

Interview Questions

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:

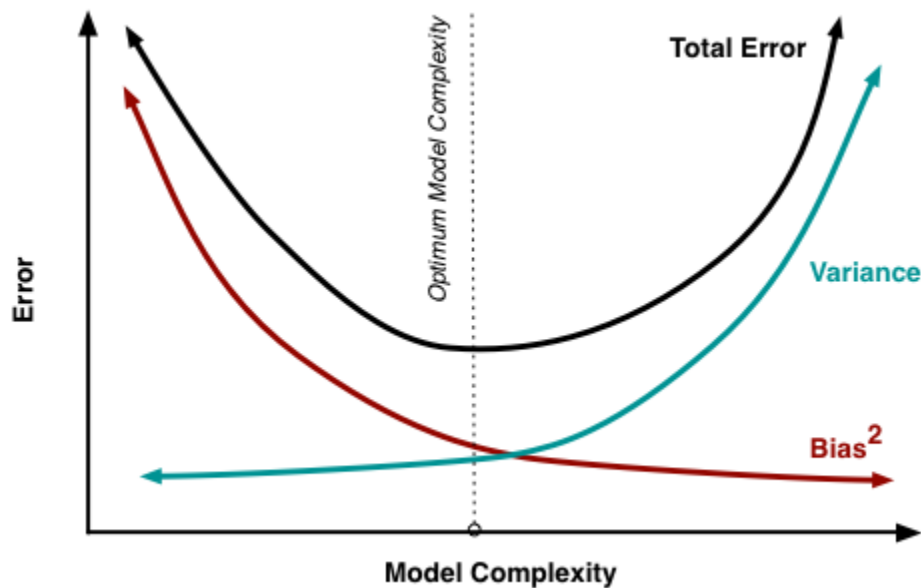
"Bias is error introduced in your model due to over simplification of machine learning algorithm." It can lead to under fitting. When you train your model at that time model makes simplified assumptions to make the target function easier to understand.

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.

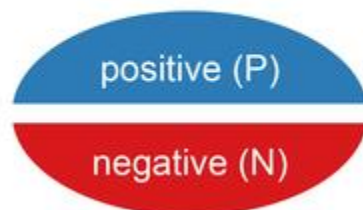
4. What is a confusion matrix ?

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)

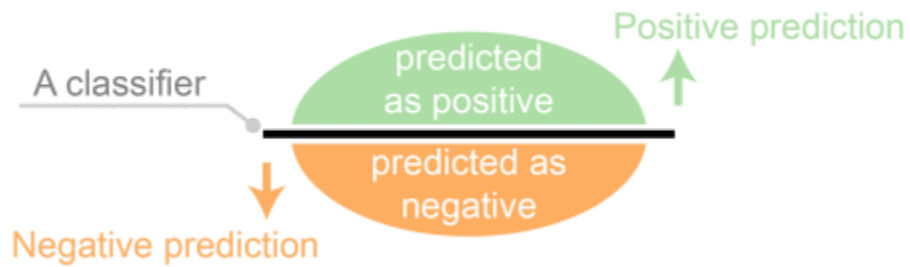
A data set used for performance evaluation is called test data set. It should contain the correct labels and predicted labels.

Two actual classes or observed labels



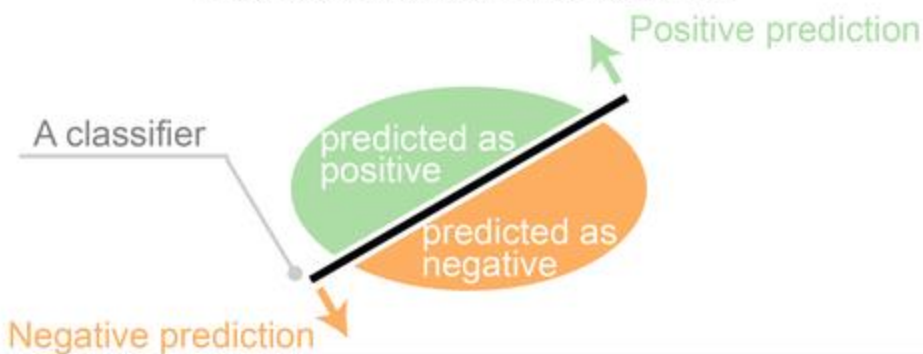
The predicted labels will exactly the same if the performance of a binary classifier is perfect.

Predicted classes of a perfect classifier



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

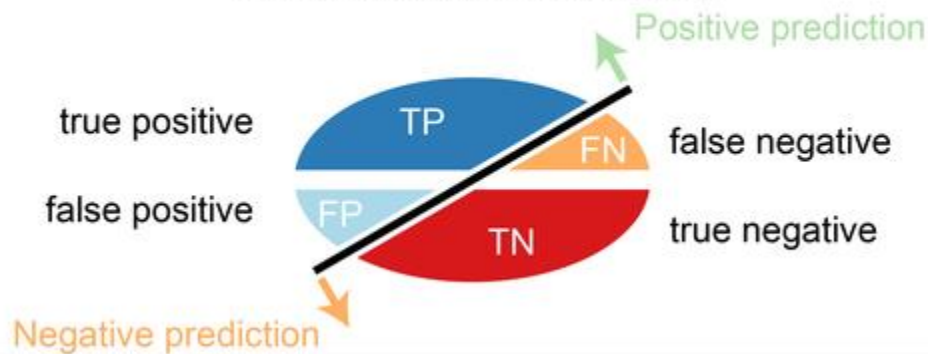
Predicted classes of a classifier



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

Four outcomes of a classifier

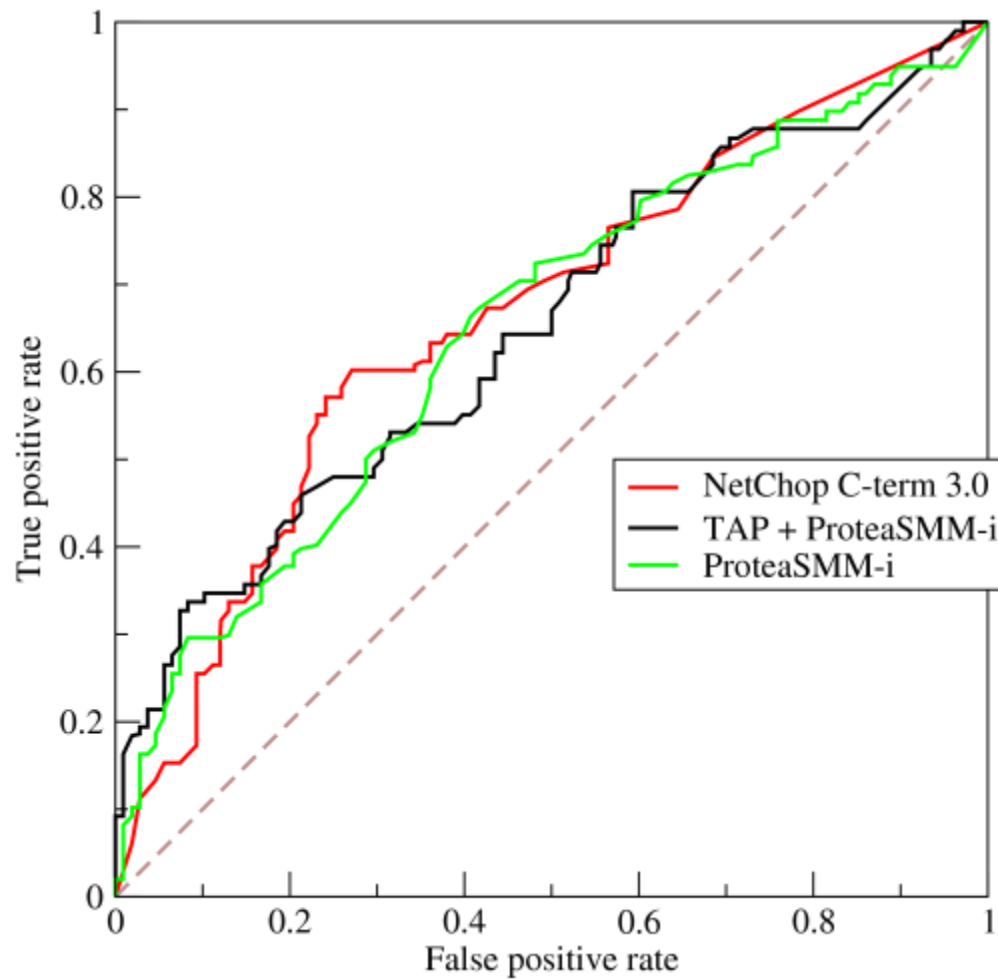


Basic measures derived from the confusion matrix

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.

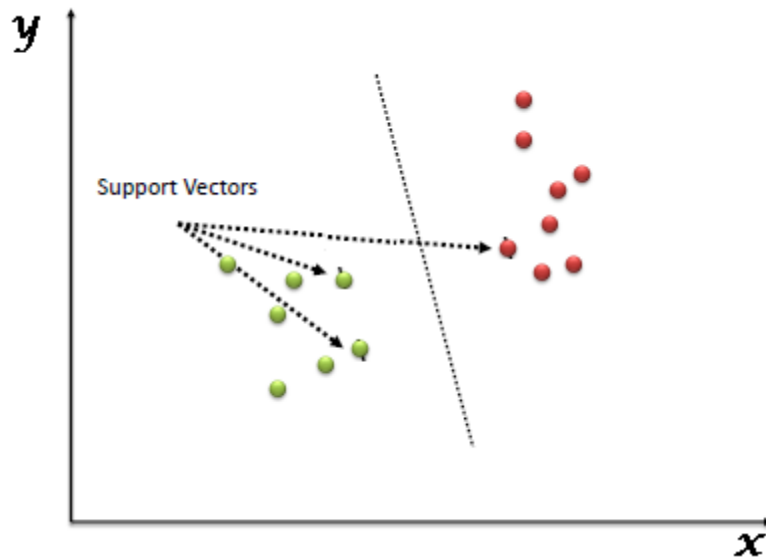


7. What is selection Bias ?

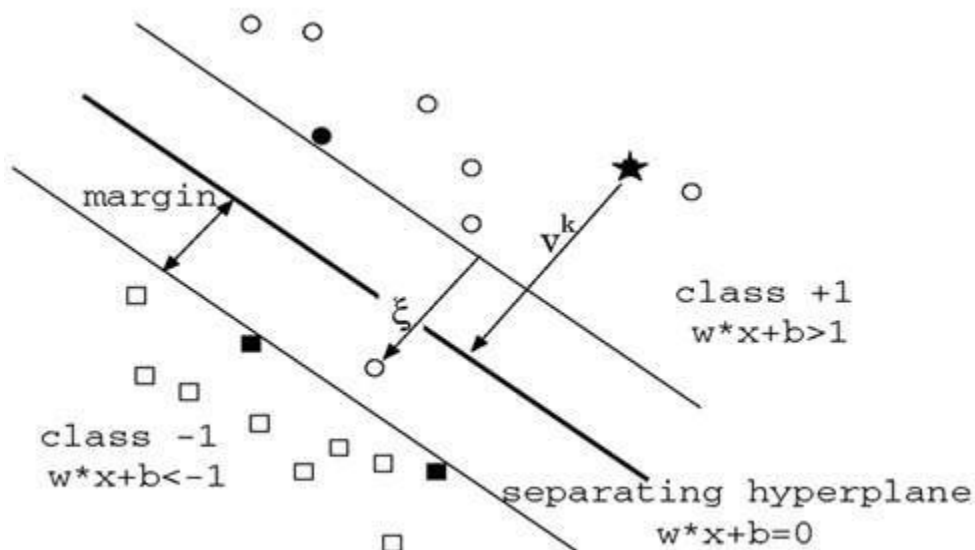
Selection bias occurs when sample obtained is not representative of the population intended to be analysed.

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.

