analysis and creating data visualizations.

#### **Roles**

- Cleaning and organizing raw data.
- Using descriptive statistics to get a big-picture view of their data.
- Analyzing interesting trends found in the data.
- Creating visualizations and dashboards to help the company interpret and make decisions with the data.
- Presenting the results of a technical analysis to business clients or internal teams.

### **Data scientist**

The data scientist is an individual who can provide immense value by tackling more open-ended questions and leveraging their knowledge of advanced statistics and



algorithms. If the analyst focuses on understanding data from the past and present perspectives, then the scientist focuses on producing reliable predictions for the future.

## **Roles**

- Evaluating statistical models to determine the validity of analyses.
- Using machine learning to build better predictive algorithms.
- Testing and continuously improving the accuracy of machine learning models.
- Building data visualizations to summarize the conclusion of an advanced analysis.

## **Data Engineer**


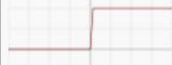


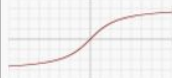




The data engineer establishes the foundation that the data analysts and scientists build upon. Data engineers are responsible for constructing data pipelines and often have to use complex tools and techniques to handle data at scale. Unlike the previous two career paths, data engineering leans a lot more toward a software development skill set.

## **Roles**

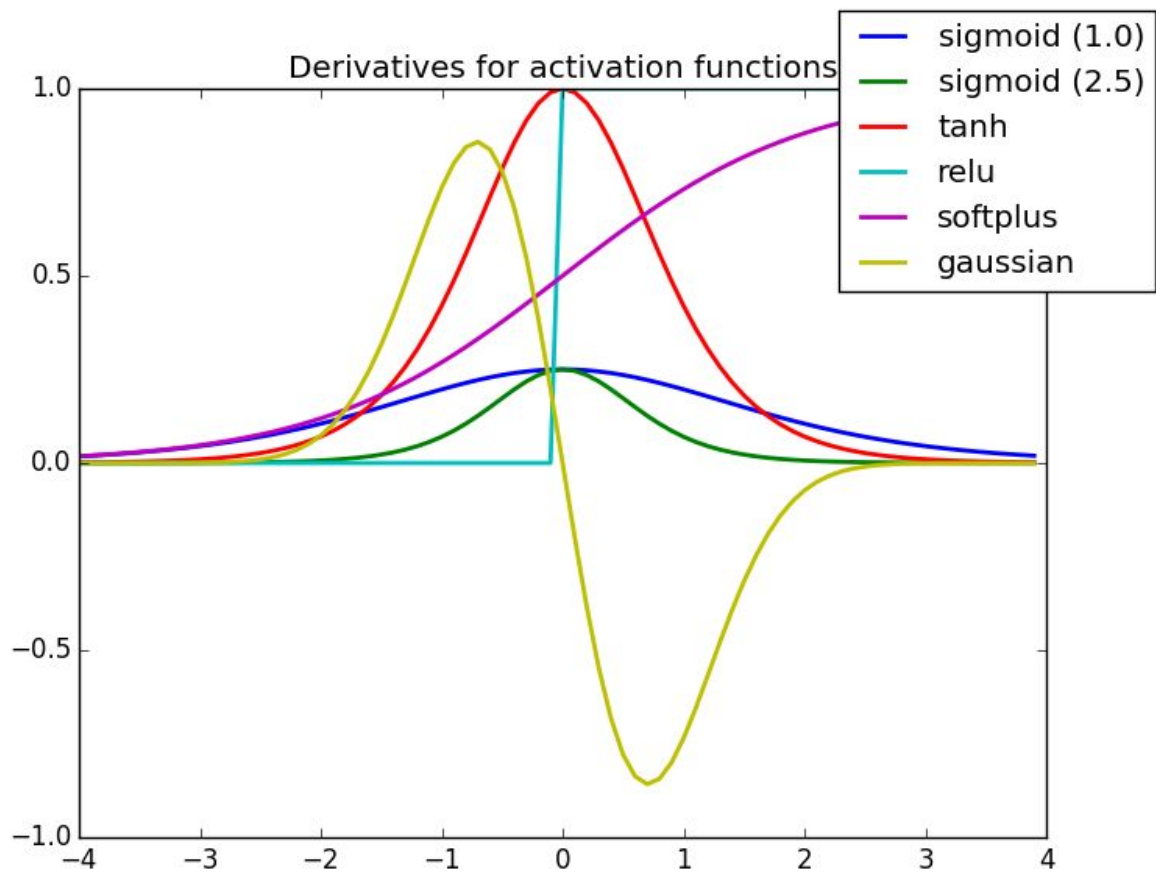
- Building APIs for data consumption.
- Integrating external or new datasets into existing data pipelines.
- Applying feature transformations for machine learning models on new data.
- Continuously monitoring and testing the system to ensure optimized performance.

