BASIC INTERVIEW Q'S ON ML

Note

This report is created based on various websites from different author point of view so please respect the author's knowledge, time and efforts and if you have any doubts regarding topics try google for the better answers.

Main motive to make this report is to make life easy for the all the 'Statistics, ML, DS, DL, NLP and Image processing ' learner's not to waste time on searching again and again for the interview questions.

Finally I am not expert in making this report or evaluating answers I just straight away taken from different websites to make one hand document for all the upcoming learners

But not the least, 'Use this document for knowledge sharing only'

Interview Questions on Machine Learning

https://www.analyticsvidhya.com/blog/2016/09/40-interview-questions-asked-at-startups-in-machine-learning-data-science/

Q1. You are given a train data set having 1000 columns and 1 million rows. The data set is based on a classification problem. Your manager has asked you to reduce the dimension of this data so that model computation time can be reduced. Your machine has memory constraints. What would you do? (You are free to make practical assumptions.)

Answer: Processing a high dimensional data on a limited memory machine is a strenuous task, your interviewer would be fully aware of that. Following are the methods you can use to tackle such situation:

1. Since we have lower RAM, we should close all other applications in our machine, including the web browser, so that most of the memory can be put to use.
2. We can randomly sample the data set. This means, we can create a smaller data set, let's say, having 1000 variables and 300000 rows and do the computations.
3. To reduce dimensionality, we can separate the numerical and categorical variables and remove the correlated variables. For numerical variables, we'll use correlation. For categorical variables, we'll use chi-square test.
4. Also, we can use PCA and pick the components which can explain the maximum variance in the data set.
5. Using online learning algorithms like Vowpal Wabbit (available in Python) is a possible option.
6. Building a linear model using Stochastic Gradient Descent is also helpful.
7. We can also apply our business understanding to estimate which all predictors can impact the response variable. But, this is an intuitive approach, failing to identify useful predictors might result in significant loss of information.

Note: For point 4 & 5, make sure you read about online learning algorithms & Stochastic Gradient Descent. These are advanced methods.

Q2. Is rotation necessary in PCA? If yes, Why? What will happen if you don't rotate the components?

Answer: Yes, rotation (orthogonal) is necessary because it maximizes the difference between variance captured by the component. This makes the components easier to interpret. Not to forget, that's the motive of doing PCA where, we aim to select fewer components (than features) which can explain the maximum variance in the data set. By doing rotation, the relative location of the components doesn't change, it only changes the actual coordinates of the points.
If we don't rotate the components, the effect of PCA will diminish and we'll have to select more number of components to explain variance in the data set.Know more: PCA

Q3. You are given a data set. The data set has missing values which spread along 1 standard deviation from the median. What percentage of data would remain unaffected? Why?

Answer: This question has enough hints for you to start thinking! Since, the data is spread across median, let's assume it's a normal distribution. We know, in a normal distribution, ~68% of the data lies in 1 standard deviation from mean (or mode, median), which leaves ~32% of the data unaffected. Therefore, ~32% of the data would remain unaffected by missing values.

Q4. You are given a data set on cancer detection. You've build a classification model and achieved an accuracy of 96%. Why shouldn't you be happy with your model performance? What can you do about it?

Answer: If you have worked on enough data sets, you should deduce that cancer detection results in imbalanced data. In an imbalanced data set, accuracy should not be used as a measure of performance because 96% (as given) might only be predicting majority class correctly, but our class of interest is minority class (4%) which is the people who actually got diagnosed with cancer. Hence, in order to evaluate model performance, we should use Sensitivity (True Positive Rate), Specificity (True Negative Rate), F measure to determine class wise performance of the classifier. If the minority class performance is found to to be poor, we can undertake the following steps:

1. We can use undersampling, oversampling or SMOTE to make the data balanced.
2. We can alter the prediction threshold value by doing probability caliberation and finding a optimal threshold using AUC-ROC curve.
3. We can assign weight to classes such that the minority classes gets larger weight.
4. We can also use anomaly detection.

Know more: Imbalanced Classification

Q5. Why is naive Bayes so 'naive'?

Answer: naive Bayes is so 'naive' because it assumes that all of the features in a data set are equally important and independent. As we know, these assumptions are rarely true in real world scenario.

Q6. Explain prior probability, likelihood and marginal likelihood in context of naiveBayes algorithm?
Answer: Prior probability is nothing but, the proportion of dependent (binary) variable in the data set. It is the closest guess you can make about a class, without any further information. For example: In a data set, the dependent variable is binary (1 and 0). The proportion of 1 (spam) is 70% and 0 (not spam) is 30%. Hence, we can estimate that there are 70% chances that any new email would be classified as spam.
Likelihood is the probability of classifying a given observation as 1 in presence of some other variable. For example: The probability that the word 'FREE' is used in previous spam message is likelihood. Marginal likelihood is, the probability that the word 'FREE' is used in any message.

Q7. You are working on a time series data set. You manager has asked you to build a high accuracy model. You start with the decision tree algorithm, since you know it works fairly well on all kinds of data. Later, you tried a time series regression model and got higher accuracy than decision tree model. Can this happen? Why?
Answer: Time series data is known to posses linearity. On the other hand, a decision tree algorithm is known to work best to detect non – linear interactions. The reason why decision tree failed to provide robust predictions because it couldn't map the linear relationship as good as a regression model did. Therefore, we learned that, a linear regression model can provide robust prediction given the data set satisfies its linearity assumptions.

Q8. You are assigned a new project which involves helping a food delivery company save more money. The problem is, company's delivery team aren't able to deliver food on time. As a result, their customers get unhappy. And, to keep them happy, they end up delivering food for free. Which machine learning algorithm can save them?
Answer: You might have started hopping through the list of ML algorithms in your mind. But, wait! Such questions are asked to test your machine learning fundamentals.
This is not a machine learning problem. This is a route optimization problem. A machine learning problem consist of three things:

1. There exists a pattern.
2. You cannot solve it mathematically (even by writing exponential equations).
3. You have data on it.

Always look for these three factors to decide if machine learning is a tool to solve a particular problem.

Q9. You came to know that your model is suffering from low bias and high variance. Which algorithm should you use to tackle it? Why?
Answer: Low bias occurs when the model's predicted values are near to actual values. In other words, the model becomes flexible enough to mimic the training data distribution. While it sounds like great achievement, but not to forget, a flexible model has no generalization capabilities. It means, when this model is tested on an unseen data, it gives disappointing results.
In such situations, we can use bagging algorithm (like random forest) to tackle high variance problem. Bagging algorithms divides a data set into subsets made with repeated randomized sampling. Then, these samples are used to generate a set of models using a single learning algorithm. Later, the model predictions are combined using voting (classification) or averaging (regression).
Also, to combat high variance, we can:

1. Use regularization technique, where higher model coefficients get penalized, hence lowering model complexity.
2. Use top n features from variable importance chart. May be, with all the variable in the data set, the algorithm is having difficulty in finding the meaningful signal.

Q10. You are given a data set. The data set contains many variables, some of which are highly correlated and you know about it. Your manager has asked you to run PCA. Would you remove correlated variables first? Why?
Answer: Chances are, you might be tempted to say No, but that would be incorrect. Discarding correlated variables have a substantial effect on PCA because, in presence of correlated variables, the variance explained by a particular component gets inflated.
For example: You have 3 variables in a data set, of which 2 are correlated. If you run PCA on this data set, the first principal component would exhibit twice the variance than it would exhibit with uncorrelated variables. Also, adding correlated variables lets PCA put more importance on those variable, which is misleading.

Q11. After spending several hours, you are now anxious to build a high accuracy model. As a result, you build 5 GBM models, thinking a boosting algorithm would do the magic. Unfortunately, neither of models could perform better than benchmark score. Finally, you decided to combine those models. Though, ensembled models are known to return high accuracy, but you are unfortunate. Where did you miss?

Answer: As we know, ensemble learners are based on the idea of combining weak learners to create strong learners. But, these learners provide superior result when the combined models are uncorrelated. Since, we have used 5 GBM models and got no accuracy improvement, suggests that the models are correlated. The problem with correlated models is, all the models provide same information.
For example: If model 1 has classified User1122 as 1, there are high chances model 2 and model 3 would have done the same, even if its actual value is 0. Therefore, ensemble learners are built on the premise of combining weak uncorrelated models to obtain better predictions.

Q12. How is kNN different from kmeans clustering?
Answer: Don't get mislead by 'k' in their names. You should know that the fundamental difference between both these algorithms is, kmeans is unsupervised in nature and kNN is supervised in nature. kmeans is a clustering algorithm. kNN is a classification (or regression) algorithm.
kmeans algorithm partitions a data set into clusters such that a cluster formed is homogeneous and the points in each cluster are close to each other. The algorithm tries to maintain enough separability between these clusters. Due to unsupervised nature, the clusters have no labels.
kNN algorithm tries to classify an unlabeled observation based on its k (can be any number) surrounding neighbors. It is also known as lazy learner because it involves minimal training of model. Hence, it doesn't use training data to make generalization on unseen data set.

Q13. How is True Positive Rate and Recall related? Write the equation.
Answer: True Positive Rate = Recall. Yes, they are equal having the formula (TP/TP + FN).
Know more: Evaluation Metrics

Q14. You have built a multiple regression model. Your model R^2 isn't as good as you wanted. For improvement, your remove the intercept term, your model R^2 becomes 0.8 from 0.3. Is it possible? How?
Answer: Yes, it is possible. We need to understand the significance of intercept term in a regression model. The intercept term shows model prediction without any independent variable i.e. mean prediction. The formula of $R^2 = 1 – \sum(y – y´)^2/\sum(y – ymean)^2$ where $y´$ is predicted value.
When intercept term is present, R^2 value evaluates your model wrt. to the mean model. In absence of intercept term (`ymean`), the model can make no such evaluation, with large denominator, $\sum(y - y´)^2/\sum(y)^2$ equation's value becomes smaller than actual, resulting in higher R^2.

Q15. After analyzing the model, your manager has informed that your regression model is suffering from multicollinearity. How would you check if he's true? Without losing any information, can you still build a better model?
Answer: To check multicollinearity, we can create a correlation matrix to identify & remove variables having correlation above 75% (deciding a threshold is subjective). In addition, we can use calculate VIF (variance inflation factor) to check the presence of multicollinearity. VIF value <= 4 suggests no multicollinearity whereas a value of >= 10 implies serious multicollinearity. Also, we can use tolerance as an indicator of multicollinearity.
But, removing correlated variables might lead to loss of information. In order to retain those variables, we can use penalized regression models like ridge or lasso regression. Also, we can add some random noise in correlated variable so that the variables become different from each other. But, adding noise might affect the prediction accuracy, hence this approach should be carefully used.
Know more: Regression

Q16. When is Ridge regression favorable over Lasso regression?
Answer: You can quote ISLR's authors Hastie, Tibshirani who asserted that, in presence of few variables with medium / large sized effect, use lasso regression. In presence of many variables with small / medium sized effect, use ridge regression.
Conceptually, we can say, lasso regression (L1) does both variable selection and parameter shrinkage, whereas Ridge regression only does parameter shrinkage and end up including all the coefficients in the model. In presence of correlated variables, ridge regression might be the preferred choice. Also, ridge regression works best in situations where the least square estimates have higher variance. Therefore, it depends on our model objective.
Know more: Ridge and Lasso Regression

Q17. Rise in global average temperature led to decrease in number of pirates around the world. Does that mean that decrease in number of pirates caused the climate change?
Answer: After reading this question, you should have understood that this is a classic case of "causation and correlation". No, we can't conclude that decrease in number of pirates caused the climate change because there might be other factors (lurking or confounding variables) influencing this phenomenon.
Therefore, there might be a correlation between global average temperature and number of pirates, but based on this information we can't say that pirated died because of rise in global average temperature.

Know more: Causation and Correlation

Q18. While working on a data set, how do you select important variables? Explain your methods.
Answer: Following are the methods of variable selection you can use:

1. Remove the correlated variables prior to selecting important variables
2. Use linear regression and select variables based on p values
3. Use Forward Selection, Backward Selection, Stepwise Selection
4. Use Random Forest, Xgboost and plot variable importance chart
5. Use Lasso Regression
6. Measure information gain for the available set of features and select top n features accordingly.

Q19. What is the difference between covariance and correlation?
Answer: Correlation is the standardized form of covariance.
Covariances are difficult to compare. For example: if we calculate the covariances of salary ($) and age (years), we'll get different covariances which can't be compared because of having unequal scales. To combat such situation, we calculate correlation to get a value between -1 and 1, irrespective of their respective scale.

Q20. Is it possible capture the correlation between continuous and categorical variable? If yes, how?
Answer: Yes, we can use ANCOVA (analysis of covariance) technique to capture association between continuous and categorical variables.

Q21. Both being tree based algorithm, how is random forest different from Gradient boosting algorithm (GBM)?
Answer: The fundamental difference is, random forest uses bagging technique to make predictions. GBM uses boosting techniques to make predictions.
In bagging technique, a data set is divided into n samples using randomized sampling. Then, using a single learning algorithm a model is build on all samples. Later, the resultant predictions are combined using voting or averaging. Bagging is done is parallel. In boosting, after the first round of predictions, the algorithm weighs misclassified predictions higher, such that they can be corrected in the succeeding round. This sequential process of giving higher weights to misclassified predictions continue until a stopping criterion is reached.
Random forest improves model accuracy by reducing variance (mainly). The trees grown are uncorrelated to maximize the decrease in variance. On the other hand, GBM improves accuracy my reducing both bias and variance in a model.
Know more: Tree based modeling

Q22. Running a binary classification tree algorithm is the easy part. Do you know how does a tree splitting takes place i.e. how does the tree decide which variable to split at the root node and succeeding nodes?
Answer: A classification trees makes decision based on Gini Index and Node Entropy. In simple words, the tree algorithm find the best possible feature which can divide the data set into purest possible children nodes.
Gini index says, if we select two items from a population at random then they must be of same class and probability for this is 1 if population is pure. We can calculate Gini as following:

1. Calculate Gini for sub-nodes, using formula sum of square of probability for success and failure (p^2+q^2).
2. Calculate Gini for split using weighted Gini score of each node of that split

Entropy is the measure of impurity as given by (for binary class):

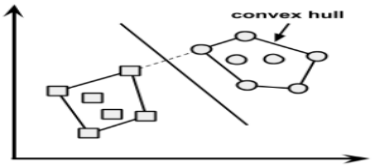
$$\text{Entropy} = -p \log_2 p - q \log_2 q$$

Here p and q is probability of success and failure respectively in that node. Entropy is zero when a node is homogeneous. It is maximum when a both the classes are present in a node at 50% – 50%. Lower entropy is desirable.

Q23. You've built a random forest model with 10000 trees. You got delighted after getting training error as 0.00. But, the validation error is 34.23. What is going on? Haven't you trained your model perfectly?
Answer: The model has overfitted. Training error 0.00 means the classifier has mimiced the training data patterns to an extent, that they are not available in the unseen data. Hence, when this classifier was run on unseen sample, it couldn't find those patterns and returned prediction with higher error. In random forest, it happens when we use larger number of trees than necessary. Hence, to avoid these situation, we should tune number of trees using cross validation.

Q24. You've got a data set to work having p (no. of variable) > n (no. of observation). Why is OLS as bad option to work with? Which techniques would be best to use? Why?

Answer: In such high dimensional data sets, we can't use classical regression techniques, since their assumptions tend to fail. When p > n, we can no longer calculate a unique least square coefficient estimate, the variances become infinite, so OLS cannot be used at all.
To combat this situation, we can use penalized regression methods like lasso, LARS, ridge which can shrink the coefficients to reduce variance. Precisely, ridge regression works best in situations where the least square estimates have higher variance.
Among other methods include subset regression, forward stepwise regression.



Q25. What is convex hull ? (Hint: Think SVM)
Answer: In case of linearly separable data, convex hull represents the outer boundaries of the two group of data points. Once convex hull is created, we get maximum margin hyperplane (MMH) as a perpendicular bisector between two convex hulls. MMH is the line which attempts to create greatest separation between two groups.
Q26. We know that one hot encoding increasing the dimensionality of a data set. But, label encoding doesn't. How Answer: Don't get baffled at this question. It's a simple question asking the difference between the two.sing one hot encoding, the dimensionality (a.k.a features) in a data set get increased because it creates a new variable for each level present in categorical variables. For example: let's say we have a variable 'color'. The variable has 3 levels namely Red, Blue and Green. One hot encoding 'color' variable will generate three new variables as `Color.Red`, `Color.Blue` and `Color.Green` containing 0 and 1 value.In label encoding, the levels of a categorical variables gets encoded as 0 and 1, so no new variable is created. Label encoding is majorly used for binary variables.
Q27. What cross validation technique would you use on time series data set? Is it k-fold or LOOCV?
Answer: Neither.
In time series problem, k fold can be troublesome because there might be some pattern in year 4 or 5 which is not in year 3. Resampling the data set will separate these trends, and we might end up validation on past years, which is incorrect. Instead, we can use forward chaining strategy with 5 fold as shown below:

- fold 1 : training [1], test [2]
- fold 2 : training [1 2], test [3]
- fold 3 : training [1 2 3], test [4]
- fold 4 : training [1 2 3 4], test [5]
- fold 5 : training [1 2 3 4 5], test [6]

where 1,2,3,4,5,6 represents "year".

Q28. You are given a data set consisting of variables having more than 30% missing values? Let's say, out of 50 variables, 8 variables have missing values higher than 30%. How will you deal with them?
Answer: We can deal with them in the following ways:

1. Assign a unique category to missing values, who knows the missing values might decipher some trend
2. We can remove them blatantly.
3. Or, we can sensibly check their distribution with the target variable, and if found any pattern we'll keep those missing values and assign them a new category while removing others.

29. 'People who bought this, also bought…' recommendations seen on amazon is a result of which algorithm?
Answer: The basic idea for this kind of recommendation engine comes from collaborative filtering.
Collaborative Filtering algorithm considers "User Behavior" for recommending items. They exploit behavior of other users and items in terms of transaction history, ratings, selection and purchase information. Other users behaviour and preferences over the items are used to recommend items to the new users. In this case, features of the items are not known.
Know more: Recommender System

Q30. What do you understand by Type I vs Type II error ?
Answer: Type I error is committed when the null hypothesis is true and we reject it, also known as a 'False Positive'. Type II error is committed when the null hypothesis is false and we accept it, also known as 'False Negative'.
In the context of confusion matrix, we can say Type I error occurs when we classify a value as positive (1) when it is actually negative (0). Type II error occurs when we classify a value as negative (0) when it is actually positive(1).

Q31. You are working on a classification problem. For validation purposes, you've randomly sampled the training data set into train and validation. You are confident that your model will work incredibly well on unseen data since your validation accuracy is high. However, you get shocked after getting poor test accuracy. What went wrong?
Answer: In case of classification problem, we should always use stratified sampling instead of random sampling. A random sampling doesn't takes into consideration the proportion of target classes. On the contrary, stratified sampling helps to maintain the distribution of target variable in the resultant distributed samples also.

Q32. You have been asked to evaluate a regression model based on R^2, adjusted R^2 and tolerance. What will be your criteria?
Answer: Tolerance (1 / VIF) is used as an indicator of multicollinearity. It is an indicator of percent of variance in a predictor which cannot be accounted by other predictors. Large values of tolerance is desirable.
We will consider adjusted R^2 as opposed to R^2 to evaluate model fit because R^2 increases irrespective of improvement in prediction accuracy as we add more variables. But, adjusted R^2 would only increase if an additional variable improves the accuracy of model, otherwise stays same. It is difficult to commit a general threshold value for adjusted R^2 because it varies between data sets. For example: a gene mutation data set might result in lower adjusted R^2 and still provide fairly good predictions, as compared to a stock market data where lower adjusted R^2 implies that model is not good.

Q33. In k-means or kNN, we use euclidean distance to calculate the distance between nearest neighbors. Why not manhattan distance ?
Answer: We don't use manhattan distance because it calculates distance horizontally or vertically only. It has dimension restrictions. On the other hand, euclidean metric can be used in any space to calculate distance. Since, the data points can be present in any dimension, euclidean distance is a more viable option.
Example: Think of a chess board, the movement made by a bishop or a rook is calculated by manhattan distance because of their respective vertical & horizontal movements.

Q34. Explain machine learning to me like a 5 year old.
Answer: It's simple. It's just like how babies learn to walk. Every time they fall down, they learn (unconsciously) & realize that their legs should be straight and not in a bend position. The next time they fall down, they feel pain. They cry. But, they learn 'not to stand like that again'. In order to avoid that pain, they try harder. To succeed, they even seek support from the door or wall or anything near them, which helps them stand firm.
This is how a machine works & develops intuition from its environment.
Note: The interview is only trying to test if have the ability of explain complex concepts in simple terms.

Q35. I know that a linear regression model is generally evaluated using Adjusted R^2 or F value. How would you evaluate a logistic regression model?
Answer: We can use the following methods:

1. Since logistic regression is used to predict probabilities, we can use AUC-ROC curve along with confusion matrix to determine its performance.
2. Also, 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.
3. Null Deviance indicates the response predicted by a model with nothing but an intercept. Lower the value, better the model. Residual deviance indicates the response predicted by a model on adding independent variables. Lower the value, better the model.

Know more: Logistic Regression

Q36. Considering the long list of machine learning algorithm, given a data set, how do you decide which one to use?
Answer: You should say, the choice of machine learning algorithm solely depends of the type of data. If you are given a data set which is exhibits linearity, then linear regression would be the best algorithm to use. If you given to work on images, audios, then neural network would help you to build a robust model.
If the data comprises of non linear interactions, then a boosting or bagging algorithm should be the choice. If the business requirement is to build a model which can be deployed, then we'll use regression or a decision tree model (easy to interpret and explain) instead of black box algorithms like SVM, GBM etc.
In short, there is no one master algorithm for all situations. We must be scrupulous enough to understand which algorithm to use.

Q37. Do you suggest that treating a categorical variable as continuous variable would result in a better predictive model?

Answer: For better predictions, categorical variable can be considered as a continuous variable only when the variable is ordinal in nature.

Q38. When does regularization becomes necessary in Machine Learning?
Answer: Regularization becomes necessary when the model begins to ovefit / underfit. This technique introduces a cost term for bringing in more features with the objective function. Hence, it tries to push the coefficients for many variables to zero and hence reduce cost term. This helps to reduce model complexity so that the model can become better at predicting (generalizing).

Q39. What do you understand by Bias Variance trade off?
Answer: The error emerging from any model can be broken down into three components mathematically. Following are these component :

$$Err(x) = \left(E[\hat{f}(x)] - f(x)\right)^2 + E\left[\hat{f}(x) - E[\hat{f}(x)]\right]^2 + \sigma_e^2$$

$$Err(x) = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

Bias error is useful to quantify how much on an average are the predicted values different from the actual value. A high bias error means we have a under-performing model which keeps on missing important trends. **Variance**on the other side quantifies how are the prediction made on same observation different from each other. A high variance model will over-fit on your training population and perform badly on any observation beyond training.
Q40. OLS is to linear regression. Maximum likelihood is to logistic regression. Explain the statement.
Answer: OLS and Maximum likelihood are the methods used by the respective regression methods to approximate the unknown parameter (coefficient) value. In simple words,
Ordinary least square(OLS) is a method used in linear regression which approximates the parameters resulting in minimum distance between actual and predicted values. Maximum Likelihood helps in choosing the the values of parameters which maximizes the likelihood that the parameters are most likely to produce observed data.

Probability

https://www.analyticsvidhya.com/blog/2017/04/40-questions-on-probability-for-all-aspiring-data-scientists/
Useful Resources
Basics of Probability for Data Science explained with examples
Introduction to Conditional Probability and Bayes theorem for data science professionals

1) Let A and B be events on the same sample space, with P (A) = 0.6 and P (B) = 0.7. Can these two events be disjoint?
(B) No
Solution: **(B)**
These two events cannot be disjoint because P(A)+P(B) >1.
P(AUB) = P(A)+P(B)-P(A∩B).
An event is disjoint if P(A∩B) = 0. If A and B are disjoint P(AUB) = 0.6+0.7 = 1.3
And Since probability cannot be greater than 1, these two mentioned events cannot be disjoint.
2) Alice has 2 kids and one of them is a girl. What is the probability that the other child is also a girl? You can assume that there are an equal number of males and females in the world.
C) 0.333
Solution: **(C)**
The outcomes for two kids can be {BB, BG, GB, GG}
Since it is mentioned that one of them is a girl, we can remove the BB option from the sample space. Therefore the sample space has 3 options while only one fits the second condition. Therefore the probability the second child will be a girl too is 1/3.

3) A fair six-sided die is rolled twice. What is the probability of getting 2 on the first roll and not getting 4 on the second roll?
C) 5/36
Solution: **(C)**
The two events mentioned are independent. The first roll of the die is independent of the second roll. Therefore the probabilities can be directly multiplied.
P(getting first 2) = 1/6
P(no second 4) = 5/6
Therefore P(getting first 2 and no second 4) = 1/6* 5/6 = 5/36
4) $P(A \cup B \cup C) = P(A \cap C^c) + P(C) + P(B \cap A^c \cap C^c)$

A) True Solution: **(A)**
P(A∩Cc) will be only P(A). P(only A)+P(C) will make it P(AUC). P(B∩Ac∩Cc) is P(only B) Therefore P(AUC) and P(only B) will make P(AUBUC)

5) Consider a tetrahedral die and roll it twice. What is the probability that the number on the first roll is strictly higher than the number on the second roll?
Note: A tetrahedral die has only four sides (1, 2, 3 and 4).
B) 3/8
Solution: **(B)**

(1,1) (2,1) (3,1) (4,1)

(1,2) (2,2) (3,2) (4,2)

(1,3) (2,3) (3,3) (4,3)

(1,4) (2,4) (3,4) (4,4)

There are 6 out of 16 possibilities where the first roll is strictly higher than the second roll.

6) Which of the following options cannot be the probability of any event?
A)-0.00001
B)0.5
C) 1.001
F) A and C
Solution: **(F)**
Probability always lie within 0 to 1.

7) Anita randomly picks 4 cards from a deck of 52-cards and places them back into the deck (Any set of 4 cards is equally likely). Then, Babita randomly chooses 8 cards out of the same deck (Any set of 8 cards is equally likely). Assume that the choice of 4 cards by Anita and the choice of 8 cards by Babita are independent. What is the probability that all 4 cards chosen by Anita are in the set of 8 cards chosen by Babita?
A)$^{48}C_4$ x $^{52}C_4$
B)$^{48}C_4$ x $^{52}C_8$
C)$^{48}C_8$ x $^{52}C_8$
D) None of the above
Solution: **(A)**
The total number of possible combination would be 52C4 (For selecting 4 cards by Anita) * 52C8 (For selecting 8 cards by Babita).
Since, the 4 cards that Anita chooses is among the 8 cards which Babita has chosen, thus the number of combinations possible is 52C4 (For selecting the 4 cards selected by Anita) * 48C4 (For selecting any other 4 cards by Babita, since the 4 cards selected by Anita are common)

Question Context 8:
A player is randomly dealt a sequence of 13 cards from a deck of 52-cards. All sequences of 13 cards are equally likely. In an equivalent model, the cards are chosen and dealt one at a time. When choosing a card, the dealer is equally likely to pick any of the cards that remain in the deck.
8) If you dealt 13 cards, what is the probability that the 13th card is a King?
B) 1/13
Solution: **(B)**
Since we are not told anything about the first 12 cards that are dealt, the probability that the 13th card dealt is a King, is the same as the probability that the first card dealt, or in fact any particular card dealt is a King, and this equals: 4/52

9) A fair six-sided die is rolled 6 times. What is the probability of getting all outcomes as unique?
A) 0.01543Solution: **(A)**
For all the outcomes to be unique, we have 6 choices for the first turn, 5 for the second turn, 4 for the third turn and so on
Therefore the probability if getting all unique outcomes will be equal to 0.01543

10) A group of 60 students is randomly split into 3 classes of equal size. All partitions are equally likely. Jack and Jill are two students belonging to that group. What is the probability that Jack and Jill will end up in the same class?

B) 19/59
Solution: **(B)**
Assign a different number to each student from 1 to 60. Numbers 1 to 20 go in group 1, 21 to 40 go to group 2, 41 to 60 go to group 3.
All possible partitions are obtained with equal probability by a random assignment if these numbers, it doesn't matter with which students we start, so we are free to start by assigning a random number to Jack and then we assign a random number to Jill. After Jack has been assigned a random number there are 59 random numbers available for Jill and 19 of these will put her in the same group as Jack. Therefore the probability is 19/59

11) We have two coins, A and B. For each toss of coin A, the probability of getting head is 1/2 and for each toss of coin B, the probability of getting Heads is 1/3. All tosses of the same coin are independent. We select a coin at random and toss it till we get a head. The probability of selecting coin A is ¼ and coin B is 3/4. What is the expected number of tosses to get the first heads?

A) 2.75
Solution: **(A)**
If coin A is selected then the number of times the coin would be tossed for a guaranteed Heads is 2, similarly, for coin B it is 3. Thus the number of times would be
Tosses = 2 * (1/4)[probability of selecting coin A] + 3*(3/4)[probability of selecting coin B]
= 2.75

12) Suppose a life insurance company sells a $240,000 one year term life insurance policy to a 25-year old female for $210. The probability that the female survives the year is .999592. Find the expected value of this policy for the insurance company.

C) $112
Solution: **(C)**
P(company loses the money) = 0.99592
P(company does not lose the money) = 0.000408
The amount of money company loses if it loses = 240,000 – 210 = 239790
While the money it gains is $210
Expected money the company will have to give = 239790*0.000408 = 97.8
Expect money company gets = 210.
Therefore the value = 210 – 98 = $112

$$P(A \cap B \cap C^c) = P(A)\, P(C^c \cap A \mid A)\, P(B \mid A \cap C^c)$$

13)
A) True
Solution: **(A)**
The above statement is true. You would need to know that
P(A/B) = P(A∩B)/P(B)
P(Cc∩A|A) = P(Cc∩A∩A)/P(A) = P(Cc∩A)/P(A)
P(B|A ∩ Cc) = P(A∩B∩Cc)/P(A ∩ Cc)
Multiplying the three we would get – P(A∩B∩Cc), hence the equations holds true

14) When an event A independent of itself?

D) If and only if P(A)=0 or 1
Solution: **(D)**
The event can only be independent of itself when either there is no chance of it happening or when it is certain to happen. Event A and B is independent when P(A∩B) = P(A)*P(B). Now if B=A, P(A∩A) = P(A) when P(A) = 0 or 1.

15) Suppose you're in the final round of "Let's make a deal" game show and you are supposed to choose from three doors – 1, 2 & 3. One of the three doors has a car behind it and other two doors have goats. Let's say you choose Door 1 and the host opens Door 3 which has a goat behind it. To assure the probability of your win, which of the following options would you choose.

A) Switch your choice
Solution: **(A)**
I would recommend reading this article for a detailed discussion of the Monty Hall's Problem.

16) Cross-fertilizing a red and a white flower produces red flowers 25% of the time. Now we cross-fertilize five pairs of red and white flowers and produce five offspring. What is the probability that there are no red flower plants in the five offspring?

A) 23.7%
Solution: **(A)**
The probability of offspring being Red is 0.25, thus the probability of the offspring not being red is 0.75. Since all the pairs are independent of each other, the probability that all the offsprings are not red would be (0.75)5 = 0.237. You can think of this as a binomial with all failures.

17) A roulette wheel has 38 slots – 18 red, 18 black, and 2 green. You play five games and always bet on red slots. How many games can you expect to win?
B) 2.3684C) 2.6316
Solution: **(B)**
The probability that it would be Red in any spin is 18/38. Now, you are playing the game 5 times and all the games are independent of each other. Thus, the number of games that you can win would be 5*(18/38) = 2.3684

18) A roulette wheel has 38 slots, 18 are red, 18 are black, and 2 are green. You play five games and always bet on red. What is the probability that you win all the 5 games?
B) 0.0238

Solution: **(B)**
The probability that it would be Red in any spin is 18/38. Now, you are playing for game 5 times and all the games are independent of each other. Thus, the probability that you win all the games is (18/38)5 = 0.0238

19) Some test scores follow a normal distribution with a mean of 18 and a standard deviation of 6. What proportion of test takers have scored between 18 and 24?
C) 34%
Solution: **(C)**
So here we would need to calculate the Z scores for value being 18 and 24. We can easily doing that by putting sample mean as 18 and population mean as 18 with σ = 6 and calculating Z. Similarly we can calculate Z for sample mean as 24.
Z= (X-μ)/σ
Therefore for 26 as X,
Z = (18-18)/6 = 0 , looking at the Z table we find 50% people have scores below 18.
For 24 as X
Z = (24-18)/6 = 1, looking at the Z table we find 84% people have scores below 24.
Therefore around 34% people have scores between 18 and 24.

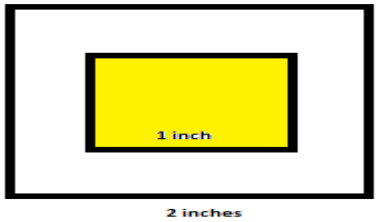
20) A jar contains 4 marbles. 3 Red & 1 white. Two marbles are drawn with replacement after each draw. What is the probability that the same color marble is drawn twice?
C) 5/8
Solution: **(C)**
If the marbles are of the same color then it will be 3/4 * 3/4 + 1/4 * 1/4 = 5/8.

21) Which of the following events is most likely?
A) At least one 6, when 6 dice are rolled
Solution: **(A)**
Probability of '6' turning up in a roll of dice is P(6) = (1/6) & P(6') = (5/6). Thus, probability of
∞ Case 1: (1/6) * (5/6)5 = 0.06698
∞ Case 2: (1/6)2 * (5/6)10 = 0.00448
∞ Case 3: (1/6)3 * (5/6)15 = 0.0003
Thus, the highest probability is Case 1

22) Suppose you were interviewed for a technical role. 50% of the people who sat for the first interview received the call for second interview. 95% of the people who got a call for second interview felt good about their first interview. 75% of people who did not receive a second call, also felt good about their first interview. If you felt good after your first interview, what is the probability that you will receive a second interview call?
B) 56%
Solution: **(B)**
Let's assume there are 100 people that gave the first round of interview. The 50 people got the interview call for the second round. Out of this 95 % felt good about their interview, which is 47.5. 50 people did not get a call for the interview; out of which 75% felt good about, which is 37.5. Thus, the total number of people that felt good after giving their interview is (37.5 + 47.5) 85. Thus, out of 85 people who felt good, only 47.5 got the call for next round. Hence, the probability of success is (47.5/85) = 0.558.

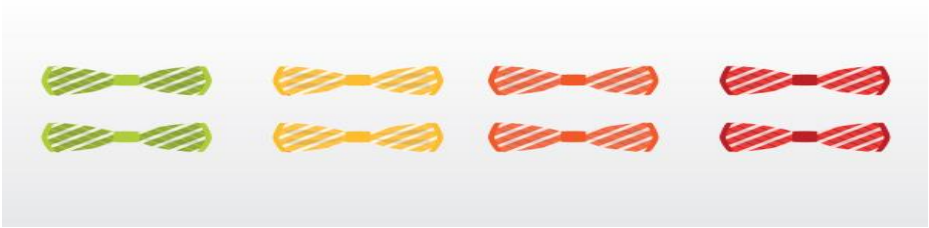
Another more accepted way to solve this problem is the Baye's theorem. I leave it to you to check for yourself.

23) A coin of diameter 1-inches is thrown on a table covered with a grid of lines each two inches apart. What is the probability that the coin lands inside a square without touching any of the lines of the grid? You can assume that the person throwing has no skill in throwing the coin and is throwing it randomly.
You can assume that the person throwing has no skill in throwing the coin and is throwing it randomly.
B) 1/4
Solution: **(B)**
Think about where all the center of the coin can be when it lands on 2 inches grid and it not touching the lines of the grid.



If the yellow region is a 1 inch square and the outside square is of 2 inches. If the center falls in the yellow region, the coin will not touch the grid line. Since the total area is 4 and the area of the yellow region is 1, the probability is ¼ .

24) There are a total of 8 bows of 2 each of green, yellow, orange & red. In how many ways can you select 1 bow?



C) 4
Solution: **(C)**
You can select one bow out of four different bows, so you can select one bow in four different ways.

25) Consider the following probability density function: What is the probability for X≤6 i.e. P(x≤6)

$$f(x) = \frac{1}{8} e^{-x/8} \text{ for } x>=0$$

What is the probability for X≤6 i.e. P(x≤6)
B) 0.5276
Solution: **(B)**
To calculate the area of a particular region of a probability density function, we need to integrate the function under the bounds of the values for which we need to calculate the probability.
Therefore on integrating the given function from 0 to 6, we get 0.5276

26) In a class of 30 students, approximately what is the probability that two of the students have their birthday on the same day (defined by same day and month) (assuming it's not a leap year)?
For example – Students with birthday 3rd Jan 1993 and 3rd Jan 1994 would be a favorable event.
C) 70%
Solution: **(C)**
The total number of combinations possible for no two persons to have the same birthday in a class of 30 is 30 * (30-1)/2 = 435.
Now, there are 365 days in a year (assuming it's not a leap year). Thus, the probability of people having a different birthday would be 364/365. Now there are 870 combinations possible. Thus, the probability that no two people have the same birthday is (364/365)^435 = 0.303.
Thus, the probability that two people would have their birthdays on the same date would be 1 – 0.303 = 0.696

27) Ahmed is playing a lottery game where he must pick 2 numbers from 0 to 9 followed by an English alphabet (from 26-letters). He may choose the same number both times.

If his ticket matches the 2 numbers and 1 letter drawn in order, he wins the grand prize and receives $10405. If just his letter matches but one or both of the numbers do not match, he wins $100. Under any other circumstance, he wins nothing. The game costs him $5 to play. Suppose he has chosen 04R to play.
What is the expected net profit from playing this ticket?
B) $2.81C) $-1.82
Solution: **(B)**
Expected value in this case
E(X) = P(grand prize)*(10405-5)+P(small)(100-5)+P(losing)*(-5)
P(grand prize)= (1/10)*(1/10)*(1/26)
P(small) = 1/26-1/2600, the reason we need to do this is we need to exclude the case where he gets the letter right and also the numbers rights. Hence, we need to remove the scenario of getting the letter right.
P(losing) = 1-1/26-1/2600
Therefore we can fit in the values to get the expected value as $2.81

28) Assume you sell sandwiches. 70% people choose egg**, and the rest choose chicken. What is the probability of selling 2 egg sandwiches to the next 3 customers?**

C) 0.147

Solution: **(C)**
The probability of selling Egg sandwich is 0.7 & that of a chicken sandwich is 0.3. Now, the probability that next 3 customers would order 2 egg sandwich is 0.7 * 0.7 *0.3 = 0.147. They can order them in any sequence, the probabilities would still be the same.

Question context: 29 – 30
HIV is still a very scary disease to even get tested for. The US military tests its recruits for HIV when they are recruited. They are tested on three rounds of Elisa(an HIV test) before they are termed to be positive.
The prior probability of anyone having HIV is 0.00148. The true positive rate for Elisa is 93% and the true negative rate is 99%.
29) What is the probability that a recruit has HIV, given he tested positive on first Elisa test? The prior probability of anyone having HIV is 0.00148. The true positive rate for Elisa is 93% and the true negative rate is 99%.
A) 12%
Solution: **(A)**
I recommend going through the Bayes updating section of this article for the understanding of the above question.

30) What is the probability of having HIV, given he tested positive on Elisa the second time as well.
The prior probability of anyone having HIV is 0.00148. The true positive rate for Elisa is 93% and the true negative rate is 99%.
C) 93%
Solution: **(C)**
I recommend going through the Bayes updating section of this article for the understanding of the above question.

31) Suppose you're playing a game in which we toss a fair coin multiple times. You have already lost thrice where you guessed heads but a tails appeared. Which of the below statements would be correct in this case?
C) You have the same probability of winning in guessing either, hence whatever you guess there is just a 50-50 chance of winning or losing
Solution: **(C)**
This is a classic problem of gambler's fallacy/monte carlo fallacy, where the person falsely starts to think that the results should even out in a few turns. The gambler starts to believe that if we have received 3 heads, you should receive a 3 tails. This is however not true. The results would even out only in infinite number of trials.

32) The inference using the frequentist approach will always yield the same result as the Bayesian approach.
B) FALSE
Solution: **(B)**
The frequentist Approach is highly dependent on how we define the hypothesis while Bayesian approach helps us update our prior beliefs. Therefore the frequentist approach might result in an opposite inference if we declare the hypothesis differently. Hence the two approaches might not yield the same results.

33) Hospital records show that 75% of patients suffering from a disease die due to that disease. What is the probability that 4 out of the 6 randomly selected patients recover?

C) 0.03295
Solution: **(C)**
Think of this as a binomial since there are only 2 outcomes, either the patient dies or he survives.
Here n =6, and x=4. p=0.25(probability if living(success)) q = 0.75(probability of dying(failure))
P(X) = nCx pxqn-x = 6C4 (0.25)4(0.75)2 = 0.03295

34) The students of a particular class were given two tests for evaluation. Twenty-five percent of the class cleared both the tests and forty-five percent of the students were able to clear the first test.
Calculate the percentage of students who passed the second test given that they were also able to pass the first test.

C) 55%
Solution: **(C)**
This is a simple problem of conditional probability. Let A be the event of passing in first test.
B is the event of passing in the second test.
P(A∩B) is passing in both the events
P(passing in second given he passed in the first one) = P(A∩B)/P(A)
= 0.25/0.45 which is around 55%

35) While it is said that the probabilities of having a boy or a girl are the same, let's assume that the actual probability of having a boy is slightly higher at 0.51. Suppose a couple plans to have 3 children. What is the probability that exactly 2 of them will be boys?

A) 0.38
Solution: **(A)**
Think of this as a binomial distribution where getting a success is a boy and failure is a girl. Therefore we need to calculate the probability of getting 2 out of three successes.
P(X) = nCx pxqn-x = 3C2 (0.51)2(0.49)1 = 0.382

36) Heights of 10 year-olds, regardless of gender, closely follow a normal distribution with mean 55 inches and standard deviation 6 inches. Which of the following is true?

D) None of these
Solution: **(D)**
None of the above statements are true.

37) About 30% of human twins are identical, and the rest are fraternal. Identical twins are necessarily the same sex, half are males and the other half are females. One-quarter of fraternal twins are both males, one-quarter both female, and one-half are mixed: one male, one female. You have just become a parent of twins and are told they are both girls. Given this information, what is the probability that they are identical?

C) 46%

Solution: **(C)**
This is a classic problem of Bayes theorem.
P(I) denoted Probability of being identical and P(~I) denotes Probability of not being identical
P(Identical) = 0.3
P(not Identical)= 0.7
P(FF|I)= 0.5
P(MM|I)= 0.5
P(MM|~I)= 0.25
P(FF|~I)= 0.25
P(FM|~I)= 0.25
P(I|FF) = 0.46

38) Rob has fever and the doctor suspects it to be typhoid. To be sure, the doctor wants to conduct the test. The test results positive when the patient actually has typhoid 80% of the time. The test gives positive when the patient does not have typhoid 10% of the time. If 1% of the population has typhoid, what is the probability that Rob has typhoid provided he tested positive?

B) 7%
Solution: **(B)**
We need to find the probability of having typhoid given he tested positive.
=P(testing +ve and having typhoid) / P(testing positive)

= = 0.074

39) Jack is having two coins in his hand. Out of the two coins, one is a real coin and the second one is a faulty one with Tails on both sides. He blindfolds himself to choose a random coin and tosses it in the air. The coin falls down with Tails facing upwards. What is the probability that this tail is shown by the faulty coin?

B) 2/3
Solution: **(B)**
We need to find the probability of the coin being faulty given that it showed tails.
P(Faulty) = 0.5
P(getting tails) = 3/4
P(faulty and tails) =0.5*1 = 0.5
Therefore the probability of coin being faulty given that it showed tails would be 2/3

40) A fly has a life between 4-6 days. What is the probability that the fly will die at exactly 5 days?

D) 0
Solution: **(D)**
Here since the probabilities are continuous, the probabilities form a mass function. The probability of a certain event is calculated by finding the area under the curve for the given conditions. Here since we're trying to calculate the probability of the fly dying at exactly 5 days – the area under the curve would be 0. Also to come to think of it, the probability if dying at exactly 5 days is impossible for us to even figure out since we cannot measure with infinite precision if it was exactly 5 days.

Correlation

https://www.analyticsvidhya.com/blog/2015/06/correlation-common-questions/

Introduction

Understanding the Mathematical formulation of Correlation coefficient

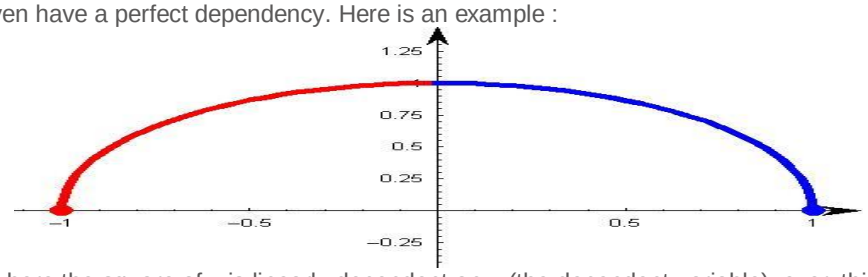
The most widely used correlation coefficient is Pearson Coefficient. Here is the mathematical formula to derive Pearson Coefficient.

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2}\sqrt{\sum_i (y_i - \overline{y})^2}}$$

Explanation: It simply is the ratio of co-variance of two variables to a product of variance (of the variables). It takes a value between +1 and -1. An extreme value on both the side means they are strongly correlated with each other. A value of zero indicates a NIL correlation but not a non-dependence. You'll understand this clearly in one of the following answers.

Answer – 1: Correlation vs. Dependency

A non-dependency between two variable means a zero correlation. However the inverse is not true. A zero correlation can even have a perfect dependency. Here is an example :



In this scenario, where the square of x is linearly dependent on y (the dependent variable), everything to the right of y axis is negative correlated and to left is positively correlated. So what will be the Pearson Correlation coefficient?

If you do the math, you will see a zero correlation between these two variables. What does that mean? For a pair of variables which are perfectly dependent on each other, can also give you a zero correlation.

Must remember tip: Correlation quantifies the linear dependence of two variables. It cannot capture non-linear relationship between two variables.

Answer – 2: Is Correlation Transitive?

Suppose that X, Y, and Z are random variables. X and Y are positively correlated and Y and Z are likewise positively correlated. Does it follow that X and Z must be positively correlated?

As we shall see by example, the answer is (perhaps surprisingly) "**No**." We may prove that if the correlations are sufficiently close to 1, then X and Z must be positively correlated.

Let's assume C(x,y) is the correlation coefficient between x and y. Like wise we have C(x,z) and C(y,z). Here is an equation which comes from solving correlation equation mathematically :

```
C(x,y) = C(y,z) * C(z,x) - Square Root ( (1 - C(y,z)^2 ) *  (1 - C(z,x)^2 ) )
```

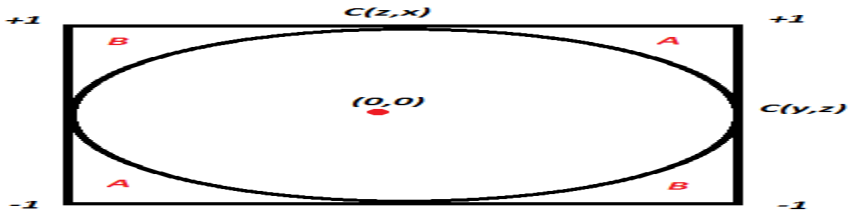
Now if we want C(x,y) to be more than zero , we basically want the RHS of above equation to be positive. Hence, you need to solve for :

```
C(y,z) * C(z,x) > Square Root ( (1 - C(y,z)^2 ) *  (1 - C(z,x)^2 ) )
```

We can actually solve the above equation for both C(y,z) > 0 and C(y,z) < 0 together by squaring both sides. This will finally give the result as C(x,y) is a non zero number if following equation holds true:

```
C(y,z) ^ 2 + C(z,x) ^ 2 > 1
```

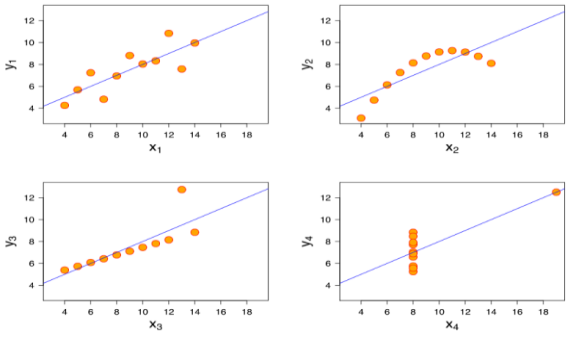
Wow, this is an equation for a circle. Hence the following plot will explain everything :



If the two known correlation are in the A zone, the third correlation will be positive. If they lie in the B zone, the third correlation will be negative. Inside the circle, we cannot say anything about the relationship. A very interesting insight here is that even if C(y,z) and C(z,x) are 0.5, C(x,y) can actually also be negative.

Answer – 3: Is Pearson coefficient sensitive to outliers?

The answer is Yes. Even a single outlier can change the direction of the coefficient. Here are a few cases, all of which have the same correlation coefficient of 0.81 :

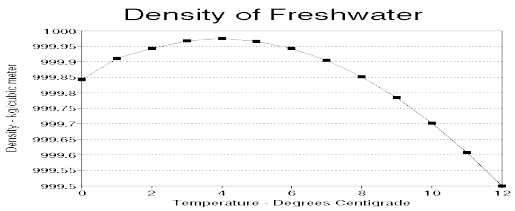


Consider the last two graphs(X 3Y3 and X 4Y4). X3Y3 is clearly a case of perfect correlation where a single outlier brings down the coefficient significantly. The last graph is complete opposite, the correlation coefficient becomes a high positive number because of a single outlier. Conclusively, this turns out to be the biggest concern with correlation coefficient, it is highly influenced by the outliers.

Answer – 4: Does causation imply correlation?

If you have read our above three answers, I am sure you will be able to answer this one. The answer is No, because causation can also lead to a non-linear relationship. Let's understand how!

Below is the graph showing density of water from 0 to 12 degree Celsius. We know that density is an effect of changing temperature. But, density can reach its maximum value at 4 degree Celsius. Therefore, it will not be linearly correlated to the temperature.



Answer – 5: Difference between Correlation and Simple Linear Regression

These two are really close. So let's start with a few things which are common for both.

- The square of Pearson's correlation coefficient is the same as the one in simple linear regression
- Neither simple linear regression nor correlation answer questions of causality directly. This point is important, because I've met people thinking that simple regression can magically allow an inference that X causes. That's preposterous belief.

What's the difference between correlation and simple linear regression?

Now let's think of few differences between the two. Simple linear regression gives much more information about the relationship than Pearson Correlation. Here are a few things which regression will give but correlation coefficient will not.

- The slope in a linear regression gives the marginal change in output/target variable by changing the independent variable by unit distance. Correlation has no slope.
- The intercept in a linear regression gives the value of target variable if one of the input/independent variable is set zero. Correlation does not have this information.
- Linear regression can give you a prediction given all the input variables. Correlation analysis does not predict anything

Answer – 6: Pearson vs. Spearman

The simplest answer here is Pearson captures how linearly dependent are the two variables whereas Spearman captures the monotonic behavior of the relation between the variables.

For instance consider following relationship :

$$y = exp\ (x)$$

Here you will find Pearson coefficient to be 0.25 but the Spearman coefficient to be 1. As a thumb rule, you should only begin with Spearman when you have some initial hypothesis of the relation being non-linear. Otherwise, we generally try Pearson first and if that is low, try Spearman. This way you know whether the variables are linearly related or just have a monotonic behavior.

Answer – 7: Correlation vs. co-variance

If you skipped the mathematical formula of correlation at the start of this article, now is the time to revisit the same.

Correlation is simply the normalized co-variance with the standard deviation of both the factors. This is done to ensure we get a number between +1 and -1. Co-variance is very difficult to compare as it depends on the units of the two variable. It might come out to be the case that marks of student is more correlated to his toe nail in mili-meters than it is to his attendance rate.

This is just because of the difference in units of the second variable. Hence, we see a need to normalize this co-variance with some spread to make sure we compare apples with apples. This normalized number is known as the correlation.

Statistics

https://www.analyticsvidhya.com/blog/2017/05/41-questions-on-statisitics-data-scientists-analysts/

Questions & Solution

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

B) Mean, Median and Mode

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.

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

D) 0

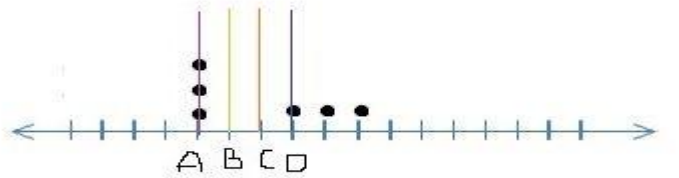
Solution: (D)

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

3) 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

Solution: (A)
X= μ+Zσ 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

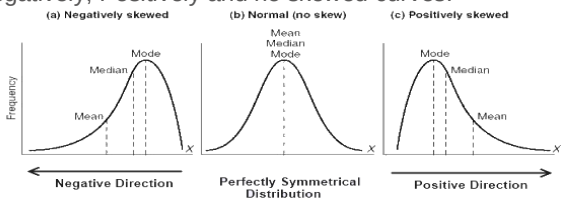
4) Which of the following measures of central tendency will always change if a single value in the data changes?
A) Mean
Solution: (A)
The mean of the dataset would always change if we change any value of the data set. Since we are summing up all the values together to get it, every value of the data set contributes to its value. Median and mode may or may not change with altering a single value in the dataset.
5) Below, we have represented six data points on a scale where vertical lines on scale represent unit.



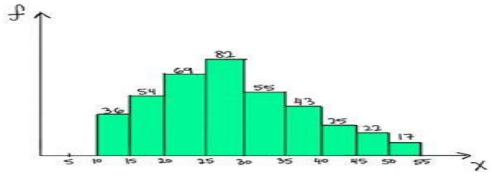
Which of the following line represents the mean of the given data points, where the scale is divided into same units?
C) C
Solution: (C)
It's a little tricky to visualize this one by just looking at the data points. We can simply substitute values to understand the mean. Let A be 1, B be 2, C be 3 and 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.

6) If a positively skewed distribution has a median of 50, which of the following statement is true?
A) Meanisgreaterthan50
C) Mode is less than 50

E) Both A and C
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.
7) Which of the following is a possible value for the median of the below distribution?



B) 26
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.

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

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.

C) Both 2 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.

9) 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

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.

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

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.

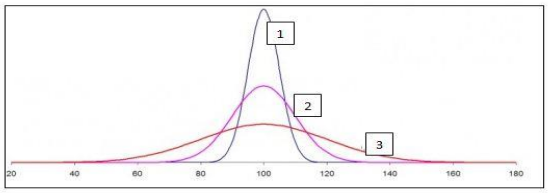
11) Standard deviation is robust to outliers?

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.

12) 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

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.

