500 Data Science Interview Questions

Vamsee Puligadda

500
Data
Science
Interview
Questions
Vamsee Puligadda

500 Data Science Interview Questions and Answers

Vamsee Puligadda

Disclaimer

Although the author and publisher have made every effort to ensure that the information in this book was correct at press time, the author and publisher do not assume and hereby disclaim any liability to any party for any loss, damage, or disruption caused by errors or omissions, whether such errors or omissions result from negligence, accident, or any other cause.

For any suggestions/concerns, please write to us: indianwolfpublications@gmail.com

Copyright 2018 Vamsee Puligadda & Indian Wolf Publications

Foreword

I am neither the owner of a famous publishing company nor am a top IT company with hundreds of in-house developers to create anything I wanted to with a great ease.

I am an independent Software developer with passion towards what I do and trust me, a lot of time, efforts were put into creating this extensive collection of questions and answers at a single place.

If it helps at least a few in their careers to achieve their dream jobs, I will be more than happy.

Thank You.

- Vamsee Puligadda.

Question 1. What Is A Recommender System?

Answer:

A recommender system is today widely deployed in multiple fields like movie recommendations, music preferences, social tags, research articles, search queries and so on. The recommender systems work as per collaborative and content-based filte

ring or by deploying a personality-based approach. This type of system works based on a person's past behavior in order to build a model for the future. This will predict the future product buying, movie viewing or book reading by people. It also creates a filtering approach using the discrete characteristics of items while recommending additional items.

Question 2. Compare Sas, R And Python Programming?

Answer:

SAS: it is one of the most widely used analytics tools used by some of the biggest companies on earth. It has some of the best statistical functions, graphical user interface, but can come with a price tag and hence it cannot be readily adopted by smaller enterprises

R: The best part about R is that it is an Open Source tool and hence used generously by academia and the research community. It is a robust tool for statistical computation, graphical representation and reporting. Due to its open source nature it is always being updated with the latest features and then readily available to everybody.

Python: Python is a powerful open source programming language that is easy to learn, works well with most other tools and technologies. The best part about Python is that it has innumerable libraries and community created modules making it very robust. It has functions for statistical operation, model building and more.

Question 3. Explain The Various Benefits Of R Language?

Answer:

The R programming language includes a set of software suite that is used for graphical representation, statistical computing, data manipulation and calculation.

Some of the highlights of R programming environment include the following:

An extensive collection of tools for data analysis

Operators for performing calculations on matrix and array

Data analysis technique for graphical representation

A highly developed yet simple and effective programming language
It extensively supports machine learning applications
It acts as a connecting link between various software, tools and datasets
Create high quality reproducible analysis that is flexible and powerful
Provides a robust package ecosystem for diverse needs
It is useful when you have to solve a data-oriented problem

Question 4. How Do Data Scientists Use Statistics?

Answer:

Statistics helps Data Scientists to look into the data for patterns, hidden insights and convert Big Data into Big insights. It helps to get a better idea of what the customers are expecting. Data Scientists can learn about the consumer behavior, interest, engagement, retention and finally conversion all through the power of insightful statistics. It helps them to build powerful data models in order to validate certain inferences and predictions. All this can be converted into a powerful business proposition by giving users what they want at precisely when they want it.

Question 5. What Is Logistic Regression?

Answer:

It is a statistical technique or a model in order to analyze a dataset and predict the binary outcome. The outcome has to be a binary outcome that is either zero or one or a yes or no.

Question 6. Why Data Cleansing Is Important In Data Analysis?

Answer:

With data coming in from multiple sources it is important to ensure that data is good enough for analysis. This is where data cleansing becomes extremely vital. Data cleansing extensively deals with the process of detecting and correcting of data records, ensuring that data is complete and accurate and the components of data that are irrelevant are deleted or modified as per the needs. This process can be deployed in concurrence with data wrangling or batch processing.

Once the data is cleaned it confirms with the rules of the data sets in the system. Data cleansing is an essential part of the data science because the data can be prone to error due to human negligence, corruption during transmission or storage among other things. Data cleansing takes a huge chunk of time and effort of a Data Scientist because of the multiple sources from which data emanates and the speed at which it comes.

Question 7. Describe Univariate, Bivariate And Multivariate Analysis.?

Answer:

As the name suggests these are analysis methodologies having a single, double or multiple variables.

So a univariate analysis will have one variable and due to this there are no relationships, causes. The major aspect of the univariate analysis is to summarize the data and find the patterns within it to make actionable decisions.

A Bivariate analysis deals with the relationship between two sets of data. These sets of paired data come from related sources, or samples. There are various tools to analyze such data including the chi-squared tests and t-tests when the data are having a correlation.

If the data can be quantified then it can analyzed using a graph plot or a scatterplot. The strength of the correlation between the two data sets will be tested in a Bivariate analysis.

Question 8. How Machine Learning Is Deployed In Real World Scenarios?

Answer:

Here are some of the scenarios in which machine learning finds applications in real world:

Ecommerce: Understanding the customer churn, deploying targeted advertising, remarketing.

Search engine: Ranking pages depending on the personal preferences of the searcher

Finance: Evaluating investment opportunities & risks, detecting fraudulent transactions

Medicare: Designing drugs depending on the patient's history and needs

Robotics: Machine learning for handling situations that are out of the ordinary

Social media: Understanding relationships and recommending connections

Extraction of information: framing questions for getting answers from databases over the web.

Question 9. What Are The Various Aspects Of A Machine Learning Process?

Answer:

In this post I will discuss the components involved in solving a problem using machine learning.

Domain knowledge:

This is the first step wherein we need to understand how to extract the various features from the data and learn more about the data that we are dealing with. It has got more to do with the type of domain that we are dealing with and

familiarizing the system to learn more about it.

Feature Selection:

This step has got more to do with the feature that we are selecting from the set of features that we have. Sometimes it happens that there are a lot of features and we have to make an intelligent decision regarding the type of feature that we want to select to go ahead with our machine learning endeavor.

Algorithm:

This is a vital step since the algorithms that we choose will have a very major impact on the entire process of machine learning. You can choose between the linear and nonlinear algorithm. Some of the algorithms used are Support Vector Machines, Decision Trees, Naïve Bayes, K-Means Clustering, etc.

Training:

This is the most important part of the machine learning technique and this is where it differs from the traditional programming. The training is done based on the data that we have and providing more real world experiences. With each consequent training step the machine gets better and smarter and able to take improved decisions.

Evaluation:

In this step we actually evaluate the decisions taken by the machine in order to decide whether it is up to the mark or not. There are various metrics that are involved in this process and we have to closed deploy each of these to decide on the efficacy of the whole machine learning endeavor.

Optimization:

This process involves improving the performance of the machine learning process using various optimization techniques. Optimization of machine learning is one of the most vital components wherein the performance of the algorithm is vastly improved. The best part of optimization techniques is that machine learning is not just a consumer of optimization techniques but it also provides new ideas for optimization too.

Testing:

Here various tests are carried out and some these are unseen set of test cases. The data is partitioned into test and training set. There are various testing techniques like cross-validation in order to deal with multiple situations.

Question 10. What Do You Understand By The Term Normal Distribution?

Answer:

It is a set of continuous variable spread across a normal curve or in the shape of a bell curve. It can be considered as a continuous probability distribution

and is useful in statistics. It is the most common distribution curve and it becomes very useful to analyze the variables and their relationships when we have the normal distribution curve.

The normal distribution curve is symmetrical. The non-normal distribution approaches the normal distribution as the size of the samples increases. It is also very easy to deploy the Central Limit Theorem. This method helps to make sense of data that is random by creating an order and interpreting the results using a bell-shaped graph.

Question 11. What Is Linear Regression?

Answer:

It is the most commonly used method for predictive analytics. The Linear Regression method is used to describe relationship between a dependent variable and one or independent variable. The main task in the Linear Regression is the method of fitting a single line within a scatter plot.

The Linear Regression consists of the following three methods:

Determining and analyzing the correlation and direction of the data

Deploying the estimation of the model

Ensuring the usefulness and validity of the model

It is extensively used in scenarios where the cause effect model comes into play. For example you want to know the effect of a certain action in order to determine the various outcomes and extent of effect the cause has in determining the final outcome.

Question 12. What Is Interpolation And Extrapolation?

Answer:

The terms of interpolation and extrapolation are extremely important in any statistical analysis. Extrapolation is the determination or estimation using a known set of values or facts by extending it and taking it to an area or region that is unknown. It is the technique of inferring something using data that is available.

Interpolation on the other hand is the method of determining a certain value which falls between a certain set of values or the sequence of values.

This is especially useful when you have data at the two extremities of a certain region but you don’t have enough data points at the specific point. This is when you deploy interpolation to determine the value that you need.

Question 13. What Is Power Analysis?

Answer:

The power analysis is a vital part of the experimental design. It is involved

with the process of determining the sample size needed for detecting an effect of a given size from a cause with a certain degree of assurance. It lets you deploy specific probability in a sample size constraint.

The various techniques of statistical power analysis and sample size estimation are widely deployed for making statistical judgment that are accurate and evaluate the size needed for experimental effects in practice.

Power analysis lets you understand the sample size estimate so that they are neither high nor low. A low sample size there will be no authentication to provide reliable answers and if it is large there will be wastage of resources.

Question 14. What Is K-means? How Can You Select K For K-means?

Answer:

K-means clustering can be termed as the basic unsupervised learning algorithm. It is the method of classifying data using a certain set of clusters called as K clusters. It is deployed for grouping data in order to find similarity in the data.

It includes defining the K centers, one each in a cluster. The clusters are defined into K groups with K being predefined. The K points are selected at random as cluster centers. The objects are assigned to their nearest cluster center. The objects within a cluster are as closely related to one another as possible and differ as much as possible to the objects in other clusters. K-means clustering works very well for large sets of data.

Question 15. How Is Data Modeling Different From Database Design?

Answer:

Data Modeling: It can be considered as the first step towards the design of a database. Data modeling creates a conceptual model based on the relationship between various data models. The process involves moving from the conceptual stage to the logical model to the physical schema. It involves the systematic method of applying the data modeling techniques.

Database Design: This is the process of designing the database. The database design creates an output which is a detailed data model of the database. Strictly speaking database design includes the detailed logical model of a database but it can also include physical design choices and storage parameters.

Question 16. What Are Feature Vectors?

Answer:

n-dimensional vector of numerical features that represent some object

Term occurrences frequencies, pixels of an image etc.

Feature space: vector space associated with these vectors

Question 17. Explain The Steps In Making A Decision Tree.?

Answer:

Take the entire data set as input

Look for a split that maximizes the separation of the classes. A split is any test that divides the data in two sets

Apply the split to the input data (divide step)

Re-apply steps 1 to 2 to the divided data

Stop when you meet some stopping criteria

This step is called pruning. Clean up the tree when you went too far doing splits.

Question 18. What Is Root Cause Analysis?

Answer:

Root cause analysis was initially developed to analyze industrial accidents, but is now widely used in other areas. It is basically a technique of problem solving 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 reoccurring.

Question 19. Explain Cross-validation.?

Answer:

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

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

Question 20. What Is Collaborative Filtering?

Answer:

The process of filtering used by most of the recommender systems to find patterns or information by collaborating perspectives, numerous data sources and several agents.

Question 21. Do Gradient Descent Methods At All Times Converge To Similar Point?

Answer:

No, they do not because in some cases it reaches a local minima or a local optima point. You will not reach the global optima point. This is governed by the data and the starting conditions.

Question 22. What Is The Goal Of A/b Testing?

Answer:

It is a statistical hypothesis testing for randomized experiment with two variables A and B. The objective of A/B Testing is to detect any changes to the web page to maximize or increase the outcome of an interest.

Question 23. What Are The Drawbacks Of Linear Model?

Answer:

Some drawbacks of the linear model are:

The assumption of linearity of the errors

It can't be used for count outcomes, binary outcomes

There are overfitting problems that it can't solve

Question 24. What Is The Law Of Large Numbers?

Answer:

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

Question 25. What Are Confounding Variables?

Answer:

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

Question 26. Explain Star Schema.?

Answer:

It is a traditional database schema with a central table. Satellite tables map ID's to physical name or description 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.

Question 27. How Regularly An Algorithm Must Be Update?

Answer:

You 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

Question 28. What Are Eigenvalue And Eigenvector?

Answer:

Eigenvectors are 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.

Question 29. Why Is Resampling Done?

Answer:

Resampling is done in one 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.

Question 30. Explain Selective Bias.?

Answer:

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

Question 31. What Are The Types Of Biases That Can Occur During Sampling?

Answer:

- Selection bias
- Under coverage bias
- Survivorship bias

Question 32. How To Work Towards A Random Forest?

Answer:

Underlying principle of this technique is that several weak learners combined provide a strong learner. The steps involved are

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

32) 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.

34) 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.

35) 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.

36) Why data cleaning plays a vital role in analysis?

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.

37) 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.

38) 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.

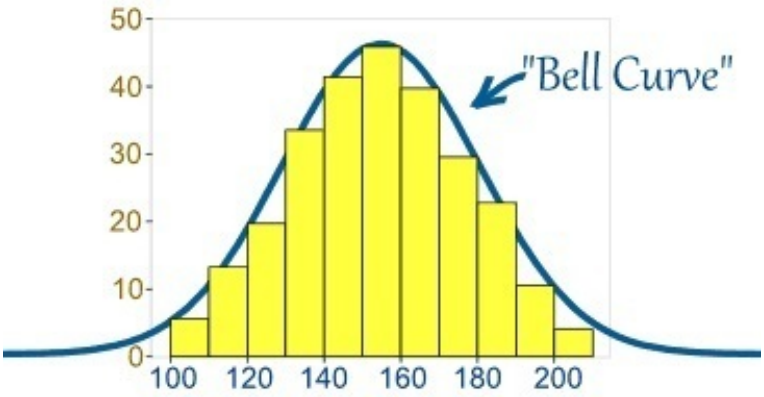


Image Credit : mathisfun.com

39) 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.

40) 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.

41) What is power analysis?

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

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

43) 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.

44) 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.

45) 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.

46) 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- Value > 0.05 denotes weak evidence against the null hypothesis which means the null hypothesis cannot be rejected.
- P-value <= 0.05 denotes strong evidence against the null hypothesis which means the null hypothesis can be rejected.
- P-value=0.05is the marginal value indicating it is possible to go either way.

47) 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

48) What are categorical variables?

49) 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.

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

51) What is the difference between Supervised Learning an 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.

52) Explain the use of Combinatorics in data science.

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

54) 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.

55) 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.

56) What is Gradient Descent?

57) 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.

58) 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.

59) 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.

60) 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.

61) During analysis, how do you treat missing values?

The extent of the missing values is identified after identifying the variables with missing values. If any patterns are identified the analyst has to concentrate on them as it could lead to interesting and meaningful business insights. If there are no patterns identified, then the missing values can be substituted with mean or median values (imputation) or they can simply be ignored.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.

62) 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-mornla 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.

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

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

64) 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 short them afterwards

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)))