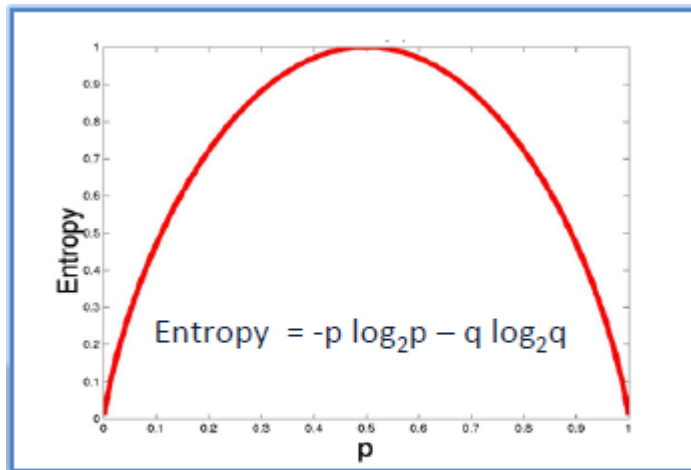


12. What is Entropy and Information gain in Decision tree algorithm ?

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 entropy 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	

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

$$\begin{aligned} G(\text{PlayGolf, Outlook}) &= E(\text{PlayGolf}) - E(\text{PlayGolf, Outlook}) \\ &= 0.940 - 0.693 = 0.247 \end{aligned}$$

13. What is pruning in Decision Tree ?

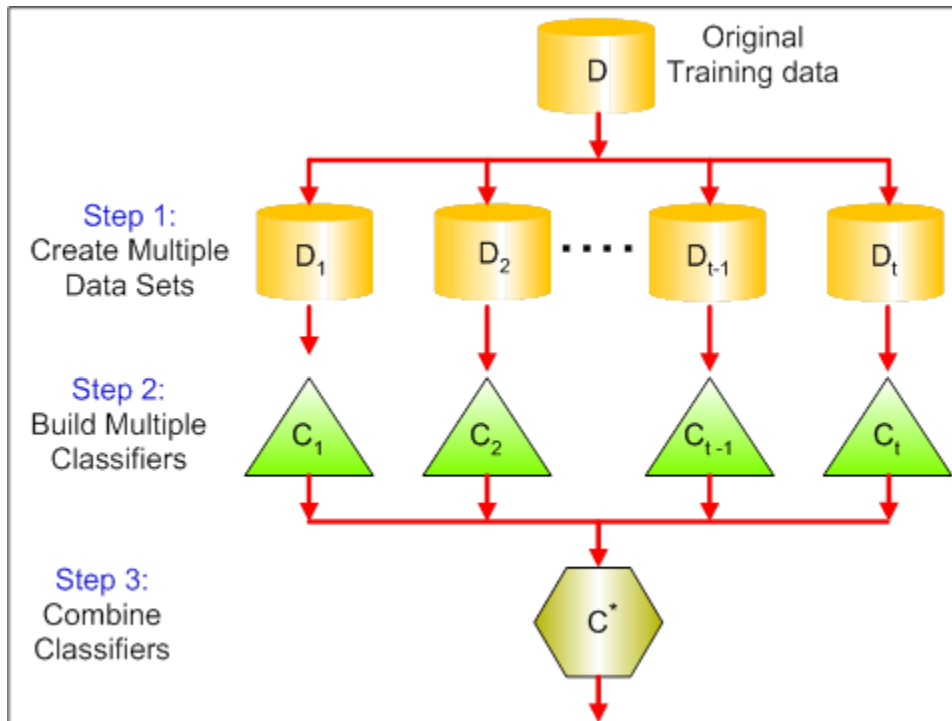
When we remove sub-nodes of a decision node, this process is called pruning or opposite process of splitting.

14. What is Ensemble Learning ?

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

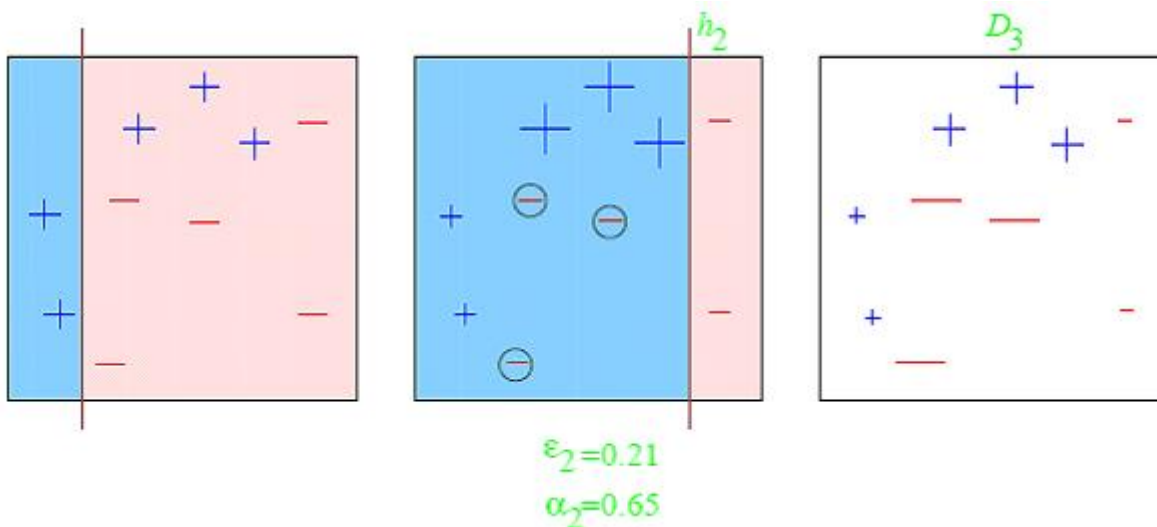
Bagging

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]

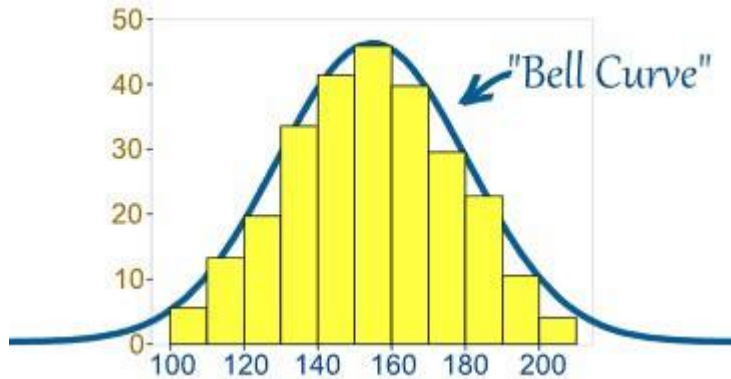
fold 1: training[1 2 3], test[4]

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.

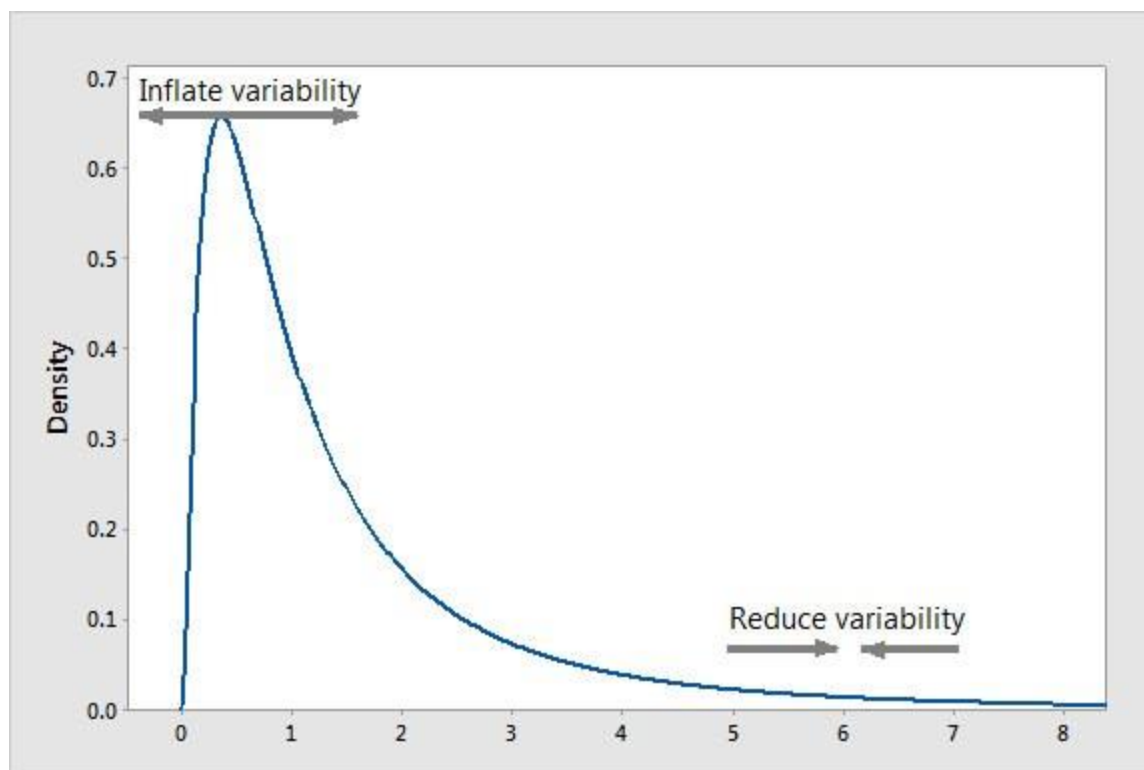
18. What do you understand by the term Normal Distribution?



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?

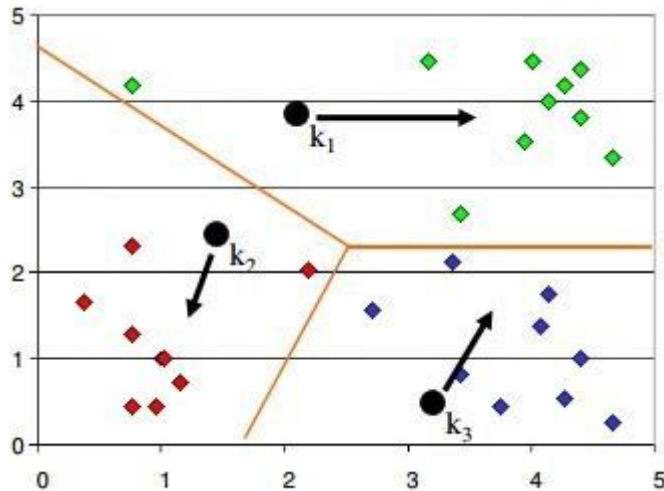
Dependent variable for a regression analysis might not satisfy one or more assumptions of an ordinary least squares regression. The residuals could either curve as the prediction increases or follow skewed distribution. In such scenarios, it is necessary to transform the response variable so that the 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.