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

A)+/-2.33

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

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

A)TRUE

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.

Context for Questions 15-17
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.
15) What is the null hypothesis in this case?
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.
16) What would be the Type I error?
B) Concluding that listening to music while studying improves memory when it actually doesn't.
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.
17) After performing the Z-test, what can we conclude ____ ?
B)Listening to music significantly improves memory at p
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

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

Z critical value for α = 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.

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

19) What happens to the confidence interval when we introduce some outliers to the data?
B) Confidence interval will increase with the introduction of outliers.
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.

Context for questions 20- 22
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.
20) What is the standard error of the mean?
B) 6
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

21) What is the probability of getting a mean of 175 or less after all the patients start dieting?
A)20%
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.

22) Which of the following statement is correct?
B) The doctor does not have enough evidence that dieting reduces blood sugar level.
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.
Question Context 23-25
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

23) What is the value of t-statistic?
A) 3.191
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

24) Is there a significant difference in the scores of the two groups?
A)Yes
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 α = 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.

25) What percentage of variability in scores is explained by the method of teaching?
A) 36.13
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.

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

26) [True or False] F statistic cannot be negative.
A) TRUE

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.

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

B)

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.

28) 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_____ ?

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.

29) 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.**

C) 2and3

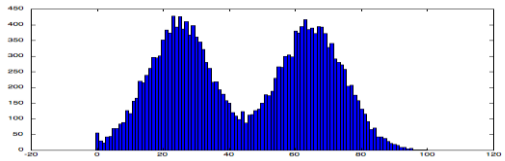
Solution: (C)
Though sometimes causation might be intuitive from a high correlation but actually correlation does not imply any causal inference. It just tells us the strength of the relationship between the two variables. If both the variables move together, there is a high correlation among them.

30) 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?

B) 74%

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.

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



B) Above histogram is bimodal

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.

32) 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?

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

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

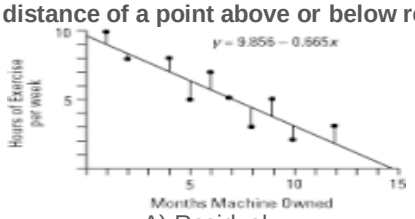
A) The r squared value may increase or remain constant, the adjusted r squared may increase or decrease.

.

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.

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



A) Residual

B) Prediction Error

D) Both A and B

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.

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

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

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.

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

B) Significance level = 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.