  }
```

65) 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 predcitions 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.

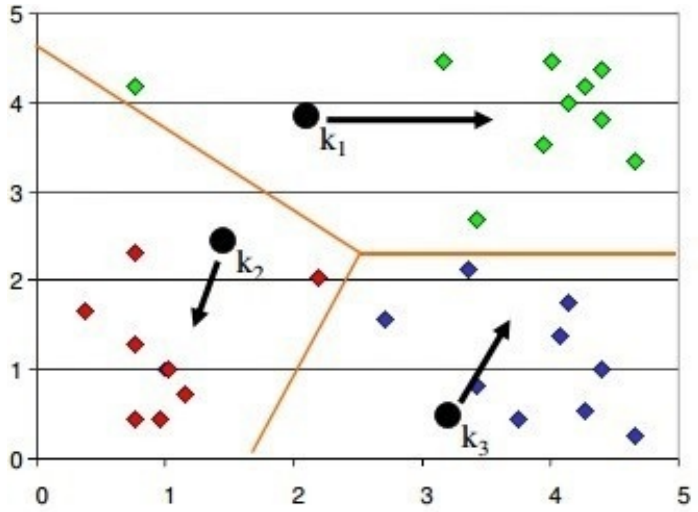
Maximum likelihood does not take prior into consideration (ignores the prior) so it is like being a Bayesian while using some kind of a flat prior.

66) What is Regularization and what kind of problems does regularization solve?
67) What is multicollinearity and how you can overcome it?
68) What is the curse of dimensionality?
69) How do you decide whether your linear regression model fits the data?
70) What is the difference between squared error and absolute error?
71) 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.

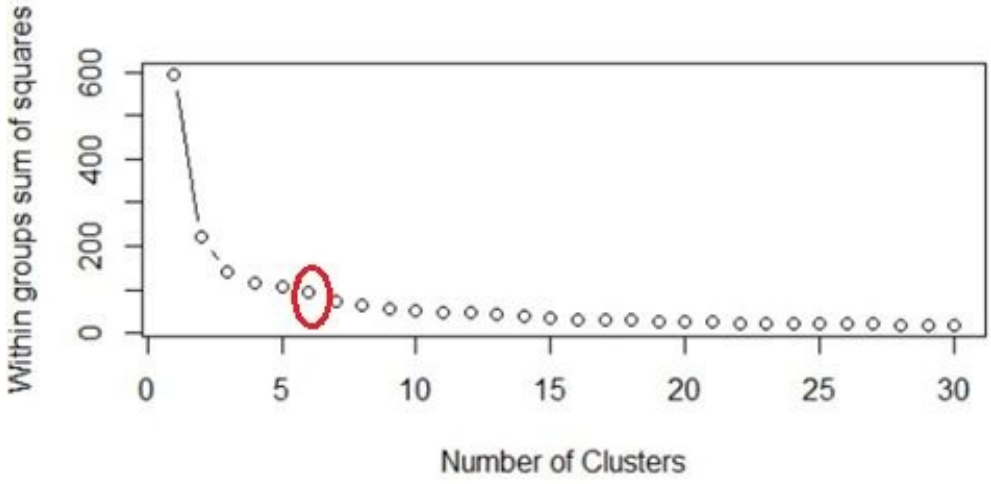
For example for the linear regression y=mx+c, we give the data for the variable x, y and the machine learns about the values of m and c from the data.

72) How are confidence intervals constructed and how will you interpret them?
73) How will you explain logistic regression to an economist, physican scientist and biologist?
74) How can you overcome Overfitting?
75) Differentiate between wide and tall data formats?
76) Is Naïve Bayes bad? If yes, under what aspects.
77) How would you develop a model to identify plagiarism?
78) 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.

79) Is it better to have too many false negatives or too many false positives?
80) Is it possible to perform logistic regression with Microsoft Excel?
It is possible to perform logistic regression with Microsoft Excel. There are two ways to do it using Excel.

a) One is to use Add-ins provided by many websites which we can use.

b) Second is to use fundamentals of logistic regression and use Excel's computational power to build a logistic regression

But when this question is being asked in an interview, interviewer is not looking for a name of Add-ins rather a method using the base excel functionalities.

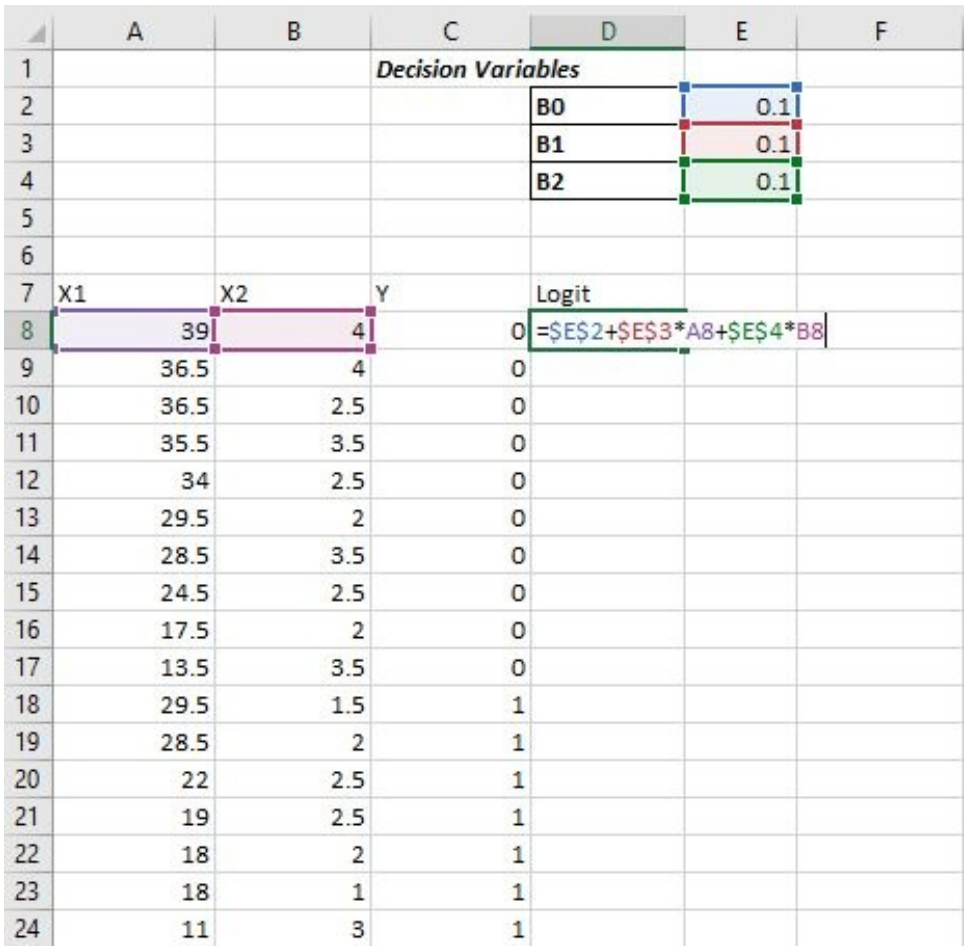
Let's use a sample data to learn about logistic regression using Excel. (Example assumes that you are familiar with basic concepts of logistic regression)

	A	B	C
6			
7	X1	X2	Y
8	39	4	0
9	36.5	4	0
10	36.5	2.5	0
11	35.5	3.5	0
12	34	2.5	0
13	29.5	2	0
14	28.5	3.5	0
15	24.5	2.5	0
16	17.5	2	0
17	13.5	3.5	0
18	29.5	1.5	1
19	28.5	2	1
20	22	2.5	1
21	19	2.5	1
22	18	2	1
23	18	1	1
24	11	3	1
25	11	2.5	1
26	7.5	2	1
27	5	3	1

Data shown above consists of three variables where X1 and X2 are independent variables and Y is a class variable. We have kept only 2 categories for our purpose of binary logistic regression classifier.

Next we have to create a logit function using independent variables, i.e.

Logit = L = $\beta_0 + \beta_1 * X_1 + \beta_2 * X_2$



	A	B	C	D	E	F
1			*Decision Variables*			
2				B0	0.1	
3				B1	0.1	
4				B2	0.1	
5						
6						
7	X1	X2	Y	Logit		
8	39	4	0	=$E$2+$E$3*A8+$E$4*B8		
9	36.5	4	0			
10	36.5	2.5	0			
11	35.5	3.5	0			
12	34	2.5	0			
13	29.5	2	0			
14	28.5	3.5	0			
15	24.5	2.5	0			
16	17.5	2	0			
17	13.5	3.5	0			
18	29.5	1.5	1			
19	28.5	2	1			
20	22	2.5	1			
21	19	2.5	1			
22	18	2	1			
23	18	1	1			
24	11	3	1			

We have kept the initial values of beta 1, beta 2 as 0.1 for now and we will use Excel Solve to optimize the beta values in order to maximize our log likelihood estimate.

Assuming that you are aware of logistic regression basics, we calculate probability values from Logit using following formula:

Probability= e^Logit/(1+ e^Logit)

e is base of natural logarithm i.e. e = 2.71828163
Let’s put it into excel formula to calculate probability values for each of the observation.

	A	B	C	D	E	F
1			*Decision Variables*			
2				**B0**	0.1	
3				**B1**	0.1	
4				**B2**	0.1	
5						
6						
7	X1	X2	Y	Logit	Probability	
8	39	4	0	4.4	=EXP(D8)/(1+EXP(D8))	
9	36.5	4	0	4.15		
10	36.5	2.5	0	4		
11	35.5	3.5	0	4		
12	34	2.5	0	3.75		
13	29.5	2	0	3.25		
14	28.5	3.5	0	3.3		
15	24.5	2.5	0	2.8		
16	17.5	2	0	2.05		
17	13.5	3.5	0	1.8		
18	29.5	1.5	1	3.2		
19	28.5	2	1	3.15		
20	22	2.5	1	2.55		
21	19	2.5	1	2.25		
22	18	2	1	2.1		
23	18	1	1	2		
24	11	3	1	1.5		

The conditional probability is the probability of Predicted Y, given set of independent variables X.

And this p can be calculated as-

P (X) ^Yactual*[1-P (X) ^(1-Yactual)]

Then we have to take natural log of the above function-

ln [P (X) ^Yactual*[1-P(X)^(1-Yactual)]]

Which turns out to be –

Yactual*ln [P(X)]*(Yactual- 1)*ln[1-P(X)]

Log likelihood function LL is the sum of above equation for all the observations

	A	B	C	D	E	F	G
1			***Decision Variables***				
2				**B0**	0.1		
3				**B1**	0.1		
4				**B2**	0.1		
5							
6							
7	X1	X2	Y	Logit	Probability	P(Y=y|X)	
8	39	4	0	4.4	0.987871565	=C8*LN(E8)+(1-C8)*LN(1-E8)	
9	36.5	4	0	4.15	0.984480243		
10	36.5	2.5	0	4	0.98201379		
11	35.5	3.5	0	4	0.98201379		
12	34	2.5	0	3.75	0.97702263		
13	29.5	2	0	3.25	0.962673113		
14	28.5	3.5	0	3.3	0.964428811		
15	24.5	2.5	0	2.8	0.942675824		
16	17.5	2	0	2.05	0.885947619		
17	13.5	3.5	0	1.8	0.858148935		
18	29.5	1.5	1	3.2	0.960834277		
19	28.5	2	1	3.15	0.958908722		
20	22	2.5	1	2.55	0.927573515		
21	19	2.5	1	2.25	0.904650535		
22	18	2	1	2.1	0.890903179		
23	18	1	1	2	0.880797078		
24	11	3	1	1.5	0.817574476		

Log likelihood LL will be sum of column G, which we just calculated

	A	B	C	D	E	F	G	H	I
1			*Decision Variables*				*Log Likelihood*		
2				**B0**	0.1			=SUM(F8:F27)	
3				**B1**	0.1				
4				**B2**	0.1				
5									
6									
7	X1	X2	Y	Logit	Probability	P(Y=y|X)			
8	39	4	0	4.4	0.987871565	-4.41220258			
9	36.5	4	0	4.15	0.984480243	-4.16564145			
10	36.5	2.5	0	4	0.98201379	-4.01814993			
11	35.5	3.5	0	4	0.98201379	-4.01814993			
12	34	2.5	0	3.75	0.97702263	-3.77324546			
13	29.5	2	0	3.25	0.962673113	-3.28804137			
14	28.5	3.5	0	3.3	0.964428811	-3.33621926			
15	24.5	2.5	0	2.8	0.942675824	-2.85903283			
16	17.5	2	0	2.05	0.885947619	-2.17109745			
17	13.5	3.5	0	1.8	0.858148935	-1.95297761			
18	29.5	1.5	1	3.2	0.960834277	-0.03995333			
19	28.5	2	1	3.15	0.958908722	-0.04195939			
20	22	2.5	1	2.55	0.927573515	-0.07518323			
21	19	2.5	1	2.25	0.904650535	-0.10020656			
22	18	2	1	2.1	0.890903179	-0.11551952			
23	18	1	1	2	0.880797078	-0.12692801			
24	11	3	1	1.5	0.817574476	-0.20141328			

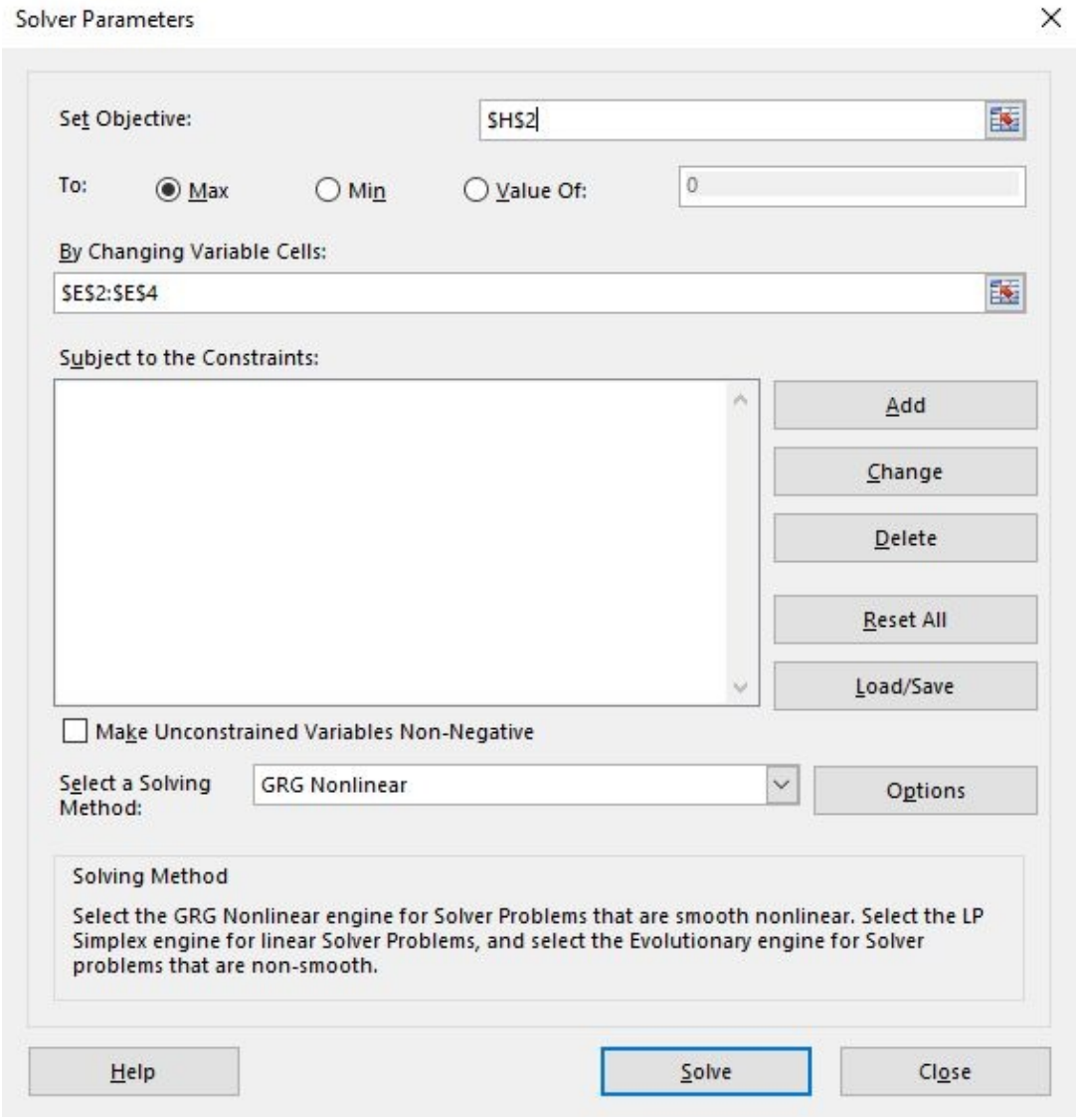
The objective is to maximize the Log Likelihood i.e. cell H2 in this example. We have to maximize H2 by optimizing B_0, B_1, and B_2.
We'll use Excel's solver add-in to achieve the same.

Excel comes with this Add-in pre-installed and you must see it under Data Tab in Excel as shown below

If you don't see it there then make sure if you have loaded it. To load an add-in in Excel,

Go to File >> Options >> Add-Ins and see if checkbox in front of required add-in is checked or not? Make sure to check it to load an add-in into Excel.
If you don't see Solver Add-in there, go to the bottom of the screen (Manage Add-Ins) and click on OK. Next you will see a popup window which should have your Solver add-in present. Check the checkbox in-front of the add-in name. If you don't see it there as well click on browse and direct it to the required folder which contains Solver Add-In.

Once you have your Solver loaded, click on Solver icon under Data tab and You will see a new window popped up like –

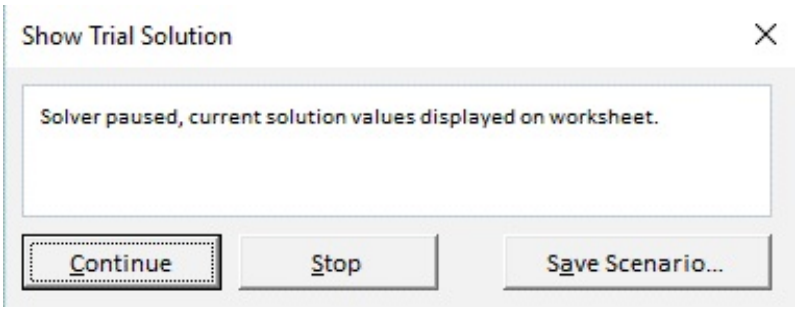


Put H2 in set objective, select max and fill cells E2 to E4 in next form field.

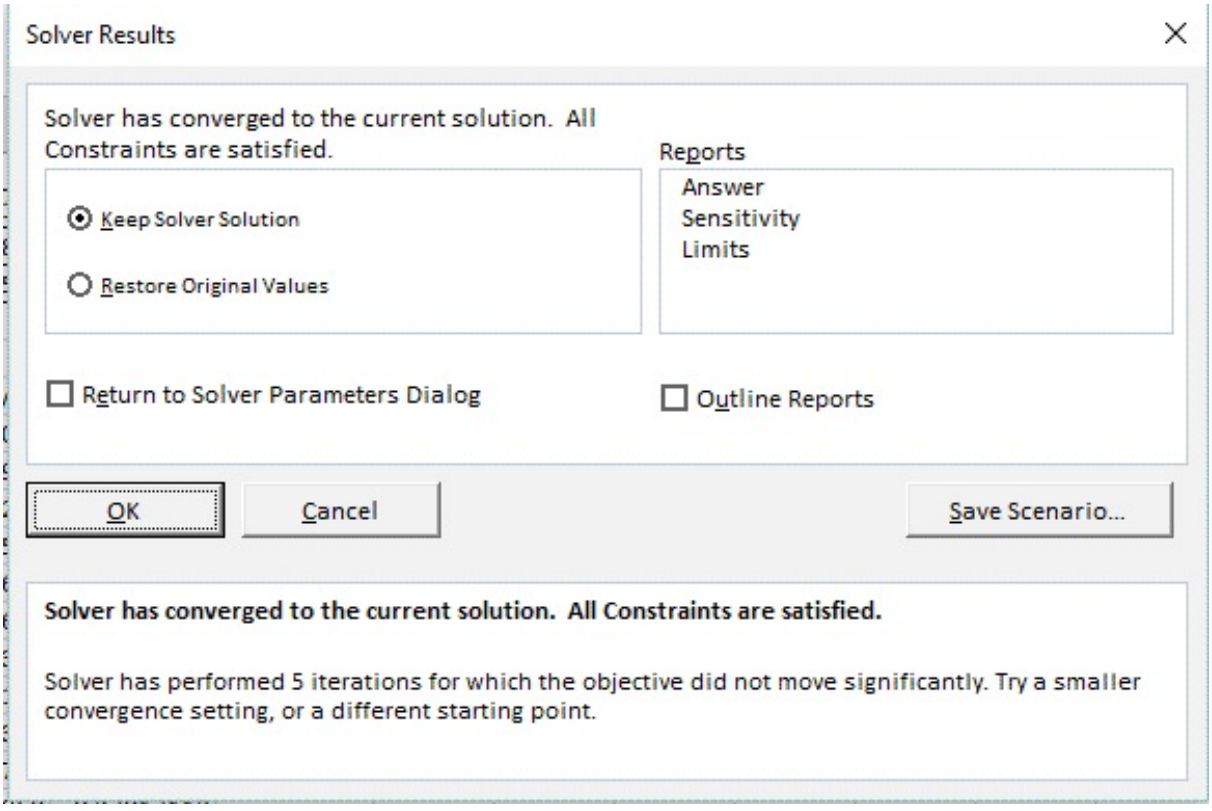
By doing this we have told Solver to Maximize H2 by changing values in cells E2 to E4.

Now click on Solve button at the bottom –

You will see a popup like below -



This shows that Solver has found a local maxima solution but we are in need of Global Maxima Output. Keep clicking on Continue until it shows the below popup



It shows that Solver was able to find and converge the solution. In case it is not able to converge it will throw an error. Select “Keep Solver Solution” and Click on OK to accept the solution provided by Solver.

Now, you can see that value of Beta coefficients from $B_{0,}$ B_1 B_2 have changed and our Log Likelihood function has been maximized.

	A	B	C	D	E	F	G	H
1			*Decision Variables*				*Log Likelihood*	
2				**B0**	12.48309171			-6.65456
3				**B1**	-0.23406877			
4				**B2**	-2.93832567			
5								
6								
7	X1	X2	Y	Logit	Probability	P(Y=y|X)		
8	39	4	0	-8.3988928	0.000225066	-0.00022509		
9	36.5	4	0	-7.8137209	0.000403988	-0.00040407		
10	36.5	2.5	0	-3.4062324	0.032101254	-0.0326278		
11	35.5	3.5	0	-6.1104893	0.002214549	-0.00221701		
12	34	2.5	0	-2.8210605	0.056196661	-0.05783746		
13	29.5	2	0	-0.2985882	0.425902646	-0.55495629		
14	28.5	3.5	0	-4.4720079	0.011295312	-0.01135959		
15	24.5	2.5	0	-0.5974072	0.354937108	-0.43840746		
16	17.5	2	0	2.510237	0.924856362	-2.58835382		
17	13.5	3.5	0	-0.9609765	0.276682734	-0.32390733		
18	29.5	1.5	1	1.1705746	0.763248868	-0.27017113		
19	28.5	2	1	-0.0645194	0.483875735	-0.72592715		
20	22	2.5	1	-0.0122353	0.496941214	-0.69928354		
21	19	2.5	1	0.689971	0.665960476	-0.40652496		

Using these values of Betas you can calculate the probability and hence response variable by deciding the probability cut-off.

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

82) 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.

83) 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

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

85) 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 suprises 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%.

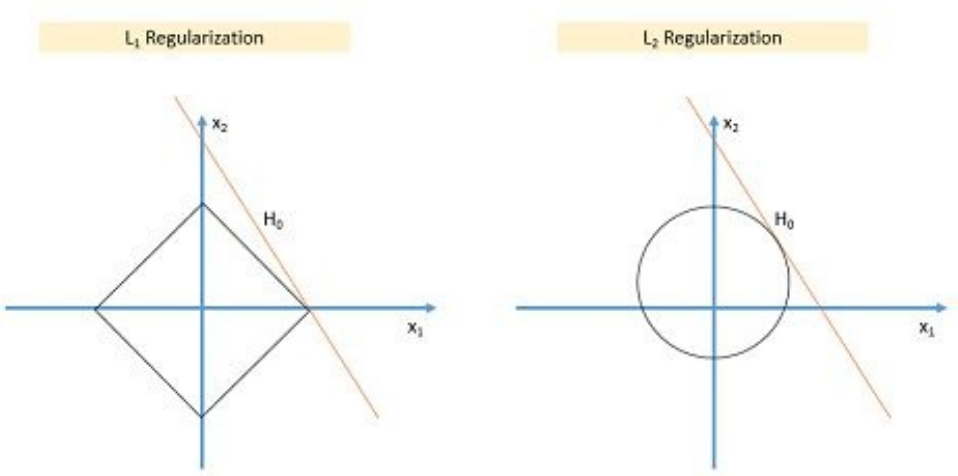
Precision is the ratio of number of events you can correctly recall to a number of all events you recall (combination of wrong and correct recalls).

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

87) 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 L2 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.

88) How can you deal with different types of seasonality in time series modelling?

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)

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

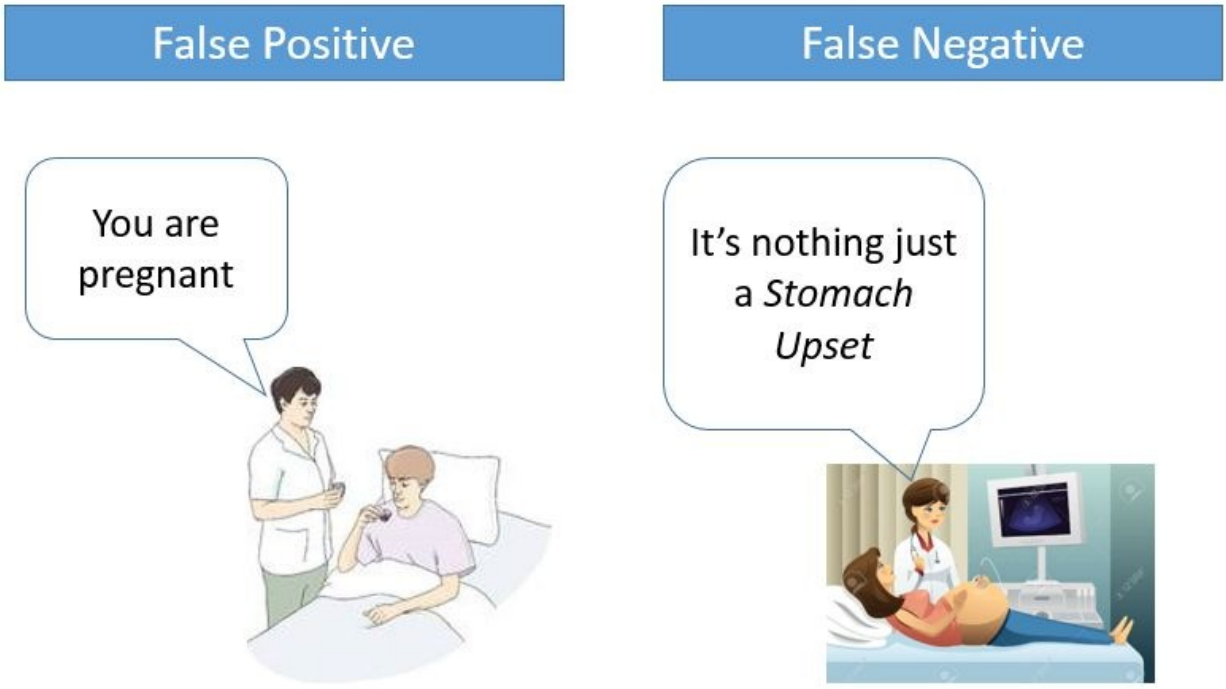
90) What do you understand by conjugate-prior with respect to Naïve Bayes?

91) 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.



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?

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

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?

93) 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 be**come 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.

94) 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 leaning 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.

95) What makes a dataset gold standard?

96) 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 senstivity is pretty straight forward-

Senstivity = True Positives /Positives in Actual Dependent Variable

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

97) What is the importance of having a selection bias?

Selection Bias occurs when there is no appropriate randomization acheived 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.

98) 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.

99) What do you understand by feature vectors?

100) 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.

Question 101. Explain About Data Import In R Language?

Answer:

R Commander is used to import data in R language. To start the R commander GUI, the user must type in the command Rcmdr into the console. There are 3 different ways in which data can be imported in R language-

Users can select the data set in the dialog box or enter the name of the data set (if they know).

Data can also be entered directly using the editor of R Commander via Data->New Data Set. However, this works well when the data set is not too large.

Data can also be imported from a URL or from a plain text file (ASCII), from any other statistical package or from the clipboard.

Question 102. How Missing Values And Impossible Values Are Represented In R Language?

Answer:

NaN (Not a Number) is used to represent impossible values whereas NA (Not Available) is used to represent missing values. The best way to answer this question would be to mention that deleting missing values is not a good idea because the probable cause for missing value could be some problem with data collection or programming or the query. It is good to find the root cause of the missing values and then take necessary steps handle them.

Question 103. R Language Has Several Packages For Solving A Particular Problem. How Do You Make A Decision On Which One Is The Best To Use?

Answer:

CRAN package ecosystem has more than 6000 packages. The best way for beginners to answer this question is to mention that they would look for a package that follows good software development principles. The next thing would be to look for user reviews and find out if other data scientists or analysts have been able to solve a similar problem.

Question 104. Which Function In R Language Is Used To Find Out Whether The Means Of 2 Groups Are Equal To Each Other Or Not?

Answer:

t.tests ()

Question 105. What Is The Best Way To Communicate The Results Of Data Analysis Using R Language?

Answer:

The best possible way to do this is combine the data, code and analysis results in a single document using knitr for reproducible research. This helps others to verify the findings, add to them and engage in discussions. Reproducible research makes it easy to redo the experiments by inserting new data and applying it to a different problem.

Question 106. How Many Data Structures Does R Language Have?

Answer:

R language has Homogeneous and Heterogeneous data structures.

Homogeneous data structures have same type of objects – Vector, Matrix ad Array.

Heterogeneous data structures have different type of objects – Data

frames and lists.

Question 107. What Is The Process To Create A Table In R Language Without Using External Files?

Answer:

MyTable= data.frame ()

edit (MyTable)

The above code will open an Excel Spreadsheet for entering data into MyTable.

Learn Data Science in R Programming to land a top gig as an Enterprise Data Scientist!

Question 108. Explain About The Significance Of Transpose In R Language?

Answer:

Transpose t () is the easiest method for reshaping the data before analysis.

Question 109. What Are With () And By () Functions Used For?

Answer:

With () function is used to apply an expression for a given dataset and BY () function is used for applying a function each level of factors.

Question 110. Dplyr Package Is Used To Speed Up Data Frame Management Code. Which Package Can Be Integrated With Dplyr For Large Fast Tables?

Answer:

data.table

Question 111. In Base Graphics System, Which Function Is Used To Add Elements To A Plot?

Answer:

boxplot () or text ()

Question 112. What Are The Different Type Of Sorting Algorithms Available In R Language?

Answer:

Bucket Sort
Selection Sort
Quick Sort
Bubble Sort
Merge Sort

Question 113. What Is The Command Used To Store R Objects In A File?

Answer:

save (x, file=”x.Rdata”)

Question 114. What Is The Best Way To Use Hadoop And R Together For Analysis?

Answer:

HDFS can be used for storing the data for long-term. MapReduce jobs submitted from either Oozie, Pig or Hive can be used to encode, improve and sample the data sets from HDFS into R. This helps to leverage complex analysis tasks on the subset of data prepared in R.

Question 115. What Will Be The Output Of Log (-5.8) When Executed On R Console?

Answer:

Executing the above on R console will display a warning sign that NaN (Not a Number) will be produced because it is not possible to take the log of negative number.

Question 116. How Is A Data Object Represented Internally In R Language?

Answer:

unclass (as.Date (“2016-10-05″))

Question 117. Which Package In R Supports The Exploratory Analysis Of Genomic Data?

Answer:

Adegenet.

Question 118. What Is The Difference Between Data Frame And A Matrix In R?

Answer:

Data frame can contain heterogeneous inputs while a matrix cannot. In matrix only similar data types can be stored whereas in a data frame there can be different data types like characters, integers or other data frames.

Question 119. How Can You Add Datasets In R?

Answer:

rbind () function can be used add datasets in R language provided the columns in the datasets should be same.

Question 120. What Are Factor Variable In R Language?

Answer:

Factor variables are categorical variables that hold either string or numeric values. Factor variables are used in various types of graphics and particularly for statistical modelling where the correct number of degrees of freedom is assigned to them.

Question 121. What Is The Memory Limit In R?

Answer:

8TB is the memory limit for 64-bit system memory and 3GB is the limit for 32-bit system memory.

Question 122. What Are The Data Types In R On Which Binary Operators Can Be Applied?

Answer:

Scalars, Matrices ad Vectors.

Question 123. How Do You Create Log Linear Models In R Language?

Answer:

Using the loglm () function

Question 124. What Will Be The Class Of The Resulting Vector If You Concatenate A Number And Na?

Answer:

number

Question 125. What Is Meant By K-nearest Neighbour?

Answer:

K-Nearest Neighbour is one of the simplest machine learning classification algorithms that is a subset of supervised learning based on lazy learning. In this algorithm the function is approximated locally and any computations are deferred until classification.

Question 126. What Will Be The Class Of The Resulting Vector If You Concatenate A Number And A Character?

Answer:

character

Question 127. If You Want To Know All The Values In C (1, 3, 5, 7, 10) That Are Not In C (1, 5, 10, 12, 14). Which In-built Function In R Can Be Used To Do This? Also, How This Can Be Achieved Without Using The In-built Function?

Answer:

Using in-built function - setdiff(c (1, 3, 5, 7, 10), c (1, 5, 10, 11, 13))

Without using in-built function - c (1, 3, 5, 7, 10) [! c (1, 3, 5, 7, 10) %in% c (1, 5, 10, 11, 13).

Question 128. How Can You Debug And Test R Programming Code?

Answer:

R code can be tested using Hadley's testthat package.

Question 129. What Will Be The Class Of The Resulting Vector If You

Concatenate A Number And A Logical?

Answer:

Number.

Question 130. Write A Function In R Language To Replace The Missing Value In A Vector With The Mean Of That Vector?

Answer:

mean impute <- function(x) {x [is.na(x)] <- mean(x, na.rm = TRUE); x}

Question 131. What Happens If The Application Object Is Not Able To Handle An Event?

Answer:

The event is dispatched to the delegate for processing.

Question 132. Differentiate Between Lapply And Sapply?

Answer:

If the programmers want the output to be a data frame or a vector, then sapply function is used whereas if a programmer wants the output to be a list then lapply is used. There one more function known as vapply which is preferred over sapply as vapply allows the programmer to specific the output type. The disadvantage of using vapply is that it is difficult to be implemented and more verbose.

Question 133. Differentiate Between Seq (6) And Seq_along (6)?

Answer:

Seq_along(6) will produce a vector with length 6 whereas seq(6) will produce a sequential vector from 1 to 6 c((1,2,3,4,5,6)).

Question 134. How Will You Read A .csv File In R Language?

Answer:

read.csv () function is used to read a .csv file in R language.

Below is a simple example –

filcontent

print (filecontent)

Question 135. How Do You Write R Commands?

Answer:

The line of code in R language should begin with a hash symbol (#).

Question 136. How Can You Verify If A Given Object "x" Is A Matric Data Object?

Answer:

If the function call is.matrix(X) returns TRUE then X can be termed as a matrix data object.

Question 137. What Do You Understand By Element Recycling In R?

Answer:

If two vectors with different lengths perform an operation –the elements of the shorter vector will be re-used to complete the operation. This is referred to as element recycling.

Example – Vector A <-c(1,2,0,4) and Vector B<-(3,6) then the result of A*B will be (3,12,0,24). Here 3 and 6 of vector B are repeated when computing the result.

Question 138. How Can You Verify If A Given Object "x" Is A Matrix Data Object?

Answer:

If the function call is.matrix(X) returns true then X can be considered as a matrix data object otheriwse not.

Question 139. How Will You Measure The Probability Of A Binary Response Variable In R Language?

Answer:

Logistic regression can be used for this and the function glm () in R language provides this functionality.

Question 140. What Is The Use Of Sample And Subset Functions In R Programming Language?

Answer:

Sample () function can be used to select a random sample of size 'n' from a huge dataset.

Subset () function is used to select variables and observations from a given dataset.

Question 141. How Can You Resample Statistical Tests In R Language?

Answer:

Coin package in R provides various options for re-randomization and permutations based on statistical tests. When test assumptions cannot be met then this package serves as the best alternative to classical methods as it does not assume random sampling from well-defined populations.

Question 142. What Is The Purpose Of Using Next Statement In R Language?

Answer:

If a developer wants to skip the current iteration of a loop in the code without terminating it then they can use the next statement. Whenever the R parser comes across the next statement in the code, it skips evaluation of the loop further and jumps to the next iteration of the loop.

Question 143. How Will You Create Scatter Plot Matrices In R Language?

Answer:

A matrix of scatter plots can be produced using pairs. Pairs function takes various parameters like formula, data, subset, labels, etc.

The two key parameters required to build a scatter plot matrix are –

formula- A formula basically like ~a+b+c . Each term gives a separate variable in the pairs plots where the terms should be numerical vectors. It basically represents the series of variables used in pairs.

data- It basically represents the dataset from which the variables have to be taken for building a scatterplot.

Question 144. How Will You Check If An Element 25 Is Present In A Vector?

Answer:

There are various ways to do this-

It can be done using the match () function- match () function returns the first appearance of a particular element.

The other is to use %in% which returns a Boolean value either true or false.

Is.element () function also returns a Boolean value either true or false based on whether it is present in a vector or not.

Question 145. What Is The Difference Between Library() And Require() Functions In R Language?

Answer:

There is no real difference between the two if the packages are not being loaded inside the function. require () function is usually used inside function and throws a warning whenever a particular package is not found. On the flip side, library () function gives an error message if the desired package cannot be loaded.

Question 146. What Are The Rules To Define A Variable Name In R Programming Language?

Answer:

A variable name in R programming language can contain numeric and alphabets along with special characters like dot (.) and underline (-). Variable names in R language can begin with an alphabet or the dot symbol. However, if the variable name begins with a dot symbol it should not be a followed by a numeric digit.

Question 147. What Do You Understand By A Workspace In R Programming Language?

Answer:

The current R working environment of a user that has user defined objects like lists, vectors, etc. is referred to as Workspace in R language.

Question 148. Which Function Helps You Perform Sorting In R Language?

Answer:

Order ()

Question 149. How Will You List All The Data Sets Available In All R Packages?

Answer:

Using the below line of code-

data(package = .packages(all.available = TRUE))

Question 150. Which Function Is Used To Create A Histogram Visualisation In R Programming Language?

Answer:

Hist()

Question 151. Write The Syntax To Set The Path For Current Working Directory In R Environment?

Answer:

Setwd("dir_path")

Question 152. What Will Be The Output Of Runif (7)?

Answer:

It will generate 7 random numbers between 0 and 1.

Question 153. What Is The Difference Between Rnorm And Runif Functions?

Answer:

rnorm function generates "n" normal random numbers based on the mean and standard deviation arguments passed to the function.

Syntax of rnorm function -

rnorm(n, mean = , sd =)

runif function generates "n" unform random numbers in the interval of minimum and maximum values passed to the function.

Syntax of runif function -

runif(n, min = , max =)

Question 157. What Will Be The Output On Executing The Following R Programming Code ?

Answer:

```
mat<-matrix(rep(c(TRUE,FALSE),8),nrow=4)
sum(mat)
 8
```

Question 158. How Will You Combine Multiple Different String Like "data", "science", "in" ,"r", "programming" As A Single String "data_science_in_r_programmming" ?

Answer:

```
paste("Data", "Science", "in" ,"R", "Programming",sep="_")
```

Question 160. Write A Function To Extract The First Name From The String "mr. Tom White"?

Answer:

```
substr ("Mr. Tom White",start=5, stop=7)
```

Question 161. Can You Tell If The Equation Given Below Is Linear Or Not ?

Answer:

Emp_sal= 2000+2.5(emp_age)2

Yes it is a linear equation as the coefficients are linear.

Question 162. What Is R Base Package?

Answer:

R Base package is the package that is loaded by default whenever R programming environent is loaded .R base package provides basic fucntionalites in R environment like arithmetic calcualtions, input/output.

Question 164. How Will You Merge Two Dataframes In R Programming Language?

Answer:

Merge () function is used to combine two dataframes and it identifies common rows or columns between the 2 dataframes. Merge () function basically finds the intersection between two different sets of data.

Merge () function in R language takes a long list of arguments as follows –

Syntax for using Merge function in R language -

merge (x, y, by.x, by.y, all.x or all.y or all)

X represents the first dataframe.

Y represents the second dataframe.

by.X- Variable name in dataframe X that is common in Y.

by.Y- Variable name in dataframe Y that is common in X.

all.x - It is a logical value that specifies the type of merge. all.X should be set

to true, if we want all the observations from dataframe X . This results in Left Join.

all.y - It is a logical value that specifies the type of merge. all.y should be set to true , if we want all the observations from dataframe Y . This results in Right Join.

all – The default value for this is set to FALSE which means that only matching rows are returned resulting in Inner join. This should be set to true if you want all the observations from dataframe X and Y resulting in Outer join.

Question 167. What Will Be The Result Of Multiplying Two Vectors In R Having Different Lengths?

Answer:

The multiplication of the two vectors will be performed and the output will be displayed with a warning message like – "Longer object length is not a multiple of shorter object length." Suppose there is a vector a<-c (1, 2, 3) and vector b <- (2, 3) then the multiplication of the vectors a*b will give the resultant as 2 6 6 with the warning message. The multiplication is performed in a sequential manner but since the length is not same, the first element of the smaller vector b will be multiplied with the last element of the larger vector a.

Question 168. R Programming Language Has Several Packages For Data Science Which Are Meant To Solve A Specific Problem, How Do You Decide Which One To Use?

Answer:

CRAN package repository in R has more than 6000 packages, so a data scientist needs to follow a well-defined process and criteria to select the right one for a specific task. When looking for a package in the CRAN repository a data scientist should list out all the requirements and issues so that an ideal R package can address all those needs and issues.

The best way to answer this question is to look for an R package that follows good software development principles and practices. For example, you might want to look at the quality documentation and unit tests. The next step is to check out how a particular R package is used and read the reviews posted by other users of the R package. It is important to know if other data scientists or data analysts have been able to solve a similar problem as that of yours. When you in doubt choosing a particular R package, I would always ask for feedback from R community members or other colleagues to ensure that I am making the right choice.

Question 189. How Can You Merge Two Data Frames In R Language?

Answer:

Data frames in R language can be merged manually using cbind () functions or by using the merge () function on common rows or columns.

Question 170. Explain The Usage Of Which() Function In R Language?

Answer:

which() function determines the position of elements in a logical vector that are TRUE. In the below example, we are finding the row number wherein the maximum value of variable v1 is recorded.

```
mydata=data.frame(v1 = c(2,4,12,3,6))
which(mydata$v1==max(mydata$v1))
```

It returns 3 as 12 is the maximum value and it is at 3rd row in the variable x=v1.

Question 170. How Will You Convert A Factor Variable To Numeric In R Language ?

Answer:

A factor variable can be converted to numeric using the as.numeric() function in R language. However, the variable first needs to be converted to character before being converted to numberic because the as.numeric() function in R does not return original values but returns the vector of the levels of the factor variable.