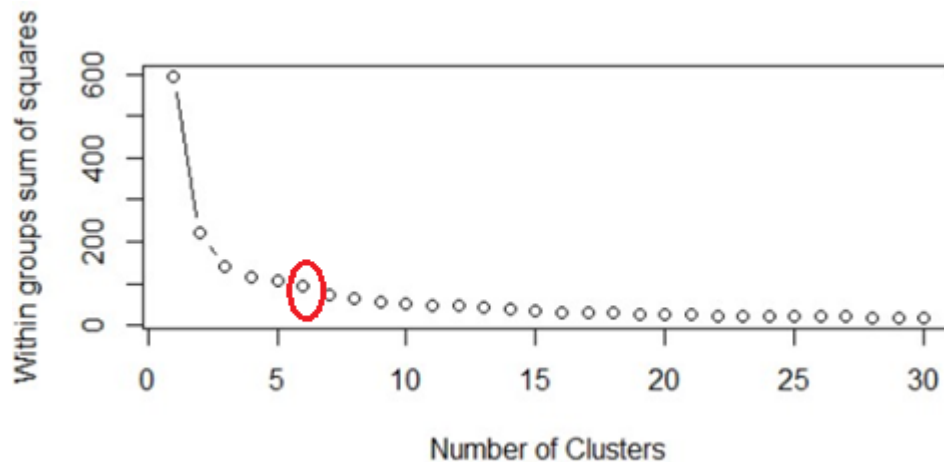
A Box-Cox transformation is a way to transform non-normal dependent variables into a normal shape. Normality is an important assumption for many statistical techniques; if your data isn't normal, applying a Box-Cox means that you are able to run a broader number of tests. The Box-Cox transformation is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



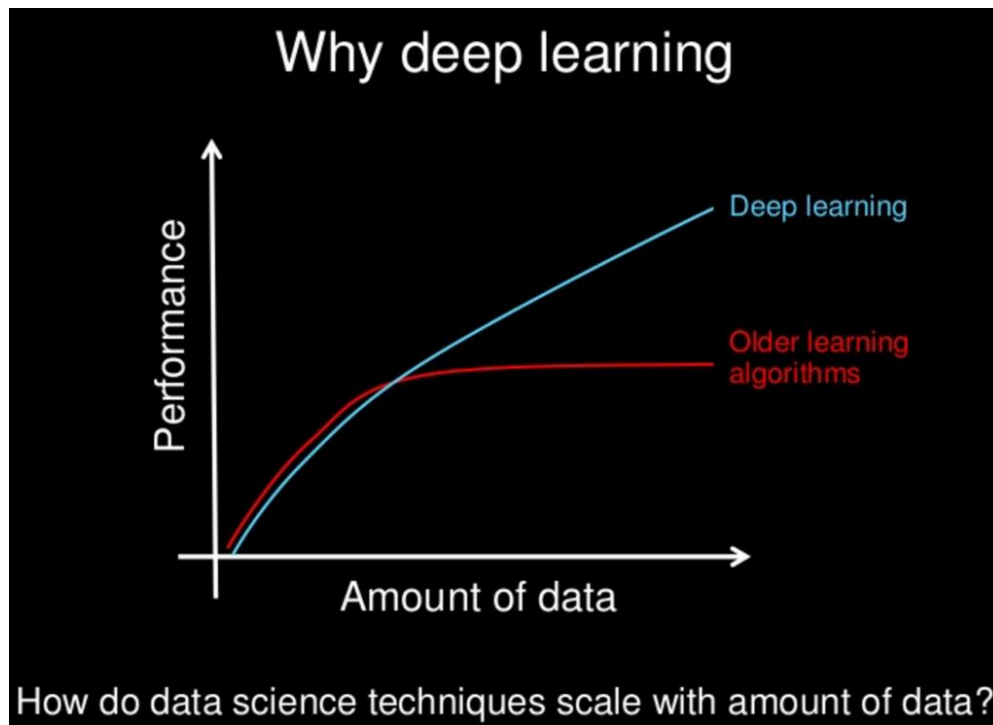
Within Sum of squares is generally used to explain the homogeneity within a cluster. If you plot WSS for a range of number of clusters, you will get the plot shown below. The Graph is generally known as Elbow Curve.



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 dendrograms and identify the distinct groups from there.

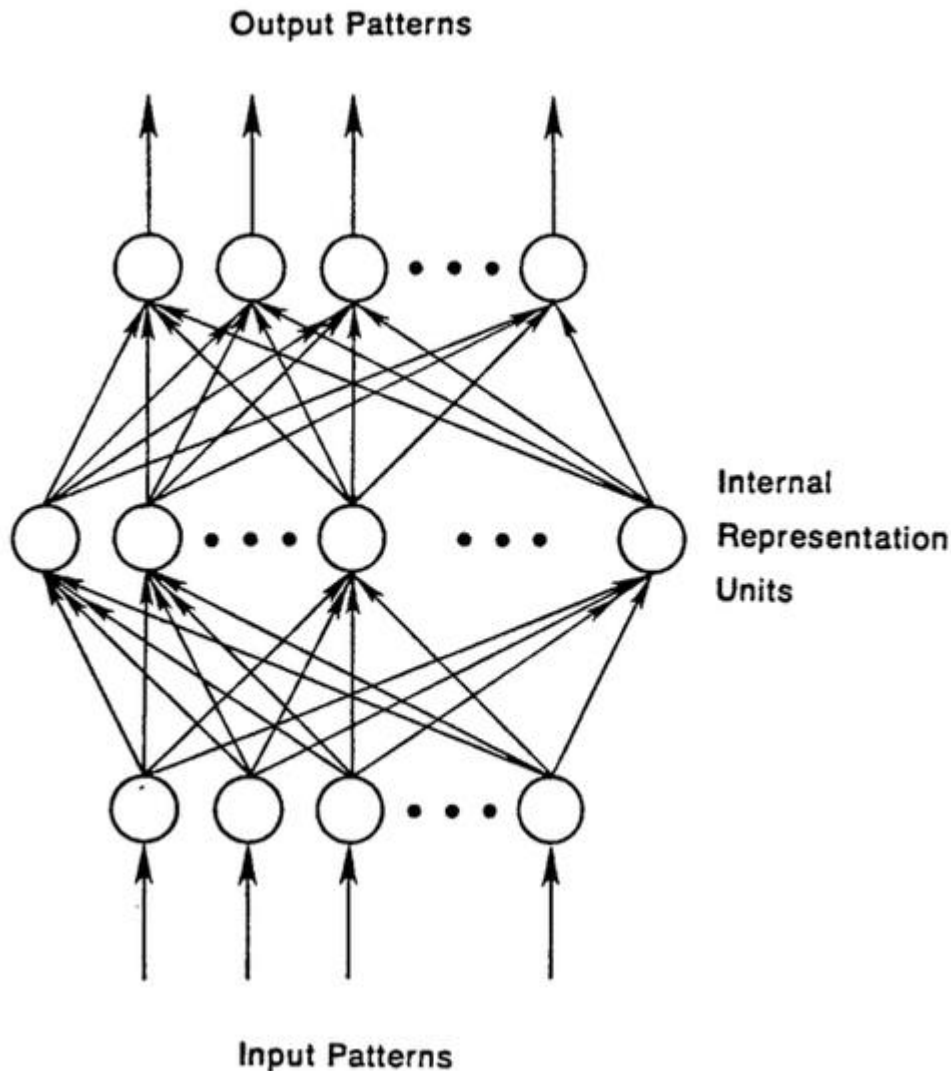
21. What is deep learning?

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.



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 operations 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.

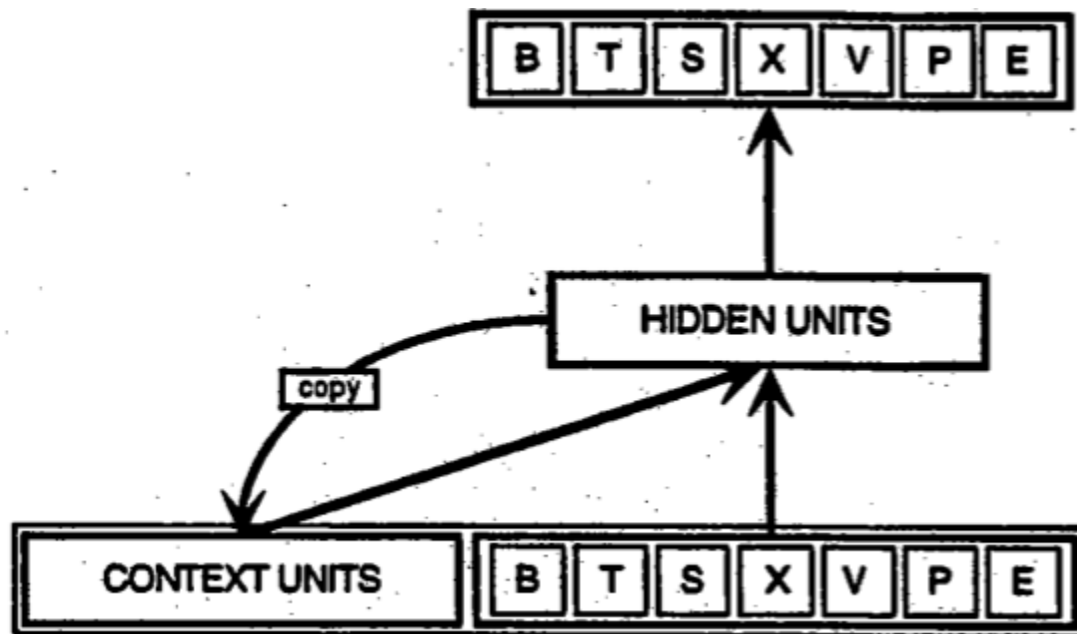


Recurrent networks on the other hand, take as their input not just the current input example they see, but also the what they have perceived previously in time. The BTSXPE at the bottom of the drawing represents the input example in the current moment, and CONTEXT UNIT represents the output of the previous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

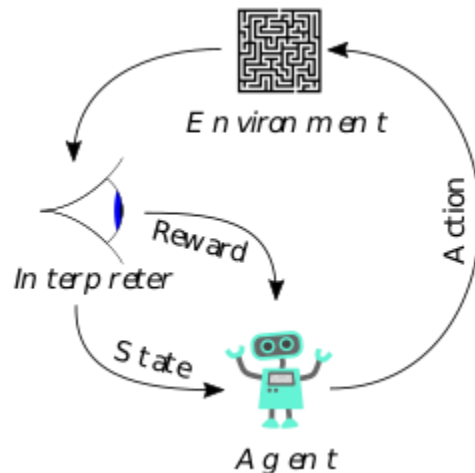
1. Supervised machine learning,
2. Unsupervised machine learning,
3. Reinforcement learning

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?

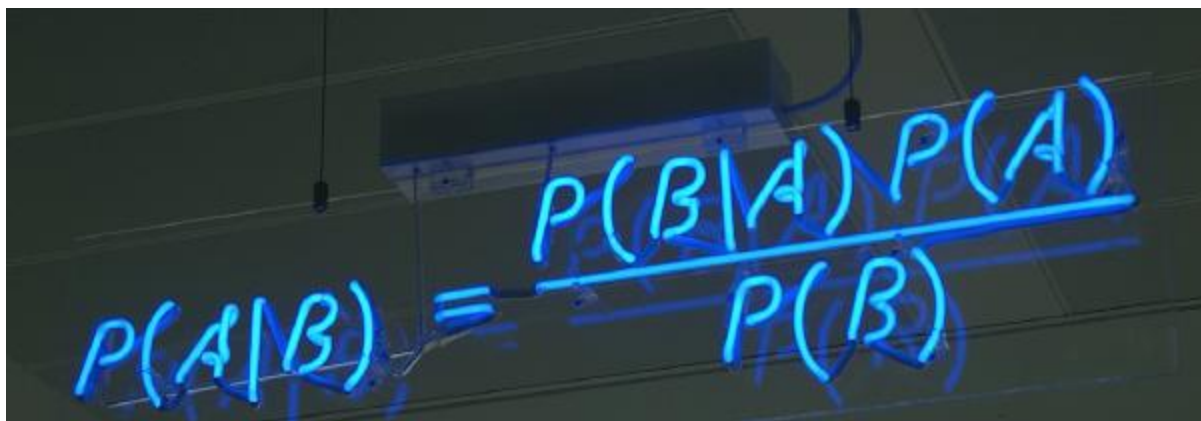
When you perform a hypothesis test in statistics, a p-value can help you determine the strength of your results. p-value is a number between 0 and 1. Based on the value it will denote the strength of the results. The claim which is on trial is called Null Hypothesis.

Low p-value (≤ 0.05) indicates strength against the null hypothesis which means we can reject the null Hypothesis. High p-value (≥ 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.


$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

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 —

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

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 (λ) 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.

<https://hackr.io/blog/data-science-interview-questions>

<https://www.dezyre.com/article/100-data-science-interview-questions-and-answers-general-for-2018/184>

<https://www.simplilearn.com/data-science-interview-questions-article>

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

The best possible answer for this would be Python because it has Pandas library that provides easy to use data structures and high performance data analysis tools.

3) **Which technique is used to predict categorical responses? ([get sample code here](#))**

Classification technique is used widely in mining for classifying data sets.

4) **What is logistic regression? Or State an example when you have used logistic regression recently. ([get sample use-case here](#))**

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.

5) **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.

6) **Why data cleaning plays a vital role in analysis? ([get sample use-case here](#))**

Cleaning data from multiple sources to transform it into a format that data analysts or data scientists can work with is a cumbersome process because - as the number of data sources increases, the time take to clean the data increases exponentially due to the number of sources and the volume of data generated in these sources. It might take up to 80% of the time for just cleaning data making it a critical part of analysis task.