37) [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)
```

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.

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

39) 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
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**

40) 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
B)increaseby5pound
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.

41) [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
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

Linear regression

https://www.analyticsvidhya.com/blog/2017/07/30-questions-to-test-a-data-scientist-on-linear-regression/

Skill test Questions and Answers

1) True-False: Linear Regression is a supervised machine learning algorithm.

A) TRUE

Solution: (A)

Yes, Linear regression is a supervised learning algorithm because it uses true labels for training. Supervised learning algorithm should have input variable (x) and an output variable (Y) for each example.

2) True-False: Linear Regression is mainly used for Regression.

A) TRUE

Solution: (A)

Linear Regression has dependent variables that have continuous values.

3) True-False: It is possible to design a Linear regression algorithm using a neural network?

A) TRUE

Solution: (A)

True. A Neural network can be used as a *universal* approximator, so it can definitely implement a linear regression algorithm.

4) Which of the following methods do we use to find the best fit line for data in Linear Regression?

A) Least Square Error

Solution: (A)

In linear regression, we try to minimize the least square errors of the model to identify the line of best fit.

5) Which of the following evaluation metrics can be used to evaluate a model while modeling a continuous output variable?

D) Mean-Squared-Error

Solution: (D)

Since linear regression gives output as continuous values, so in such case we use mean squared error metric to evaluate the model performance. Remaining options are use in case of a classification problem use UC-ROC,ACCURACY,LOGLOSS.

6) True-False: Lasso Regularization can be used for variable selection in Linear Regression.

A) TRUE

Solution: (A)

True, In case of lasso regression we apply absolute penalty which makes some of the coefficients zero.

7) Which of the following is true about Residuals?

Solution: (A)

Residuals refer to the error values of the model. Therefore lower residuals are desired.

8) Suppose that we have N independent variables (X1,X2… Xn) and dependent variable is Y. Now Imagine that you are applying linear regression by fitting the best fit line using least square error on this data.
You found that correlation coefficient for one of it's variable(Say X1) with Y is -0.95.
Which of the following is true for X1?

A) Relation between the X1 and Y is weak
B) Relation between the X1 and Y is strong
C) Relation between the X1 and Y is neutral
D) Correlation can't judge the relationship

Solution: (B)

The absolute value of the correlation coefficient denotes the strength of the relationship. Since absolute correlation is very high it means that the relationship is strong between X1 and Y.

9) Looking at above two characteristics, which of the following option is the correct for Pearson correlation between V1 and V2?

If you are given the two variables V1 and V2 and they are following below two characteristics.
1. If V1 increases then V2 also increases
2. If V1 decreases then V2 behavior is unknown
will be close to 0
D) None of these

Solution: (D)

We cannot comment on the correlation coefficient by using only statement 1. We need to consider the both of these two statements. Consider V1 as x and V2 as |x|. The correlation coefficient would not be close to 1 in such a case.

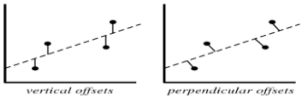
10) Suppose Pearson correlation between V1 and V2 is zero. In such case, is it right to conclude that V1 and V2 do not have any relation between them?

B) FALSE

Solution: (B)

Pearson correlation coefficient between 2 variables might be zero even when they have a relationship between them. If the correlation coefficient is zero, it just means that that they don't move together. We can take examples like y=|x| or y=x^2.

11) Which of the following offsets, do we use in linear regression's least square line fit? Suppose horizontal axis is independent variable and vertical axis is dependent variable.



A) Vertical offset

Solution: (A)

We always consider residuals as vertical offsets. We calculate the direct differences between actual value and the Y labels. Perpendicular offset are useful in case of PCA.

12) True- False: Overfitting is more likely when you have huge amount of data to train?

B) FALSE

Solution: (B)

With a small training dataset, it's easier to find a hypothesis to fit the training data exactly i.e. overfitting.

13) We can also compute the coefficient of linear regression with the help of an analytical method called "Normal Equation". Which of the following is/are true about Normal Equation?

1. We don't have to choose the learning rate
2. It becomes slow when number of features is very large
3. Thers is no need to iterate

D) 1,2 and 3

Solution: (D)

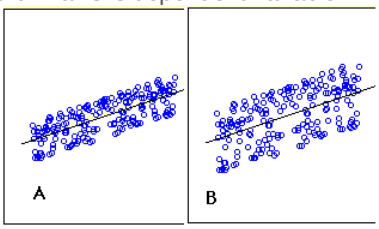
Instead of gradient descent, Normal Equation can also be used to find coefficients. Refer this article for read more about normal equation.

14) Which of the following statement is true about sum of residuals of A and B?

Below graphs show two fitted regression lines (A & B) on randomly generated data. Now, I want to find the sum of residuals in both cases A and B.

Note:

1. Scale is same in both graphs for both axis.
2. X axis is independent variable and Y-axis is dependent variable.



C) Both have same sum of residuals

Solution: (C)

Sum of residuals will always be zero, therefore both have same sum of residuals

Question Context 15-17:

Suppose you have fitted a complex regression model on a dataset. Now, you are using Ridge regression with penality x.

15) Choose the option which describes bias in best manner.

A) In case of very large x; bias is low

B) In case of very large x; bias is high

C) We can't say about bias

D) None of these

Solution: (B)

If the penalty is very large it means model is less complex, therefore the bias would be high.

16) What will happen when you apply very large penalty?

B) Some of the coefficient will approach zero but not absolute zero

Solution: (B)

In lasso some of the coefficient value become zero, but in case of Ridge, the coefficients become close to zero but not zero.

17) What will happen when you apply very large penalty in case of Lasso?

A) Some of the coefficient will become zero

Solution: (A)

As already discussed, lasso applies absolute penalty, so some of the coefficients will become zero.

18) Which of the following statement is true about outliers in Linear regression?

A) Linear regression is sensitive to outliers

Solution: (A)

The slope of the regression line will change due to outliers in most of the cases. So Linear Regression is sensitive to outliers.

19) Suppose you plotted a scatter plot between the residuals and predicted values in linear regression and you found that there is a relationship between them. Which of the following conclusion do you make about this situation?

A) Since the there is a relationship means our model is not good

Solution: (A)

There should not be any relationship between predicted values and residuals. If there exists any relationship between them,it means that the model has not perfectly captured the information in the data.

Question Context 20-22:

Suppose that you have a dataset D1 and you design a linear regression model of degree 3 polynomial and you found that the training and testing error is "0" or in another terms it perfectly fits the data.

20) What will happen when you fit degree 4 polynomial in linear regression?

A) There are high chances that degree 4 polynomial will over fit the data

Solution: (A)

Since is more degree 4 will be more complex(overfit the data) than the degree 3 model so it will again perfectly fit the data. In such case training error will be zero but test error may not be zero.

21) What will happen when you fit degree 2 polynomial in linear regression?

B) It is high chances that degree 2 polynomial will under fit the data

Solution: (B)

If a degree 3 polynomial fits the data perfectly, it's highly likely that a simpler model(degree 2 polynomial) might under fit the data.

22) In terms of bias and variance. Which of the following is true when you fit degree 2 polynomial?

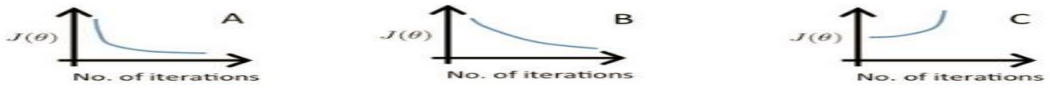
C) Bias will be high, variance will be low

Solution: (C)

Since a degree 2 polynomial will be less complex as compared to degree 3, the bias will be high and variance will be low.

Question Context 23:

Which of the following is true about below graphs(A,B, C left to right) between the cost function and Number of iterations?



23) Suppose l1, l2 and l3 are the three learning rates for A,B,C respectively. Which of the following is true about l1,l2 and l3?

A) l2 < l1 < l3

Solution: (A)

In case of high learning rate, step will be high, the objective function will decrease quickly initially, but it will not find the global minima and objective function starts increasing after a few iterations.

In case of low learning rate, the step will be small. So the objective function will decrease slowly

Question Context 24-25:

We have been given a dataset with n records in which we have input attribute as x and output attribute as y. Suppose we use a linear regression method to model this data. To test our linear regressor, we split the data in training set and test set randomly.

24) Now we increase the training set size gradually. As the training set size increases, what do you expect will happen with the mean training error?

D) Can't Say

Solution: (D)

Training error may increase or decrease depending on the values that are used to fit the model. If the values used to train contain more outliers gradually, then the error might just increase.

25) What do you expect will happen with bias and variance as you increase the size of training data?

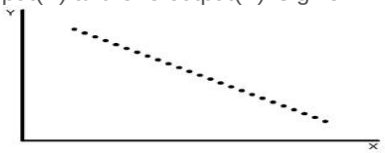
D) Bias increases and Variance decreases

Solution: (D)

As we increase the size of the training data, the bias would increase while the variance would decrease.

Question Context 26:

Consider the following data where one input(X) and one output(Y) is given.



26) What would be the root mean square training error for this data if you run a Linear Regression model of the form (Y = A0+A1X)?

C) Equal to 0

Solution: (C)

We can perfectly fit the line on the following data so mean error will be zero.

Question Context 27-28:

Suppose you have been given the following scenario for training and validation error for Linear Regression.

Scenario	Learning Rate	Number of iterations	Training Error	Validation Error
1	0.1	1000	100	110
2	0.2	600	90	105
3	0.3	400	110	110
4	0.4	300	120	130
5	0.4	250	130	150

27) Which of the following scenario would give you the right hyper parameter?

B) 2

Solution: (B)

Option B would be the better option because it leads to less training as well as validation error.

28) Suppose you got the tuned hyper parameters from the previous question. Now, Imagine you want to add a variable in variable space such that this added feature is important. Which of the following thing would you observe in such case?

D) Training Error will decrease and Validation error will decrease

Solution: (D)

If the added feature is important, the training and validation error would decrease.

Question Context 29-30:

Suppose, you got a situation where you find that your linear regression model is under fitting the data.

29) In such situation which of the following options would you consider?

1. I will add more variables
2. I will start introducing polynomial degree variables
3. I will remove some variables

A) 1 and 2

Solution: (A)

In case of under fitting, you need to induce more variables in variable space or you can add some polynomial degree variables to make the model more complex to be able to fir the data better.

30) Now situation is same as written in previous question(under fitting).Which of following regularization L1 or L2 algorithm would you prefer?

D) None of these

Solution: (D)

I won't use any regularization methods because regularization is used in case of overfitting.

Logistics Regression

https://www.analyticsvidhya.com/blog/2017/08/skilltest-logistic-regression/

Skill test Questions and Answers

1) True-False: Is Logistic regression a supervised machine learning algorithm?
A) TRUE
Solution: A
True, Logistic regression is a supervised learning algorithm because it uses true labels for training. Supervised learning algorithm should have input variables (x) and an target variable (Y) when you train the model .
2) True-False: Is Logistic regression mainly used for Regression?
B) FALSE
Solution: B
Logistic regression is a classification algorithm, don't confuse with the name regression.
3) True-False: Is it possible to design a logistic regression algorithm using a Neural Network Algorithm?
A) TRUE
Solution: A
True, Neural network is a is a *universal* approximator so it can implement linear regression algorithm.
4) True-False: Is it possible to apply a logistic regression algorithm on a 3-class Classification problem?
A) TRUE
Solution: A
Yes, we can apply logistic regression on 3 classification problem, We can use One Vs all method for 3 class classification in logistic regression.
5) Which of the following methods do we use to best fit the data in Logistic Regression?
B) Maximum Likelihood
Solution: B
Logistic regression uses maximum likely hood estimate for training a logistic regression.
6) Which of the following evaluation metrics can not be applied in case of logistic regression output to compare with target?
A) AUC-ROC
B) Accuracy
C) Logloss
D) Mean-Squared-Error
Solution: D
Since, Logistic Regression is a classification algorithm so it's output can not be real time value so mean squared error can not use for evaluating it
7) One of the very good methods to analyze the performance of Logistic Regression is AIC, which is similar to R-Squared in Linear Regression. Which of the following is true about AIC?
A) We prefer a model with minimum AIC value
Solution: A
We select the best model in logistic regression which can least AIC. For more information refer this source: http://www4.ncsu.edu/~shu3/Presentation/AIC.pdf
8) [True-False] Standardisation of features is required before training a Logistic Regression.
B) FALSE
Solution: B
Standardization isn't required for logistic regression. The main goal of standardizing features is to help convergence of the technique used for optimization.

9) Which of the following algorithms do we use for Variable Selection?
A) LASSO
Solution: A
In case of lasso we apply a absolute penality, after increasing the penality in lasso some of the coefficient of variables may become zero.
Context: 10-11
Consider a following model for logistic regression: P (y =1|x, w)= g(w0 + w1x)
where g(z) is the logistic function.
In the above equation the P (y =1|x; w) , viewed as a function of x, that we can get by changing the parameters w.
10) What would be the range of p in such case?
C) (0, 1)
Solution: C
For values of *x* in the range of real number from −∞ to +∞ Logistic function will give the output between (0,1)
11) In above question what do you think which function would make p between (0,1)?
A) logistic function
Solution: A
Explanation is same as question number 10
Context: 12-13
Suppose you train a logistic regression classifier and your hypothesis function H is

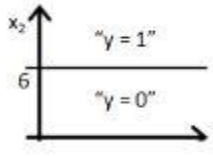
$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

Where

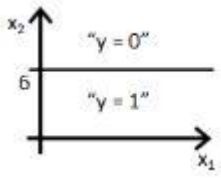
$$\theta_0 = 6, \theta_1 = 0, \theta_2 = -1.$$

12) Which of the following figure will represent the decision boundary as given by above classifier?

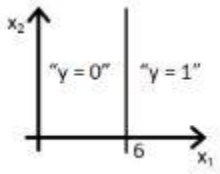
A)



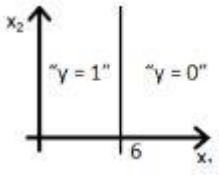
B)



C)



D

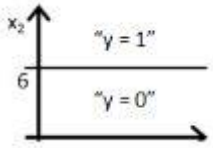


Solution: B

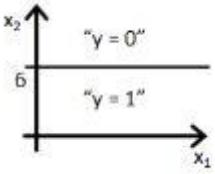
Option B would be the right answer. Since our line will be represented by y = g(-6+x2) which is shown in the option A and option B. But option B is the right answer because when you put the value x2 = 6 in the equation then y = g(0) you will get that means y= 0.5 will be on the line, if you increase the value of x2 greater then 6 you will get negative values so output will be the region y =0.

13) If you replace coefficient of x1 with x2 what would be the output figure?

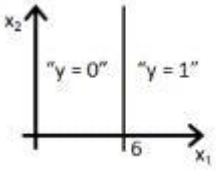
A)



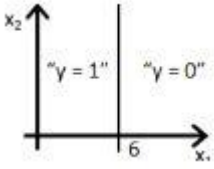
B)



C)



D)



Solution: D

Same explanation as in previous question.

14) Suppose you have been given a fair coin and you want to find out the odds of getting heads. Which of the following option is true for such a case?

C) odds will be 1

Solution: C

Odds are defined as the ratio of the probability of success and the probability of failure. So in case of fair coin probability of success is 1/2 and the probability of failure is 1/2 so odd would be 1

15) The logit function(given as l(x)) is the log of odds function. What could be the range of logit function in the domain x=[0,1]?

A) (– ∞ , ∞)

Solution: A

For our purposes, the odds function has the advantage of transforming the probability function, which has values from 0 to 1, into an equivalent function with values between 0 and ∞. When we take the natural log of the odds function, we get a range of values from -∞ to ∞.

16) Which of the following option is true?

A) Linear Regression errors values has to be normally distributed but in case of Logistic Regression it is not the case

Solution:A

Only A is true. Refer this tutorial https://czep.net/stat/mlelr.pdf

17) Which of the following is true regarding the logistic function for any value "x"?

Note:

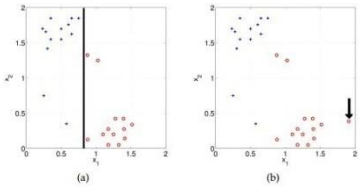
Logistic(x): is a logistic function of any number "x"
Logit(x): is a logit function of any number "x"
Logit_inv(x): is a inverse logit function of any number "x"

A) Logistic(x) = Logit(x)
B) Logistic(x) = Logit_inv(x)
C) Logit_inv(x) = Logit(x)
D) None of these

Solution: B

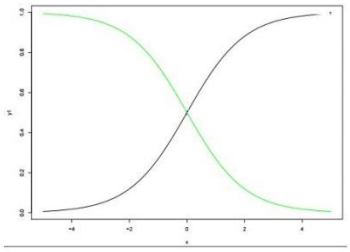
Refer this link for the solution: https://en.wikipedia.org/wiki/Logit

18) How will the bias change on using high(infinite) regularisation?
Suppose you have given the two scatter plot "a" and "b" for two classes(blue for positive and red for negative class). In scatter plot "a", you correctly classified all data points using logistic regression (black line is a decision boundary).



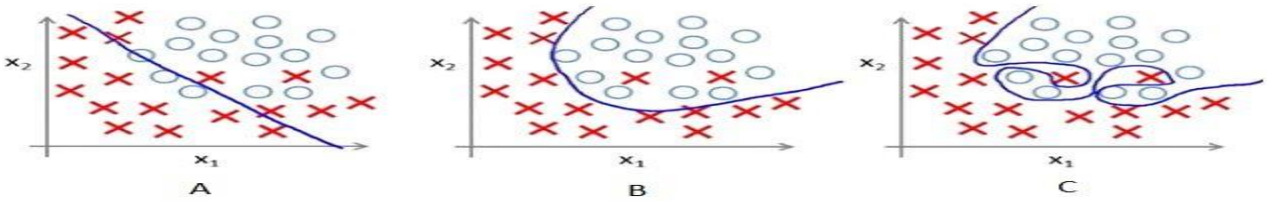
A) Bias will be high
Solution: A
Model will become very simple so bias will be very high.
19) Suppose, You applied a Logistic Regression model on a given data and got a training accuracy X and testing accuracy Y. Now, you want to add a few new features in the same data. Select the option(s) which is/are correct in such a case.
Note: Consider remaining parameters are same.
A) Training accuracy increases
D) Testing accuracy increases or remains the same
Solution: A and D
Adding more features to model will increase the training accuracy because model has to consider more data to fit the logistic regression. But testing accuracy increases if feature is found to be significant
20) Choose which of the following options is true regarding One-Vs-All method in Logistic Regression.
A) We need to fit n models in n-class classification problem
Solution: A
If there are n classes, then n separate logistic regression has to fit, where the probability of each category is predicted over the rest of the categories combined.

21) Below are two different logistic models with different values for β0 and β1.



Which of the following statement(s) is true about β0 and β1 values of two logistics models (Green, Black)?
Note: consider Y = β0 + β1*X. Here, β0 is intercept and β1 is coefficient.
A) β1 for Green is greater than Black
B) β1 for Green is lower than Black
C) β1 for both models is same
D) Can't Say
Solution: B
β0 and β1: β0 = 0, β1 = 1 is in X1 color(black) and β0 = 0, β1 = −1 is in X4 color (green)

Context 22-24
Below are the three scatter plot(A,B,C left to right) and hand drawn decision boundaries for logistic regression.



22) Which of the following above figure shows that the decision boundary is overfitting the training data?
C) C
Solution: C
Since in figure 3, Decision boundary is not smooth that means it will over-fitting the data.

23) What do you conclude after seeing this visualization?

1. The training error in first plot is maximum as compare to second and third plot.
2. The best model for this regression problem is the last (third) plot because it has minimum training error (zero).
3. The second model is more robust than first and third because it will perform best on unseen data.
4. The third model is overfitting more as compare to first and second.
5. All will perform same because we have not seen the testing data.

C) 1, 3 and 4

Solution: C

The trend in the graphs looks like a quadratic trend over independent variable X. A higher degree(Right graph) polynomial might have a very high accuracy on the train population but is expected to fail badly on test dataset. But if you see in left graph we will have training error maximum because it underfits the training data

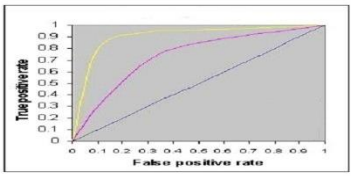
24) Suppose, above decision boundaries were generated for the different value of regularization. Which of the above decision boundary shows the maximum regularization?

A) A

Solution: A

Since, more regularization means more penality means less complex decision boundry that shows in first figure A.

25) The below figure shows AUC-ROC curves for three logistic regression models. Different colors show curves for different hyper parameters values. Which of the following AUC-ROC will give best result?



A) Yellow

Solution: A

The best classification is the largest area under the curve so yellow line has largest area under the curve.

26) What would do if you want to train logistic regression on same data that will take less time as well as give the comparatively similar accuracy(may not be same)?

Suppose you are using a Logistic Regression model on a huge dataset. One of the problem you may face on such huge data is that Logistic regression will take very long time to train.

D) Increase the learning rate and decrease the number of iteration

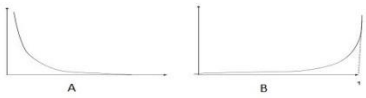
Solution: D

If you decrease the number of iteration while training it will take less time for surly but will not give the same accuracy for getting the similar accuracy but not exact you need to increase the learning rate.

27) Which of the following image is showing the cost function for y =1.

Following is the loss function in logistic regression(Y-axis loss function and x axis log probability) for two class classification problem.

Note: Y is the target class

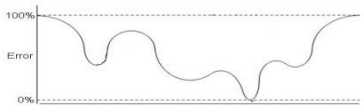


A) A

Solution: A

A is the true answer as loss function decreases as the log probability increases

28) Suppose, Following graph is a cost function for logistic regression.



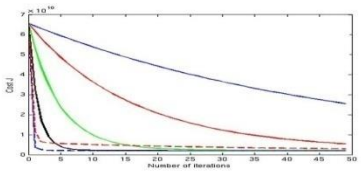
Now, How many local minimas are present in the graph?

C) 3

Solution: C

There are three local minima present in the graph

29) Imagine, you have given the below graph of logistic regression which is shows the relationships between cost function and number of iteration for 3 different learning rate values (different colors are showing different curves at different learning rates).



Suppose, you save the graph for future reference but you forgot to save the value of different learning rates for this graph. Now, you want to find out the relation between the leaning rate values of these curve. Which of the following will be the true relation?

Note:

1. The learning rate for blue is l1
2. The learning rate for red is l2
3. The learning rate for green is l3

C) l1 < l2 < l3

Solution: C

If you have low learning rate means your cost function will decrease slowly but in case of large learning rate cost function will decrease very fast.

30) Can a Logistic Regression classifier do a perfect classification on the below data?

Note: You can use only X1 and X2 variables where X1 and X2 can take only two binary values(0,1).

B) FALSE

Solution: B

No, logistic regression only forms linear decision surface, but the examples in the figure are not linearly separable.

https://www.cs.cmu.edu/~tom/10701_sp11/midterm_sol.pdf

40 Questions to test a data scientist on Machine Learning

https://www.analyticsvidhya.com/blog/2017/04/40-questions-test-data-scientist-machine-learning-solution-skillpower-machine-learning-datafest-2017/

Question Context

A feature F1 can take certain value: A, B, C, D, E, & F and represents grade of students from a college.

1) Which of the following statement is true in following case?.

B) Feature F1 is an example of ordinal variable.

Solution: (B)

Ordinal variables are the variables which has some order in their categories. For example, grade A should be consider as high grade than grade B.

2) Which of the following is an example of a deterministic algorithm?

A) PCA

Solution: (A)

A deterministic algorithm is that in which output does not change on different runs. PCA would give the same result if we run again, but not k-means.

3) [True or False] A Pearson correlation between two variables is zero but, still their values can still be related to each other.

A) TRUE

Solution: (A)

*Y=X*2. Note that, they are not only associated, but one is a function of the other and Pearson correlation between them is 0.

4) Which of the following statement(s) is / are true for Gradient Decent (GD) and Stochastic Gradient Decent (SGD)?

1. **In GD and SGD, you update a set of parameters in an iterative manner to minimize the error function.**
2. **In SGD, you have to run through all the samples in your training set for a single update of a parameter in each iteration.**
3. **In GD, you either use the entire data or a subset of training data to update a parameter in each iteration.**

A) Only 1
Solution: (A)
In SGD for each iteration you choose the batch which is generally contain the random sample of data But in case of GD each iteration contain the all of the training observations.

5) Which of the following hyper parameter(s), when increased may cause random forest to over fit the data?

1. **Number of Trees**
2. **Depth of Tree**
3. **Learning Rate**

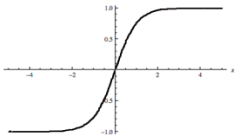
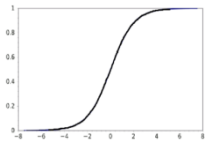
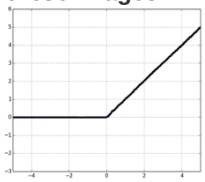
B) Only 2
Solution: (B)
Usually, if we increase the depth of tree it will cause overfitting. Learning rate is not an hyperparameter in random forest. Increase in the number of tree will cause under fitting.

6) Imagine, you are working with "Analytics Vidhya" and you want to develop a machine learning algorithm which predicts the number of views on the articles.
Your analysis is based on features like author name, number of articles written by the same author on Analytics Vidhya in past and a few other features. Which of the following evaluation metric would you choose in that case?

1. **Mean Square Error**
2. **Accuracy**
3. **F1 Score**

A) Only 1
Solution:(A)
You can think that the number of views of articles is the continuous target variable which fall under the regression problem. So, mean squared error will be used as an evaluation metrics.

7) Given below are three images (1,2,3). Which of the following option is correct for these images?

A)  B)  C) 

.
D) 1 is tanh, 2 is SIGMOID and 3 is ReLU activation functions.
Solution: (D)
The range of SIGMOID function is [0,1].The range of the tanh function is [-1,1].The range of the RELU function is [0, infinity].
So Option D is the right answer.

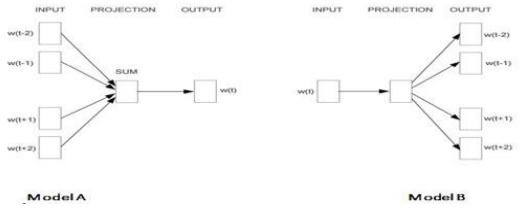
8) Below are the 8 actual values of target variable in the train file.
[0,0,0,1,1,1,1,1]
What is the entropy of the target variable?
A) -(5/8 log(5/8) + 3/8 log(3/8))
Solution: (A)

$$-\sum p(x) * log p(x)$$

The formula for entropy is
So the answer is A.

9) Let's say, you are working with categorical feature(s) and you have not looked at the distribution of the categorical variable in the test data.
You want to apply one hot encoding (OHE) on the categorical feature(s). What challenges you may face if you have applied OHE on a categorical variable of train dataset?
A) All categories of categorical variable are not present in the test dataset.
B) Frequency distribution of categories is different in train as compared to the test dataset.
D) Both A and B
Solution: (D)
Both are true, The OHE will fail to encode the categories which are present in test but not in train so it could be one of the main challenges while applying OHE. The challenge given in option B is also true you need to more careful while applying OHE if frequency distribution doesn't same in train and test.

10) Skip gram model is one of the best models used in Word2vec algorithm for words embedding. Which one of the following models depict the skip gram model?



B) B

Solution: (B)

Both models (model1 and model2) are used in Word2vec algorithm. The model1 represent a CBOW model where as Model2 represent the Skip gram model.

11) Let's say, you are using activation function X in hidden layers of neural network. At a particular neuron for any given input, you get the output as "-0.0001". Which of the following activation function could X represent?

B) tanh

Solution: (B)

The function is a tanh because the this function output range is between (-1,-1).

12) [True or False] LogLoss evaluation metric can have negative values.

B) FALSE

Solution: (B)

Log loss cannot have negative values.

13) Which of the following statements is/are true about "Type-1" and "Type-2" errors?

1. **Type1 is known as false positive and Type2 is known as false negative.**
2. **Type1 is known as false negative and Type2 is known as false positive.**
3. **Type1 error occurs when we reject a null hypothesis when it is actually true.**

E) 1 and 3

Solution: (E)

In statistical hypothesis testing, a type I error is the incorrect rejection of a true null hypothesis (a "false positive"), while a type II error is incorrectly retaining a false null hypothesis (a "false negative").

14) Which of the following is/are one of the important step(s) to pre-process the text in NLP based projects?

1. **Stemming**
2. **Stop word removal**
3. **Object Standardization**

D) 1,2 and 3

Solution: (D)

Stemming is a rudimentary rule-based process of stripping the suffixes ("ing", "ly", "es", "s" etc) from a word.
Stop words are those words which will have not relevant to the context of the data for example is/am/are.
Object Standardization is also one of the good way to pre-process the text.

15) Suppose you want to project high dimensional data into lower dimensions. The two most famous dimensionality reduction algorithms used here are PCA and t-SNE. Let's say you have applied both algorithms respectively on data "X" and you got the datasets "X_projected_PCA" , "X_projected_tSNE". Which of the following statements is true for "X_projected_PCA" & "X_projected_tSNE" ?

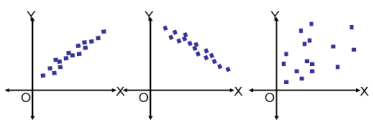
B) X_projected_tSNE will have interpretation in the nearest neighbour space.

Solution: (B)

t-SNE algorithm consider nearest neighbour points to reduce the dimensionality of the data. So, after using t-SNE we can think that reduced dimensions will also have interpretation in nearest neighbour space. But in case of PCA it is not the case.

Context: 16-17

Given below are three scatter plots for two features (Image 1, 2 & 3 from left to right).



16) In the above images, which of the following is/are example of multi-collinear features?

D) Features in Image 1 & 2

Solution: (D)

In Image 1, features have high positive correlation where as in Image 2 has high negative correlation between the features so in both images pair of features are the example of multicollinear features.

17) In previous question, suppose you have identified multi-collinear features. Which of the following action(s) would you perform next?

1. **Remove both collinear variables.**
2. **Instead of removing both variables, we can remove only one variable.**
3. **Removing correlated variables might lead to loss of information. In order to retain those variables, we can use penalized regression models like ridge or lasso regression.**

E) Either 2 or 3

Solution: (E)

You cannot remove the both features because after removing the both features you will lose all of the information so you should either remove the only 1 feature or you can use the regularization algorithm like L1 and L2.

18) Adding a non-important feature to a linear regression model may result in.

1. **Increase in R-square**
2. **Decrease in R-square**

A) Only 1 is correct

Solution: (A)

After adding a feature in feature space, whether that feature is important or unimportant features the R-squared always increase.

19) Suppose, you are given three variables X, Y and Z. The Pearson correlation coefficients for (X, Y), (Y, Z) and (X, Z) are C1, C2 & C3 respectively.
Now, you have added 2 in all values of X (i.enew values become X+2), subtracted 2 from all values of Y (i.e. new values are Y-2) and Z remains the same. The new coefficients for (X,Y), (Y,Z) and (X,Z) are given by D1, D2 & D3 respectively. How do the values of D1, D2 & D3 relate to C1, C2 & C3?

E) D1 = C1, D2 = C2, D3 = C3

Solution: (E)

Correlation between the features won't change if you add or subtract a value in the features.

20) Imagine, you are solving a classification problems with highly imbalanced class. The majority class is observed 99% of times in the training data.
Your model has 99% accuracy after taking the predictions on test data. Which of the following is true in such a case?

1. Accuracy metric is not a good idea for imbalanced class problems.
2. Accuracy metric is a good idea for imbalanced class problems.
3. Precision and recall metrics are good for imbalanced class problems.
4. Precision and recall metrics aren't good for imbalanced class problems.

A) 1 and 3

Solution: (A)

Refer the question number 4 from in this article.

21) In ensemble learning, you aggregate the predictions for weak learners, so that an ensemble of these models will give a better prediction than prediction of individual models.
Which of the following statements is / are true for weak learners used in ensemble model?

1. **They don't usually overfit.**
2. **They have high bias, so they cannot solve complex learning problems**
3. **They usually overfit.**

A) 1 and 2

Solution: (A)

Weak learners are sure about particular part of a problem. So, they usually don't overfit which means that weak learners have low variance and high bias.

22) Which of the following options is/are true for K-fold cross-validation?

1. **Increase in K will result in higher time required to cross validate the result.**
2. **Higher values of K will result in higher confidence on the cross-validation result as compared to lower value of K.**
3. **If K=N, then it is called Leave one out cross validation, where N is the number of observations.**

D) 1,2 and 3

Solution: (D)

Larger k value means less bias towards overestimating the true expected error (as training folds will be closer to the total dataset) and higher running time (as you are getting closer to the limit case: Leave-One-Out CV). We also need to consider the variance between the k folds accuracy while selecting the k.

Question Context 23-24

Cross-validation is an important step in machine learning for hyper parameter tuning. Let's say you are tuning a hyper-parameter "max_depth" for GBM by selecting it from 10 different depth values (values are greater than 2) for tree based model using 5-fold cross validation.
Time taken by an algorithm for training (on a model with max_depth 2) 4-fold is 10 seconds and for the prediction on remaining 1-fold is 2 seconds.
Note: Ignore hardware dependencies from the equation.

23) Which of the following option is true for overall execution time for 5-fold cross validation with 10 different values of "max_depth"?

D) More than or equal to 600 seconds

Solution: (D)

Each iteration for depth "2" in 5-fold cross validation will take 10 secs for training and 2 second for testing. So, 5 folds will take 12*5 = 60 seconds. Since we are searching over the 10 depth values so the algorithm would take 60*10 = 600 seconds. But training and testing a model on depth greater than 2 will take more time than depth "2" so overall timing would be greater than 600.

24) In previous question, if you train the same algorithm for tuning 2 hyper parameters say "max_depth" and "learning_rate".

You want to select the right value against "max_depth" (from given 10 depth values) and learning rate (from given 5 different learning rates). In such cases, which of the following will represent the overall time?

D) None of these

Solution: (D)

Same as question number 23.

25) Given below is a scenario for training error TE and Validation error VE for a machine learning algorithm M1. You want to choose a hyperparameter (H) based on TE and VE.

H	TE	VE
1	105	90
2	200	85
3	250	96
4	105	85
5	300	100

Which value of H will you choose based on the above table?

D) 4

Solution: (D)

Looking at the table, option D seems the best

26) What would you do in PCA to get the same projection as SVD?

A) Transform data to zero mean

Solution: (A)

When the data has a zero mean vector PCA will have same projections as SVD, otherwise you have to centre the data first before taking SVD.

Question Context 27-28

Assume there is a black box algorithm, which takes training data with multiple observations (t1, t2, t3,…….. tn) and a new observation (q1). The black box outputs the nearest neighbor of q1 (say ti) and its corresponding class label ci.

You can also think that this black box algorithm is same as 1-NN (1-nearest neighbor).

27) It is possible to construct a k-NN classification algorithm based on this black box alone.

Note: Where n (number of training observations) is very large compared to k.

A) TRUE

Solution: (A)

In first step, you pass an observation (q1) in the black box algorithm so this algorithm would return a nearest observation and its class.

In second step, you through it out nearest observation from train data and again input the observation (q1). The black box algorithm will again return the a nearest observation and it's class.

You need to repeat this procedure k times

28) Instead of using 1-NN black box we want to use the j-NN (j>1) algorithm as black box. Which of the following option is correct for finding k-NN using j-NN?

1. **J must be a proper factor of k**
2. **J > k**
3. **Not possible**

A) 1

Solution: (A)

Same as question number 27

29) Suppose you are given 7 Scatter plots 1-7 (left to right) and you want to compare Pearson correlation coefficients between variables of each scatterplot.

Which of the following is in the right order?

1. 1<2<3<4
2. 1>2>3 > 4
3. 7<6<5<4
4. 7>6>5>4

B) 2 and 3

Solution: (B)

From image 1to 4 correlation is decreasing (absolute value).But from image 4 to 7 correlation is increasing but values are negative (for example, 0, -0.3, -0.7, -0.99).

30) You can evaluate the performance of a binary class classification problem using different metrics such as accuracy, log-loss, F-Score. Let's say, you are using the log-loss function as evaluation metric.
Which of the following option is / are true for interpretation of log-loss as an evaluation metric?

$$logLoss = \frac{-1}{N}\sum_{i=1}^{N}(y_i(logp_i) + (1 - y_i)log(1 - p_i))$$

1. If a classifier is confident about an incorrect classification, then log-loss will penalise it heavily.
2. For a particular observation, the classifier assigns a very small probability for the correct class then the corresponding contribution to the log-loss will be very large.
3. Lower the log-loss, the better is the model.

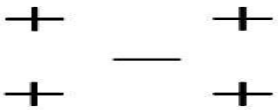
D) 1,2 and 3

Solution: (D)

Options are self-explanatory.

Question 31-32

Below are five samples given in the dataset.



Note: Visual distance between the points in the image represents the actual distance.

31) Which of the following is leave-one-out cross-validation accuracy for 3-NN (3-nearest neighbor)?

C) 0.8

Solution: (C)

In Leave-One-Out cross validation, we will select (n-1) observations for training and 1 observation of validation. Consider each point as a cross validation point and then find the 3 nearest point to this point. So if you repeat this procedure for all points you will get the correct classification for all positive class given in the above figure but negative class will be misclassified. Hence you will get 80% accuracy.

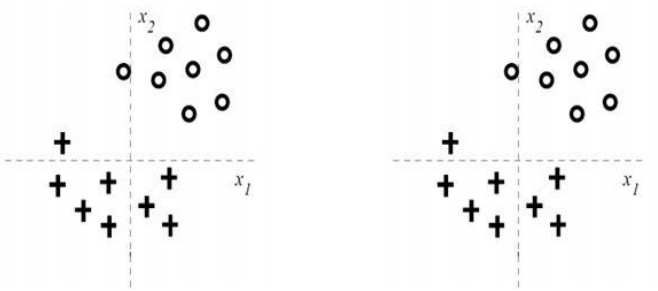
32) Which of the following value of K will have least leave-one-out cross validation accuracy?

A) 1NN

Solution: (A)

Each point which will always be misclassified in 1-NN which means that you will get the 0% accuracy.

33) Suppose you are given the below data and you want to apply a logistic regression model for classifying it in two given classes.



You are using logistic regression with L1 regularization.

$$\sum_{i=1}^{n} \log P(y_i|x_i, w_0, w_1, w_2) - C(|w_1| + |w_2|).$$

Where C is the regularization parameter and w1 & w2 are the coefficients of x1 and x2.
Which of the following option is correct when you increase the value of C from zero to a very large value?
B) First w1 becomes zero and then w2 becomes zero
Solution: (B)
By looking at the image, we see that even on just using x2, we can efficiently perform classification. So at first w1 will become 0. As regularization parameter increases more, w2 will come more and closer to 0.
34) Suppose we have a dataset which can be trained with 100% accuracy with help of a decision tree of depth 6. Now consider the points below and choose the option based on these points.
Note: All other hyper parameters are same and other factors are not affected.

1. **Depth 4 will have high bias and low variance**
2. **Depth 4 will have low bias and low variance**

A) Only 1
Solution: (A)
If you fit decision tree of depth 4 in such data means it will more likely to underfit the data. So, in case of underfitting you will have high bias and low variance.
35) Which of the following options can be used to get global minima in k-Means Algorithm?

1. **Try to run algorithm for different centroid initialization**
2. **Adjust number of iterations**
3. **Find out the optimal number of clusters**

D) All of above
Solution: (D)
All of the option can be tuned to find the global minima.
36) Imagine you are working on a project which is a binary classification problem. You trained a model on training dataset and get the below confusion matrix on validation dataset.

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

Based on the above confusion matrix, choose which option(s) below will give you correct predictions?

1. **Accuracy is ~0.91**
2. **Misclassification rate is ~ 0.91**
3. **False positive rate is ~0.95**
4. **True positive rate is ~**0.95

C) 1 and 4
Solution: (C)
The Accuracy (correct classification) is (50+100)/165 which is nearly equal to 0.91.
The true Positive Rate is how many times you are predicting positive class correctly so true positive rate would be 100/105 = 0.95 also known as "Sensitivity" or "Recall"
37) For which of the following hyperparameters, higher value is better for decision tree algorithm?

1. **Number of samples used for split**
2. **Depth of tree**
3. **Samples for leaf**

E) Can't say
Solution: (E)
For all three options A, B and C, it is not necessary that if you increase the value of parameter the performance may increase. For example, if we have a very high value of depth of tree, the resulting tree may overfit the data, and would not generalize well. On the other hand, if we have a very low value, the tree may underfit the data. So, we can't say for sure that "higher is better".
Context 38-39
Imagine, you have a 28 * 28 image and you run a 3 * 3 convolution neural network on it with the input depth of 3 and output depth of 8.
Note: Stride is 1 and you are using same padding.
38) What is the dimension of output feature map when you are using the given parameters.
A) 28 width, 28 height and 8 depth
Solution: (A)
The formula for calculating output size is

output size = (N – F)/S + 1
where, N is input size, F is filter size and S is stride.
Read this article to get a better understanding.

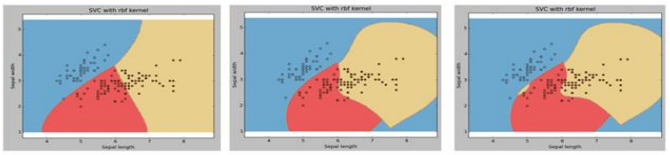
39) What is the dimensions of output feature map when you are using following parameters.

B) 13 width, 13 height and 8 depth

Solution: (B)

Same as above

40) Suppose, we were plotting the visualization for different values of C (Penalty parameter) in SVM algorithm. Due to some reason, we forgot to tag the C values with visualizations. In that case, which of the following option best explains the C values for the images below (1,2,3 left to right, so C values are C1 for image1, C2 for image2 and C3 for image3) in case of rbf kernel.



C) C1 < C2 < C3

Solution: (C)

Penalty parameter C of the error term. It also controls the trade-off between smooth decision boundary and classifying the training points correctly. For large values of C, the optimization will choose a smaller-margin hyperplane. Read more here.

NLP

https://www.analyticsvidhya.com/blog/2017/07/30-questions-test-data-scientist-natural-language-processing-solution-skilltest-nlp/

Skill Test Questions and Answers

Q1 Which of the following techniques can be used for the purpose of keyword normalization, the process of converting a keyword into its base form?

1. Lemmatization
2. Levenshtein
3. Stemming
4. Soundex

C) 1 and 3

Solution: (C)

Lemmatization and stemming are the techniques of keyword normalization, while Levenshtein and Soundex are techniques of string matching.

2) N-grams are defined as the combination of N keywords together. How many bi-grams can be generated from given sentence:

"Analytics Vidhya is a great source to learn data science"

C) 9

Solution: (C)

Bigrams: Analytics Vidhya, Vidhya is, is a, a great, great source, source to, To learn, learn data, data science

3) How many trigrams phrases can be generated from the following sentence, after performing following text cleaning steps:

- Stopword Removal
- Replacing punctuations by a single space

"#Analytics-vidhya is a great source to learn @data_science."

C) 5

Solution: (C)

After performing stopword removal and punctuation replacement the text becomes: "Analytics vidhya great source learn data science"

Trigrams – Analytics vidhya great, vidhya great source, great source learn, source learn data, learn data science

4) Which of the following regular expression can be used to identify date(s) present in the text object:

"The next meetup on data science will be held on 2017-09-21, previously it happened on 31/03, 2016"

D) None of the above

Solution: (D)

None if these expressions would be able to identify the dates in this text object.

Question Context 5-6:

You have collected a data of about 10,000 rows of tweet text and no other information. You want to create a tweet classification model that categorizes each of the tweets in three buckets – positive, negative and neutral.

5) Which of the following models can perform tweet classification with regards to context mentioned above?
C) None of the above
Solution: (C)
Since, you are given only the data of tweets and no other information, which means there is no target variable present. One cannot train a supervised learning model, both svm and naive bayes are supervised learning techniques.

6) You have created a document term matrix of the data, treating every tweet as one document. Which of the following is correct, in regards to document term matrix?

1. Removal of stopwords from the data will affect the dimensionality of data
2. Normalization of words in the data will reduce the dimensionality of data
3. Converting all the words in lowercase will not affect the dimensionality of the data

D) 1 and 2
Solution: (D)
Choices A and B are correct because stopword removal will decrease the number of features in the matrix, normalization of words will also reduce redundant features, and, converting all words to lowercase will also decrease the dimensionality.

7) Which of the following features can be used for accuracy improvement of a classification model?
A) Frequency count of terms
B) Vector Notation of sentence
C) Part of Speech Tag
D) Dependency Grammar
E) All of these
Solution: (E)
All of the techniques can be used for the purpose of engineering features in a model.

8) What percentage of the total statements are correct with regards to Topic Modeling?

1. It is a supervised learning technique
2. LDA (Linear Discriminant Analysis) can be used to perform topic modeling
3. Selection of number of topics in a model does not depend on the size of data
4. Number of topic terms are directly proportional to size of the data

A) 0
Solution: (A)
LDA is unsupervised learning model, LDA is latent Dirichlet allocation, not Linear discriminant analysis. Selection of the number of topics is directly proportional to the size of the data, while number of topic terms is not directly proportional to the size of the data. Hence none of the statements are correct.

9) In Latent Dirichlet Allocation model for text classification purposes, what does alpha and beta hyperparameter represent-
D) Alpha: density of topics generated within documents, beta: density of terms generated within topics True
Solution: (D)
Option D is correct

10) Solve the equation according to the sentence "I am planning to visit New Delhi to attend Analytics Vidhya Delhi Hackathon".
A = (# of words with Noun as the part of speech tag)
B = (# of words with Verb as the part of speech tag)
C = (# of words with frequency count greater than one)
What are the correct values of A, B, and C?
D) 7, 4, 2
Solution: (D)
Nouns: I, New, Delhi, Analytics, Vidhya, Delhi, Hackathon (7)
Verbs: am, planning, visit, attend (4)
Words with frequency counts > 1: to, Delhi (2)
Hence option D is correct.

11) In a corpus of N documents, one document is randomly picked. The document contains a total of T terms and the term "data" appears K times.
What is the correct value for the product of TF (term frequency) and IDF (inverse-document-frequency), if the term "data" appears in approximately one-third of the total documents?
B) K * Log(3) / T
Solution: (B)
formula for TF is K/T
formula for IDF is log(total docs / no of docs containing "data")
= log(1 / (⅓))
= log (3)
Hence correct choice is Klog(3)/T

Question Context 12 to 14:

Refer the following document term matrix

Term	Document						
	d1	*d2*	*d3*	*d4*	*d5*	*d6*	*d7*
t1	2	1	0	0	0	0	0
t2	1	2	0	0	0	0	1
t3	3	1	0	0	1	1	0
t4	0	0	1	2	1	1	1
t5	0	0	1	1	1	1	1
t6	0	0	1	1	0	0	0

12) Which of the following documents contains the same number of terms and the number of terms in the one of the document is not equal to least number of terms in any document in the entire corpus.

C) d2 and d4

Solution: (C)

Both of the documents d2 and d4 contains 4 terms and does not contain the least number of terms which is 3.

13) Which are the most common and the rarest term of the corpus?

A) t4, t6

Solution: (A)

T5 is most common terms across 5 out of 7 documents, T6 is rare term only appears in d3 and d4

14) What is the term frequency of a term which is used a maximum number of times in that document?

B) t3 – 3/6

Solution: (B)

t3 is used max times in entire corpus = 3, tf for t3 is 3/6

15) Which of the following technique is not a part of flexible text matching?

D) Keyword Hashing

Solution: (D)

Except Keyword Hashing all other (soundex,metaphone,editdistance) are the techniques used in flexible string matching

16) True or False: Word2Vec model is a machine learning model used to create vector notations of text objects. Word2vec contains multiple deep neural networks

B) FALSE

Solution: (B)

Word2vec also contains preprocessing model which is not a deep neural network

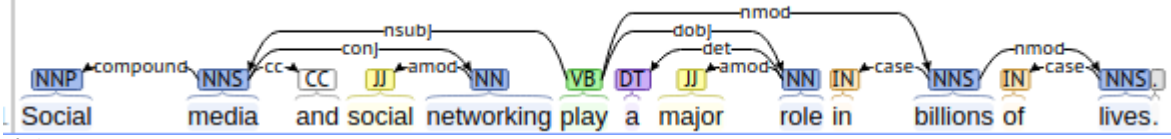
17) Which of the following statement is(are) true for Word2Vec model?

C) Skip-gram is a deep neural network model

Solution: (C)

Word2vec contains the Continuous bag of words and skip-gram models, which are deep neural nets.

18) With respect to this context-free dependency graphs, how many sub-trees exists in the sentence?



D) 6

Solution: (D)

Subtrees in the dependency graph can be viewed as nodes having an outward link, for example:
Media, networking, play, role, billions, and lives are the roots of subtrees

19) What is the right order for a text classification model components

1. Text cleaning
2. Text annotation
3. Gradient descent
4. Model tuning
5. Text to predictors

C) 12534

Solution: (C)

A right text classification model contains – cleaning of text to remove noise, annotation to create more features, converting text-based features into predictors, learning a model using gradient descent and finally tuning a model.

20) Polysemy is defined as the coexistence of multiple meanings for a word or phrase in a text object. Which of the following models is likely the best choice to correct this problem?

B) Convolutional Neural Networks
Solution: (B)
CNNs are popular choice for text classification problems because they take into consideration left and right contexts of the words as features which can solve the problem of polysemy

21) Which of the following models can be used for the purpose of document similarity?
A) Training a word 2 vector model on the corpus that learns context present in the document
B) Training a bag of words model that learns occurrence of words in the document
C) Creating a document-term matrix and using cosine similarity for each document
D) All of the above
Solution: (D)
word2vec model can be used for measuring document similarity based on context. Bag Of Words and document term matrix can be used for measuring similarity based on terms.

22) What are the possible features of a text corpus

1. Count of word in a document
2. Boolean feature – presence of word in a document
3. Vector notation of word
4. Part of Speech Tag
5. Basic Dependency Grammar
6. Entire document as a feature

E) 12345
Solution: (E)
Except for entire document as the feature, rest all can be used as features of text classification learning model.

23) While creating a machine learning model on text data, you created a document term matrix of the input data of 100K documents. Which of the following remedies can be used to reduce the dimensions of data –

1. Latent Dirichlet Allocation
2. Latent Semantic Indexing
3. Keyword Normalization

D) 1, 2, 3
Solution: (D)
All of the techniques can be used to reduce the dimensions of the data.

24) Google Search's feature – "Did you mean", is a mixture of different techniques. Which of the following techniques are likely to be ingredients?

1. Collaborative Filtering model to detect similar user behaviors (queries)
2. Model that checks for Levenshtein distance among the dictionary terms
3. Translation of sentences into multiple languages

C) 1, 2
Solution: (C)
Collaborative filtering can be used to check what are the patterns used by people, Levenshtein is used to measure the distance among dictionary terms.

25) While working with text data obtained from news sentences, which are structured in nature, which of the grammar-based text parsing techniques can be used for noun phrase detection, verb phrase detection, subject detection and object detection.
B) Dependency Parsing and Constituency Parsing
Solution: (B)
Dependency and constituent parsing extract these relations from the text

26) Social Media platforms are the most intuitive form of text data. You are given a corpus of complete social media data of tweets. How can you create a model that suggests the hashtags?
A) Perform Topic Models to obtain most significant words of the corpus
B) Train a Bag of Ngrams model to capture top n-grams – words and their combinations
C) Train a word2vector model to learn repeating contexts in the sentences
D) All of these
Solution: (D)
All of the techniques can be used to extract most significant terms of a corpus.

27) While working with context extraction from a text data, you encountered two different sentences: The tank is full of soldiers. The tank is full of nitrogen. Which of the following measures can be used to remove the problem of word sense disambiguation in the sentences?
A) Compare the dictionary definition of an ambiguous word with the terms contained in its neighborhood
Solution: (A)
Option 1 is called Lesk algorithm, used for word sense disambiguation, rest others cannot be used.

28) Collaborative Filtering and Content Based Models are the two popular recommendation engines, what role does NLP play in building such algorithms.

A) Feature Extraction from text
B) Measuring Feature Similarity
C) Engineering Features for vector space learning model
D) All of these
Solution: (D)
NLP can be used anywhere where text data is involved – feature extraction, measuring feature similarity, create vector features of the text.

29) Retrieval based models and Generative models are the two popular techniques used for building chatbots. Which of the following is an example of retrieval model and generative model respectively.
B) Rule-based learning and Sequence to Sequence model
Solution: (B)
choice 2 best explains examples of retrieval based models and generative models

30) What is the major difference between CRF (Conditional Random Field) and HMM (Hidden Markov Model)?
B) CRF is Discriminative whereas HMM is Generative model
Solution: (B)
Option B is correct

Tree based models

https://www.analyticsvidhya.com/blog/2017/09/30-questions-test-tree-based-models/

Skill test Questions and Answers

1) Which of the following is/are true about bagging trees?
1. In bagging trees, individual trees are independent of each other
2. Bagging is the method for improving the performance by aggregating the results of weak learners

C) 1 and 2
Solution: C
Both options are true. In Bagging, each individual trees are independent of each other because they consider different subset of features and samples.

2) Which of the following is/are true about boosting trees?
1. In boosting trees, individual weak learners are independent of each other
2. It is the method for improving the performance by aggregating the results of weak learners

A) 1
B) 2
C) 1 and 2
D) None of these
Solution: B
In boosting tree individual weak learners are not independent of each other because each tree correct the results of previous tree. Bagging and boosting both can be consider as improving the base learners results.

3) Which of the following is/are true about Random Forest and Gradient Boosting ensemble methods?
1. Both methods can be used for classification task
4 Both methods can be used for regression task

E) 1 and 4
Solution: E
Both algorithms are design for classification as well as regression task.

4) In Random forest you can generate hundreds of trees (say T1, T2 …..Tn) and then aggregate the results of these tree. Which of the following is true about individual(Tk) tree in Random Forest?
1. Individual tree is built on a subset of the feature
2. Individual tree is built on a subset of observations

A) 1 and 2
Solution: A
Random forest is based on bagging concept, that consider faction of sample and faction of feature for building the individual trees.

5) Which of the following is true about "max_depth" hyperparameter in Gradient Boosting?
1Lower is better parameter in case of same validation accuracy
3 Increase the value of max_depth may overfit the data

A) 1 and 3
Solution: A
Increase the depth from the certain value of depth may overfit the data and for 2 depth values validation accuracies are same we always prefer the small depth in final model building.

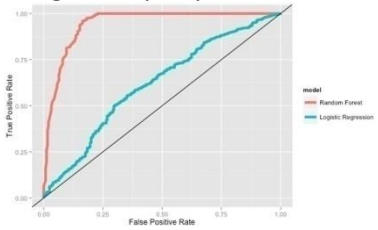
6) Which of the following algorithm doesn't uses learning Rate as of one of its hyperparameter?
2Extra Trees
4Random Forest

D) 2 and 4
Solution: D

Random Forest and Extra Trees don't have learning rate as a hyperparameter.

7) Which of the following algorithm would you take into the consideration in your final model building on the basis of performance?

Suppose you have given the following graph which shows the ROC curve for two different classification algorithms such as Random Forest(Red) and Logistic Regression(Blue)



A) Random Forest

Solution: A

Since, Random forest has largest AUC given in the picture so I would prefer Random Forest

8) Which of the following is true about training and testing error in such case?

Suppose you want to apply AdaBoost algorithm on Data D which has T observations. You set half the data for training and half for testing initially. Now you want to increase the number of data points for training T1, T2 … Tn where T1 < T2…. Tn-1 < Tn

B) The difference between training error and test error decreases as number of observations increases

Solution: B

As we have more and more data, training error increases and testing error de-creases. And they all converge to the true error.

9) In random forest or gradient boosting algorithms, features can be of any type. For example, it can be a continuous feature or a categorical feature. Which of the following option is true when you consider these types of features?

A) Only Random forest algorithm handles real valued attributes by discretizing them

B) Only Gradient boosting algorithm handles real valued attributes by discretizing them

C) Both algorithms can handle real valued attributes by discretizing them

Solution: C

Both can handle real valued features.

10) Which of the following algorithm are not an example of ensemble learning algorithm?

E) Decision Trees

Solution: E

Decision trees doesn't aggregate the results of multiple trees so it is not an ensemble algorithm.

11) Suppose you are using a bagging based algorithm say a RandomForest in model building. Which of the following can be true?

1. Number of tree should be as large as possible
2. You will have interpretability after using RandomForest

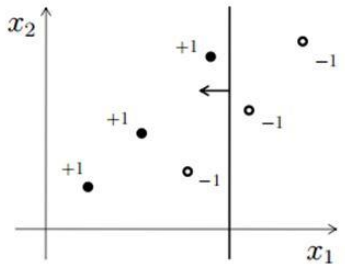
A) 1

Solution: A

Since Random Forest aggregate the result of different weak learners, If It is possible we would want more number of trees in model building. Random Forest is a black box model you will lose interpretability after using it.

Context 12-15

Consider the following figure for answering the next few questions. In the figure, X1 and X2 are the two features and the data point is represented by dots (-1 is negative class and +1 is a positive class). And you first split the data based on feature X1(say splitting point is x11) which is shown in the figure using vertical line. Every value less than x11 will be predicted as positive class and greater than x will be predicted as negative class.



12) How many data points are misclassified in above image?

A) 1

Solution: A

Only one observation is misclassified, one negative class is showing at the left side of vertical line which will be predicting as a positive class.

13) Which of the following splitting point on feature x1 will classify the data correctly?

A) Greater than x11
B) Less than x11
C) Equal to x11
D) None of above

Solution: D

If you search any point on X1 you won't find any point that gives 100% accuracy.

14) If you consider only feature X2 for splitting. Can you now perfectly separate the positive class from negative class for any one split on X2?

B) No

Solution: B

It is also not possible.

15) Now consider only one splitting on both (one on X1 and one on X2) feature. You can split both features at any point. Would you be able to classify all data points correctly?

B) FALSE

Solution: B

You won't find such case because you can get minimum 1 misclassification.

Context 16-17

Suppose, you are working on a binary classification problem with 3 input features. And you chose to apply a bagging algorithm(X) on this data. You chose max_features = 2 and the n_estimators =3. Now, Think that each estimators have 70% accuracy.

Note: Algorithm X is aggregating the results of individual estimators based on maximum voting

16) What will be the maximum accuracy you can get?

D) 100%

Solution: D

Refer below table for models M1, M2 and M3.

Actual predictions	M1	M2	M3	Output
1	1	0	1	1
1	1	0	1	1
1	1	0	1	1
1	0	1	1	1
1	0	1	1	1
1	0	1	1	1
1	1	1	1	1
1	1	1	0	1
1	1	1	0	1
1	1	1	0	1

17) What will be the minimum accuracy you can get?

C) It can be less than 70%

Solution: C

Refer below table for models M1, M2 and M3.

Actual predictions	M1	M2	M3	Output
1	1	0	0	0
1	1	1	1	1
1	1	0	0	0
1	0	1	0	0

1	0	1	1	1
1	0	0	1	0
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1

18) Suppose you are building random forest model, which split a node on the attribute, that has highest information gain. In the below image, select the attribute which has the highest information gain?

A) Outlook

Solution: A

Information gain increases with the average purity of subsets. So option A would be the right answer.

19) Which of the following is true about the Gradient Boosting trees?

1. In each stage, introduce a new regression tree to compensate the shortcomings of existing model
2. We can use gradient decent method for minimize the loss function

C) 1 and 2

Solution: C

Both are true and self explanatory

20) True-False: The bagging is suitable for high variance low bias models?

A) TRUE

Solution: A

The bagging is suitable for high variance low bias models or you can say for complex models.

21) Which of the following is true when you choose fraction of observations for building the base learners in tree based algorithm?

A) Decrease the fraction of samples to build a base learners will result in decrease in variance

Solution: A

Answer is self explanatory

Context 22-23

Suppose, you are building a Gradient Boosting model on data, which has millions of observations and 1000's of features. Before building the model you want to consider the difference parameter setting for time measurement.

22) Consider the hyperparameter "number of trees" and arrange the options in terms of time taken by each hyperparameter for building the Gradient Boosting model?

Note: remaining hyperparameters are same

1. Number of trees = 100
2. Number of trees = 500
3. Number of trees = 1000

B) 1<2<3

Solution: B

The time taken by building 1000 trees is maximum and time taken by building the 100 trees is minimum which is given in solution B

23) Now, Consider the learning rate hyperparameter and arrange the options in terms of time taken by each hyperparameter for building the Gradient boosting model?

Note: Remaining hyperparameters are same

1. learning rate = 1
2. learning rate = 2
3. learning rate = 3

A) 1~2~3

Solution: A

Since learning rate doesn't affect time so all learning rates would take equal time.

24) In greadient boosting it is important use learning rate to get optimum output. Which of the following is true abut choosing the learning rate?

C) Learning Rate should be low but it should not be very low

Solution: C

Learning rate should be low but it should not be very low otherwise algorithm will take so long to finish the training because you need to increase the number trees.

25) [True or False] Cross validation can be used to select the number of iterations in boosting; this procedure may help reduce overfitting.

A) TRUE

Solution: A

26) When you use the boosting algorithm you always consider the weak learners. Which of the following is the main reason for having weak learners?

1. To prevent overfitting
2. To prevent under fitting

A) 1

Solution: A

To prevent overfitting, since the complexity of the overall learner increases at each step. Starting with weak learners implies the final classifier will be less likely to overfit.

27) To apply bagging to regression trees which of the following is/are true in such case?

1. We build the N regression with N bootstrap sample
2. We take the average the of N regression tree
3. Each tree has a high variance with low bia

D) 1,2 and 3

Solution: D

All of the options are correct and self explanatory

28) How to select best hyperparameters in tree based models?

B) Measure performance over validation data

Solution: B

We always consider the validation results to compare with the test result.

29) In which of the following scenario a gain ratio is preferred over Information Gain?

A) When a categorical variable has very large number of category

Solution: A

When high cardinality problems, gain ratio is preferred over Information Gain technique.

30) Suppose you have given the following scenario for training and validation error for Gradient Boosting. Which of the following hyper parameter would you choose in such case?

Scenario	Depth	Training Error	Validation Error
1	2	100	110
2	4	90	105
3	6	50	100
4	8	45	105
5	10	30	150

B) 2

Solution: B

Scenario 2 and 4 has same validation accuracies but we would select 2 because depth is lower is better hyper parameter.

KNN

https://www.analyticsvidhya.com/blog/2017/09/30-questions-test-k-nearest-neighbors-algorithm/

Skill test Questions and Answers

1) [True or False] k-NN algorithm does more computation on test time rather than train time.

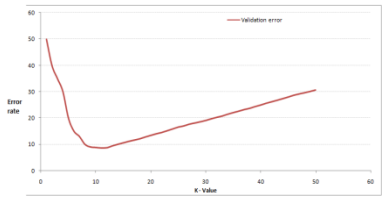
A) TRUE

Solution: A

The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples.

In the testing phase, a test point is classified by assigning the label which are most frequent among the *k* training samples nearest to that query point – hence higher computation.

2) In the image below, which would be the best value for k assuming that the algorithm you are using is k-Nearest Neighbor.



B) 10
Solution: B
Validation error is the least when the value of k is 10. So it is best to use this value of k
3) Which of the following distance metric can not be used in k-NN?
A) ManhattanB) MinkowskiC) TanimotoD) JaccardE) MahalanobisF) All can be used
Solution: F
All of these distance metric can be used as a distance metric for k-NN.
4) Which of the following option is true about k-NN algorithm?
C) It can be used in both classification and regression
Solution: C
We can also use k-NN for regression problems. In this case the prediction can be based on the mean or the median of the k-most similar instances.
5) Which of the following statement is true about k-NN algorithm?

1. k-NN performs much better if all of the data have the same scale
2. k-NN works well with a small number of input variables (p), but struggles when the number of inputs is very large
3. k-NN makes no assumptions about the functional form of the problem being solved

D) All of the above
Solution: D
The above mentioned statements are assumptions of kNN algorithm

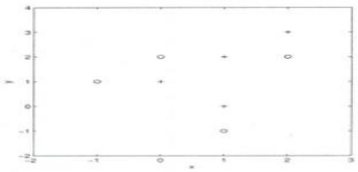
6) Which of the following machine learning algorithm can be used for imputing missing values of both categorical and continuous variables?
A) K-NN
Solution: A
k-NN algorithm can be used for imputing missing value of both categorical and continuous variables.
7) Which of the following is true about Manhattan distance?
A) It can be used for continuous variables
Solution: A
Manhattan Distance is designed for calculating the distance between real valued features.
8) Which of the following distance measure do we use in case of categorical variables in k-NN?

1. Hamming Distance

A) 1
Solution: A
Both Euclidean and Manhattan distances are used in case of continuous variables, whereas hamming distance is used in case of categorical variable.
9) Which of the following will be Euclidean Distance between the two data point A(1,3) and B(2,3)?
A) 1
Solution: A
sqrt((1-2)^2 + (3-3)^2) = sqrt(1^2 + 0^2) = 1
10) Which of the following will be Manhattan Distance between the two data point A(1,3) and B(2,3)?
A) 1
Solution: A
sqrt(mod((1-2)) + mod((3-3))) = sqrt(1 + 0) = 1
Context: 11-12
Suppose, you have given the following data where x and y are the 2 input variables and Class is the dependent variable.

x	y	Class
−1	1	−
0	1	+
0	2	−
1	−1	−
1	0	+
1	2	+
2	2	−
2	3	+

Below is a scatter plot which shows the above data in 2D space.



11) Suppose, you want to predict the class of new data point x=1 and y=1 using eucludian distance in 3-NN. In which class this data point belong to?

A) + Class

Solution: A

All three nearest point are of +class so this point will be classified as +class.

12) In the previous question, you are now want use 7-NN instead of 3-KNN which of the following x=1 and y=1 will belong to?

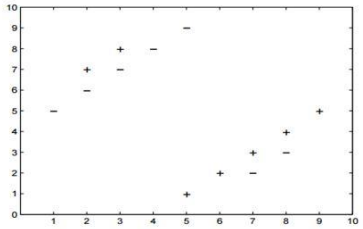
B) – Class

Solution: B

Now this point will be classified as – class because there are 4 – class and 3 +class point are in nearest circle.

Context 13-14:

Suppose you have given the following 2-class data where "+" represent a postive class and "" is represent negative class.



13) Which of the following value of k in k-NN would minimize the leave one out cross validation accuracy?

B) 5

Solution: B

5-NN will have least leave one out cross validation error.

14) Which of the following would be the leave on out cross validation accuracy for k=5?

E) None of the above

Solution: E

In 5-NN we will have 10/14 leave one out cross validation accuracy.

15) Which of the following will be true about k in k-NN in terms of Bias?

A) When you increase the k the bias will be increases

Solution: A

large K means simple model, simple model always condider as high bias

16) Which of the following will be true about k in k-NN in terms of variance?

B) When you decrease the k the variance will increases

Solution: B

Simple model will be consider as less variance model

17) The following two distances(Eucludean Distance and Manhattan Distance) have given to you which generally we used in K-NN algorithm. These distance are between two points A(x1,y1) and B(x2,Y2).
Your task is to tag the both distance by seeing the following two graphs. Which of the following option is true about below graph ?

B) Left is Euclidean Distance and right is Manhattan Distance

Solution: B

Left is the graphical depiction of how euclidean distance works, whereas right one is of Manhattan distance.

18) When you find noise in data which of the following option would you consider in k-NN?

A) I will increase the value of k

Solution: A

To be more sure of which classifications you make, you can try increasing the value of k.

19) In k-NN it is very likely to overfit due to the curse of dimensionality. Which of the following option would you consider to handle such problem?

1. Dimensionality Reduction
2. Feature selection

C) 1 and 2

Solution: C

In such case you can use either dimensionality reduction algorithm or the feature selection algorithm

20) Below are two statements given. Which of the following will be true both statements?

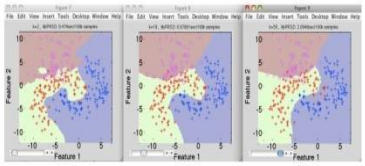
1. k-NN is a memory-based approach is that the classifier immediately adapts as we collect new training data.
2. The computational complexity for classifying new samples grows linearly with the number of samples in the training dataset in the worst-case scenario.

C) 1 and 2

Solution: C

Both are true and self explanatory

21) Suppose you have given the following images(1 left, 2 middle and 3 right), Now your task is to find out the value of k in k-NN in each image where k1 is for 1st, k2 is for 2nd and k3 is for 3rd figure.



D) None of these

Solution: D

Value of k is highest in k3, whereas in k1 it is lowest

22) Which of the following value of k in the following graph would you give least leave one out cross validation accuracy?

B) 2

Solution: B

If you keep the value of k as 2, it gives the lowest cross validation accuracy. You can try this out yourself.