```
X <- factor(c(4, 5, 6, 6, 4))
X1 = as.numeric(as.character(X))
```

171. What is the difference between supervised and unsupervised machine learning?

Supervised Machine learning:

Supervised machine learning requires training labeled data.

Unsupervised Machine learning:

Unsupervised machine learning doesn’t required labeled data.

172. 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 underfitting. When you train your model at

that time model makes simplified assumptions to make the target function easier to understand.

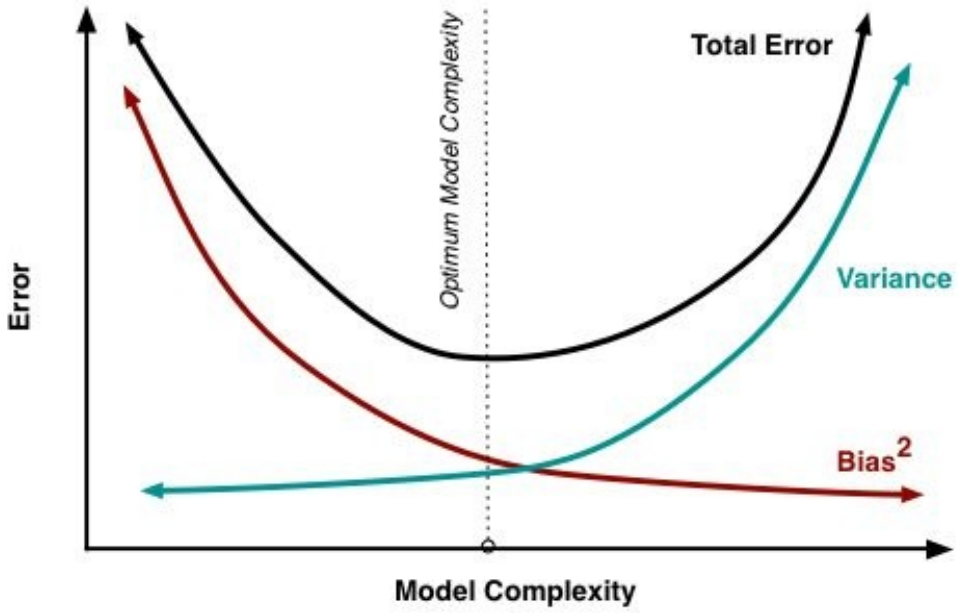
Low bias machine learning algorithms - Decision Trees, k-NN and SVM

Hight bias machine learning algorithms - Liear 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 dataset and performs bad on test dataset.” It can lead high sensitivity and overfitting.

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 achive good prediction performance.

1. The k-nearest neighbors 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 neighbors 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.

173. What is exploding gradients ?

“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.

Gradient:

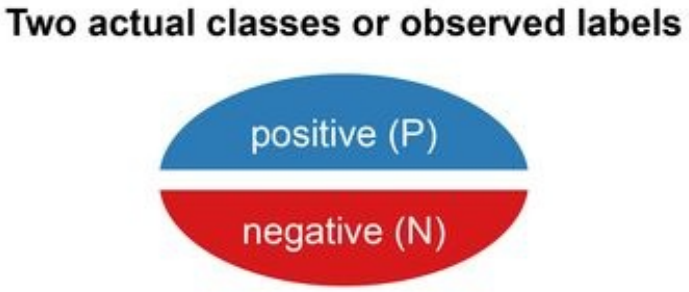
Gradient is the **direction and magnitude** calculated during training of a neural network that is used to update the network weights in the right direction and by the right amount.

174. What is a confusion matrix ?

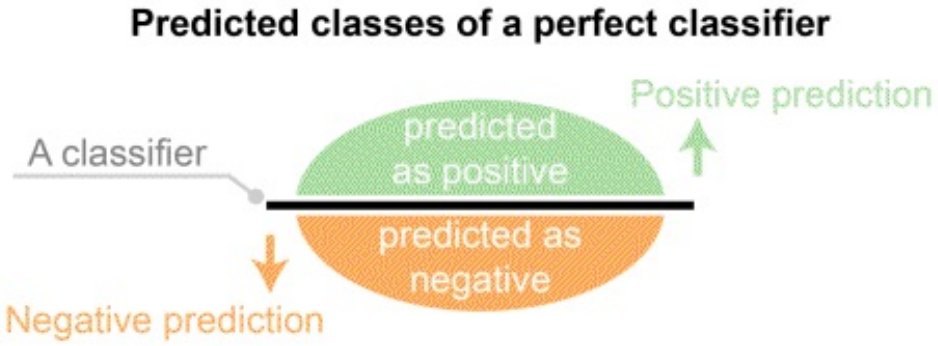
The confusion matrix is a 2X2 table that contains 4 outputs provided by the **binary classifier**. Various measures, such as error-rate, accuracy, specificity, sensitivity, precision and recall are derived from it. *Confusion Matrix*

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

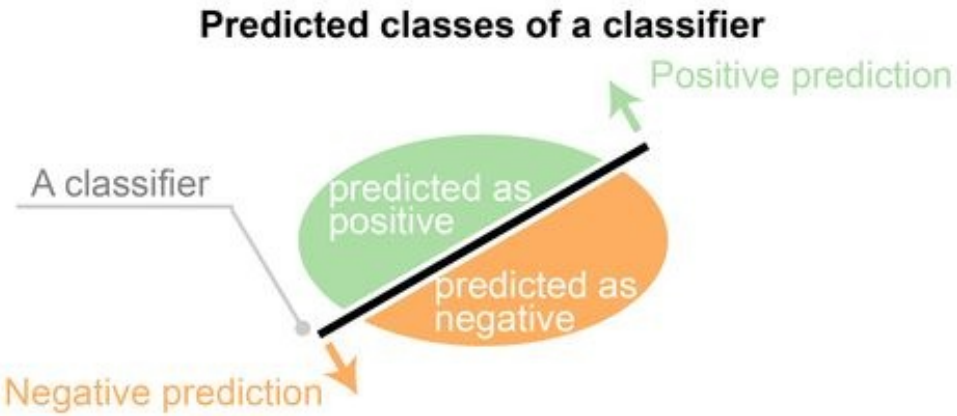
A dataset used for performance evaluation is called test dataset. It should contain the correct labels and predicted labels.



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

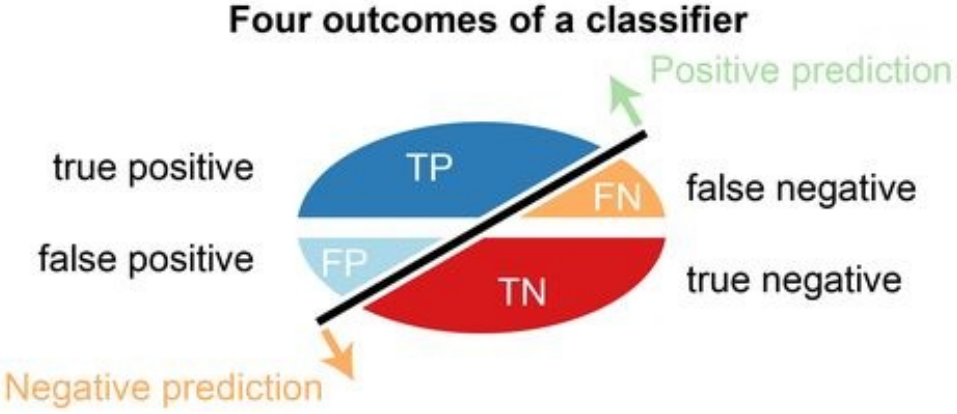


The predicted labels usually match with part of the observed labels in real world scenarios.



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

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



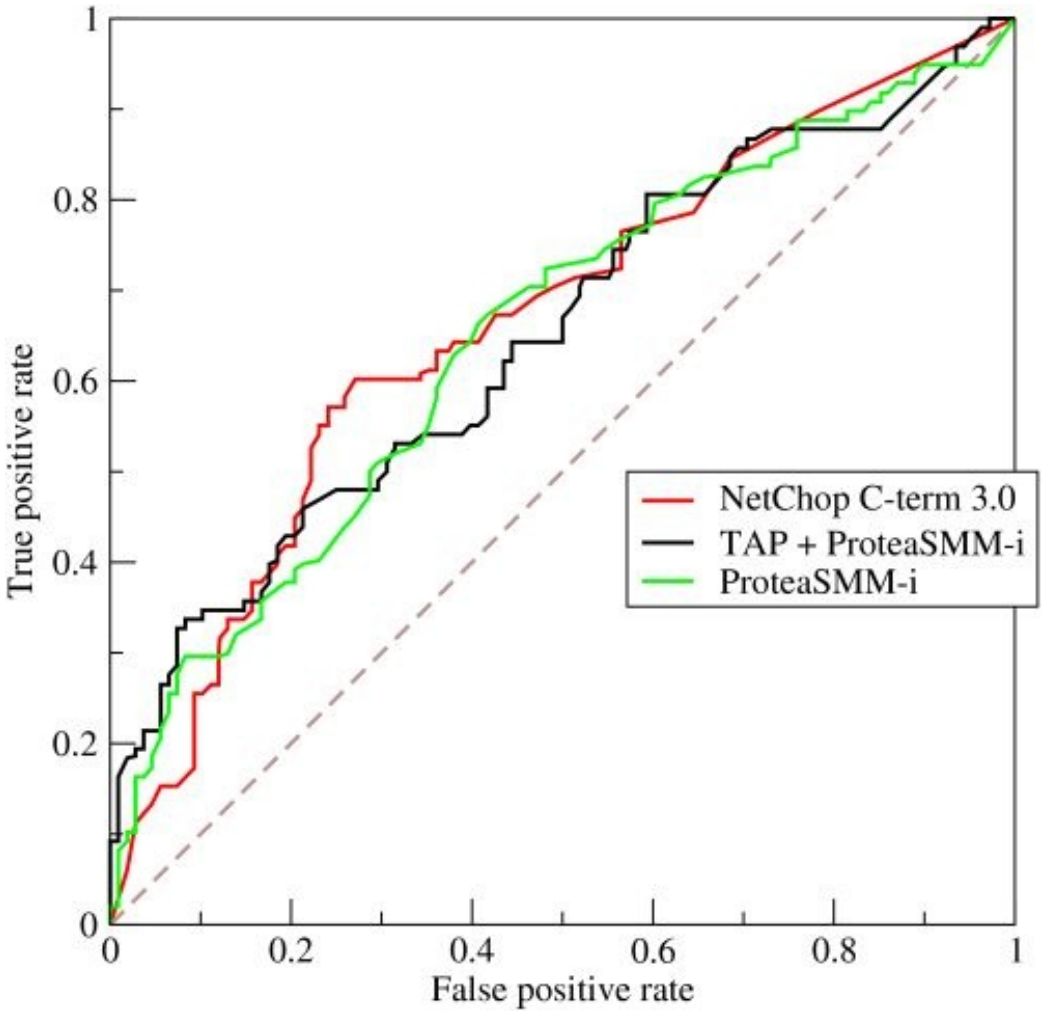
Basic measures derived from the confusion matrix

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

176. 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.

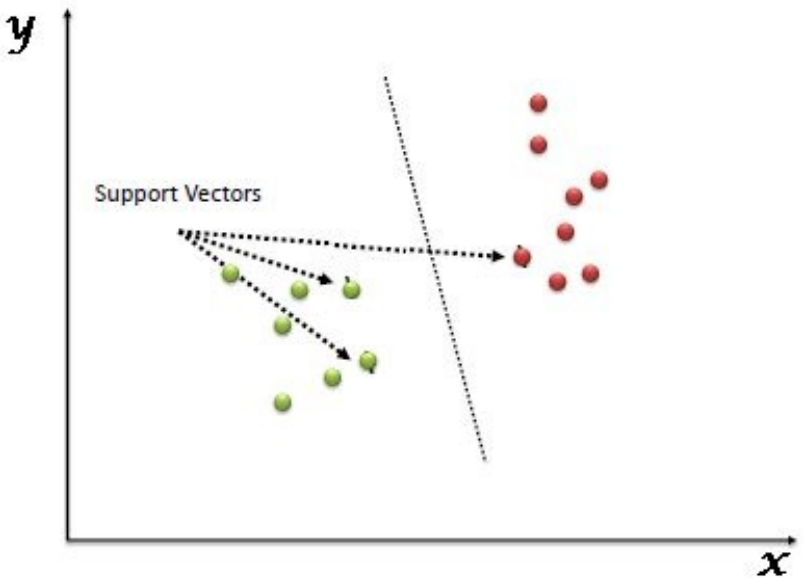


177. What is selection Bias ?

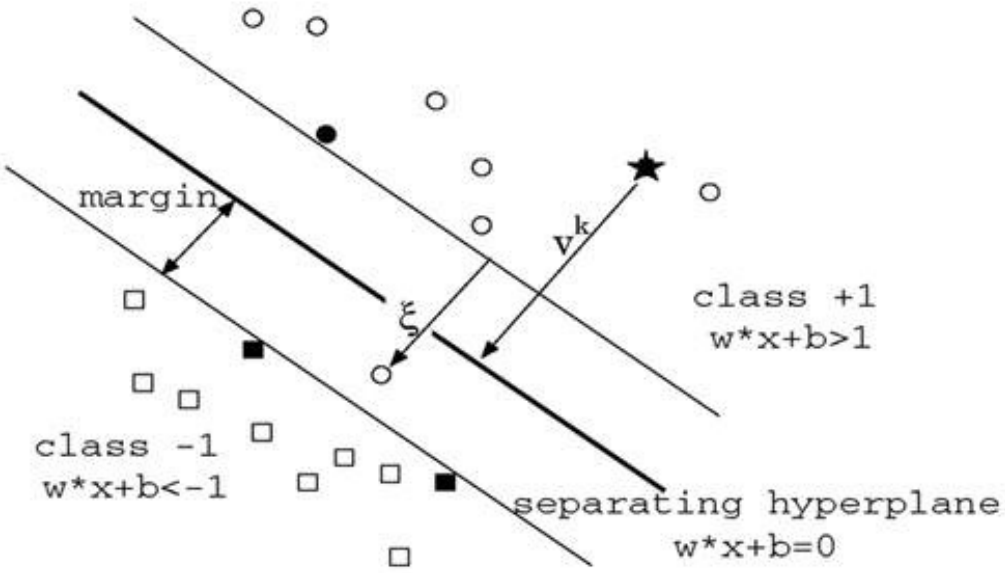
Selection bias occurs when sample obtained is not represantative of the population intended to be analyzed.

178. 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 dataset, SVM tries to plot it in n-dimentional space with the value of each feature being the value of a particular coordinate. SVM uses hyper planes to seperate out different classes based on the provided kernel function.



179. 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.

180. 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

181. Explain Decision Tree algorithm in detail.

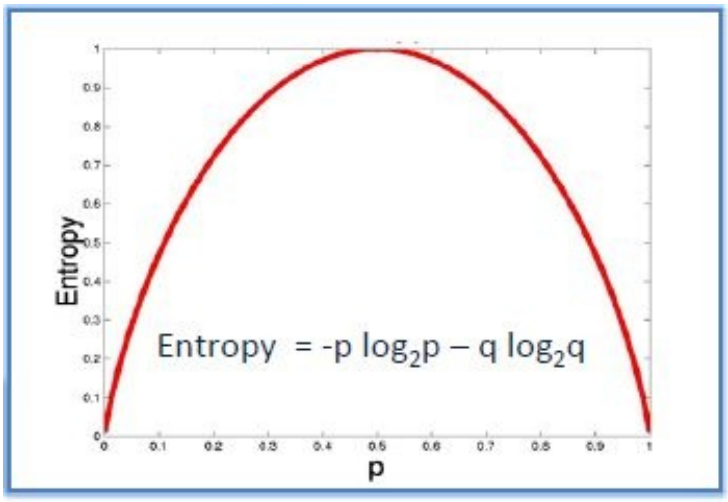
Decision tree is a supervised machine learning algorithm mainly used for the **Regression and Classification**.It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf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

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

The core algorithm for building decision tree is called **ID3**. **ID3** uses **Enteropy** 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 homogenious subsets. **ID3** uses enteropy to check the homogeneity of a sample. If the sample is completely homogenious then entropy is zero and if the sample is an equally divided it has entropy of one.



Entropy = $-0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$

Information Gain

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

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

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

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

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

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

G(PlayGolf, Outlook) = **E**(PlayGolf) – **E**(PlayGolf, Outlook)

= 0.940 – 0.693 = 0.247

183. What is pruning in Decision Tree ?

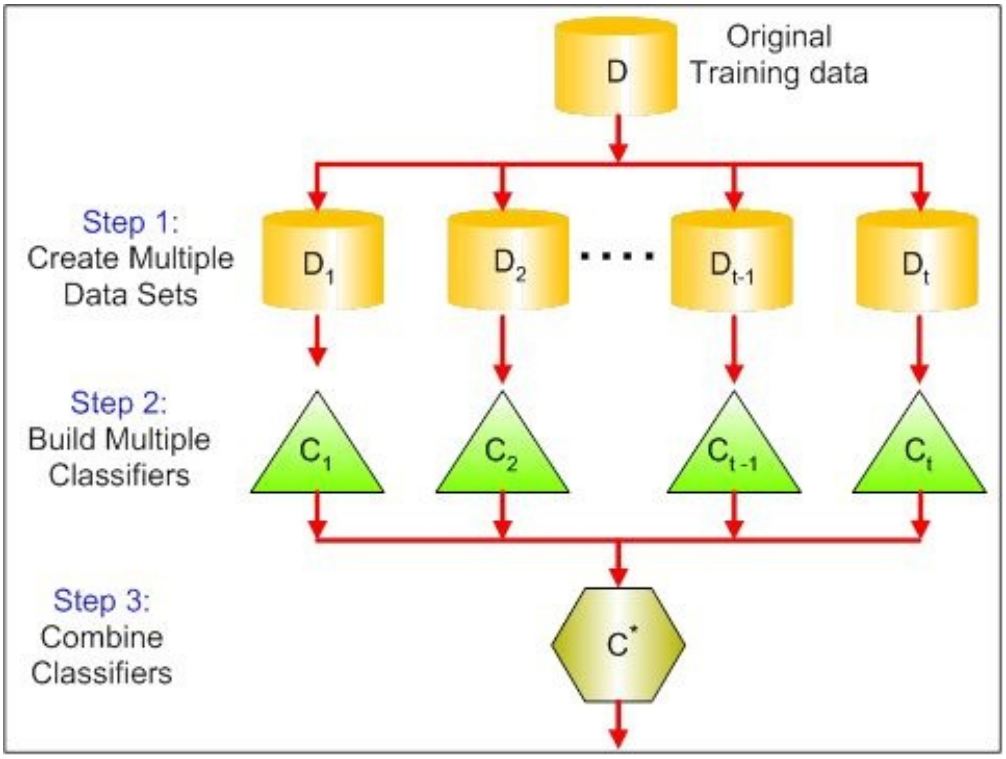
When we remove sub-nodes of a decision node, this procsss is called pruning or opposite process of splitting.

184. What is Ensemble Learning ?

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

Bagging

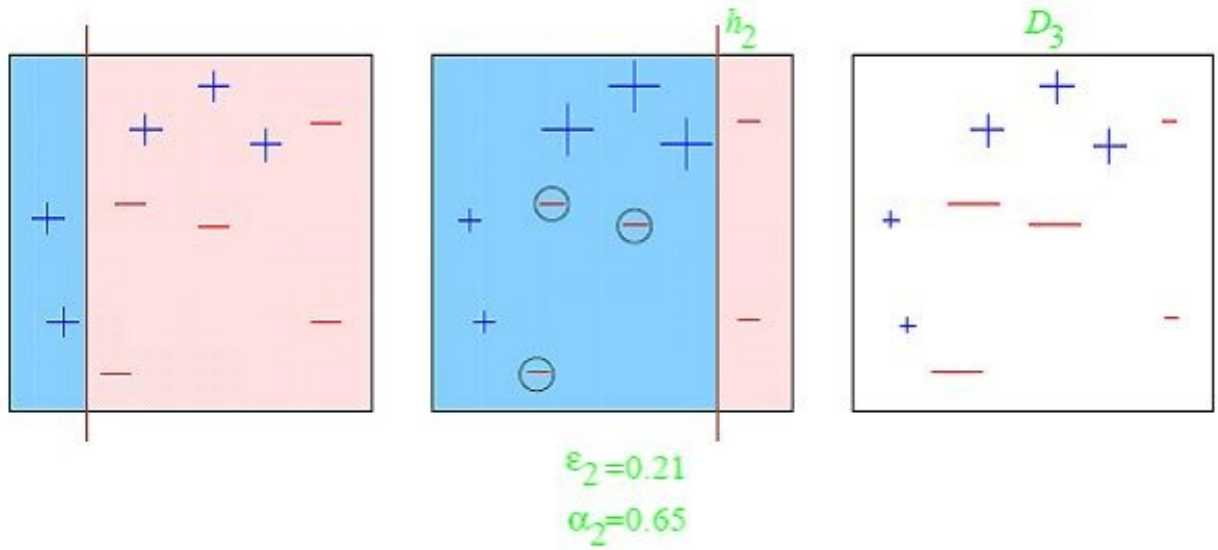
Bagging tries to implement similar learners on small sample populations and then takes a mean of all the predictions. In generalized bagging, you can use different learners on different population. As you expect this helps us to reduce the variance error.



Boosting

Boosting is an iterative technique which adjust the weight of an observation based on the last classification. If an observation was classfied 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 overfit on the training data.



185. 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 dimentionality 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.

186. What cross-validation technique would you use on a time series dataset.

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]

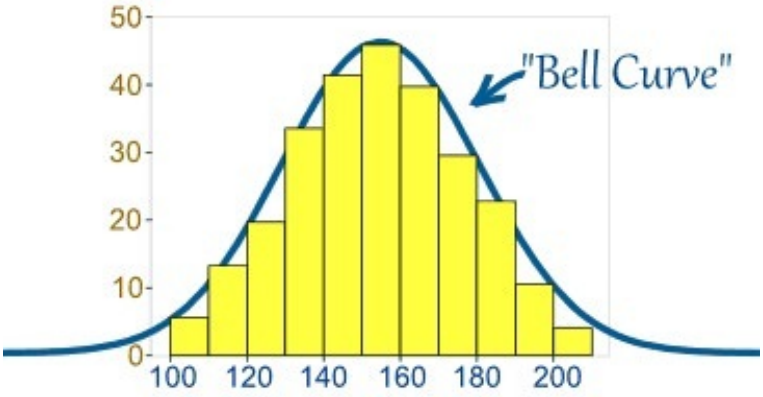
fold 1: training[1 2 3], test[4]

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

187. 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.

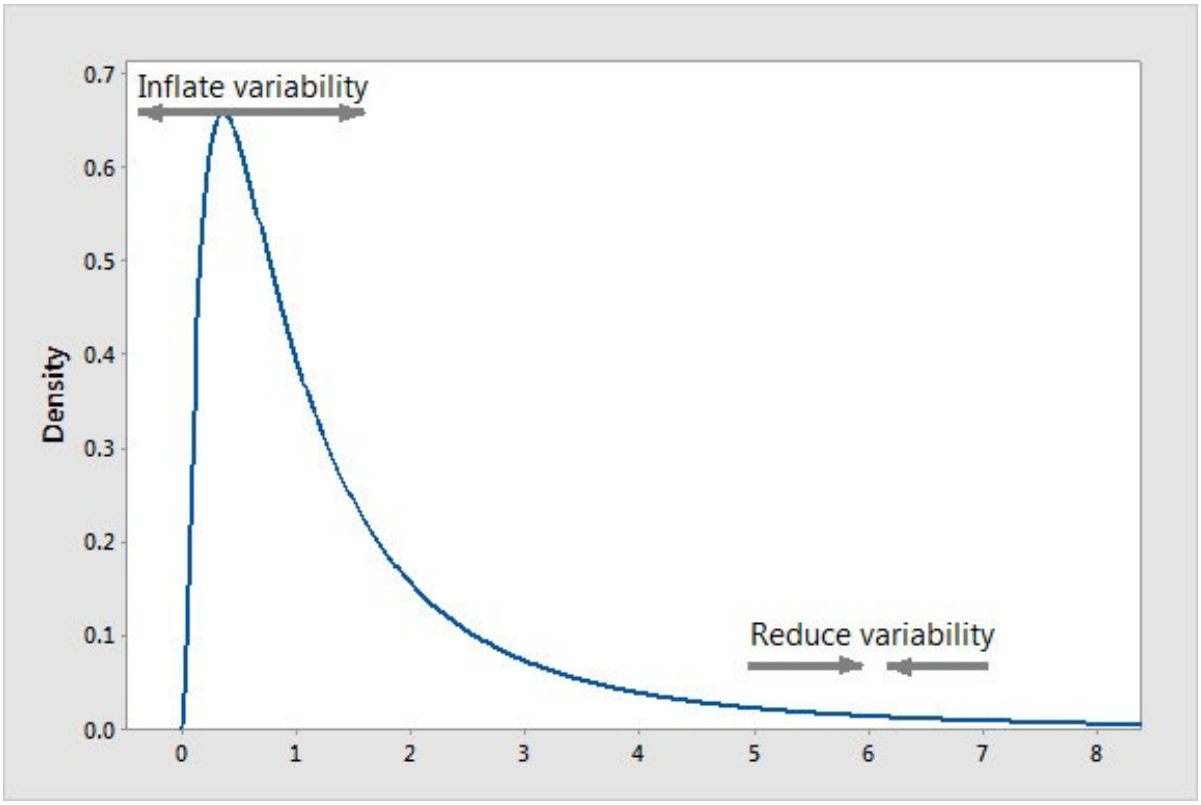
188. What do you understand by the term Normal Distribution?



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.

189. What is a Box Cox Transformation?

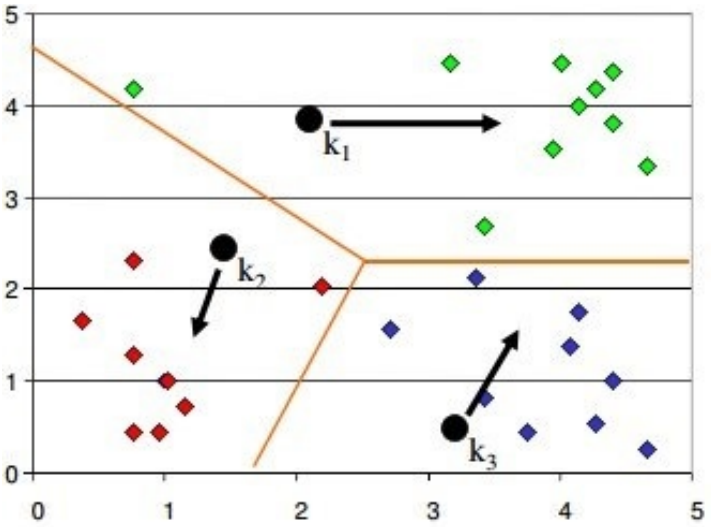
Dependent variable for a regression analysis might not satisfy one or more assumptions of an ordinary least squares regression. The residuals could either curve as the prediction increases or follow skewed distribution. In such scenarios, it is necessary to transform the response variable so that the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



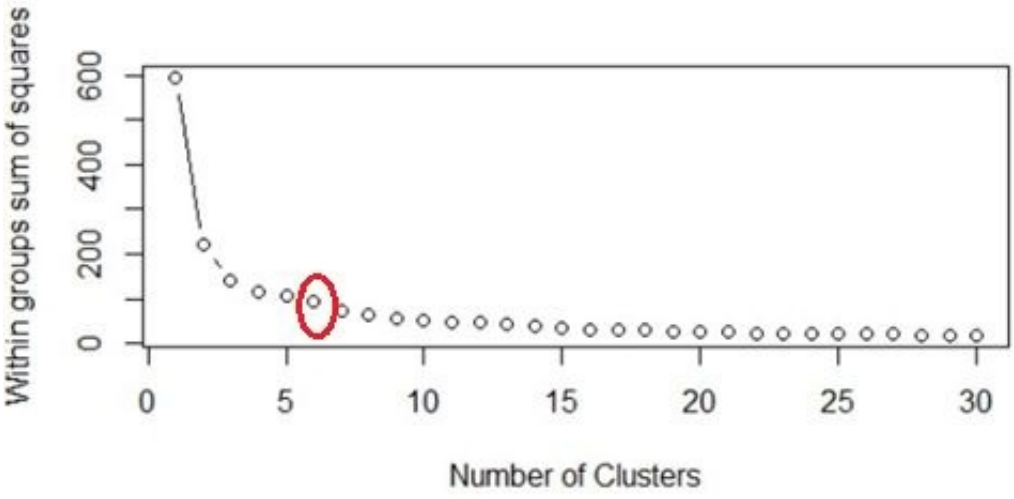
A Box Cox transformation is a way to transform non-normal dependent variables into a normal shape. Normality is an important assumption for many statistical techniques, if your data isn't normal, applying a Box-Cox means that you are able to run a broader number of tests. The Box Cox transformation is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

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

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



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



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

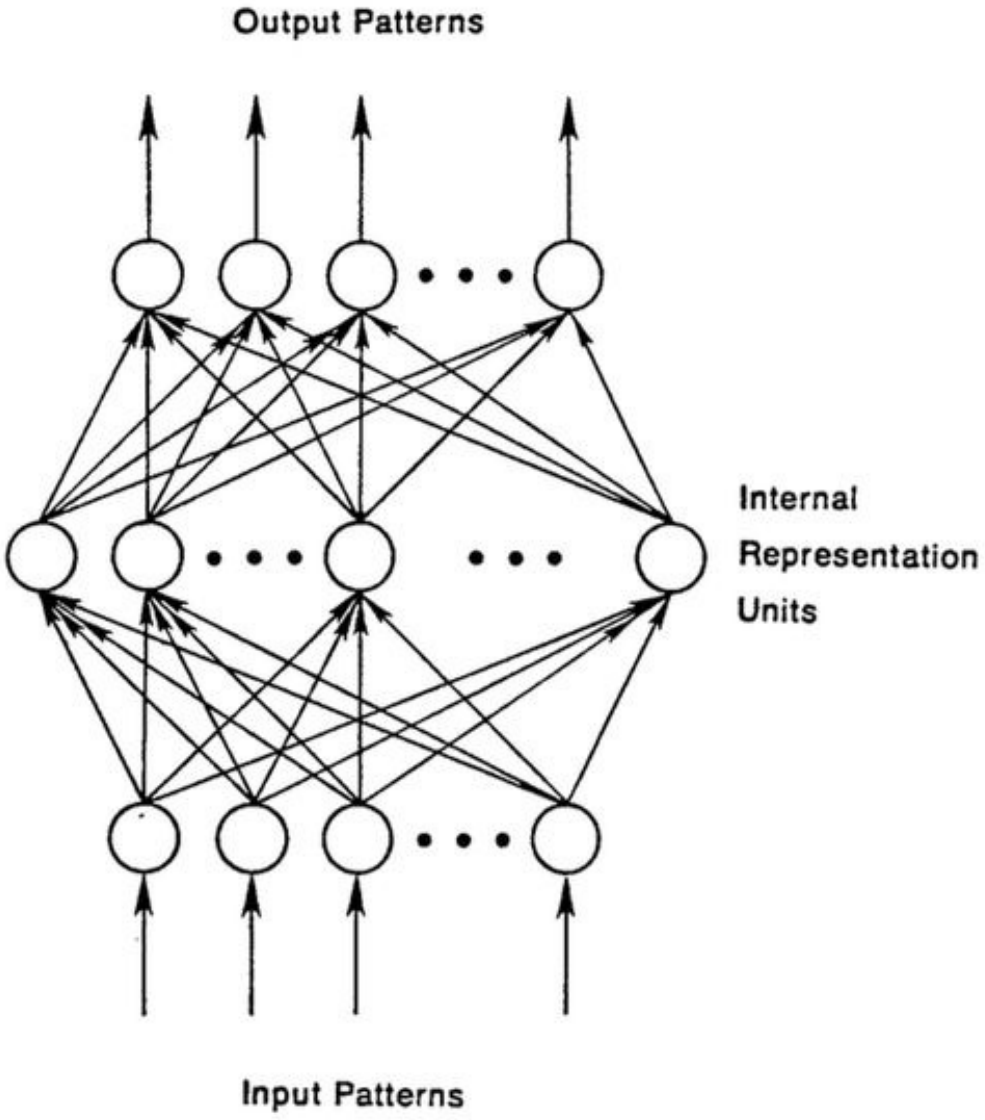
191. What is deep learning?

Deep learning is subfield 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 extention 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 latyers to better understand the input output relationship.

192. What are Recurrent Neural Networks(RNNs) ?

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

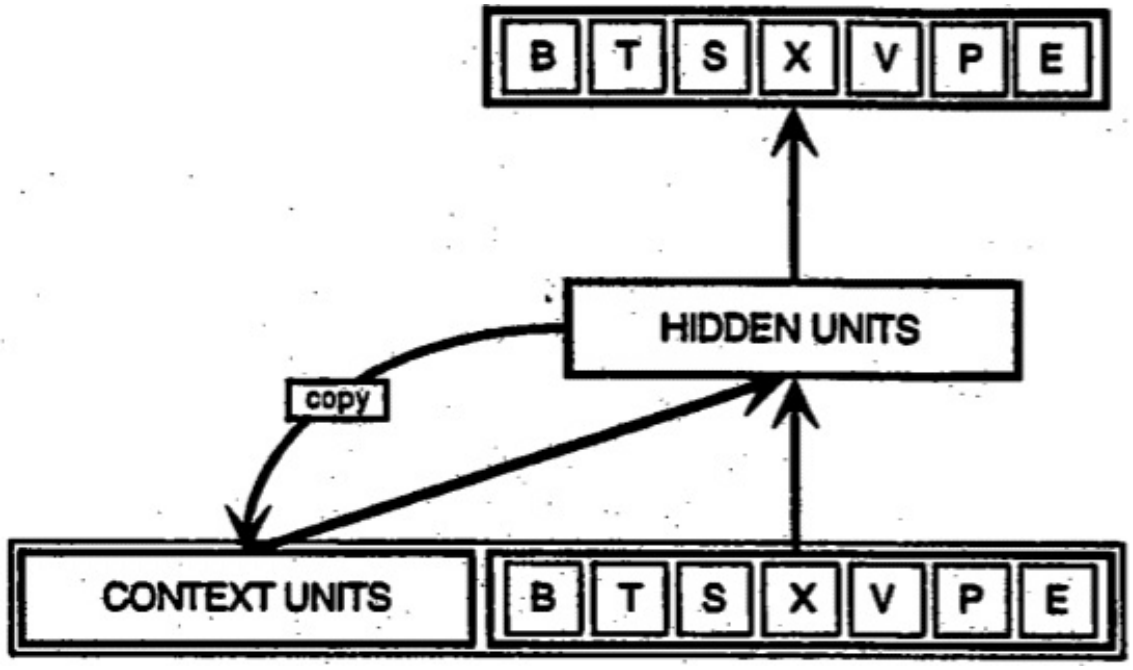


Recurrent networks on the other hand, take as their input not just the current input example they see, but also the what they have percieved previously in time. The BTSXPE at the bottom of the drawing represents the input example in the current moment, and CONTEXT UNIT represents the output of the previous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

Backpropagation in feedforward 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 backpropagation called backpropagation 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 backpropagation needs to work.



193. 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 categorized in following three categories.

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

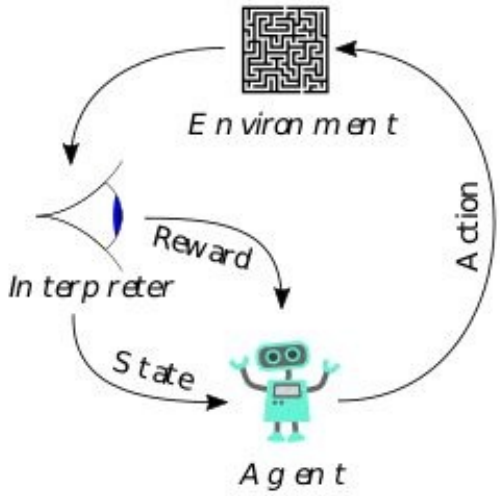
Deep learning:

Deep Learning is a subfield of machine learning concerned with algorithms

inspired by the structure and function of the brain called artificial neural networks.

194. 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 maximize 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/panelity mechanism.

195. What is selection bias ?

Selection Bias

Selection bias is the bias introduced by the selection of individuals, groups or data for analysis in such a way that proper randomization is not achieved, thereby ensuring that the sample obtained is not representative of the population intended to be analyzed. 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.

196. Explain what regularization is and why it is useful.

Regularization

Regularization is the process of adding tunning 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.

197. 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.

198. 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.

199. 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 labeled dataset, 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 discreate values then it will 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.

200. If you are having 4GB RAM in your machine and you want to train your model on 10GB dataset. 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 dataset, it doesn’t load complete dataset 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 dataset in small size datasets.
2. Use partialfit method of SVM, it requires subset of complete dataset.
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.

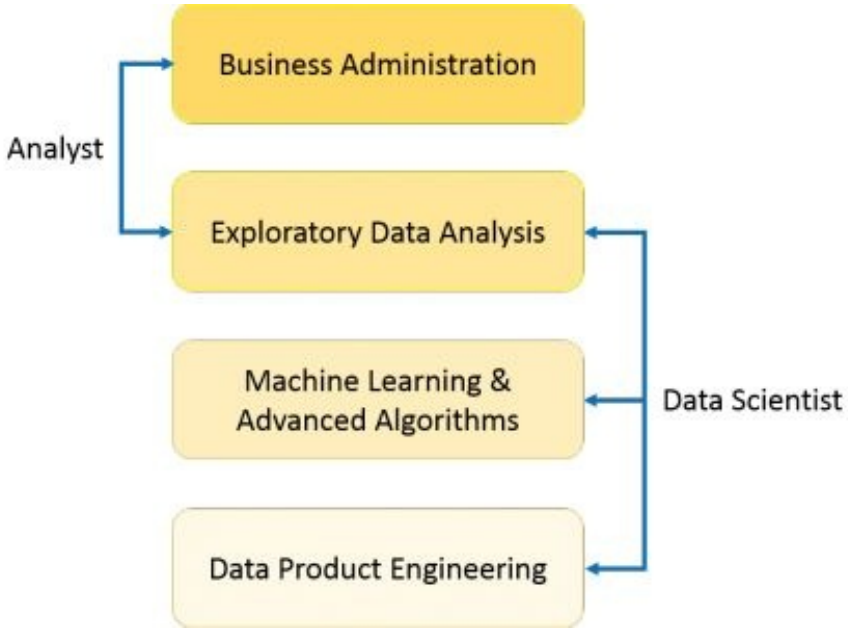
32. What is 'Naive' in a Naive Bayes ?

The Naive Bayes Algorithm is based on the Bayes Theoram. Bayes' theoram describes the probablitiy of an event, based on prior knowledge of conditions that might be related to the event.

201. What is Data Science? Also, list the differences between supervised and unsupervised learning.

Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data. How is this different from what statisticians have been doing for years?

The answer lies in the difference between explaining and predicting.



Supervised Learning vs Unsupervised Learning	
Supervised Learning	**Unsupervised Learning**
1. Input data is labeled.	1. Input data is unlabeled.
2. Uses training dataset.	2. Uses the input data set.
3. Used for prediction.	3. Used for analysis.
4. Enables classification and regression.	4. Enables Classification, Density Estimation, & Dimension Reduction

202. What are the important skills to have in Python with regard to data analysis?

The following are some of the important skills to possess which will come handy when performing data analysis using Python.

- Good understanding of the built-in data types especially lists, dictionaries, tuples, and sets.
- Mastery of N-dimensional NumPy Arrays.
- Mastery of Pandas dataframes.
- Ability to perform element-wise vector and matrix operations on NumPy arrays.
- Knowing that you should use the Anaconda distribution and the conda package manager.

- Familiarity with Scikit-learn. ****Scikit-Learn Cheat Sheet****
- Ability to write efficient list comprehensions instead of traditional for loops.
- Ability to write small, clean functions (important for any developer), preferably pure functions that don't alter objects.
- Knowing how to profile the performance of a Python script and how to optimize bottlenecks.

The following will help to tackle any problem in data analytics and machine learning.

203. What is Selection Bias?

Selection bias is a kind of error that occurs when the researcher decides who is going to be studied. It is usually associated with research where the selection of participants isn't random. It is sometimes referred to as the selection effect. It is the distortion of 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.

The types of selection bias include:

1. **Sampling bias**: It is a systematic error due to a non-random sample of a population causing some members of the population to be less likely to be included than others resulting in a biased sample.
2. **Time interval**: A trial may be terminated early at an extreme value (often for ethical reasons), but the extreme value is likely to be reached by the variable with the largest variance, even if all variables have a similar mean.
3. **Data**: When specific subsets of data are chosen to support a conclusion or rejection of bad data on arbitrary grounds, instead of according to previously stated or generally agreed criteria.
4. **Attrition**: Attrition bias is a kind of selection bias caused by attrition (loss of participants) discounting trial subjects/tests that did not run to completion.

Data Scientist Masters

204. What is the difference between "long" and "wide" format data?

In the **wide** format, a subject's repeated responses will be in a single row, and

each response is in a separate column. In the **long** format, each row is a one-time point per subject. You can recognize data in wide format by the fact that columns generally represent groups.

205. 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 a symmetrical bell-shaped curve.

Properties of Nornal Distribution:

1. Unimodal -one mode
2. Symmetrical -left and right halves are mirror images
3. Bell-shaped -maximum height (mode) at the mean
4. Mean, Mode, and Median are all located in the center
5. Asymptotic

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

It is a statistical hypothesis testing for a 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. A/B testing is a fantastic method for figuring out the best online promotional and marketing strategies for your business. It can be used to test everything from website copy to sales emails to search ads

An example of this could be identifying the click-through rate for a banner ad.

207. 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, Random Forest 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 seasonality is pretty straightforward.

Seasonality = (**True Positives**) / (**Positives in Actual Dependent Variable**)

*where true positives are positive events which are correctly classified as positives.

208. What are the differences between overfitting and underfitting?

In statistics and machine learning, one of the most common tasks is to fit a model to a set of training data, so as to be able to make reliable predictions on general untrained data.

In overfitting, a statistical model describes random error or noise instead of the underlying relationship. Overfitting occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. A model that has been overfit has poor predictive performance, as it overreacts to minor fluctuations in the training data.

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

We will prefer Python because of the following reasons:

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

Python vs R

210. How does data cleaning plays a vital role in analysis?

Data cleaning can help in analysis because:

- Cleaning data from multiple sources helps to transform it into a format that data analysts or data scientists can work with.
- Data Cleaning helps to increase the accuracy of the model in machine learning.
- It is a cumbersome process because as the number of data sources increases, the time taken to clean the data increases exponentially due to the number of sources and the volume of data generated by these sources.

- It might take up to 80% of the time for just cleaning data making it a critical part of analysis task.

211. Differentiate between univariate, bivariate and multivariate analysis.

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

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

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

212. What is Cluster Sampling?

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

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

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

213. What is Systematic Sampling?

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

214. What are Eigenvectors and Eigenvalues?

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

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

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