## **72. Why derivative/differentiation is used ?**

When updating the curve, to know in which direction and how much to change or update the curve depending upon the slope. That is why we use differentiation in almost every part of Machine Learning and Deep Learning.

Name	Plot	Equation	Derivative
Identity		$f(x) = x$	$f'(x) = 1$
Binary step		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for } x \neq 0 \\ ? & \text{for } x = 0 \end{cases}$
Logistic (a.k.a. Soft step)		$f(x) = \frac{1}{1 + e^{-x}}$	$f'(x) = f(x)(1 - f(x))$
Tanh		$f(x) = \tanh(x) = \frac{2}{1 + e^{-2x}} - 1$	$f'(x) = 1 - f(x)^2$
ArcTan		$f(x) = \tan^{-1}(x)$	$f'(x) = \frac{1}{x^2 + 1}$
Rectified Linear Unit (ReLU)		$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$
Parameteric Rectified Linear Unit (PReLU) [2]		$f(x) = \begin{cases} \alpha x & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} \alpha & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$
Exponential Linear Unit (ELU) [3]		$f(x) = \begin{cases} \alpha(e^x - 1) & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$	$f'(x) = \begin{cases} f(x) + \alpha & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$
SoftPlus		$f(x) = \log_e(1 + e^x)$	$f'(x) = \frac{1}{1 + e^{-x}}$

**Fig: Activation Function Cheetsheet**



**Fig: Derivative of Activation Functions**

### 73) Mention the difference between Data Mining and Machine learning?

Machine learning relates with the study, design and development of the algorithms that give computers the capability to learn without being explicitly programmed. While, data mining can be defined as the process in which the unstructured data tries to extract knowledge or unknown interesting patterns. During this process machine, learning algorithms are used.

#### **74) What is 'Overfitting' in Machine learning?**

In machine learning, when a statistical model describes random error or noise instead of underlying relationship 'overfitting' occurs. When a model is excessively complex, overfitting is normally observed, because of having too many parameters with respect to the number of training data types. The model exhibits poor performance which has been overfit.

#### **75) Why overfitting happens?**

The possibility of overfitting exists as the criteria used for training the model is not the same as the criteria used to judge the efficacy of a model.

#### **76) How can you avoid overfitting ?**

By using a lot of data overfitting can be avoided, overfitting happens relatively as you have a small dataset, and you try to learn from it. But if you have a small database and you are forced to come with a model based on that. In such situation, you can use a technique known as cross validation. In this method the dataset splits into two section, testing and training datasets, the testing dataset will only test the model while, in training dataset, the datapoints will come up with the model.

In this technique, a model is usually given a dataset of a known data on which training (training data set) is run and a dataset of unknown data against which the model is tested. The idea of cross validation is to define a dataset to "test" the model in the training phase.

#### **77) What is inductive machine learning?**

The inductive machine learning involves the process of learning by examples, where a system, from a set of observed instances tries to induce a general rule.

78) What are the five popular algorithms of Machine Learning?