23) A company has build a kNN classifier that gets 100% accuracy on training data. When they deployed this model on client side it has been found that the model is not at all accurate. Which of the following thing might gone wrong?

Note: Model has successfully deployed and no technical issues are found at client side except the model performance

A) It is probably a overfitted model

Solution: A

In an overfitted module, it seems to be performing well on training data, but it is not generalized enough to give the same results on a new data.

24) You have given the following 2 statements, find which of these option is/are true in case of k-NN?

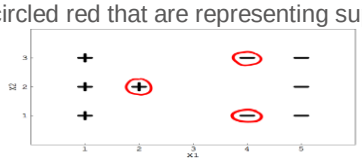
1. In case of very large value of k, we may include points from other classes into the neighborhood.
2. In case of too small value of k the algorithm is very sensitive to noise

C) 1 and 2
Solution: C
Both the options are true and are self explanatory.
25) Which of the following statements is true for k-NN classifiers?
A) The classification accuracy is better with larger values of k
B) The decision boundary is smoother with smaller values of k
C) The decision boundary is linear
D) k-NN does not require an explicit training step
Solution: D
Option A: This is not always true. You have to ensure that the value of k is not too high or not too low.
Option B: This statement is not true. The decision boundary can be a bit jagged
Option C: Same as option B
Option D: This statement is true
26) True-False: It is possible to construct a 2-NN classifier by using the 1-NN classifier?
A) TRUE
Solution: A
You can implement a 2-NN classifier by ensembling 1-NN classifiers
27) In k-NN what will happen when you increase/decrease the value of k?
A) The boundary becomes smoother with increasing value of K
Solution: A
The decision boundary would become smoother by increasing the value of K
28) Following are the two statements given for k-NN algorthm, which of the statement(s) is/are true?

1. We can choose optimal value of k with the help of cross validation
2. Euclidean distance treats each feature as equally important

C) 1 and 2
Solution: C
Both the statements are true
Context 29-30:
Suppose, you have trained a k-NN model and now you want to get the prediction on test data. Before getting the prediction suppose you want to calculate the time taken by k-NN for predicting the class for test data.
Note: Calculating the distance between 2 observation will take D time.
29) What would be the time taken by 1-NN if there are N(Very large) observations in test data?
A) N*D
Solution: A
The value of N is very large, so option A is correct
30) What would be the relation between the time taken by 1-NN,2-NN,3-NN.
C) 1-NN ~ 2-NN ~ 3-NN
Solution: C
The training time for any value of k in kNN algorithm is the same.

Svm
https://www.analyticsvidhya.com/blog/2017/10/svm-skilltest/
Question Context: 1 – 2
Suppose you are using a Linear SVM classifier with 2 class classification problem. Now you have been given the following data in which some points are circled red that are representing support vectors.



1) If you remove the following any one red points from the data. Does the decision boundary will change?
A) Yes
Solution: A
These three examples are positioned such that removing any one of them introduces slack in the constraints. So the decision boundary would completely change.
2) [True or False] If you remove the non-red circled points from the data, the decision boundary will change?
B) False
Solution: B
On the other hand, rest of the points in the data won't affect the decision boundary much.
3) What do you mean by generalization error in terms of the SVM?

B) How accurately the SVM can predict outcomes for unseen data

Solution: B

Generalisation error in statistics is generally the out-of-sample error which is the measure of how accurately a model can predict values for previously unseen data.

4) When the C parameter is set to infinite, which of the following holds true?

A) The optimal hyperplane if exists, will be the one that completely separates the data

Solution: A

At such a high level of misclassification penalty, soft margin will not hold existence as there will be no room for error.

5) What do you mean by a hard margin?

A) The SVM allows very low error in classification

Solution: A

A hard margin means that an SVM is very rigid in classification and tries to work extremely well in the training set, causing overfitting.

6) The minimum time complexity for training an SVM is O(n2). According to this fact, what sizes of datasets are not best suited for SVM's?

A) Large datasets

Solution: A

Datasets which have a clear classification boundary will function best with SVM's.

7) The effectiveness of an SVM depends upon:

A) Selection of Kernel B) Kernel Parameters C) Soft Margin Parameter C D) All of the above

Solution: D

The SVM effectiveness depends upon how you choose the basic 3 requirements mentioned above in such a way that it maximises your efficiency, reduces error and overfitting.

8) Support vectors are the data points that lie closest to the decision surface.

A) TRUE

Solution: A

They are the points closest to the hyperplane and the hardest ones to classify. They also have a direct bearing on the location of the decision surface.

9) The SVM's are less effective when:

C) The data is noisy and contains overlapping points

Solution: C

When the data has noise and overlapping points, there is a problem in drawing a clear hyperplane without misclassifying.

10) Suppose you are using RBF kernel in SVM with high Gamma value. What does this signify?

B) The model would consider only the points close to the hyperplane for modeling

Solution: B

The gamma parameter in SVM tuning signifies the influence of points either near or far away from the hyperplane.
For a low gamma, the model will be too constrained and include all points of the training dataset, without really capturing the shape.
For a higher gamma, the model will capture the shape of the dataset well.

11) The cost parameter in the SVM means:

C) The tradeoff between misclassification and simplicity of the model

Solution: C

The cost parameter decides how much an SVM should be allowed to "bend" with the data. For a low cost, you aim for a smooth decision surface and for a higher cost, you aim to classify more points correctly. It is also simply referred to as the cost of misclassification.

12) Suppose you are building a SVM model on data X. The data X can be error prone which means that you should not trust any specific data point too much. Now think that you want to build a SVM model which has quadratic kernel function of polynomial degree 2 that uses Slack variable C as one of it's hyper parameter. Based upon that give the answer for following question.

What would happen when you use very large value of C(C->infinity)?
Note: For small C was also classifying all data points correctly

A) We can still classify data correctly for given setting of hyper parameter C

Solution: A

For large values of C, the penalty for misclassifying points is very high, so the decision boundary will perfectly separate the data if possible.

13) What would happen when you use very small C (C~0)?

A) Misclassification would happen

Solution: A

The classifier can maximize the margin between most of the points, while misclassifying a few points, because the penalty is so low.

14) If I am using all features of my dataset and I achieve 100% accuracy on my training set, but ~70% on validation set, what should I look out for?
C) Overfitting
Solution: C
If we're achieving 100% training accuracy very easily, we need to check to verify if we're overfitting our data.
15) Which of the following are real world applications of the SVM?
A) Text and Hypertext Categorization
B) Image Classification
C) Clustering of News Articles
D) All of the above
Solution: D
SVM's are highly versatile models that can be used for practically all real world problems ranging from regression to clustering and handwriting recognitions.
Question Context: 16 – 18
Suppose you have trained an SVM with linear decision boundary after training SVM, you correctly infer that your SVM model is under fitting.
16) Which of the following option would you more likely to consider iterating SVM next time?
C) You will try to calculate more variables
Solution: C
The best option here would be to create more features for the model.
17) Suppose you gave the correct answer in previous question. What do you think that is actually happening?
1. We are lowering the bias
2. We are lowering the variance
3. We are increasing the bias
4. We are increasing the variance
C) 1 and 4
Solution: C
Better model will lower the bias and increase the variance

18) In above question suppose you want to change one of it's(SVM) hyperparameter so that effect would be same as previous questions i.e model will not under fit?
A) We will increase the parameter C
Solution: A
Increasing C parameter would be the right thing to do here, as it will ensure regularized model
19) We usually use feature normalization before using the Gaussian kernel in SVM. What is true about feature normalization?
1. We do feature normalization so that new feature will dominate other
2. Some times, feature normalization is not feasible in case of categorical variables
3. Feature normalization always helps when we use Gaussian kernel in SVM
B) 1 and 2
Solution: B
Statements one and two are correct.
Question Context: 20-22
Suppose you are dealing with 4 class classification problem and you want to train a SVM model on the data for that you are using One-vs-all method. Now answer the below questions?
20) How many times we need to train our SVM model in such case?
D) 4
Solution: D
For a 4 class problem, you would have to train the SVM at least 4 times if you are using a one-vs-all method.
21) Suppose you have same distribution of classes in the data. Now, say for training 1 time in one vs all setting the SVM is taking 10 second. How many seconds would it require to train one-vs-all method end to end?
B) 40
Solution: B
It would take 10×4 = 40 seconds
22) Suppose your problem has changed now. Now, data has only 2 classes. What would you think how many times we need to train SVM in such case?
A) 1
Solution: A
Training the SVM only one time would give you appropriate results
Question context: 23 – 24

Suppose you are using SVM with linear kernel of polynomial degree 2, Now think that you have applied this on data and found that it perfectly fit the data that means, Training and testing accuracy is 100%.

23) Now, think that you increase the complexity(or degree of polynomial of this kernel). What would you think will happen?

A) Increasing the complexity will overfit the data

Solution: A

Increasing the complexity of the data would make the algorithm overfit the data.

24) In the previous question after increasing the complexity you found that training accuracy was still 100%. According to you what is the reason behind that?

1. Since data is fixed and we are fitting more polynomial term or parameters so the algorithm starts memorizing everything in the data
2. Since data is fixed and SVM doesn't need to search in big hypothesis space

C) 1 and 2

Solution: C

Both the given statements are correct.

25) What is/are true about kernel in SVM?

1. Kernel function map low dimensional data to high dimensional space
2. It's a similarity function

C) 1 and 2

Solution: C

Both the given statements are correct.

Dimensionality Reduction techniques

https://www.analyticsvidhya.com/blog/2017/03/questions-dimensionality-reduction-data-scientist/

Questions & Answers

1) Imagine, you have 1000 input features and 1 target feature in a machine learning problem. You have to select 100 most important features based on the relationship between input features and the target features. Do you think, this is an example of dimensionality reduction?

A. Yes

Solution: **(A)**

2) [True or False] It is not necessary to have a target variable for applying dimensionality reduction algorithms.

A. TRUE

Solution: **(A)**

LDA is an example of supervised dimensionality reduction algorithm.

3) I have 4 variables in the dataset such as – A, B, C & D. I have performed the following actions:

Step 1: Using the above variables, I have created two more variables, namely E = A + 3 * B and F = B + 5 * C + D.

Step 2: Then using only the variables E and F I have built a Random Forest model.

Could the steps performed above represent a dimensionality reduction method?

A. True

Solution: **(A)**

Yes, Because Step 1 could be used to represent the data into 2 lower dimensions.

4) Which of the following techniques would perform better for reducing dimensions of a data set?

A. Removing columns which have too many missing values

Solution: **(A)**

If a columns have too many missing values, (say 99%) then we can remove such columns.

5) [True or False] Dimensionality reduction algorithms are one of the possible ways to reduce the computation time required to build a model.

A. TRUE

Solution: **(A)**

Reducing the dimension of data will take less time to train a model.

6) Which of the following algorithms cannot be used for reducing the dimensionality of data?

A. t-SNE

B. PCA

C. LDA False

D. None of these

Solution: **(D)**

All of the algorithms are the example of dimensionality reduction algorithm.

7) [True or False] PCA can be used for projecting and visualizing data in lower dimensions.

A. TRUE

Solution: **(A)**

Sometimes it is very useful to plot the data in lower dimensions. We can take the first 2 principal components and then visualize the data using scatter plot.

8) The most popularly used dimensionality reduction algorithm is Principal Component Analysis (PCA). Which of the following is/are true about PCA?

1. PCA is an unsupervised method
2. It searches for the directions that data have the largest variance
3. Maximum number of principal components <= number of features
4. All principal components are orthogonal to each other

F. All of the above

Solution: **(F)**

All options are self explanatory.

9) Suppose we are using dimensionality reduction as pre-processing technique, i.e, instead of using all the features, we reduce the data to k dimensions with PCA. And then use these PCA projections as our features. Which of the following statement is correct?

B. Higher 'k' means less regularization

Solution: **(B)**

Higher k would lead to less smoothening as we would be able to preserve more characteristics in data, hence less regularization.

10) In which of the following scenarios is t-SNE better to use than PCA for dimensionality reduction while working on a local machine with minimal computational power?

C. Dataset with 10,000 entries and 8 features

Solution: **(C)**

t-SNE has quadratic time and space complexity. Thus it is a very heavy algorithm in terms of system resource utilization.

11) Which of the following statement is true for a t-SNE cost function?

B. It is symmetric in nature.

Solution: **(B)**

Cost function of SNE is asymmetric in nature. Which makes it difficult to converge using gradient decent. A symmetric cost function is one of the major differences between SNE and t-SNE.

Question 12

Imagine you are dealing with text data. To represent the words you are using word embedding (Word2vec). In word embedding, you will end up with 1000 dimensions. Now, you want to reduce the dimensionality of this high dimensional data such that, similar words should have a similar meaning in nearest neighbor space.In such case, which of the following algorithm are you most likely choose?

A. t-SNE

Solution: **(A)**

t-SNE stands for t-Distributed Stochastic Neighbor Embedding which consider the nearest neighbours for reducing the data.

13) [True or False] t-SNE learns non-parametric mapping.

A. TRUE

Solution: **(A)**

t-SNE learns a non-parametric mapping, which means that it does not learn an explicit function that maps data from the input space to the map. For more information read from this link.

14) Which of the following statement is correct for t-SNE and PCA?

D. t-SNE is nonlinear whereas PCA is linear

Solution: **(D)**

Option D is correct. Read the explanation from this link

15) In t-SNE algorithm, which of the following hyper parameters can be tuned?

A. Number of dimensions

B. Smooth measure of effective number of neighbours

C. Maximum number of iterations

D. All of the above

Solution: **(D)**

All of the hyper-parameters in the option can tuned.

16) What is of the following statement is true about t-SNE in comparison to PCA?

A. When the data is huge (in size), t-SNE may fail to produce better results.

Solution: **(A)**

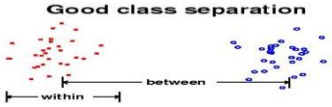
Option A is correct

17) Xi and Xj are two distinct points in the higher dimension representation, where as Yi & Yj are the representations of Xi and Xj in a lower dimension.

1. The similarity of datapoint Xi to datapoint Xj is the conditional probability p (j|i) .

2. The similarity of datapoint Yi to datapoint Yj is the conditional probability q (j|i) .
Which of the following must be true for perfect representation of xi and xj in lower dimensional space?
C. p (j|i) = q (j|i)
Solution: **(C)**
The conditional probabilities for similarity of two points must be equal because similarity between the points must remain unchanged in both higher and lower dimension for them to be perfect representations.

18) Which of the following is true about LDA?



A. LDA aims to maximize the distance between class and minimize the within class distance
Solution: **(A)**
Option A is correct.

19) In which of the following case LDA will fail?
A. If the discriminatory information is not in the mean but in the variance of the data
Solution: **(A)**
Option A is correct

20) Which of the following comparison(s) are true about PCA and LDA?

1. Both LDA and PCA are linear transformation techniques
2. LDA is supervised whereas PCA is unsupervised
3. PCA maximize the variance of the data, whereas LDA maximize the separation between different classes,

E. 1, 2 and 3
Solution: **(E)**
All of the options are correct

21) What will happen when eigenvalues are roughly equal?
B. PCA will perform badly
Solution: **(B)**
When all eigen vectors are same in such case you won't be able to select the principal components because in that case all principal components are equal.

22) PCA works better if there is?

1. A linear structure in the data
2. If the data lies on a curved surface and not on a flat surface
3. If variables are scaled in the same unit

C. 1 and 3
Solution: **(C)**
Option C is correct

23) What happens when you get features in lower dimensions using PCA?

1. The features will still have interpretability
2. The features will lose interpretability
3. The features must carry all information present in data
4. The features may not carry all information present in data

D. 2 and 4
Solution: **(D)**
When you get the features in lower dimensions then you will lose some information of data most of the times and you won't be able to interpret the lower dimension data.

24) Imagine, you are given the following scatterplot between height and weight.

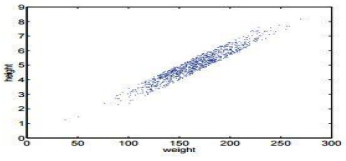


Figure 1: Height vs weight.

Select the angle which will capture maximum variability along a single axis?
B. ~ 45 degree
Solution: **(B)**
Option B has largest possible variance in data.

25) Which of the following option(s) is / are true?

1. You need to initialize parameters in PCA
2. You don't need to initialize parameters in PCA

3. PCA can be trapped into local minima problem
4. PCA can't be trapped into local minima problem

D. 2 and 4
Solution: **(D)**
PCA is a deterministic algorithm which doesn't have parameters to initialize and it doesn't have local minima problem like most of the machine learning algorithms has.

Question Context 26
The below snapshot shows the scatter plot of two features (X1 and X2) with the class information (Red, Blue). You can also see the direction of PCA and LDA.

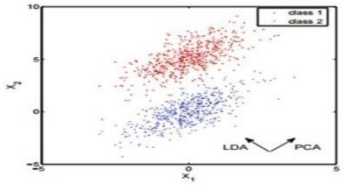


Figure 3: PCA vs LDA.

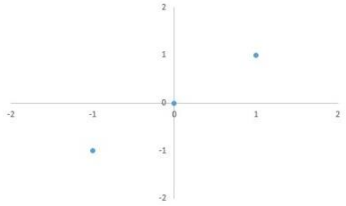
26) Which of the following method would result into better class prediction?
B. Building a classification algorithm with LDA
Solution: **(B)**
If our goal is to classify these points, PCA projection does only more harm than good—the majority of blue and red points would land overlapped on the first principal component.hence PCA would confuse the classifier.

27) Which of the following options are correct, when you are applying PCA on a image dataset?
1.
1. It can be used to effectively detect deformable objects.
2. It is invariant to affine transforms.
3. It can be used for lossy image compression.
4. It is not invariant to shadows.

C. 3 and 4
Solution: **(C)**
Option C is correct

28) Under which condition SVD and PCA produce the same projection result?
B. When data has zero mean
Solution: **(B)**
When the data has a zero mean vector, otherwise you have to center the data first before taking SVD.

Question Context 29
Consider 3 data points in the 2-d space: (-1, -1), (0,0), (1,1).



29) What will be the first principal component for this data?
1. [√ 2 /2 , √ 2/ 2]
2. (1/ √ 3, 1/ √ 3)
3. ([-√ 2/ 2 , √ 2/ 2])
4. (- 1/ √ 3, – 1/ √ 3)

C. 1 and 3
Solution: **(C)**
The first principal component is v = [√ 2 /2 , √ 2/ 2] T (you shouldn't really need to solve any SVD or eigenproblem to see this). Note that the principal component should be normalized to have unit length. (The negation v = [− √ 2/ 2 , − √ 2/ 2] T is also correct.)

30) If we project the original data points into the 1-d subspace by the principal component [√ 2 /2, √ 2 /2] T. What are their coordinates in the 1-d subspace?
A. (− √ 2), (0), (√ 2)
Solution: **(A)**

The coordinates of three points after projection should be z1 = x T 1 v = [−1, −1][√ 2/ 2 , √ 2 /2] T = − √ 2, z2 = x T 2 v = 0, z3 = x T 3 v = √ 2.

31) For the projected data you just obtained projections ((− √ 2), (0), (√ 2)). Now if we represent them in the original 2-d space and consider them as the reconstruction of the original data points, what is the reconstruction error? Context: 29-31:

A. 0%

Solution: **(A)**

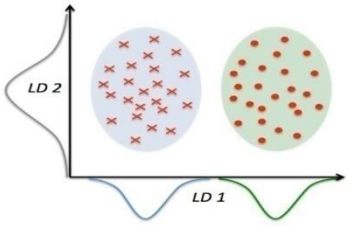
The reconstruction error is 0, since all three points are perfectly located on the direction of the first principal component. Or, you can actually calculate the reconstruction: z1 ·v.

xˆ1 = − √ 2·[√ 2/ 2 , √ 2/2] T = [−1, −1]T

xˆ2 = 0*[0, 0]T = [0,0] xˆ3 = √ 2* [1, 1]T = [1,1]

which are exactly x1, x2, x3.

32) In LDA, the idea is to find the line that best separates the two classes. In the given image which of the following is a good projection?



A. LD1

Solution: **(A)**

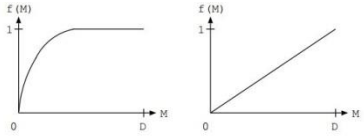
LD1 Is a good projection because it best separates the class.

Question Context 33

PCA is a good technique to try, because it is simple to understand and is commonly used to reduce the dimensionality of the data. Obtain the eigenvalues λ1 ≥ λ2 ≥ • • • ≥ λN and plot.

$$f(M) = \sum_{i=1}^{M} \lambda_i / \sum_{i=1}^{N} \lambda_i$$

To see how f(M) increases with M and takes maximum value 1 at M = D. We have two graph given below:



33) Which of the above graph shows better performance of PCA? Where M is first M principal components and D is total number of features?

A. Left

Solution: **(A)**

PCA is good if f(M) asymptotes rapidly to 1. This happens if the first eigenvalues are big and the remainder are small. PCA is bad if all the eigenvalues are roughly equal. See examples of both cases in figure.

34) Which of the following option is true?

A. LDA explicitly attempts to model the difference between the classes of data. PCA on the other hand does not take into account any difference in class.

Solution: **(A)**

Options are self explanatory.

35) Which of the following can be the first 2 principal components after applying PCA?

1. (0.5, 0.5, 0.5, 0.5) and (0.71, 0.71, 0, 0)
2. (0.5, 0.5, 0.5, 0.5) and (0, 0, -0.71, -0.71)
3. (0.5, 0.5, 0.5, 0.5) and (0.5, 0.5, -0.5, -0.5)
4. (0.5, 0.5, 0.5, 0.5) and (-0.5, -0.5, 0.5, 0.5)

D. 3 and 4

Solution: **(D)**

For the first two choices, the two loading vectors are not orthogonal.

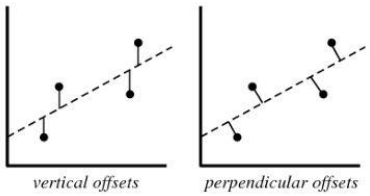
36) Which of the following gives the difference(s) between the logistic regression and LDA?

1. If the classes are well separated, the parameter estimates for logistic regression can be unstable.
2. If the sample size is small and distribution of features are normal for each class. In such case, linear discriminant analysis is more stable than logistic regression.

C. 1 and 2

Solution: **(C)**

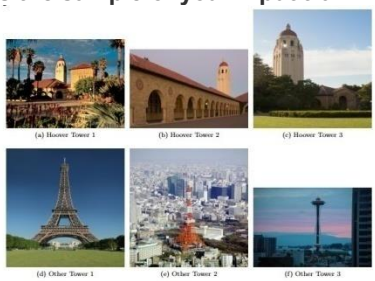
Refer this video



37) Which of the following offset, do we consider in PCA?
B. Perpendicular offset
Solution: **(B)**
We always consider residual as vertical offsets. Perpendicular offset are useful in case of PCA

38) Imagine you are dealing with 10 class classification problem and you want to know that at most how many discriminant vectors can be produced by LDA. What is the correct answer?
B. 9
Solution: **(B)**
LDA produces at most c − 1 discriminant vectors. You may refer this link for more information.

Question Context 39
The given dataset consists of images of "Hoover Tower" and some other towers. Now, you want to use PCA (Eigenface) and the nearest neighbour method to build a classifier that predicts whether new image depicts "Hoover tower" or not. The figure gives the sample of your input training images.

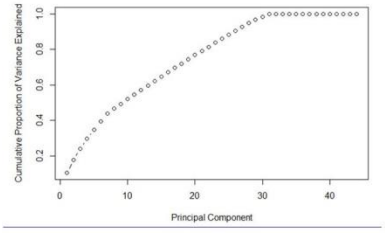


39) In order to get reasonable performance from the "Eigenface" algorithm, what pre-processing steps will be required on these images?

1. Align the towers in the same position in the image.
2. Scale or crop all images to the same size.

C. 1 and 2
Solution: **(C)**
Both the statements are correct.

40) What are the optimum number of principle components in the below figure ?



B. 30
Solution: **(B)**
We can see in the above figure that the number of components = 30 is giving highest variance with lowest number of components. Hence option 'B' is the right answer.

Clustering Techniques

https://www.analyticsvidhya.com/blog/2017/02/test-data-scientist-clustering/

Questions & Answers

Q1. Movie Recommendation systems are an example of:

1. **Classification**
2. **Clustering**
3. **Reinforcement Learning**
4. **Regression**

E. 2 and 3

Solution: (E)

Generally, movie recommendation systems cluster the users in a finite number of similar groups based on their previous activities and profile. Then, at a fundamental level, people in the same cluster are made similar recommendations.

In some scenarios, this can also be approached as a classification problem for assigning the most appropriate movie class to the user of a specific group of users. Also, a movie recommendation system can be viewed as a reinforcement learning problem where it learns by its previous recommendations and improves the future recommendations.

Q2. Sentiment Analysis is an example of:

1. **Regression**
2. **Classification**
3. **Clustering**
4. **Reinforcement Learning**

Options:

E. 1, 2 and 4

Solution: (E)

Sentiment analysis at the fundamental level is the task of classifying the sentiments represented in an image, text or speech into a set of defined sentiment classes like happy, sad, excited, positive, negative, etc. It can also be viewed as a regression problem for assigning a sentiment score of say 1 to 10 for a corresponding image, text or speech.

Another way of looking at sentiment analysis is to consider it using a reinforcement learning perspective where the algorithm constantly learns from the accuracy of past sentiment analysis performed to improve the future performance.

Q3. Can decision trees be used for performing clustering?

A. True

Solution: (A)

Decision trees can also be used to for clusters in the data but clustering often generates natural clusters and is not dependent on any objective function.

Q4. Which of the following is the most appropriate strategy for data cleaning before performing clustering analysis, given less than desirable number of data points:

1. **Capping and flouring of variables**
2. **Removal of outliers**

Options:

A. 1 only

Solution: (A)

Removal of outliers is not recommended if the data points are few in number. In this scenario, capping and flouring of variables is the most appropriate strategy.

Q5. What is the minimum no. of variables/ features required to perform clustering?

B. 1

Solution: (B)

At least a single variable is required to perform clustering analysis. Clustering analysis with a single variable can be visualized with the help of a histogram.

Q6. For two runs of K-Mean clustering is it expected to get same clustering results?

B. No

Solution: (B)

K-Means clustering algorithm instead converses on local minima which might also correspond to the global minima in some cases but not always. Therefore, it's advised to run the K-Means algorithm multiple times before drawing inferences about the clusters.

However, note that it's possible to receive same clustering results from K-means by setting the same seed value for each run. But that is done by simply making the algorithm choose the set of same random no. for each run.

Q7. Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means

A. Yes

Solution: (A)

When the K-Means algorithm has reached the local or global minima, it will not alter the assignment of data points to clusters for two successive iterations.

Q8. Which of the following can act as possible termination conditions in K-Means?

1. **For a fixed number of iterations.**
2. **Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum.**
3. **Centroids do not change between successive iterations.**

4. **Terminate when RSS falls below a threshold.**

Options:

D. All of the above

Solution: (D)

All four conditions can be used as possible termination condition in K-Means clustering:

1. This condition limits the runtime of the clustering algorithm, but in some cases the quality of the clustering will be poor because of an insufficient number of iterations.
2. Except for cases with a bad local minimum, this produces a good clustering, but runtimes may be unacceptably long.
3. This also ensures that the algorithm has converged at the minima.
4. Terminate when RSS falls below a threshold. This criterion ensures that the clustering is of a desired quality after termination. Practically, it's a good practice to combine it with a bound on the number of iterations to guarantee termination.

Q9. Which of the following clustering algorithms suffers from the problem of convergence at local optima?

1. **K- Means clustering algorithm**
2. **Agglomerative clustering algorithm**
3. **Expectation-Maximization clustering algorithm**
4. **Diverse clustering algorithm**

Options:

D. 1 and 3

Solution: (D)

Out of the options given, only K-Means clustering algorithm and EM clustering algorithm has the drawback of converging at local minima.

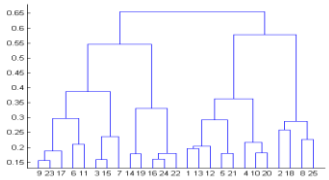
Q10. Which of the following algorithm is most sensitive to outliers?

A. K-means clustering algorithm

Solution: (A)

Out of all the options, K-Means clustering algorithm is most sensitive to outliers as it uses the mean of cluster data points to find the cluster center.

Q11. After performing K-Means Clustering analysis on a dataset, you observed the following dendrogram. Which of the following conclusion can be drawn from the dendrogram?



D. The above dendrogram interpretation is not possible for K-Means clustering analysis

Solution: (D)

A dendrogram is not possible for K-Means clustering analysis. However, one can create a cluster gram based on K-Means clustering analysis.

Q12. How can Clustering (Unsupervised Learning) be used to improve the accuracy of Linear Regression model (Supervised Learning):

1. **Creating different models for different cluster groups.**
2. **Creating an input feature for cluster ids as an ordinal variable.**
3. **Creating an input feature for cluster centroids as a continuous variable.**
4. **Creating an input feature for cluster size as a continuous variable.**

Options:

F. All of the above

Solution: (F)

Creating an input feature for cluster ids as ordinal variable or creating an input feature for cluster centroids as a continuous variable might not convey any relevant information to the regression model for multidimensional data. But for clustering in a single dimension, all of the given methods are expected to convey meaningful information to the regression model. For example, to cluster people in two groups based on their hair length, storing clustering ID as ordinal variable and cluster centroids as continuous variables will convey meaningful information.

Q13. What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithm for the same dataset?

A. Proximity function used

B. of data points used

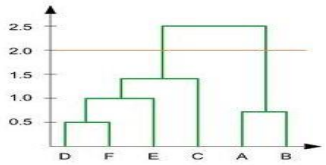
C. of variables used

D. B and c only

E. All of the above

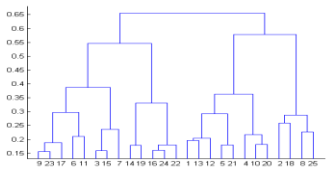
Solution: (E)
Change in either of Proximity function, no. of data points or no. of variables will lead to different clustering results and hence different dendrograms.

Q14. In the figure below, if you draw a horizontal line on y-axis for y=2. What will be the number of clusters formed?

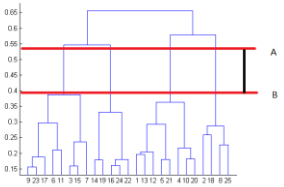


B. 2
Solution: (B)
Since the number of vertical lines intersecting the red horizontal line at y=2 in the dendrogram are 2, therefore, two clusters will be formed.

Q15. What is the most appropriate no. of clusters for the data points represented by the following dendrogram:



B. 4
Solution: (B)
The decision of the no. of clusters that can best depict different groups can be chosen by observing the dendrogram. The best choice of the no. of clusters is the no. of vertical lines in the dendrogram cut by a horizontal line that can transverse the maximum distance vertically without intersecting a cluster.

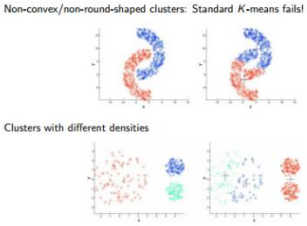


In the above example, the best choice of no. of clusters will be 4 as the red horizontal line in the dendrogram below covers maximum vertical distance AB.

Q16. In which of the following cases will K-Means clustering fail to give good results?

1. **Data points with outliers**
2. **Data points with different densities**
3. **Data points with round shapes**
4. **Data points with non-convex shapes**

Options:
D. 1, 2 and 4
Solution: (D)
K-Means clustering algorithm fails to give good results when the data contains outliers, the density spread of data points across the data space is different and the data points follow non-convex shapes.



Q17. Which of the following metrics, do we have for finding dissimilarity between two clusters in hierarchical clustering?

1. **Single-link**
2. **Complete-link**
3. **Average-link**

Options:

D. 1, 2 and 3

Solution: (D)

All of the three methods i.e. single link, complete link and average link can be used for finding dissimilarity between two clusters in hierarchical clustering.

Q18. Which of the following are true?

1. **Clustering analysis is negatively affected by multicollinearity of features**
2. **Clustering analysis is negatively affected by heteroscedasticity**

Options:

A. 1 only

Solution: (A)

Clustering analysis is not negatively affected by heteroscedasticity but the results are negatively impacted by multicollinearity of features/ variables used in clustering as the correlated feature/ variable will carry extra weight on the distance calculation than desired.

Q19. Given, six points with the following attributes:

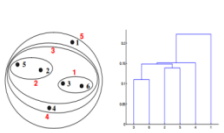
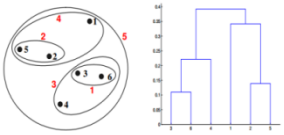
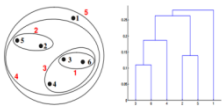
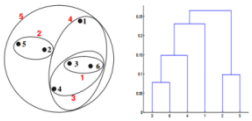
point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

Table : X-Y coordinates of six points.

	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

Which of the following clustering representations and dendrogram depicts the use of MIN or Single link proximity function in hierarchical clustering:

A.  B.  C.  D. 

Solution: (A)

For the single link or MIN version of hierarchical clustering, the proximity of two clusters is defined to be the minimum of the distance between any two points in the different clusters. For instance, from the table, we see that the distance between points 3 and 6 is 0.11, and that is the height at which they are joined into one cluster in the dendrogram. As another example, the distance between clusters {3, 6} and {2, 5} is given by dist({3, 6}, {2, 5}) = min(dist(3, 2), dist(6, 2), dist(3, 5), dist(6, 5)) = min(0.1483, 0.2540, 0.2843, 0.3921) = 0.1483.

Q20 Given, six points with the following attributes:

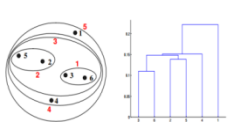
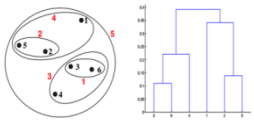
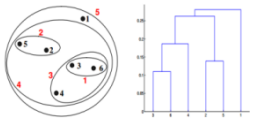
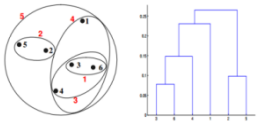
point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

Table : X-Y coordinates of six points.

	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

Which of the following clustering representations and dendrogram depicts the use of MAX or Complete link proximity function in hierarchical clustering:

A.  B.  C.  D. 

Solution: (B)

For the single link or MAX version of hierarchical clustering, the proximity of two clusters is defined to be the maximum of the distance between any two points in the different clusters. Similarly, here points 3 and 6 are merged first. However, {3, 6} is merged with {4}, instead of {2, 5}. This is because the dist({3, 6}, {4}) = max(dist(3, 4), dist(6, 4)) = max(0.1513, 0.2216) = 0.2216, which is smaller than dist({3, 6}, {2, 5}) = max(dist(3, 2), dist(6, 2), dist(3, 5), dist(6, 5)) = max(0.1483, 0.2540, 0.2843, 0.3921) = 0.3921 and dist({3, 6}, {1}) = max(dist(3, 1), dist(6, 1)) = max(0.2218, 0.2347) = 0.2347.

Q21 Given, six points with the following attributes:

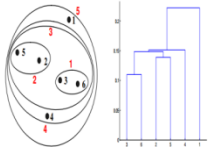
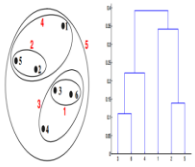
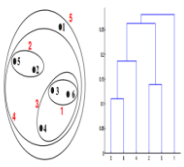
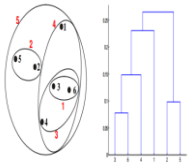
point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

Table : X-Y coordinates of six points.

	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

Which of the following clustering representations and dendrogram depicts the use of Group average proximity function in hierarchical clustering:

A.  B.  C.  D. 

Solution: (C)

For the group average version of hierarchical clustering, the proximity of two clusters is defined to be the average of the pairwise proximities between all pairs of points in the different clusters. This is an intermediate approach between MIN and MAX. This is expressed by the following equation:

$$proximity(cluster_1, cluster_2) = \sum_{\substack{p_1 \in cluster_1 \\ p_2 \in cluster_2}} \frac{proximity(p_1, p_2)}{size(cluster_1) * size(cluster_2)}$$

Here, the distance between some clusters. dist({3, 6, 4}, {1}) = (0.2218 + 0.3688 + 0.2347)/(3 ∗ 1) = 0.2751. dist({2, 5}, {1}) = (0.2357 + 0.3421)/(2 ∗ 1) = 0.2889. dist({3, 6, 4}, {2, 5}) = (0.1483 + 0.2843 + 0.2540 + 0.3921 + 0.2042 + 0.2932)/(6∗1) = 0.2637. Because dist({3, 6, 4}, {2, 5}) is smaller than dist({3, 6, 4}, {1}) and dist({2, 5}, {1}), these two clusters are merged at the fourth stage

Q22. Given, six points with the following attributes:

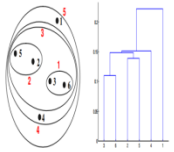
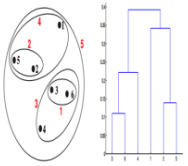
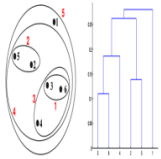
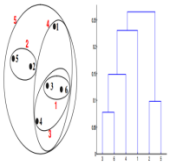
point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

Table : X-Y coordinates of six points.

	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

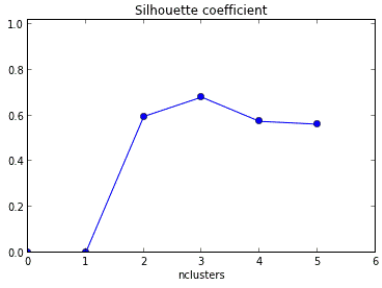
Which of the following clustering representations and dendrogram depicts the use of Ward's method proximity function in hierarchical clustering:

A.  B.  C.  D. 

Solution: (D)

Ward method is a centroid method. Centroid method calculates the proximity between two clusters by calculating the distance between the centroids of clusters. For Ward's method, the proximity between two clusters is defined as the increase in the squared error that results when two clusters are merged. The results of applying Ward's method to the sample data set of six points. The resulting clustering is somewhat different from those produced by MIN, MAX, and group average.

Q23. What should be the best choice of no. of clusters based on the following results:



C. 3

Solution: (C)

The silhouette coefficient is a measure of how similar an object is to its own cluster compared to other clusters. Number of clusters for which silhouette coefficient is highest represents the best choice of the number of clusters.

Q24. Which of the following is/are valid iterative strategy for treating missing values before clustering analysis?

A. Imputation with mean

B. Nearest Neighbor assignment

C. Imputation with Expectation Maximization algorithm
D. All of the above
Solution: (C)
All of the mentioned techniques are valid for treating missing values before clustering analysis but only imputation with EM algorithm is iterative in its functioning.

Q25. K-Mean algorithm has some limitations. One of the limitation it has is, it makes hard assignments(A point either completely belongs to a cluster or not belongs at all) of points to clusters.
Note: Soft assignment can be consider as the probability of being assigned to each cluster: say K = 3 and for some point xn, p1 = 0.7, p2 = 0.2, p3 = 0.1)
Which of the following algorithm(s) allows soft assignments?

1. **Gaussian mixture models**
2. **Fuzzy K-means**

Options:
A. 1 only
B. 2 only
C. 1 and 2
D. None of these
Solution: (C)
Both, Gaussian mixture models and Fuzzy K-means allows soft assignments.

Q26. Assume, you want to cluster 7 observations into 3 clusters using K-Means clustering algorithm. After first iteration clusters, C1, C2, C3 has following observations:
C1: {(2,2), (4,4), (6,6)}
C2: {(0,4), (4,0)}
C3: {(5,5), (9,9)}
What will be the cluster centroids if you want to proceed for second iteration?
A. C1: (4,4), C2: (2,2), C3: (7,7)
B. C1: (6,6), C2: (4,4), C3: (9,9)
C. C1: (2,2), C2: (0,0), C3: (5,5)
D. None of these
Solution: (A)
Finding centroid for data points in cluster C1 = ((2+4+6)/3, (2+4+6)/3) = (4, 4)
Finding centroid for data points in cluster C2 = ((0+4)/2, (4+0)/2) = (2, 2)
Finding centroid for data points in cluster C3 = ((5+9)/2, (5+9)/2) = (7, 7)
Hence, C1: (4,4), C2: (2,2), C3: (7,7)

Q27. Assume, you want to cluster 7 observations into 3 clusters using K-Means clustering algorithm. After first iteration clusters, C1, C2, C3 has following observations:
C1: {(2,2), (4,4), (6,6)}
C2: {(0,4), (4,0)}
C3: {(5,5), (9,9)}
What will be the Manhattan distance for observation (9, 9) from cluster centroid C1. In second iteration.
A. 10
B. 5*sqrt(2)
C. 13*sqrt(2)
D. None of these
Solution: (A)
Manhattan distance between centroid C1 i.e. (4, 4) and (9, 9) = (9-4) + (9-4) = 10

Q28. If two variables V1 and V2, are used for clustering. Which of the following are true for K means clustering with k =3?

1. **If V1 and V2 has a correlation of 1, the cluster centroids will be in a straight line**
2. **If V1 and V2 has a correlation of 0, the cluster centroids will be in straight line**

Options:
A. 1 only
B. 2 only
C. 1 and 2
D. None of the above
Solution: (A)
If the correlation between the variables V1 and V2 is 1, then all the data points will be in a straight line. Hence, all the three cluster centroids will form a straight line as well.

Q29. Feature scaling is an important step before applying K-Mean algorithm. What is reason behind this?
A. In distance calculation it will give the same weights for all features
Solution; (A)
Feature scaling ensures that all the features get same weight in the clustering analysis. Consider a scenario of clustering people based on their weights (in KG) with range 55-110 and height (in inches) with range 5.6 to 6.4. In this case, the clusters produced without scaling can be very misleading as the range of weight is much higher than that of height. Therefore, its necessary to bring them to same scale so that they have equal weightage on the clustering result.

Q30. Which of the following method is used for finding optimal of cluster in K-Mean algorithm?
A. Elbow method
Solution: (A)
Out of the given options, only elbow method is used for finding the optimal number of clusters. The elbow method looks at the percentage of variance explained as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data.

Q31. What is true about K-Mean Clustering?

1. **K-means is extremely sensitive to cluster center initializations**
2. **Bad initialization can lead to Poor convergence speed**
3. **Bad initialization can lead to bad overall clustering**

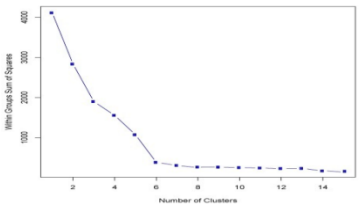
Options:
D. 1, 2 and 3
Solution: (D)
All three of the given statements are true. K-means is extremely sensitive to cluster center initialization. Also, bad initialization can lead to Poor convergence speed as well as bad overall clustering.

Q32. Which of the following can be applied to get good results for K-means algorithm corresponding to global minima?

1. **Try to run algorithm for different centroid initialization**
2. **Adjust number of iterations**
3. **Find out the optimal number of clusters**

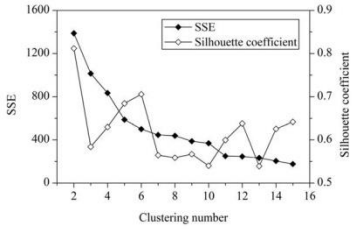
Options:
D. All of above
Solution: (D)
All of these are standard practices that are used in order to obtain good clustering results.

Q33. What should be the best choice for number of clusters based on the following results:



B. 6
Solution: (B)
Based on the above results, the best choice of number of clusters using elbow method is 6.

Q34. What should be the best choice for number of clusters based on the following results:



A. 2
B. 4
C. 6
D. 8
Solution: (C)

Generally, a higher average silhouette coefficient indicates better clustering quality. In this plot, the optimal clustering number of grid cells in the study area should be 2, at which the value of the average silhouette coefficient is highest. However, the SSE of this clustering solution (k = 2) is too large. At k = 6, the SSE is much lower. In addition, the value of the average silhouette coefficient at k = 6 is also very high, which is just lower than k = 2. Thus, the best choice is k = 6.

Q35. Which of the following sequences is correct for a K-Means algorithm using Forgy method of initialization?

1. **Specify the number of clusters**
2. **Assign cluster centroids randomly**
3. **Assign each data point to the nearest cluster centroid**
4. **Re-assign each point to nearest cluster centroids**
5. **Re-compute cluster centroids**

Options:

A. 1, 2, 3, 5, 4

Solution: (A)

The methods used for initialization in K means are Forgy and Random Partition. The Forgy method randomly chooses k observations from the data set and uses these as the initial means. The Random Partition method first randomly assigns a cluster to each observation and then proceeds to the update step, thus computing the initial mean to be the centroid of the cluster's randomly assigned points.

Q36. If you are using Multinomial mixture models with the expectation-maximization algorithm for clustering a set of data points into two clusters, which of the assumptions are important:

C. All the data points follow two multinomial distribution

Solution: (C)

In EM algorithm for clustering its essential to choose the same no. of clusters to classify the data points into as the no. of different distributions they are expected to be generated from and also the distributions must be of the same type.

Q37. Which of the following is/are not true about Centroid based K-Means clustering algorithm and Distribution based expectation-maximization clustering algorithm:

1. **Both starts with random initializations**
2. **Both are iterative algorithms**
3. **Both have strong assumptions that the data points must fulfill**
4. **Both are sensitive to outliers**
5. **Expectation maximization algorithm is a special case of K-Means**
6. **Both requires prior knowledge of the no. of desired clusters**
7. **The results produced by both are non-reproducible.**

Options:

B. 5 only

Solution: (B)

All of the above statements are true except the 5th as instead K-Means is a special case of EM algorithm in which only the centroids of the cluster distributions are calculated at each iteration.

Q38. Which of the following is/are not true about DBSCAN clustering algorithm:

1. **For data points to be in a cluster, they must be in a distance threshold to a core point**
2. **It has strong assumptions for the distribution of data points in** dataspace
3. **It has substantially high time complexity of order O(n^3)**
4. **It does not require prior knowledge of the no. of desired clusters**
5. **It is robust to outliers**

Options:

A. 1 only
B. 2 only
C. 4 only
D. 2 and 3
E. 1 and 5
F. 1, 3 and 5

Solution: (D)

- DBSCAN can form a cluster of any arbitrary shape and does not have strong assumptions for the distribution of data points in the dataspace.
- DBSCAN has a low time complexity of order O(n log n) only.

Q39. Which of the following are the high and low bounds for the existence of F-Score?

A. [0,1]
B. (0,1)
C. [-1,1]

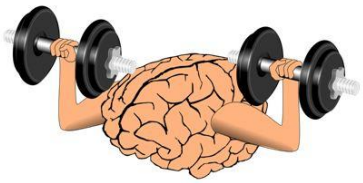
D. None of the above
Solution: (A)
The lowest and highest possible values of F score are 0 and 1 with 1 representing that every data point is assigned to the correct cluster and 0 representing that the precession and/ or recall of the clustering analysis are both 0. In clustering analysis, high value of F score is desired.

Q40. Following are the results observed for clustering 6000 data points into 3 clusters: A, B and C:

		Actual			
		A	B	C	SUM
Predicted	A	600	400	200	1200
	B	1000	1200	200	2400
	C	400	400	1600	2400
	SUM	2000	2000	2000	

What is the F_1-Score with respect to cluster B?
A. 3
B. 4
C. 5
D. 6
Solution: (D)
Here,
True Positive, TP = 1200
True Negative, TN = 600 + 1600 = 2200
False Positive, FP = 1000 + 200 = 1200
False Negative, FN = 400 + 400 = 800
Therefore,
Precision = TP / (TP + FP) = 0.5
Recall = TP / (TP + FN) = 0.6
Hence,
F_1 = 2 * (Precision * Recall)/ (Precision + recall) = 0.54 ~ 0.5

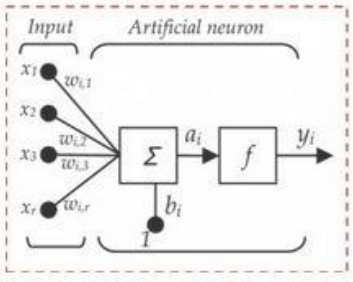
Questions and Answers
Q1. A neural network model is said to be inspired from the human brain.



The neural network consists of many neurons, each neuron takes an input, processes it and gives an output. Here's a diagrammatic representation of a real neuron.

Which of the following statement(s) correctly represents a real neuron?
A. A neuron has a single input and a single output only
B. A neuron has multiple inputs but a single output only
C. A neuron has a single input but multiple outputs
D. A neuron has multiple inputs and multiple outputs
E. All of the above statements are valid
Solution: (E)
A neuron can have a single Input / Output or multiple Inputs / Outputs.

Q2. Below is a mathematical representation of a neuron.



The different components of the neuron are denoted as:

- **x1, x2,…, xN: These are inputs to the neuron. These can either be the actual observations from input layer or an intermediate value from one of the hidden layers.**
- **w1, w2,…,wN: The Weight of each input.**
- **bi: Is termed as Bias units. These are constant values added to the input of the activation function corresponding to each weight. It works similar to an intercept term.**
- **a: Is termed as the activation of the neuron which can be represented as**
- **and y: is the output of the neuron**

$$a = f(\sum_{i=0}^{N} w_i x_i)$$

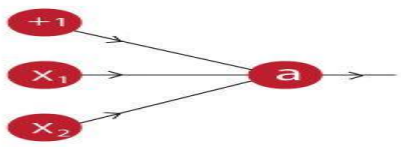
Considering the above notations, will a line equation (y = mx + c) fall into the category of a neuron?
A. Yes
B. No
Solution: (A)
A single neuron with no non-linearity can be considered as a linear regression function.

Q3. Let us assume we implement an AND function to a single neuron. Below is a tabular representation of an AND function:

X1	X2	X1 AND X2
0	0	0
0	1	0
1	0	0
1	1	1

The activation function of our neuron is denoted as:

$$f(x) = \begin{cases} 0, & for\ x < 0 \\ 1, & for\ x \geq 0 \end{cases}$$

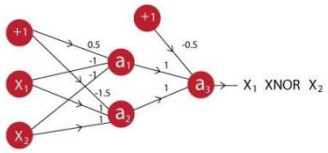


What would be the weights and bias?
(Hint: For which values of w1, w2 and b does our neuron implement an AND function?)
A. Bias = -1.5, w1 = 1, w2 = 1
B. Bias = 1.5, w1 = 2, w2 = 2
C. Bias = 1, w1 = 1.5, w2 = 1.5
D. None of these
Solution: (A)
A.

1. f(-1.5*1 + 1*0 + 1*0) = f(-1.5) = 0
2. f(-1.5*1 + 1*0 + 1*1) = f(-0.5) = 0
3. f(-1.5*1 + 1*1 + 1*0) = f(-0.5) = 0
4. f(-1.5*1 + 1*1+ 1*1) = f(0.5) = 1

Therefore option A is correct

Q4. A network is created when we multiple neurons stack together. Let us take an example of a neural network simulating an XNOR function.



You can see that the last neuron takes input from two neurons before it. The activation function for all the neurons is given by:

$$f(x) = \begin{cases} 0, & for\ x < 0 \\ 1, & for\ x \geq 0 \end{cases}$$

Suppose X1 is 0 and X2 is 1, what will be the output for the above neural network?

A. 0
B. 1

Solution: (A)

Output of a1: f(0.5*1 + -1*0 + -1*1) = f(-0.5) = 0
Output of a2: f(-1.5*1 + 1*0 + 1*1) = f(-0.5) = 0
Output of a3: f(-0.5*1 + 1*0 + 1*0) = f(-0.5) = 0
So the correct answer is A

Q5. In a neural network, knowing the weight and bias of each neuron is the most important step. If you can somehow get the correct value of weight and bias for each neuron, you can approximate any function. What would be the best way to approach this?

A. Assign random values and pray to God they are correct
B. Search every possible combination of weights and biases till you get the best value
C. Iteratively check that after assigning a value how far you are from the best values, and slightly change the assigned values values to make them better
D. None of these

Solution: (C)

Option C is the description of gradient descent.

Q6. What are the steps for using a gradient descent algorithm?

1. **Calculate error between the actual value and the predicted value**
2. **Reiterate until you find the best weights of network**
3. **Pass an input through the network and get values from output layer**
4. **Initialize random weight and bias**
5. **Go to each neurons which contributes to the error and change its respective values to reduce the error**

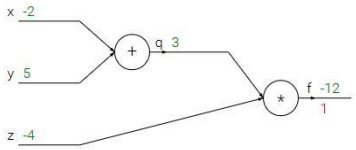
A. 1, 2, 3, 4, 5
B. 5, 4, 3, 2, 1
C. 3, 2, 1, 5, 4
D. 4, 3, 1, 5, 2

Solution: (D)

Option D is correct

Q7. Suppose you have inputs as x, y, and z with values -2, 5, and -4 respectively. You have a neuron 'q' and neuron 'f' with functions:

q = x + y
f = q * z
Graphical representation of the functions is as follows:



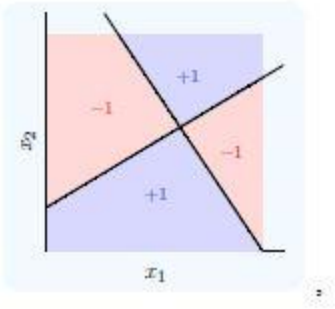
What is the gradient of F with respect to x, y, and z?
(HINT: To calculate gradient, you must find (df/dx), (df/dy) and (df/dz))
A. (-3,4,4)
B. (4,4,3)
C. (-4,-4,3)
D. (3,-4,-4)
Solution: (C)
Option C is correct.

Q8. Now let's revise the previous slides. We have learned that:

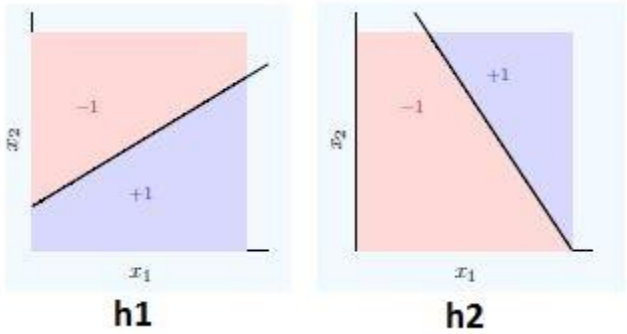
- **A neural network is a (crude) mathematical representation of a brain, which consists of smaller components called neurons.**
- **Each neuron has an input, a processing function, and an output.**
- **These neurons are stacked together to form a network, which can be used to approximate any function.**
- **To get the best possible neural network, we can use techniques like gradient descent to update our neural network model.**

Given above is a description of a neural network. When does a neural network model become a deep learning model?
A. When you add more hidden layers and increase depth of neural network
B. When there is higher dimensionality of data
C. When the problem is an image recognition problem
D. None of these
Solution: (A)
More depth means the network is deeper. There is no strict rule of how many layers are necessary to make a model deep, but still if there are more than 2 hidden layers, the model is said to be deep.

Q9. A neural network can be considered as multiple simple equations stacked together. Suppose we want to replicate the function for the below mentioned decision boundary.



Using two simple inputs h1 and h2



What will be the final equation?

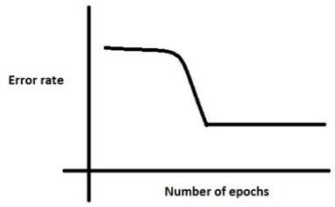
A. (h1 AND NOT h2) OR (NOT h1 AND h2)
B. (h1 OR NOT h2) AND (NOT h1 OR h2)
C. (h1 AND h2) OR (h1 OR h2)
D. None of these
Solution: (A)
As you can see, combining h1 and h2 in an intelligent way can get you a complex equation easily. Refer Chapter 9 of this book

Q10. "Convolutional Neural Networks can perform various types of transformation (rotations or scaling) in an input". Is the statement correct True or False?
A. True
B. False
Solution: (B)
Data Preprocessing steps (viz rotation, scaling) is necessary before you give the data to neural network because neural network cannot do it itself.

Q11. Which of the following techniques perform similar operations as dropout in a neural network?
A. Bagging
B. Boosting
C. Stacking
D. None of these
Solution: (A)
Dropout can be seen as an extreme form of bagging in which each model is trained on a single case and each parameter of the model is very strongly regularized by sharing it with the corresponding parameter in all the other models. Refer here

Q 12. Which of the following gives non-linearity to a neural network?
A. Stochastic Gradient Descent
B. Rectified Linear Unit
C. Convolution function
D. None of the above
Solution: (B)
Rectified Linear unit is a non-linear activation function.

Q13. In training a neural network, you notice that the loss does not decrease in the few starting epochs.



The reasons for this could be:
1. **The learning is rate is low**
2. **Regularization parameter is high**
3. **Stuck at local minima**

What according to you are the probable reasons?
A. 1 and 2
B. 2 and 3
C. 1 and 3
D. Any of these
Solution: (D)
The problem can occur due to any of the reasons mentioned.
Q14. Which of the following is true about model capacity (where model capacity means the ability of neural network to approximate complex functions) ?
A. As number of hidden layers increase, model capacity increases
Solution: (A)
Only option A is correct.

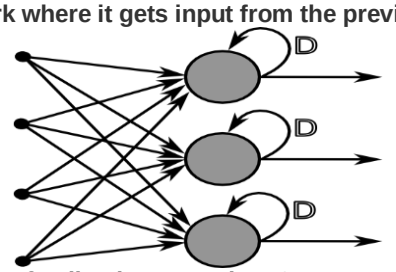
Q15. If you increase the number of hidden layers in a Multi Layer Perceptron, the classification error of test data always decreases. True or False?

B. False

Solution: (B)

This is not always true. Overfitting may cause the error to increase.

Q16. You are building a neural network where it gets input from the previous layer as well as from itself.



Which of the following architecture has feedback connections?

A. Recurrent Neural network

Solution: (A)

Option A is correct.

Q17. What is the sequence of the following tasks in a perceptron?

1. **Initialize weights of perceptron randomly**
2. **Go to the next batch of dataset**
3. **If the prediction does not match the output, change the weights**
4. **For a sample input, compute an output**

D. 1, 4, 3, 2

Solution: (D)

Sequence D is correct.

Q18. Suppose that you have to minimize the cost function by changing the parameters. Which of the following technique could be used for this?

A. Exhaustive Search

B. Random Search

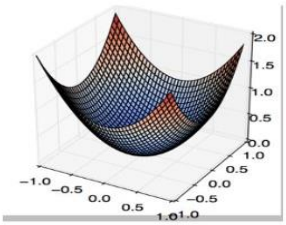
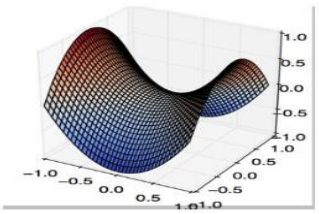
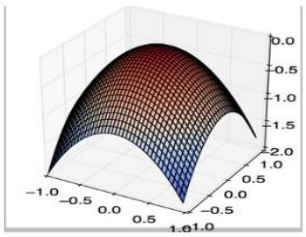
C. Bayesian Optimization

D. Any of these

Solution: (D)

Any of the above mentioned technique can be used to change parameters.

Q19. First Order Gradient descent would not work correctly (i.e. may get stuck) in which of the following graphs?

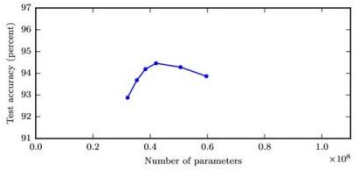
A.  B.  C. 

D. None of these

Solution: (B)

This is a classic example of saddle point problem of gradient descent.

Q20. The below graph shows the accuracy of a trained 3-layer convolutional neural network vs the number of parameters (i.e. number of feature kernels).

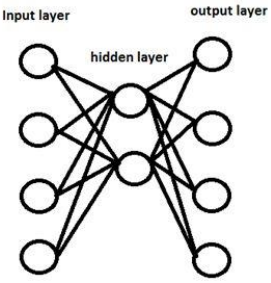


The trend suggests that as you increase the width of a neural network, the accuracy increases till a certain threshold value, and then starts decreasing.

What could be the possible reason for this decrease?

C. As the number of kernels increase, they start to correlate with each other which in turn helps overfitting
Solution: (C)
As mentioned in option C, the possible reason could be kernel correlation.

Q21. Suppose we have one hidden layer neural network as shown above. The hidden layer in this network works as a dimensionality reductor. Now instead of using this hidden layer, we replace it with a dimensionality reduction technique such as PCA.



Would the network that uses a dimensionality reduction technique always give same output as network with hidden layer?
A. Yes
B. No
Solution: (B)
Because PCA works on correlated features, whereas hidden layers work on predictive capacity of features.

Q22. Can a neural network model the function (y=1/x)?
A. Yes
B. No
Solution: (A)
Option A is true, because activation function can be reciprocal function.

Q23. In which neural net architecture, does weight sharing occur?
A. Convolutional neural Network
B. Recurrent Neural Network
C. Fully Connected Neural Network
D. Both A and B
Solution: (D)
Option D is correct.

Q24. Batch Normalization is helpful because
A. It normalizes (changes) all the input before sending it to the next layer
B. It returns back the normalized mean and standard deviation of weights
C. It is a very efficient backpropagation technique
D. None of these
Solution: (A)
To read more about batch normalization, see refer this video

Q25. Instead of trying to achieve absolute zero error, we set a metric called bayes error which is the error we hope to achieve. What could be the reason for using bayes error?
A. Input variables may not contain complete information about the output variable
B. System (that creates input-output mapping) may be stochastic
C. Limited training data
D. All the above
Solution: (D)
In reality achieving accurate prediction is a myth. So we should hope to achieve an "achievable result".

Q26. The number of neurons in the output layer should match the number of classes (Where the number of classes is greater than 2) in a supervised learning task. True or False?
A. True
B. False
Solution: (B)
It depends on output encoding. If it is one-hot encoding, then its true. But you can have two outputs for four classes, and take the binary values as four classes(00,01,10,11).

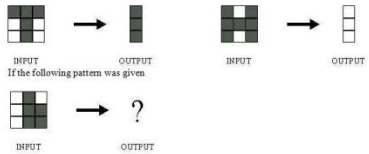
Q27. In a neural network, which of the following techniques is used to deal with overfitting?
A. Dropout
B. Regularization
C. Batch Normalization
D. All of these
Solution: (D)
All of the techniques can be used to deal with overfitting.

Q28. Y = ax^2 + bx + c (polynomial equation of degree 2)
Can this equation be represented by a neural network of single hidden layer with linear threshold?
A. Yes
B. No
Solution: (B)
The answer is no because having a linear threshold restricts your neural network and in simple terms, makes it a consequential linear transformation function.

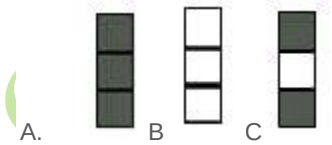
Q29. What is a dead unit in a neural network?
A. A unit which doesn't update during training by any of its neighbour
B. A unit which does not respond completely to any of the training patterns
C. The unit which produces the biggest sum-squared error
D. None of these
Solution: (A)
Option A is correct.
Q30. Which of the following statement is the best description of early stopping?
A. Train the network until a local minimum in the error function is reached
B. Simulate the network on a test dataset after every epoch of training. Stop training when the generalization error starts to increase
C. Add a momentum term to the weight update in the Generalized Delta Rule, so that training converges more quickly
D. A faster version of backpropagation, such as the `Quickprop' algorithm
Solution: (B)
Option B is correct.

Q31. What if we use a learning rate that's too large?
A. Network will converge
B. Network will not converge
C. Can't Say
Solution: B
Option B is correct because the error rate would become erratic and explode.

Q32. The network shown in Figure 1 is trained to recognize the characters H and T as shown below:



What would be the output of the network?



D Could be A or B depending on the weights of neural network
Solution: (D)
Without knowing what are the weights and biases of a neural network, we cannot comment on what output it would give.

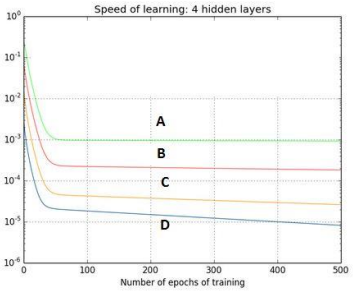
Q33. Suppose a convolutional neural network is trained on ImageNet dataset (Object recognition dataset). This trained model is then given a completely white image as an input.The output probabilities for this input would be equal for all classes. True or False?
A. True
B. False

Solution: (B)
There would be some neurons which are do not activate for white pixels as input. So the classes wont be equal.

Q34. When pooling layer is added in a convolutional neural network, translation in-variance is preserved. True or False?
A. True
B. False
Solution: (A)
Translation invariance is induced when you use pooling.

Q35. Which gradient technique is more advantageous when the data is too big to handle in RAM simultaneously?
A. Full Batch Gradient Descent
B. Stochastic Gradient Descent
Solution: (B)
Option B is correct.

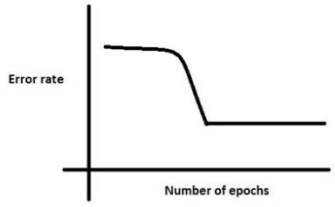
Q36. The graph represents gradient flow of a four-hidden layer neural network which is trained using sigmoid activation function per epoch of training. The neural network suffers with the vanishing gradient problem.



Which of the following statements is true?
A. Hidden layer 1 corresponds to D, Hidden layer 2 corresponds to C, Hidden layer 3 corresponds to B and Hidden layer 4 corresponds to A
B. Hidden layer 1 corresponds to A, Hidden layer 2 corresponds to B, Hidden layer 3 corresponds to C and Hidden layer 4 corresponds to D
Solution: (A)
This is a description of a vanishing gradient problem. As the backprop algorithm goes to starting layers, learning decreases.

Q37. For a classification task, instead of random weight initializations in a neural network, we set all the weights to zero. Which of the following statements is true?
A. There will not be any problem and the neural network will train properly
B. The neural network will train but all the neurons will end up recognizing the same thing
C. The neural network will not train as there is no net gradient change
D. None of these
Solution: (B)
Option B is correct.

Q38. There is a plateau at the start. This is happening because the neural network gets stuck at local minima before going on to global minima.

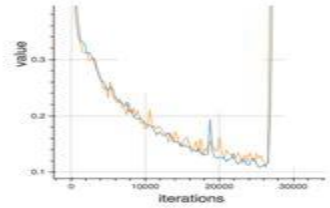


To avoid this, which of the following strategy should work?
A. Increase the number of parameters, as the network would not get stuck at local minima
B. Decrease the learning rate by 10 times at the start and then use momentum

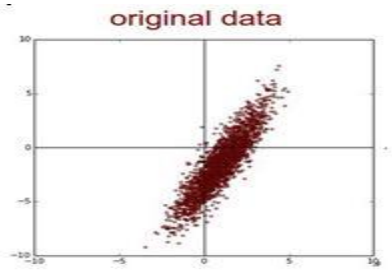
C. Jitter the learning rate, i.e. change the learning rate for a few epochs
D. None of these
Solution: (C)
Option C can be used to take a neural network out of local minima in which it is stuck.

Q39. For an image recognition problem (recognizing a cat in a photo), which architecture of neural network would be better suited to solve the problem?
A. Multi Layer Perceptron
B. Convolutional Neural Network
C. Recurrent Neural network
D. Perceptron
Solution: (B)
Convolutional Neural Network would be better suited for image related problems because of its inherent nature for taking into account changes in nearby locations of an image

Q40. Suppose while training, you encounter this issue. The error suddenly increases after a couple of iterations.

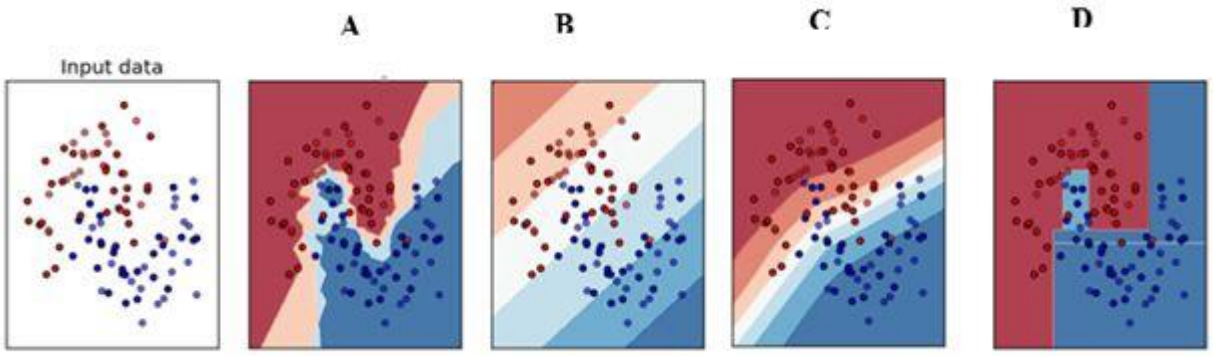


You determine that there must a problem with the data. You plot the data and find the insight that, original data is somewhat skewed and that may be causing the problem.



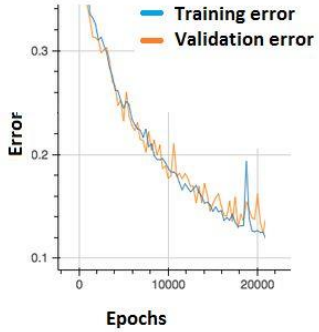
What will you do to deal with this challenge?
A. Normalize
B. Apply PCA and then Normalize
C. Take Log Transform of the data
D. None of these
Solution: (B)
First you would remove the correlations of the data and then zero center it.

Q41. Which of the following is a decision boundary of Neural Network?



A) B
B) A
C) D
D) C
E) All of these
Solution: (E)
A neural network is said to be a universal function approximator, so it can theoretically represent any decision boundary.

Q42. In the graph below, we observe that the error has many "ups and downs"



Should we be worried?
A. Yes, because this means there is a problem with the learning rate of neural network.
B. No, as long as there is a cumulative decrease in both training and validation error, we don't need to worry.
Solution: (B)
Option B is correct. In order to decrease these "ups and downs" try to increase the batch size.

Q43. What are the factors to select the depth of neural network?
1. Type of neural network (eg. MLP, CNN etc)
2. Input data
3. Computation power, i.e. Hardware capabilities and software capabilities
4. Learning Rate
5. The output function to map

A. 1, 2, 4, 5
B. 2, 3, 4, 5
C. 1, 3, 4, 5
D. All of these
Solution: (D)
All of the above factors are important to select the depth of neural network

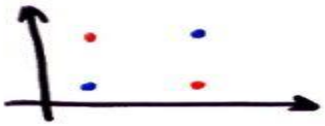
Q44. Consider the scenario. The problem you are trying to solve has a small amount of data. Fortunately, you have a pre-trained neural network that was trained on a similar problem. Which of the following methodologies would you choose to make use of this pre-trained network?
A. Re-train the model for the new dataset

B. Assess on every layer how the model performs and only select a few of them
C. Fine tune the last couple of layers only
D. Freeze all the layers except the last, re-train the last layer
Solution: (D)
If the dataset is mostly similar, the best method would be to train only the last layer, as previous all layers work as feature extractors.

Q45. Increase in size of a convolutional kernel would necessarily increase the performance of a convolutional network.
A. True
B. False
Solution: (B)
Increasing kernel size would not necessarily increase performance. This depends heavily on the dataset.

Questions to test a Data Scientist on Deep Learning

1) Is the data linearly separable?



B) No
Solution: B
If you can draw a line or plane between the data points, it is said to be linearly separable.

2) Which of the following are universal approximators?
A) Kernel SVM
B) Neural Networks
C) Boosted Decision Trees
D) All of the above
Solution: D
All of the above methods can approximate any function.

3) In which of the following applications can we use deep learning to solve the problem?
A) Protein structure prediction
B) Prediction of chemical reactions
C) Detection of exotic particles
D) All of these
Solution: D
We can use neural network to approximate any function so it can theoretically be used to solve any problem.

4) Which of the following statements is true when you use 1×1 convolutions in a CNN?
A) It can help in dimensionality reduction
B) It can be used for feature pooling
C) It suffers less overfitting due to small kernel size
D) All of the above
Solution: D
1×1 convolutions are called bottleneck structure in CNN.

5) Question Context:
Statement 1: It is possible to train a network well by initializing all the weights as 0
Statement 2: It is possible to train a network well by initializing biases as 0
Which of the statements given above is true?
A) Statement 1 is true while Statement 2 is false
B) Statement 2 is true while statement 1 is false
C) Both statements are true
D) Both statements are false
Solution: B
Even if all the biases are zero, there is a chance that neural network may learn. On the other hand, if all the weights are zero; the neural neural network may never learn to perform the task.

6) The number of nodes in the input layer is 10 and the hidden layer is 5. The maximum number of connections from the input layer to the hidden layer are

A) 50
B) Less than 50
C) More than 50
D) It is an arbitrary value
Solution: A
Since MLP is a fully connected directed graph, the number of connections are a multiple of number of nodes in input layer and hidden layer.

7) The input image has been converted into a matrix of size 28 X 28 and a kernel/filter of size 7 X 7 with a stride of 1. What will be the size of the convoluted matrix?
A) 22 X 22
B) 21 X 21
C) 28 X 28
D) 7 X 7
Solution: A
The size of the convoluted matrix is given by C=((I-F+2P)/S)+1, where C is the size of the Convoluted matrix, I is the size of the input matrix, F the size of the filter matrix and P the padding applied to the input matrix. Here P=0, I=28, F=7 and S=1. There the answer is 22.

8) In a simple MLP model with 8 neurons in the input layer, 5 neurons in the hidden layer and 1 neuron in the output layer. What is the size of the weight matrices between hidden output layer and input hidden layer?
A) [1 X 5] , [5 X 8]
B) [8 X 5] , [1 X 5]
C) [8 X 5] , [5 X 1]
D) [5 x 1] , [8 X 5]
Solution: D
The size of weights between any layer 1 and layer 2 Is given by [nodes in layer 1 X nodes in layer 2]

9) Given below is an input matrix named I, kernel F and Convoluted matrix named C. Which of the following is the correct option for matrix C with stride =2 ?

I

1	0	0	1	1	0	1
0	0	1	1	1	0	1
1	1	1	0	1	0	1
1	1	0	1	0	0	0
1	0	1	0	1	1	0
0	1	1	0	0	1	1
0	1	1	1	0	1	1

F

1	0	0
0	1	1
1	1	0

A)

4	4	3	3	3
4	2	3	3	2
3	3	3	1	3
3	4	2	3	2
4	3	3	2	4

B)

4	4	3	3	3
4	2	3	2	2
3	2	3	3	3
3	4	2	3	2
4	3	2	2	4

C)

4	3	3
3	3	3
4	3	4

D)

4	3	3
3	2	2
3	3	4

Solution: C
1 and 2 are automatically eliminated since they do not conform to the output size for a stride of 2. Upon calculation option 3 is the correct answer.

10) Given below is an input matrix of shape 7 X 7. What will be the output on applying a max pooling of size 3 X 3 with a stride of 2?

1	2	4	1	4	0	1
0	0	1	6	1	5	5
1	4	4	5	1	4	1
4	1	5	1	6	5	0
1	0	6	5	1	1	8
2	3	1	8	5	8	1
0	9	1	2	3	1	4

A) | 4 | 6 | 5 |
|---|---|---|
| 6 | 6 | 8 |
| 9 | 8 | 8 |

B) | 4 | 5 | 5 |
|---|---|---|
| 6 | 6 | 8 |
| 9 | 8 | 6 |

C) | 4 | 5 | 6 |
|---|---|---|
| 3 | 6 | 8 |
| 9 | 9 | 6 |

D) | 4 | 3 | 3 |
|---|---|---|
| 3 | 3 | 3 |
| 4 | 3 | 4 |

Solution: A
Max pooling takes a 3 X 3 matrix and takes the maximum of the matrix as the output. Slide it over the entire input matrix with a stride of 2 and you will get option (1) as the answer.

11) Which of the following functions can be used as an activation function in the output layer if we wish to predict the probabilities of n classes (p1, p2..pk) such that sum of p over all n equals to 1?
A) Softmax
B) ReLu
C) Sigmoid
D) Tanh
Solution: A
Softmax function is of the form in which the sum of probabilities over all k sum to 1.

12) Assume a simple MLP model with 3 neurons and inputs= 1,2,3. The weights to the input neurons are 4,5 and 6 respectively. Assume the activation function is a linear constant value of 3. What will be the output ?
A) 32
B) 643
C) 96
D) 48
Solution: C
The output will be calculated as 3(1*4+2*5+6*3) = 96

13) Which of following activation function can't be used at output layer to classify an image ?
A) sigmoid
B) Tanh
C) ReLU
D) If(x>5,1,0)
E) None of the above

Solution: C
ReLU gives continuous output in range 0 to infinity. But in output layer, we want a finite range of values. So option C is correct.

14) [True | False] In the neural network, every parameter can have their different learning rate.
A) TRUE
B) FALSE
Solution: A
Yes, we can define the learning rate for each parameter and it can be different from other parameters.

15) Dropout can be applied at visible layer of Neural Network model?
A) TRUE
B) FALSE
Solution: A
Look at the below model architecture, we have added a new Dropout layer between the input (or visible layer) and the first hidden layer. The dropout rate is set to 20%, meaning one in 5 inputs will be randomly excluded from each update cycle.