7) Differentiate between univariate, bivariate and multivariate analysis.

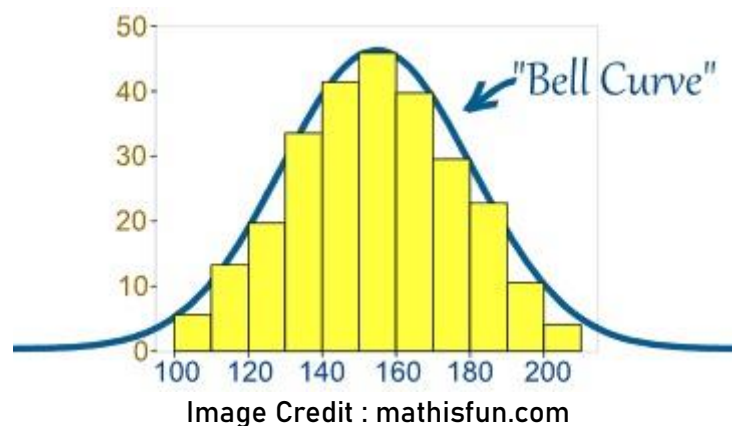
These are descriptive statistical analysis techniques which can be differentiated based on the number of variables involved at a given point of time. For example, the pie charts of sales based on territory involve only one variable and can be referred to as univariate analysis.

If the analysis attempts to understand the difference between 2 variables at time as in a scatterplot, then it is referred to as bivariate analysis. For example, analysing the volume of sale and a spending can be considered as an example of bivariate analysis.

Analysis that deals with the study of more than two variables to understand the effect of variables on the responses is referred to as multivariate analysis.

8) What do you understand by the term Normal Distribution? ([get sample code here](#))

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.



9) What is Linear Regression?

Linear regression is a statistical technique where the score of a variable Y is predicted from the score of a second variable X. X is referred to as the predictor variable and Y as the criterion variable.

10) What is Interpolation and Extrapolation?

Estimating a value from 2 known values from a list of values is Interpolation. Extrapolation is approximating a value by extending a known set of values or facts.

11) What is power analysis?

An experimental design technique for determining the effect of a given sample size.



12) What is K-means? How can you select K for K-means?

13) 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.

14) What is the difference between Cluster and Systematic 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. 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 for systematic sampling is equal probability method.

15) Are expected value and mean value different?

They are not different but the terms are used in different contexts. Mean is generally referred when talking about a probability distribution or sample population whereas expected value is generally referred in a random variable context.

For Sampling Data

Mean value is the only value that comes from the sampling data.

Expected Value is the mean of all the means i.e. the value that is built from multiple samples. Expected value is the population mean.

For Distributions

Mean value and Expected value are same irrespective of the distribution, under the condition that the distribution is in the same population.

16) What does P-value signify about the statistical data?

P-value is used to determine the significance of results after a hypothesis test in statistics. P-value helps the readers to draw conclusions and is always between 0 and 1.

- $P\text{-Value} > 0.05$ denotes weak evidence against the null hypothesis which means the null hypothesis cannot be rejected.
- $P\text{-value} \leq 0.05$ denotes strong evidence against the null hypothesis which means the null hypothesis can be rejected.
- $P\text{-value} = 0.05$ is the marginal value indicating it is possible to go either way.

[Get hands-on experience for your interviews with free access to solved code examples found here \(these are ready-to-use for your projects\)](#)

17) Do gradient descent methods always converge to same point?

No, they do not because in some cases it reaches a local minima or a local optima point. You don't reach the global optima point. It depends on the data and starting conditions

18) What are categorical variables?



19) A test has a true positive rate of 100% and false positive rate of 5%. There is a population with a 1/1000 rate of having the condition the test identifies. Considering a positive test, what is the probability of having that condition?

Let's suppose you are being tested for a disease, if you have the illness the test will end up saying you have the illness. However, if you don't have the illness- 5% of the times the test will end up saying you have the illness and 95% of the times the test will give accurate result that you don't have the illness. Thus there is a 5% error in case you do not have the illness.

Out of 1000 people, 1 person who has the disease will get true positive result.

Out of the remaining 999 people, 5% will also get true positive result.

Close to 50 people will get a true positive result for the disease.

This means that out of 1000 people, 51 people will be tested positive for the disease even though only one person has the illness. There is only a 2% probability of you having the disease even if your reports say that you have the disease.

20) How you can make data normal using Box-Cox transformation?

21) What is the difference between Supervised Learning and Unsupervised Learning?

If an algorithm learns something from the training data so that the knowledge can be applied to the test data, then it is referred to as Supervised Learning. Classification is an example for Supervised Learning. If the algorithm does not learn anything beforehand because there is no response variable or any training data, then it is referred to as unsupervised learning. Clustering is an example for unsupervised learning.

22) Explain the use of Combinatorics in data science.

23) Why is vectorization considered a powerful method for optimizing numerical code?

24) What is the goal of A/B Testing?

It is a statistical hypothesis testing for randomized experiment with two variables A and B. The goal of A/B Testing is to identify any changes to the web page to maximize or increase the outcome of an interest. An example for this could be identifying the click through rate for a banner ad.

25) What is an Eigenvalue and Eigenvector?

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.

[The #1 question in your interview is "What experience do you have?". Get hands-on experience with free access to code examples solved by industry experts. Click here \(these are ready-to-use for your projects\)](#)

26) What is Gradient Descent?

27) 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 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 in within a range

2) To just remove the value.

28) How can you assess a good logistic model?

There are various methods to assess the results of a logistic regression analysis-

- Using Classification Matrix to look at the true negatives and false positives.
- Concordance that helps identify the ability of the logistic model to differentiate between the event happening and not happening.
- Lift helps assess the logistic model by comparing it with random selection.

29) What are various steps involved in an analytics project?

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

30) How can you iterate over a list and also retrieve element indices at the same time?

This can be done using the enumerate function which takes every element in a sequence just like in a list and adds its location just before it.

31) During analysis, how do you treat missing values? ([get sample code here](#))

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. There are various factors to be considered when answering this question-

- Understand the problem statement, understand the data and then give the answer. 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.
- Should we even treat missing values is another important point to consider? 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.

32) Explain about the box cox transformation in regression models.

For some reason or the other, the response 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.

33) Can you use machine learning for time series analysis?

Yes, it can be used but it depends on the applications.

34) Write a function that takes in two sorted lists and outputs a sorted list that is their union.

First solution which will come to your mind is to merge two lists and sort them afterwards

[Would you like to rapidly solve such coding problems in your interview? Get access to 100+ solved code examples.](#)

[Click here \(these are ready-to-use for your projects\)](#)

Python code-

```
def return_union(list_a, list_b):  
    return sorted(list_a + list_b)
```

R code-

```
return_union <- function(list_a, list_b)  
{  
    list_c<-list(c(unlist(list_a),unlist(list_b)))  
    return(list(list_c[[1]][order(list_c[[1]])]))  
}
```

Generally, the tricky part of the question is not to use any sorting or ordering function. In that case you will have to write your own logic to answer the question and impress your interviewer.

Python code-

```
def return_union(list_a, list_b):  
    len1 = len(list_a)  
    len2 = len(list_b)  
    final_sorted_list = []  
    j = 0  
    k = 0  
  
    for i in range(len1+len2):  
        if k == len1:  
            final_sorted_list.extend(list_b[j:])
```

```

        break
    elif j == len2:
        final_sorted_list.extend(list_a[k:])
        break
    elif list_a[k] < list_b[j]:
        final_sorted_list.append(list_a[k])
        k += 1
    else:
        final_sorted_list.append(list_b[j])
        j += 1
return final_sorted_list

```

Similar function can be returned in R as well by following the similar steps.

```

return_union <- function(list_a,list_b)
{
#Initializing length variables
len_a <- length(list_a)
len_b <- length(list_b)
len <- len_a + len_b

```

```

#initializing counter variables

```

```

j=1
k=1

```

```

#Creating an empty list which has length equal to sum of both the lists

```

```

list_c <- list(rep(NA,len))

```

```

#Here goes our for loop

```

```

for(i in 1:len)
{
    if(j>len_a)
    {
        list_c[i:len] <- list_b[k:len_b]
        break
    }
    else if(k>len_b)
    {
        list_c[i:len] <- list_a[j:len_a]
        break
    }
    else if(list_a[[j]] <= list_b[[k]])
    {
        list_c[[i]] <- list_a[[j]]
        j <- j+1
    }
}

```



```
else if(list_a[[j]] > list_b[[k]])
{
  list_c[[i]] <- list_b[[k]]
  k <- k+1
}
}
return(list(unlist(list_c)))

}
```

35) What is the difference between Bayesian Estimate and Maximum Likelihood Estimation (MLE)?

In bayesian estimate we have some knowledge about the data/problem (prior) .There may be several values of the parameters which explain data and hence we can look for multiple parameters like 5 gammas and 5 lambdas that do this. As a result of Bayesian Estimate, we get multiple models for making multiple predictions i.e. one for each pair of parameters but with the same prior. So, if a new example need to be predicted than computing the weighted sum of these predictions serves the purpose.

Maximum likelihood does not take prior into consideration (ignores the prior) so it is like being a Bayesian while using some kind of a flat prior.

36) What is Regularization and what kind of problems does regularization solve?

37) What is multicollinearity and how you can overcome it?

38) What is the curse of dimensionality?

39) How do you decide whether your linear regression model fits the data?

40) What is the difference between squared error and absolute error?

41) What is Machine Learning?

The simplest way to answer this question is – we give the data and equation to the machine. Ask the machine to look at the data and identify the coefficient values in an equation.

For example for the linear regression $y=mx+c$, we give the data for the variable x , y and the machine learns about the values of m and c from the data.

42) How are confidence intervals constructed and how will you interpret them?

43) How will you explain logistic regression to an economist, physican scientist and biologist?

44) How can you overcome Overfitting?

45) Differentiate between wide and tall data formats?

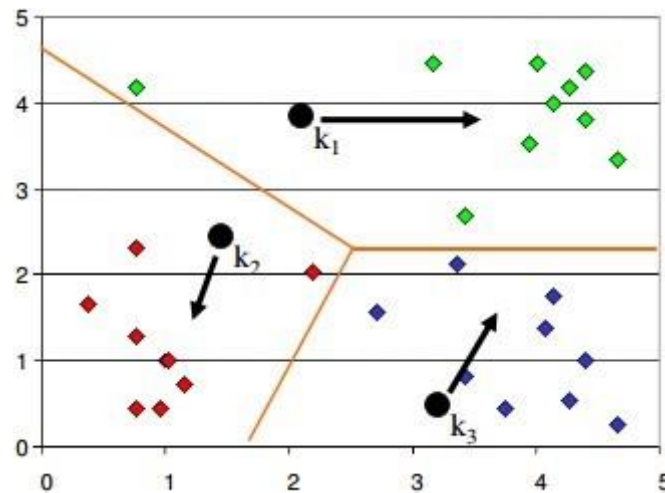
46) Is Naïve Bayes bad? If yes, under what aspects.

47) How would you develop a model to identify plagiarism?

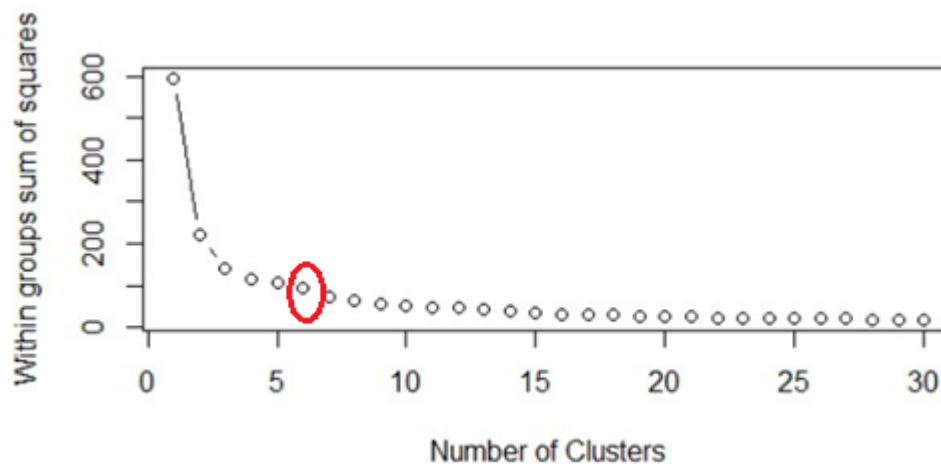
48) How will you define the number of clusters in a clustering algorithm?