- a) Decision Trees
- b) Neural Networks (back propagation)
- c) Probabilistic networks
- d) Nearest Neighbor
- e) Support vector machines

79) **What are the different Algorithm techniques in Machine Learning?**

The different types of techniques in Machine Learning are

- a) Supervised Learning
- b) Unsupervised Learning
- c) Semi-supervised Learning
- d) Reinforcement Learning
- e) Transduction
- f) Learning to Learn

80) **What are the three stages to build the hypotheses or model in machine learning?**

- a) Model building
- b) Model testing
- c) Applying the model

**81) What is the standard approach to supervised learning?**

The standard approach to supervised learning is to split the set of example into the training set and the test.

**82) What is 'Training set' and 'Test set'?**

In various areas of information science like machine learning, a set of data is used to discover the potentially predictive relationship known as 'Training Set'. Training set is an examples given to the learner, while Test set is used to test the accuracy of the hypotheses generated by the learner, and it is the set of example held back from the learner. Training set are distinct from Test set.

**83) List down various approaches for machine learning?**

The different approaches in Machine Learning are

- a) Concept Vs Classification Learning
- b) Symbolic Vs Statistical Learning
- c) Inductive Vs Analytical Learning

**84) What is not Machine Learning?**

- a) Artificial Intelligence

- b) Rule based inference

**85) Explain what is the function of 'Unsupervised Learning'?**

- a) Find clusters of the data
- b) Find low-dimensional representations of the data
- c) Find interesting directions in data
- d) Interesting coordinates and correlations
- e) Find novel observations/ database cleaning

**86) Explain what is the function of 'Supervised Learning'?**

- a) Classifications
- b) Speech recognition
- c) Regression
- d) Predict time series
- e) Annotate strings

**87) What is algorithm independent machine learning?**

Machine learning in where mathematical foundations is independent of any particular classifier or learning algorithm is referred as algorithm independent machine learning?

**88) What is the difference between artificial learning and machine learning?**



Designing and developing algorithms according to the behaviours based on empirical data are known as Machine Learning. While artificial intelligence in addition to machine learning, it also covers other aspects like knowledge representation, natural language processing, planning, robotics etc.

**89) What is classifier in machine learning?**

A classifier in a Machine Learning is a system that inputs a vector of discrete or continuous feature values and outputs a single discrete value, the class.

**90) What are the advantages of Naive Bayes?**

In Naïve Bayes classifier will converge quicker than discriminative models like logistic regression, so you need less training data. The main advantage is that it can't learn interactions between features.

**91) In what areas Pattern Recognition is used?**

Pattern Recognition can be used in

- a) Computer Vision
- b) Speech Recognition
- c) Data Mining
- d) Statistics
- e) Informal Retrieval
- f) Bio-Informatics

**92) What is Genetic Programming?**

Genetic programming is one of the two techniques used in machine learning. The model is based on the testing and selecting the best choice among a set of results.

**93) What is Inductive Logic Programming in Machine Learning?**

Inductive Logic Programming (ILP) is a subfield of machine learning which uses logical programming representing background knowledge and examples.

**94) What is Model Selection in Machine Learning?**

The process of selecting models among different mathematical models, which are used to describe the same data set is known as Model Selection. Model selection is applied to the fields of statistics, machine learning and data mining.

**95) What are the two methods used for the calibration in Supervised Learning?**

The two methods used for predicting good probabilities in Supervised Learning are

- a) Platt Calibration
- b) Isotonic Regression

These methods are designed for binary classification, and it is not trivial.

**96) Which method is frequently used to prevent overfitting?**

When there is sufficient data 'Isotonic Regression' is used to prevent an overfitting issue.

**97) What is the difference between heuristic for rule learning and heuristics for decision trees?**

The difference is that the heuristics for decision trees evaluate the average quality of a number of disjointed sets while rule learners only evaluate the quality of the set of instances that is covered with the candidate rule.

**98) What is Perceptron in Machine Learning?**

In Machine Learning, Perceptron is an algorithm for supervised classification of the input into one of several possible non-binary outputs.

**99) Explain the two components of Bayesian logic program?**

Bayesian logic program consists of two components. The first component is a logical one ; it consists of a set of Bayesian Clauses, which captures the qualitative structure of the domain. The second component is a quantitative one, it encodes the quantitative information about the domain.

**100) What are Bayesian Networks (BN) ?**

Bayesian Network is used to represent the graphical model for probability relationship among a set of variables .