```
def create_model():

    # create model

    model = Sequential()

    model.add(Dropout(0.2, input_shape=(60,)))
```

```
    model.add(Dense(60, activation='relu'))

    model.add(Dense(1, activation='sigmoid'))

    # Compile model  sgd = SGD(lr=0.1)

    model.compile(loss='binary_crossentropy', optimizer=sgd, metrics=['accuracy'])

    return model
```

16) I am working with the fully connected architecture having one hidden layer with 3 neurons and one output neuron to solve a binary classification challenge. Below is the structure of input and output:
Input dataset: [[1,0,1,0] , [1,0,1,1] , [0,1,0,1]]
Output: [[1] , [1] , [0]]
To train the model, I have initialized all weights for hidden and output layer with 1.
What do you say model will able to learn the pattern in the data?
A) Yes
B) No
Solution: B
As all the weights of the neural network model are same, so all the neurons will try to do the same thing and the model will never converge.

17) Which of the following neural network training challenge can be solved using batch normalization?
A) Overfitting
B) Restrict activations to become too high or low
C) Training is too slow
D) Both B and C
E) All of the above
Solution: D
Batch normalization restricts the activations and indirectly improves training time.

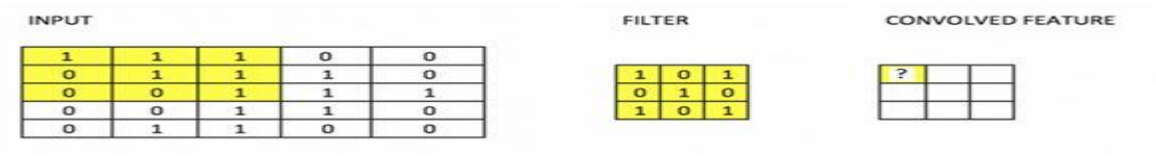
18) Which of the following would have a constant input in each epoch of training a Deep Learning model?
A) Weight between input and hidden layer
B) Weight between hidden and output layer
C) Biases of all hidden layer neurons
D) Activation function of output layer
E) None of the above
Solution: A
Weights between input and hidden layer are constant.

19) True/False: Changing Sigmoid activation to ReLu will help to get over the vanishing gradient issue?
A) TRUE
B) FALSE
Solution: A
ReLU can help in solving vanishing gradient problem.

20) In CNN, having max pooling always decrease the parameters?
A) TRUE
B) FALSE
Solution: B
This is not always true. If we have a max pooling layer of pooling size as 1, the parameters would remain the same.

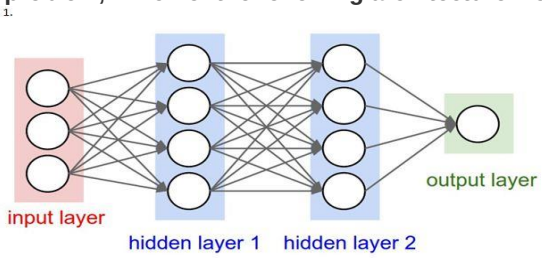
21) [True or False] BackPropogation cannot be applied when using pooling layers
A) TRUE
B) FALSE
Solution: B
BackPropogation can be applied on pooling layers too.

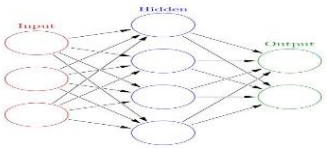
22) What value would be in place of question mark?



Here we see a convolutional function being applied to input.
A) 3
B) 4
C) 5
D) 6
Solution: B
Option B is correct

23) For a binary classification problem, which of the following architecture would you choose?

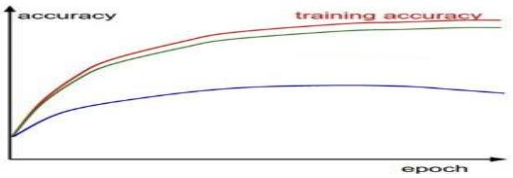




A) 1
B) 2
C) Any one of these
D) None of these
Solution: C
We can either use one neuron as output for binary classification problem or two separate neurons.

24) Suppose there is an issue while training a neural network. The training loss/validation loss remains constant. What could be the possible reason?
A) Architecture is not defined correctly
B) Data given to the model is noisy
C) Both of these
Solution: C
Both architecture and data could be incorrect. Refer this article https://www.analyticsvidhya.com/blog/2017/07/debugging-neural-network-with-tensorboard/

25)



The red curve above denotes training accuracy with respect to each epoch in a deep learning algorithm. Both the green and blue curves denote validation accuracy.
Which of these indicate overfitting?
A) Green Curve
B) Blue Curve
Solution: B
Blue curve shows overfitting, whereas green curve is generalized.

26) Which of the following statement is true regrading dropout?
1: Dropout gives a way to approximate by combining many different architectures
2: Dropout demands high learning rates
3: Dropout can help preventing overfitting
A) Both 1 and 2
B) Both 1 and 3
C) Both 2 and 3
D) All 1, 2 and 3
Solution: B
Statements 1 and 3 are correct, statement 2 is not always true. Even after applying dropout and with low learning rate, a neural network can learn.

27) Gated Recurrent units can help prevent vanishing gradient problem in RNN.
A) True
B) False
Solution: A
Option A is correct. This is because it has implicit memory to remember past behavior.

28) Suppose you are using early stopping mechanism with patience as 2, at which point will the neural network model stop training?

Sr. No.	Training Loss	Validation Loss
1	1.0	1.1
2	0.9	1.0
3	0.8	1.0
4	0.7	1.0
5	0.6	1.1

A) 2
B) 3
C) 4
D) 5
Solution: C
As we have set patience as 2, the network will automatically stop training after epoch 4.

29) [True or False] Sentiment analysis using Deep Learning is a many-to one prediction task
A) TRUE
B) FALSE
Solution: A
Option A is correct. This is because from a sequence of words, you have to predict whether the sentiment was positive or negative.

30) What steps can we take to prevent overfitting in a Neural Network?
A) Data Augmentation
B) Weight Sharing
C) Early Stopping
D) Dropout
E) All of the above
Solution: E
All of the above mentioned methods can help in preventing overfitting problem.

40 Questions to test a data scientist on Deep Learning
https://www.analyticsvidhya.com/blog/2017/04/40-questions-test-data-scientist-deep-learning/

1) The difference between deep learning and machine learning algorithms is that there is no need of feature engineering in machine learning algorithms, whereas, it is recommended to do feature engineering first and then apply deep learning.
A) TRUE

B) FALSE
Solution: **(B)**
Deep learning itself does feature engineering whereas machine learning requires manual feature engineering.

2) Which of the following is a representation learning algorithm?
A) Neural network
B) Random Forest
C) k-Nearest neighbor
D) None of the above
Solution: **(A)**
Neural network converts data in such a form that it would be better to solve the desired problem. This is called representation learning.

3) Which of the following option is correct for the below-mentioned techniques?

1. **AdaGrad uses first order differentiation**
2. **L-BFGS uses second order differentiation**
3. **AdaGrad uses second order differentiation**
4. **L-BFGS uses first order differentiation**

A) 1 and 2
Solution: **(A)**
Option A is correct.

4) Increase in size of a convolutional kernel would necessarily increase the performance of a convolutional neural network.
A) TRUE
B) FALSE
Solution: **(B)**
Kernel size is a hyperparameter and therefore by changing it we can increase or decrease performance.

Question Context
Suppose we have a deep neural network model which was trained on a vehicle detection problem. The dataset consisted of images on cars and trucks and the aim was to detect name of the vehicle (the number of classes of vehicles are 10).
Now you want to use this model on different dataset which has images of only Ford Mustangs (aka car) and the task is to locate the car in an image.

5) Which of the following categories would be suitable for this type of problem?
A) Fine tune only the last couple of layers and change the last layer (classification layer) to regression layer
Solution: **(A)**

6) Suppose you have 5 convolutional kernel of size 7 x 7 with zero padding and stride 1 in the first layer of a convolutional neural network. You pass an input of dimension 224 x 224 x 3 through this layer. What are the dimensions of the data which the next layer will receive?
C) 218 x 218 x 5
Solution: **(C)**

7) Suppose we have a neural network with ReLU activation function. Let's say, we replace ReLu activations by linear activations.
Would this new neural network be able to approximate an XNOR function?
Note: The neural network was able to approximate XNOR function with activation function ReLu.
A) Yes
B) No
Solution: **(B)**
If ReLU activation is replaced by linear activation, the neural network loses its power to approximate non-linear function.

8) Suppose we have a 5-layer neural network which takes 3 hours to train on a GPU with 4GB VRAM. At test time, it takes 2 seconds for single data point.
Now we change the architecture such that we add dropout after 2nd and 4th layer with rates 0.2 and 0.3 respectively.
What would be the testing time for this new architecture?
A) Less than 2 secs
B) Exactly 2 secs
C) Greater than 2 secs
D) Can't Say
Solution: **(B)**
The changes is architecture when we add dropout only changes in the training, and not at test time.

9) Which of the following options can be used to reduce overfitting in deep learning models?

1. **Add more data**

2. **Use data augmentation**
3. **Use architecture that generalizes well**
4. **Add regularization**
5. **Reduce architectural complexity**

A) 1, 2, 3
B) 1, 4, 5
C) 1, 3, 4, 5
D) All of these
Solution: **(D)**
All of the above techniques can be used to reduce overfitting.

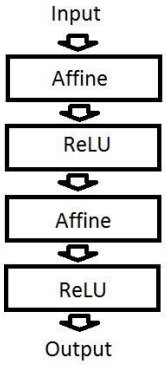
10) Perplexity is a commonly used evaluation technique when applying deep learning for NLP tasks. Which of the following statement is correct?
A) Higher the perplexity the better
B) Lower the perplexity the better
Solution: **(B)**

11) Suppose an input to Max-Pooling layer is given above. The pooling size of neurons in the layer is (3, 3).

3	4	5
4	5	6
5	6	7

What would be the output of this Pooling layer?
A) 3
B) 5
C) 5.5
D) 7
Solution: **(D)**
Max pooling works as follows, it first takes the input using the pooling size we defined, and gives out the highest activated input.

12) Suppose there is a neural network with the below configuration.
If we remove the ReLU layers, we can still use this neural network to model non-linear functions.



A) TRUE
B) FALSE
Solution: **(B)**

13) Deep learning can be applied to which of the following NLP tasks?
A) Machine translation
B) Sentiment analysis
C) Question Answering system
D) All of the above
Solution: **(D)**
Deep learning can be applied to all of the above-mentioned NLP tasks.

14) Scenario 1: You are given data of the map of Arcadia city, with aerial photographs of the city and its outskirts. The task is to segment the areas into industrial land, farmland and natural landmarks like river, mountains, etc.

Scenario 2: You are given data of the map of Arcadia city, with detailed roads and distances between landmarks. This is represented as a graph structure. The task is to find out the nearest distance between two landmarks.
Deep learning can be applied to Scenario 1 but not Scenario 2.
A) TRUE
B) FALSE
Solution: **(B)**
Scenario 1 is on Euclidean data and scenario 2 is on Graphical data. Deep learning can be applied to both types of data.

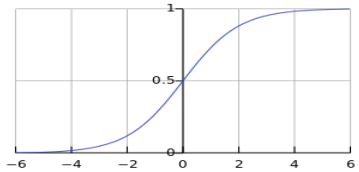
15) Which of the following is a data augmentation technique used in image recognition tasks?
1. **Horizontal flipping**
2. **Random cropping**
3. **Random scaling**
4. **Color jittering**
5. **Random translation**
6. **Random shearing**

A) 1, 2, 4
B) 2, 3, 4, 5, 6
C) 1, 3, 5, 6
D) All of these
Solution: **(D)**

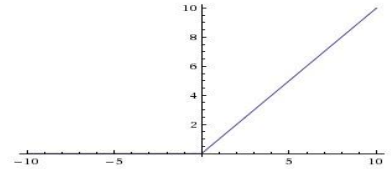
16) Given an n-character word, we want to predict which character would be the n+1th character in the sequence. For example, our input is "predictio" (which is a 9 character word) and we have to predict what would be the 10th character.
Which neural network architecture would be suitable to complete this task?
A) Fully-Connected Neural Network
B) Convolutional Neural Network
C) Recurrent Neural Network
D) Restricted Boltzmann Machine
Solution: **(C)**
Recurrent neural network works best for sequential data. Therefore, it would be best for the task.

17) What is generally the sequence followed when building a neural network architecture for semantic segmentation for image?
A) Convolutional network on input and deconvolutional network on output
B) Deconvolutional network on input and convolutional network on output
Solution: **(A)**

18) Sigmoid was the most commonly used activation function in neural network, until an issue was identified. The issue is that when the gradients are too large in positive or negative direction, the resulting gradients coming out of the activation function get squashed. This is called saturation of the neuron.



That is why ReLU function was proposed, which kept the gradients same as before in the positive direction.



A ReLU unit in neural network never gets saturated.
A) TRUE
B) FALSE
Solution: **(B)**
ReLU can get saturated too. This can be on the negative side of x-axis.

19) What is the relationship between dropout rate and regularization?
Note: we have defined dropout rate as the probability of keeping a neuron active?
A) Higher the dropout rate, higher is the regularization
B) Higher the dropout rate, lower is the regularization
Solution: **(B)**
Higher dropout rate says that more neurons are active. So there would be less regularization.

20) What is the technical difference between vanilla backpropagation algorithm and backpropagation through time (BPTT) algorithm?
A) Unlike backprop, in BPTT we sum up gradients for corresponding weight for each time step
B) Unlike backprop, in BPTT we subtract gradients for corresponding weight for each time step
Solution: **(A)**
BPTT is used in context of recurrent neural networks. It works by summing up gradients for each time step

21) Exploding gradient problem is an issue in training deep networks where the gradient getS so large that the loss goes to an infinitely high value and then explodes.
What is the probable approach when dealing with "Exploding Gradient" problem in RNNs?
A) Use modified architectures like LSTM and GRUs
B) Gradient clipping
C) Dropout
D) None of these
Solution: **(B)**
To deal with exploding gradient problem, it's best to threshold the gradient values at a specific point. This is called gradient clipping.

22) There are many types of gradient descent algorithms. Two of the most notable ones are l-BFGS and SGD. l-BFGS is a second order gradient descent technique whereas SGD is a first order gradient descent technique.
In which of the following scenarios would you prefer l-BFGS over SGD?
1. Data is sparse
2. Number of parameters of neural network are small

A) Both 1 and 2
B) Only 1
C) Only 2
D) None of these
Solution: **(A)**
l-BFGS works best for both of the scenarios.

23) Which of the following is not a direct prediction technique for NLP tasks?
A) Recurrent Neural Network
B) Skip-gram model
C) PCA
D) Convolutional neural network
Solution: **(C)**

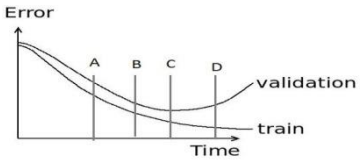
24) Which of the following would be the best for a non-continuous objective during optimization in deep neural net?
A) L-BFGS
B) SGD
C) AdaGrad
D) Subgradient method
Solution: **(D)**
Other optimization algorithms might fail on non-continuous objectives, but sub-gradient method would not.

25) Which of the following is correct?
1. **Dropout randomly masks the input weights to a neuron**
2. **Dropconnect randomly masks both input and output weights to a neuron**

A) 1 is True and 2 is False
B) 1 is False and 2 is True
C) Both 1 and 2 are True
D) Both 1 and 2 are False

Solution: **(D)**
In dropout, neurons are dropped; whereas in dropconnect; connections are dropped. So both input and output weights will be rendered in useless, i.e. both will be dropped for a neuron. Whereas in dropconnect, only one of them should be dropped

26) While training a neural network for image recognition task, we plot the graph of training error and validation error for debugging.



What is the best place in the graph for early stopping?
A) A
B) B
C) C
D) D
Solution: **(C)**
You would "early stop" where the model is most generalized. Therefore option C is correct.

27) Research is going on to solve image inpainting problem using deep learning. For this, which loss function would be appropriate for computing the pixel-wise region to be inpainted?

Image inpainting is one of those problems which requires human expertise for solving it. It is particularly useful to repair damaged photos or videos. Below is an example of input and output of an image inpainting example.
A) Euclidean loss
B) Negative-log Likelihood loss
C) Any of the above
Solution: **(C)**
Both A and B can be used as a loss function for image inpainting problem.

28) Backpropagation works by first calculating the gradient of ___ and then propagating it backwards.
A) Sum of squared error with respect to inputs
B) Sum of squared error with respect to weights
C) Sum of squared error with respect to outputs
D) None of the above
Solution: **(C)**
29) Mini-Batch sizes when defining a neural network are preferred to be multiple of 2's such as 256 or 512. What is the reason behind it?
A) Gradient descent optimizes best when you use an even number
B) Parallelization of neural network is best when the memory is used optimally
C) Losses are erratic when you don't use an even number
D) None of these
Solution: **(B)**

30) Xavier initialization is most commonly used to initialize the weights of a neural network. Below is given the formula for initialization.

$$\mathrm{Var}(W) = \frac{2}{n_{\mathrm{in}} + n_{\mathrm{out}}}$$

1. **If weights at the start are small, then signals reaching the end will be too tiny.**
2. **If weights at the start are too large, signals reaching the end will be too large.**

3. **Weights from Xavier's init are drawn from the Gaussian distribution.**

Xavier's init helps reduce vanishing gradient problem.

Xavier's init is used to help the input signals reach deep into the network. Which of the following statements are true?

A) 1, 2, 4
B) 2, 3, 4
C) 1, 3, 4
D) 1, 2, 3
E) 1, 2, 3, 4
Solution: **(D)**
All of the above statements are true.

31) As the length of sentence increases, it becomes harder for a neural translation machine to perform as sentence meaning is represented by a fixed dimensional vector. To solve this, which of the following could we do?

A) Use recursive units instead of recurrent
B)Use attention mechanism
C) Use character level translation
D) None of these
Solution: **(B)**

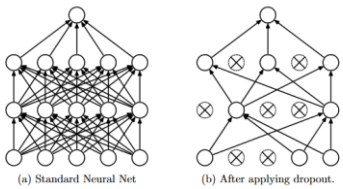
32) A recurrent neural network can be unfolded into a full-connected neural network with infinite length.

A) TRUE
B) FALSE
Solution: **(A)**
Recurrent neuron can be thought of as a neuron sequence of infinite length of time steps.

33) Which of the following is a bottleneck for deep learning algorithm?

A) Data related to the problem
B) CPU to GPU communication
C) GPU memory
D) All of the above
Solution: **(D)**
Along with having the knowledge of how to apply deep learning algorithms, you should also know the implementation details. Therefore you should know that all the above mentioned problems are a bottleneck for deep learning algorithm.

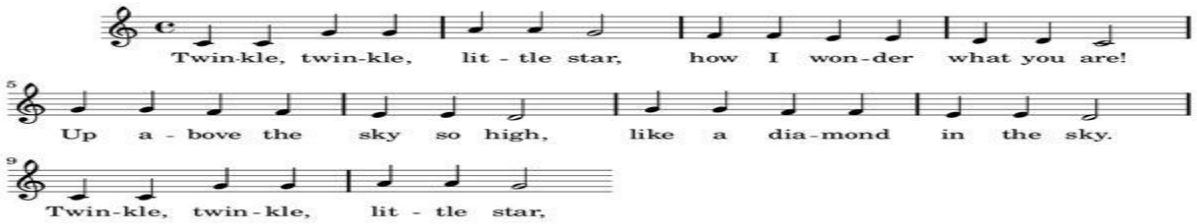
34) Dropout is a regularization technique used especially in the context of deep learning. It works as following, in one iteration we first randomly choose neurons in the layers and masks them. Then this network is trained and optimized in the same iteration. In the next iteration, another set of randomly chosen neurons are selected and masked and the training continues.


(a) Standard Neural Net (b) After applying dropout.

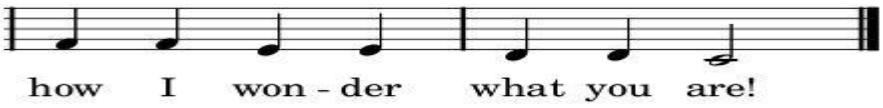
Dropout technique is not an advantageous technique for which of the following layers?

A) Affine layer
B) Convolutional layer
C) RNN layer
D) None of these
Solution: **(C)**
Dropout does not work well with recurrent layer. You would have to modify dropout technique a bit to get good results.

35) Suppose your task is to predict the next few notes of song when you are given the preceding segment of the song.
For example:
The input given to you is an image depicting the music symbols as given below,



Your required output is an image of succeeding symbols.



Which architecture of neural network would be better suited to solve the problem?
A) End-to-End fully connected neural network
B) Convolutional neural network followed by recurrent units
C) Neural Turing Machine
D) None of these
Solution: **(B)**
CNN work best on image recognition problems, whereas RNN works best on sequence prediction. Here you would have to use best of both worlds!

36) When deriving a memory cell in memory networks, we choose to read values as vector values instead of scalars. Which type of addressing would this entail?
A) Content-based addressing
B) Location-based addressing
Solution: **(A)**

37) It is generally recommended to replace pooling layers in generator part of convolutional generative adversarial nets with _____ ?
A) Affine layer
B) Strided convolutional layer
C) Fractional strided convolutional layer
D) ReLU layer
Solution: **(C)**
Option C is correct. Go through this link.

Question Context 38-40
GRU is a special type of Recurrent Neural Networks proposed to overcome the difficulties of classical RNNs. This is the paper in which they were proposed: "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. Read the full paper here.
38) Which of the following statements is true with respect to GRU?
1. **Units with short-term dependencies have reset gate very active.**
2. **Units with long-term dependencies have update gate very active**

A) Only 1
B) Only 2
C) None of them
D) Both 1 and 2
Solution: **(D)**

39) If calculation of reset gate in GRU unit is close to 0, which of the following would occur?
A) Previous hidden state would be ignored
B) Previous hidden state would be not be ignored
Solution: **(A)**

40) If calculation of update gate in GRU unit is close to 1, which of the following would occur?
A) Forgets the information for future time steps
B) Copies the information through many time steps
Solution: **(B)**

Image Processing

Skill test Questions and Answers

1) Match the following image formats to their correct number of channels

- GrayScale
- RGB

I. 1 channel
II. 2 channels
III. 3 channels
IV. 4 channels

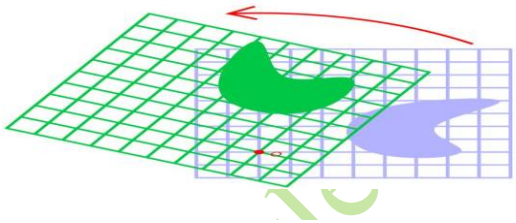
None

A) RGB -> I, GrayScale-> III
B) RGB -> IV, GrayScale-> II
C) RGB -> III, GrayScale -> I
D) RGB -> II, GrayScale -> I

Solution: C

Grayscale images have one number per pixel, and are stored as an m × n matrix, whereas Color images have 3 numbers per pixel – red, green, and blue brightness (RGB)

2)Suppose you have to rotate an image. Image rotation is nothing but multiplication of image by a specific matrix to get a new transformed image.



For simplicity, we consider one point in the image to rotate with co-ordinates as (1, 0) to a co-ordinate of (0, 1), which of the following matrix would we have to multiply with?

A) $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ B) $\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$ C) $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ D) $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

Solution: C

The calculation of would be like this;

[[0], [1]] = [[0, -1], [1, 0]] x [1, 0]

3) [True or False] To blur an image, you can use a linear filter

A) TRUE
B) FALSE

Solution: B

Blurring compares neighboring pixels in a filter and smooth them. For this, you cannot use a linear filter.

4) Which of the following is a challenge when dealing with computer vision problems?

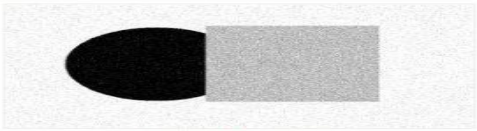
A) Variations due to geometric changes (like pose, scale etc)
B) Variations due to photometric factors (like illumination, appearance etc)
C) Image occlusion
D) All of the above

Solution: D

All the above mentioned options are challenges in computer vision

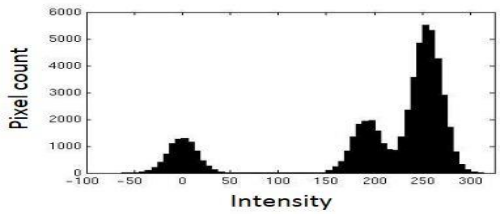
5)

Suppose we have an image given below.



Input image

Our task is to segment the objects in the image. A simple way to do this is to represent the image in terms of the intensity of pixels and the cluster them according to the values. On doing this, we got this type of structure.



Suppose we choose k-means clustering to solve the problem, what would be the appropriate value of k from just a visual inspection of the intensity graph?
A) 1
B) 2
C) 3
D) 4
Solution: C
Three clusters will be formed; points in the circle, points in the square and the points excluding both of these objects

6)

In this image, you can find an edge labelled in the red region. Which form of discontinuity creates this kind of edge?
A) Depth Discontinuity
B) Surface color Discontinuity
C) Illumination discontinuity
D) None of the above
Solution: A
The chair and wall are far from each other, causing an edge in the image.

7) Finite difference filters in image processing are very susceptible to noise. To cope up with this, which of the following methods can you use so that there would be minimal distortions by noise?
A) Downsample the image
B) Convert the image to grayscale from RGB
C) Smooth the image
D) None of the above
Solution: C
Smoothing helps in reducing noise by forcing pixels to be more like their neighbours

8) Consider and image with width and height as 100×100. Each pixel in the image can have a color from Grayscale, i.e. values. How much space would this image require for storing?
Note: No compression is done.
A) 2,56,00,000
B) 25,60,000
C) 2,56,000

D) 8,00,000
E) 80,000
F) 8,000
Solution: E
The answer will be 8x100x100 because 8 bits will be required to represent a number from 0-256

9) [True or False] Quantizing an image will reduce the amount of memory required for storage.
A) TRUE
B) FALSE
Solution: A
The statement given is true.

10) Suppose we have a grayscale image, with most of the values of pixels being same. What can we use to compress the size of image?
A) Encode the pixels with same values in a dictionary

Solution: A
Encoding same values of pixels will greatly reduce the size for storage

11) [True or False] JPEG is a lossy image compression technique
A) TRUE
Solution: A
The reason for JPEG being a lossy compression technique is because of the use of quantization.

12) Given an image with only 2 pixels and 3 possible values for each pixel, what is the number of possible image histograms that can be formed?
C) 9
Solution: C
The permutations possible of the histograms would be 9.

13) Suppose we have a 1D image with values as

[2, 5, 8, 5, 2]
Now we apply average filter on this image of size 3. What would be the value of the last second pixel?

A) The value would remain the same
B) The value would increase by 2
C) The value would decrease by 2
D) None of the above

Solution: A
(8+5+2)/3 will become 5. So there will be no change.

14) fMRI (Functional magnetic resonance imaging) is a technology where volumetric scans of the brain are acquired while the subject is performing some cognitive tasks over time. What is the dimensionality of fMRI output signals?
A) 1D
B) 2D
C) 3D
D) None of the above
Solution: D
The question itself mentions "volumetric scans" over time, so it would be a series of 3D scans

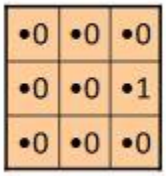
15) Which of the following methods is used as a model fitting method for edge detection?
A) SIFT
B) Difference of Gaussian detector
C) RANSAC
D) None of the above
Solution: C
RANSAC is used to find the best fit line in edge detection

16)

Suppose we have an image which is noisy. This type of noise in the image is called salt-and-pepper noise

[True or False] Median filter technique is the best way to denoise this image
A) TRUE
B) FALSE
Solution: A
Median filter technique helps reduce noise to a good enough extent

17) If we convolve an image with the matrix given below, what would be the relation between the original and modified image?



A) The image will be shifted to the right by 1 pixel
B) The image will be shifted down by 1 pixel
C) The image will be shifted to the left by 1 pixel
D) The image will be shifted up by 1 pixel
Solution: A
I would suggest you to try this yourself and see the result!

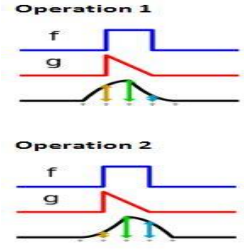
18) Which of the following is a correct way to sharpen an image?
A)
1. Convolve the image with identity matrix
2. Subtract this resulting image from the original
3. Add this subtracted result back to the original image

B)
1. Smooth the image
2. Subtract this smoothed image from the original
3. Add this subtracted result back to the original image

C)
1. Smooth the image
2. Add this smoothed image back to the original image

D) None of the above
Solution: B
Option B gives a correct way to sharpen an image

19) Below given images are two operations performed on a signal. Can you identify which is which?



A) Operation 1 is cross correlation between signal f and signal g, whereas operation 2 is convolution function applied

to signal f and signal g
B) Operation 2 is cross correlation between signal f and signal g, whereas operation 1 is convolution function applied to signal f and signal g
Solution: A
Correlation and convolution are two different methods with give different result. Convolution defines how much the signals overlap, whereas correlation tries to find the relation between the signals

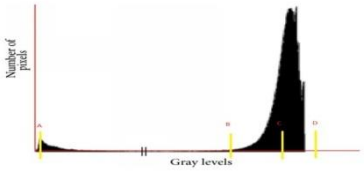
20) [True or False] By using template matching along with cross correlation, you can build a vision system for TV remote control
A) TRUE
B) FALSE
Solution: A
This is a excellent example of cross correlation in computer vision. Refer paper "Computer Vision for Interactive Computer Graphics," W.Freeman et al, IEEE Computer Graphics and Applications

21) Suppose you are creating a face detector in the wild. Which of the following features would you select for creating a robust facial detector?
1. Location of iris, eyebrow and chin
2. Boolean feature: Is the person smiling or not
3. Angle of orientation of face
4. Is the person sitting or standing
A) 1, 2
B) 1, 3
C) 1, 2, 3
D) 1, 2, 3, 4
Solution: B
Options 1, 3 would be relevant features for the problem, but 2, 4 may not be

22) Which of the following is example of low level feature in an image?
A) HOG
B) SIFT
C) HAAR features
D) All of the above
Solution: D
All the above are examples of low-level features

23) In RGBA mode of color representation, what does A represent?
A) Depth of an image
B) Intensity of colors
C) Opacity of an image
D) None of the above
Solution: C
Opacity can be mentioned by introducing it as a fourth parameter in RGB

24) In Otsu thresholding technique, you remove the noise by thresholding the points which are irrelevant and keeping those which do not represent noise.

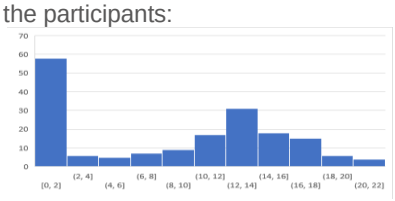


In the image given, at which point would you threshold on?
A) A
B) B
C) C
D) D
Solution: B
Line B would catch most of the noise in the image.

25) Which of the following data augmentation technique would you prefer for an object recognition problem?
`

A) Horizontal flipping
B) Rescaling
C) Zooming in the image
D) All of the above
Solution: D
All the mentioned techniques can be used for data augmentation.

Overall Distribution
Below is the distribution of the scores of the participants:



You can access the scores here. More than a hundered people participated in the skill test and the highest score obtained was a 22.

Frequently Asked Questions on Deep Learning

What is Deep Learning and why is it so popular these days?
Deep Learning is nothing but a paradigm of machine learning which has shown incredible promise in the recent years. This is because of the fact that Deep Learning shows great analogy with the functioning of the human brain. The superiority of the human brain is an evident fact, and it is considered to be the most versatile and efficient self-learning model that has ever been created.
Let us understand the functioning of a deep learning model with an example:

What do you see in the above image?
The most obvious answer would be "*a car*", right? Despite the fact, that there is sand, greenery, clouds and a lot of other things, our brain tags this image as one of a car. This is because our brain has learnt to identify the primary subject of an image.
This ability of deriving useful information from a lot of extraneous data is what makes deep learning special. With the amount of data that is being generated these days, we want our models to be better with more of this data being fed into it. While deep learning models get better with the increase in the amount of data.
Now although Deep Learning has been around for many years, the major breakthroughs from these techniques came just in the recent years. This is because of two main reasons – the first and foremost, as we saw before, is the increase of data generated through various sources. The infographic below succinctly visualizes this trend. The second is the growth in hardware resources required to run these models. GPUs, which are becoming a requirement to run deep learning models, are multiple times faster and they help us build bigger and deeper deep learning models in comparatively less time than we required previously.

This is the reason that Deep Learning has become a major buzz word in the data science industry.

Is Deep Learning just a hype or does it have real-life applications?
Deep Learning has found many practical applications in the recent past. From Netflix's famous movie recommendation system to Google's self-driving cars, deep learning is already transforming a lot of businesses and is expected to bring about a revolution in almost all industries. Deep learning models are being used from diagnosing cancer to winning presidential elections, from creating art and writing literature to making real life money. Thus it would be wrong to say that it is just a hyped topic anymore.
Some major applications of deep learning that are being employed by technology companies are:

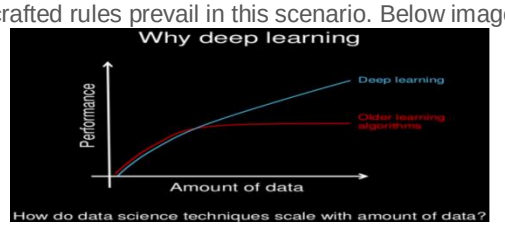
- Google and Facebook are **translating text** into hundreds of languages at a time. This is being done through some deep learning models being applied to NLP tasks and is a major success story.

- Conversational agents like Siri, Alexa, Cortana basically work on **simplifying the speech recognition**techniques through LSTMs and RNNs. Voice commands have added a whole new domain to the possibilities of a machine.
- Deep learning is being used in **impactful computer vision applications** such as OCR (Optical Character Recognition) and real time language translation
- Multimedia sharing apps like Snapchat and Instagram apply **facial feature detection** which is another application of deep learning.
- Deep Learning is being used **in Healthcare domain to locate malignant cells** and other foreign bodies in order to detect complex diseases.

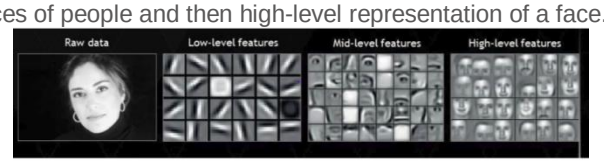
However, some people develop a thinking that deep learning is overhyped because of the fact that labeled data required for training deep learning models is not readily available. Even if the data is available, the computational power required to train such models does not come cheap. Hence, due to these barriers, people are not able to experience the power of deep learning and term it as just hype.
Go through the following blog to build some real life deep learning applications yourself:

- 6 Deep Learning Applications a beginner can build in minutes (using Python)

What is the difference between Deep Learning and Machine Learning?
This is one of the most important questions that most of us need to understand. The comparison can be done mainly on the below three verticals:
Data dependencies
The most important difference between deep learning and traditional machine learning is its performance as the scale of data increases. When the data is small, deep learning algorithms don't perform that well. This is because deep learning algorithms need a large amount of data to understand it perfectly. On the other hand, traditional machine learning algorithms with their handcrafted rules prevail in this scenario. Below image summarizes this fact.



Feature engineering
Feature engineering is a process of putting domain knowledge into the creation of feature extractors to reduce the complexity of the data and make patterns more visible to learning algorithms to work. This process is difficult and expensive in terms of time and expertise.
In Machine learning, most of the applied features need to be identified by an expert and then hand-coded as per the domain and data type.
For example, features can be pixel values, shape, textures, position and orientation. The performance of most of the Machine Learning algorithm depends on how accurately the features are identified and extracted.
Deep learning algorithms try to learn high-level features from data. This is a very distinctive part of Deep Learning and a major step ahead of traditional Machine Learning. Therefore, deep learning reduces the task of developing new feature extractor for every problem. Like, Convolutional NN will try to learn low-level features such as edges and lines in early layers then parts of faces of people and then high-level representation of a face.



Interpretability
Last but not the least, we have interpretability as a factor for comparison of machine learning and deep learning. This factor is the main reason deep learning is still thought 10 times before its use in industry.
Let's take an example. Suppose we use deep learning to give automated scoring to essays. The performance it gives in scoring is quite excellent and is near human performance. But there's is an issue. It does not reveal why it has given that score. Indeed mathematically you can find out which nodes of a deep neural network were activated, but we don't know what there neurons were supposed to model and what these layers of neurons were doing collectively. So we fail to interpret the results.
On the other hand, machine learning algorithms like decision trees give us crisp rules as to why it chose what it chose, so it is particularly easy to interpret the reasoning behind it. Therefore, algorithms like decision trees and linear/logistic regression are primarily used in industry for interpretability.

If you would like to learn about a more in-depth comparison between machine learning and deep learning, I recommend you go through the following blog:

- Deep Learning vs. Machine Learning – the essential differences you need to know!

What are the prerequisites for starting out in Deep Learning?

Starting out in deep learning is not as difficult as people might make you believe. There are a few elementary basics that you should cover before diving into deep learning. Deep learning requires knowledge of the following topics:

- **Mathematics**: You should be comfortable with probability, derivatives, linear algebra and a few other basic topics. Khan Academy offers a decent course covering almost all the above topics here.
- **Statistics:** The basics of statistics are required for going forward with any machine learning problem. Understanding the concepts of statistics are essential because most of the deep learning concepts are derived from assimilating the concepts of statistics. You can check the online courses available here.
- **Tool:** A decent level of coding skills are required for implementing deep learning into real life problems. Coursera's, Introduction to Data Science in Python is a decent course to start off with Python as a tool.
- **Machine Learning**: Machine learning is the base for deep learning. One can not start learning deep learning without understanding the concepts of machine learning. You could go through Intro to Machine Learning or Andrew Ng's course Machine Learning for a theoretical base.

For a more detailed understanding about the prerequisites please follow:

- A Complete Guide on Getting Started with Deep Learning in Python

Which Tools/Languages should I prefer to build Deep learning models?

I would recommend you use Python, because of its robust ecosystem for machine learning. The python ecosystem comprises of developers and coders who are providing open source libraries and support for the community of python users. This makes the task of writing complex codes for various algorithms much easier and the techniques easier to implement and experiment with.

Also, Python being a more generalized programming language, can be used for both the development and implementation. This greatly simplifies the transition from development to operations. That is, a deep learning product that can predict the price of flight tickets, can not only be developed in python but can also be attached with your website in the same form. This is what makes Python a universal language.

Besides this, I would suggest that beginner's use high level libraries like Keras. This makes experimentation easier by providing abstraction to the unnecessary information that is hidden under the algorithms. And giving access to the parameters that can be tweaked to enhance the performance of such models. Let us understand this with an example:

When you press the buttons on a television remote, do you need to care about the background processes that are happening inside the remote? Do you need to know about what signal is being sent out for that key, or how is it being amplified?

No, right?

Because maybe an understanding of these processes is required for a physicist but for a lame man sitting in his bedroom, it is just an information overload.

There are also other contenders apart from Python in the deep learning space such as R, Julia, C++, and Java. For alternatives of libraries, you can check out TensorFlow, Pytorch, Caffe2, DL4J, etc. We should stay updated with their developments as well.

If you are not well versed with programming, there are also a few GUI based softwares, that require no coding, to build deep learning models, such as Lobe or Google's AutoML, among others.

Why are GPUs necessary for building Deep Learning models?

When you train a deep learning model, two main operations are performed:

- Forward Pass
- Backward Pass

In forward pass, input is passed through the neural network and after processing the input, an output is generated. Whereas in backward pass, we update the weights of neural network on the basis of error we get in forward pass.

Both of these operations are essentially matrix multiplications. A simple matrix multiplication can be represented by the image below

Here, we can see that each element in one row of first array is multiplied with one column of second array. So in a neural network, we can consider first array as input to the neural network, and the second array can be considered as weights of the network.

This seems to be a simple task. Now just to give you a sense of what kind of scale deep learning – VGG16 (a convolutional neural network of 16 hidden layers which is frequently used in deep learning applications) has ~140 million parameters; aka weights and biases. Now think of all the matrix multiplications you would have to do to pass just one input to this network! It would take years to train this kind of systems if we take traditional approaches.

We saw that the computationally intensive part of neural network is made up of multiple matrix multiplications. So how can we make it faster?

We can simply do this by performing all the operations at the same time instead of doing it one after the other. This, in a nutshell, is why we use GPU (graphics processing units) instead of a CPU (central processing unit) for training a neural network.

When (and where) to apply Neural Networks ?

Deep Learning have been in the spotlight for quite some time now. Its "deeper" versions are making tremendous breakthroughs in many fields such as image recognition, speech and natural language processing etc.

Now that we know it is so impactful; the main question that arises is when to and when not to apply neural networks? This field is like a gold mine right now, with many discoveries uncovered everyday. And to be a part of this "gold rush", you have to keep a few things in mind:

- **Firstly, deep learning models require clear and informative data (and mostly big data) to train.** Try to imagine deep learning model as a child. It first observes how its parent walks. Then it tries to walk on its own, and with its every step, the child learns how to perform a particular task. It may fall a few times, but after few unsuccessful attempts, it learns how to walk. If you don't let it, it might not ever learn how to walk. The more exposure you can provide to the child, the better it is.
- **It is prudent to use Deep Learning for complex problems such as image processing.** Deep Learning algorithms belong to a class of algorithms called representation learning algorithms. These algorithms break down complex problems into simpler form so that they become understandable (or "representable"). Think of it as chewing food before you gulp. This would be harder for traditional (non-representation learning) algorithms.
- **When you have an appropriate type of deep learning to solve the problem.** Each problem has its own twists. So the data decides the way you solve the problem. For example, if the problem is of sequence generation, recurrent neural networks are more suitable. Whereas, if it is image related problem, you would probably be better of taking convolutional neural networks for a change.
- **Last but not the least, hardware requirements are essential for running a deep neural network model.** Neural nets were "discovered" long ago, but they are shining in the recent years for the main reason that computational resources are better and more powerful. If you want to solve a real life problem with these networks, get ready to buy some high-end hardware!

Do we need a lot of data to train deep learning models?

It is true that we need a large amount of data to train a typical deep learning model. But we can generally overcome this by using something called transfer learning. Let me explain thoroughly.

One of the barrier for using deep learning models for industry applications is where the data is not in huge amount. A few examples of data needed to train some of the popular deep learning models are:

	Google's Neural Machine Translation	VGG Network	DeepVideo
Objective	Text Translation	Image Category Classification	Video Category Classification
Data Size	6M pairs of English-French sentences	1.2M images with labeled categories	1.1M videos with labeled categories
Parameters	380M	140M	About 100M

However, a deep learning model trained on a specific task can be reused for different problem in the same domain even if the amount of data is not that huge. This technique is known as **Transfer Learning**.

For instance, we have a set of 1000 images of cats and dogs labeled as 1 and 0 (1 for cat and 0 for dog) and we have another set of 500 test images that we need to classify. So, instead of training a deep learning model on the data of 1000 images, we can use a **pre-trained** VGGNet model and retrain it on our data and use it to classify the unlabeled set of images. A pre-trained model may not be 100% accurate in your application, but it saves huge efforts required to reinvent the wheel.

You may have a look at this article to get a better intuition of using a pre-trained model.

Where can I find basic project ideas in order to practice deep learning?

To practice deep learning, ideas alone will not help. We also need labeled data to test our ideas using deep learning.

1. For beginners, getting started with the MNIST data is highly recommended. The dataset contains handwritten digits with their actual labels, i.e., numbers from 0 to 9. You can even compete in the Identify the Digits competition to evaluate your models
2. For intermediate users, this Age Detection challenge is a nice project to work on. The dataset consists of facial images of Indian movie actors. The task is to predict the age of a person from his or her facial attributes. For simplicity, the problem has been converted to a multiclass problem with classes as Young, Middle and Old.

You can also refer this list of exciting deep learning datasets and problems.

What are some Deep Learning interview questions?

Some common questions that may be asked on deep learning are:

1. How do deep learning models learn?
2. What are some limitations of a deep learning model?
3. What are the differences between feedforward neural networks and recurrent neural networks?
4. What are activation functions and why are they required?
5. What is a CNN and what are its applications?
6. What is pooling? How does it work?
7. What is a dropout layer and why is it used?
8. What is the vanishing gradient problem and how do we overcome that?
9. What are optimization functions? Name a few of the common optimization functions.

What is the future of Deep learning?

Deep learning has come a long way in recent years, but still has a lot of untapped potential. We are still in the nascent stages of this field, with new breakthroughs happening seemingly every day. One of the use-cases that we can definitely see in the suture is of automobile industry, where Deep Learning can revolutionize it by making self-driving cars a reality. While we don't have a crystal ball to predict the future, we can see deep learning models requiring less and less involvement from human data scientists and researchers.

In the immediate future, we can definitely see a trend where the knowledge of deep learning will be a skill required by every Data Science practitioner. In fact, you must have caught sight of a job position spurn out recently, called a "Deep Learning Engineer". This person is responsible to deploy and maintain Deep Learning models used by various departments of that company. Needless to say, there will be a huge demand of such people in the industry.

Currently, one of the limitations of DL is that it does what a human asks of it. It requires tons of data to learn it's target objective, and replicates that. This has induced bias in certain applications. We can see this improving over time such that the bias is eliminated in the training process.

We might even stop differentiating deep learning from the other types of learning, with time. It is primed to become a popular and commonly used field and will not require special branding efforts to market or sell.

There are a lot of cases still where researchers, after training a DL model, are unable to explain the 'why' behind it. "It's producing great results but why did you tune a hyperparameter a certain way?" Hopefully with the rapid advancement in DL, we will see this black box concept becoming history, and we can explain the intuition behind the decision it takes.

40 Questions to test your skill in Python for Data Science

https://www.analyticsvidhya.com/blog/2017/05/questions-python-for-data-science/

Question Context 1

You must have seen the show "How I met your mother". Do you remember the game where they played, in which each person drinks a shot whenever someone says "but, um". I thought of adding a twist to the game. What if you could use your technical skills to play this game?

To identify how many shots a person is having in the entire game, you are supposed to write a code.

Below is the subtitle sample script.

Note: Python regular expression library has been imported as re.

```
txt = '''450

 00:17:53,457 --> 00:17:56,175

 Okay, but, um,

 thanks for being with us.



451

 00:17:56,175 --> 00:17:58,616

 But, um, if there's any

 college kids watching,
```

```
452

 00:17:58,616 --> 00:18:01,610

 But, um, but, um, but, um,

 but, um, but, um,



453

 00:18:01,610 --> 00:18:03,656

 We have to drink, professor.

454

 00:18:03,656 --> 00:18:07,507

 It's the rules.

 She said "But, um"



455

 00:18:09,788 --> 00:18:12,515

 But, um, but, um, but, um...

 god help us all.

 '''
```

1) Which of the following codes would be appropriate for this task?
A) len(re.findall('But, um', txt))
B) re.search('But, um', txt).count()
C) len(re.findall('[B,b]ut, um', txt))
D) re.search('[B,b]ut, um', txt)).count()
Solution: (C)
You have to find both capital and small versions of "but" So option C is correct.

Question Context 2
Suppose you are given the below string
str = """Email_Address,Nickname,Group_Status,Join_Year
aa@aaa.com,aa,Owner,2014
bb@bbb.com,bb,Member,2015
cc@ccc.com,cc,Member,2017
dd@ddd.com,dd,Member,2016
ee@eee.com,ee,Member,2020
"""

In order to extract only the domain names from the email addresses from the above string (for eg. "aaa", "bbb"..) you write the following code:

```
for i in re.finditer('([a-zA-Z]+)@([a-zA-Z]+).(com)', str):

    print i.group(__)
```

2) What number should be mentioned instead of "__" to index only the domains?
Note: Python regular expression library has been imported as re.

A) 0
B) 1
C) 2
D) 3
Solution: (C)
Read syntax of regular expression re.

Question Context 3
Your friend has a hypothesis – "*All those people who have names ending with the sound of "y" (Eg: Hollie) are intelligent people.*" Please note: The name should end with the sound of 'y' but not end with alphabet 'y'. Now you being a data freak, challenge the hypothesis by scraping data from your college's website. Here's data you have collected.

Name	Marks
Andy	0
Mandi	10
Sandy	20
Hollie	18
Molly	19
Dollie	15

You want to make a list of all people who fall in this category. You write following code do to the same:

```
temp = []

 for i in re.finditer(pattern, str):

     temp.append(i.group(1))
```

3) What should be the value of "pattern" in regular expression?
Note: Python regular expression library has been imported as re.
A) pattern = '(i|ie)(,)'
B) pattern = '(i$|ie$)(,)'
C) pattern = '([a-zA-Z]+i|[a-zA-Z]+ie)(,)'
D) None of these
Solution: (B)
You have to find the pattern the end in either "i" or "ie". So option B is correct.

Question Context 4
Assume, you are given two lists:
a = [1,2,3,4,5]
b = [6,7,8,9]
The task is to create a list which has all the elements of a and b in one dimension.
Output:
a = [1,2,3,4,5,6,7,8,9]
4) Which of the following option would you choose?

A) a.append(b)
B) a.extend(b)
C) Any of the above
D) None of these
Solution: (B)
Option B is correct

5) You have built a machine learning model which you wish to freeze now and use later. Which of the following command can perform this task for you?
Note: Pickle library has been imported as pkl.
A) push(model, "file")
B) save(model, "file")
C) dump(model, "file")
D) freeze(model, "file")
Solution: (C)
Option C is correct

Question Context 6
We want to convert the below string in date-time value:

```
import time

str = '21/01/2017'

datetime_value = time.strptime(str,date_format)
```

6) To convert the above string, what should be written in place of *date_format*?
A) "%d/%m/%y"
B) "%D/%M/%Y"
C) "%d/%M/%y"
D) "%d/%m/%Y"
Solution: (D)
Option D is correct

Question Context 7
I have built a simple neural network for an image recognition problem. Now, I want to test if I have assigned the weights & biases for the hidden layer correctly. To perform this action, I am giving an identity matrix as input. Below is my identity matrix:
A = [1, 0, 0
0, 1, 0
0, 0, 1]
7) How would you create this identity matrix in python?
Note: Library numpy has been imported as np.
A) np.eye(3)
B) identity(3)
C) np.array([1, 0, 0], [0, 1, 0], [0, 0, 1])
D) All of these
Solution: (A)
Option B does not exist (it should be np.identity()). And option C is wrong, because the syntax is incorrect. So the answer is option A

8) To check whether the two arrays occupy same space, what would you do?
I have two numpy arrays "e" and "f".
You get the following output when you print "e" & "f"

```
print e

[1, 2, 3, 2, 3, 4, 4, 5, 6]

print f
```

```
[[1, 2, 3], [2, 3, 4], [4, 5, 6]]
```

When you change the values of the first array, the values for the second array also changes. This creates a problem while processing the data.
For example, if you set the first 5 values of e as 0; i.e.

```
print e[:5]



0
```

the final values of e and f are

```
print e



[0, 0, 0, 0, 0, 4, 4, 5, 6]



print f



[[0, 0, 0], [0, 0, 4], [4, 5, 6]]
```

You surmise that the two arrays must have the same space allocated.
A) Check memory of both arrays, if they match that means the arrays are same.
B) Do "np.array_equal(e, f)" and if the output is "True" then they both are same
C) Print flags of both arrays by e.flags and f.flags; check the flag "OWNDATA". If one of them is False, then both the arrays have same space allocated.
D) None of these
Solution: (C)
Option C is correct

Question Context 9
Suppose you want to join train and test dataset (both are two numpy arrays train_set and test_set) into a resulting array (resulting_set) to do data processing on it simultaneously. This is as follows:

```
train_set = np.array([1, 2, 3])

test_set = np.array([[0, 1, 2], [1, 2, 3]])

resulting_set --> [[1, 2, 3], [0, 1, 2], [1, 2, 3]]
```

9) How would you join the two arrays?
Note: Numpy library has been imported as np
A) resulting_set = train_set.append(test_set)
B) resulting_set = np.concatenate([train_set, test_set])
C) resulting_set = np.vstack([train_set, test_set])
D) None of these
Solution: (C)
Both option A and B would do horizontal stacking, but we would like to have vertical stacking. So option C is correct

Question Context 10
Suppose you are tuning hyperparameters of a random forest classifier for the Iris dataset.

Sepal_length	Sepal_width	Petal_length	Petal_width	Species
4.6	3.2	1.4	0.2	Iris-setosa
5.3	3.7	1.5	0.2	Iris-setosa
5.0	3.3	1.4	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
6.4	3.2	4.5	1.5	Iris-versicolor

10) What would be the best value for "random_state (Seed value)"?
A) np.random.seed(1)
B) np.random.seed(40)
C) np.random.seed(32)
D) Can't say
Solution: (D)
There is no best value for seed. It depends on the data.

Question 11
While reading a csv file with numpy, you want to automatically fill missing values of column "Date_Of_Joining" with date "01/01/2010".

Name	Age	Date_Of_Joining	Total_Experience
Andy	20	01/02/2013	0
Mandy	30	01/05/2014	10
Sandy	10		0
Bandy	40	01/10/2009	20

11) Which command will be appropriate to fill missing value while reading the file with numpy?
Note: numpy has been imported as np
A) filling_values = ("-", 0, 01/01/2010, 0)
temp = np.genfromtxt(filename, filling_values=filling_values)
B) filling_values = ("-", 0, 01/01/2010, 0)
temp = np.loadtxt(filename, filling_values=filling_values)
C) filling_values = ("-", 0, 01/01/2010, 0)
temp = np.gentxt(filename, filling_values=filling_values)
D) None of these
Solution: (A)
Option A is correct

12) How would you import a decision tree classifier in sklearn?
A) from sklearn.decision_tree import DecisionTreeClassifier
B) from sklearn.ensemble import DecisionTreeClassifier
C) from sklearn.tree import DecisionTreeClassifier
D) None of these
Solution: (C)
Option C is correct

13) You have uploaded the dataset in csv format on google spreadsheet and shared it publicly. You want to access it in python, how can you do this?
Note: Library StringIO has been imported as StringIO.

```
A)

link = https://docs.google.com/spreadsheets/d/...source =

StringIO.StringIO(requests.get(link).content))
```

```
data = pd.read_csv(source)

B)

link = https://docs.google.com/spreadsheets/d/...source = StringIO(request.get(link).content))

data = pd.read_csv(source)
```

C)

```
link = https://docs.google.com/spreadsheets/d/...source =

StringIO(requests.get(link).content))

data = pd.read_csv(source)
```

D) None of these
Solution: (A)
Option A is correct

Question Context 14
Imagine, you have a dataframe train file with 2 columns & 3 rows, which is loaded in pandas.
import pandas as pd

```
train  = pd.DataFrame({'id':[1,2,4],'features':[["A","B","C"],["A","D","E"],["C","D","F"]]})
```

Now you want to apply a lambda function on "features" column:

```
train['features_t'] = train["features"].apply(lambda x: " ".join(["_".join(i.split(" ")) for i

in x]))
```

14) What will be the output of following print command?

```
print train['features_t']
```

A)
0 A B C
1 A D E
2 C D F

B)
0 AB
1 ADE
2 CDF
C) Error
D) None of these
Solution: (A)
Option A is correct

Question Context 15
We have a multi-class classification problem for predicting quality of wine on the basis of its attributes. The data is loaded in a dataframe "df"

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	Alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25	67	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15	54	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17	60	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5

The quality column currently has values 1 to 10, but we want to substitute this by a binary classification problem. You want to keep the threshold for classification to 5, such that if the class is greater than 5, the output should be 1, else output should be 0.
15) Which of the following codes would help you perform this task?
Note: Numpy has been imported as np and dataframe is set as df.
A)

```
Y = df[quality].values

Y = np.array([1 if y >= 6 else 0 for y in Y])
```

B)

```
Y = df[quality].values()

Y = np.array([0 if y >= 6 else 1 for y in Y])
```

C)

```
Y = df[quality]

Y = np.array([0 if y >= 6 else 1 for y in Y])
```

D)None of these
Solution: (A)
Option A is correct

Question Context 16
Suppose we make a dataframe as

```
df = pd.DataFrame(['ff', 'gg', 'hh', 'yy'],

              [24, 12, 48, 30],

               columns = ['Name', 'Age'])
```

16) What is the difference between the two data series given below?

1. **df['Name'] and**
2. **df.loc[:, 'Name']**

Note: Pandas has been imported as pd
A) 1 is view of original dataframe and 2 is a copy of original dataframe.
B) 2 is view of original dataframe and 1 is a copy of original dataframe.
C) Both are copies of original dataframe.
D) Both are views of original dataframe
Solution: (B)
Option B is correct. Refer the official docs of pandas library.

Question Context 17
Consider a function "fun" which is defined below:

```
def fun(x):

    x[0] = 5

    return x
```

Now you define a list which has three numbers in it.
g = [10,11,12]
17) Which of the following will be the output of the given print statement:

```
print  fun(g), g
```

A) [5, 11, 12] [5, 11, 12]
B) [5, 11, 12] [10, 11, 12]
C) [10, 11, 12] [10, 11, 12]
D) [10, 11, 12] [5, 11, 12]
Solution: (A)
Option A is correct

Question Context 18
Sigmoid function is usually used for creating a neural network activation function. A sigmoid function is denoted as

```
def sigmoid(x):

   return (1 / (1 + math.exp(-x)))
```

18) It is necessary to know how to find the derivatives of sigmoid, as it would be essential for backpropagation. Select the option for finding derivative?
A)

```
import scipy



Dv = scipy.misc.derive(sigmoid)
```

B)

```
from sympy import *



x = symbol(x)
```

```
y = sigmoid(x)



Dv = y.differentiate(x)
```

C)

```
Dv = sigmoid(x) * (1 - sigmoid(x))
```

D) None of these
Solution: (C)
Option C is correct

Question Context 19
Suppose you are given a monthly data and you have to convert it to daily data.
For example,

ID	Electricity_Usage	Month
1	2000	1
2	20	2
3	4000	3
4	40	4

ID	Electricity_Usage	Date	Month
1	100	1	1
1	100	2	1
1	100	3	1
1	100	4	1
1	100	5	1

For this, first you have to expand the data for every month (considering that every month has 30 days)
19) Which of the following code would do this?
Note: Numpy has been imported as np and dataframe is set as df.
A) new_df = pd.concat([df]*30, index = False)
B) new_df = pd.concat([df]*30, ignore_index=True)
C) new_df = pd.concat([df]*30, ignore_index=False)
D) None of these
Solution: (B)
Option B is correct

Context: 20-22
Suppose you are given a dataframe df.

```
df = pd.DataFrame({'Click_Id':['A','B','C','D','E'],'Count':[100,200,300,400,250]})
```

20) Now you want to change the name of the column 'Count' in df to 'Click_Count'. So, for performing that action you have written the following code.

```
df.rename(columns = {'Count':'Click_Count'})
```

What will be the output of print statement below?

```
print df.columns
```

Note: Pandas library has been imported as pd.
A) ['Click_Id', 'Click_Count']
B) ['Click_Id', 'Count']
C) Error
D) None of these

Solution: (B)
Option B is correct

Context: 20-22
Suppose you are given a data frame df.

```
df = pd.DataFrame({'Click_Id':['A','B','C','D','E'],'Count':[100,200,300,400,250]})
```

21) In many data science projects, you are required to convert a dataframe into a dictionary. Suppose you want to convert "df" into a dictionary such that 'Click_Id' will be the key and 'Count' will be the value for each key. Which of the following options will give you the desired result?
Note: Pandas library has been imported as pd
A) set_index('Click_Id')['Count'].to_dict()
B) set_index('Count')['Click_Id'].to_dict()
C) We cannot perform this task since dataframe and dictionary are different data structures
D) None of these
Solution: (A)
Option A is correct

22) In above dataframe df. Suppose you want to assign a df to df1, so that you can recover original content of df in future using df1 as below.

```
df1 = df
```

Now you want to change some values of "Count" column in df.

```
df.loc[df.Click_Id == 'A', 'Count'] += 100
```

Which of the following will be the right output for the below print statement?

```
print df.Count.values,df1.Count.values
```

Note: Pandas library has been imported as pd.
A) [200 200 300 400 250] [200 200 300 400 250]
B) [100 200 300 400 250] [100 200 300 400 250]
C) [200 200 300 400 250] [100 200 300 400 250]
D) None of these
Solution: (A)
Option A is correct

23) You write a code for preprocessing data, and you notice it is taking a lot of time. To amend this, you put a bookmark in the code so that you come to know how much time is spent on each code line. To perform this task, which of the following actions you would take?

1. **You put bookmark as time.sleep() so that you would know how much the code has "slept" literally**
2. **You put bookmark as time.time() and check how much time elapses in each code line**
3. **You put bookmark as datetime.timedelta(), so that you would find out differences of execution times**
4. **You copy whole code in an Ipython / Jupyter notebook, with each code line as a separate block and write magic function %%timeit in each block**

A) 1 & 2
B) 1,2 & 3
C) 1,2 & 4
D) All of the above
Solution: (C)
Option C is correct

24) How would you read data from the file using pandas by skipping the first three lines?
Note: pandas library has been imported as pd In the given file (email.csv), the first three records are empty.

```
,,,
```

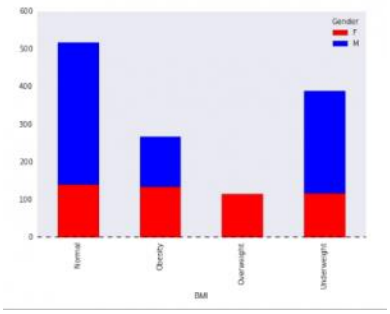
```
,,,

,,,

Email_Address,Nickname,Group_Status,Join_Year

aa@aaa.com,aa,Owner,2014

bb@bbb.com,bb,Member,2015

cc@ccc.com,cc,Member,2017

dd@ddd.com,dd,Member,2016
```

A) read_csv('email.csv', skip_rows=3)
B) read_csv('email.csv', skiprows=3)
C) read_csv('email.csv', skip=3)
D) None of these
Solution: (B)
Option B is correct

25) What should be written in-place of "method" to produce the desired outcome?
Given below is dataframe "df":

EMPID	Gender	Age	Sales	BMI	Income
E001	M	34	123	Normal	350
E002	F	40	114	Overweight	450
E003	F	37	135	Obesity	169
E004	M	30	139	Underweight	189
E005	F	44	117	Underweight	183
E006	M	36	121	Normal	80
E007	M	32	133	Obesity	166
E008	F	26	140	Normal	120
E009	M	32	133	Normal	75
E010	M	36	133	Underweight	40

Now, you want to know whether BMI and Gender would influence the sales.
For this, you want to plot a bar graph as shown below:



The code for this is:

```
var = df.groupby(['BMI','Gender']).Sales.sum()



var.unstack().plot(kind='bar', method,  color=['red','blue'], grid=False)
```

A) stacked=True
B) stacked=False
C) stack=False
D) None of these
Solution: (A)
It's a stacked bar chart.

26) Suppose, you are given 2 list – City_A and City_B.
City_A = ['1','2','3','4']
City_B = ['2','3','4','5']
In both cities, some values are common. Which of the following code will find the name of all cities which are present in "City_A" but not in "City_B".
A) [i for i in City_A if i not in City_B]
B) [i for i in City_B if i not in City_A]
C) [i for i in City_A if i in City_B]
D) None of these
Solution: (A)
Option A is correct

Question Context 27
Suppose you are trying to read a file "temp.csv" using pandas and you get the following error.

```
Traceback (most recent call last):

File "<input>", line 1, in<module>

UnicodeEncodeError: 'ascii' codec can't encode character.
```

27) Which of the following would likely correct this error?
Note: pandas has been imported as pd
A) pd.read_csv("temp.csv", compression='gzip')
B) pd.read_csv("temp.csv", dialect='str')
C) pd.read_csv("temp.csv", encoding='utf-8')
D) None of these
Solution: (C)
Option C is correct, because encoding should be 'utf-8'

28) Suppose you are defining a tuple given below:
tup = (1, 2, 3, 4, 5)
Now, you want to update the value of this tuple at 2nd index to 10. Which of the following option will you choose?

A) tup(2) = 10
B) tup[2] = 10
C) tup{2} = 10
D) None of these
Solution: (D)
A tuple cannot be updated.

29) You want to read a website which has url as "www.abcd.org". Which of the following options will perform this task?
A) urllib2.urlopen(www.abcd.org)
B) requests.get(www.abcd.org)
C) Both A and B
D) None of these
Solution: (C)
Option C is correct

Question Context 30
Suppose you are given the below web page

```
html_doc = """
 <!DOCTYPE html>
 <htmllang="en">
 <head>
 <metacharset="utf-8">
 <metaname="viewport" content="width=device-width">
 <title>udacity/deep-learning:   Repo   for   the   Deep   Learning   Nanodegree   Foundations
program.</title>
 <linkrel="search"     type="application/opensearchdescription+xml"     href="/opensearch.xml"
title="GitHub">
 <linkrel="fluid-icon" href="https://github.com/fluidicon.png" title="GitHub">
 <metaproperty="fb:app_id" content="1401488693436528">
 <linkrel="assets" href="https://assets-cdn.github.com/">
 ...


"""
```

30) To read the title of the webpage you are using BeautifulSoup. What is the code for this?
Hint: You have to extract text in title tag

A. from bs4 import BeautifulSoup
soup =BeautifulSoup(html_doc,'html.parser')
print soup.title.name

B. from bs4 import BeautifulSoup
soup =BeautifulSoup(html_doc,'html.parser')
print soup.title.string
C. from bs4 import BeautifulSoup
soup=BeautifulSoup(html_doc,'html.parser')
print soup.title.get_text
D. None of these

Solution: (B)
Option B is correct
Question Context 31
Imagine, you are given a list of items in a DataFrame as below.
D = ['A','B','C','D','E','AA','AB']
Now, you want to apply label encoding on this list for importing and transforming, using LabelEncoder.

```
from sklearn.preprocessing import LabelEncoder



le = LabelEncoder()
```

31) What will be the output of the print statement below ?

```
print le.fit_transform(D)
```

A. array([0, 2, 3, 4, 5, 6, 1])
B. array([0, 3, 4, 5, 6, 1, 2])
C. array([0, 2, 3, 4, 5, 1, 6])
D. Any of the above

Solution: (D)
Option D is correct

32) Which of the following will be the output of the below print statement?

```
print df.val == np.nan
```

Assume, you have defined a data frame which has 2 columns.

```
import numpy as np



df = pd.DataFrame({'Id':[1,2,3,4],'val':[2,5,np.nan,6]})
```

A) 0 False
1 False
2 False
3 False
B) 0 False
1 False
2 True
3 False
C) 0 True
1 True
2 True
3 True
D) None of these
Solution: (A)
Option A is correct

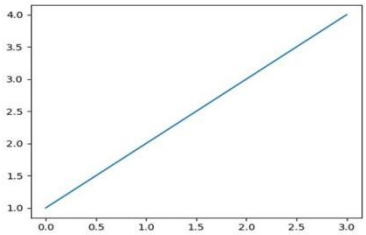
33) Suppose the data is stored in HDFS format and you want to find how the data is structured. For this, which of the following command would help you find out the names of HDFS keys?
Note: HDFS file has been loaded by h5py as hf.
A) hf.key()
B) hf.key
C) hf.keys()
D) None of these
Solution: (C)
Option C is correct

Question Context 34
You are given reviews for movies below:
reviews = ['movie is unwatchable no matter how decent the first half is . ', 'somewhat funny and well paced action thriller that has jamie foxx as a hapless fast talking hoodlum who is chosen by an overly demanding', 'morse is okay as the agent who comes up with the ingenious plan to get whoever did it at all cost .']
Your task is to find sentiments from the review above. For this, you first write a code to find count of individual words in all the sentences.

```
counts = Counter()



for i in range(len(reviews)):

    for word in reviews[i].split(value):

        counts[word] += 1
```

34)What value should we split on to get individual words?
A. ' '
B. ','
C. '.'
D. None of these
Solution: (A)
Option A is correct

35) How to set a line width in the plot given below?



For the above graph, the code for producing the plot was

```
import matplotlib.pyplot as plt

plt.plot([1,2,3,4])

plt.show()
```

A. In line two, write plt.plot([1,2,3,4], width=3)
B. In line two, write plt.plot([1,2,3,4], line_width=3

C. In line two, write plt.plot([1,2,3,4], lw=3)
D. None of these

Solution: (C)
Option C is correct

36) How would you reset the index of a dataframe to a given list? The new index is given as: new_index=['Safari','Iceweasel','Comodo Dragon','IE10','Chrome']
Note: df is a pandas dataframe

	http_status	response_time
Firefox	200	0.04
Chrome	200	0.02
Safari	404	0.07
IE10	404	0.08
Konqueror	301	1.00

A) df.reset_index(new_index,)
B) df.reindex(new_index,)
C) df.reindex_like(new_index,)
D) None of these

Solution: (A)
Option A is correct

37) Determine the proportion of passengers survived based on their passenger class.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th…	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

A. crosstab(df_train['Pclass'], df_train['Survived'])

Solution: (A)

Option A is correct

38) You want to write a generic code to calculate n-gram of the text. The 2-gram of this sentence would be [["this, "is"], ["is", "a"], ["a, "sample"], ["sample", "text"]]
Which of the following code would be correct?
For a given a sentence:
'this is a sample text'.

A. def generate_ngrams(text, n):
words = text.split()
output = [] for i in range(len(words)-n+1):
append(words[i:i+n])
return output

Solution: (B)

Option B is correct

39) Which of the following code will export dataframe (df) in CSV file, encoded in UTF-8 after hiding index & header labels.

A. df_1.to_csv('../data/file.csv',encoding='utf-8′,index=False,header=False)

Solution: (C)

Option C is correct

40) Which of the following is a correct implementation of mean squared error (MSE) metric?

Note: numpy library has been imported as np.

A. def MSE(real_target, predicted_target):
return np.sqrt(np.mean((np.square(real_target) – np.square(predicted_target))))

Solution: (B)

Option B is correct

Sql

https://www.analyticsvidhya.com/blog/2017/05/questions-sql-for-all-aspiring-data-scientists/

Questions & Solution

1) Which of the following option(s) is/are correct?

1. **SQL is case insensitive**
2. **SQL is a programming language**

A) 1

Solution: (A)

SQL is a querying language and it is not case sensitive.

2) What is true for a Null value in SQL?

1. **Null +1 = Null**
2. **Null + 1 = 1**
3. **Null * 2 = Null**
4. **Null * 2 = 0**

A) 1 and 3

Solution: (A)

NULL represents a unknown value so adding anything to null value will give a null value in result.

3) Which of the following is not an aggregate function in SQL?

C) DISTINCT()

Solution: (C)

All of the functions except DISTINCT function given in the question is an example of aggregate function.

4) Which of the following option is true for the below queries?

```
Query1: select 1 in (1,2,1,2);



Query2: select 1 in (1,2);
```

B) Both queries will give same outputs.

Solution: (B)
Both query will return the same output.

5) Which of the following cannot be a superkey in a relational schema with attributes A,B,C,D,E and primary key AD?
C) A B C E

Solution: (C)
The attributes "A", "D" should be present in the super key. Option C doen't have "D" attribute so C would be the right answer.

Questions Context 6-7
You run the following Queries in the following order:
Create a table "Me" using the below SQL query.

```
Query1: Create table Me(name varchar(20), salary int);
```

Next, you create a view based on "Me" table by using the following query.

```
Query2: Create view me_view as select name from me;
```

Finally, you run the following query:

```
Query3: Drop table Me;
```

6) Which of the following statements are true?
1. **Query3 will give an error**
2. **Query3 will run smoothly**
3. **Query2 will give an error**
4. **Query2 will run smoothly**

B) 1 and 4

Solution: (B)
Query 2 is used for creating the view on table "Me" so it would run fine but if you run the Query3 it will generate the below error.
ERROR: cannot drop table me because other objects depend on it
DETAIL: view me_view depends on table me
HINT: Use DROP … CASCADE to drop the dependent objects too.

7) Now, you have changed the 'Query3' as below.