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. 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. 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.

49) Is it better to have too many false negatives or too many false positives?

51) What do you understand by Fuzzy merging ? Which language will you use to handle it?

52) What is the difference between skewed and uniform distribution?

When the observations in a dataset are spread equally across the range of distribution, then it is referred to as uniform distribution. There are no clear perks in an uniform distribution. Distributions that have more observations on one side of the graph than the other are referred to as skewed distribution. Distributions with fewer observations on the left (towards lower values) are said to be skewed left and distributions with fewer observation on the right (



towards higher values) are said to be skewed right.

[The #1 question in your interview will be "What experience do you have? Get hands-on experience with free access to 100+ code examples solved by industry experts. Click here \(you can rapidly get some project experience before your interviews\)](#)

53) You created a predictive model of a quantitative outcome variable using multiple regressions. What are the steps you would follow to validate the model?

Since the question asked, is about post model building exercise, we will assume that you have already tested for null hypothesis, multi collinearity and Standard error of coefficients.

Once you have built the model, you should check for following –

- Global F-test to see the significance of group of independent variables on dependent variable
- R^2
- Adjusted R^2
- RMSE, MAPE

In addition to above mentioned quantitative metrics you should also check for-

- Residual plot
- Assumptions of linear regression

54) What do you understand by Hypothesis in the content of Machine Learning?

55) What do you understand by Recall and Precision?

Recall measures "Of all the actual true samples how many did we classify as true?"

Precision measures "Of all the samples we classified as true how many are actually true?"

We will explain this with a simple example for better understanding –

Imagine that your wife gave you surprises every year on your anniversary in last 12 years. One day all of a sudden your wife asks – "Darling, do you remember all anniversary surprises from me?".

This simple question puts your life into danger. To save your life, you need to Recall all 12 anniversary surprises from your memory. Thus, Recall(R) is the ratio of number of events you can correctly recall to the number of all correct events. If you can recall all the 12 surprises correctly then the recall ratio is 1 (100%) but if you can recall only 10 surprises correctly of the 12 then the recall ratio is 0.83 (83.3%).

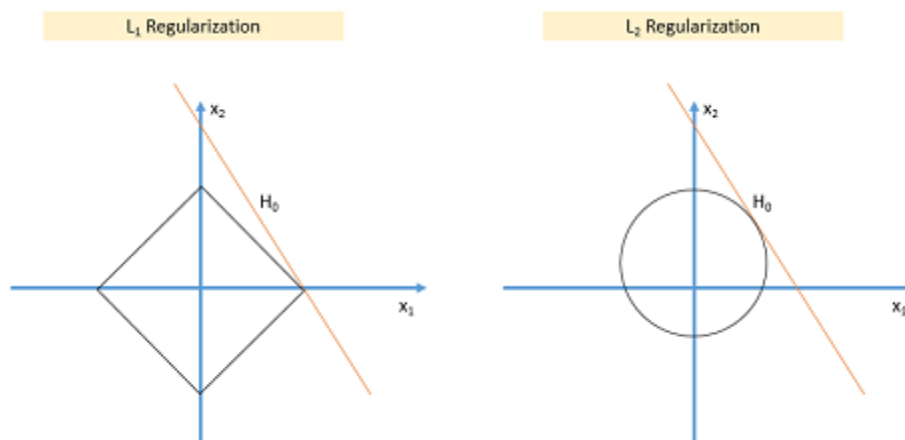
However, you might be wrong in some cases. For instance, you answer 15 times, 10 times the surprises you guess are correct and 5 wrong. This implies that your recall ratio is 100% but the precision is 66.67%.

Precision is the ratio of number of events you can correctly recall to a number of all events you recall (combination of wrong and correct recalls).

56) How will you find the right K for K-means?

57) Why L1 regularizations causes parameter sparsity whereas L2 regularization does not?

Regularizations in statistics or in the field of machine learning is used to include some extra information in order to solve a problem in a better way. L1 & L2 regularizations are generally used to add constraints to optimization problems.



In the example shown above H_0 is a hypothesis. If you observe, in L_1 there is a high likelihood to hit the corners as solutions while in L_2 , it doesn't. So in L_1 variables are penalized more as compared to L_2 which results into sparsity.

In other words, errors are squared in L_2 , so model sees higher error and tries to minimize that squared error.

58) How can you deal with different types of seasonality in time series modelling? ([get 100+ solved code examples here](#))

Seasonality in time series occurs when time series shows a repeated pattern over time. E.g., stationary sales decreases during holiday season, air conditioner sales increases during the summers etc. are few examples of seasonality in a time series.

Seasonality makes your time series non-stationary because average value of the variables at different time periods. Differentiating a time series is generally known as the best method of removing seasonality from a time series. Seasonal differencing can be defined as a numerical difference between a particular value and a value with a periodic lag (i.e. 12, if monthly seasonality is present)

59) In experimental design, is it necessary to do randomization? If yes, why?

60) What do you understand by conjugate-prior with respect to Naïve Bayes?

61) Can you cite some examples where a false positive is important than a false negative?

Before we start, let us understand what are false positives and what are false negatives.



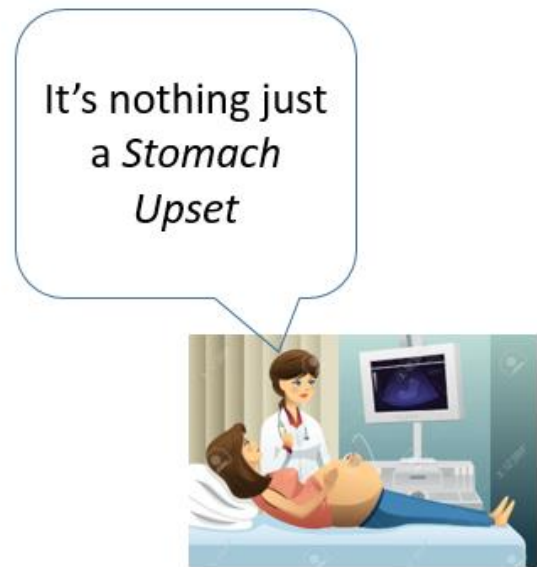
False Positives are the cases where you wrongly classified a non-event as an event a.k.a Type I error.

And, False Negatives are the cases where you wrongly classify events as non-events, a.k.a Type II error.

False Positive



False Negative



In medical field, assume you have to give chemo therapy to patients. Your lab tests patients for certain vital information and based on those results they decide to give radiation therapy to a patient.

Assume a patient comes to that hospital and he is tested positive for cancer (But he doesn't have cancer) based on lab prediction. What will happen to him? (Assuming Sensitivity is 1)

One more example might come from marketing. Let's say an ecommerce company decided to give \$1000 Gift voucher to the customers whom they assume to purchase at least \$5000 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 5K.

Now what if they have sent it to false positive cases?

62) Can you cite some examples where a false negative important than a false positive? ([get 100+ solved code examples here](#))

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 shortage of staff they decided to scan passenger 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?

Another example can be judicial system. What if Jury or judge decide to make a criminal go free?

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

63) 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 of time they don't want to acquire bad customers. In this scenario both the false positives and false negatives become **very important to measure**.

These days we hear many cases of players using steroids during sport competitions Every player has to go through a steroid test before the game starts. A false positive can ruin the career of a Great sportsman and a false negative can make the game unfair.

[Get hands-on experience for your interviews with free access to solved code examples found here \(these are ready-to-use for your projects\)](#)

64) Can you explain the difference between a Test Set and a Validation Set?

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, 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.
- Test Set is to assess the performance of the model i.e. evaluating the predictive power and generalization.
- Validation set is to tune the parameters.

65) What makes a dataset gold standard?

66) What do you understand by statistical power of sensitivity and how do you calculate it?

Sensitivity is commonly used to validate the accuracy of a classifier (Logistic, SVM, RF etc.). Sensitivity is nothing but "Predicted TRUE events/ Total events". True events here are the events which were true and model also predicted them as true.

Calculation of sensitivity is pretty straight forward-

Sensitivity = True Positives /Positives in Actual Dependent Variable

Where, True positives are Positive events which are correctly classified as Positives.

67) What is the importance of having a selection bias? [\(get 100+ solved code examples here\)](#)



Selection Bias occurs when there is no appropriate randomization achieved while selecting individuals, groups or data to be analysed. Selection bias implies that the obtained sample does not exactly represent the population that was actually intended to be analyzed. Selection bias consists of Sampling Bias, Data, Attribute and Time Interval.

68) Give some situations where you will use an SVM over a RandomForest Machine Learning algorithm and vice-versa.

SVM and Random Forest are both used in classification problems.

- a) If you are sure that your data is outlier free and clean then go for SVM. It is the opposite - if your data might contain outliers then Random forest would be the best choice
- b) Generally, SVM consumes more computational power than Random Forest, so if you are constrained with memory go for Random Forest [machine learning algorithm](#).
- c) Random Forest gives you a very good idea of variable importance in your data, so if you want to have variable importance then choose Random Forest machine learning algorithm.
- d) Random Forest machine learning algorithms are preferred for multiclass problems.
- e) SVM is preferred in multi-dimensional problem set - like text classification

but as a good data scientist, you should experiment with both of them and test for accuracy or rather you can use ensemble of many Machine Learning techniques.

69) What do you understand by feature vectors?

70) How do data management procedures like missing data handling make selection bias worse?

Missing value treatment is one of the primary tasks which a data scientist is supposed to do before starting data analysis. There are multiple methods for missing value treatment. If not done properly, it could potentially result into selection bias. Let see few missing value treatment examples and their impact on selection-

Complete Case Treatment: Complete case treatment is when you remove entire row in data even if one value is missing. You could achieve a selection bias if your values are not missing at random and they have some pattern. Assume you are conducting a survey and few people didn't specify their gender. Would you remove all those people? Can't it tell a different story?

Available case analysis: Let say you are trying to calculate correlation matrix for data so you might remove the missing values from variables which are needed for that particular correlation coefficient. In this case your values will not be fully correct as they are coming from population sets.

Mean Substitution: In this method missing values are replaced with mean of other available values. This might make your distribution biased e.g., standard deviation, correlation and regression are mostly dependent on the mean value of variables.

Hence, various data management procedures might include selection bias in your data if not chosen correctly.

71) What are the advantages and disadvantages of using regularization methods like Ridge Regression?

72) What do you understand by long and wide data formats?



73) What do you understand by outliers and inliers? What would you do if you find them in your dataset?

74) Write a program in Python which takes input as the diameter of a coin and weight of the coin and produces output as the money value of the coin.

75) What are the basic assumptions to be made for linear regression? ([get sample code here](#))
Normality of error distribution, statistical independence of errors, linearity and additivity.

76) Can you write the formula to calculate R-square?

R-Square can be calculated using the below formula -

$1 - (\text{Residual Sum of Squares} / \text{Total Sum of Squares})$

77) What is the advantage of performing dimensionality reduction before fitting an SVM?

Support Vector Machine Learning Algorithm performs better in the reduced space. It is beneficial to perform dimensionality reduction before fitting an SVM if the number of features is large when compared to the number of observations.

78) How will you assess the statistical significance of an insight whether it is a real insight or just by chance?

Statistical importance of an insight can be assessed using Hypothesis Testing.

79) How would you create a taxonomy to identify key customer trends in unstructured data?

[Tweet: Data Science Interview questions #1 - How would you create a taxonomy to identify key customer trends in unstructured data? - http://ctt.ec/sdqZ0+](#)

The best way to approach this question is to mention that it is good to check with the business owner and understand their objectives before categorizing the data. Having done this, it is always good to follow an iterative approach by pulling new data samples and improving the model accordingly by validating it for accuracy by soliciting feedback from the stakeholders of the business. This helps ensure that your model is producing actionable results and improving over the time.