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

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

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

216. Can you cite some examples where a false negative important than a false positive?

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

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

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

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

In the **Banking** industry giving loans is the primary source of making money but

at the same time if your repayment rate is not good you will not make any profit, rather you will risk huge losses.

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

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

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

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

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

219. Explain cross-validation.

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

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

220. What is Machine Learning?

Machine Learning explores the study and construction of algorithms that can learn from and make predictions on data. Closely related to computational statistics. Used to devise complex models and algorithms that lend themselves to a prediction which in commercial use is known as predictive analytics.

Figure: Applications of Machine Learning

221. What is the Supervised Learning?

Supervised learning is the machine learning task of inferring a function from labeled training data. The training data consist of a set of training examples.

Algorithms: Support Vector Machines, Regression, Naive Bayes, Decision Trees, K-nearest Neighbor Algorithm and Neural Networks

E.g. If you built a fruit classifier, the labels will be “this is an orange, this is an apple and this is a banana”, based on showing the classifier examples of apples, oranges and bananas.

222. What is Unsupervised learning?

Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses.

Algorithms: Clustering, Anomaly Detection, Neural Networks and Latent Variable Models

E.g. In the same example, a fruit clustering will categorize as “fruits with soft skin and lots of dimples”, “fruits with shiny hard skin” and “elongated yellow fruits”.

223. What are the various classification algorithms?

The below diagram lists the most important classification algorithms.

224. What is logistic regression? 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.

225. What are Recommender Systems?

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

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

226. 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.

227. What is Collaborative filtering?

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

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

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

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

228. How can outlier values be treated?

Outlier values can be identified by using univariate or any other graphical analysis method. If the number of outlier values is few then they can be assessed

individually but for a large number of outliers, the values can be substituted with either the 99th or the 1st percentile values.

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

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

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

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

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

230. During analysis, how do you treat missing values?

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

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

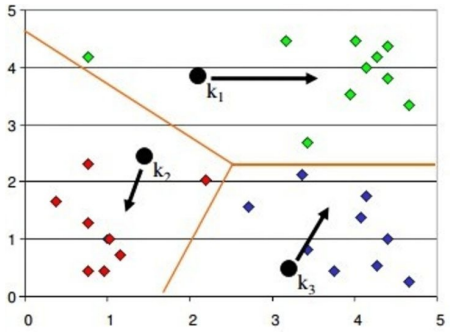
If it is a categorical variable, the default value is assigned. The missing value is assigned a default value. If you have a distribution of data coming, for normal distribution give the mean value.

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

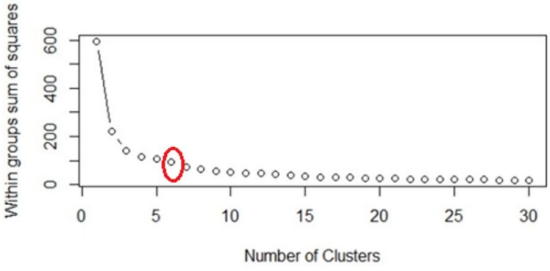
231. How will you define the number of clusters in a clustering algorithm?

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

For example, the following image shows three different groups.



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



- 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 the **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.

Now that we have seen the Machine Learning Questions, Let’s continue our Data Science Interview Questions blog with some Probability questions.

232. In any 15-minute interval, there is a 20% probability that you will see at least one shooting star. What is the probability that you see at least one shooting star in the period of an hour?

Probability of not seeing any shooting star in 15 minutes is

= 1 – P(Seeing one shooting star)
= 1 – 0.2 = 0.8

Probability of not seeing any shooting star in the period of one hour

= (0.8) ^ 4 = 0.4096

Probability of seeing at least one shooting star in the one hour

= 1 – P(Not seeing any star)
= 1 – 0.4096 = 0.5904

233. How can you generate a random number between 1 – 7 with only a die?

- Any die has six sides from 1-6. There is no way to get seven equal outcomes from a single rolling of a die. If we roll the die twice and consider the event of two rolls, we now have 36 different outcomes.
- To get our 7 equal outcomes we have to reduce this 36 to a number divisible by 7. We can thus consider only 35 outcomes and exclude the other one.
- A simple scenario can be to exclude the combination (6,6), i.e., to roll the die again if 6 appears twice.
- All the remaining combinations from (1,1) till (6,5) can be divided into 7 parts of 5 each. This way all the seven sets of outcomes are equally likely.

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

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

BB, BG, GB and GG;

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

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

Thus, P(Having two girls given one girl) = **1 / 3**

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

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

Probability of selecting fair coin = 999/1000 = **0.999**
Probability of selecting unfair coin = 1/1000 = **0.001**

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

an unfair coin

P (A) = 0.999 * (1/2)^5 = 0.999 * (1/1024) = **0.000976**
P (B) = 0.001 * 1 = **0.001**
P(A / A + B) = 0.000976 / (0.000976 + 0.001) = **0.4939**
P(B / A + B) = 0.001 / 0.001976 = **0.5061**

Probability of selecting another head = P(A/A+B) * 0.5 + P(B/A+B) * 1 = 0.4939 * 0.5 + 0.5061 = **0.7531**

236. What do you mean by Deep Learning and Why has it become popular now?

Deep Learning is nothing but a paradigm of machine learning which has shown incredible promise in recent years. This is because of the fact that Deep Learning shows a great analogy with the functioning of the human brain.

Now although Deep Learning has been around for many years, the major breakthroughs from these techniques came just in recent years. This is because of two main reasons:

- The increase in the amount of data generated through various sources
- The growth in hardware resources required to run these models

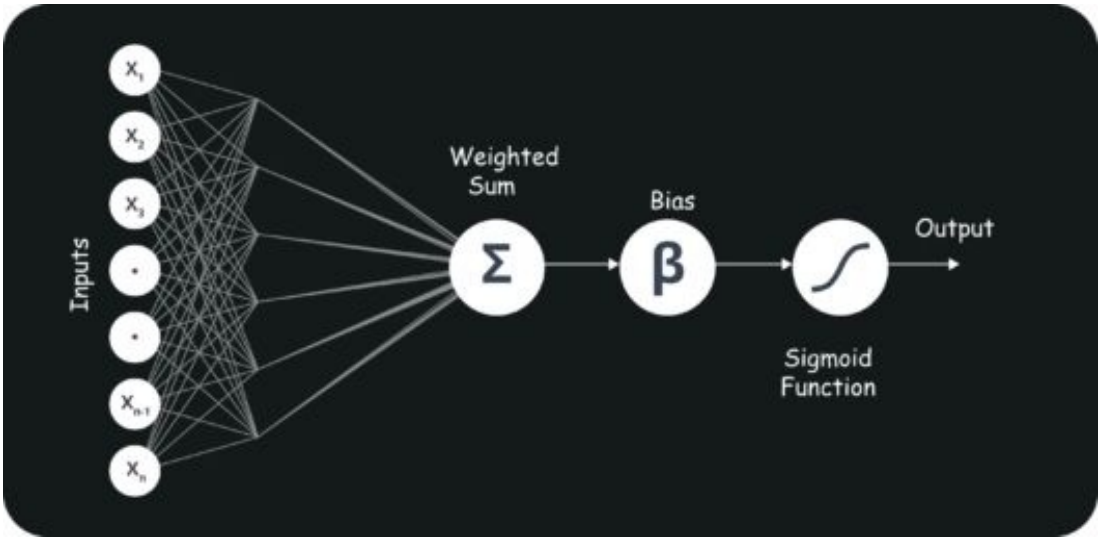
GPUs are multiple times faster and they help us build bigger and deeper deep learning models in comparatively less time than we required previously

237. What are Artificial Neural Networks?

Artificial Neural networks are a specific set of algorithms that have revolutionized machine learning. They are inspired by biological neural networks. **Neural Networks** can adapt to changing input so the network generates the best possible result without needing to redesign the output criteria.

238. Describe the structure of Artificial Neural Networks?

Artificial Neural Networks works on the same principle as a biological Neural Network. It consists of inputs which get processed with weighted sums and Bias, with the help of Activation Functions.



239. Explain Gradient Descent.

To Understand Gradient Descent, Let's understand what is a Gradient first.

A **gradient** measures how much the output of a function changes if you change the inputs a little bit. It simply measures the change in all weights with regard to the change in error. You can also think of a gradient as the slope of a function.

Gradient Descent can be thought of climbing down to the bottom of a valley, instead of climbing up a hill. This is because it is a minimization algorithm that minimizes a given function (**Activation Function**).

240. What is Back Propagation and Explain it's Working.

Backpropagation is a training algorithm used for multilayer neural network. In this method, we move the error from an end of the network to all weights inside the network and thus allowing efficient computation of the gradient.

It has the following steps:

- Forward Propagation of Training Data
- Derivatives are computed using output and target
- Back Propagate for computing derivative of error wrt output activation
- Using previously calculated derivatives for output
- Update the Weights

241. What are the variants of Back Propagation?

- **Stochastic Gradient Descent:** We use only single training example for calculation of gradient and update parameters.
- **Batch Gradient Descent:** We calculate the gradient for the whole dataset and perform the update at each iteration.
- **Mini-batch Gradient Descent**: It's one of the most popular optimization algorithms. It's a variant of Stochastic Gradient Descent and here instead of single training example, mini-batch of samples is used.

242. What are the different Deep Learning Frameworks?

- Pytorch
- TensorFlow
- Microsoft Cognitive Toolkit
- Keras
- Caffe
- Chainer

243. What is the role of Activation Function?

The Activation function is used to introduce non-linearity into the neural network helping it to learn more complex function. Without which the neural network would be only able to learn linear function which is a linear combination of its input data. An activation function is a function in an artificial neuron that delivers an output based on inputs

244. What is an Auto-Encoder?

Autoencoders are simple learning networks that aim to transform inputs into outputs with the minimum possible error. This means that we want the output to be as close to input as possible. We add a couple of layers between the input and the output, and the sizes of these layers are smaller than the input layer. The autoencoder receives unlabeled input which is then encoded to reconstruct the input.

245. What is a Boltzmann Machine?

Boltzmann machines have a simple learning algorithm that allows them to discover interesting features that represent complex regularities in the training data. The Boltzmann machine is basically used to optimize the weights and the quantity for the given problem. The learning algorithm is very slow in networks with many layers of feature detectors. "**Restricted Boltzmann Machines**" algorithm has a single layer of feature detectors which makes it faster than the

rest.

251) Which of these measures are used to analyze the central tendency of data?

A) Mean and Normal Distribution

B) Mean, Median and Mode

C) Mode, Alpha & Range

D) Standard Deviation, Range and Mean

E) Median, Range and Normal Distribution

Solution: (B)

The mean, median, mode are the three statistical measures which help us to analyze the central tendency of data. We use these measures to find the central value of the data to summarize the entire data set.

252) Five numbers are given: (5, 10, 15, 5, 15). Now, what would be the sum of deviations of individual data points from their mean?

A) 10

B)25

C) 50

D) 0

E) None of the above

Solution: (D)

The sum of deviations of the individual will always be 0.

253) A test is administered annually. The test has a mean score of 150 and a standard deviation of 20. If Ravi's z-score is 1.50, what was his score on the

test?

A) 180
B) 130
C) 30
D) 150
E) None of the above

Solution: (A)

$X = \mu + Z\sigma$ where μ is the mean, σ is the standard deviation and X is the score we're calculating. Therefore $X = 150+20*1.5 = 180$

254) Which of the following measures of central tendency will always change if a single value in the data changes?

A) Mean

B) Median

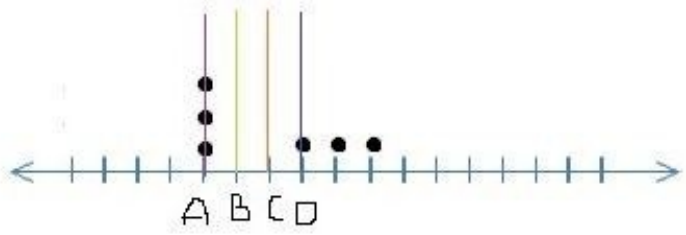
C) Mode

D) All of these

Solution: (A)

The mean of the dataset would always change if we change any value of the data set. Since we are summing up all the values together to get it, every value of the data set contributes to its value. Median and mode may or may not change with altering a single value in the dataset.

255) Below, we have represented six data points on a scale where vertical lines on scale represent unit.



Which of the following line represents the mean of the given data points, where the scale is divided into same units?

A) A
B) B
C) C
D) D

Solution: (C)

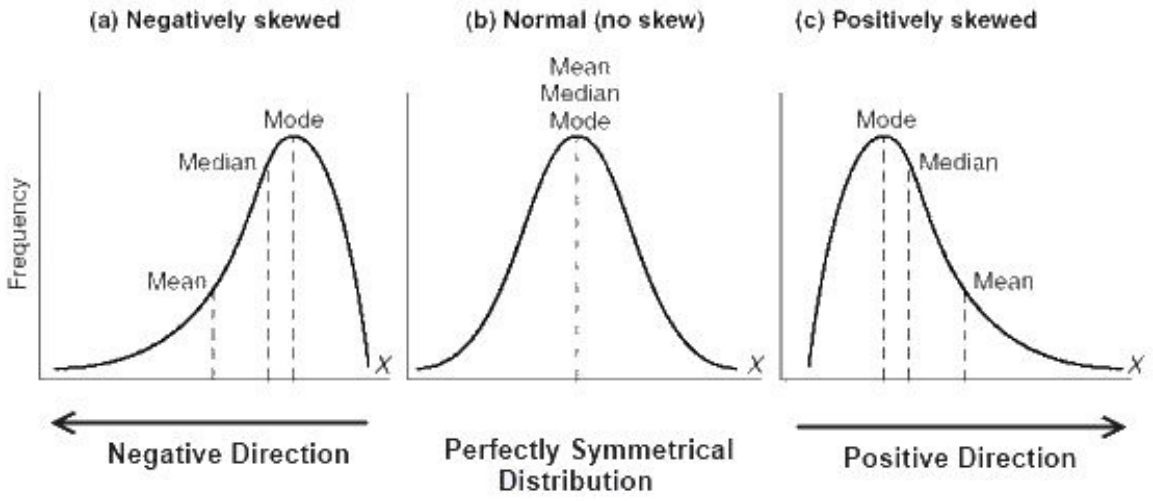
It's a little tricky to visualize this one by just looking at the data points. We can simply substitute values to understand the mean. Let A be 1, B be 2, C be 3 and so on. The data values as shown will become {1,1,1,4,5,6} which will have mean to be 18/6 = 3 i.e. C.

256) If a positively skewed distribution has a median of 50, which of the following statement is true?

A) Mean is greater than 50
B) Mean is less than 50
C) Mode is less than 50
D) Mode is greater than 50
E) Both A and C
F) Both B and D

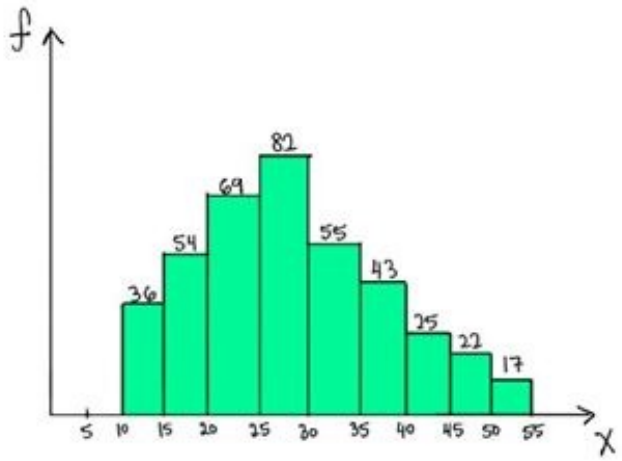
Solution: (E)

Below are the distributions for Negatively, Positively and no skewed curves.



As we can see for a positively skewed curve, Mode<Median<Mean. So if median is 50, mean would be more than 50 and mode will be less than 50.

257) Which of the following is a possible value for the median of the below distribution?



A) 32
B) 26
C) 17
D) 40

Solution: (B)

To answer this one we need to go to the basic definition of a median. Median is the value which has roughly half the values before it and half the values after.

The number of values less than 25 are (36+54+69 = 159) and the number of values greater than 30 are (55+43+25+22+17= 162). So the median should lie somewhere between 25 and 30. Hence 26 is a possible value of the median.

258) Which of the following statements are true about Bessels Correction while calculating a sample standard deviation?

1. **Bessels correction is always done when we perform any operation on a sample data.**
2. **Bessels correction is used when we are trying to estimate population standard deviation from the sample.**
3. **Bessels corrected standard deviation is less biased.**

A) Only 2

B) Only 3

C) Both 2 and 3

D) Both 1 and 3

Solution: (C)

Contrary to the popular belief Bessel's correction should not be always done. It's basically done when we're trying to estimate the population standard deviation using the sample standard deviation. The bias is definitely reduced as the standard deviation will now(after correction) be depicting the dispersion of the population more than that of the sample.

259) If the variance of a dataset is correctly computed with the formula using (n – 1) in the denominator, which of the following option is true?

A) Dataset is a sample
B) Dataset is a population
C) Dataset could be either a sample or a population
D) Dataset is from a census
E) None of the above

Solution: (A)

If the variance has n-1 in the formula, it means that the set is a sample. We try to estimate the population variance by dividing the sum of squared difference with the mean with n-1.

When we have the actual population data we can directly divide the sum of squared differences with n instead of n-1.

260) [True or False] Standard deviation can be negative.

A) TRUE

B) FALSE

Solution: (B)

Below is the formula for standard deviation

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Since the differences are squared, added and then rooted, negative standard deviations are not possible.

261) Standard deviation is robust to outliers?

A) True

B) False

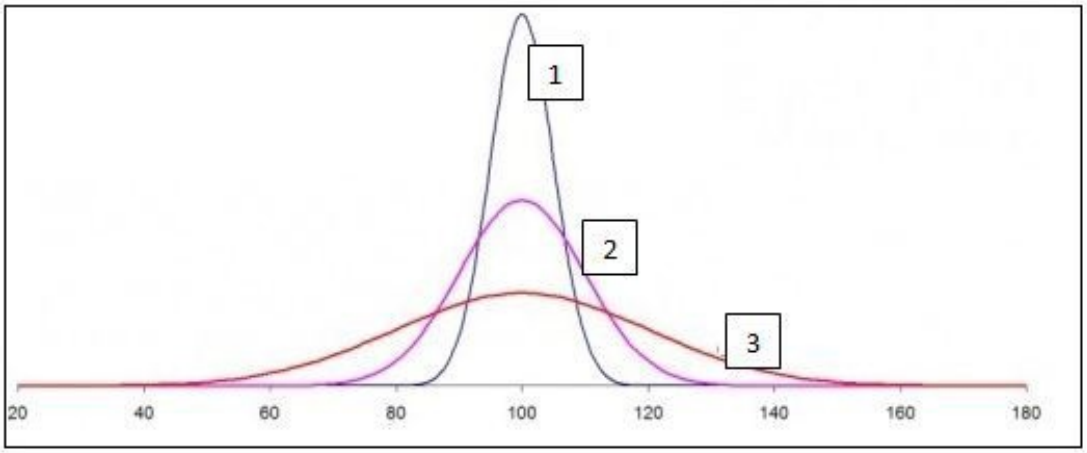
Solution: (B)