```
Query3: DROP TABLE Me CASCADE;
```

And, you also want to run below query on the same table.

```
Query4: select * from me_view;
```

Which of the following statements are true for such cases?
1. **Query3 will give an error**
2. **Query3 will run smoothly**
3. **Query4 will give an error**
4. **Query4 will run smoothly**

C) 2 and 3

Solution: (C)
If you drop the base table using cascade it will drop the base table as well as view table also so Query 3 will run fine but Query 4 will give an error.

8) Imagine, you have a column 'A' in table1 which is a primary key and it refers to column 'B' in table2. Further, column 'A' has only three values (1,2,3). Which of the following options is / are correct?
1. **Inserting a value 4 in column 'A' of table1 will result in an error**
2. **Inserting a value 4 in column 'B' of table2 will result in an error**
3. **Inserting a value 4 in column 'A' of table1 will be successful**
4. **Inserting a value 4 in column 'B' of table2 will be successful**

A) 1 and 2
B) 2 and 3

C) 1 and 2
D) 3 and 4
Solution: (B)
You can insert any value except the duplicate values in column A in table 1 but you cannot insert the values other then 1,2 and 3 in columns B in table2 due to foreign key integrity because it is referenced by the column A.

9) Consider a table T1 which contains 'Column A' in it. Column A is declared as a primary key. Which of the following is true for column "A" in T1?

1. **Column A can contain values [1,2,3,4,1]**
2. **Column A cannot contain values [1,2,3,4,1]**
3. **Column A can contain values [1,2,3,4,NULL]**
4. **Column A cannot contain values [1,2,3,4,NULL]**

A) 1 and 4
B) 2 and 4
C) 1 and 3
D) 2 and 3
Solution: (B)
A primary key column cannot contain duplicate and null values.

10) Imagine you have a table "T1" which has three columns "A", "B" and "C" where A is a "primary key". Which of the following query will return number of rows present in the table T1

```
Query1: SELECT COUNT(A) FROM T1;


Query2: SELECT COUNT(*) FROM T1;


Query3: SELECT COUNT(A,B) FROM T1;
```

A) 1 and 2
Solution: (A)
Query1 and Query2 will return the same output.

11) Which of the following statement describes the capabilities of UPDATE command most appropriately?
D) It can update multiple values of multiple columns
Solution: (D)

12) What is true about indexing in a database?
A) Search will be faster after you have indexed the database
Solution: (A)
Option A is correct. Read more here.

13) Consider three tables T1, T2 and T3. Table T1, T2 and T3 have 10, 20, 30 number of records respectively. Further, there are some records which are common in all three tables.
You want to apply a cartesian product on these three tables. How many rows will be available in cartesian product of these tables?
A) 6000
Solution: (A)

14) Tables A, B have three columns (namely: 'id', 'age', 'name') each. These tables have no null values and there are 100 records in each of the table.
Here are two queries based on these two tables A and B.

```
Query1: SELECT A.id FROM A WHERE A.age > ALL (SELECT B.age FROM B WHERE B.name = 'Ankit')


Query2: SELECT A.id FROM A WHERE A.age > ANY (SELECT B.age FROM B WHERE B.name = 'Ankit')
```

Which of the following statement is correct for the output of each query?
C) The number of tuples in the output Query 1 will be less than or equal to the output of Query 2

Solution: (C)
Answer C is correct because natural join always give either same or less number of rows if you compare it with cartesian product. To know more read from this tutorial.

15) What will be the output of the following query in PostgreSQL?

```
Query 1: SELECT DATE_PART('year', '2012-01-01'::date) - DATE_PART('year', '2011-10-02'::date);
```

A)

?column?

365

(1 row)

B)

?column?

9

(1 row)

C)

?column?

1

(1 row)

D)

?column?

305

(1 row)

Solution: (C)
It will give the year differece in output so answer C is correct.

16) Imagine you are given the following table named "AV".

ID	NAME	DOB
1	ANKIT	1990-09-19
2	FAIZAN	1993-01-01
3	SUNIL	1985-11-02
4	SAURABH	1994-11-12
5	KUNAL	1983-11-12

And you want to run the following queries Q1, Q2 and Q3 given below.

```
Q1: DROP TABLE AV;


Q2: DELETE FROM AV;
```

```
Q3: SELECT * FROM AV;
```

Which sequence for the three queries will not result in an error?
D) Q2 -> Q3 -> Q1
Solution: (D)
"DROP TABLE" will drop the table as well as it's reference. So, you can't access the table once you have dropped it. But in case of "DELETE TABLE" reference will not be droped so you can still access the table if you use "DELETE TABLE" command.

17) Imagine you are given the following table named "AV".

id	name	dob	sal
1	ANKIT	1990-09-19	200
2	FAIZAN	1993-01-01	300
3	SUNIL	1985-11-02	500
4	SAURABH	1994-11-12	350
5	KUNAL	1983-11-12	600

You apply the following query Q1 on AV, which is given below:

```
Q1: SELECT * FROM AV WHERE SAL BETWEEN 200 AND 500;
```

What will be the output for query Q1?
A)

Id	Name	Dob	sal
1	ANKIT	1990-09-19	200
2	FAIZAN	1993-01-01	300
3	SUNIL	1985-11-02	500
4	SAURABH	1994-11-12	350

Solution: (A)
The boundary salaries (200 and 500) will also be in the out so A is the right answer.

18) Imagine you are given the following table named "AV".

id	name	dob	sal
1	ANKIT	1990-09-19	200
2	FAIZAN	1993-01-01	300
3	SUNIL	1985-11-02	500
4	SAURABH	1994-11-12	350
5	KUNAL	1983-11-12	600

What would be the output for the following query?

```
Query: SELECT ID, SUBSTRING(NAME,2,5) "sub_name" FROM AV;
```

B)

Id	Sub_name
1	NKIT
2	AIZAN
3	UNIL

4	AURAB
5	UNAL

Solution: (B)

Question Context 19-21

Assume you are given the two tables AV1 and AV2 which represent two different departments of AV.

AV1 TABLE

Id	name
1	ANKIT
2	FAIZAN
3	SUNIL
4	SAURABH
5	KUNAL

AV2 TABLE

Id	name
1	DEEPAK
2	SWATI
3	DEEPIKA
4	PRANAV
5	KUNAL
5	SUNIL

19) Now, you want the names of all people who work in both the departments. Which of the following SQL query would you write?

A) SELECT NAME FROM AV1 INTERSECT SELECT NAME FROM AV2;

Solution: (A)

INTERSECT would be used for such output.

20) What is the output for the below query?

```
Query: SELECT NAME FROM AV1 EXCEPT SELECT NAME FROM AV2;
```

A)

name

FAIZAN

SAURABH

ANKIT

B)

Solution: (A)

This query will give the names in AV1 which are not present in AV2.

21) What will be the output of below query?

```
Query: SELECT NAME FROM AV1 NATURAL JOIN AV2;
```

A)

name

SUNIL

KUNAL

B)

name

SUNIL

C) None of these

Solution: (B)

Question Context 22-24

Suppose you are given the below table called A_V.

Id	Name	Sal	dept
1	ANKIT	100	DS
2	FAIZAN	200	DS
3	SUNIL	800	ALL
4	SAURABH		INTERN
5	KUNAL	1000	ALL

22) What is the output for the below query?

```
Query: SELECT DEPT, AVG(SAL) FROM A_V GROUP BY DEPT,NAME;
```

A)

dept	avg
DS	100.0000000000000000
ALL	800.0000000000000000
ALL	1000.0000000000000000
DS	200.0000000000000000

B)

dept	avg
INTERN	
DS	100.0000000000000000
ALL	800.0000000000000000
ALL	1000.0000000000000000
DS	200.0000000000000000

C) ERROR

D) None of these

Solution: (B)

23) What is the output for the below query?

```
Query: SELECT COALESCE(sal,2)+100 AS sal FROM A_V;
```

A)

Sal
202

302

902

Null

1102

B)

Sal

200

300

900

102

1100

C)

Sal

202

302

902

102

1102

D) None of these

Solution: (B)

First replace null value will be replaced to 2 using COALESCE then 100 will be added.

24) What is the output for the below query?

```
Query: SELECT * FROM a_v WHERE name In ('Ankit', 'Faizan');
```

A)

Id	Name	Sal	Dept
1	ANKIT	100	DS
2	FAIZAN	200	DS

B) Empty output

Solution: (B)

SQL is not case sensitive but when you search for something in a string column it becomes case sensitive. So output will have zero rows because 'Ankit' != 'ANKIT' and 'Faizan' != 'FAIZAN'.

25) You are given a string " AnalyticsVidhya ". The string contains two unnecessary spaces – one at the start and another at the end. You find out the length of this string by applying the below queries.

```
Query1: SELECT length(rtrim(' AnalyticsVidhya '));


Query2: SELECT length(ltrim(' AnalyticsVidhya '));
```

```
Query3: SELECT length(rtrim(ltrim(' AnalyticsVidhya ')));


Query4: SELECT length(ltrim(rtrim(' AnalyticsVidhya ')));
```

If op1, op2, op3, op4 are the output of the Query 1, 2, 3 and 4 respectively, what will be the correct relation between these four queries?

1. **op1 = op2 and op3 = op4**
2. **op1< op3 and op2 > op4**
3. **op1 > op3 and op2< op4**
4. **op1 > op3 and op2 > op4**

D) 1 and 4
Solution: (D)
Option D is correct. For more information read from this tutorial.

Questions Context 26-27
Below you are given a table "split".

uk	id
ANKIT-001-1000-AV1	1
SUNIL-002-2000-AV2	2
FAIZY-007-3000-AV1	3

26) Now, you want to apply a query on this.

```
Query: SELECT SPLIT_PART(uk, '-', 0) FROM SPLIT;
```

What is the output for the above query?
E) Error
Solution:(E)
The query will give the below error.
ERROR: field position must be greater than zero
27) In the above table "split", you want to replace some characters using "translate" command. Which of the following will be the output of the following query?

```
Query: SELECT TRANSLATE(UK, 'ANK', '123') FROM SPLIT;
```

A)

translate
123IT-001-1000-1V1
SU2IL-002-2000-1V2
F1IZY-007-3000-1V1

Solution: (A)
In the above query character "A" will replace to "1", "B" to 2 and "C" to 3.

28) Which of the following query will list all station names which contain their respective city names. For example, station "Mountain View Caltrain Station" is for city "Mountain View".
Refer to the table below this question.

Index	Station_name	City
1	Mountain View Caltrain Station	Mountain View
2	Dlf Square Phase 2	Dlf Square
3	Sikandarpur Metro Gurgaon	Gurgaon

4	Akola Station	Akola

A) select * from station where station_name like '%' || city || '%';
Solution: (A)

29) Consider the following legal instance of a relational schema S with attributes ABC.
Which of following functional dependencies is/are not possible?

1. A->B
2. C->A
3. B->A

D) None of above
Solution: (D)
Read from this tutorial.
30) Suppose you have a table called "Student" and this table has a column named "marks". Now, you apply Query1 on "Student" table.

```
Query 1: SELECT * FROM Student where marks * 100 > 70;
```

After this, you create an index on column "marks" and then you re-run Query 2 (same as Query 1).

```
Query 2: SELECT * FROM Student where marks * 100 > 70;
```

If Query 1 is taking time T1 and Query 2 is taking time T2.
Which of the following is true for the query time?
C) T1 ~ T2
Solution: (C)
To search fast you need to create the index on marks*100 but in the question we have created the index on marks.
31) Suppose you have 1000 records in a table called "Customers". You want to select top 100 records from it. Which of the below commands can you use?

```
1. SELECT TOP 100 * FROM Customers;2. SELECT TOP 10 PERCENT * FROM Customers;
```

C) 1 and 2
Solution: (C)
Both query can be used to get the desired output.

32) Which of the following is the outcome of the following query?

```
Query: SELECT REPLACE( 'Faizan and Ankit are close friends', 'Faizan', 'Ankit' )
```

Ankit and Ankit are close friends
Solution: (B)
"Faizan" will be replaced by "Ankit".

33) Which one of the following queries always gives the same answer as the nested "Query" shown below.

```
Query: select * from R where a in (select S.a from S)
```

C) select R.* from R,(select distinct a from S) as S1 where R.a=S1.a
Solution: (C)
Option C is correct.

Question Context 34-35
Consider the following table "avian" (id, name, sal).

Id	Name	Sal
1	Ankit	20
1	Faizan	10
2	Faizan	10

3	Faizan	20
1	Sunil	10
2	Sunil	20
1	Kunal	10
10	Nikam	30

34) Which of the following options will be required at the end of the following SQL query?

```
Query: SELECT P1.name FROM avian P1
```

So that the appended query finds out the name of the employee who has the maximum salary?
D) WHERE P1.sal >= Any (select max(P2.sal) from avian P2)
Solution: (D)
B – Returns the addresses of all theaters.
C – Returns null set. max() returns a single value and there won't be any value > max.
D – Returns null set. Same reason as C. All and ANY works the same here as max returns a single value.

35) Which of the following options can be used to find the name of the person with second highest salary?
A) select max(sal) from avian where sal < (select max(sal) from avian)
B) Both
Solution: (B)
Query in the option B
"(select max(sal) from avian)"
first return the highest salary(say H) then the query
"(select max(sal) from avian where sal < H)"
will search for highest salary which is less then H.

Question Context 36-39
Suppose you are given a database of bike sharing which has three tables: Station, Trip and Weather.
Station Table

station_id	station_name	city	zip_code
2	M	S1	95113
3	N	S2	95112
4	L	S3	95114
5	G	S4	95115
6	O	San Jose	95115
1	K	San Jose	95115

Trip Table

Id	Duration	start_time	start_station_name	start_station_id	end_time	end_station_name	end_station_name	bike_id
5081	183	2013-08-29 22:08:00	M	2	2013-08-29 22:12:00	M	2	309
5082	100	2013-08-01 22:08:00	N	3	2013-08-01 22:12:00	L	4	301

5083	283	2013-08-02 22:08:00	O	6	2013-08-02 22:12:00	G	5	303
5084	23	2013-08-09 22:08:00	M	2	2013-08-10 22:12:00	O	7	305

Weather Table

zip_code	max_temp	min_temp
95113	74	61
95112	70	21
95115	91	40

36) Imagine, you run following query on above schema.

```
Query: select  city , count( station_id ) as cnt from station group by city order by cnt desc

, city asc;
```

Which of the following option is correct for this query?
A) This query will print city name and number of stations sorted by number of stations in increasing magnitude. If numbers of stations are same, it will print by decreasing order of city name.
Solution: (A)
A is correct answer.

37) Which of the following query will find the percentage (round to 5 decimal places) of self-loop trips (which start and end at the same station) among all trips?
A)

```
select round(self_loop_cnt.cnt * 1.0/trip_cnt.cnt,5) as percentage from (select count(*) as

cnt from trip where start_station_id = end_station_id) as self_loop_cnt ,(select count(*) as

cnt from trip) as trip_cnt;
```

Solution: (A)
Query in option A will give the desired result

38 Which of the following statements is / are true for the below query?

```
Query: select station_name from station where zip_code = (select zip_code from weather where

max_temp = (select max(max_temp) from weather))
```

Note: All the zip_code are present in table weather also present in station table

1. **The query will return names of all stations where maximum temperature was found across all cities.**
2. **This query will always give more than zero records.**

A) 1 and 2
Solution: (A)

39) What will be the output of the following query?

```
Query: select end_time , (select sum(duration) from trip as i where i.bike_id = 301 and
i.end_time <= o.end_time ) as ac from trip as o where o.bike_id = 301 order by ac asc ;
B.
 end_time       | ac
---------------------+-----
2013-08-01 22:12:00 | 100
2013-08-09 22:12:00 | 653
```

Solution: (B)
This query will find a cumulative traveling durations of bike 301.

Question Context 40-42
Suppose you are given 4 tables: Team, Player, Game and GameStats. Below are the SQL statements which create these tables.

```
CREATE TABLE Team (
    name varchar(50) PRIMARY KEY,
    city varchar(50));


CREATE TABLE Player (
    playerID integer PRIMARY KEY,
    name varchar(50),
    position varchar(10),
    height integer,
    weight integer,
    team varchar(50) REFERENCES Team(name),
    CHECK (position='Guard' OR position='Center' OR position='Forward'));;


CREATE TABLE Game (
    gameID integer PRIMARY KEY,
    hometeam varchar(50) REFERENCES Team(name) NOT NULL,
```

```
    awayteam varchar(50) REFERENCES Team(name) NOT NULL,
    homescore integer,
    awayscore integer,
    CHECK (hometeam <> awayteam));


CREATE TABLE GameStats (
    playerID integer REFERENCES Player(playerID) NOT NULL,
    gameID integer REFERENCES Game(gameID) NOT NULL,
    points integer,
    assists integer,
    rebounds integer,
    PRIMARY KEY (playerID, gameID)
```

40) Which of the following query will return distinct names of the players who play at "Guard" Position and their name contains "Jo". (ORDER BY A)
B) SELECT name FROM player WHERE position='Guard' AND name LIKE '%Jo%' ORDER BY name
Solution: (B)
This query Finds any values that have "Jo" in any position using '%jo%' expression in command. Notice that 'Jo' is different then 'jo' because expression in like operator is case sensitive.

41) What will be the output for the below query?

```
Query:  SELECT COUNT(*) AS num_of_games    FROM player p1, player p2, gamestats g1, gamestats
g2     WHERE   p1.name='Saurabh'   AND   p2.name='Faizan'   AND   g1.playerid=p1.playerid   AND
          g2.playerid=p2.playerid AND g1.gameid=g2.gameid AND g1.points > g2.points
```

A) Return the number of games where 'Saurabh' scored more points than 'Faizan'
Solution: (A)

42) What is the expected output of the following query?