80) How will you find the correlation between a categorical variable and a continuous variable?

You can use the analysis of covariance technique to find the correlation between a categorical variable and a continuous variable.

Q: What do you understand by the Selection Bias? What are its various types?

A: Selection bias is typically associated with research that doesn't have a random selection of participants. It is a type of error that occurs when a researcher decides who is going to be studied. On some occasions, selection bias is also referred to as the selection effect.

In other words, selection bias is a distortion of statistical analysis that results from the sample collecting method. When selection bias is not taken into account, some conclusions made by a research study might not be accurate. Following are the various types of selection bias:



- **Sampling Bias** – A systematic error resulting due to a non-random sample of a TOP MENTOR populace causing certain members of the same to be less likely included than others that results in a biased sample.
- **Time Interval** – A trial might be ended at an extreme value, usually due to ethical reasons, but the extreme value is most likely to be reached by the variable with the most variance, even though all variables have a similar mean.
- **Data** – Results when specific data subsets are selected for supporting a conclusion or rejection of bad data arbitrarily.
- **Attrition** – Caused due to attrition, i.e. loss of participants, discounting trial subjects or tests that didn't run to completion.

Q: Please explain the goal of A/B Testing.

A: A/B Testing is a statistical hypothesis testing meant for a randomized experiment with two variables, A and B. The goal of A/B Testing is to maximize the likelihood of an outcome of some interest by identifying any changes to a webpage.

A highly reliable method for finding out the best online marketing and promotional strategies for a business, A/B Testing can be employed for testing everything, ranging from sales emails to search ads and website copy.

Q: How will you calculate the Sensitivity of machine learning models?

A: In machine learning, Sensitivity is used for validating the accuracy of a classifier, such as Logistic, Random Forest, and SVM. It is also known as REC (recall) or TPR (true positive rate).

Sensitivity can be defined as the ratio of predicted true events and total events i.e.:

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{Positives in Actual Dependent Variable}}$$

Here, true events are those events that were true as predicted by a machine learning model. The best sensitivity is 1.0 and the worst sensitivity is 0.0.

Q: Could you draw a comparison between overfitting and underfitting?

A: In order to make reliable predictions on general untrained data in machine learning and statistics, it is required to fit a (machine learning) model to a set of training data. Overfitting and underfitting are two of the most common modeling errors that occur while doing so.

Following are the various differences between overfitting and underfitting:

- **Definition** – A statistical model suffering from overfitting describes some random error or noise in place of the underlying relationship. When underfitting occurs, a statistical model or machine learning algorithm fails in capturing the underlying trend of the data.
- **Occurrence** – When a statistical model or machine learning algorithm is excessively complex, it can result in overfitting. Example of a complex model is one having too many parameters when compared to the total number of observations. Underfitting occurs when trying to fit a linear model to non-linear data.



- **Poor Predictive Performance** – Although both overfitting and underfitting yield poor predictive performance, the way in which each one of them does so is different. While the overfitted model overreacts to minor fluctuations in the training data, the underfit model under-reacts to even bigger fluctuations.

Q: Between Python and R, which one would you pick for text analytics and why?

A: For text analytics, [Python](#) will gain an upper hand over R due to these reasons:

- The Pandas library in Python offers easy-to-use data structures as well as high-performance data analysis tools
- Python has a faster performance for all types of text analytics
- [R](#) is a best-fit for machine learning than mere text analysis

Q: Please explain the role of data cleaning in data analysis.

A: Data cleaning can be a daunting task due to the fact that with the increase in the number of data sources, the time required for cleaning the data increases at an exponential rate.

This is due to the vast volume of data generated by additional sources. Also, data cleaning can solely take up to 80% of the total time required for carrying out a data analysis task.

Nevertheless, there are several reasons for using data cleaning in data analysis. Two of the most important ones are:

- Cleaning data from different sources helps in transforming the data into a format that is easy to work with
- Data cleaning increases the accuracy of a machine learning model

Q: What do you mean by cluster sampling and systematic sampling?

A: When studying the target population spread throughout a wide area becomes difficult and applying simple random sampling becomes ineffective, the technique of cluster sampling is used. A cluster sample is a probability sample, in which each of the sampling units is a collection or cluster of elements.

Following the technique of systematic sampling, elements are chosen from an ordered sampling frame. The list is advanced in a circular fashion. This is done in such a way so that once the end of the list is reached, the same is progressed from the start, or top, again.

Q: Please explain Eigenvectors and Eigenvalues.

A: Eigenvectors help in understanding linear transformations. They are calculated typically for a correlation or covariance matrix in data analysis.



In other words, eigenvectors are those directions along which some particular linear transformation acts by compressing, flipping, or stretching.

Eigenvalues can be understood either as the strengths of the transformation in the direction of the eigenvectors or the factors by which the compressions happens.

Q: Can you compare the validation set with the test set?

A: A validation set is part of the training set used for parameter selection as well as for avoiding overfitting of the machine learning model being developed. On the contrary, a test set is meant for evaluating or testing the performance of a trained machine learning model.

Q: What do you understand by linear regression and logistic regression?

A: Linear regression is a form of statistical technique in which the score of some variable Y is predicted on the basis of the score of a second variable X, referred to as the predictor variable. The Y variable is known as the criterion variable.

Also known as the logit model, logistic regression is a statistical technique for predicting the binary outcome from a linear combination of predictor variables.

Q: Please explain Recommender Systems along with an application.

A: Recommender Systems is a subclass of information filtering systems, meant for predicting the preferences or ratings awarded by a user to some product.

An application of a recommender system is the product recommendations section in Amazon. This section contains items based on the user's search history and past orders.

Q: What are outlier values and how do you treat them?

A: Outlier values, or simply outliers, are data points in statistics that don't belong to a certain population. An outlier value is an abnormal observation that is very much different from other values belonging to the set.

Identification of outlier values can be done by using univariate or some other graphical analysis method. Few outlier values can be assessed individually but assessing a large set of outlier values require the substitution of the same with either the 99th or the 1st percentile values.

There are two popular ways of treating outlier values:

1. To change the value so that it can be brought within a range
2. To simply remove the value

Note: - Not all extreme values are outlier values.

Q: Please enumerate the various steps involved in an analytics project.

A: Following are the numerous steps involved in an analytics project:

- Understanding the business problem
- Exploring the data and familiarizing with the same
- Preparing the data for modeling by means of detecting outlier values, transforming variables, treating missing values, et cetera
- Running the model and analyzing the result for making appropriate changes or modifications to the model (an iterative step that repeats until the best possible outcome is gained)
- Validating the model using a new dataset
- Implementing the model and tracking the result for analyzing the performance of the same

Q: Could you explain how to define the number of clusters in a clustering algorithm?

A: The primary objective of clustering is to group together similar identities in such a way that while entities within a group are similar to each other, the groups remain different from one another.

Generally, [Within Sum of Squares](#) is used for explaining the homogeneity within a cluster. For defining the number of clusters in a clustering algorithm, WSS is plotted for a range pertaining to a number of clusters. The resultant graph is known as the Elbow Curve.

The Elbow Curve graph contains a point that represents the point post in which there aren't any decrements in the WSS. This is known as the bending point and represents K in K-Means.

Although the aforementioned is the widely-used approach, another important approach is the Hierarchical clustering. In this approach, dendrograms are created first and then distinct groups are identified from there.

Q: What do you understand by Deep Learning?

A: Deep Learning is a paradigm of machine learning that displays a great degree of analogy with the functioning of the human brain. It is a neural network method based on convolutional neural networks (CNN).

Deep learning has a wide array of uses, ranging from social network filtering to medical image analysis and speech recognition. Although Deep Learning has been present for a long time, it's only recently that it has gained worldwide acclaim. This is mainly due to:

- An increase in the amount of data generation via various sources
- The growth in hardware resources required for running Deep Learning models

Caffe, Chainer, Keras, Microsoft Cognitive Toolkit, Pytorch, and TensorFlow are some of the most popular Deep Learning frameworks as of today.

Q: Please explain Gradient Descent.

A: The degree of change in the output of a function relating to the changes made to the inputs is known as a gradient. It measures the change in all weights with respect to the change in error. A gradient can also be comprehended as the slope of a function.

Gradient Descent refers to escalating down to the bottom of a valley. Simply, consider this something as opposed to climbing up a hill. It is a minimization algorithm meant for minimizing a given activation function.

Q: How does Backpropagation work? Also, it state its various variants.

A: Backpropagation refers to a training algorithm used for multilayer neural networks. Following the backpropagation algorithm, the error is moved from an end of the network to all weights inside the network. Doing so allows for efficient computation of the gradient.

Backpropagation works in the following way:

- Forward propagation of training data
- Output and target is used for computing derivatives
- Backpropagate for computing the derivative of the error with respect to the output activation
- Using previously calculated derivatives for output generation
- Updating the weights

Following are the various variants of Backpropagation:

- **Batch Gradient Descent** – The gradient is calculated for the complete dataset and update is performed on each iteration
- **Mini-batch Gradient Descent** – Mini-batch samples are used for calculating gradient and updating parameters (a variant of the Stochastic Gradient Descent approach)
- **Stochastic Gradient Descent** – Only a single training example is used to calculate gradient and updating parameters

Q: What do you know about Autoencoders?

A: Autoencoders are simplistic learning networks used for transforming inputs into outputs with minimum possible error. It means that the outputs resulted are very close to the inputs.

A couple of layers are added between the input and the output with the size of each layer smaller than the size pertaining to the input layer. An autoencoder receives unlabeled input that is encoded for reconstructing the output.

Q: Please explain the concept of a Boltzmann Machine.



A: A Boltzmann Machine features a simple learning algorithm that enables the same to discover fascinating features representing complex regularities present in the training data. It is basically used for optimizing the quantity and weight for some given problem.

The simple learning algorithm involved in a Boltzmann Machine is very slow in networks that have many layers of feature detectors.

That completes the list of the 20 essential data science interview questions. I hope you will find it useful to prepare well for your upcoming data science job interview(s). Wish you good luck!

1. From the below given 'diamonds' dataset, extract only those rows where the 'price' value is greater than 1000 and the 'cut' is ideal.

carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75
0.24	Very Good	J	VVS2	62.8	57.0	336	3.94	3.96	2.48
0.24	Very Good	I	VVS1	62.3	57.0	336	3.95	3.98	2.47
0.26	Very Good	H	SI1	61.9	55.0	337	4.07	4.11	2.53
0.22	Fair	E	VS2	65.1	61.0	337	3.87	3.78	2.49

First, we will load the **ggplot2** package:

```
library(ggplot2)
```

Next, we will use the **dplyr** package:

```
library(dplyr)// It is based on the grammar of data manipulation.
```

To extract those particular records, use the below command:

```
diamonds %>% filter(price>1000 & cut=="Ideal")-> diamonds_1000_idea
```

2. Make a scatter plot between 'price' and 'carat' using ggplot. 'Price' should be on y-axis, 'carat' should be on x-axis, and the 'color' of the points should be determined by 'cut.'

We will implement the scatter plot using **ggplot**.



The ggplot is based on the grammar of data visualization, and it helps us stack multiple layers on top of each other.

So, we will start with the data layer, and on top of the data layer we will stack the aesthetic layer. Finally, on top of the aesthetic layer we will stack the geometry layer.

Code:

```
>ggplot(data=diamonds, aes(x=carat, y=price, col=cut))+geom_point()
```

3. Introduce 25 percent missing values in this 'iris' dataset and impute the 'Sepal.Length' column with 'mean' and the 'Petal.Length' column with 'median.'