**101) Why instance based learning algorithm sometimes referred as Lazy learning algorithm?**

Instance based learning algorithm is also referred as Lazy learning algorithm as they delay the induction or generalization process until classification is performed.

**102) What are the two classification methods that SVM ( Support Vector Machine) can handle?**

- a) Combining binary classifiers
- b) Modifying binary to incorporate multiclass learning

**103) What is ensemble learning?**

To solve a particular computational program, multiple models such as classifiers or experts are strategically generated and combined. This process is known as ensemble learning.

**104) Why ensemble learning is used?**

Ensemble learning is used to improve the classification, prediction, function approximation etc of a model.

**105) When to use ensemble learning?**

Ensemble learning is used when you build component classifiers that are more accurate and independent from each other.

**106) What are the two paradigms of ensemble methods?**

The two paradigms of ensemble methods are

- a) Sequential ensemble methods
- b) Parallel ensemble methods

**107) What is the general principle of an ensemble method and what is bagging and boosting in ensemble method?**

The general principle of an ensemble method is to combine the predictions of several models built with a given learning algorithm in order to improve robustness over a single model. Bagging is a method in ensemble for improving unstable estimation or classification schemes. While boosting method are used sequentially to reduce the bias of the combined model. Boosting and Bagging both can reduce errors by reducing the variance term.

**108) What is bias-variance decomposition of classification error in ensemble method?**

The expected error of a learning algorithm can be decomposed into bias and variance. A bias term measures how closely the average classifier produced by the learning algorithm matches the target function. The variance term measures how much the learning algorithm's prediction fluctuates for different training sets.

**109) What is an Incremental Learning algorithm in ensemble?**

Incremental learning method is the ability of an algorithm to learn from new data that may be available after classifier has already been generated from already available dataset.

**110) What is PCA, KPCA and ICA used for?**

PCA (Principal Components Analysis), KPCA (Kernel based Principal Component Analysis) and ICA (Independent Component Analysis) are important feature extraction techniques used for dimensionality reduction.

**111) What is dimension reduction in Machine Learning?**

In Machine Learning and statistics, dimension reduction is the process of reducing the number of random variables under considerations and can be divided into feature selection and feature extraction

**112) What are support vector machines?**

Support vector machines are supervised learning algorithms used for classification and regression analysis.

**113) What are the components of relational evaluation techniques?**

The important components of relational evaluation techniques are

- a) Data Acquisition
- b) Ground Truth Acquisition
- c) Cross Validation Technique
- d) Query Type
- e) Scoring Metric
- f) Significance Test

**114) What are the different methods for Sequential Supervised Learning?**

The different methods to solve Sequential Supervised Learning problems are

- a) Sliding-window methods
- b) Recurrent sliding windows
- c) Hidden Markow models
- d) Maximum entropy Markow models
- e) Conditional random fields
- f) Graph transformer networks

**115) What are the areas in robotics and information processing where sequential prediction problem arises?**

The areas in robotics and information processing where sequential prediction problem arises are

- a) Imitation Learning
- b) Structured prediction
- c) Model based reinforcement learning

**116) What is batch statistical learning?**

Statistical learning techniques allow learning a function or predictor from a set of observed data that can make predictions about unseen or future data. These

techniques provide guarantees on the performance of the learned predictor on the future unseen data based on a statistical assumption on the data generating process.

**117) What is PAC Learning?**

PAC (Probably Approximately Correct) learning is a learning framework that has been introduced to analyze learning algorithms and their statistical efficiency.

**118) What are the different categories you can categorized the sequence learning process?**

- a) Sequence prediction
- b) Sequence generation
- c) Sequence recognition
- d) Sequential decision

**119) What is sequence learning?**

Sequence learning is a method of teaching and learning in a logical manner.

**120) What are two techniques of Machine Learning ?**

The two techniques of Machine Learning are

- a) Genetic Programming
- b) Inductive Learning

**121) Give a popular application of machine learning that you see on day to day basis?**



The recommendation engine implemented by major ecommerce websites uses  
Machine Learning