If you look at the formula for standard deviation above, a very high or a very low value would increase standard deviation as it would be very different from the mean. Hence outliers will effect standard deviation.

262) For the below normal distribution, which of the following option holds true ?

σ1, σ2 and σ3 represent the standard deviations for curves 1, 2 and 3

respectively.



A) σ1> σ2> σ3

B) σ1< σ2< σ3

C) σ1= σ2= σ3

D) None

Solution: (B)

From the definition of normal distribution, we know that the area under the curve is 1 for all the 3 shapes. The curve 3 is more spread and hence more dispersed (most of values being within 40-160). Therefore it will have the highest standard deviation. Similarly, Curve 1 has a very low range and all the values are in a small range of 80-120. Hence, curve 1 has the least standard deviation.

263) What would be the critical values of Z for 98% confidence interval for a two-tailed test ?

A) +/- 2.33
B) +/- 1.96
C) +/- 1.64
D) +/- 2.55

Solution: (A)

We need to look at the z table for answering this. For a 2 tailed test, and a 98%

confidence interval, we should check the area before the z value as 0.99 since 1% will be on the left side of the mean and 1% on the right side. Hence we should check for the z value for area>0.99. The value will be +/- 2.33

264) [True or False] The standard normal curve is symmetric about 0 and the total area under it is 1.

A)TRUE

B) FALSE

Solution: (A)

By the definition of the normal curve, the area under it is 1 and is symmetric about zero. The mean, median and mode are all equal and 0. The area to the left of mean is equal to the area on the right of mean. Hence it is symmetric.

Studies show that listening to music while studying can improve your memory. To demonstrate this, a researcher obtains a sample of 36 college students and gives them a standard memory test while they listen to some background music. Under normal circumstances (without music), the mean score obtained was 25 and standard deviation is 6. The mean score for the sample after the experiment (i.e With music) is 28.

265) What is the null hypothesis in this case?

A) Listening to music while studying will not impact memory.
B) Listening to music while studying may worsen memory.
C) Listening to music while studying may improve memory.
D) Listening to music while studying will not improve memory but can make it worse.

Solution: (D)

The null hypothesis is generally assumed statement, that there is no relationship in the measured phenomena. Here the null hypothesis would be that there is no relationship between listening to music and improvement in memory.

266) What would be the Type I error?

A) Concluding that listening to music while studying improves memory, and it's right.
B) Concluding that listening to music while studying improves memory when it actually doesn't.
C) Concluding that listening to music while studying does not improve memory but it does.

Solution: (B)

Type 1 error means that we reject the null hypothesis when its actually true. Here the null hypothesis is that music does not improve memory. Type 1 error would be that we reject it and say that music does improve memory when it actually doesn't.

267) After performing the Z-test, what can we conclude ____ ?

A) Listening to music does not improve memory.

B)Listening to music significantly improves memory at p

C) The information is insufficient for any conclusion.

D) None of the above

Solution: (B)

Let's perform the Z test on the given case. We know that the null hypothesis is that listening to music does not improve memory.

Alternate hypothesis is that listening to music does improve memory.

In this case the standard error i.e. $\frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{36}} = 1$

The Z score for a sample mean of 28 from this population is

$$Z = \frac{sample\ mean - population\ mean}{\text{standard error}} = \frac{28-25}{1} = 3$$

Z critical value for $\alpha = 0.05$ (one tailed) would be 1.65 as seen from the z table.

Therefore since the Z value observed is greater than the Z critical value, we can reject the null hypothesis and say that listening to music does improve the memory with 95% confidence.

268) A researcher concludes from his analysis that a placebo cures AIDS. What type of error is he making?

A) Type 1 error

B) Type 2 error

C) None of these. The researcher is not making an error.

D) Cannot be determined

Solution: (D)

By definition, type 1 error is rejecting the null hypothesis when its actually true and type 2 error is accepting the null hypothesis when its actually false. In this case to define the error, we need to first define the null and alternate hypothesis.

269) What happens to the confidence interval when we introduce some outliers to the data?

A) Confidence interval is robust to outliers

B) Confidence interval will increase with the introduction of outliers.

C) Confidence interval will decrease with the introduction of outliers.

D) We cannot determine the confidence interval in this case.

Solution: (B)

We know that confidence interval depends on the standard deviation of the data. If we introduce outliers into the data, the standard deviation increases, and hence the confidence interval also increases.

A medical doctor wants to reduce blood sugar level of all his patients by

altering their diet. He finds that the mean sugar level of all patients is 180 with a standard deviation of 18. Nine of his patients start dieting and the mean of the sample is observed to 175. Now, he is considering to recommend all his patients to go on a diet.

Note: He calculates 99% confidence interval.

270) What is the standard error of the mean?

A) 9
B) 6
C) 7.5
D) 18

Solution: (B)

The standard error of the mean is the standard deviation by the square root of the number of values. i.e.

Standard error = $18/\sqrt{9}$ = 6

271) What is the probability of getting a mean of 175 or less after all the patients start dieting?

A) 20%
B) 25%
C) 15%
D) 12%

Solution: (A)

This actually wants us to calculate the probability of population mean being 175 after the intervention. We can calculate the Z value for the given mean.

$$Z = \frac{\text{sample mean} - \text{population}}{\text{standard error}} = \frac{175-180}{6}$$

$$Z = -\frac{5}{6} = -0.833$$

If we look at the z table, the corresponding value for z = -0.833 ~ 0.2033.

Therefore there is around 20% probability that if everyone starts dieting, the population mean would be 175.

272) Which of the following statement is correct?

A) The doctor has a valid evidence that dieting reduces blood sugar level.

B) The doctor does not have enough evidence that dieting reduces blood sugar level.

C) If the doctor makes all future patients diet in a similar way, the mean blood pressure will fall below 160.

Solution: (B)

We need to check if we have sufficient evidence to reject the null. The null hypothesis is that dieting has no effect on blood sugar. This is a two tailed test. The z critical value for a 2 tailed test would be ±2.58.

The z value as we have calculated is -0.833.

Since Z value < Z critical value, we do not have enough evidence that dieting reduces blood sugar.

A researcher is trying to examine the effects of two different teaching methods. He divides 20 students into two groups of 10 each. For group 1, the teaching method is using fun examples. Where as for group 2 the teaching method is using software to help students learn. After a 20 minutes lecture of both groups, a test is conducted for all the students.

We want to calculate if there is a significant difference in the scores of both the groups.

It is given that:

- Alpha=0.05, two tailed.
- Mean test score for group 1 = 10
- Mean test score for group 2 = 7
- Standard error = 0.94

273) What is the value of t-statistic?

A) 3.191
B) 3.395
C) Cannot be determined.
D) None of the above

Solution: (A)

The t statistic of the given group is nothing but the difference between the group means by the standard error.

=(10-7)/0.94 = 3.191

274) Is there a significant difference in the scores of the two groups?

A) Yes
B) No

Solution: (A)

The null hypothesis in this case would be that there is no difference between the groups, while the alternate hypothesis would be that the groups are significantly different.

The t critical value for a 2 tailed test at $\alpha = 0.05$ is ± 2.101. The t statistic obtained is 3.191. Since the t statistic is more than the critical value of t, we can reject the null hypothesis and say that the two groups are significantly different with 95% confidence.

275) What percentage of variability in scores is explained by the method of teaching?

A) 36.13
B) 45.21
C) 40.33
D) 32.97

Solution: (A)

The % variability in scores is given by the R^2 value. The formula for R^2 given by

$$R^2 = \frac{\text{t square}}{\text{t square} + \text{degree of freedom}}$$

The degrees of freedom in this case would be 10+10 -2 since there are two groups with size 10 each. The degree of freedom is 18.

$$R^2 = \frac{3.191 * 3.191}{(3.191 * 3.191) + 18} = 36.13$$

276) [True or False] F statistic cannot be negative.

A) TRUE

B) FALSE

Solution: (A)

F statistic is the value we receive when we run an ANOVA test on different groups to understand the differences between them. The F statistic is given by the ratio of between group variability to within group variability

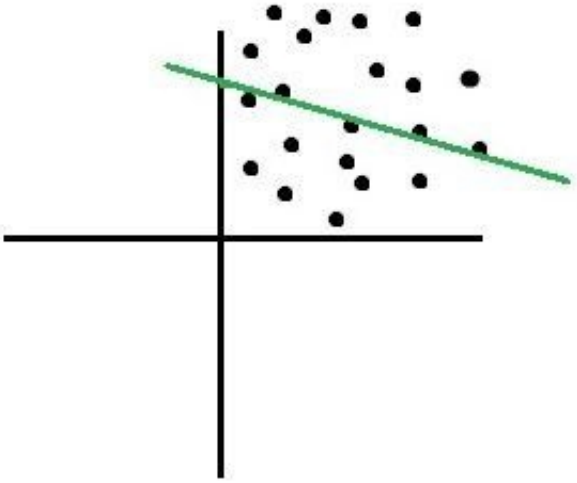
Below is the formula for f Statistic.

$$\frac{\text{Sum of squared error for between group/degree of freedom of between group}}{\text{Sum of squared error for within group/degree of freedom of within group}}$$

Since both the numerator and denominator possess square terms, F statistic cannot be negative.

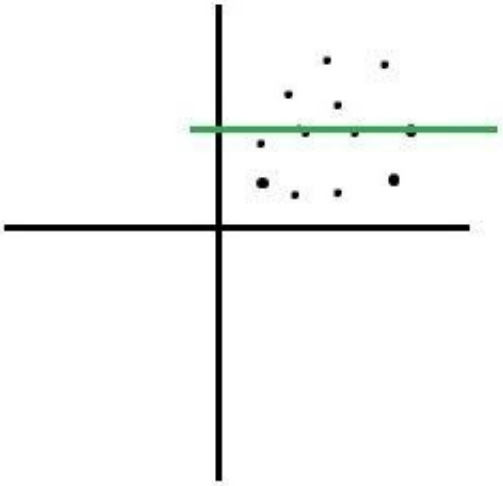
277) Which of the graph below has very strong positive correlation?

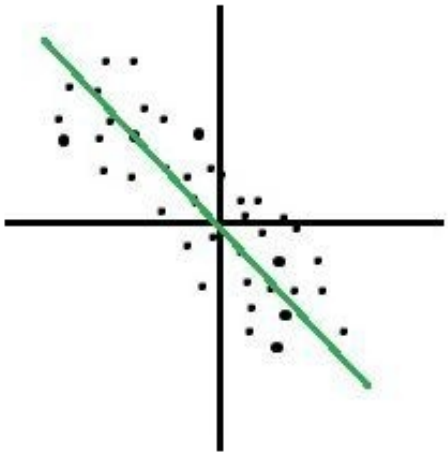
A)



B)

C)





D)

Solution: (B)

A strong positive correlation would occur when the following condition is met. If x increases, y should also increase, if x decreases, y should also decrease. The slope of the line would be positive in this case and the data points will show a clear linear relationship. Option B shows a strong positive relationship.

278) Correlation between two variables (Var1 and Var2) is 0.65. Now, after adding numeric 2 to all the values of Var1, the correlation co-efficient will_____ ?

A) Increase
B) Decrease
C) None of the above

Solution: (C)

If a constant value is added or subtracted to either variable, the correlation coefficient would be unchanged. It is easy to understand if we look at the formula for calculating the correlation.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

If we add a constant value to all the values of x, the x_i and will change by the same number, and the differences will remain the same. Hence, there is no change in the correlation coefficient.

279) It is observed that there is a very high correlation between math test scores and amount of physical exercise done by a student on the test day. What can you infer from this?

1. **High correlation implies that after exercise the test scores are high.**
2. **Correlation does not imply causation.**
3. **Correlation measures the strength of linear relationship between amount of exercise and test scores.**

A) Only 1
B) 1 and 3
C) 2 and 3
D) All the statements are true

Solution: (C)

Though sometimes causation might be intuitive from a high correlation but actually correlation does not imply any causal inference. It just tells us the strength of the relationship between the two variables. If both the variables move together, there is a high correlation among them.

280) If the correlation coefficient (r) between scores in a math test and amount of physical exercise by a student is 0.86, what percentage of variability in math test is explained by the amount of exercise?

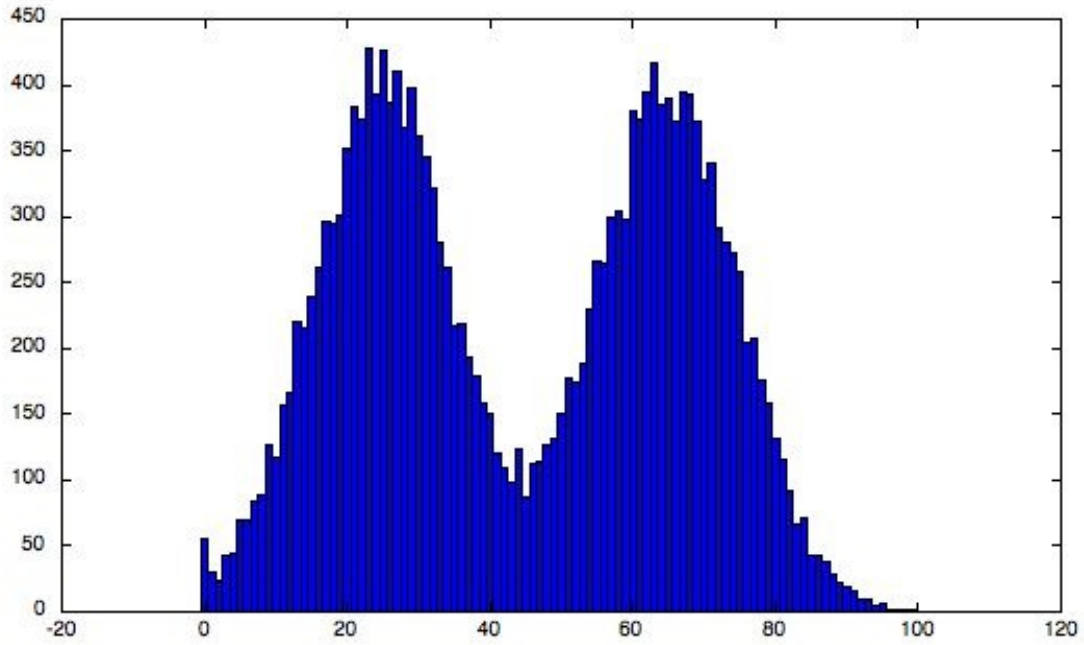
A) 86%
B) 74%
C) 14%
D) 26%

Solution: (B)

The % variability is given by r^2, the square of the correlation coefficient. This value represents the fraction of the variation in one variable that may be explained by the other variable. Therefore % variability explained would be

0.86^2.

281) Which of the following is true about below given histogram?



A) Above histogram is unimodal

B) Above histogram is bimodal

C) Given above is not a histogram

D) None of the above

Solution: (B)

The above histogram is bimodal. As we can see there are two values for which we can see peaks in the histograms indicating high frequencies for those values. Therefore the histogram is bimodal.

282) Consider a regression line y=ax+b, where a is the slope and b is the intercept. If we know the value of the slope then by using which option can we always find the value of the intercept?

A) Put the value (0,0) in the regression line True

B) Put any value from the points used to fit the regression line and compute the value of b False

C) Put the mean values of x & y in the equation along with the value a to get b False

D) None of the above can be used False

Solution: (C)

In case of ordinary least squares regression, the line would always pass through the mean values of x and y. If we know one point on the line and the value of slope, we can easily find the intercept.

283) What happens when we introduce more variables to a linear regression model?

A) The r squared value may increase or remain constant, the adjusted r squared may increase or decrease.

B) The r squared may increase or decrease while the adjusted r squared always increases.

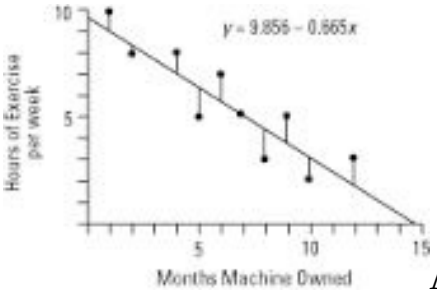
C) Both r square and adjusted r square always increase on the introduction of new variables in the model.

D) Both might increase or decrease depending on the variables introduced.

Solution: (A)

The R square always increases or at least remains constant because in case of ordinary least squares the sum of square error never increases by adding more variables to the model. Hence the R squared does not decrease. The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance.

284) In a scatter diagram, the vertical distance of a point above or below regression line is known as ____ ?



A) Residual
B) Prediction Error
C) Prediction
D) Both A and B
E) None of the above

Solution: (D)

The lines as we see in the above plot are the vertical distance of points from the regression line. These are known as the residuals or the prediction error.

285) In univariate linear least squares regression, relationship between correlation coefficient and coefficient of determination is _____ ?

A) Both are unrelated False

B) The coefficient of determination is the coefficient of correlation squared True

C) The coefficient of determination is the square root of the coefficient of correlation False

D) Both are same F

Solution: (B)

The coefficient of determination is the R squared value and it tells us the amount of variability of the dependent variable explained by the independent variable. This is nothing but correlation coefficient squared. In case of multivariate regression the r squared value represents the ratio of the sum of explained variance to the sum of total variance.

286) What is the relationship between significance level and confidence level?

A) Significance level = Confidence level
B) Significance level = 1- Confidence level
C) Significance level = 1/Confidence level
D) Significance level = sqrt (1 – Confidence level)

Solution: (B)

Significance level is 1-confidence interval. If the significance level is 0.05, the corresponding confidence interval is 95% or 0.95. The significance level is the probability of obtaining a result as extreme as, or more extreme than, the result actually obtained when the null hypothesis is true. The confidence interval is the range of likely values for a population parameter, such as the population mean. For example, if you compute a 95% confidence interval for the average price of an ice cream, then you can be 95% confident that the interval contains the true average cost of all ice creams.

The significance level and confidence level are the complementary portions in the normal distribution.

287) [True or False] Suppose you have been given a variable V, along with its mean and median. Based on these values, you can find whether the variable "V" is left skewed or right skewed for the condition