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa

To introduce missing values, we will be using the **missForest** package:

```
library(missForest)
```

Using the prodNA function, we will be introducing 25 percent of missing values:

```
Iris.mis<-prodNA(iris,noNA=0.25)
```

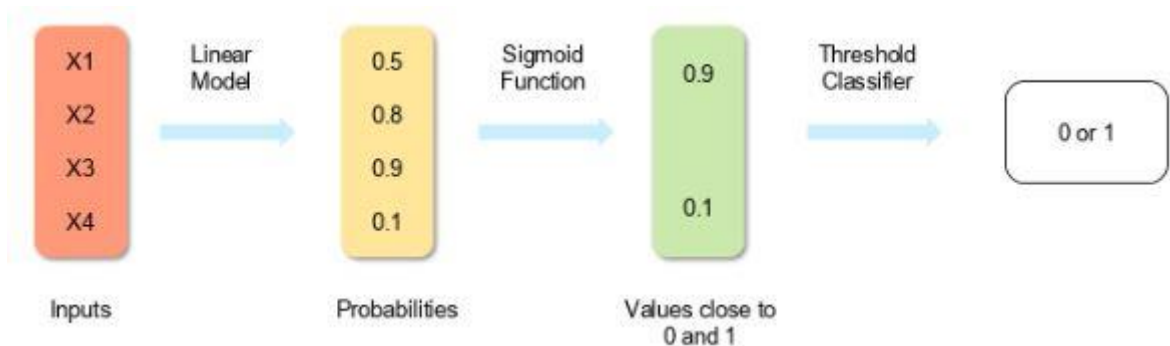
For imputing the 'Sepal.Length' column with 'mean' and the 'Petal.Length' column with 'median,' we will be using the Hmisc package and the impute function:

```
library(Hmisc)
iris.mis$Sepal.Length<-with(iris.mis, impute(Sepal.Length,mean))
iris.mis$Petal.Length<-with(iris.mis, impute(Petal.Length,median))
```

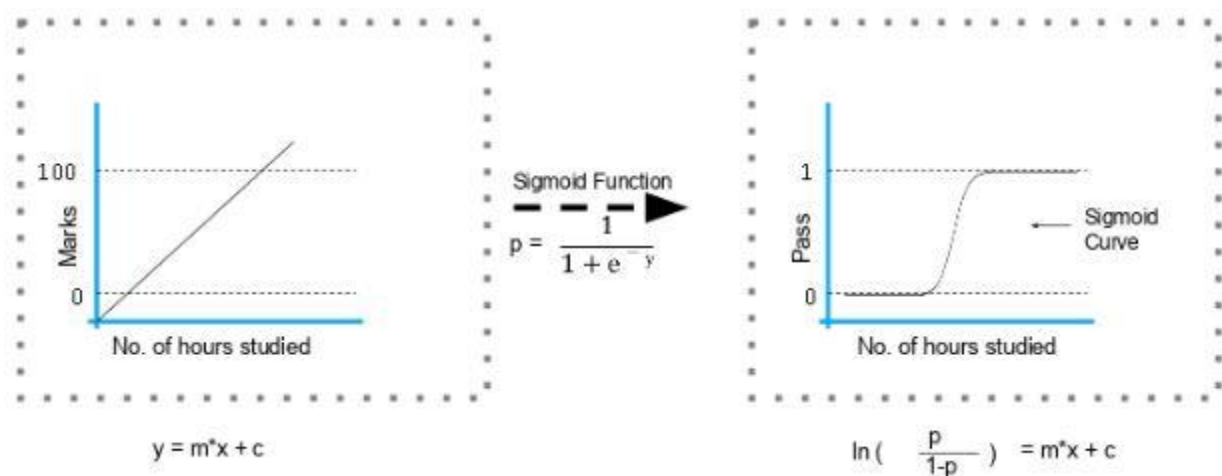
2. How is logistic regression done?

Logistic regression measures the relationship between the dependent variable (our label of what we want to predict) and one or more independent variables (our features) by estimating probability using its underlying logistic function (sigmoid).

The image shown below depicts how logistic regression works:



The formula and graph for the sigmoid function are as shown:



3. Explain the steps in making a decision tree.

1. Take the entire data set as input
2. Calculate entropy of the target variable, as well as the predictor attributes
3. Calculate your information gain of all attributes (we gain information on sorting different objects from each other)

4. Choose the attribute with the highest information gain as the root node
5. Repeat the same procedure on every branch until the decision node of each branch is finalized

For example, let's say you want to build a decision tree to decide whether you should accept or decline a job offer. The decision tree for this case is as shown:



It is clear from the decision tree that an offer is accepted if:

- Salary is greater than \$50,000
- The commute is less than an hour
- Incentives are offered

4. How do you build a random forest model?

A [random forest](#) is built up of a number of decision trees. If you split the data into different packages and make a decision tree in each of the different groups of data, the random forest brings all those trees together.

Steps to build a random forest model:

1. Randomly select 'k' features from a total of 'm' features where $k \ll m$
2. Among the 'k' features, calculate the node D using the best split point
3. Split the node into daughter nodes using the best split
4. Repeat steps two and three until leaf nodes are finalized



5. Build forest by repeating steps one to four for 'n' times to create 'n' number of trees

5. How can you avoid the overfitting of your model?

Overfitting refers to a model that is only set for a very small amount of data and ignores the bigger picture. There are three main methods to avoid overfitting:

1. Keep the model simple—take fewer variables into account, thereby removing some of the noise in the training data
2. Use cross-validation techniques, such as k folds cross-validation
3. Use regularization techniques, such as LASSO, that penalize certain model parameters if they're likely to cause overfitting

6. Differentiate between univariate, bivariate, and multivariate analysis.

Univariate

Univariate data contains only one variable. The purpose of the univariate analysis is to describe the data and find patterns that exist within it.

Example: height of students

Height (in cm)
164
167.3
170

174.2

178

180

The patterns can be studied by drawing conclusions using mean, median, mode, dispersion or range, minimum, maximum, etc.

Bivariate

Bivariate data involves two different variables. The analysis of this type of data deals with causes and relationships and the analysis is done to determine the relationship between the two variables.

Example: temperature and ice cream sales in the summer season

Temperature (in Celcius)	Sales
20	2,000
25	2,100
26	2,300

28	2,400
30	2,600
36	3,100

Here, the relationship is visible from the table that temperature and sales are directly proportional to each other. The hotter the temperature, the better the sales.

Multivariate

Multivariate data involves three or more variables, it is categorized under multivariate. It is similar to a bivariate, but contains more than one dependent variable.

Example: data for house price prediction

No. of rooms	Floors	Area (sq ft)	Price
2	0	900	\$4000,00
3	2	1,100	\$600,000
3.5	5	1,500	\$900,000



4	3	2,100	\$1,200,000
---	---	-------	-------------

The patterns can be studied by drawing conclusions using mean, median, and mode, dispersion or range, minimum, maximum, etc. You can start describing the data and using it to guess what the price of the house will be.

7. What are the feature selection methods used to select the right variables?

There are two main methods for feature selection:

Filter Methods

This involves:

- Linear discrimination analysis
- ANOVA
- Chi-Square

The best analogy for selecting features is "bad data in, bad answer out." When we're limiting or selecting the features, it's all about cleaning up the data coming in.

Wrapper Methods

This involves:

- Forward Selection: We test one feature at a time and keep adding them until we get a good fit
- Backward Selection: We test all the features and start removing them to see what works better
- Recursive Feature Elimination: Recursively looks through all the different features and how they pair together

Wrapper methods are very labor-intensive, and high-end computers are needed if a lot of data analysis is performed with the wrapper method.



8. In your choice of language, write a program that prints the numbers ranging from one to 50. **INTOR**

But for multiples of three, print "Fizz" instead of the number and for the multiples of five, print "Buzz." For numbers which are multiples of both three and five, print "FizzBuzz"

The code is shown below:

```
for fizzbuzz in range(51):
    if fizzbuzz % 3 == 0 and fizzbuzz % 5 == 0:
        print("fizzbuzz")
        continue
    elif fizzbuzz % 3 == 0:
        print("fizz")
        continue
    elif fizzbuzz % 5 == 0:
        print("buzz")
        continue
    print(fizzbuzz)
```

Note that the range mentioned is 51, which means zero to 50. However, the range asked in the question is one to 50. Therefore, in the above code, you can include the range as (1,51).

The output of the above code is as shown:

```
fizzbuzz
1
2
fizz
4
buzz
fizz
7
8
fizz
buzz
11
fizz
13
14
fizzbuzz
16
17
fizz
19
buzz
fizz
22
23
fizz
buzz
26
...
46
47
fizz
49
buzz
```



9. You are given a data set consisting of variables with more than 30 percent missing values. How will you deal with them?

The following are ways to handle missing data values:

If the data set is large, we can just simply remove the rows with missing data values. It is the quickest way; we use the rest of the data to predict the values.

For smaller data sets, we can substitute missing values with the mean or average of the rest of the data using pandas data frame in python. There are different ways to do so, such as `df.mean()`, `df.fillna(mean)`.

10. For the given points, how will you calculate the Euclidean distance in Python?

```
plot1 = [1,3]
```

```
plot2 = [2,5]
```

The Euclidean distance can be calculated as follows:

```
euclidean_distance = sqrt( (plot1[0]-plot2[0])**2 + (plot1[1]-plot2[1])**2 )
```

11. What are dimensionality reduction and its benefits?

Dimensionality reduction refers to the process of converting a data set with vast dimensions into data with fewer dimensions (fields) to convey similar information concisely.

This reduction helps in compressing data and reducing storage space. It also reduces computation time as fewer dimensions lead to less computing. It removes redundant features; for example, there's no point in storing a value in two different units (meters and inches).

12. How will you calculate eigenvalues and eigenvectors of the following 3x3 matrix?

-2	-4	2
----	----	---

-2	1	2
4	2	5

The characteristic equation is as shown:

Expanding determinant:

$$(-2 - \lambda) [(1-\lambda) (5-\lambda) - 2 \times 2] + 4[(-2) \times (5-\lambda) - 4 \times 2] + 2[(-2) \times 2 - 4(1-\lambda)] = 0$$

$$-\lambda^3 + 4\lambda^2 + 27\lambda - 90 = 0,$$

$$\lambda^3 - 4\lambda^2 - 27\lambda + 90 = 0$$

Here we have an algebraic equation built from the eigenvectors.

By hit and trial:

$$3^3 - 4 \times 3^2 - 27 \times 3 + 90 = 0$$

Hence, $(\lambda - 3)$ is a factor:

$$\lambda^3 - 4\lambda^2 - 27\lambda + 90 = (\lambda - 3) (\lambda^2 - \lambda - 30)$$

Eigenvalues are 3, -5, 6:

$$(\lambda - 3) (\lambda^2 - \lambda - 30) = (\lambda - 3) (\lambda + 5) (\lambda - 6),$$

Calculate eigenvector for $\lambda = 3$

For $X = 1$,

$$-5 - 4Y + 2Z = 0,$$



$$-2 - 2Y + 2Z = 0$$

Subtracting the two equations:

$$3 + 2Y = 0,$$

Subtracting back into second equation:

$$Y = -(3/2)$$

$$Z = -(1/2)$$

Similarly, we can calculate the eigenvectors for -5 and 6 .

13. How should you maintain a deployed model?

The steps to maintain a deployed model are:

Monitor

Constant monitoring of all models is needed to determine their performance accuracy. When you change something, you want to figure out how your changes are going to affect things. This needs to be monitored to ensure it's doing what it's supposed to do.

Evaluate

Evaluation metrics of the current model are calculated to determine if a new algorithm is needed.

Compare

The new models are compared to each other to determine which model performs the best.

Rebuild

The best performing model is re-built on the current state of data.



14. What are recommender systems?

A recommender system predicts what a user would rate a specific product based on their preferences. It can be split into two different areas:

Collaborative filtering

As an example, Last.fm recommends tracks that other users with similar interests play often. This is also commonly seen on Amazon after making a purchase; customers may notice the following message accompanied by product recommendations: "Users who bought this also bought..."

Content-based filtering

As an example: Pandora uses the properties of a song to recommend music with similar properties. Here, we look at content, instead of looking at who else is listening to music.

15. How do you find RMSE and MSE in a linear regression model?

RMSE and MSE are two of the most common measures of accuracy for a linear regression model.

RMSE indicates the Root Mean Square Error.

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (Predicted_i - Actual_i)^2}{N}}$$

MSE indicates the Mean Square Error.

$$MSE = \frac{\sum_{i=1}^N (Predicted_i - Actual_i)^2}{N}$$



16. How can you select k for k-means?

We use the elbow method to select k for k-means clustering. The idea of the elbow method is to run k-means clustering on the data set where 'k' is the number of clusters.