```
mean(V) > median(V)
```

A) True
B) False

Solution: (B)

Since, its no where mentioned about the type distribution of the variable V, we cannot say whether it is left skewed or right skewed for sure.

288) The line described by the linear regression equation (OLS) attempts to ____ ?

A) Pass through as many points as possible.

B) Pass through as few points as possible

C) Minimize the number of points it touches

D) Minimize the squared distance from the points

Solution: (D)

The regression line attempts to minimize the squared distance between the points and the regression line. By definition the ordinary least squares regression tries to have the minimum sum of squared errors. This means that the sum of squared residuals should be minimized. This may or may not be achieved by passing through the maximum points in the data. The most common case of not passing through all points and reducing the error is when the data has a lot of outliers or is not very strongly linear.

289) We have a linear regression equation (Y = 5X +40) for the below table.

X	Y
5	45
6	76
7	78
8	87
9	79

Which of the following is a MAE (Mean Absolute Error) for this linear model?

A) 8.4
B) 10.29
C) 42.5
D) None of the above

Solution: (A)

To calculate the mean absolute error for this case, we should first calculate the values of y with the given equation and then calculate the absolute error with respect to the actual values of y. Then the average value of this absolute error would be the mean absolute error. The below table summarises these values.

X	Y	5X+40	Absolute Error
5	45	65	20
6	76	70	6
7	78	75	3
8	87	80	7
9	79	85	6
		Mean error	**8.4**

290) A regression analysis between weight (y) and height (x) resulted in the following least squares line: y = 120 + 5x. This implies that if the height is increased by 1 inch, the weight is expected to

A) increase by 1 pound
B) increase by 5 pound
C) increase by 125 pound
D) None of the above

Solution: (B)

Looking at the equation given y=120+5x. If the height is increased by 1 unit, the weight will increase by 5 pounds. Since 120 will be the same in both cases and will go off in the difference.

291) [True or False] Pearson captures how linearly dependent two variables are whereas Spearman captures the monotonic behaviour of the relation between the variables.

A)TRUE

B) FALSE

Solution: (A)

The statement is true. Pearson correlation evaluated the linear relationship between two continuous variables. A relationship is linear when a change in one variable is associated with a proportional change in the other variable.

The spearman evaluates a monotonic relationship. A monotonic relationship is one where the variables change together but not necessarily at a constant rate.

292) What do you understand by long and wide data formats?
293) What do you understand by outliers and inliers? What would you do if you find them in your dataset?
294) 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.
295) What are the basic assumptions to be made for linear regression?
Normality of error distribution, statistical independence of errors, linearity and additivity.

296) Can you write the formula to calculat R-square?
R-Square can be calculated using the below formular -

1 - (Residual Sum of Squares/ Total Sum of Squares)

297) 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.

298) 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 accessed using Hypothesis Testing.

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

300) How will you find the correlation between a categorical variable and a continuous variable ?
You can use the analysis of covariance technqiue to find the correlation between a categorical variable and a continuous variable.

Q301. What are the different sampling methods?

- Random Sampling
- Systematic Sampling

- Stratified Sampling
- Quota Sampling

Q302. Common Data Quality Issues

- Missing Values
- Noise in the Data Set
- Outliers
- Mixture of Different Languages (like English and Chinese)
- Range Constraints

Q303. What is the difference between supervised learning and un-supervised learning?

Supervised learning: Target variable is available and the algorithm learns for the train data

And applies to test data (unseen data).

Unsupervised learning: Target variable is not available and the algorithm does not need to learn

Anything beforehand.

Q304. What is Imbalanced Data Set and how to handle them? Name Few Examples?

- Fraud detection
- Disease screening

Imbalanced Data Set means that the population of one class is extremely large than the other

(Eg: Fraud – 99% and Non-Fraud – 1%)

Imbalanced dataset can be handled by either oversampling, undersampling and penalized Machine Learning Algorithm.

Q305. If you are dealing with 10M Data, then will you go for Machine learning (or) Deep learning Algorithm?

- Machine learning algorithmsuits well for small data and it might take huge amount of time to train for large data.
- Whereas Deep learning algorithm takes less amount of data to train due to the help of GPU(Parallel Processing).

Q306. Examples of Supervised learning algorithm?

- Linear Regression and Logistic Regression
- Decision Trees and Random Forest
- SVM
- Naïve Bayes
- XGBoost

Q307. In Logistic Regression, if you want to know the best features in your dataset then what you would do?

Apply step function, which calculates the AIC for different permutation and combination of features and provides the best features for the dataset.

Q308. What is Feature Engineering? Explain with Example?

Feature engineering is the process of using domain knowledge of the data to create features for machine learning algorithm to work

- Adding more columns (or) removing columns from the existing column
- Outlier Detection
- Normalization etc

Q309. How to select the important features in the given data set?

- In Logistic Regression, we can use step() which gives AIC score of set of features
- In Decision Tree, We can use information gain(which internally uses entropy)
- In Random Forest, We can use varImpPlot

Q310. When does multicollinearity problem occur and how to handle it?

It exists when 2 or more predictors are highly correlated with each other.

Example: In the Data Set if you have grades of 2nd PUC and marks of 2nd PUC, Then both gives the same trend to capture, which might internally hamper the speed and time.so we need to check if the multi collinearity exists by using VIF(variance Inflation Factor).

Note: if the Variance Inflation Factor is more than 4, then multi collinearity problem exists.

Q311. What is Variance inflation Factors (VIF)

Measure how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related.

Q312. Examples of Parametric machine learning algorithm and non-parametric machine learning algorithm

- Parametric machine learning algorithm– Linear Regression, Logistic Regression
- Non-Parametric machine learning algorithm – Decision Trees, SVM, Neural Network

Q313. What are parametric and non-parametric machine learning algorithm? And their importance

Algorithm which does not make strong assumptions are non-parametric algorithm and they are free to learn from training data. Algorithm that makes

strong assumptions are parametric and it involves

1. select the form for the function and
2. learn the coefficients for the function from training data.

Q314. When does linear and logistic regression performs better, generally?

It works better when we remove the attributes which are unrelated to the output variable and highly co-related variable to each other.

Q315. Why you call naïve bayes as "naïve" ?

Reason: It assumes that the input variable is independent, but in real world it is unrealistic, since all the features would be dependent on each other.

Q316. Give some example for false positive, false negative, true positive, true negative

- False Positive – A cancer screening test comes back positive, but you don't have cancer
- False Negative – A cancer screening test comes back negative, but you have cancer
- True Positive – A Cancer Screening test comes back positive, and you have cancer
- True Negative – A Cancer Screening test comes back negative, and you don't have cancer

Q317. What is Sensitivity and Specificity?

Sensitivity means "proportion of actual positives that are correctly classified" in other words "True Positive"

Specificity means "proportion of actual negatives that are correctly classified" "True Negative"

Q318. When to use Logistic Regression and when to use Linear Regression?

If you are dealing with a classification problem like (Yes/No, Fraud/Non Fraud, Sports/Music/Dance) then use Logistic Regression.

If you are dealing with continuous/discrete values, then go for Linear Regression.

Q319. What are the different imputation algorithm available?

Imputation algorithm means "replacing the Blank values by some values)

- Mean imputation
- Median Imputation
- MICE
- miss forest
- Amelia

Q320. What is AIC(Akaike Information Criteria)

The analogous metric of adjusted R^2 in logistic regression is AIC.

AIC is the measure of fit which penalizes model for the number of model coefficients. Therefore, we always prefer model with minimum AIC value.

Q321. Suppose you have 10 samples, where 8 are positive and 2 are negative, how to calculate Entropy (important to know)

E(S) = 8/10log(8/10) – 2/10log(2/10)

Note: Log is à base 2

Q322. What is perceptron in Machine Leaning?

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

Q323. How to ensure we are not over fitting the model?

- Keep the attributes/Columns which are really important
- Use K-Fold cross validation techniques
- Make use of drop-put incase of neural network

Q324.How the root node is predicted in Decision Tree Algorithm?

Mathematical Formula “Entropy” is utilized for predicting the root node of the tree.

Q325. What are the different Backend Process available in Keras?

- TensorFlow
- Theano
- CNTK

Q326. Name Few Deep Learning Algorithm

- TensorFlow
- Theano
- Lasagne
- mxnet
- blocks
- Keras
- CNTK
- TFLearn

Q327. How to split the data with equal set of classes in both training and testing data?

Using Stratified Shuffle package

Q328. What do you mean by giving “epoch = 1” in neural network?

It means that “traversing the data set one time

Q329. What do you mean by Ensemble Model? When to use?

Ensemble Model is a combination of Different Models to predict correctly and

with good accuracy.

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

Q330. When will you use SVM and when to use Random Forest?

- SVM can be used if the data is outlier free whereas Naïve Bayes can be used even if it has outliers (since it has built in package to take care).
- SVM suits best for Text Classification Model and Random Forest suits for Binomial/Multinomial Classification Problem.
- Random Forest takes care of over fitting problem with the help of tree pruning

Q331. Applications of Machine Learning?

- Self Driving Cars
- Image Classification
- Text Classification
- Search Engine
- Banking, Healthcare Domain

Q332. If you are given with a use case – 'Predict whether the transaction is fraud (or) not fraud", which algorithm would you choose

Logistic Regression

Q333. If you are given with a use case – 'Predict the house price range in the coming years", which algorithm would you choose

Linear Regression

Q334. What is the underlying mathematical knowledge behind Naïve Bayes?

Bayes Theorem

Q335. When to use Random Forest and when to Use XGBoost?

If you want all core processors in your system to be utilized, then go for XGBoost(since it supports parallel processing) and if your data is small then go for random forest.

Q336. If you are training model gives 90% accuracy and test model gives 60% accuracy? Then what problem you are facing with?

Overfitting.

Overfitting and can be reduced by many methods like (Tree Pruning, Removing the minute information provided in the data set).

Q337. In Google if you type "How are "it gives you the recommendation as "How are you "/"How do you do", this is based on what?

This kind of recommendation engine comes from collaborative filtering.

Q338. What is margin, kernels, Regularization in SVM?

- Margin – Distance between the hyper plane and closest data points is referred as “margin”
- Kernels – there are three types of kernel which determines the type of data you are dealing with i) Linear, ii) Radial, iii) Polynomial
- Regularization – The Regularization parameter (often termed as C parameter in python’s sklearn library) tells the SVM optimization how much you want to avoid misclassifying each training example

Q339. What is Boosting? Explain how Boosting works?

Boosting is a Ensemble technique that attempts to create strong classifier from a number of weak classifiers

- After the first tree is created, the performance of the tree on each training instance is used to weight how much attention the next tree that is created should pay attention to each training instance by giving more weights to the misclassified one.
- Models are created one after the other, each updating the weights on the training instance

Q340. What is Null Deviance and Residual Deviance (Logistic Regression Concept?)

Null Deviance indicates the response predicted by a model with nothing but an intercept

Residual deviance indicates the response predicted by a model on adding independent variables

Note:

Lower the value, better the model

Q341. What are the different method to split the tree in decision tree?

Information gain and gini index

Q342. What is the weakness for Decision Tree Algorithm?

Not suitable for continuous/Discrete variable

Performs poorly on small data

Q343. Why do we use PCA(Principal Components Analysis) ?

These are important feature extraction techniques used for dimensionality reduction.

Q344. During Imbalanced Data Set, will you

- Calculate the Accuracy only? (or)
- Precision, Recall, F1 Score separately

We need to calculate precision, Recall separately

Q345.How to ensure we are not over fitting the model?

- Keep the attributes/Columns which are really important
- Use K-Fold cross validation techniques
- make use of drop-put in case of neural network

Q346. Steps involved in Decision Tree and finding the root node for the tree

Step 1:- How to find the Root Node

Use Information gain to understand the each attribute information w.r.t target variable and place the attribute with the highest information gain as root node.

Step 2:- How to Find the Information Gain

Please apply the entropy (Mathematical Formulae) to calculate Information Gain. Gain (T,X) = Entropy(T) – Entropy(T,X) here represent target variable and X represent features.

Step3: Identification of Terminal Node

Based on the information gain value obtained from the above steps, identify the second most highest information gain and place it as the terminal node.

Step 4: Predicted Outcome

Recursively iterate the step4 till we obtain the leaf node which would be our predicted target variable.

Step 5: Tree Pruning and optimization for good results

It helps to reduce the size of decision trees by removing sections of the tree to avoid over fitting.

Q347. What is hyper plane in SVM?

It is a line that splits the input variable space and it is selected to best separate the points in the input variable space by their class(0/1,yes/no).

Q348. Explain Bigram with an Example?

Eg: I Love Data Science

Bigram – (I Love) (Love Data) (Data Science)

Q349. What are the different activation functions in neural network?

Relu, Leaky Relu , Softmax, Sigmoid

Q350. Which Algorithm Suits for Text Classification Problem?

SVM, Naïve Bayes, Keras, Theano, CNTK, TFLearn(Tensorflow)

Q351. You are given a train data set having lot of columns and rows. How do you reduce the dimension of this data?

- Principal Component Analysis(PCA) would help us here which can explain the maximum variance in the data set.
- We can also check the co-relation for numerical data and remove the

problem of multi-collinearity(if exists) and remove some of the columns which may not impact the model.
- We can create multiple dataset and execute them batch wise.

Q352. You are given a data set on fraud detection. Classification model achieved accuracy of 95%.Is it good?

Accuracy of 96% is good. But we may have to check the following items:

- what was the dataset for the classification problem
- Is Sensitivity and Specificity are acceptable
- if there are only less negative cases, and all negative cases are not correctly classified, then it might be a problem

In-Addition it is related to fraud detection, hence needs to be careful here in prediction (i.e not wrongly predicting the fraud as non-fraud patient.

Q353. What is prior probability and likelihood?

Prior probability:
The proportion of dependent variable in the data set.

Likelihood:
It is the probability of classifying a given observation as '1' in the presence of some other variable.

Q354. How can we know if your data is suffering from low bias and high variance?

Random Forest Algorithm can be used to tackle high variance problem.in the cases of low bias and high variance L1,L2 regularization can help.

Q355. How is kNN different from kmeans clustering?

Kmeans partitions a data set into clusters, which is homogeneous and points in the cluster are close to each other. Whereas KNN tries to classify unlabelled observation based on its K surrounding neighbours.

Q356. Random Forest has 1000 trees, Training error: 0.0 and validation error is 20.00.What is the issue here?

It is the classical example of over fitting. It is not performing well on the unseen data. We may have to tune our model using cross validation and other techniques to overcome over fitting

Q357. Data set consisting of variables having more than 30% missing values? How will you deal with them?

We can remove them, if it does not impact our model
We can apply imputation techniques (like MICE, MISSFOREST,AMELIA) to avoid missing values

Q358. What do you understand by Type I vs. Type II error?

Type I error occurs when – “we classify a value as positive, when the actual value is negative”

(False Positive)

Type II error occurs when – “we classify a value as negative, when the actual value if positive”

(False Negative)

Q359. Based on the dataset, how will you know which algorithm to apply ?

- If it is classification related problem,then we can use logistic,decision trees etc…
- If it is Regression related problem, then we can use Linear Regression.
- If it is Clustering based, we can use KNN.
- We can also apply XGB, RF for better accuracy.

Q360. Why normalization is important?

Data Set can have one column in the range (10,000/20,000) and other column might have data in the range (1, 2, 3).clearly these two columns are in different range and cannot accurately analyse the trend. So we can apply normalization here by using min-max normalization (i.e to convert it into 0-1 scale).

Q361. What is Data Science?

Formally, It’s the way to Quantify your intuitions.
Technically, Data Science is a combination of Machine Learning, Deep Learning & Artificial
Intelligence. Where Deep Learning is the subset of AI.

Q362. What is Machine Learning?

Machine learning is the process of generating the predictive power using past data(memory). It is a
one-time process where the predictions can fail in the future (if your data distribution changes).

Q363. What is Deep Learning?

Deep Learning is the process of adding one more logic to the machine learning, where it iterates
itself with the new data and will not fail in future, even though your data distribution changes. The
more it iterates, more it works better.

Q364. Where to use R & Python?

R can be used whenever the data is structed. Python is efficient to handle unstructured data. R can’t

handle high volume data. Python backend working with Theano/tensor made it easy to perform it as
fast comparing with R.

Q365. Which Algorithms are used to do a Binary classification?

Logistic Regression, KNN, Random Forest, CART, C50 are few algorithms which can perform Binary
classification.

Q366. Which Algorithms are used to do a Multinomial classification?

Naïve Bayes, Random Forest are widely used for multinomial classification.

Q367. What is LOGIT function?

LOGIT function is Log of ODDS ratio. ODDS ratio can be termed as the Probability of success divided
by Probability of failure. Which is the final probability value of your binary classification, where we
use ROC curve to get the cut-Off value of the probability.

Q368. What are all the pre-processing steps that are highly recommended?

- Structural Analysis
- Outlier Analysis
- Missing value treatments
- Feature engineering

Q369. What is Normal Distribution?

Whenever data that defines with having Mean = Median = Mode, then the data is called as normally
distributed data.

Q370. What is empirical Rule?

Empirical Rule says that whenever data is normally distributed, your data should be having the
distribution in a way of,
68 percent of your data spread is within Plus or Minus 1 standard deviation
95 percent of your data spread is within Plus or Minus 2 standard deviation
99.7 percent of your data spread is within Plus or Minus 3 standard deviation

Question 371. What Is Bayesian?

Answer:

Bayesians condition on the data actually observed and consider the probability distribution on the hypotheses.

Question 372. What Is Frequentist?

Answer:

Frequentists condition on a hypothesis of choice and consider the probability distribution on the data, whether observed or not.

Question 373. What Is Likelihood?

Answer:

The probability of some observed outcomes given a set of parameter values is regarded as the likelihood of the set of parameter values given the observed outcomes.

Question 374. What Is P-value?

Answer:

In statistical significance testing, the p-value is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. If the p-value is less than 0.05 or 0.01, corresponding respectively to a 5% or 1% chance of rejecting the null hypothesis when it is true.

Question 375. Give An Example Of P-value?

Answer:

Suppose that the experimental results show the coin turning up heads 14 times out of 20 total flips

> null hypothesis (H0): fair coin;
> observation O: 14 heads out of 20 flips; and
> p-value of observation O given H0 = Prob(≥ 14 heads or ≥ 14 tails) = 0.115.

The calculated p-value exceeds 0.05, so the observation is consistent with the null hypothesis - that the observed result of 14 heads out of 20 flips can be ascribed to chance alone - as it falls within the range of what would happen 95% of the time were this in fact the case. In our example, we fail to reject the null hypothesis at the 5% level. Although the coin did not fall evenly, the deviation from expected outcome is small enough to be reported as being "not statistically significant at the 5% level".

Question 376. What Is Sampling?

Answer:

Sampling is that part of statistical practice concerned with the selection of an unbiased or random subset of individual observations within a population of

individuals intended to yield some knowledge about the population of concern.

Question 377. What Are Sampling Methods?

Answer:

There are four sampling methods:

Simple Random (purely random),
Systematic(every kth member of population),
Cluster (population divided into groups or clusters)
Stratified (divided by exclusive groups or strata, sample from each group) samplings.

Question 378. What Is Mode?

Answer:

The mode of a data sample is the element that occurs most often in the collection.

x=[1 2 3 3 3 4 4]

mode(x) % return 3, happen most.

Question 379. What Is Median?

Answer:

Median is described as the numeric value separating the higher half of a sample, a population, or a probability distribution, from the lower half. The median of a finite list of numbers can be found by arranging all the observations from lowest value to highest value and picking the middle one

median(x) % return 3.

Question 380. What Is Quartile?

Answer:

second quartile (50th percentile) .
third quartile (75th percentile) .
kth percentile.
prctile(x, 25) % 25th percentile, return 2.25.
prctile(x, 50) % 50th percentile, return 3, i.e. median.

Question 381. What Is Skewness?

Answer:

Skewness is a measure of the asymmetry of the data around the sample mean. If skewness is negative, the data are spread out more to the left of the mean than to the right. If skewness is positive, the data are spread out more to the right.

Skewness(x) % return-0.5954

Question 382. What Is Variance?

Answer:

variance describes how far values lie from the mean.

var(x) %return 1.1429

Question 383. What Is Kurtosis?

Answer:

Kurtosis is a measure of how outlier-prone a distribution is.

kurtosis(x) % return2.3594

Question 384. What Is Moment?

Answer:

Quantitative measure of the shape of a set of points.

moment(x, 2); %return second moment

Question 385. What Is Covariance?

Answer:

Measure of how much two variables change together.

y2=[1 3 4 5 6 7 8]

cov(x,y2) %return 2*2 matrix, diagonal represents variance.

Question 386. What Is One Sample T-test?

Answer:

T-test is any statistical hypothesis test in which the test statistic follows a Student's t distribution if the null hypothesis is supported.

[h,p,ci] = ttest(y2,0)% return 1 0.0018 ci =2.6280 7.0863

Question 387. What Is Alternative Hypothesis?

Answer 9:

The Alternative hypothesis (denoted by H1) is the statement that must be true if the null hypothesis is false.

Question 388. What Is Significance Level?

Answer:

The probability of rejecting the null hypothesis when it is called the significance level α , and very common choices are $\alpha = 0.05$ and $\alpha = 0.01$.

Question 389. Give Example Of Central Limit Theorem?

Answer:

Given that the population of men has normally distributed weights, with a mean of 173 lb and a standard deviation of 30 lb, find the probability that

a. if 1 man is randomly selected, his weight is greater than 180 lb.

b. if 36 different men are randomly selected, their mean weight is greater that

180 lb.

Solution: a) $z = (x - \mu)/ \sigma = (180-173)/30 = 0.23$

For normal distribution $P(Z>0.23) = 0.4090$

b) $\sigma \bar{x} = \sigma/\sqrt{n} = 20/\sqrt{36} = 5$

$z= (180-173)/5 = 1.40$

$P(Z>1.4) = 0.0808$

Question 390. What Is Binomial Probability Formula?

Answer:

$P(x)= p \times q^{n-x} n!/[(n-x)!x!]$

where n = number of trials.

x = number of successes among n trials.

p = probability of success in any one trial.

q = 1 -p.

Question 391. Do You Know What Is Binary Search?

Answer:

For binary search, the array should be arranged in ascending or descending order. In each step, the algorithm compares the search key value with the key value of the middle element of the array. If the keys match, then a matching element has been found and its index, or position, is returned. Otherwise, if the search key is less than the middle element's key, then the algorithm repeats its action on the sub-array to the left of the middle element or, if the search key is greater, on the sub-array to the right.

Question 392. Explain Hash Table?

Answer:

A hash table is a data structure used to implement an associative array, a structure that can map keys to values. A hash table uses a hash function to compute an index into an array of buckets or slots, from which the correct value can be found.

Question 393. Explain Central Limit Theorem?

Answer:

As the sample size increases, the sampling distribution of sample means approaches a normal distribution.

If all possible random samples of size n are selected from a population with mean μ and standard deviation σ, the mean of the sample means is denoted by $\mu \bar{x}$, so,

$\mu \bar{x} = \mu$

the standard deviation of the sample means is:

$\sigma \bar{x} = \sigma/\sqrt{n}$

Question 394. What Is Null Hypothesis?

Answer:

The null hypothesis (denote by H0) is a statement about the value of a population parameter (such as mean), and it must contain the condition of equality and must be written with the symbol =, ≤, or ≤.

Question 395. What Is Linear Regression?

Answer:

Modeling the relationship between a scalar variable y and one or more variables denoted X. In linear regression, models of the unknown parameters are estimated from the data using linear functions.

polyfit(x,y2,1) %return 2.1667 -1.3333, i.e 2.1667x-1.3333

Question 396. When You Are Creating A Statistical Model How Do You Prevent Over-fitting?

Answer:

Over-fitting can be prevented by cross-validation.

Question 397. What Is Descriptive Statistics?

Answer:

We study in descriptive statistics the methods for organizing, displaying, and describing data.

Question 398. What Is A Sample?

Answer:

When data are collected in a statistical study for only a portion or subset of all elements of interest we are using a Sample.

Question 399. Give An Example Of Inferential Statistics?

Answer:

Example of Inferential Statistic :

You asked five of your classmates about their height. On the basis of this information, you stated that the average height of all students in your university or college is 67 inches.

Question 400. A Normal Population Distribution Is Needed For The Which Of The Statistical Tests:

Answer:

variance estimation.
standard error of the mean.
Student's t-test.

Q401. (Given a Dataset) Analyze this dataset and give me a model that can predict this response variable.

Start by fitting a simple model (multivariate regression, logistic regression), do some feature engineering accordingly, and then try some complicated models.
Always split the dataset into train, validation, test dataset and use cross validation to check their performance.
Determine if the problem is classification or regression
Favor simple models that run quickly and you can easily explain.
Mention cross validation as a means to evaluate the model.
Plot and visualize the data.

Q402. What could be some issues if the distribution of the test data is significantly different than the distribution of the training data?

The model that has high training accuracy might have low test accuracy. Without further knowledge, it is hard to know which dataset represents the population data and thus the generalizability of the algorithm is hard to measure. This should be mitigated by repeated splitting of train vs test dataset (as in cross validation).
When there is a change in data distribution, this is called the dataset shift. If the train and test data has a different distribution, then the classifier would likely overfit to the train data.
This issue can be overcome by using a more general learning method.
This can occur when:
P(y|x) are the same but P(x) are different. (covariate shift)
P(y|x) are different. (concept shift)
The causes can be:
Training samples are obtained in a biased way. (sample selection bias)
Train is different from test because of temporal, spatial changes. (non-stationary environments)
Solution to covariate shift
importance weighted cv

Q403. What are some ways I can make my model more robust to outliers?

We can have regularization such as L1 or L2 to reduce variance (increase bias).
Changes to the algorithm:
Use tree-based methods instead of regression methods as they are more resistant to outliers. For statistical tests, use non parametric tests instead of parametric ones.
Use robust error metrics such as MAE or Huber Loss instead of MSE.

Changes to the data:
Winsorizing the data
Transforming the data (e.g. log)
Remove them only if you're certain they're anomalies not worth predicting

Q404. What are some differences you would expect in a model that minimizes squared error, versus a model that minimizes absolute error? In which cases would each error metric be appropriate?

MSE is more strict to having outliers. MAE is more robust in that sense, but is harder to fit the model for because it cannot be numerically optimized. So when there are less variability in the model and the model is computationally easy to fit, we should use MAE, and if that's not the case, we should use MSE.
MSE: easier to compute the gradient, MAE: linear programming needed to compute the gradient
MAE more robust to outliers. If the consequences of large errors are great, use MSE
MSE corresponds to maximizing likelihood of Gaussian random variables

Q405. What error metric would you use to evaluate how good a binary classifier is? What if the classes are imbalanced? What if there are more than 2 groups?

Accuracy: proportion of instances you predict correctly. Pros: intuitive, easy to explain, Cons: works poorly when the class labels are imbalanced and the signal from the data is weak
AUROC: plot fpr on the x axis and tpr on the y axis for different threshold. Given a random positive instance and a random negative instance, the AUC is the probability that you can identify who's who. Pros: Works well when testing the ability of distinguishing the two classes, Cons: can't interpret predictions as probabilities (because AUC is determined by rankings), so can't explain the uncertainty of the model
logloss/deviance: Pros: error metric based on probabilities, Cons: very sensitive to false positives, negatives
When there are more than 2 groups, we can have k binary classifications and add them up for logloss. Some metrics like AUC is only applicable in the binary case.

Q406. What are various ways to predict a binary response variable? Can you compare two of them and tell me when one would be more appropriate? What's the difference between these? (SVM, Logistic Regression, Naive Bayes, Decision Tree, etc.)

Things to look at: N, P, linearly seperable?, features independent?, likely to

overfit?, speed, performance, memory usage
Logistic Regression:
features roughly linear, problem roughly linearly separable
robust to noise, use l1,l2 regularization for model selection, avoid overfitting
the output come as probabilities
efficient and the computation can be distributed
can be used as a baseline for other algorithms
(-) can hardly handle categorical features
SVM:
with a nonlinear kernel, can deal with problems that are not linearly separable
(-) slow to train, for most industry scale applications, not really efficient
Naive Bayes:
computationally efficient when P is large by alleviating the curse of dimensionality
works surprisingly well for some cases even if the condition doesn't hold
with word frequencies as features, the independence assumption can be seen reasonable. So the algorithm can be used in text categorization
(-) conditional independence of every other feature should be met
Tree Ensembles:
good for large N and large P, can deal with categorical features very well
non parametric, so no need to worry about outliers
GBT's work better but the parameters are harder to tune
RF works out of the box, but usually performs worse than GBT
Deep Learning:
works well for some classification tasks (e.g. image)
used to squeeze something out of the problem

Q407. What is regularization and where might it be helpful? What is an example of using regularization in a model?

Regularization is useful for reducing variance in the model, meaning avoiding overfitting . For example, we can use L1 regularization in Lasso regression to penalize large coefficients.

Q408. Why might it be preferable to include fewer predictors over many?

When we add irrelevant features, it increases model's tendency to overfit because those features introduce more noise. When two variables are correlated, they might be harder to interpret in case of regression, etc.
curse of dimensionality
adding random noise makes the model more complicated but useless
computational cost

Ask someone for more details.

Q409. Given training data on tweets and their retweets, how would you predict the number ofretweets of a given tweet after 7 days after only observing 2 days worth of data?

Build a time series model with the training data with a seven day cycle and then use that for a new data with only 2 days data.
Ask someone for more details.
Build a regression function to estimate the number of retweets as a function of time t
to determine if one regression function can be built, see if there are clusters in terms of the trends in the number of retweets
if not, we have to add features to the regression function
features + # of retweets on the first and the second day -> predict the seventh day
https://en.wikipedia.org/wiki/Dynamic_time_warping

Q410. How could you collect and analyze data to use social media to predict the weather?

We can collect social media data using twitter, Facebook, instagram API's. Then, for example, for twitter, we can construct features from each tweet, e.g. the tweeted date, number of favorites, retweets, and of course, the features created from the tweeted content itself. Then use a multi variate time series model to predict the weather.
Ask someone for more details. Get Data Science Training in Kalayan Nagar Bangalore.

Q411. How would you construct a feed to show relevant content for a site that involves userinteractions with items?

We can do so using building a recommendation engine. The easiest we can do is to show contents that are popular other users, which is still a valid strategy if for example the contents are news articles. To be more accurate, we can build a content based filtering or collaborative filtering. If there's enough user usage data, we can try collaborative filtering and recommend contents other similar users have consumed. If there isn't, we can recommend similar items based on vectorization of items (content based filtering).

Q412. How would you design the people you may know feature on LinkedIn or Facebook?

Find strong unconnected people in weighted connection graph
Define similarity as how strong the two people are connected
Given a certain feature, we can calculate the similarity based on
friend connections (neighbors)

Check-in’s people being at the same location all the time.
same college, workplace
Have randomly dropped graphs test the performance of the algorithm
ref. News Feed Optimization
Affinity score: how close the content creator and the users are
Weight: weight for the edge type (comment, like, tag, etc.). Emphasis on features the company wants to promote
Time decay: the older the less important

Q413. How would you predict who someone may want to send a Snapchat or Gmail to?

for each user, assign a score of how likely someone would send an email to
the rest is feature engineering:
number of past emails, how many responses, the last time they exchanged an email, whether the last email ends with a question mark, features about the other users, etc.
Ask someone for more details.
People who someone sent emails the most in the past, conditioning on time decay.

Q414. How would you suggest to a franchise where to open a new store?

build a master dataset with local demographic information available for each location.
local income levels, proximity to traffic, weather, population density, proximity to other businesses
a reference dataset on local, regional, and national macroeconomic conditions (e.g. unemployment, inflation, prime interest rate, etc.)
any data on the local franchise owner-operators, to the degree the manager
identify a set of KPIs acceptable to the management that had requested the analysis concerning the most desirable factors surrounding a franchise
quarterly operating profit, ROI, EVA, pay-down rate, etc.
run econometric models to understand the relative significance of each variable
run machine learning algorithms to predict the performance of each location candidate

Q415. In a search engine, given partial data on what the user has typed, how would you predict the user’s eventual search query?

Based on the past frequencies of words shown up given a sequence of words, we can construct conditional probabilities of the set of next sequences of words that can show up (n-gram). The sequences with highest conditional probabilities can show up as top candidates.

To further improve this algorithm,
we can put more weight on past sequences which showed up more recently and near your location to account for trends
show your recent searches given partial data

Q416. Given a database of all previous alumni donations to your university, how would you predict which recent alumni are most likely to donate?

Based on frequency and amount of donations, graduation year, major, etc, construct a supervised regression (or binary classification) algorithm.

Q417. You're Uber and you want to design a heatmap to recommend to drivers where to wait for a passenger. How would you approach this?

Based on the past pickup location of passengers around the same time of the day, day of the week (month, year), construct
Ask someone for more details.
Based on the number of past pickups
account for periodicity (seasonal, monthly, weekly, daily, hourly)
special events (concerts, festivals, etc.) from tweets

Q418. How would you build a model to predict a March Madness bracket?

One vector each for team A and B. Take the difference of the two vectors and use that as an input to predict the probability that team A would win by training the model. Train the models using past tournament data and make a prediction for the new tournament by running the trained model for each round of the tournament
Some extensions:
Experiment with different ways of consolidating the 2 team vectors into one (e.g concantenating, averaging, etc)
Consider using a RNN type model that looks at time series data.

Q419. You want to run a regression to predict the probability of a flight delay, but there are flights with delays of up to 12 hours that are really messing up your model. How can you address this?

This is equivalent to making the model more robust to outliers.

Probability

Q421. Bobo the amoeba has a 25%, 25%, and 50% chance of producing 0, 1, or 2 o spring, respectively. Each of Bobo's descendants also have the same probabilities. What is the probability that Bobo's lineage dies out?

p=1/4+1/4p+1/2p^2 => p=1/2

Q422. In any 15-minute interval, there is a 20% probability that you will see at least one shooting star. What is the proba- bility that you see at least one

shooting star in the period of an hour?
1-(0.8)^4. Or, we can use Poisson processes

Q424. How can you get a fair coin toss if someone hands you a coin that is weighted to come up heads more often than tails?
Flip twice and if HT then H, TH then T.

Q425. You have an 50-50 mixture of two normal distributions with the same standard deviation. How far apart do the means need to be in order for this distribution to be bimodal?
more than two standard deviations

Q426. Given draws from a normal distribution with known parameters, how can you simulate draws from a uniform distribution?
plug in the value to the CDF of the same random variable

Q427. A certain couple tells you that they have two children, at least one of which is a girl. What is the probability that they have two girls?
1/3

Q428. You have a group of couples that decide to have children until they have their first girl, afterwhich they stop having children. What is the expected gender ratio of the children that are born?What is the expected number of children each couple will have?
gender ratio is 1:1. Expected number of children is 2. let X be the number of children until getting a female (happens with prob 1/2). this follows a geometric distribution with probability 1/2

Q429. How many ways can you split 12 people into 3 teams of 4?
the outcome follows a multinomial distribution with n=12 and k=3. but the classes are indistinguishable

Q430. Your hash function assigns each object to a number between 1:10, each with equal probability. With 10 objects, what is the probability of a hash collision? What is the expected number of hash collisions? What is the expected number of hashes that are unused?
the probability of a hash collision: 1-(10!/10^10)
the expected number of hash collisions: 1-10*(9/10)^10
the expected number of hashes that are unused: 10*(9/10)^10

Q431. You call 2 UberX's and 3 Lyfts. If the time that each takes to reach you is IID, what is theprobability that all the Lyfts arrive first? What is the probability that all the UberX's arrive first?
Lyfts arrive first: 2!*3!/5!

Ubers arrive first: same

Q432. I write a program should print out all the numbers from 1 to 300, but prints out Fizz instead if the number is divisible by 3, Buzz instead if the number is divisible by 5, and FizzBuzz if the number is divisible by 3 and 5. What is the total number of numbers that is either Fizzed, Buzzed, or FizzBuzzed?

100+60-20=140

Q433. On a dating site, users can select 5 out of 24 adjectives to describe themselves. A match isdeclared between two users if they match on at least 4 adjectives. If Alice and Bob randomly pick adjectives, what is the probability that they form a match?

24C5*(1+5(24-5))/24C5*24C5 = 4/1771

Q434. A lazy high school senior types up application and envelopes to n different colleges, but puts the applications randomly into the envelopes. What is the expected number of applications that went to the right college?

1

Q435. Let's say you have a very tall father. On average, what would you expect the height of his son to be? Taller, equal, or shorter? What if you had a very short father?

Shorter. Regression to the mean

Q436. What's the expected number of coin flips until you get two heads in a row?

the expected number of coin flips until you get two tails in a row.

Q437. Let's say we play a game where I keep flipping a coin until I get heads. If the first time I get heads is on the nth coin, then I pay you 2n-1 dollars. How much would you pay me to play this game?

less than $3

Q438. You have two coins, one of which is fair and comes up heads with a probability 1/2, and the other which is biased and comes up heads with probability 3/4. You randomly pick coin and flip it twice, and get heads both times. What is the probability that you picked the fair coin?

4/13

Data Analysis

Q439. Let's say you're building the recommended music engine at Spotify to recommend peoplemusic based on past lis- tening history. How would you approach this problem?

collaborative filtering

Q440. What is R2? What are some other metrics that could be better than R2 and why?

goodness of fit measure. variance explained by the regression / total variance
the more predictors you add the higher R^2 becomes.
hence use adjusted R^2 which adjusts for the degrees of freedom
or train error metrics

Q441. What is the curse of dimensionality?

High dimensionality makes clustering hard, because having lots of dimensions means that everything is “far away” from each other.
For example, to cover a fraction of the volume of the data we need to capture a very wide range for each variable as the number of variables increases
All samples are close to the edge of the sample. And this is a bad news because prediction is much more difficult near the edges of the training sample.
The sampling density decreases exponentially as p increases and hence the data becomes much more sparse without significantly more data.
We should conduct PCA to reduce dimensionality

Q442. Is more data always better?

Statistically,
It depends on the quality of your data, for example, if your data is biased, just getting more data won’t help.
It depends on your model. If your model suffers from high bias, getting more data won’t improve your test results beyond a point. You’d need to add more features, etc.
Practically,
Also there’s a tradeoff between having more data and the additional storage, computational power, memory it requires. Hence, always think about the cost of having more data.

Q443. What are advantages of plotting your data before performing analysis?

Data sets have errors. You won’t find them all but you might find some. That 212 year old man. That 9 foot tall woman.
Variables can have skewness, outliers etc. Then the arithmetic mean might not be useful. Which means the standard deviation isn’t useful.
Variables can be multimodal! If a variable is multimodal then anything based on its mean or median is going to be suspect.

Q444. How can you make sure that you don’t analyze something that ends up meaningless?

Proper exploratory data analysis.
In every data analysis task, there's the exploratory phase where you're just graphing things, testing things on small sets of the data, summarizing simple statistics, and getting rough ideas of what hypotheses you might want to pursue further.

Then there's the exploitatory phase, where you look deeply into a set of hypotheses.

The exploratory phase will generate lots of possible hypotheses, and the exploitatory phase will let you really understand a few of them. Balance the two and you'll prevent yourself from wasting time on many things that end up meaningless, although not all.

Q445. What is the role of trial and error in data analysis? What is the role of making a hypothesisbefore diving in?

data analysis is a repetition of setting up a new hypothesis and trying to refute the null hypothesis.

The scientific method is eminently inductive: we elaborate a hypothesis, test it and refute it or not. As a result, we come up with new hypotheses which are in turn tested and so on. This is an iterative process, as science always is.

Q446. How can you determine which features are the most important in your model?

run the features though a Gradient Boosting Machine or Random Forest to generate plots of relative importance and information gain for each feature in the ensembles.

Look at the variables added in forward variable selection

Q447. How do you deal with some of your predictors being missing?

Remove rows with missing values – This works well if 1) the values are missing randomly (see Vinay Prabhu's answer for more details on this) 2) if you don't lose too much of the dataset after doing so.

Build another predictive model to predict the missing values – This could be a whole project in itself, so simple techniques are usually used here.

Use a model that can incorporate missing data – Like a random forest, or any tree-based method.

Q448. You have several variables that are positively correlated with your response, and you thinkcombining all of the variables could give you a good prediction of your response. However, you see that in the multiple linear regression, one of the weights on the predictors is negative. What could be the issue?

Multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression model are highly linearly related.
Leave the model as is, despite multicollinearity. The presence of multicollinearity doesn't affect the efficiency of extrapolating the fitted model to new data provided that the predictor variables follow the same pattern of multicollinearity in the new data as in the data on which the regression model is based.
principal component regression

Q449. Let's say you're given an unfeasible amount of predictors in a predictive modeling task. What are some ways to make the prediction more feasible?

PCA

Q450. Now you have a feasible amount of predictors, but you're fairly sure that you don't need all of them. How would you perform feature selection on the dataset?

ridge / lasso / elastic net regression
Univariate Feature Selection where a statistical test is applied to each feature individually. You retain only the best features according to the test outcome scores
"Recursive Feature Elimination":
First, train a model with all the feature and evaluate its performance on held out data.
Then drop let say the 10% weakest features (e.g. the feature with least absolute coefficients in a linear model) and retrain on the remaining features.
Iterate until you observe a sharp drop in the predictive accuracy of the model.

Q451. Your linear regression didn't run and communicates that there are an infinite number of best estimates for the regression coefficients. What could be wrong?

$p > n$.
If some of the explanatory variables are perfectly correlated (positively or negatively) then the coefficients would not be unique.

Q452. You run your regression on different subsets of your data, and that in each subset, the betavalue for a certain variable varies wildly. What could be the issue here?

The dataset might be heterogeneous. In which case, it is recommended to cluster datasets into different subsets wisely, and then draw different models for different subsets. Or, use models like non parametric models (trees) which can deal with heterogeneity quite nicely.

What is the main idea behind ensemble learning? If I had many different models that predicted the same response variable, what might I want to do to incorporate all of the models? Would you expect this to perform better than an individual model or worse?
The assumption is that a group of weak learners can be combined to form a strong learner.
Hence the combined model is expected to perform better than an individual model.
Assumptions:
average out biases
reduce variance
Bagging works because some underlying learning algorithms are unstable: slightly different inputs leads to very different outputs. If you can take advantage of this instability by running multiple instances, it can be shown that the reduced instability leads to lower error. If you want to understand why, the original bagging paper(http://www.springerlink.com/cont…) has a section called "why bagging works"
Boosting works because of the focus on better defining the "decision edge". By reweighting examples near the margin (the positive and negative examples) you get a reduced error (see http://citeseerx.ist.psu.edu/vie…)
Use the outputs of your models as inputs to a meta-model.
For example, if you're doing binary classification, you can use all the probability outputs of your individual models as inputs to a final logistic regression (or any model, really) that can combine the probability estimates.

One very important point is to make sure that the output of your models are out-of-sample predictions. This means that the predicted value for any row in your dataframe should NOT depend on the actual value for that row.

Q453. Given that you have wi data in your o ce, how would you determine which rooms and areasare underutilized and overutilized?

If the data is more used in one room, then that one is over utilized! Maybe account for the room capacity and normalize the data.

Q454. How would you quantify the influence of a Twitter user?

like page rank with each user corresponding to the web pages and linking to the page equivalent to following.

Q455. You have 100 mathletes and 100 math problems. Each mathlete gets to choose 10 problems to solve. Given data on who got what problem correct, how would you rank the problems in terms of difficulty?

One way you could do this is by storing a "skill level" for each user and a

“difficulty level” for each problem. We assume that the probability that a user solves a problem only depends on the skill of the user and the difficulty of the problem.* Then we maximize the likelihood of the data to find the hidden skill and difficulty levels.
The Rasch model for dichotomous data takes the form:
{\displaystyle \Pr\{X_{ni}=1\}={\frac {\exp({\beta _{n}}-{\delta _{i}})}{1+\exp({\beta _{n}}-{\delta _{i}})}},}
where is the ability of person and is the difficulty of item}.

Q456. You have 5000 people that rank 10 sushis in terms of salt- iness. How would you aggregate this data to estimate the true saltiness rank in each sushi?

Some people would take the mean rank of each sushi. If I wanted something simple, I would use the median, since ranks are (strictly speaking) ordinal and not interval, so adding them is a bit risque (but people do it all the time and you probably won’t be far wrong).

Q457. Given data on congressional bills and which congressio- nal representatives co-sponsored the bills, how would you determine which other representatives are most similar to yours in voting behavior? How would you evaluate who is the most liberal? Most republican? Most bipartisan?

collaborative filtering. you have your votes and we can calculate the similarity for each representatives and select the most similar representative
for liberal and republican parties, find the mean vector and find the representative closest to the center point

Q458. How would you come up with an algorithm to detect plagiarism in online content?

reduce the text to a more compact form (e.g. fingerprinting,
bag of wor
ds) then compare those with other texts by calculating the similarity

Q459. You have data on all purchases of customers at a grocery store. Describe to me how you would program an algorithm that would cluster the customers into groups. How would you determine the appropriate number of clusters include?

KNN
choose a small value of k that still has a low SSE (elbow method)
https://bl.ocks.org/rpgove/0060ff3b656618e9136b
Statistical Inference

Q460. In an A/B test, how can you check if assignment to the various buckets was truly random?

Plot the distributions of multiple features for both A and B and make sure that they have the same shape. More rigorously, we can conduct a permutation test to see if the distributions are the same.
MANOVA to compare different means

Q461. What might be the benefits of running an A/A test, where you have two buckets who areexposed to the exact same product?

Verify the sampling algorithm is random.

Q462. What would be the hazards of letting users sneak a peek at the other bucket in an A/B test?

The user might not act the same suppose had they not seen the other bucket. You are essentially adding additional variables of whether the user peeked the other bucket, which are not random across groups.

Q463. What would be some issues if blogs decide to cover one of your experimental groups?

Same as the previous question. The above problem can happen in larger scale.

Q464. How would you conduct an A/B test on an opt-in feature?

Ask someone for more details.

Q465. How would you run an A/B test for many variants, say 20 or more?

one control, 20 treatment, if the sample size for each group is big enough.
Ways to attempt to correct for this include changing your confidence level (e.g. Bonferroni Correction) or doing family-wide tests before you dive in to the individual metrics (e.g. Fisher's Protected LSD).

Q466. How would you run an A/B test if the observations are extremely right-skewed?

lower the variability by modifying the KPI
cap values
percentile metrics
log transform
https://www.quora.com/How-would-you-run-an-A-B-test-if-the-observations-are-extremely-right-skewed

Q467. I have two different experiments that both change the sign-up button to my website. I want to test them at the same time. What kinds of things should I keep in mind?

exclusive -> ok

Q468. What is a p-value? What is the difference between type-1 and type-2

error?

type-1 error: rejecting Ho when Ho is a true
type-2 error: not rejecting Ho when Ha is true
[toggle_content title="Q49. You are AirBnB and you want to test the hypothesis that a greater number of photographs increases the chances that a buyer selects the listing. How would you test this hypothesis?

For randomly selected listings with more than 1 pictures, hide 1 random picture for group A, and show all for group B. Compare the booking rate for the two groups.
Ask someone for more details.

Q469. How would you design an experiment to determine the impact of latency on userengagement?

The best way I know to quantify the impact of performance is to isolate just that factor using a slowdown experiment, i.e., add a delay in an A/B test.

Q470. What is maximum likelihood estimation? Could there be any case where it doesn't exist?

A method for parameter optimization (fitting a model). We choose parameters so as to maximize the likelihood function (how likely the outcome would happen given the current data and our model).
maximum likelihood estimation (MLE) is a method of estimating the parameters of a statistical model given observations, by finding the parameter values that maximize the likelihood of making the observations given the parameters. MLE can be seen as a special case of the maximum a posteriori estimation (MAP) that assumes a uniform prior distribution of the parameters, or as a variant of the MAP that ignores the prior and which therefore is unregularized.
for Gaussian mixtures, non-parametric models, it doesn't exist

Q471. What's the difference between a MAP, MOM, MLE estimator? In which cases would you want to use each?

MAP estimates the posterior distribution given the prior distribution and data which maximizes the likelihood function. MLE is a special case of MAP where the prior is uninformative uniform distribution.
MOM sets moment values and solves for the parameters. MOM has not used much anymore because maximum likelihood estimators have higher probability of being close to the quantities to be estimated and are more often unbiased.

Q472. What is a confidence interval and how do you interpret it?

For example, 95% confidence interval is an interval that when constructed for a set of samples each sampled in the same way, the constructed intervals include the true mean 95% of the time.

if confidence intervals are constructed using a given confidence level in an infinite number of independent experiments, the proportion of those intervals that contain the true value of the parameter will match the confidence level.

Q473. What is unbiasedness as a property of an estimator? Is this always a desirable property when performing inference? What about in data analysis or predictive modeling?

Unbiasedness means that the expectation of the estimator is equal to the population value we are estimating. This is desirable in inference because the goal is to explain the dataset as accurately as possible. However, this is not always desirable for data analysis or predictive modeling as there is the bias variance tradeoff. We sometimes want to prioritize the generalizability and avoid overfitting by reducing variance and thus increasing bias.

OTHER Important Data Science Interview Questions and Answers

Q474. What is the difference between population and sample in data?

Sample is the set of people who participated in your study whereas the population is the set of people to whom you want to generalize the results. For example – If you want to study the obesity among the children in India and you study 1000 children then those 1000 became sample whereas the all the children in the country is the population.

Sample is the subset of population.

Q475. What is the difference sample and sample frame?

Sample frame is the number of people who wanted to study whereas sample is the actual number of people who participated in your study. Ex – If you sent a marketing survey link to 300 people through email and only 100 participated in the survey then 300 is the sample survey and 100 is the sample.

Sample is the subset of sample frame. Both Sample and Sample Frame are subset of population.

Q476. What is the difference between univariate, bivariate and multivariate analysis?

Univariate analysis is performed on one variable, bivariate on two variable and multivariate analysis on two or more variables

Q477. What is difference between interpolation and extrapolation?

Extrapolation is the estimation of future values based on the observed trend on the past. Interpolation is the estimation of missing past values within two values in a sequence of values

Q478. What is precision and recall?

Precision is the percentage of correct predictions you have made and recall is the

percentage of predictions that actually turned out to be true

Q479. What is confusion matrix?

- Confusion matrix is a table which contains information about predicted values and actual values in a classification model
- It has four parts namely true positive ,true negative, false positive and false negative
- It can be used to calculate accuracy, precision and recall

Q480. What is hypothesis testing?

While performing the an experiment hypothesis testing to is used to analyze the various factors that are assumed to have an impact on the outcome of experiment

An hypothesis is some kind of assumption and hypothesis testing is used to determine whether the stated hypothesis is true or not

Initial assumption is called null hypothesis and the opposite alternate hypothesis

Q481. What is a p-value in statistics?

In hypothesis testing, p value helps to arrive at a conclusion. When p -value is too small then null hypothesis is rejected and alternate is accepted. When p-value is large then null hypothesis is accepted.

Q482. What is difference between Type-I error and Type-II error in hypothesis testing?

Type-I error is we reject the null hypothesis which was supposed to be accepted. It represents false positive

Type-II error represents we accept the null hypothesis which was supposed to be rejected. It represents false negative.

Q483. QWhat are the different types of missing value treatment?

- Deletion of values
- Guess the value
- Average Substitution
- Regression based substitution
- Multiple Imputation

Q484. What is gradient descent?

When building a statistical model the objective is reduce the value of the cost function that is associated with the model. Gradient descent is an iterative optimization technique used to determine the minima of the cost function

Q485. What is difference between supervised and unsupervised learning algorithms?

Supervised learning are the class of algorithms in which model is trained by explicitly labelling the outcome. Ex. Regression, Classification

Unsupervised learning no output is given and the algorithm is made to learn the outcomes implicity Ex. Association, Clustering

Q486. What is the need for regularization in model building?

Regularization is used to penalize the model when it overfits the model. It predominantly helps in solving the overfitting problem.

Q487. Difference between bias and variance tradeoff?

High Bias is an underlying error wrong assumption that makes the model to underfit. High Variance in a model means noise in data has been too taken seriously by the model which will result in overfitting.

Typically we would like to have a model with low bias and low variance

Q488. How to solve overfitting?

- Introduce Regularization
- Perform Cross Validation
- Reduce the number of features
- Increase the number of entries
- Ensembling

Q489. How will you detect the presence of overfitting?

When you build a model which has very high model accuracy on train data set and very low prediction accuracy in test data set then it is a indicator of overfitting

Q490. How do you determine the number of clusters in k-means clustering?

Elbow method (Plotting the percentage of variance explained w.r.t to number of clusters)
Gap Statistic
Silhouette method

Q491. What is the difference between causality and correlation?

Correlation is the measure that helps us understand the relationship between two or more variables
Causation represents that causal relationship between two events. It is also known to represent cause and effect
Causation means there is correlation but correlation doesn’t necessarily mean causation

Q492. Explain normal distribution?

Normal distribution is a bell shaped curve that represents distribution of data around its mean. Any normal process would follow the normal distribution. Most of data points tend to concentrated around the mean. If a point is further away from the mean then it is less likely to appear

Q493. What are the different ways of performing aggregation in python using pandas?

Group by function
Pivot function
Aggregate function

Q494. What are merge two list and get only unique values?

List a = [1,2,3,4] List b= [1,2,5,6] A = list(set(a+b))

Q495. How to save and retrieve model objects in python?

By using a library called pickle you can train any model and store the object in a pickle file.
When needed in future you can retrieve the object and use the model for prediction.
[toggle_content title="Q96. What is an anomaly and how is it different from outliers?

Anomaly detection is identification of items or events that didn't fit to the exact pattern or other items in a dataset. Outliers are valid data points that are outside the norm whereas anomaly are invalid data points that are created by process that is different from process that created the other data points

Q497. What is an ensemble learning?

Ensemble learning is the art of combining more than one model to predict the final outcome of an experiment. Commonly used ensemble techniques bagging, boosting and stacking

Q498. Name few libraries that is used in python for data analysis?

Numpy
Scipy
Pandas
Scikit learn
Matplotlib\ seaborn

Q499. What are the different types of data?

Data is broadly classified into two types 1) Numerical 2) Categorical
Numerical variables is further classified into discrete and continuous data
Categorical variables
Systematic Sampling
Stratified Sampling
Quota Sampling are further classified into Binary, Nominal and Ordinal data

Q500. What is a lambda function in python?

Lambda function are used to create small, one-time anonymous function in

python. It enables the programmer to create functions without a name and almost instantly