Within the sum of squares (WSS), it is defined as the sum of the squared distance between each member of the cluster and its centroid.

17. What is the significance of p-value?

p-value typically ≤ 0.05

This indicates strong evidence against the null hypothesis; so you reject the null hypothesis.

p-value typically > 0.05

This indicates weak evidence against the null hypothesis, so you accept the null hypothesis.

p-value at cutoff 0.05

This is considered to be marginal, meaning it could go either way.

18. How can outlier values be treated?

You can drop outliers only if it is a garbage value.

Example: height of an adult = abc ft. This cannot be true, as the height cannot be a string value. In this case, outliers can be removed.

If the outliers have extreme values, they can be removed. For example, if all the data points are clustered between zero to 10, but one point lies at 100, then we can remove this point.

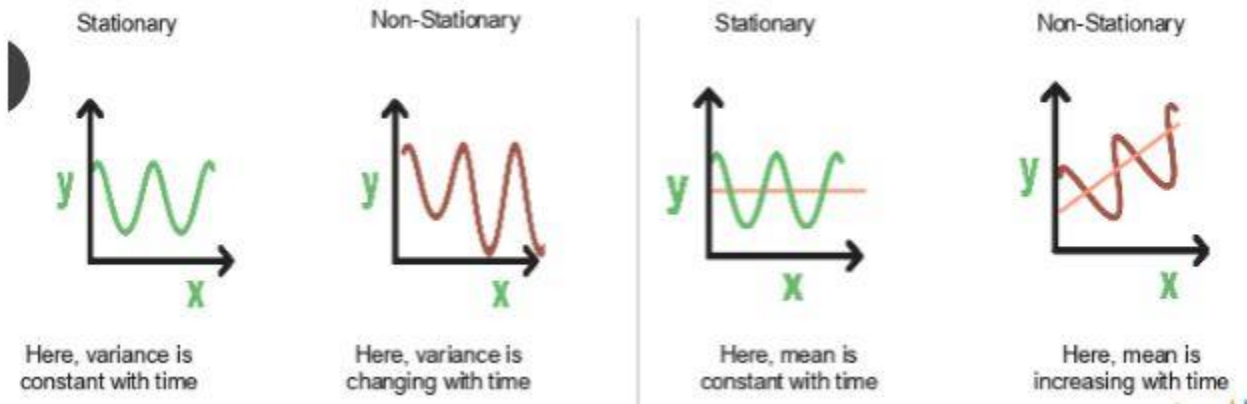
If you cannot drop outliers, you can try the following:

- Try a different model. Data detected as outliers by linear models can be fit by nonlinear models. Therefore, be sure you are choosing the correct model.
- Try normalizing the data. This way, the extreme data points are pulled to a similar range.
- You can use algorithms that are less affected by outliers; an example would be random forests.

19. How can a time-series data be declared as stationary?

It is stationary when the variance and mean of the series are constant with time.

Here is a visual example:



In the first graph, the variance is constant with time. Here, X is the time factor and Y is the variable. The value of Y goes through the same points all the time; in other words, it is stationary.

In the second graph, the waves get bigger, which means it is non-stationary and the variance is changing with time.

20. How can you calculate accuracy using a confusion matrix?

Consider this confusion matrix:

Total=650		actual	
		p	n
predicted	P	262	15
	N	26	347

True Positive (arrow from 262)
 False Negative (arrow from 26)
 False Positive (arrow from 15)
 True Negative (arrow from 347)

You can see the values for total data, actual values, and predicted values.

The formula for accuracy is:

Accuracy = (True Positive + True Negative) / Total Observations

$$= (262 + 347) / 650$$

$$= 609 / 650$$

$$= 0.93$$

As a result, we get an accuracy of 93 percent.

21. Write the equation and calculate the precision and recall rate.

Consider the same confusion matrix used in the previous question.

Total=650		actual	
		p	n
predicted	P	262	15
	N	26	347

True Positive (arrow from 262)
 False Negative (arrow from 26)
 False Positive (arrow from 15)
 True Negative (arrow from 347)

Precision = (True positive) / (True Positive + False Positive)



$$= 262 / 277$$

$$= 0.94$$

Recall Rate = (True Positive) / (Total Positive + False Negative)

$$= 262 / 288$$

$$= 0.90$$

22. 'People who bought this also bought...' recommendations seen on Amazon are a result of which algorithm?

The recommendation engine is accomplished with collaborative filtering. Collaborative filtering explains the behavior of other users and their purchase history in terms of ratings, selection, etc.

The engine makes predictions on what might interest a person based on the preferences of other users. In this algorithm, item features are unknown.

For example, a sales page shows that a certain number of people buy a new phone and also buy tempered glass at the same time. Next time, when a person buys a phone, he or she may see a recommendation to buy tempered glass as well.

23. Write a basic SQL query that lists all orders with customer information.

Usually, we have order tables and customer tables that contain the following columns:

Order Table

Orderid

customerId

OrderNumber



TotalAmount

Customer Table

Id

FirstName

LastName

City

Country

The SQL query is:

```
SELECT OrderNumber, TotalAmount, FirstName, LastName, City, Country
```

```
FROM Order
```

```
JOIN Customer
```

```
ON Order.CustomerId = Customer.Id
```

24. You are given a dataset on cancer detection. You have built a classification model and achieved an accuracy of 96 percent. Why shouldn't you be happy with your model performance? What can you do about it?

Cancer detection results in imbalanced data. In an imbalanced dataset, accuracy should not be based as a measure of performance. It is important to focus on the remaining four percent, which represents the patients who were wrongly diagnosed. Early diagnosis is crucial when it comes to cancer detection, and can greatly improve a patient's prognosis.

Hence, to evaluate model performance, we should use Sensitivity (True Positive Rate), Specificity (True Negative Rate), F measure to determine the class wise performance of the classifier.



25. Which of the following machine learning algorithms can be used for inputting missing values of both categorical and continuous variables?

- K-means clustering
- Linear regression
- K-NN (k-nearest neighbor)
- Decision trees

The K nearest neighbor algorithm can be used because it can compute the nearest neighbor and if it doesn't have a value, it just computes the nearest neighbor based on all the other features.

When you're dealing with K-means clustering or linear regression, you need to do that in your pre-processing, otherwise, they'll crash. Decision trees also have the same problem, although there is some variance.

26. Below are the eight actual values of the target variable in the train file. What is the entropy of the target variable?

[0, 0, 0, 1, 1, 1, 1, 1]

Choose the correct answer.

1. $-(5/8 \log(5/8) + 3/8 \log(3/8))$
2. $5/8 \log(5/8) + 3/8 \log(3/8)$
3. $3/8 \log(5/8) + 5/8 \log(3/8)$
4. $5/8 \log(3/8) - 3/8 \log(5/8)$

The target variable, in this case, is 1.

The formula for calculating the entropy is:

Putting $p=5$ and $n=8$, we get

Entropy = $A = -(5/8 \log(5/8) + 3/8 \log(3/8))$



27. We want to predict the probability of death from heart disease based on three risk factors: age, gender, and blood cholesterol level. What is the most appropriate algorithm for this case?

Choose the correct option:

1. Logistic Regression
2. Linear Regression
3. K-means clustering
4. Apriori algorithm

The most appropriate algorithm for this case is A, logistic regression.

28. After studying the behavior of a population, you have identified four specific individual types that are valuable to your study. You would like to find all users who are most similar to each individual type. Which algorithm is most appropriate for this study?

Choose the correct option:

1. K-means clustering
2. Linear regression
3. Association rules
4. Decision trees

As we are looking for grouping people together specifically by four different similarities, it indicates the value of k. Therefore, K-means clustering (answer A) is the most appropriate algorithm for this study.

29. You have run the association rules algorithm on your dataset, and the two rules {banana, apple} => {grape} and {apple, orange} => {grape} have been found to be relevant. What else must be true?

Choose the right answer:

1. {banana, apple, grape, orange} must be a frequent itemset
2. {banana, apple} => {orange} must be a relevant rule
3. {grape} => {banana, apple} must be a relevant rule



4. {grape, apple} must be a frequent itemset

The answer is A: {grape, apple} must be a frequent itemset

30. Your organization has a website where visitors randomly receive one of two coupons. It is also possible that visitors to the website will not receive a coupon. You have been asked to determine if offering a coupon to website visitors has any impact on their purchase decisions. Which analysis method should you use?

1. One-way ANOVA
2. K-means clustering
3. Association rules
4. Student's t-test

The answer is A: One-way ANOVA

Additional Questions on Basic Data Science Concepts

31. What are the feature vectors?

A feature vector is an n-dimensional vector of numerical features that represent an object. In machine learning, feature vectors are used to represent numeric or symbolic characteristics (called features) of an object in a mathematical way that's easy to analyze.

32. What are the steps in making a decision tree?

1. Take the entire data set as input.
2. Look for a split that maximizes the separation of the classes. A split is any test that divides the data into two sets.
3. Apply the split to the input data (divide step).
4. Re-apply steps one and two to the divided data.
5. Stop when you meet any stopping criteria.
6. This step is called pruning. Clean up the tree if you went too far doing splits.



33. What is root cause analysis?

Root cause analysis was initially developed to analyze industrial accidents but is now widely used in other areas. It is a problem-solving technique used for isolating the root causes of faults or problems. A factor is called a root cause if its deduction from the problem-fault-sequence averts the final undesirable event from recurring.

34. What is logistic regression?

Logistic regression is also known as the logit model. It is a technique used to forecast the binary outcome from a linear combination of predictor variables.

35. 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.

36. Explain cross-validation.

Cross-validation is a model validation technique for evaluating how the outcomes of a statistical analysis will generalize to an independent data set. It is mainly used in backgrounds where the objective is to 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) to limit problems like overfitting and gain insight into how the model will generalize to an independent data set.

37. What is collaborative filtering?

Most recommender systems use this filtering process to find patterns and information by collaborating perspectives, numerous data sources, and several agents.

38. Do gradient descent methods always converge to similar points?

They do not, because in some cases, they reach a local minima or a local optima point. You would not reach the global optima point. This is governed by the data and the starting conditions.



39. What is the goal of A/B Testing?

This is statistical hypothesis testing for randomized experiments with two variables, A and B. The objective of A/B testing is to detect any changes to a web page to maximize or increase the outcome of a strategy.

40. What are the drawbacks of the linear model?

- The assumption of linearity of the errors
- It can't be used for count outcomes or binary outcomes
- There are overfitting problems that it can't solve

41. What is the law of large numbers?

It is a theorem that describes the result of performing the same experiment very frequently. This theorem forms the basis of frequency-style thinking. It states that the sample mean, sample variance and sample standard deviation converge to what they are trying to estimate.

42. What are the confounding variables?

These are extraneous variables in a statistical model that correlates directly or inversely with both the dependent and the independent variable. The estimate fails to account for the confounding factor.

43. What is 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.

44. How regularly must an algorithm be updated?

You will want to update an algorithm when:

- You want the model to evolve as data streams through infrastructure



- The underlying data source is changing
- There is a case of non-stationarity

45. What are eigenvalue and eigenvector?

Eigenvalues are the directions along which a particular linear transformation acts by flipping, compressing, or stretching.

Eigenvectors are for understanding linear transformations. In data analysis, we usually calculate the eigenvectors for a correlation or covariance matrix.

46. Why is resampling done?

Resampling is done in any of these cases:

- Estimating the accuracy of sample statistics by using subsets of accessible data, or drawing randomly with replacement from a set of data points
- Substituting labels on data points when performing significance tests
- Validating models by using random subsets (bootstrapping, cross-validation)

47. What is selection bias?

Selection bias, in general, is a problematic situation in which error is introduced due to a non-random population sample.

48. What are the types of biases that can occur during sampling?

1. Selection bias
2. Undercoverage bias
3. Survivorship bias

49. What is survivorship bias?

Survivorship bias is the logical error of focusing aspects that support surviving a process and casually overlooking those that did not because of their lack of prominence. This can lead to wrong conclusions in numerous ways.



50. How do you work towards a random forest?

The underlying principle of this technique is that several weak learners combine to provide a strong learner. The steps involved are:

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