```
Query: SELECT s.playerid, AVG(s.points) AS average_points FROM (SELECT st.playerid, st.points
FROM player p, game g, gamestats st WHERE st.gameid=g.gameid AND p.team=g.hometeam AND
p.playerid= st.playerid) s GROUP BY s.playerid ORDER BY s.playerid
```

A) List all players' playerIDs and their average points in all home games that they played in (ORDER BY Players' playerID)
Solution: (A)

Clearly explain dimensionality reduction stating its usability and benefits?

The process where the number of featured variables is minimized, taking into account a set of principal variables, can be termed as dimensionality reduction. Now it can be said that dimensionality reduction technique can be used in order to know how much a variable can contribute to representing the information. The technique that is mostly preferred and also used to know the contribution of a variable are nonlinear and trial and error technique. Some of the benefits of this process are known to be speeding computation, minimizing storage space and reduction in data dimension.

How can a user handle missing or corrupted data in a data set?

The best possible way to find a corrupted data in a data set is by replacing the variable with another value or by introducing new column and rows. Now it has noted that are few other techniques to find the missing data are known as the fillna() method and other is known as the dropna() and insull() method.

What is clustering algorithm?

Clustering algorithm can be defined as the unsupervised learning technique which used for finding out the structure of an unlabelled data. This clustering could be defined as the data which is similar in their orientation but dissimilar when compared to other clusters.

How can exploratory data analysis or EDA can be performed?

The main goal of this algorithm is to find out the information about the data before it is being applied to any model. Basically when EDA is performed the IT professionals look for some global insights which is to check out the mean variable of each specific class. After this action is performed then the IT professionals run a panda known df.info () to check for any of the variables are categorical or continuous like int, float or string.

How to decide on which machine learning model to use?

In deciding which machine learning model to use one should always keep the no freelunchtheorem at the back of their mind. Now if the user wants to estimate a direct relationship between the output variable and single variable then choosing a single regression model or multiple regression model is the best choice. Now if the user wants to determine complex nonlinear relationships then choosing neutral network model is the ideal choice.

How to use convolutions of pictures instead of FC layers?

This can be explained in two parts, firstly the users need to derive the information from the image since FC will have no actual information. The second part is using convolution neural networks which is useful since the FC acts as its own detector.

What makes CNN translation invariant?

Now it has to be noted that each convolution acts on its own way or acts as its own feature detector meaning if the user wants to perform image detection then convolution acts as a own feature detector. Now it is irrelevant where the image is since convolution will be acted in the entire image.

Why is there max polling classification in CNN's?

CNN's contains max pooling classification because it has the ability to minimize the computation process since the feature maps tend to be smaller in size than that of pooling. In addition, with the help of max poling classification of more translation can be found invariance.

Why does CNN's have encoder-decoder style or structure?

CNN's have the encoder-decoder structure for two reasons, firstly the encoder is helpful in extracting the feature network and the decoder is used to decode the image in segments and thenupscale it back to its original size

What is the importance of residual networks?

One of the major importance of residual networks is that it allows access from the past or previous layers of data. This access allows the flow of information to be smooth throughout the network.

What is batch normalization and how does it work

The technique where each input layer gets modified as the previous layers tend to change is known as the batch modification. The batch normalization mainly works by making a standard deviation to be 1 and the output to be zero.

How to handle imbalance data sheet?

Datasheet could be handled with the few basic steps. Some of these include:

- Using class weights.
- Using the training examples again and gain.
- Avoid any under sample if the data is too large.
- Use data augmentation.

Why machine learning is using small kernels instead of large kernels?

The use of small kernels is due to the fact that, with smaller kernels proper receptive field can be known. Since smaller kernels use small computations and fewer parameters it is possible to get more mapping functions and even more filters.

Can there be any other projects which can be related?

In order to draw relations with some other projects, the user doesn't need to think a lot. The user just have to think over the facts which connect the research to business.

Explain the current master's research? What worked? What did not? Future directions?

Current master's research basically means which algorithms can be used to determine the value of coefficients and which model is best suitable for use. The use of machine learning algorithms worked a great deal but the single regression technique did not give the values correctly. Future directions would taking the time and doing research first before jumping to anyconclusion.

Machine Learning Interview Questions: Algorithms/Theory

These algorithms questions will test your grasp of the theory behind machine learning.

Q1- What's the trade-off between bias and variance?

More reading: Bias-Variance Tradeoff (Wikipedia)

Bias is error due to erroneous or overly simplistic assumptions in the learning algorithm you're using. This can lead to the model underfitting your data, making it hard for it to have high predictive accuracy and for you to generalize your knowledge from the training set to the test set.

Variance is error due to too much complexity in the learning algorithm you're using. This leads to the algorithm being highly sensitive to high degrees of variation in your training data, which can lead your model to overfit the data. You'll be carrying too much noise from your training data for your model to be very useful for your test data.

The bias-variance decomposition essentially decomposes the learning error from any algorithm by adding the bias, the variance and a bit of irreducible error due to noise in the underlying dataset. Essentially, if you make the model more complex and add more variables, you'll lose bias but gain some variance — in order to get the optimally reduced amount of error, you'll have to tradeoff bias and variance. You don't want either high bias or high variance in your model.

Q2- What is the difference between supervised and unsupervised machine learning?

More reading: What is the difference between supervised and unsupervised machine learning? (Quora)

Supervised learning requires training labeled data. For example, in order to do classification (a supervised learning task), you'll need to first label the data you'll use to train the model to classify data into your labeled groups. Unsupervised learning, in contrast, does not require labeling data explicitly.

Q3- How is KNN different from k-means clustering?

More reading: How is the k-nearest neighbor algorithm different from k-means clustering? (Quora)

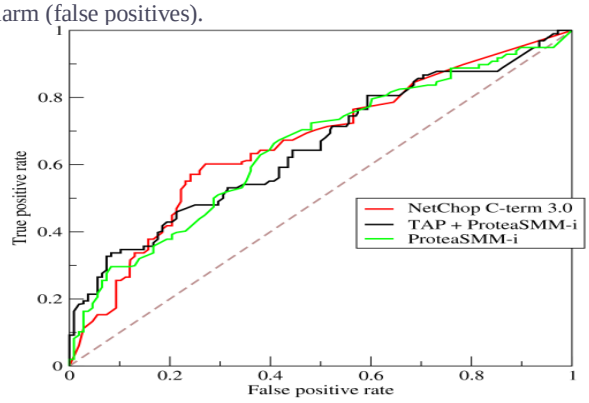
K-Nearest Neighbors is a supervised classification algorithm, while k-means clustering is an unsupervised clustering algorithm. While the mechanisms may seem similar at first, what this really means is that in order for K-Nearest Neighbors to work, you need labeled data you want to classify an unlabeled point into (thus the nearest neighbor part). K-means clustering requires only a set of unlabeled points and a threshold: the algorithm will take unlabeled points and gradually learn how to cluster them into groups by computing the mean of the distance between different points.

The critical difference here is that KNN needs labeled points and is thus supervised learning, while k-means doesn't — and is thus unsupervised learning.

Q4- Explain how a ROC curve works.

More reading: Receiver operating characteristic (Wikipedia)

The ROC curve is a graphical representation of the contrast between true positive rates and the false positive rate at various thresholds. It's often used as a proxy for the trade-off between the sensitivity of the model (true positives) vs the fall-out or the probability it will trigger a false alarm (false positives).



Q5- Define precision and recall.

More reading: Precision and recall (Wikipedia)

Recall is also known as the true positive rate: the amount of positives your model claims compared to the actual number of positives there are throughout the data. Precision is also known as the positive predictive value, and it is a measure of the amount of accurate positives your model claims compared to the number of positives it actually claims. It can be easier to think of recall and precision in the context of a case where you've predicted that there were 10 apples and 5 oranges in a case of 10 apples. You'd have perfect recall (there are actually 10 apples, and you predicted there would be 10) but 66.7% precision because out of the 15 events you predicted, only 10 (the apples) are correct.

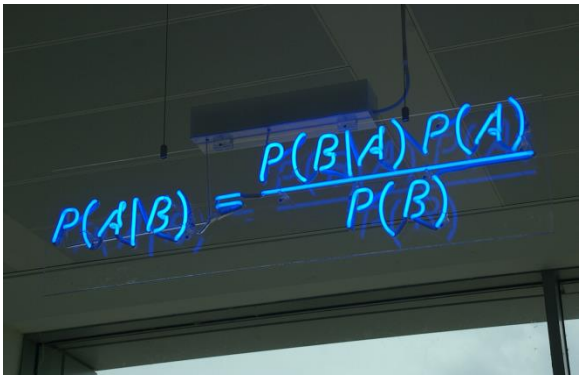
Q6- What is Bayes' Theorem? How is it useful in a machine learning context?

More reading: An Intuitive (and Short) Explanation of Bayes' Theorem (BetterExplained)

Bayes' Theorem gives you the posterior probability of an event given what is known as prior knowledge.

Mathematically, it's expressed as the true positive rate of a condition sample divided by the sum of the false positive rate of the population and the true positive rate of a condition. Say you had a 60% chance of actually having the flu after a flu test, but out of people who had the flu, the test will be false 50% of the time, and the overall population only has a 5% chance of having the flu. Would you actually have a 60% chance of having the flu after having a positive test?

Bayes' Theorem says no. It says that you have a (.6 * 0.05) (True Positive Rate of a Condition Sample) / (.6*0.05)(True Positive Rate of a Condition Sample) + (.5*0.95) (False Positive Rate of a Population) = 0.0594 or 5.94% chance of getting a flu.



Bayes' Theorem is the basis behind a branch of machine learning that most notably includes the Naive Bayes classifier. That's something important to consider when you're faced with machine learning interview questions.

Q7- Why is "Naive" Bayes naive?

More reading: Why is "naive Bayes" naive? (Quora)

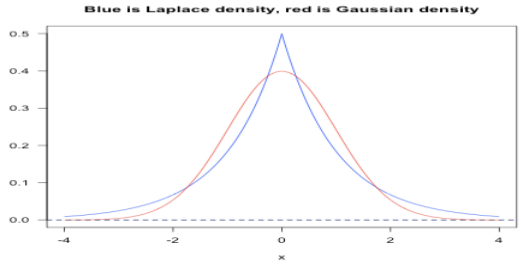
Despite its practical applications, especially in text mining, Naive Bayes is considered "Naive" because it makes an assumption that is virtually impossible to see in real-life data: the conditional probability is calculated as the pure product of the individual probabilities of components. This implies the absolute independence of features — a condition probably never met in real life.

As a Quora commenter put it whimsically, a Naive Bayes classifier that figured out that you liked pickles and ice cream would probably naively recommend you a pickle ice cream.

Q8- Explain the difference between L1 and L2 regularization.

More reading: What is the difference between L1 and L2 regularization? (Quora)

L2 regularization tends to spread error among all the terms, while L1 is more binary/sparse, with many variables either being assigned a 1 or 0 in weighting. L1 corresponds to setting a Laplacean prior on the terms, while L2 corresponds to a Gaussian prior.



Q9- What's your favorite algorithm, and can you explain it to me in less than a minute?

This type of question tests your understanding of how to communicate complex and technical nuances with poise and the ability to summarize quickly and efficiently. Make sure you have a choice and make sure you can explain different algorithms so simply and effectively that a five-year-old could grasp the basics!

Q10- What's the difference between Type I and Type II error?

More reading: Type I and type II errors (Wikipedia)

Don't think that this is a trick question! Many machine learning interview questions will be an attempt to lob basic questions at you just to make sure you're on top of your game and you've prepared all of your bases.

Type I error is a false positive, while Type II error is a false negative. Briefly stated, Type I error means claiming something has happened when it hasn't, while Type II error means that you claim nothing is happening when in fact something is.

A clever way to think about this is to think of Type I error as telling a man he is pregnant, while Type II error means you tell a pregnant woman she isn't carrying a baby.

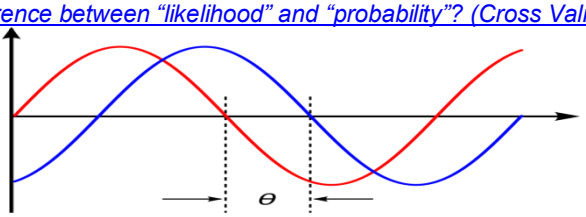
Q11- What's a Fourier transform?

More reading: Fourier transform (Wikipedia)

A Fourier transform is a generic method to decompose generic functions into a superposition of symmetric functions. Or as this more intuitive tutorial puts it, given a smoothie, it's how we find the recipe. The Fourier transform finds the set of cycle speeds, amplitudes and phases to match any time signal. A Fourier transform converts a signal from time to frequency domain — it's a very common way to extract features from audio signals or other time series such as sensor data.

Q12- What's the difference between probability and likelihood?

More reading: What is the difference between "likelihood" and "probability"? (Cross Validated)



Q13- What is deep learning, and how does it contrast with other machine learning algorithms?

More reading: Deep learning (Wikipedia)

Deep learning is a subset of machine learning that is concerned with neural networks: how to use backpropagation and certain principles from neuroscience to more accurately model large sets of unlabelled or semi-structured data. In that sense, deep learning represents an unsupervised learning algorithm that learns representations of data through the use of neural nets.

Q14- What's the difference between a generative and discriminative model?

More reading: What is the difference between a Generative and Discriminative Algorithm? (Stack Overflow)

A generative model will learn categories of data while a discriminative model will simply learn the distinction between different categories of data. Discriminative models will generally outperform generative models on classification tasks.

Q15- What cross-validation technique would you use on a time series dataset?

More reading: Using k-fold cross-validation for time-series model selection (CrossValidated)

Instead of using standard k-folds cross-validation, you have to pay attention to the fact that a time series is not randomly distributed data — it is inherently ordered by chronological order. If a pattern emerges in later time periods for example, your model may still pick up on it even if that effect doesn't hold in earlier years!

You'll want to do something like forward chaining where you'll be able to model on past data then look at forward-facing data.

- fold 1 : training [1], test [2]
- fold 2 : training [1 2], test [3]
- fold 3 : training [1 2 3], test [4]
- fold 4 : training [1 2 3 4], test [5]
- fold 5 : training [1 2 3 4 5], test [6]

Q16- How is a decision tree pruned?

More reading: Pruning (decision trees)

Pruning is what happens in decision trees when branches that have weak predictive power are removed in order to reduce the complexity of the model and increase the predictive accuracy of a decision tree model. Pruning can happen bottom-up and top-down, with approaches such as reduced error pruning and cost complexity pruning.

Reduced error pruning is perhaps the simplest version: replace each node. If it doesn't decrease predictive accuracy, keep it pruned. While simple, this heuristic actually comes pretty close to an approach that would optimize for maximum accuracy.

Q17- Which is more important to you– model accuracy, or model performance?

More reading: Accuracy paradox (Wikipedia)

This question tests your grasp of the nuances of machine learning model performance! Machine learning interview questions often look towards the details. There are models with higher accuracy that can perform worse in predictive power — how does that make sense?
Well, it has everything to do with how model accuracy is only a subset of model performance, and at that, a sometimes misleading one. For example, if you wanted to detect fraud in a massive dataset with a sample of millions, a more accurate model would most likely predict no fraud at all if only a vast minority of cases were fraud. However, this would be useless for a predictive model — a model designed to find fraud that asserted there was no fraud at all! Questions like this help you demonstrate that you understand model accuracy isn't the be-all and end-all of model performance.

Q18- What's the F1 score? How would you use it?

More reading: F1 score (Wikipedia)

The F1 score is a measure of a model's performance. It is a weighted average of the precision and recall of a model, with results tending to 1 being the best, and those tending to 0 being the worst. You would use it in classification tests where true negatives don't matter much.

Q19- How would you handle an imbalanced dataset?

More reading: 8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset (Machine Learning Mastery)

An imbalanced dataset is when you have, for example, a classification test and 90% of the data is in one class. That leads to problems: an accuracy of 90% can be skewed if you have no predictive power on the other category of data! Here are a few tactics to get over the hump:

1- Collect more data to even the imbalances in the dataset.
2- Resample the dataset to correct for imbalances.
3- Try a different algorithm altogether on your dataset.

What's important here is that you have a keen sense for what damage an unbalanced dataset can cause, and how to balance that.

Q20- When should you use classification over regression?

More reading: Regression vs Classification (Math StackExchange)

Classification produces discrete values and dataset to strict categories, while regression gives you continuous results that allow you to better distinguish differences between individual points. You would use classification over regression if you wanted your results to reflect the belongingness of data points in your dataset to certain explicit categories (ex: If you wanted to know whether a name was male or female rather than just how correlated they were with male and female names.)

Q21- Name an example where ensemble techniques might be useful.

More reading: Ensemble learning (Wikipedia)

Ensemble techniques use a combination of learning algorithms to optimize better predictive performance. They typically reduce overfitting in models and make the model more robust (unlikely to be influenced by small changes in the training data).

You could list some examples of ensemble methods, from bagging to boosting to a "bucket of models" method and demonstrate how they could increase predictive power.

Q22- How do you ensure you're not overfitting with a model?

More reading: How can I avoid overfitting? (Quora)

This is a simple restatement of a fundamental problem in machine learning: the possibility of overfitting training data and carrying the noise of that data through to the test set, thereby providing inaccurate generalizations.

There are three main methods to avoid overfitting:

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

Q23- What evaluation approaches would you work to gauge the effectiveness of a machine learning model?

More reading: How to Evaluate Machine Learning Algorithms (Machine Learning Mastery)

You would first split the dataset into training and test sets, or perhaps use cross-validation techniques to further segment the dataset into composite sets of training and test sets within the data. You should then implement a choice selection of performance metrics: here is a fairly comprehensive list. You could use measures such as the F1 score, the accuracy, and the confusion matrix. What's important here is to demonstrate that you understand the nuances of how a model is measured and how to choose the right performance measures for the right situations.

Q24- How would you evaluate a logistic regression model?

More reading: Evaluating a logistic regression (CrossValidated)

A subsection of the question above. You have to demonstrate an understanding of what the typical goals of a logistic regression are (classification, prediction etc.) and bring up a few examples and use cases.

Q25- What's the "kernel trick" and how is it useful?

More reading: Kernel method (Wikipedia)

The Kernel trick involves kernel functions that can enable in higher-dimension spaces without explicitly calculating the coordinates of points within that dimension: instead, kernel functions compute the inner products between the images of all pairs of data in a feature space. This allows them the very useful attribute of calculating the coordinates of higher dimensions while being computationally cheaper than the explicit calculation of said coordinates. Many algorithms can be expressed in terms of inner products. Using the kernel trick enables us effectively run algorithms in a high-dimensional space with lower-dimensional data.

Machine Learning Interview Questions: Programming

These machine learning interview questions test your knowledge of programming principles you need to implement machine learning principles in practice. Machine learning interview questions tend to be technical questions that test your logic and programming skills: this section focuses more on the latter.

Q26- How do you handle missing or corrupted data in a dataset?

More reading: Handling missing data (O'Reilly)

You could find missing/corrupted data in a dataset and either drop those rows or columns, or decide to replace them with another value.

In Pandas, there are two very useful methods: isnull() and dropna() that will help you find columns of data with missing or corrupted data and drop those values. If you want to fill the invalid values with a placeholder value (for example, 0), you could use the fillna() method.

Q27- Do you have experience with Spark or big data tools for machine learning?

More reading: 50 Top Open Source Tools for Big Data (Datamation)

You'll want to get familiar with the meaning of big data for different companies and the different tools they'll want. Spark is the big data tool most in demand now, able to handle immense datasets with speed. Be honest if you don't have experience with the tools demanded, but also take a look at job descriptions and see what tools pop up: you'll want to invest in familiarizing yourself with them.

Q28- Pick an algorithm. Write the psuedo-code for a parallel implementation.

More reading: Writing pseudocode for parallel programming (Stack Overflow)

This kind of question demonstrates your ability to think in parallelism and how you could handle concurrency in programming implementations dealing with big data. Take a look at pseudocode frameworks such as Peril-L and visualization tools such as Web Sequence Diagrams to help you demonstrate your ability to write code that reflects parallelism.

Q29- What are some differences between a linked list and an array?

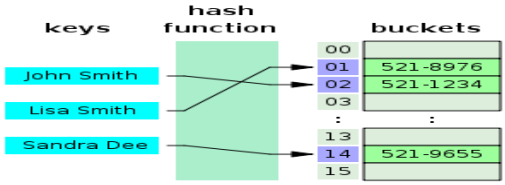
More reading: Array versus linked list (Stack Overflow)

An array is an ordered collection of objects. A linked list is a series of objects with pointers that direct how to process them sequentially. An array assumes that every element has the same size, unlike the linked list. A linked list can more easily grow organically: an array has to be pre-defined or re-defined for organic growth. Shuffling a linked list involves changing which points direct where — meanwhile, shuffling an array is more complex and takes more memory.

Q30- Describe a hash table.

More reading: Hash table (Wikipedia)

A hash table is a data structure that produces an associative array. A key is mapped to certain values through the use of a hash function. They are often used for tasks such as database indexing.



Q31- Which data visualization libraries do you use? What are your thoughts on the best data visualization tools?

More reading: 31 Free Data Visualization Tools (Springboard)

What's important here is to define your views on how to properly visualize data and your personal preferences when it comes to tools. Popular tools include R's ggplot, Python's seaborn and matplotlib, and tools such as Plot.ly and Tableau.

***Related**: 20 Python Interview Questions*

Q32- How would you implement a recommendation system for our company's users?

More reading: How to Implement A Recommendation System? (Stack Overflow)

A lot of machine learning interview questions of this type will involve implementation of machine learning models to a company's problems. You'll have to research the company and its industry in-depth, especially the revenue drivers the company has, and the types of users the company takes on in the context of the industry it's in.

Q33- How can we use your machine learning skills to generate revenue?

More reading: Startup Metrics for Startups (500 Startups)

This is a tricky question. The ideal answer would demonstrate knowledge of what drives the business and how your skills could relate. For example, if you were interviewing for music-streaming startup Spotify, you could remark that your skills at developing a better recommendation model would increase user retention, which would then increase revenue in the long run.
The startup metrics Slideshare linked above will help you understand exactly what performance indicators are important for startups and tech companies as they think about revenue and growth.

Q34- What do you think of our current data process?

More reading: The Data Science Process Email Course – Springboard

This kind of question requires you to listen carefully and impart feedback in a manner that is constructive and insightful. Your interviewer is trying to gauge if you'd be a valuable member of their team and whether you grasp the nuances of why certain things are set the way they are in the company's data process based on company- or industry-specific conditions. They're trying to see if you can be an intellectual peer. Act accordingly.

Q35- What are the last machine learning papers you've read?

More reading: What are some of the best research papers/books for machine learning?

Keeping up with the latest scientific literature on machine learning is a must if you want to demonstrate interest in a machine learning position. This overview of deep learning in Nature by the scions of deep learning themselves (from Hinton to Bengio to LeCun) can be a good reference paper and an overview of what's happening in deep learning — and the kind of paper you might want to cite.

Q36- Do you have research experience in machine learning?

Related to the last point, most organizations hiring for machine learning positions will look for your formal experience in the field. Research papers, co-authored or supervised by leaders in the field, can make the difference between you being hired and not. Make sure you have a summary of your research experience and papers ready — and an explanation for your background and lack of formal research experience if you don't.

Q37- What are your favorite use cases of machine learning models?

More reading: What are the typical use cases for different machine learning algorithms? (Quora)

The Quora thread above contains some examples, such as decision trees that categorize people into different tiers of intelligence based on IQ scores. Make sure that you have a few examples in mind and describe what resonated with you. It's important that you demonstrate an interest in how machine learning is implemented.

Q38- How would you approach the "Netflix Prize" competition?

More reading: Netflix Prize (Wikipedia)

The Netflix Prize was a famed competition where Netflix offered $1,000,000 for a better collaborative filtering algorithm. The team that won called BellKor had a 10% improvement and used an ensemble of different methods to win. Some familiarity with the case and its solution will help demonstrate you've paid attention to machine learning for a while.

Q39- Where do you usually source datasets?

More reading: 19 Free Public Data Sets For Your First Data Science Project (Springboard)

Machine learning interview questions like these try to get at the heart of your machine learning interest. Somebody who is truly passionate about machine learning will have gone off and done side projects on their own, and have a good idea of what great datasets are out there. If you're missing any, check out Quandl for economic and financial data, and Kaggle's Datasets collection for another great list.

Q40- How do you think Google is training data for self-driving cars?

More reading: Waymo Tech

Machine learning interview questions like this one really test your knowledge of different machine learning methods, and your inventiveness if you don't know the answer. Google is currently using recaptcha to source labeled data on storefronts and traffic signs. They are also building on training data collected by Sebastian Thrun at GoogleX — some of which was obtained by his grad students driving buggies on desert dunes!

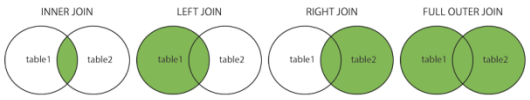
Q41- How would you simulate the approach AlphaGo took to beat Lee Sidol at Go?

More reading: Mastering the game of Go with deep neural networks and tree search (Nature)

AlphaGo beating Lee Sidol, the best human player at Go, in a best-of-five series was a truly seminal event in the history of machine learning and deep learning. The Nature paper above describes how this was accomplished with "Monte-Carlo tree search with deep neural networks that have been trained by supervised learning, from human expert games, and by reinforcement learning from games of self-play."

Difference between joins

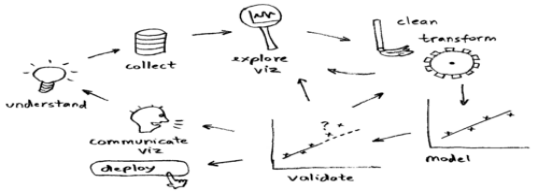
- **(INNER) JOIN**: Returns records that have matching values in both tables
- **LEFT (OUTER) JOIN**: Return all records from the left table, and the matched records from the right table
- **RIGHT (OUTER) JOIN**: Return all records from the right table, and the matched records from the left table
- **FULL (OUTER) JOIN**: Return all records when there is a match in either left or right table



Project Workflow

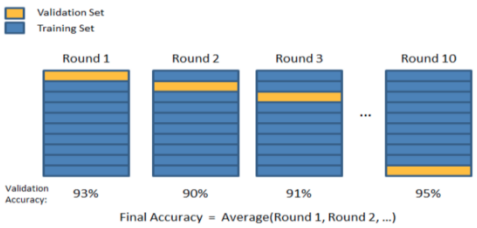
Given a data science / machine learning project, what steps should we follow? Here's how I would tackle it:

- **Specify business objective.** Are we trying to win more customers, achieve higher satisfaction, or gain more revenues?
- **Define problem.** What is the specific gap in your ideal world and the real one that requires machine learning to fill? Ask questions that can be addressed using your data and predictive modeling (ML algorithms).
- **Create a common sense baseline.** But before you resort to ML, set up a baseline to solve the problem as if you know zero data science. You may be amazed at how effective this baseline is. It can be as simple as recommending the top N popular items or other rule-based logic. This baseline can also server as a good benchmark for ML algorithms.
- **Review ML literatures.** To avoid reinventing the wheel and get inspired on what techniques / algorithms are good at addressing the questions using our data.
- **Set up a single-number metric.** What it means to be successful - high accuracy, lower error, or bigger AUC - and how do you measure it? The metric has to align with high-level goals, most often the success of your business. Set up a single-number against which all models are measured.
- **Do exploratory data analysis (EDA).** Play with the data to get a general idea of data type, distribution, variable correlation, facets etc. This step would involve a lot of plotting.
- **Partition data.** Validation set should be large enough to detect differences between the models you are training; test set should be large enough to indicate the overall performance of the final model; training set, needless to say, the larger the merrier.
- **Preprocess.** This would include data integration, cleaning, transformation, reduction, discretization and more.
- **Engineer features.** Coming up with features is difficult, time-consuming, requires expert knowledge. Applied machine learning is basically feature engineering. This step usually involves feature selection and creation, using domain knowledge. Can be minimal for deep learning projects.
- **Develop models.** Choose which algorithm to use, what hyperparameters to tune, which architecture to use etc.
- **Ensemble.** Ensemble can usually boost performance, depending on the correlations of the models/features. So it's always a good idea to try out. But be open-minded about making tradeoff - some ensemble are too complex/slow to put into production.
- **Deploy model.** Deploy models into production for inference.
- **Monitor model.** Monitor model performance, and collect feedbacks.
- **Iterate.** Iterate the previous steps. Data science tends to be an iterative process, with new and improved models being developed over time.



Cross Validation

Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a validation set to evaluate it. For example, a k-fold cross validation divides the data into k folds (or partitions), trains on each k-1 fold, and evaluate on the remaining 1 fold. This results to k models/evaluations, which can be averaged to get a overall model performance.



Feature Importance

- In linear models, feature importance can be calculated by the scale of the coefficients
- In tree-based methods (such as random forest), important features are likely to appear closer to the root of the tree. We can get a feature's importance for random forest by computing the averaging depth at which it appears across all trees in the forest.

Mean Squared Error vs. Mean Absolute Error

- **Similarity**: both measure the average model prediction error; range from 0 to infinity; the lower the better
- Mean Squared Error (MSE) gives higher weights to large error (e.g., being off by 10 just MORE THAN TWICE as bad as being off by 5), whereas Mean Absolute Error (MAE) assign equal weights (being off by 10 is just twice as bad as being off by 5)
- MSE is continuously differentiable, MAE is not (where y_pred == y_true)

L1 vs L2 regularization

- **Similarity**: both L1 and L2 regularization **prevent overfitting** by shrinking (imposing a penalty) on the coefficients
- **Difference**: L2 (Ridge) shrinks all the coefficient by the same proportions but eliminates none, while L1 (Lasso) can shrink some coefficients to zero, performing variable selection.
- **Which to choose**: If all the features are correlated with the label, ridge outperforms lasso, as the coefficients are never zero in ridge. If only a subset of features are correlated with the label, lasso outperforms ridge as in lasso model some coefficient can be shrunken to zero.
- In Graph (a), the black square represents the feasible region of the L1 regularization while graph (b) represents the feasible region for L2 regularization. The contours in the plots represent different loss values (for the unconstrained regression model). The feasible point that minimizes the loss is more likely to happen on the coordinates on graph (a) than on graph (b) since graph (a) is more **angular**. This effect amplifies when your number of coefficients increases, i.e. from 2 to 200. The implication of this is that the L1 regularization gives you sparse estimates. Namely, in a high dimensional space, you got mostly zeros and a small number of non-zero coefficients.

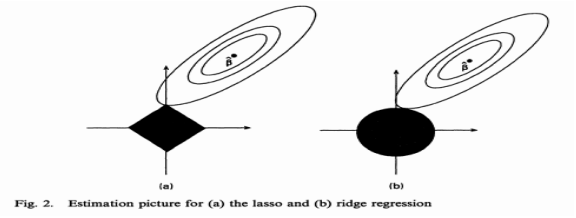


Fig. 2. Estimation picture for (a) the lasso and (b) ridge regression

Correlation vs Covariance

- Both determine the relationship and measure the dependency between two random variables
- Correlation is when the change in one item may result in the change in the another item, while covariance is when two items vary together (joint variability)
- Covariance is nothing but a measure of correlation. On the contrary, correlation refers to the scaled form of covariance
- Range: correlation is between -1 and +1, while covariance lies between negative infinity and infinity.

Would adding more data address underfitting

Underfitting happens when a model is not complex enough to learn well from the data. It is the problem of model rather than data size. So a potential way to address underfitting is to increase the model complexity (e.g., to add higher order coefficients for linear model, increase depth for tree-based methods, add more layers / number of neurons for neural networks etc.

Activation Function

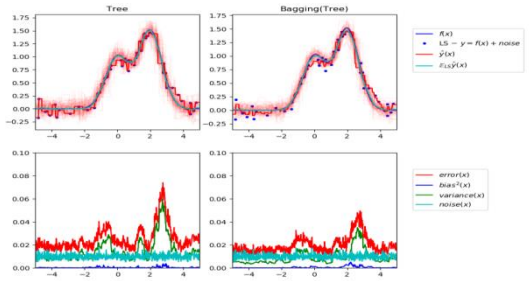
For neural networks

- Non-linearity: ReLU is often used. Use Leaky ReLU (a small positive gradient for negative input, say, `y = 0.01x` when x < 0) to address dead ReLU issue

- Multi-class: softmax
- Binary: sigmoid
- Regression: linear

Bagging

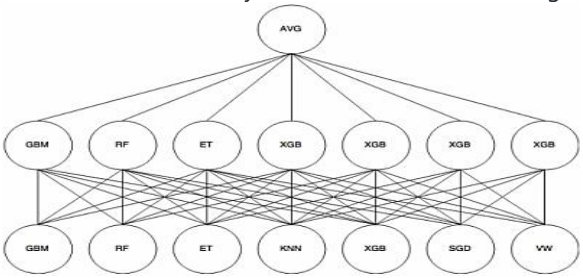
To address overfitting, we can use an ensemble method called bagging (bootstrap aggregating), which reduces the variance of the meta learning algorithm. Bagging can be applied to decision tree or other algorithms.
Here is a great illustration of a single estimator vs. bagging.



- Bagging is when samlping is performed *with* replacement. When sampling is performed *without* replacement, it's called pasting.
- Bagging is popular due to its boost for performance, but also due to that individual learners can be trained in parallel and scale well
- Ensemble methods work best when the learners are as independent from one another as possible
- Voting: soft voting (predict probability and average over all individual learners) often works better than hard voting
- out-of-bag instances can act validation set for bagging

Stacking

- Instead of using trivial functions (such as hard voting) to aggregate the predictions from individual learners, train a model to perform this aggregation
- First split the training set into two subsets: the first subset is used to train the learners in the first layer
- Next the first layer learners are used to make predictions (meta features) on the second subset, and those predictions are used to train another models (to obtain the weigts of different learners) in the second layer
- We can train multiple models in the second layer, but this entails subsetting the original dataset into 3 parts



Generative vs discriminative

- Discriminative algorithms model *p(y|x; w)*, that is, given the dataset and learned parameter, what is the probability of y belonging to a specific class. A discriminative algorithm doesn't care about how the data was generated, it simply categorizes a given example
- Generative algorithms try to model *p(x|y)*, that is, the distribution of features given that it belongs to a certain class. A generative algorithm models how the data was generated.

Given a training set, an algorithm like logistic regression or the perceptron algorithm (basically) tries to find a straight line—that is, a decision boundary—that separates the elephants and dogs. Then, to classify a new animal as either an elephant or a dog, it checks on which side of the decision boundary it falls, and makes its prediction accordingly.
Here's a different approach. First, looking at elephants, we can build a model of what elephants look like. Then, looking at dogs, we can build a separate model of what dogs look like. Finally, to classify a new animal, we can match

the new animal against the elephant model, and match it against the dog model, to see whether the new animal looks more like the elephants or more like the dogs we had seen in the training set.

Parametric vs Nonparametric

- A learning model that summarizes data with a set of parameters of fixed size (independent of the number of training examples) is called a parametric model.
- A model where the number of parameters is not determined prior to training. Nonparametric does not mean that they have no parameters. On the contrary, nonparametric models (can) become more and more complex with an increasing amount of data.

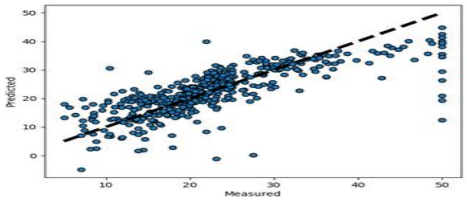
back to top

Recommender System

- I put recommend system here since technically it falls neither under supervised nor unsupervised learning
- A recommender system seeks to predict the 'rating' or 'preference' a user would give to items and then recommend items accordingly
- Content based recommender systems recommends items similar to those a given user has liked in the past, based on either explicit (ratings, like/dislike button) or implicit (viewed/finished an article) feedbacks. Content based recommenders work solely with the past interactions of a given user and do not take other users into consideration.
- Collaborative filtering is based on past interactions of the whole user base. There are two Collaborative filtering approaches: **item-based** or **user-based**
    - item-based: for user u, a score for an unrated item is produced by combining the ratings of users similar to u.
    - user-based: a rating (u, i) is produced by looking at the set of items similar to i (interaction similarity), then the ratings by u of similar items are combined into a predicted rating
- In recommender systems traditionally matrix factorization methods are used, although we recently there are also deep learning based methods
- Cold start and sparse matrix can be issues for recommender systems
- Widely used in movies, news, research articles, products, social tags, music, etc.

Linear regression

- How to learn the parameter: minimize the cost function
- How to minimize cost function: gradient descent
- Regularization:
    - L1 (Lasso): can shrink certain coef to zero, thus performing feature selection
    - L2 (Ridge): shrink all coef with the same proportion; almost always outperforms L1
    - Elastic Net: combined L1 and L2 priors as regularizer
- Assumes linear relationship between features and the label
- Can add polynomial and interaction features to add non-linearity



Logistic regression

- Generalized linear model (GLM) for binary classification problems
- Apply the sigmoid function to the output of linear models, squeezing the target to range [0, 1]
- Threshold to make prediction: usually if the output > .5, prediction 1; otherwise prediction 0
- A special case of softmax function, which deals with multi-class problems

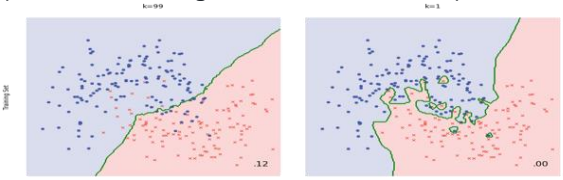
Naive Bayes

- Naive Bayes (NB) is a supervised learning algorithm based on applying Bayes' theorem

- It is called naive because it builds the naive assumption that each feature are independent of each other
- NB can make different assumptions (i.e., data distributions, such as Gaussian, Multinomial, Bernoulli)
- Despite the over-simplified assumptions, NB classifier works quite well in real-world applications, especially for text classification (e.g., spam filtering)
- NB can be extremely fast compared to more sophisticated methods

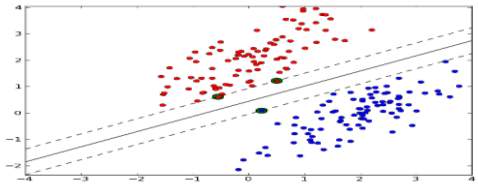
KNN

- Given a data point, we compute the K nearest data points (neighbors) using certain distance metric (e.g., Euclidean metric). For classification, we take the majority label of neighbors; for regression, we take the mean of the label values.
- Note for KNN we don't train a model; we simply compute during inference time. This can be computationally expensive since each of the test example need to be compared with every training example to see how close they are.
- There are approximation methods can have faster inference time by partitioning the training data into regions (e.g., annoy)
- When K equals 1 or other small number the model is prone to overfitting (high variance), while when K equals number of data points or other large number the model is prone to underfitting (high bias)



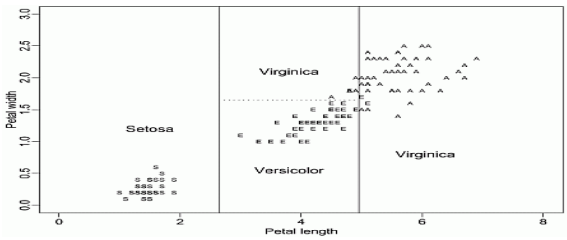
SVM

- Can perform linear, nonlinear, or outlier detection (unsupervised)
- Large margin classifier: using SVM we not only have a decision boundary, but want the boundary to be as far from the closest training point as possible
- The closest training examples are called support vectors, since they are the points based on which the decision boundary is drawn
- SVMs are sensitive to feature scaling



Decision tree

- Non-parametric, supervised learning algorithms
- Given the training data, a decision tree algorithm divides the feature space into regions. For inference, we first see which region does the test data point fall in, and take the mean label values (regression) or the majority label value (classification).
- **Construction**: top-down, chooses a variable to split the data such that the target variables within each region are as homogeneous as possible. Two common metrics: gini impurity or information gain, won't matter much in practice.
- Advantage: simply to understand & interpret, mirrors human decision making
- Disadvantage:
    - can overfit easily (and generalize poorly) if we don't limit the depth of the tree
    - can be non-robust: A small change in the training data can lead to a totally different tree
    - instability: sensitive to training set rotation due to its orthogonal decision boundaries



Random forest

Random forest improves bagging further by adding some randomness. In random forest, only a subset of features are selected at random to construct a tree (while often not subsample instances). The benefit is that random forest **decorrelates** the trees.

For example, suppose we have a dataset. There is one very predicative feature, and a couple of moderately predicative features. In bagging trees, most of the trees will use this very predicative feature in the top split, and therefore making most of the trees look similar, **and highly correlated**. Averaging many highly correlated results won't lead to a large reduction in variance compared with uncorrelated results. In random forest for each split we only consider a subset of the features and therefore reduce the variance even further by introducing more uncorrelated trees.

I wrote a notebook to illustrate this point.

In practice, tuning random forest entails having a large number of trees (the more the better, but always consider computation constraint). Also, `min_samples_leaf` (The minimum number of samples at the leaf node)to control the tree size and overfitting. Always cross validate the parameters.

Boosting Tree

How it works

Boosting builds on weak learners, and in an iterative fashion. In each iteration, a new learner is added, while all existing learners are kept unchanged. All learners are weighted based on their performance (e.g., accuracy), and after a weak learner is added, the data are re-weighted: examples that are misclassified gain more weights, while examples that are correctly classified lose weights. Thus, future weak learners focus more on examples that previous weak learners misclassified.

Difference from random forest (RF)

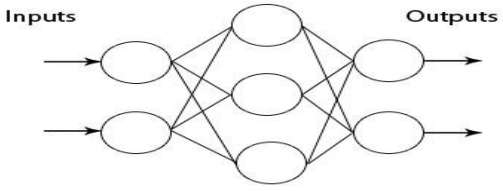
- RF grows trees **in parallel**, while Boosting is sequential
- RF reduces variance, while Boosting reduces errors by reducing bias

XGBoost (Extreme Gradient Boosting)

XGBoost uses a more regularized model formalization to control overfitting, which gives it better performance

MLP

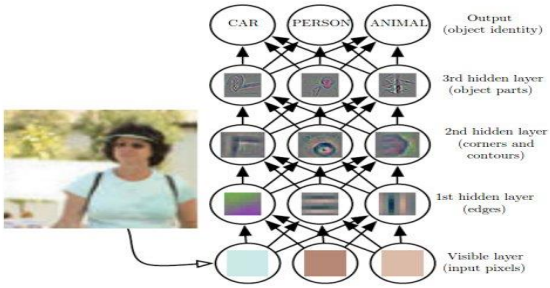
A feedforward neural network of multiple layers. In each layer we can have multiple neurons, and each of the neuron in the next layer is a linear/nonlinear combination of the all the neurons in the previous layer. In order to train the network we back propagate the errors layer by layer. In theory MLP can approximate any functions.



CNN

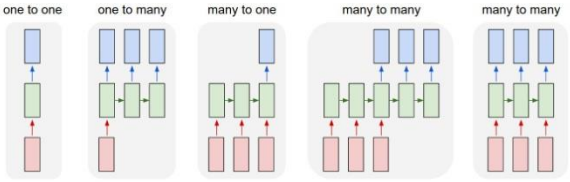
The Conv layer is the building block of a Convolutional Network. The Conv layer consists of a set of learnable filters (such as 5 * 5 * 3, width * height * depth). During the forward pass, we slide (or more precisely, convolve) the filter across the input and compute the dot product. Learning again happens when the network back propagate the error layer by layer.

Initial layers capture low-level features such as angle and edges, while later layers learn a combination of the low-level features and in the previous layers and can therefore represent higher level feature, such as shape and object parts.



RNN and LSTM

RNN is another paradigm of neural network where we have difference layers of cells, and each cell not only takes as input the cell from the previous layer, but also the previous cell within the same layer. This gives RNN the power to model sequence.

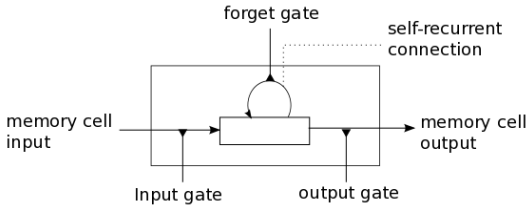


This seems great, but in practice RNN barely works due to exploding/vanishing gradient, which is cause by a series of multiplication of the same matrix. To solve this, we can use a variation of RNN, called long short-term memory (LSTM), which is capable of learning long-term dependencies.

The math behind LSTM can be pretty complicated, but intuitively LSTM introduce

- input gate
- output gate
- forget gate
- memory cell (internal state)

LSTM resembles human memory: it forgets old stuff (old internal state * forget gate) and learns from new input (input node * input gate)

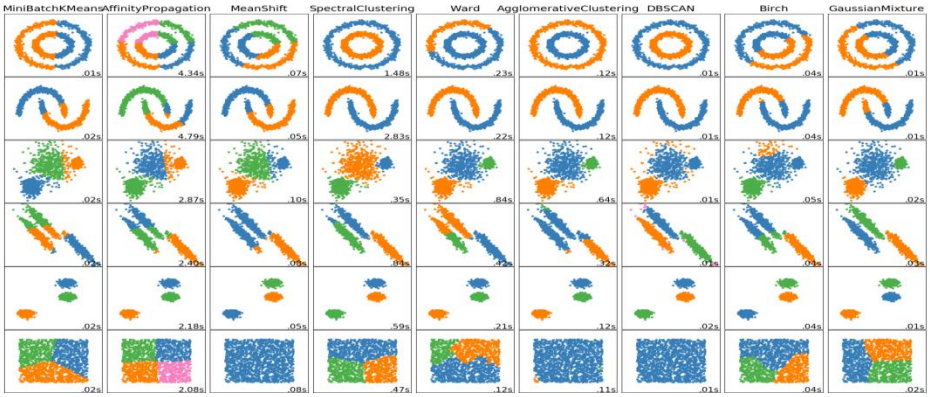


Unsupervised Learning

Clustering

- Clustering is a unsupervised learning algorithm that groups data in such a way that data points in the same group are more similar to each other than to those from other groups
- Similarity is usually defined using a distance measure (e.g, Euclidean, Cosine, Jaccard, etc.)
- The goal is usually to discover the underlying structure within the data (usually high dimensional)
- The most common clustering algorithm is K-means, where we define K (the number of clusters) and the algorithm iteratively finds the cluster each data point belongs to

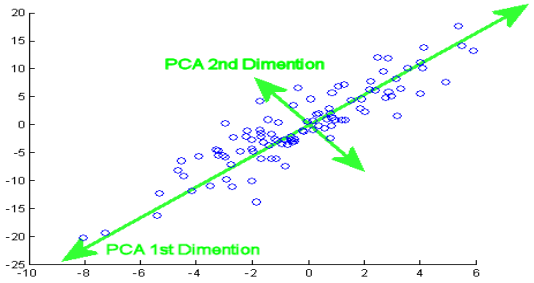
scikit-learn implements many clustering algorithms. Below is a comparison adopted from its page.



Principal Component Analysis

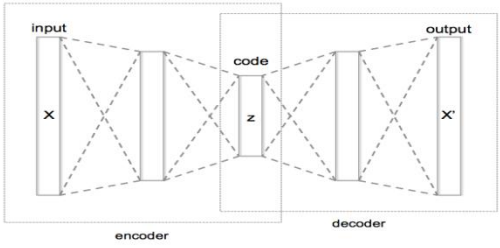
- Principal Component Analysis (PCA) is a dimension reduction technique that projects the data into a lower dimensional space
- PCA uses Singular Value Decomposition (SVD), which is a matrix factorization method that decomposes a matrix into three smaller matrices (more details of SVD here)
- PCA finds top N principal components, which are dimensions along which the data vary (spread out) the most. Intuitively, the more spread out the data along a specific dimension, the more information is contained, thus the more important this dimension is for the pattern recognition of the dataset
- PCA can be used as pre-step for data visualization: reducing high dimensional data into 2D or 3D. An alternative dimensionality reduction technique is t-SNE

Here is a visual explanation of PCA



Autoencoder

- The aim of an autoencoder is to learn a representation (encoding) for a set of data
- An autoencoder always consists of two parts, the encoder and the decoder. The encoder would find a lower dimension representation (latent variable) of the original input, while the decoder is used to reconstruct from the lower-dimension vector such that the distance between the original and reconstruction is minimized
- Can be used for data denoising and dimensionality reduction

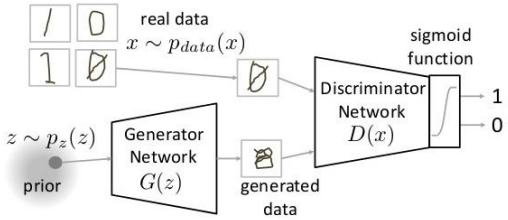


Generative Adversarial Network

- Generative Adversarial Network (GAN) is an unsupervised learning algorithm that also has supervised flavor: using supervised loss as part of training
- GAN typically has two major components: the **generator** and the **discriminator**. The generator tries to generate "fake" data (e.g, images or sentences) that fool the discriminator into thinking that they're real, while the discriminator tries to distinguish between real and generated data. It's a fight between the two

players thus the name adversarial, and this fight drives both sides to improve until "fake" data are indistinguishable from the real data

- How does it work, intuitively
  - The generator takes a **random** input and generates a sample of data
  - The discriminator then either takes the generated sample or a real data sample, and tries to predict whether the input is real or generated (i.e., solving a binary classification problem)
  - Given a truth score range of [0, 1], ideally the we'd love to see discriminator give low score to generated data but high score to real data. On the other hand, we also wanna see the generated data fool the discriminator. And this paradox drives both sides become stronger
- How does it work, from a training perspective
  - Without training, the generator creates 'garbage' data only while the discriminator is too 'innocent' to tell the difference between fake and real data
  - Usually we would first train the discriminator with both real (label 1) and generated data (label 0) for N epochs so it would have a good judgement of what is real vs. fake
  - Then we **set the discriminator non-trainable**, and train the generator. Even though the discriminator is non-trainable at this stage, we still use it as a classifier so that **error signals can be back propagated and therefore enable the generator to learn**
  - The above two steps would continue in turn until both sides cannot be improved further
- Here are some tips and tricks to make GANs work
- One Caveat is that the **adversarial part is only auxiliary: The end goal of using GAN is to generate data that even experts cannot tell if it's real or fake**



Reinforcement Learning
Natural Language Processing

Tokenization

- Tokenization is the process of converting a sequence of characters into a sequence of tokens
- Consider this example: `The quick brown fox jumped over the lazy dog`. In this case each word (separated by space) would be a token
- Sometimes tokenization doesn't have a definitive answer. For instance, `O'Neill` can be tokenized to `o` and `neill`, `oneill`, or `o'neill`.
- In some cases tokenization requires language-specific knowledge. For example, it doesn't make sense to tokenize `aren't` into `aren` and `t`
- For a more detailed treatment of tokenization please check here

Stemming and lemmatization

- The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form
- Stemming usually refers to a crude heuristic process that chops off the ends of words
- Lemmatization usually refers to doing things properly with the use of a vocabulary and morphological analysis of words
- If confronted with the token `saw`, stemming might return just `s`, whereas lemmatization would attempt to return either `see` or `saw` depending on whether the use of the token was as a verb or a noun
- For a more detailed treatment please check here

N gram

- n-gram is a contiguous sequence of n items from a given sample of text or speech
- An n-gram of size 1 is referred to as a "unigram"; size 2 is a "bigram" size 3 is a "trigram". Larger sizes are sometimes referred to by the value of n in modern language, e.g., "four-gram", "five-gram", and so on.

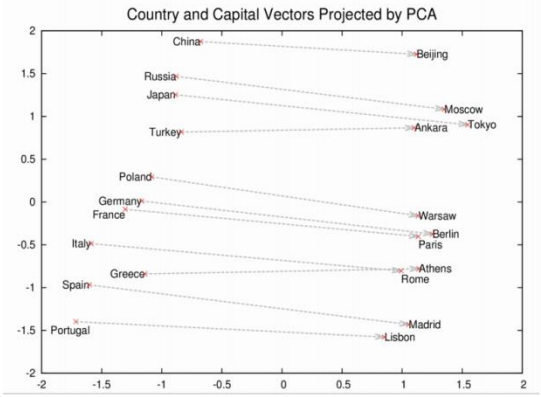
- Consider this example: `The quick brown fox jumped over the lazy dog.`
  - bigram would be `the quick`, `quick brown`, `brown fox`, ..., i.e, every two consecutive words (or tokens)
  - trigram would be `the quick brown`, `quick brown fox`, `brown fox jumped`, ..., i.e., every three consecutive words (or tokens)
- ngram model models sequence, i.e., predicts next word (n) given previous words (1, 2, 3, ..., n-1)
- multiple gram (bigram and above) captures **context**
- to choose n in n-gram requires experiments and making tradeoff between stability of the estimate against its appropriateness. Rule of thumb: trigram is a common choice with large training corpora (millions of words), whereas a bigram is often used with smaller ones.
- n-gram can be used as features for machine learning and downstream NLP tasks

Bag of Words

- Why? Machine learning models cannot work with raw text directly; rather, they take numerical values as input.
- Bag of words (BoW) builds a **vocabulary** of all the unique words in our dataset, and associate a unique index to each word in the vocabulary
- It is called a "bag" of words, because it is a representation that completely ignores the order of words
- Consider this example of two sentences: (1) `John likes to watch movies, especially horor movies.`, (2) `Mary likes movies too.` We would first build a vocabulary of unique words (all lower cases and ignoring punctuations): `[john, likes, to, watch, movies, especially, horor, mary, too]`. Then we can represent each sentence using term frequency, i.e, the number of times a term appears. So (1) would be `[1, 1, 1, 1, 2, 1, 1, 0, 0]`, and (2) would be `[0, 1, 0, 0, 1, 0, 0, 1, 1]`
- A common alternative to the use of dictionaries is the hashing trick, where words are directly mapped to indices with a hashing function
- As the vocabulary grows bigger (tens of thousand), the vector to represent short sentences / document becomes sparse (almost all zeros)

word2vec

- Shallow, two-layer neural networks that are trained to construct linguistic context of words
- Takes as input a large corpus, and produce a vector space, typically of several hundred dimension, and each word in the corpus is assigned a vector in the space
- The key idea is **context**: words that occur often in the same context should have same/opposite meanings.
- Two flavors
  - continuous bag of words (CBOW): the model predicts the current word given a window of surrounding context words
  - skip gram: predicts the surrounding context words using the current word



https://www.dezyre.com/article/top-machine-learning-interview-questions-and-answers-for-2018/357

1) What is the difference between inductive machine learning and deductive machine learning?

In inductive machine learning, the model learns by examples from a set of observed instances to draw a generalized conclusion whereas in deductive learning the model first draws the conclusion and then the conclusion is drawn. Let's understand this with an example, for instance, if you have to explain to a kid that playing with fire can cause burns. There are two ways you can explain this to kids, you can show them training examples of various fire accidents or images with burnt people and label them as

"Hazardous". In this case the kid will learn with the help of examples and not play with fire. This is referred to as Inductive machine learning. The other way is to let your kid play with fire and wait to see what happens. If the kid gets a burn they will learn not to play with fire and whenever they come across fire, they will avoid going near it. This is referred to as deductive learning.

2) How will you know which machine learning algorithm to choose for your classification problem?

If accuracy is a major concern for you when deciding on a machine learning algorithm then the best way to go about it is test a couple of different ones (by trying different parameters within each algorithm) and choose the best one by cross-validation. A general rule of thumb to choose a good enough machine learning algorithm for your classification problem is based on how large your training set is. If the training set is small then using low variance/high bias classifiers like Naïve Bayes is advantageous over high variance/low bias classifiers like k-nearest neighbour algorithms as it might overfit the model. High variance/low bias classifiers tend to win when the training set grows in size.

3) Why is Naïve Bayes machine learning algorithm naïve?

Naïve Bayes machine learning algorithm is considered Naïve because the assumptions the algorithm makes are virtually impossible to find in real-life data. Conditional probability is calculated as a pure product of individual probabilities of components. This means that the algorithm assumes the presence or absence of a specific feature of a class is not related to the presence or absence of any other feature (absolute independence of features), given the class variable. For instance, a fruit may be considered to be a banana if it is yellow, long and about 5 inches in length. However, if these features depend on each other or are based on the existence of other features, a naïve Bayes classifier will assume all these properties to contribute independently to the probability that this fruit is a banana. Assuming that all features in a given dataset are equally important and independent rarely exists in the real-world scenario.

4) How will you explain machine learning in to a layperson?

Machine learning is all about making decisions based on previous experience with a task with the intent of improving its performance. There are multiple examples that can be given to explain machine learning to a layperson –

- Imagine a curious kid who sticks his palm
- You have observed from your connections that obese people often tend to get heart diseases thus you make the decision that you will try to remain thin otherwise you might suffer from a heart disease. You have observed a ton of data and come up with a general rule of classification.
- You are playing blackjack and based on the sequence of cards you see, you decide whether to hit or to stay. In this case based on the previous information you have and by looking at what happens, you make a decision quickly.

5) List out some important methods of reducing dimensionality.

- Combine features with feature engineering.
- Use some form of algorithmic dimensionality reduction like ICA or PCA.
- Remove collinear features to reduce dimensionality.

6) You are given a dataset where the number of variables (p) is greater than the number of observations (n) (p>n). Which is the best technique to use and why?

When the number of variables is greater than the number of observations, it represents a high dimensional dataset. In such cases, it is not possible to calculate a unique least square coefficient estimate. Penalized regression methods like LARS, Lasso or Ridge seem work well under these circumstances as they tend to shrink the coefficients to reduce variance. Whenever the least square estimates have higher variance, Ridge regression technique seems to work best.

7) "People who bought this, also bought…." recommendations on Amazon are a result of which machine learning algorithm?

Recommender systems usually implement the collaborative filtering machine learning algorithm that considers user behaviour for recommending products to users. Collaborative filtering machine learning algorithms exploit the behaviour of users and products through ratings, reviews, transaction history, browsing history, selection and purchase information.

.8) Name some feature extraction techniques used for dimensionality reduction.

- Independent Component Analysis
- Principal Component Analysis
- Kernel Based Principal Component Analysis

9) List some use cases where classification machine learning algorithms can be used.

- Natural language processing (Best example for this is Spoken Language Understanding)
- Market Segmentation
- Text Categorization (Spam Filtering)
- Bioinformatics (Classifying proteins according to their function)
- Fraud Detection
- Face detection

10) What kind of problems does regularization solve?

Regularization is used to address overfitting problems as it penalizes the loss function by adding a multiple of an L1 (LASSO) or an L2 (Ridge) norm of your weights vector w.

11) How much data will you allocate for your training, validation and test sets?

There is no to the point answer to this question but there needs to be a balance/equilibrium when allocating data for training, validation and test sets.
If you make the training set too small, then the actual model parameters might have high variance. Also, if the test set is too small, there are chances of unreliable estimation of model performance. A general thumb rule to follow is to use 80: 20 train/test spilt. After this the training set can be further split into validation sets.

12) Which one would you prefer to choose – model accuracy or model performance?
Model accuracy is just a subset of model performance but is not the be-all and end-all of model performance. This question is asked to test your knowledge on how well you can make a perfect balance between model accuracy and model performance.
13) What is the most frequent metric to assess model accuracy for classification problems?
Percent Correct Classification (PCC) measures the overall accuracy irrespective of the kind of errors that are made, all errors are considered to have same weight.
14) When will you use classification over regression?
Classification is about identifying group membership while regression technique involves predicting a response. Both techniques are related to prediction, where classification predicts the belonging to a class whereas regression predicts the value from a continuous set. Classification technique is preferred over regression when the results of the model need to return the belongingness of data points in a dataset to specific explicit categories. (For instance, when you want to find out whether a name is male or female instead of just finding it how correlated they are with male and female names.
15) Why is Manhattan distance not used in kNN machine learning algorithm to calculate the distance between nearest neighbours?
Manhattan distance has restrictions on dimensions and calculates the distance either vertically or horizontally. Euclidean distance is better option in kNN to calculate the distance between nearest neighbours because the data points can be represented in any space without any dimension restriction.

Role Specific Open Ended Machine Learning Interview Questions
1) Given a particular machine learning model, what type of problems does it solve, what are the assumptions the model makes about the data and why it is best fit for a particular kind of problem?
2) Is the given machine learning model prove to over fitting? If so, what can you do about this?
3) What type of data does the machine learning model handle –categorical, numerical, etc.?
4) How interpretable is the given machine learning model?
5) What will you do if training results in very low accuracy?
6) Does the developed machine learning model have convergence problems?
7) Which is your favourite machine learning algorithm? Why it is your favourite and how will you explain about that machine learning algorithm to a layperson?
8) What approach will you follow to suggest followers on Twitter?
9) How will generate related searches on Bing?
10) Which tools and environments have you used to train and assess machine learning models?
11) If you claim in your resume that you know about recommender systems, then you might be asked to explain about PLSI and SVD models in detail.
12) How will you apply machine learning to images?

Machine Learning Interview Questions asked at Top Tech Companies

Machine Learning Interview Questions asked at Amazon
1) How will you weigh 9 marbles 3 times on a balance scale to find the heaviest one?
2) Why is gradient checking important?
3) What is loss function in a Neural Network?
4) Which one is better – random weight assignment or assigning same weights to the units in the hidden layer?
5) How will you design a spam filter?
6) Explain the difference between MLE and MAP inference.
7) Given a number, how will you find the closest number in a series of floating point data?
8) What is boosting?

Machine Learning Interview Questions asked at Baidu
1) What are the reasons for gradient descent to converge slow or not converge in various machine learning algorithms?
2) Given an objective function, calculate the range of its learning rate.
3) If the gradient descent does not converge, what could be the problem?
4) How will you check for a valid binary search tree?

Machine Learning Interview Questions asked at Spotify
1) Explain BFS (Breadth First Search algorithm)
2) How will you tell if a song in our catalogue is a duplicate or not?
3) What is your favourite machine learning algorithm and why?

4) Given the song list and metadata, disambiguate artists having same names.
5) How will you sample a stream of data to match the distribution with real data?

Machine Learning Interview Questions asked at Capital One
1) Given two years of transaction history, what features will you use to predict the credit risk?
2) Differentiate between gradient boosted tree and random forest machine learning algorithm.
3) How will you use existing features to add new features?
4) Considering that you have 100 data points and you have to predict the gender of a customer. What are the difficulties that could arise?
5) How will you set the threshold for credit card fraud detection model?
A machine learning interview is a compound process and the final result of the interview is determined by multiple factors and not just by looking at the number of right answers given by the candidate. If you really want that machine learning job, it's going to take time and dedication as you practice multiple ways to answer the above listed machine learning interview questions, but hopefully it is the enjoyable kind. You will learn a lot and get a good deal of knowledge preparing for your next machine learning interview with the help of these questions.

What is the difference between Data Mining and Data Analysis?

Data Mining	**Data Analysis**
Data mining usually does not require any hypothesis.	Data analysis begins with a question or an assumption.
Data Mining depends on clean and well-documented data.	Data analysis involves data cleaning.
Results of data mining are not always easy to interpret.	Data analysts interpret the results and convey the to the stakeholders.
Data mining algorithms automatically develop equations.	Data analysts have to develop their own equations based on the hypothesis.

2) Explain the typical data analysis process.
Data analysis deals with collecting, inspecting, cleansing, transforming and modelling data to glean valuable insights and support better decision making in an organization. The various steps involved in the data analysis process include –
Data Exploration –
Having identified the business problem, a data analyst has to go through the data provided by the client to analyse the root cause of the problem.
Data Preparation
This is the most crucial step of the data analysis process wherein any data anomalies (like missing values or detecting outliers) with the data have to be modelled in the right direction.
Data Modelling
The modelling step begins once the data has been prepared. Modelling is an iterative process wherein the model is run repeatedly for improvements. Data modelling ensures that the best possible result is found for a given business problem.
Validation
In this step, the model provided by the client and the model developed by the data analyst are validated against each other to find out if the developed model will meet the business requirements.
Implementation of the Model and Tracking
This is the final step of the data analysis process wherein the model is implemented in production and is tested for accuracy and efficiency.

3) What is the difference between Data Mining and Data Profiling?

Data Profiling, also referred to as Data Archeology is the process of assessing the data values in a given dataset for uniqueness, consistency and logic. Data profiling cannot identify any incorrect or inaccurate data but can detect only business rules violations or anomalies. The main purpose of data profiling is to find out if the existing data can be used for various other purposes.

Data Mining refers to the analysis of datasets to find relationships that have not been discovered earlier. It focusses on sequenced discoveries or identifying dependencies, bulk analysis, finding various types of attributes, etc.

4) How often should you retrain a data model?

A good data analyst is the one who understands how changing business dynamics will affect the efficiency of a predictive model. You must be a valuable consultant who can use analytical skills and business acumen to find the root cause of business problems.

The best way to answer this question would be to say that you would work with the client to define a time period in advance. However, I would refresh or retrain a model when the company enters a new market, consummate an acquisition or is facing emerging competition. As a data analyst, I would retrain the model as quick as possible to adjust with the changing behaviour of customers or change in market conditions.

5) What is data cleansing? Mention few best practices that you have followed while data cleansing.

From a given dataset for analysis, it is extremely important to sort the information required for data analysis. Data cleaning is a crucial step in the analysis process wherein data is inspected to find any anomalies, remove repetitive data, eliminate any incorrect information, etc. Data cleansing does not involve deleting any existing information from the database, it just enhances the quality of data so that it can be used for analysis.

Some of the best practices for data cleansing include –

- Developing a data quality plan to identify where maximum data quality errors occur so that you can assess the root cause and design the plan according to that.
- Follow a standard process of verifying the important data before it is entered into the database.
- Identify any duplicates and validate the accuracy of the data as this will save lot of time during analysis.
- Tracking all the cleaning operations performed on the data is very important so that you repeat or remove any operations as necessary.

6) How will you handle the QA process when developing a predictive model to forecast customer churn?

Data analysts require inputs from the business owners and a collaborative environment to operationalize analytics. To create and deploy predictive models in production there should be an effective, efficient and repeatable process. Without taking feedback from the business owner, the model will just be a one-and-done model.

The best way to answer this question would be to say that you would first partition the data into 3 different sets Training, Testing and Validation. You would then show the results of the validation set to the business owner by eliminating biases from the first 2 sets. The input from the business owner or the client will give you an idea on whether you model predicts customer churn with accuracy and provides desired results.

7) Mention some common problems that data analysts encounter during analysis.

- Having a poor formatted data file. For instance, having CSV data with un-escaped newlines and commas in columns.
- Having inconsistent and incomplete data can be frustrating.
- Common Misspelling and Duplicate entries are a common data quality problem that most of the data analysts face.
- Having different value representations and misclassified data.

8) What are the important steps in data validation process?

Data Validation is performed in 2 different steps-

Data Screening – In this step various algorithms are used to screen the entire data to find any erroneous or questionable values. Such values need to be examined and should be handled.

Data Verification- In this step each suspect value is evaluated on case by case basis and a decision is to be made if the values have to be accepted as valid or if the values have to be rejected as invalid or if they have to be replaced with some redundant values.

9) How will you create a classification to identify key customer trends in unstructured data?

A model does not hold any value if it cannot produce actionable results, an experienced data analyst will have a varying strategy based on the type of data being analysed. For example, if a customer complain was retweeted then should that data be included or not. Also, any sensitive data of the customer needs to be protected, so it is also advisable to consult with the stakeholder to ensure that you are following all the compliance regulations of the organization and disclosure laws, if any.

You can answer this question by stating that you would first consult with the stakeholder of the business to understand the objective of classifying this data. Then, you would use an iterative process by pulling new data samples and modifying the model accordingly and evaluating it for accuracy. You can mention that you would follow a basic process of mapping the data, creating an algorithm, mining the data, visualizing it and so on. However, you would accomplish this in multiple segments by considering the feedback from stakeholders to ensure that you develop an enriching model that can produce actionable results.

10) What is the criteria to say whether a developed data model is good or not?

- The developed model should have predictable performance.
- A good data model can adapt easily to any changes in business requirements.
- Any major data changes in a good data model should be scalable.

- A good data model is one that can be easily consumed for actionable results.

11) According to you what are the qualities/skills that a data analyst must posses to be successful at this position.
Problem Solving and Analytical thinking are the two important skills to be successful as a data analyst. One needs to skilled ar formatting data so that the gleaned information is available in a easy-to-read manner. Not to forget technical proficiency is of significant importance. You can also talk about other skills that the interviewer expects in an ideal candidate for the job position based on the given job description.

12) You are assigned a new data anlytics project. How will you begin with and what are the steps you will follow?
The purpose of asking this question is that the interviewer wants to understand how you approach a given data problem and what is the though process you follow to ensure that you are organized. You can start answering this question by saying that you will start with finding the objective of the given problem and defining it so that there is solid direction on what need to be done. The next step would be to do data exploration and familiarise myself with the entire dataset which is very important when working with a new dataset.The next step would be to prepare the data for modelling which would including finding outliers, handling missing values and validating the data. Having validated the data, I will start data modelling untill I discover any meaningfuk insights. After this the final step would be to implement the model and track the output results.
This is the generic data analysis process that we have explained in this answer, however, the answer to your question might slightly change based on the kind of data problem and the tools available at hand.

13) What do you know about interquartile range as data analyst?
A measure of the dispersion of data that is shown in a box plot is referred to as the interquartile range. It is the difference between the upper and the lower quartile.

Top 100 Data Scientist Interview Questions and Answers

2) Python or R – Which one would you prefer for text analytics?
The best possible answer for this would be Python because it has Pandas library that provides easy to use data structures and high performance data analysis tools.

3) Which technique is used to predict categorical responses?
Classification technique is used widely in mining for classifying data sets.

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

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

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

7) Differentiate between univariate, bivariate and multivariate analysis.
These are descriptive statistical analysis techniques which can be differentiated based on the number of variables involved at a given point of time. For example, the pie charts of sales based on territory involve only one variable and can be referred to as univariate analysis.
If the analysis attempts to understand the difference between 2 variables at time as in a scatterplot, then it is referred to as bivariate analysis. For example, analysing the volume of sale and a spending can be considered as an example of bivariate analysis.
Analysis that deals with the study of more than two variables to understand the effect of variables on the responses is referred to as multivariate analysis.

8) What do you understand by the term Normal Distribution?
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.

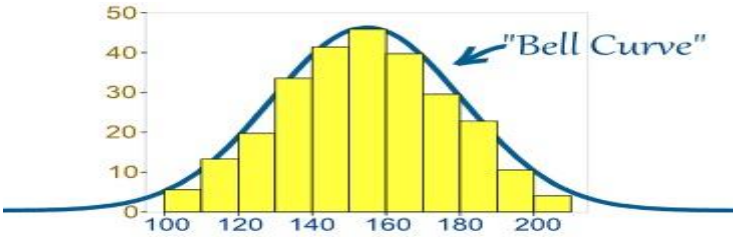


Image Credit : mathisfun.com

9) What is Linear Regression?

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

10) What is Interpolation and Extrapolation?

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

11) What is power analysis?

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

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

13) What is Collaborative filtering?

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

14) What is the difference between Cluster and Systematic Sampling?

Cluster sampling is a technique used when it becomes difficult to study the target population spread across a wide area and simple random sampling cannot be applied. Cluster Sample is a probability sample where each sampling unit is a collection, or cluster of elements. Systematic sampling is a statistical technique where elements are selected from an ordered sampling frame. In systematic sampling, the list is progressed in a circular manner so once you reach the end of the list,it is progressed from the top again. The best example for systematic sampling is equal probability method.

15) Are expected value and mean value different?

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

For Sampling Data

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

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

For Distributions

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

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

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

- P- 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.

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

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

18) What are categorical variables?

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

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

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

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

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

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

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

21) What is the difference between Supervised Learning 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.

22) Explain the use of Combinatorics in data science.

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

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

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

25) What is an Eigenvalue and Eigenvector?

Eigenvectors are used for understanding linear transformations. In data analysis, we usually calculate the eigenvectors for a correlation or covariance matrix. Eigenvectors are the directions along which a particular linear transformation acts by flipping, compressing or stretching. Eigenvalue can be referred to as the strength of the transformation in the direction of eigenvector or the factor by which the compression occurs.

26) What is Gradient Descent?

27) How can outlier values be treated?

Outlier values can be identified by using univariate or any other graphical analysis method. If the number of outlier values is few then they can be assessed individually but for large number of outliers the values can be substituted with either the 99th or the 1st percentile values. All extreme values are not outlier values.The most common ways to treat outlier values –

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

28) How can you assess a good logistic model?

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

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

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

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

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

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

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

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

- Understand the problem statement, understand the data and then give the answer.Assigning a default value which can be mean, minimum or maximum value. Getting into the data is important.
- If it is a categorical variable, the default value is assigned. The missing value is assigned a default value.
- If you have a distribution of data coming, for normal distribution give the mean value.
- Should we even treat missing values is another important point to consider? If 80% of the values for a variable are missing then you can answer that you would be dropping the variable instead of treating the missing values.

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

For some reason or the other, the response variable for a regression analysis might not satisfy one or more assumptions of an ordinary least squares regression. The residuals could either curve as the prediction increases or

follow skewed distribution. In such scenarios, it is necessary to transform the response variable so that the data meets the required assumptions. A Box cox transformation is a statistical technique to transform non-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.

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

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

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

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

In bayesian estimate we have some knowledge about the data/problem (prior) .There may be several values of the parameters which explain data and hence we can look for multiple parameters like 5 gammas and 5 lambdas that do this. As a result of Bayesian Estimate, we get multiple models for making multiple 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.

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

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

38) What is the curse of dimensionality?

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

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

41) What is Machine Learning?

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

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

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

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

44) How can you overcome Overfitting?

45) Differentiate between wide and tall data formats?

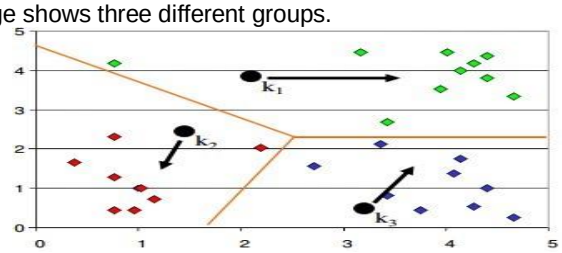
46) Is Naïve Bayes bad? If yes, under what aspects.

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

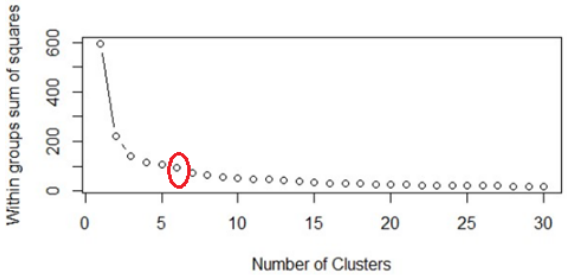
48) How will you define the number of clusters in a clustering algorithm?

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

For example, the following image shows three different groups.



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



Red circled point in above graph i.e. Number of Cluster =6 is the point after which you don't see any decrement in WSS. This point is known as bending point and taken as K in K – Means.

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

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

50) Is it possible to perform logistic regression with Microsoft Excel?

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

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

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

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

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

- Once you have built the model, you should check for following –
  - Global F-test to see the significance of group of independent variables on dependent variable
  - R^2
  - Adjusted R^2
  - RMSE, MAPE
- In addition to above mentioned quantitative metrics you should also check for-
  - Residual plot
  - Assumptions of linear regression

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

55) What do you understand by Recall and Precision?

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

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

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

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

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

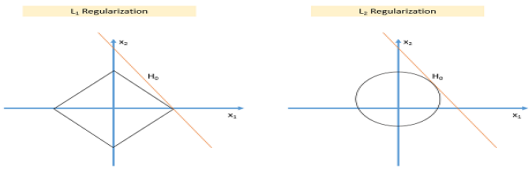
However , you might be wrong in some cases. For instance, you answer 15 times, 10 times the surprises you guess are correct and 5 wrong. This implies that your recall ratio is 100% but the precision is 66.67%.

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).

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

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

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



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

58) 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)

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

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

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

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

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

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

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

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

One more example might come from marketing. Let's say an ecommerce company decided to give $1000 Gift voucher to the customers whom they assume to purchase at least $5000 worth of items. They send free voucher mail directly to 100 customers without any minimum purchase condition because they assume to make at least 20% profit on sold items above 5K.

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

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

Assume there is an airport 'A' which has received high security threats and based on certain characteristics they identify whether a particular passenger can be a threat or not. Due to shortage of staff they decided to scan passenger being predicted as risk positives by their predictive model.

What will happen if a true threat customer is being flagged as non-threat by airport model?

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

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

63) Can you cite some examples where both false positive and false negatives are equally important?

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

Banks don't want to lose good customers and at the same point of time they don't want to acquire bad customers. In this scenario both the false positives and false negatives 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.

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

Validation set can be considered as a part of the training set as it is used for parameter selection and to avoid Overfitting of the model being built. On the other hand, test set is used for testing or evaluating the performance of a trained machine 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.

65) What makes a dataset gold standard?

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

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

Calculation of 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.

67) 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.

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

SVM and Random Forest are both used in classification problems.

a) If you are sure that your data is outlier free and clean then go for SVM. It is the opposite - if your data might contain outliers then Random forest would be the best choice

b) Generally, SVM consumes more computational power than Random Forest, so if you are constrained with memory go for Random Forest machine learning algorithm.
c) Random Forest gives you a very good idea of variable importance in your data, so if you want to have variable importance then choose Random Forest machine learning algorithm.
d) Random Forest machine learning algorithms are preferred for multiclass problems.
e) SVM is preferred in multi-dimensional problem set - like text classification
but as a good data scientist, you should experiment with both of them and test for accuracy or rather you can use ensemble of many Machine Learning techniques.

69) What do you understand by feature vectors?

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

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

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

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

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

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

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

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

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

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

75) What are the basic assumptions to be made for linear regression?

Normality of error distribution, statistical independence of errors, linearity and additivity.

76) 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)

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

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

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

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

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

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

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

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

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

81)

Puzzle based Data Science Interview Questions asked in Data Scientist Job Interviews

1) How many Piano Tuners are there in Chicago?

To solve this kind of a problem, we need to know –
Can you tell if the equation given below is linear or not ?
Emp_sal= $2000+2.5(emp\_age)^2$
Yes it is a linear equation as the coefficients are linear.
What will be the output of the following R programming code ?

```
var2<- c("I","Love,"DeZyre")
var2
```

It will give an error.

How many Pianos are there in Chicago?
How often would a Piano require tuning?
How much time does it take for each tuning?
We need to build these estimates to solve this kind of a problem. Suppose, let's assume Chicago has close to 10 million people and on an average there are 2 people in a house. For every 20 households there is 1 Piano. Now the question how many pianos are there can be answered. 1 in 20 households has a piano, so approximately 250,000 pianos are there in Chicago.
Now the next question is-"How often would a Piano require tuning? There is no exact answer to this question. It could be once a year or twice a year. You need to approach this question as the interviewer is trying to test your knowledge on whether you take this into consideration or not. Let's suppose each piano requires tuning once a year so on the whole 250,000 piano tunings are required.
Let's suppose that a piano tuner works for 50 weeks in a year considering a 5 day week. Thus a piano tuner works for 250 days in a year. Let's suppose tuning a piano takes 2 hours then in an 8 hour workday the piano tuner would be able to tune only 4 pianos. Considering this rate, a piano tuner can tune 1000 pianos a year.
Thus, 250 piano tuners are required in Chicago considering the above estimates.

2) There is a race track with five lanes. There are 25 horses of which you want to find out the three fastest horses. What is the minimal number of races needed to identify the 3 fastest horses of those 25?

Divide the 25 horses into 5 groups where each group contains 5 horses. Race between all the 5 groups (5 races) will determine the winners of each group. A race between all the winners will determine the winner of the winners and must be the fastest horse. A final race between the 2nd and 3rd place from the winners group along with the 1st and 2nd place of thee second place group along with the third place horse will determine the second and third fastest horse from the group of 25.

3) Estimate the number of french fries sold by McDonald's everyday.

4) How many times in a day does a clock's hand overlap?

5) You have two beakers. The first beaker contains 4 litre of water and the second one contains 5 litres of water.How can you our exactly 7 litres of water into a bucket?

6) A coin is flipped 1000 times and 560 times heads show up. Do you think the coin is biased?

7) Estimate the number of tennis balls that can fit into a plane.

8) How many haircuts do you think happen in US every year?

9) In a city where residents prefer only boys, every family in the city continues to give birth to children until a boy is born. If a girl is born, they plan for another child. If a boy is born, they stop. Find out the proportion of boys to girls in the city.

Probability Interview Questions for Data Science

1. There are two companies manufacturing electronic chip. Company A is manufactures defective chips with a probability of 20% and good quality chips with a probability of 80%. Company B manufactures defective chips with a probability of 80% and good chips with a probability of 20%.If you get just one electronic chip, what is the probability that it is a good chip?
2. Suppose that you now get a pack of 2 electronic chips coming from the same company either A or B. When you test the first electronic chip it appears to be good. What is the probability that the second electronic chip you received is also good?
3. A dating site allows users to select 6 out of 25 adjectives to describe their likes and preferences. A match is said to be found between two users on the website if the match on atleast 5 adjectives. If Steve and On a dating site, users can select 5 out of 24 adjectives to describe themselves. A match is declared between two users if they match on at least 4 adjectives. If Brad and Angelina randomly pick adjectives, what is the probability that they will form a match?
4. A coin is tossed 10 times and the results are 2 tails and 8 heads. How will you analyse whether the coin is fair or not? What is the p-value for the same?
5. Continuation to the above question, if each coin is tossed 10 times (100 tosses are made in total). Will you modify your approach to the test the fairness of the coin or continue with the same?
6. An ant is placed on an infinitely long twig. The ant can move one step backward or one step forward with same probability during discrete time steps. Find out the probability with which the ant will return to the starting point.

Statistics Interview Questions for Data Science

1. Explain the central limit theorem.
2. What is the relevance of central limit theorem to a class of freshmen in the social sciences who hardly have any knowledge about statistics?
3. Given a dataset, show me how Euclidean Distance works in three dimensions.
4. How will you prevent overfitting when creating a statistical model ?

Machine Learning Interview Questions for Data Scientists

1. Which is your favourite machine learning algorithm and why?
2. In which libraries for Data Science in Python and R, does your strength lie?
3. What kind of data is important for specific business requirements and how, as a data scientist will you go about collecting that data?

4. Tell us about the biggest data set you have processed till date and for what kind of analysis.
5. Which data scientists you admire the most and why?
6. Suppose you are given a data set, what will you do with it to find out if it suits the business needs of your project or not.
7. What were the business outcomes or decisions for the projects you worked on?
8. What unique skills you think can you add on to our data science team?
9. Which are your favorite data science startups?
10. Why do you want to pursue a career in data science?
11. What have you done to upgrade your skills in analytics?
12. What has been the most useful business insight or development you have found?
13. How will you explain an A/B test to an engineer who does not know statistics?
14. When does parallelism helps your algorithms run faster and when does it make them run slower?
15. How can you ensure that you don't analyse something that ends up producing meaningless results?
16. How would you explain to the senior management in your organization as to why a particular data set is important?
17. Is more data always better?
18. What are your favourite imputation techniques to handle missing data?
19. What are your favorite data visualization tools?
20. Explain the life cycle of a data science project.

How can you ensure that you don't analyse something that ends up producing meaningless results?

- Understanding whether the model chosen is correct or not.Start understanding from the point where you did Univariate or Bivariate analysis, analysed the distribution of data and correlation of variables and built the linear model.Linear regression has an inherent requirement that the data and the errors in the data should be normally distributed. If they are not then we cannot use linear regression. This is an inductive approach to find out if the analysis using linear regression will yield meaningless results or not.
- Another way is to train and test data sets by sampling them multiple times. Predict on all those datasets to find out whether or not the resultant models are similar and are performing well.
- By looking at the p-value, by looking at r square values, by looking at the fit of the function and analysing as to how the treatment of missing value could have affected- data scientists can analyse if something will produce meaningless results or not.

1) How can you build a simple logistic regression model in Python?

2) How can you train and interpret a linear regression model in SciKit learn?

3) Name a few libraries in Python used for Data Analysis and Scientific computations.

NumPy, SciPy, Pandas, SciKit, Matplotlib, Seaborn

4) Which library would you prefer for plotting in Python language: Seaborn or Matplotlib?

Matplotlib is the python library used for plotting but it needs lot of fine-tuning to ensure that the plots look shiny. Seaborn helps data scientists create statistically and aesthetically appealing meaningful plots. The answer to this question varies based on the requirements for plotting data.

5) What is the main difference between a Pandas series and a single-column DataFrame in Python?

6) Write code to sort a DataFrame in Python in descending order.

7) How can you handle duplicate values in a dataset for a variable in Python?

8) Which Random Forest parameters can be tuned to enhance the predictive power of the model?

9) Which method in pandas.tools.plotting is used to create scatter plot matrix?

Scatter_matrix

10) How can you check if a data set or time series is Random?

To check whether a dataset is random or not use the lag plot. If the lag plot for the given dataset does not show any structure then it is random.

11) Can we create a DataFrame with multiple data types in Python? If yes, how can you do it?

12) Is it possible to plot histogram in Pandas without calling Matplotlib? If yes, then write the code to plot the histogram?

13) What are the possible ways to load an array from a text data file in Python? How can the efficiency of the code to load data file be improved?

numpy.loadtxt ()

14) Which is the standard data missing marker used in Pandas?

NaN

15) Why you should use NumPy arrays instead of nested Python lists?

16) What is the preferred method to check for an empty array in NumPy?

17) List down some evaluation metrics for regression problems.

18) Which Python library would you prefer to use for Data Munging?

Pandas

19) Write the code to sort an array in NumPy by the nth column?

Using argsort () function this can be achieved. If there is an array X and you would like to sort the nth column then code for this will be x[x [: n-1].argsort ()]

20) How are NumPy and SciPy related?

21) Which python library is built on top of matplotlib and Pandas to ease data plotting?
Seaborn
22) Which plot will you use to access the uncertainty of a statistic?
Bootstrap
23) What are some features of Pandas that you like or dislike?
24) Which scientific libraries in SciPy have you worked with in your project?
25) What is pylab?
A package that combines NumPy, SciPy and Matplotlib into a single namespace.
26) Which python library is used for Machine Learning?
SciKit-Learn
27) How can you copy objects in Python?
The functions used to copy objects in Python are-
1) Copy.copy () for shallow copy
2) Copy.deepcopy () for deep copy
However, it is not possible to copy all objects in Python using these functions. For instance, dictionaries have a separate copy method whereas sequences in Python have to be copied by 'Slicing'.

28) What is the difference between tuples and lists in Python?
Tuples can be used as keys for dictionaries i.e. they can be hashed. Lists are mutable whereas tuples are immutable - they cannot be changed. Tuples should be used when the order of elements in a sequence matters. For example, set of actions that need to be executed in sequence, geographic locations or list of points on a specific route.
29) What is PEP8?
PEP8 consists of coding guidelines for Python language so that programmers can write readable code making it easy to use for any other person, later on.
30) Is all the memory freed when Python exits?
No it is not, because the objects that are referenced from global namespaces of Python modules are not always de-allocated when Python exits.
31) What does _init_.py do?
init.py is an empty py file used for importing a module in a directory. _init_.py provides an easy way to organize the files. If there is a module maindir/subdir/module.py,_init_.py is placed in all the directories so that the module can be imported using the following command-
import maindir.subdir.module
32) What is the different between range () and xrange () functions in Python?
range () returns a list whereas xrange () returns an object that acts like an iterator for generating numbers on demand.
33) How can you randomize the items of a list in place in Python?
Shuffle (lst) can be used for randomizing the items of a list in Python
34) What is a pass in Python?
Pass in Python signifies a no operation statement indicating that nothing is to be done.
35) If you are gives the first and last names of employees, which data type in Python will you use to store them?
You can use a list that has first name and last name included in an element or use Dictionary.
36) What happens when you execute the statement mango=banana in Python?
A name error will occur when this statement is executed in Python.
37) Write a sorting algorithm for a numerical dataset in Python.
38) Optimize the below python code-
word = 'word'
print word.__len__ ()
Answer: print 'word'._len_ ()
39) What is monkey patching in Python?
Monkey patching is a technique that helps the programmer to modify or extend other code at runtime. Monkey patching comes handy in testing but it is not a good practice to use it in production environment as debugging the code could become difficult.
40) Which tool in Python will you use to find bugs if any?
Pylint and Pychecker. Pylint verifies that a module satisfies all the coding standards or not. Pychecker is a static analysis tool that helps find out bugs in the course code.
41) How are arguments passed in Python- by reference or by value?
The answer to this question is neither of these because passing semantics in Python are completely different. In all cases, Python passes arguments by value where all values are references to objects.
42) You are given a list of N numbers. Create a single list comprehension in Python to create a new list that contains only those values which have even numbers from elements of the list at even indices. For instance if list[4] has an even value the it has be included in the new output list because it has an even index but if list[5] has an even value it should not be included in the list because it is not at an even index.

[x for x in list [1::2] if x%2 == 0]
The above code will take all the numbers present at even indices and then discard the odd numbers.

43) Explain the usage of decorators.

Decorators in Python are used to modify or inject code in functions or classes. Using decorators, you can wrap a class or function method call so that a piece of code can be executed before or after the execution of the original code. Decorators can be used to check for permissions, modify or track the arguments passed to a method, logging the calls to a specific method, etc.

44) How can you check whether a pandas data frame is empty or not?

The attribute df.empty is used to check whether a data frame is empty or not.

45) What will be the output of the below Python code –

```
def multipliers ():
   return [lambda x: i * x for i in range (4)]
   print [m (2) for m in multipliers ()]
```

The output for the above code will be [6, 6,6,6]. The reason for this is that because of late binding the value of the variable i is looked up when any of the functions returned by multipliers are called.

46) What do you mean by list comprehension?

The process of creating a list while performing some operation on the data so that it can be accessed using an iterator is referred to as List Comprehension.

Example:

```
[ord (j) for j in string.ascii_uppercase]
    [65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90]
```

47) What will be the output of the below code

```
word = 'aeioubcdfg'
print word [:3] + word [3:]
```

The output for the above code will be: 'aeioubcdfg'.

In string slicing when the indices of both the slices collide and a "+" operator is applied on the string it concatenates them.

48) list= ['a','e','i','o','u']

```
print list [8:]
```

The output for the above code will be an empty list []. Most of the people might confuse the answer with an index error because the code is attempting to access a member in the list whose index exceeds the total number of members in the list. The reason being the code is trying to access the slice of a list at a starting index which is greater than the number of members in the list.

49) What will be the output of the below code:

```
def foo (i= []):
   i.append (1)
   return i
>>> foo ()
>>> foo ()
```

The output for the above code will be-

[1]
[1, 1]

Argument to the function foo is evaluated only once when the function is defined. However, since it is a list, on every all the list is modified by appending a 1 to it.

50) Can the lambda forms in Python contain statements?

No, as their syntax is restrcited to single expressions and they are used for creating function objects which are returned at runtime.

This list of questions for Python interview questions and answers is not an exhaustive one and will continue to be a work in progress. Let us know in comments below if we missed out on any important question that needs to be up here.

Q1). What is Bias Error in machine learning algorithm?

Bias is the common error in the machine learning algorithm due to simplistic assumptions. It may undermine your data and does not allow you to achieve maximum accuracy. Further generalizing the knowledge from the training set to the test sets would be highly difficult for you.

Q2). What do you understand by Variance Error in machine learning algorithm?

Variance error is common in machine learning when the algorithm is highly complex and difficult to understand as well. It may lead high degree of variation to your training data that can lead the model to overfit the data. Also, there could be so much noise for the training data that is not necessary in case of the test data.

Q3). What is the bias-variance trade-off?

The bias-variance trade-off is able to handle the learning errors effectively and manages noise too that happens due to underlying

data, Essentially, this trade-off will make the model more complex than usual but errors are reduced optimally.

Q4). How will you differentiate the supervised and unsupervised machine learning?

Supervised learning needs data in the labeled form. For example, if you wanted to classify the data then you should first label the data then classify it into groups. On the other hand, unsupervised does not need any data labeling explicitly.

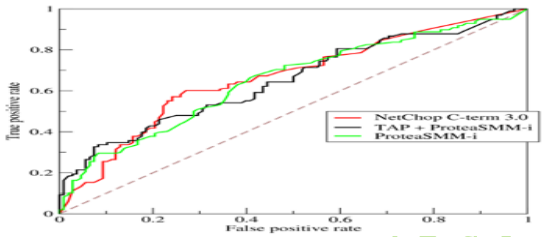
Q5). How is the k-nearest algorithm different from the KNN clustering?

K-nearest algorithm is the supervised learning while the k-means algorithm is assigned under the unsupervised learning. While these two techniques look similar at the first glance, still there is a lot of difference between the two. Supervised learning needs data in the labeled form.

For example, if you wanted to classify the data then you should first label the data then classify it into groups. On the other hand, unsupervised does not need any data labeling explicitly. The application of both the techniques depends on project needs.

Q6). What is ROC (Receiver operating characteristic) Curve? Explain the working of ROC.

A ROC curve is the pictorial representation of the contrast between true positive rates and the false positive rates calculated at multiple thresholds. It is used as the proxy to measure the trade-offs and sensitivity of the model. Based on the observation, it will trigger the false alarms.



Q7). What do you mean by the precision and the recall?

The Recall is the measure of true positive rates claimed against the total number of datasets. Precision is the prediction of positive values that your model claims compared to the number of positives it actually claims. It can be taken a special case of probability as well in case of mathematics.

Q8). What is the significance of Bayes' theorem in the context of the machine learning algorithm?

With the Bayes' Theorem, you could measure the posterior probability of an event based on your prior knowledge. It will tell you the true positive rate of a condition when divided by the sum of total false rates.

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

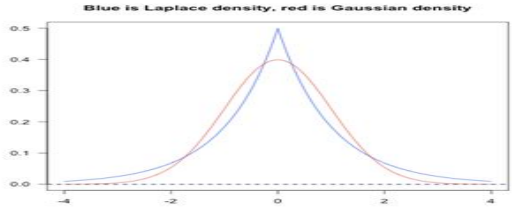
Bayes Theorem is also named as the Bayes Rule in mathematics, and it is popular for calculating the conditional probability. The name of the theorem was given after a popular mathematician Thomas Bayes. The concept of Bayes theorem is confusing sometimes but a depth understanding helps you to gain meaningful insights over the topic.

Q9). What is Naïve Bayes in machine learning?

Naïve is the word used to define the things that are virtually impossible in the real-life. Here, also you need to calculate the conditional probability as the pure product of individual probabilities of different components. This is the absolute condition that could never meet in the real-life. Have you ever heard of a pickle ice cream in actual?

Q10). How will you differentiate the L1 and L2 regularization?

L2 regularization trends to spread error among multiple terms while L! is more specific to binary variables where either 0 or 1 is assigned based on requirements. L1 tends to set a Laplacian prior on terms, but L2 tends to settings a Gaussian prior on terms.



Machine Learning Interview Questions Answers for Experienced

Q11). What is your favorite algorithm? Explain in less than a minute based on your past experiences.

The answer to this question will vary based on the projects you worked on earlier. Also, which algorithm assured better outcomes

as compared to other.

Q12). Have you ever worked on type 1 or Type 2 errors?

This is a tricky question usually asked by experienced candidates only. If you would be able to answer this question then make sure that you are at the top of the game. Type 1 error is the false positive and Type 2 error is a false negative. Type 1 error signifies something has happened even if it does not exist in real while Type 2 error means you claim something is happening in real.

Q13). How will you explain the Fourier Transformation in Machine Learning?

A Fourier Transformation is the generic method that helps in decomposing functions into a series of symmetric functions. It helps you in finding the set of cycle speeds, phases, and amplitude to match the particular time signal. It has the capability to convert the signal into frequency domain like sensor data or more.

Q14). How will you differentiate the machine learning and deep learning algorithms?

The deep learning is a part of machine learning that is usually connected with the neural networks. This is a popular technique from neuroscience to model a set of labeled and structured data more precisely. In brief, deep learning is an unsupervised learning algorithm that represents data with the help of neural nets.

Q15). How will you differentiate the generic model from the discriminative model?

A generic model will explain the multiple categories of data while the discriminative model simply tells the difference between data categories. They are used in classification tasks and need to understand deeply before you actually implement them.

Q16). What seems more important either model accuracy or performance of a model?

Well, model accuracy is just a subset of the model performance parameter. For a model who is performing excellent, there are chances of more accuracy than others.

Q17). What is the F1 score and explain its uses too?

The F1 score is used to check the performance of a model or this is the average of precision and recall of a model where 1 means the best and 0 means the worst.

Q18). Is it possible to manage imbalanced datasets in machine learning?

Collect more data, manage the imbalanced data, try a different algorithm to work on imbalanced datasets in machine learning.

Q19). Why is classification better than regression for machine learning experts?

Classification gives you discrete results while regression works on continuous results more. To become more specific with data points, you are always recommended using classification over regression in machine learning.

Q20). How would you check the effectiveness of machine learning model?

For this purpose, you can always check the F1 score to make sure either machine learning model is working effectively or needs improvement.

Q1). What is Artificial Intelligence?

AI is an area in the field of computer science, which emphasis on the implication of the human brain's cognitive functions into a machine/system, thereby making it work and act like humans. Some of the activities that can be carried out using computers infused with AI include:

Learning and planning

Speech recognition

Problem-solving

Q2). List some applications of AI/ What are the various areas where AI (Artificial Intelligence) can be used?

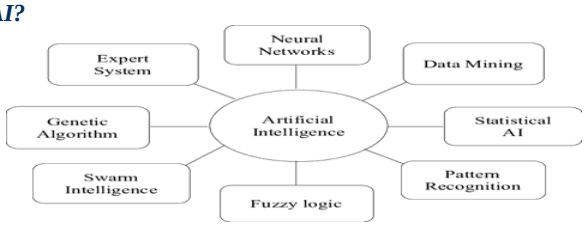
AI has a wide scope for implementation and it can be practically applied in fields of extreme diversity such as:

The linguistic field for Processing natural language

Customer support field as Chatbots, Humanoid customer support robots, sentiment analysis bots

IT field such as sales prediction, computing, computer software etc.

Q3). What Are the Branches Of AI?



Neural networks: An artificial neuron network (ANN) is an arithmetic model, based on the edifice and functionality of various biological neural networks.

Data mining: Artificial intelligence and data mining techniques have been in combination to solve issues related to diagnosis,

segmentation, classification and predictions.

Statistical AI: This branch of AI basically deals with domain models that disport both uncertainty and complexity in rational structures.

Pattern recognition: This primarily rivets on the apprehension of regularities and patterns in data.

Fuzzy logic: A type of multi-valued logic, where the truth values of the variables may vary anywhere between 0 and 1.

Swarm Intelligence: A biologically-inspired AI model that's based on the behavior of common insects such as bees, ants etc.

Genetic algorithm: A systematic combination of adaptive self-learning algorithm that gets its base from the concept of genetics.

Expert system: This is implemented in designing systems that is capable of emulating human's decision-making ability.

Q4). Explain the types of Artificial Intelligence.

Artificial intelligence can be classified into two main categories:

Strong artificial intelligence: This is basically creating real intelligence in an artificial way using the principle that even machines can be made sentimental. There are two types of strong AI: Human-like and Non-human like.

Weak artificial intelligence: These types of AI systems are created only to solve real-life problems and doesn't deal with the creation of extremely efficient human like intelligence.

Q5). What are the advantages of Fuzzy Logic Systems?

The Fuzzy logic system has the following key advantages:

The leverage to take inaccurate, malformed and clangorous input information.

Extremely easy to understandable and effortlessly constructible logics.

The flexibility to add and delete the rules as per our convenience in the FLS system.

Q6). What is an alternate key in AI?

Alternate Key: All the candidate keys except the primary keys are known as Alternate Keys.

Q7). What is the artificial key in AI?

Artificial Key: Creating a key artificially by assigning a number to individual record, in the absence of a standalone key.

Q8). What is compound key in AI?

Compound Key: Integration of various elements to generate an exceptional identified, in the absence of any data elements that specifically defines the subsistence within a construct.

Q9). What is are the core differences between supervised, unsupervised and reinforcement learning?

The difference can be the best explained using the following diagram:

Q10). How Game theory and AI related?

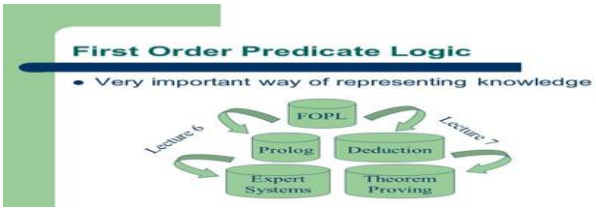
Artificial intelligence system makes use of the game theory for the purpose of enhancement as the requirement is always more than one participant. Hence, the relation between game theory and AI can be explained using the following two points:

Participant Design: Game theory is used to achieve maximum utility by enhancing a participant's decision

Mechanism Design: This is basically a type of Inverse game theory, where games a specifically designed focusing a group of ultra-smart participants.

Q11). What is FOPL?

FOPL is the abbreviation for First-order Predicate logic, which is a congregation of formal systems, with the statement being divided into two parts: a predicate and a subject. The predicate holds the potential to define or modify the subject's properties.



Q13). What is simulated annealing Algorithm?

Annealing is basically the process is of heating a metal and then immediately cooling it to make changes to its internal structure. The same principle is applied in computing where a probabilistic technique is put into practical use for the approximation of a given function's global optimum.

Q14). What is Greedy Best First Search Algorithm?

This is the algorithm process where the node closest to the goal will be expanded first. The default explanation of nodes goes by f(n) = h(n). This technique is applied at a later stage, where priority queue will come into the picture.

Q15). Share your previous project works based on your experience?

How is ML different from artificial intelligence?

AI involves machines that execute tasks which are programmed and based on human intelligence, whereas ML is a subset application of AI where machines are made to learn information. They gradually perform tasks and can automatically build models from the learnings.

Differentiate between statistics and ML.

In statistics, the relationships between relevant data (variables) is established; but in ML, the algorithms rely on data regardless of their statistical influence. In other words, statistics is concerned about inferences in the data whereas ML looks at optimisation.

What are neural networks and where do they find their application in ML? Elaborate.

Neural networks are information processing models that derive their functions based on biological neurons found in the human brain. The reason they are the choice of technique in ML is because, they help discover patterns in data that are sometimes too complex to comprehend by humans.

Differentiate between a parameter and a hyperparameter?

Parameters are attributes in training data that can be estimated during ML. Hyperparameters are attributes that cannot be determined beforehand in the training data. Example: Learning rate in neural networks.

What is 'tuning' in ML?

Generally, the goal of ML is to automatically provide accurate output from the vast amounts of input data without human intervention. Tuning is a process which makes this possible and it involves optimising hyperparameters for an algorithm or a ML model to make them perform correctly.

What is optimisation in ML?

Optimisation in general refers to minimising or maximising an objective function (in linear programming). In the context of ML, optimisation refers to tuning of hyperparameters which result in minimising the error function (or loss function).

What is the use of gradient descent?

The use of gradient descent plainly lies with the fact that it is easy to implement and is compatible with most of the ML algorithms when it comes to optimisation. This technique works on the principle of cost function.

Explain any data preprocessing technique for ML.

Standardisation: It is mainly used for algorithms following a Gaussian distribution. It can be done through scikit learn Standardscaler class (for Python).

What is dimensionality reduction? Explain in detail.

The process of reducing variables in a ML classification scenario is called Dimensionality reduction. The process is segregated into sub-processes called feature extraction and feature selection. Dimensionality reduction is done to enhance visualisation of training data. It finds the appropriate set of variables known as principal variables.

Explain Principal Component Analysis (PCA).

PCA is a dimensionality-reduction technique which mathematically transforms a set of correlated variables into a smaller set of uncorrelated variables called principal components.

What value do you optimise when using a support vector machine (SVM)?

For a linear function, SVM optimises the product of input vectors as well as the coefficients. In other words, the algorithm with the linear function can be restructured into a dot-product.

On what basis do you choose a classifier?

Classifiers must be chosen based on the accuracy it provides on the trained data. Also, the size of the dataset sometimes affects accuracy. For example, Naive Bayes classifiers suit smaller datasets in terms of accuracy due to higher asymptotic errors.

Which is better for image classification? Supervised or unsupervised classification. Justify.

In a supervised classification, the images are interpreted manually by the ML expert to create feature classes whereas this is not the case in unsupervised classification wherein the ML software creates feature classes based on image pixel values. Therefore, it is better to opt for supervised classification for image classification in terms of accuracy.

Mention key business metrics that help ML (company-specific).

Identify the key services/products/functions that holds good for ML. For example, if you consider a commercial bank, metrics such as number of new accounts, type of accounts, leads generated and so on, can be evaluated through ML methods.

These questions are usually encountered when you face an ML job interview. It is also suggested to go through all topics related to ML since it spans a large number of concepts and techniques.

Interview Questions on Machine Learning

Q1) What are the basic differences between Machine Learning and Deep Learning

Machine Learning Vs Deep Learning

	Machine Learning	**Deep Learning**
Definition	Sub-discipline of AI	Subset of machine learning
Data	Parses the data	Creates an artificial neural network
Accuracy	Requires manual intervention means decreased accuracy	Self-learning capabilities mean higher accuracy
Interpretability	Machine Learning is Faster	10 Times Faster than ML
Output	ML models produce a numerical output	DL algorithms can range from an image to text or even an audio
Data dependencies	High	Low
Hardware dependencies	Can work on low-end machines.	Heavily depend on high-end machines
Future	Effective with image recognition and face recognition in mobiles	Not much effective due to data processing limitations

Q2) What is the difference between Bias and Variance?

Bias:

Bias can be defined as a situation where an error has occurred due to use of assumptions in the learning algorithm.

Variance:

Variance is an error caused because of the complexity of the algorithm that is been used to analyze the data.

Q3) What is the difference between supervised and unsupervised machine learning?

A Supervised learning is a process where it requires training labeled data. When it comes to Unsupervised learning it doesn't require data labeling.

Q4) How is KNN different from K-means clustering?

KNN stands for K- Nearest Neighbours, it is classified as a supervised algorithm.

K-means is an unsupervised cluster algorithm.

Q5) Comparision between Machine Learning and Big Data

Machine Learning Vs Big Data		
Feature	**Machine Learning**	**Big Data**
Data Use	Technology that helps in reducing human intervention.	Data research, especially when working with huge data.
Operations	Existing data helps to tech machine what can be done further	Design patterns with analytics on existing data in terms of decision making.
Pattern Recognition	Similar to Big Data, existing data helps in pattern recognition.	Sequence and classification analysis helps in pattern recognition.
Data Volume	Best performance, while working with small-datasets.	Datasets helps in understanding and solving problems associated with large data volumes.
Application	Read existing data to predict future	Storing and analysing patterns within huge data

	information.	volumes.

Q6) Explain what is precision and Recall?

Recall:

It is known as a true positive rate. The number of positives that your model has claimed compared to the actual defined number of positives available throughout the data.

Precision:

It is also known as a positive predicted value. This is more based on the prediction. It is a measure of a number of accurate positives that the model claims when compared to the number of positives it actually claims.

Q7) What is your favorite algorithm and also explain the algorithm in briefly in a minute?

This type of questions is very common and asked by the interviewers to understand the candidate skills and assess how well he can communicate complex theories in the simplest language.

This one is a tough question and usually, individuals are not at all prepared for this situation so please be prepared and have a choice of algorithms and make sure you practice a lot before going into any sort of interviews.

Q8) What is the difference between Type 1 and Type 2 errors?

Type 1 error is classified as a false positive. I.e. This error claims that something has happened but the fact is nothing has happened. It is like a false fire alarm. The alarm rings but there is no fire.

Type 2 error is classified as a false negative. I.e. This error claims that nothing has happened but the fact is that actually, something happened at the instance.

The best way to differentiate a type 1 vs type 2 error is:

Calling a man to be pregnant- This is Type 1 example

Calling pregnant women and telling that she isn't carrying any baby- This is type 2 example

Q9) Define what is Fourier Transform in a single sentence?

A process of decomposing generic functions into a superposition of symmetric functions is considered to be a Fourier Transform.

Q10) What is deep learning?

Deep learning is a process where it is considered to be a subset of machine learning process.

Q11) What is the F1 score?

The F1 score is defined as a measure of a model's performance.

Q12) How is F1 score is used?

The average of Precision and Recall of a model is nothing but F1 score measure. Based on the results, the F1 score is 1 then it is classified as best and 0 being the worst.

Q13) How can you ensure that you are not overfitting with a particular model?

In Machine Learning concepts, they are three main methods or processes to avoid overfitting:

Firstly, keep the model simple

Must and should use cross validation techniques

It is mandatory to use regularization techniques, for example, LASSO.

Q14) How to handle or missing data in a dataset?

An individual can easily find missing or corrupted data in a data set either by dropping the rows or columns. On contrary, they can decide to replace the data with another value.

In Pandas they are two ways to identify the missing data, these two methods are very useful.

isnull() and dropna().

Q15) Do you have any relevant experience on Spark or any of big data tools that are used for Machine Learning?

Well, this sort of question is tricky to answer and the best way to respond back is, to be honest. Make sure you are familiar with Big data is and the different tools that are available. If you know about Spark then it is always good to talk about it and if you are unsure then it is best, to be honest and let the interviewer know about it.

So for this, you have to prepare what is Spark and its good to prepare other available Big data tools that are used for Machine learning.

Q16) Pick an algorithm and write a Pseudocode for the same?

This question depicts your understanding of the algorithm. This is something that one has to be very creative and also should have in-depth knowledge about the algorithms and first and foremost the individual should have a good understanding of the algorithms. Best way to answer this question would be start off with Web Sequence Diagrams.

Q17) What is the difference between an array and Linked list?

An array is an ordered fashion of collection of objects.

A linked list is a series of objects that are processed in a sequential order.

Q18): Define a hash table?

They are generally used for database indexing.

A hash table is nothing but a data structure that produces an associative array.

Q19) Mention any one of the data visualization tools that you are familiar with?

This is another question where one has to be completely honest and also giving out your personal experience with these type of tools are really important. Some of the data visualization tools are Tableau, Plot.ly, and matplotlib.

Q20) What is your opinion on our current data process?

This type of questions are asked and the individuals have to carefully listen to their use case and at the same time, the reply should be in a constructive and insightful manner. Based on your responses, the interviewer will have a chance to review and understand whether you are a value add to their team or not.

Q21) Please let us know what was your last read book or learning paper on Machine Learning?

This type of question is asked to see whether the individual has a keen interest towards learning and also he is up to the latest market standards. This is something that every candidate should be looking out for and it is vital for individuals to read through the latest publishings.

Q22) What is your favorite use case for machine learning models?

The decision tree is one of my favorite use case for machine learning models.

Q23) Is rotation necessary in PCA?

Yes, the rotation is definitely necessary because it maximizes the differences between the variance captured by the components.

Q24) What happens if the components are not rotated in PCA?

It is a straight effect. If the components are not rotated then it will diminish eventually and one has to use a lot of various components to explain the data set variance.

Q25) Explain why Navie Bayes is so Naive?

It is based on an assumption that all of the features in the data set are important, equal and independent.

Q25) How Recall and True positive rate are related?

The relation is

True Positive Rate = Recall.

Q27) Assume that you are working on a data set, explain how would you select important variables?

The following are few methods can be used to select important variables:

1. Use of Lasso Regression method.
2. Using Random Forest, plot variable importance chart.
3. Using Linear regression.

Q28) Explain how we can capture the correlation between continuous and categorical variable?

Yes, it is possible by using ANCOVA technique. It stands for Analysis of Covariance.

It is used to calculate the association between continuous and categorical variables.

Q29) Explain the concept of machine learning and assume that you are explaining this to a 5-year-old baby?

Yes, the question itself is the answer.

Machine learning is exactly the same way how babies do their day to day activities, the way they walk or sleep etc. It is a common fact that babies cannot walk straight away and they fall and then they get up again and then try. This is the same thing when it comes to machine learning, it is all about how the algorithm is working and at the same time redefining every time to make sure the end result is as perfect as possible.

One has to take real time examples while explaining these questions.

Q30) What is the difference between Machine learning and Data Mining?

Data mining is about working on unstructured data and then extract it to a level where the interesting and unknown patterns are identified.

Machine learning is a process or a study whether it closely relates to design, development of the algorithms that provide an ability to the machines to capacity to learn.

Q31) What is inductive machine learning?

Inductive machine learning is all about a process of learning by live examples.

Q32) Please state few popular Machine Learning algorithms?

1. Nearest Neighbour
2. Neural Networks
3. Decision Trees etc
4. Support vector machines

Q33) What are the different types of algorithm techniques are available in machine learning?

Some of them are :

1. Supervised learning
2. Unsupervised learning
3. Semi-supervised learning
4. Transduction
5. Learning to learn

Q34) What are the three stages to build the model in machine learning:

1. Model building
2. Model testing
3. Applying the model

1) What is Machine learning?

Machine learning is a branch of computer science which deals with system programming in order to automatically learn and improve with experience. For example: Robots are programed so that they can perform the task based on data they gather from sensors. It automatically learns programs from data.

2) Mention the difference between Data Mining and Machine learning?

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

3) What is 'Overfitting' in Machine learning?

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

4) Why overfitting happens?

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

5) How can you avoid overfitting ?

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

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

6) What is inductive machine learning?

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

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

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

8) What are the different Algorithm techniques in Machine Learning?

The different types of techniques in Machine Learning are

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

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

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

10) What is the standard approach to supervised learning?

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

11) What is 'Training set' and 'Test set'?

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

12) List down various approaches for machine learning?

The different approaches in Machine Learning are

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

13) What is not Machine Learning?

a) Artificial Intelligence
b) Rule based inference

14) Explain what is the function of 'Unsupervised Learning'?

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

15) Explain what is the function of 'Supervised Learning'?

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

16) What is algorithm independent machine learning?

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

17) What is the difference between artificial learning and machine learning?

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

18) What is classifier in machine learning?

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

19) What are the advantages of Naive Bayes?

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

20) In what areas Pattern Recognition is used?

Pattern Recognition can be used in

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

21) What is Genetic Programming?

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

22) What is Inductive Logic Programming in Machine Learning?

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

23) What is Model Selection in Machine Learning?

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

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

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

a) Platt Calibration
b) Isotonic Regression

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

25) Which method is frequently used to prevent overfitting?

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

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

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

27) What is Perceptron in Machine Learning?

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

28) Explain the two components of Bayesian logic program?

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

29) What are Bayesian Networks (BN) ?

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

30) Why instance based learning algorithm sometimes referred as Lazy learning algorithm?

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

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

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

32) What is ensemble learning?

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

33) Why ensemble learning is used?

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

34) When to use ensemble learning?

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

35) What are the two paradigms of ensemble methods?

The two paradigms of ensemble methods are

a) Sequential ensemble methods
b) Parallel ensemble methods

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

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

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

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

38) What is an Incremental Learning algorithm in ensemble?

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

39) What is PCA, KPCA and ICA used for?

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

40) What is dimension reduction in Machine Learning?

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

41) What are support vector machines?

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

42) What are the components of relational evaluation techniques?

The important components of relational evaluation techniques are

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

43) What are the different methods for Sequential Supervised Learning?

The different methods to solve Sequential Supervised Learning problems are

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

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

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

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

45) What is batch statistical learning?

Statistical learning techniques allow learning a function or predictor from a set of observed data that can make predictions about unseen or future data. These techniques provide guarantees on the performance of the learned predictor on the future unseen data based on a statistical assumption on the data generating process.

46) What is PAC Learning?

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

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

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

48) What is sequence learning?

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

49) What are two techniques of Machine Learning ?

The two techniques of Machine Learning are

a) Genetic Programming
b) Inductive Learning

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

The recommendation engine implemented by major ecommerce websites uses Machine Learning

What is the difference between supervised and unsupervised machine learning?

Supervised learning requires training using labelled data. For example, in order to do classification, which is a supervised learning task, you'll first need to label the data you'll use to train the model to classify data into your labelled groups. Unsupervised learning, in divergence, does not require labeling data explicitly.

What's the trade-off between bias and variance?

Bias is error due to over simplistic assumptions in the learning algorithm that you are using, which can lead to model *under fitting* your data and making it hard for model to give accurate predictions.

Variance, on the other hand, is error due to way too much complexity in your learning algorithm. Due to this complexity, the algorithm is highly sensitive to high degrees of variation, which can lead your model to *over fit* the data. In addition, you will be carrying too much noise from your training data for your model to be useful.

The bias-variance decomposition essentially decomposes the learning error from any algorithm by adding the bias, the variance and a bit of irreducible error due to noise in the underlying data-set. Essentially, if you make the model more complex and add more variables, you'll lose bias but gain some variance — in order to get the optimally reduced amount of error, you'll have to trade-off bias and variance. You don't want either high bias or high variance in your model.

How KNN is different from k-means clustering?

The crucial difference between both is, K-Nearest Neighbor is a *supervised classification algorithm,* whereas k-means is an *unsupervised clustering algorithm.* While the procedure may seem similar at first, what it really means is that in order to K-Nearest Neighbors to work, you need labelled data which you want to classify an unlabeled point into. In k-means clustering it requires set of unlabeled points and a threshold only. The algorithm will take that unlabeled data and will learn how to cluster them into groups by computing the mean of the distance between different points.

What is Bayes' Theorem? How it is useful?

Bayes' Theorem gives you the posterior probability of an event given what is known as prior knowledge. Also, it is the basis behind Naive Bayes classifier. That's something important to know when you're faced with machine learning interview questions and answers. For detailed and simple explanation of Naïve Bayes visit here.

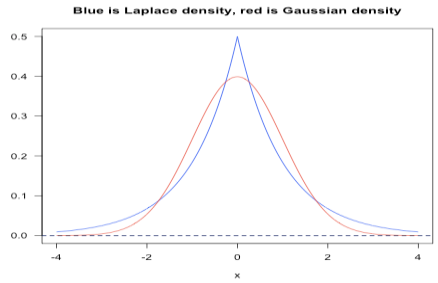
$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Likelihood — Class Prior Probability — Posterior Probability — Predictor Prior Probability

$$P(c \mid \mathbf{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

What is the difference between L1 and L2 regularization?

First, regularization is the technique which helps to solve over fitting problem in Machine Learning. L2 regularization incline to spread error among all the terms, while L1 is more binary, with most variables either being assigned a 1 or 0 in weighting. L1 corresponds to setting a Laplacean prior on the terms, while L2 corresponds to a Gaussian prior. I would say as a rule of thumb, one should always go with L2 in practice.



What is the difference between Type I and Type II error?

Don't think this as something high level stuff, interviewers ask questions in such terms just to know that you have all the bases and you are on the top.

Type I error is *false positive*, while Type II is false negative. Type I error is claiming something has happened when it hasn't. For instance, telling a man he is pregnant. On the other hand, Type II error means you claim nothing is happened but in fact something is. To exemplify, you tell a pregnant lady she isn't carrying baby.

What is the difference between Probability and Likelihood?

Not going too deep in technical, Probability quantifies prediction of outcome, likelihood quantifies trust in model. For instance, someone challenges us to a 'profitable gambling game'. Then, probabilities will serve us to compute things like the expected profile of your gains and losses. In contrast, likelihood will serve us to quantify whether we trust those probabilities in the first place; or whether we *smell a rat*.

What Deep Learning is exactly?

Most people don't know this but Machine Learning and Deep Learning is not two different things, but Deep learning is a *subset* of Machine learning. It mostly deals with *neural networks:* how to use back propagation and other certain principles from neuroscience to more accurately model large sets of unlabeled data. In a nutshell, Deep Learning represents unsupervised learning algorithm that learns data representation mainly through neural networks. Explore a little about neural nets to answer deep learning interview questions effectively.

What's the difference between a generative and discriminative model?

A discriminative model will learn the distinction between different categories of data, while A generative model will learn categories of data. Discriminative models will predominantly outperform generative models on classification tasks.

What is Time Series Analysis/Forecasting?

A Machine Learning data-set is a collection of observations. For example,

- Observation 1
- Observation 2
- Observation 3

But, a Time series data-set is different. Time series adds an explicit order dependence between observations: a time dimension. This additional dimension is both a constraint and a structure that provides a source of additional information.

- Time 1, Observation
- Time 2, Observation
- Time 3, Observation

How would you handle an imbalanced data-set?

Imbalanced data is, for example, you have 90% of the data in one class and 10% in other. Which leads to problems such as, no predictive power on the other category of data. Here are few techniques to get over it,

- Obviously collect more data to balance

- Try different algorithm (Not going to work effectively)
- Correct the imbalance in data-set

Explain Pruning in Decision trees.

Pruning is you remove branches that have weak predictive power in order to reduce the complexity of the model and in addition increase the predictive accuracy of a decision tree model. There are several flavors which includes, bottom-up and top-down pruning, with approaches such as reduced error pruning and cost complexity pruning.

In your opinion which one is more important: Model accuracy or Model Performance?

Questions like these tests your grasp over Machine Learning model performance and often look towards details. There are models with higher accuracy that can perform worse in predictive power, how does that make sense?
Well, it has everything to do with how model accuracy is only a subset of model performance, and at that, a sometimes misleading one. For example, if you wanted to detect fraud in a massive data-set with a sample of millions, a more accurate model would most likely predict no fraud at all if only a vast minority of cases were fraud. However, this would be useless for a predictive model — a model designed to find fraud that asserted there was no fraud at all! Questions like this help you demonstrate that you understand model accuracy isn't the be-all and end-all of model performance.

What's the F1 score?

It is a measure of model's performance. More technically, it is a weighted average of the precision and recall of the model, with results 1 being the best and 0 being the worst.

When should you use classification over regression?

Classification and Regression are both different in meaning. Classification produces discreate values while Regression gives you continuous results. You would use classification over regression, for example, when you wanted to know whether a name was male or female rather than just how correlated they were with male and female names.

What is convex hull?

Convex hull represents the *outer boundaries* of the two group of data points. Once convex hull is created, we get maximum margin hyperplane (MMH), which attempts to create the greatest separation between two groups, as a perpendicular bisector between two convex hulls.

General Questions

Do you have research experience in machine learning?

Machine Learning is emerging and no one wants novice players in their teams. Most employers hiring for Machine Learning position will look for your experience in the field. Research papers, co-authored or supervised by leaders in the field, can set you apart from the herd. Make sure you are ready with all the summary and justification of the work you have done in the past years.

What are the last Machine Learning papers you read? Why you think that was important?

As this field is emerging day by day, it is crucial to keep up with the latest scientific literatures to show that you are really into Machine Learning and not here just because it is the latest buzzword. Some good books to start with includes Deep Learning by Ian Goodfellow.

How would you approach the "Netflix Prize" competition?

The Netflix Prize was a famed competition where Netflix offered $1,000,000 for a better collaborative filtering algorithm. The team that won called BellKor had a 10% improvement and used an ensemble of different methods to win. Some familiarity with the case and its solution will help demonstrate you've paid attention to machine learning for a while.

What's your favorite algorithm, and can you explain it to me in less than a minute?

This type of question mainly tests your ability of communicating complex and technical nuances with poise and the ability to summarize quickly and efficiently. Make sure you have a choice of algorithm which you can explain easily. Try to explain different algorithms so simply and effectively that a five-year-old could grasp the basics.

Where do you usually source data-sets?

This type of questions are the real tie-breakers. If someone is going for an interview, he/she must know the drill of some related question. It is questions like this which purely illustrates your interest in Machine Learning. See my post for detailed answer on where to find machine learning data-sets.

How do you think Google is training data for self-driving cars?

Questions like this check your understanding of current affairs in the industry and how things at certain level works. Google is currently using *recaptcha* to source labelled data on storefronts and traffic signs. They are also building on training data collected by Sebastian Thrun at GoogleX.

Industry Specific Questions

How would you implement a recommendation system for our company's users?

There will be a lot of questions like this which will involve implementation of machine learning models to their company's problems. You should definitely study company's profile and its products before going in. In addition,

factors such as, financials of the company, in which the company operates, what are their users will help you get a clearer picture.

How can we use your machine learning skills to generate revenue?

This is a tricky question, I would say. The ideal answer would demonstrate knowledge of what drives the business and how your skills could relate. To exemplify, if you were interviewing for Spotify, you could remark that your skills at developing a better recommendation model would remarkably increase user retention, which would then increase revenue in the long run. Or something like that.

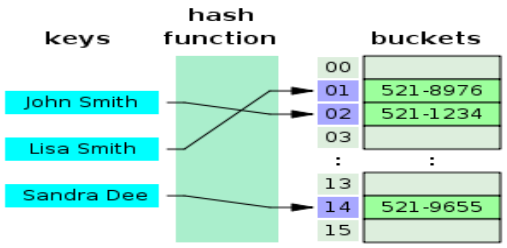
Practical/Programming Questions

How will you handle missing data?

One can find missing data in a data-set and either drop those rows or columns, or decide to replace them with another value. In python library *Pandas* there are two useful functions which will be helpful, isnull() and dropna().

Describe a hash table.

A hash table is a data structure that produces an associative array. A key is mapped to certain values through the use of a hash function. They are often used for tasks such as database indexing.



Which data visualization libraries do you use and why they are useful?

What's important here is to define your views on how to properly visualize data and your personal preferences when it comes to tools. Popular tools include R's ggplot, Python's seaborn and matplotlib, and tools such as Plot.ly and Tableau.

Do you have experience with Spark or big data tools for machine learning?

Spark is the big data tool most in demand now, able to handle immense data-sets with speed. Be honest if you don't have experience with the tools demanded, but also take a look at job descriptions and see what tools pop up: you'll want to invest in familiarizing yourself with them.

Considering the long list of machine learning algorithm, given a data set, how do you decide which one to use?

This is one more tricky question. Given what type of data there is, discrete, time series, continuous, you should give your answers.

Q1. What is Tensorflow?

TensorFlow is a machine learning library created by the Brain Team of Google and made open source in 2015. Basically, Tensorflow is a low-level toolkit for doing complicated math and it offers the users customizability to build experimental learning architectures, to work with them and to turn them into running software.

Follow the link to learn more about TensorFlow

Q2. What does the latest release of TensorFlow have to offer?

The latest release of TensorFlow is 1.7.0 and is available on www.tensorflow.org. It has been designed with deep learning in mind but applicable to a much wider range of problems.

Q3. What are Tensors?

Tensors are higher dimensional arrays which are used in computer programming to represent a multitude of data in the form of numbers. There are other n-d array libraries available on the internet like Numpy but TensorFlow stands apart from them as it offers methods to create tensor functions and automatically compute derivatives.

Q4. What is a TensorBoard?

TensorBoard, a suit of visualizing tools, is an easy solution to Tensorflow offered by the creators that lets you visualize the graphs, plot quantitative metrics about the graph with additional data like images to pass through it.

Follow the link to learn more about TensorBoard

Q5. What are the features of TensorFlow?

Tensorflow has APIs for Matlab, and C++ and has a wide language support. With each day passing by, researchers working on making it more better and recently in the latest Tensorflow Summit, tensorflow.js, a javascript library for training and deploying machine learning models has been introduced.

Q6. List a few advantages of TensorFlow?

- It has platform flexibility
- It is easily trainable on CPU as well as GPU for distributed computing.
- TensorFlow has auto differentiation capabilities
- It has advanced support for threads, asynchronous computation, and queue es.
- It is a customizable and open source.

Q7. List a few limitations of Tensorflow.

- Has GPU memory conflicts with Theano if imported in the same scope.
- No support for OpenCL
- Requires prior knowledge of advanced calculus and linear algebra along with a pretty good understanding of machine learning.

Q8. What are TensorFlow servables?

These are the central rudimentary units in TensorFlow Serving. Objects that clients use to perform the computation are called Servables.

The size of a servable is flexible. A single servable might include anything from a lookup table to a single model to a tuple of inference models.

Q9. What do the TensorFlow managers do?

Tensorflow Managers handle the full lifecycle of Servables, including:

- Loading Servables
- Serving Servables
- Unloading Servables

Q10. What are TensorFlow loaders?

Tensorflow Loaders are used for adding algorithms and data backends one of which is tensorflow itself. For example, a loader can be implemented to load, access and unload a new type of servable machine learning model.

TensorFlow Interview Questions and Answers for Freshers. Q- 1,2,3,4,5,6,7,8

TensorFlow Interview Questions and Answers for Experience. Q- 9, 10

Q11. What is deep speech?

Deep Speech developed by Mozilla is a TesnsorFlow implementation motivated by Baidu's Deep Speech architecture.

Q12.What do you mean by sources in TensorFlow?

Sources are in simple terms, modules that find and provide servables. Each Source provides zero or more servable streams. One Loader is supplied for each servable version it makes available to be loaded.

Q13. How does TensorFlow make use of the python API?

Python is the most recognisable and "the main" language when it comes to TensorFlow and its development. The first language supported by TensorFlow and still supports most of the features. It seems as TensorFlow's functionality first define in Python and then moved to C++.

Q14. What are the APIs inside the TensorFlow project?

The API's inside TensorFlow are still Python-based and they have low-level options for its users such as tf.manual or tf.nnrelu which use to build neural network architecture. These APIs also use to aid designing deep neural network having higher levels of abstraction.

Q15. What are the APIs outside TensorFlow project?

- **TFLearn:** This API shouldn't be seen as TF Learn, which is TensorFlow's tf.contrib.learn. It is a separate Python package.
- **TensorLayer:** It comes as a separate package and is different from what TensorFlow's layers API has in its bag.
- **Pretty Tensor:** It is actually a Google project which offers a fluent interface with chaining.
- **Sonnet:** It is a project of Google's DeepMind which features a modular approach.

Q16. How does TensorFlow use the C++ API?.

The runtime of TensorFlow is written in C++ and mostly C++ is connected to TensorFlow through header files in tensorflow/cc. C++ API still is in experimental stages of development but Google is committed to working with C++.

Q17. In TensorFlow, what exactly Bias and Variance are? Do you find any similarity between them?

In the learning algorithms, Biases can consider as errors which can result in failure of entire model and can alter the accuracy. Some experts believe these errors are essential to enable leaner's gain knowledge from a training point of view.

Q18. Can TensorFlow be deployed in container software?

Tensorflow can also use with containerization tools such as docker, for instance, it could use to deploy a sentiment analysis model which uses character level ConvNet networks for text classification.

Q19. What exactly Neural Networks are? What are the types of same you are familiar with?

Neural networks as the name suggests are a network of elemental processing entities that together make a complex form. There can be Artificial Neural Networks and Biological Neural Networks. The use of artificial neural networks is more common as they try to imitate the mechanics of the human brain.

Have a look at Recurrent Neural Network

Q20. What are the general advantages of using the Artifical Neural Networks?

They use to find answers to complex problems in a stepwise manner. All the information that a network can receive can easily be in any format. They also make use of real-time operations along with a good tolerance capability.

Let's revise Convolutional Neural Network

TensorFlow Interview Questions and Answers for Freshers. Q- 11,12,14,15,16,19

TensorFlow Interview Questions and Answers for Experience. Q- 13,17,18,20

Q21. What exactly do you know about Recall and Precision?

The other name of Recall is the true positive rate. It is the overall figure of positiveness a model can generally claim. The predictive value which is generally positive in nature is Precision. The difference between the true positive rate and claimed positive rate can be defined with the help of both these options.

Q23. What are some advantages of TensorFlow over other libraries?

Debugging facility, scalability, visualization of data, pipelining and many more.

Q24. How can you make sure that overfitting situation is not arriving with a model you are using?

Users need to make sure that their model is simple and not have any complex statement. Variance takes into the account and the noise eliminates from the model data. Techniques like k-fold and LASSO can also help.

Q25. What exactly do you know about a ROC curve and its working?

ROC or region of convergence used to reflect data rates which classify as true positive and false positive. Represented in the form of graphs, it can use as a proximity to swap operations related to different algorithms.

Q26. In the machine learning context, how useful and reliable Bayes' theorem is?

Baye's theorem is useful for determining the probability of an event, obtained by dividing the actual positive rate by the false positive rate. Some of the very complex questions and challenges can solve using a simple approach and eliminated with the help of this theorem.

Q27. What difference do you find in Type I and Type II errors?

Type I error means a false positive value while Type II error means a false negative.

Q28. What would be your strategy to handle a situation indicating an imbalanced dataset?

This usually occurs when a vast set of data keep in a single class. Sampling the dataset again is one of the possible solutions and the other one being the migration of data to parallel classes. The dataset should not be damaged.

Q29. What do you know about supervised and unsupervised machine learning?

Supervised learning consists of labelled data which is not necessarily present in unsupervised learning.

Q30. In machine learning based on TensorFlow, what is more important among the performance or the accuracy of a model?

It depends on the overall experience. Both of them have an equal weightage although accuracy is most important in most of the models.

Q1. How you place operations on a particular device?
You should create the operations within a with tf.device(name): context to place them on a particular device.
Q2. Which client languages are supported in TensorFlow?
TensorFlow supports multiple client languages, the best language being **Python**. There are experimental interfaces that are available for C++ **Java** and Go. Bindings for various other languages (such as C#, Julia, Ruby and **Scala**) are created and supported by the opensource community.
Q3. Do Sessions have a lifetime? What about intermediate tensors?
Resources like as tf.Variable, tf.QueueBase, and tf.ReaderBase; own by a session and may use a significant amount of memory which are released when the session is terminated with tf.Session.close.
Q4. What's the deal with feeding and placeholders?
Feeding is a phenomenon that allows you to substitute different values for one or more Tensors at the runtime. The feed_dict argument is used to map tf.Tensorobjects to numpy arrays for further executions.
Q5. Why does Session.run() hang when using a reader or a queue?
The tf.ReaderBase and tf.QueueBase classes provide special operations that become blocked since the input isn't available. They allow building clear input pipelines, by making the computation a little more complicated.
Q6. What is the lifetime of a variable?
A variable is created when you first run the tf.Variable.initializer operation for that variable in a session. It gets destroyed when that tf.Session.close.
Q7. How do variables behave when they are concurrently accessed?
Variables allow concurrency in read/write ops. The variable value may change when the concurrently updates. By default, there is no mutex (mutual exclusion).
Q8. What is the simplest way to send data to TensorBoard?
First of all, you should add summary operations to your graph, and then log them in a log directory. Then, TensorBoard should be started using:
python tensorflow/tensorboard/tensorboard.py –logdir=path/to/log-directory
Follow this link to learn more about TensorBoard
Q9. What exactly do you know about Bias-Variance decomposition?
It generally uses to decompose problems such as errors that occur during learning in different algorithms. Bias keeps reducing if the data is to be made complex. Trading off the Variance and Bias are very essential to get results that are totally free from errors.
Q10. How is k-means clustering different from KNN?
It is an unsupervised learning algorithm used for clustering. On the other hand, the KNN is a structured clustering algorithm. They both share some similarities but users need to label the data in the KNN which is not required in k-means clustering.
Read Distributed TensorFlow | TensorFlow Clustering
TensorFlow Interview Questions and Answers for Freshers. Q- 1,2,4,6,8,9
TensorFlow Interview Questions and Answers for Experience. Q- 3,5,7,10
Q11. What exactly Neural Networks are? What are the types of same you are familiar with?
Basically a connection of processing elements which can very large or very small depending on the application, it deployed for. These elements called neurons and generally, two types of networks can be seen in this category. They are Artificial Neural Networks and Biological Neural Networks. The use of artificial neural networks is more common and generally, they are considered for creating machines which are equally powerful to human brains.
Let's revise Recurrent Neural Network TensorFlow | LSTM Neural Network
Q12. How do you import Tensorflow?
import TensorFlow as tf
Q13. What are word embeddings used for and can they be used in TensorFlow?
Word embeddings usually use in Natural Language Processing as a representation of words and they can use in TensorFlow where it also call as word2vec.
Follow this link to learn more about word embedding.

Q14. Name the two models used in word embeddings?
The Continuous Bag of Words (CBOW) model and the skip-gram model
Q15. Write a code to start a simple session for the training?

```
with tf.Session() as sess:
```

Q16. Explain the following example.

```
for the epoch in      range(training_epochs):
                                                    for (x, y) in zip(train_X, train_Y):
                                                sess.run(optimizer, feed_dict={X: x, Y: y})
```

Here, the initializer is run and all the training data fit by running a loop for all the epochs

Q17. How do you see the charts and graphs for your model and what is the URL?

You can view the charts and graphs using TensorBoard by browsing to **https://localhost:6006** in your browser.

Q18. What is the confusion matrix?

A confusing matrix comprising of discrete values where each column contains a set of samples that estimated to be a keyword in your training model.

To learn more about the confusion matrix look at audio recognition using TensorFlow.

Q19. Describe the steps to configure a wide and deep model in TensorFlow?

- **Wide model features:** Choosing the base columns and crossed columns.
- **Deep model features:** Choosing the continuous columns, the dimension for each categorical column, and hidden layer sizes.

Combining these into a single model with ***DNNLinearCombinedClassifier***

Q20. Write a code to display the evaluated values while training your model in TensorFlow.

```
print('Results at epoch', (n + 1) * FLAGS.epochs_per_eval)
print('-' * 30)
for key in sorted(results):
print('%s: %s' % (key, results[key]))
```

TensorFlow Interview Questions and Answers for Freshers. Q- 11,13,14,17,18,19
TensorFlow Interview Questions and Answers for Experience. Q- 12,15,16,20

Q21. What are the imports needed for visualizing the Mandelbrot set in TensorFlow?

import PIL.Image
from io import BytesIO
from IPython.display import Image, display

Follow this link to learn more about Mandelbrot Set

Q22. How do you report a vulnerability in TensorFlow?

The reports about any security issues can send directly to security@tensorflow.org. The report to this email delivered to the security team at TensorFlow. The emails then acknowledged within 24 hours and detailed response is provided within a week along with the next steps.

Q23. What do you use for deploying a lite model file in TensorFlow?

- **Java API:** A wrapper around C++ API on Android.
- **C++ API:** It loads the TensorFlow Lite model and calls the interpreter.
- **Interpreter:** It can use to execute the model. It uses selective kernel loading which is a unique feature of TensorFlow Lite.

You can also implement custom kernels using the C++ API.

Let's discuss TensorFlow Mobile | TensorFlow Lite: A Learning Solution

Q24. What are placeholders in TensorFlow?

It is an assurity to the **TensorFlow** that an external value will be provided later.

Q25. What is tf.contrib.learn?

tf.contrib.learn is a TensorFlow library for simplifying the working of machine learning, and it includes:

- managing data sets
- managing feeding

Q26. What is input pipeline optimization?

The process flow of your model includes the loading of the image from the disk, converting it to a tensor followed by manipulating the tensor by cropping, padding and then making a batch. The process flow described above is called input pipeline.

Q27. List the two configurations needed to optimize CPU performance?

Intra_op_parallelism and iner_op_parallelism
Read TensorFlow Performance Optimization | Optimize GPU & CPU

Q28. How do you define a cluster in TensorFlow?
cluster = tf.train.ClusterSpec({"local": ["localhost:2222", "localhost:2223"]})

Q29. What is the MNIST dataset?
It is a dataset containing information of handwritten digits.
Follow this link for a deep understanding of MNIST Dataset.

Q30. What are the different dashboards in TensorFlow?
Below mentioned are different types of dashboards in TensorFlow:

1. 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.

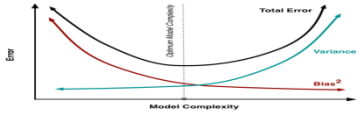
2. What is bias, variance trade off ?

Bias:
"Bias is error introduced in your model due to over simplification of machine learning algorithm." It can lead to 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.

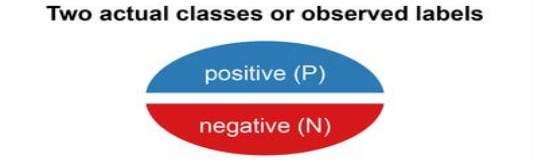
3. What is exploding gradients ?
"Exploding gradients are a problem where **large error gradients** accumulate and result in very large updates to neural network model weights during training." At an extreme, the values of weights can become so large as to overflow and result in NaN values.
This has 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.

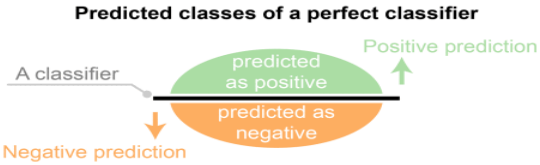
4. What is a confusion matrix ?
The confusion matrix is a 2X2 table that contains 4 outputs provided by the **binary classifier**. Various measures, such as error-rate, accuracy, specificity, sensitivity, precision and recall are derived from it. *Confusion Matrix*

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

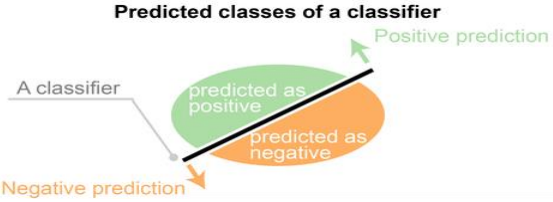
A dataset used for performance evaluation is called test dataset. It should contain the correct labels and predicted labels.



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

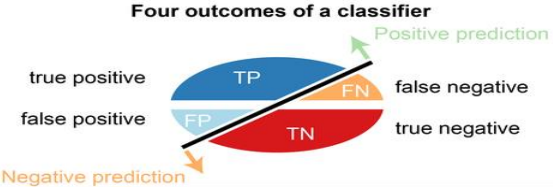


The predicted labels usually match with part of the observed labels in real world scenarios.



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

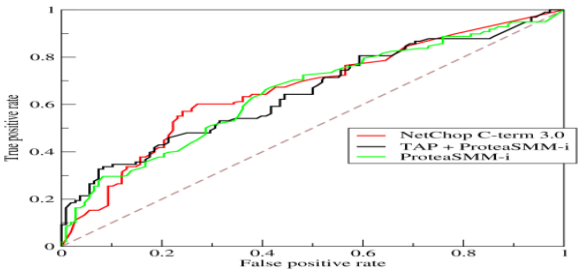


Basic measures derived from the confusion matrix

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.

6. Explain how a ROC curve works ?

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

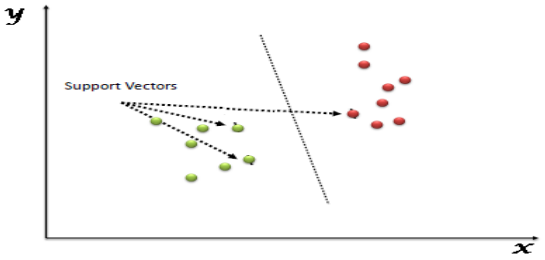


7. What is selection Bias ?

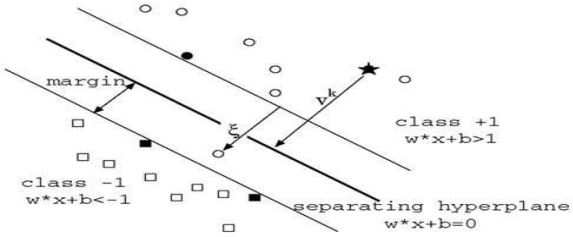
Selection bias occurs when sample obtained is not represantative of the population intended to be analyzed.

8. Explain SVM machine learning algorithm in detail.

SVM stands for support vector machine, it is a supervised machine learning algorithm which can be used for both **Regression and Classification**. If you have n features in your training 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.



9. What are support vectors in SVM.



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

10. What are the different kernels functions in SVM ?

There are four types of kernels in SVM.

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

11. Explain Decision Tree algorithm in detail.

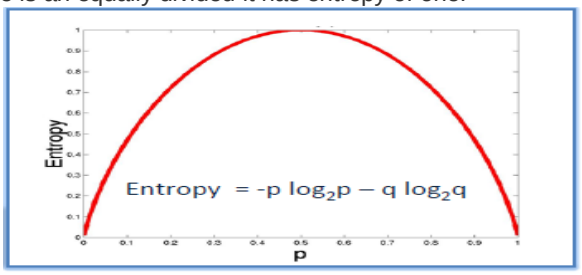
Decision tree is a supervised machine learning algorithm mainly used for the **Regression and Classification**.It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf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

12. 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

13. What is pruning in Decision Tree ?

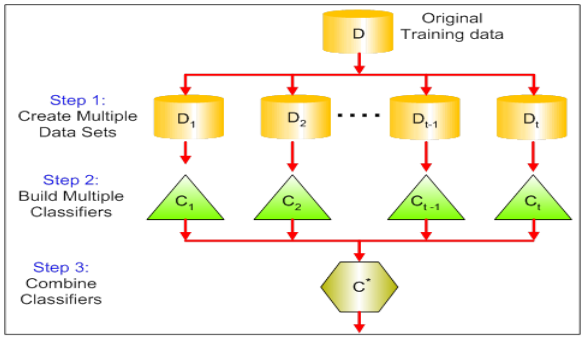
When we remove sub-nodes of a decision node, this procsss is called pruning or opposite process of splitting.

14. 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.

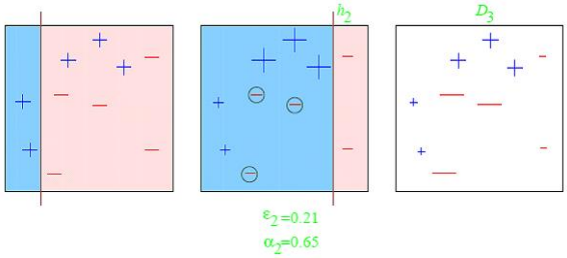
Bagging

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.



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

Random forest is a versatile machine learning method capable of performing both regression and classification tasks. It is also used for 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.

16. 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.

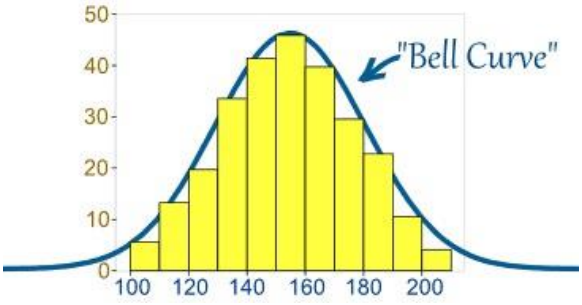
In case of time series data, you should use techniques like forward chaining – Where you will be model on past data then look at forward-facing data.

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]

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

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

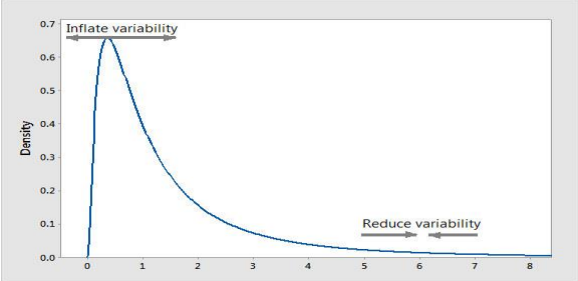
18. What do you understand by the term Normal Distribution?



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

19. What is a Box Cox Transformation?

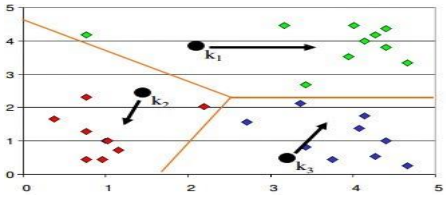
Dependent variable for a regression analysis might not satisfy one or more assumptions of an ordinary least squares regression. The residuals could either curve as the prediction increases or follow skewed distribution. In such scenarios, it is necessary to transform the response variable so that the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



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

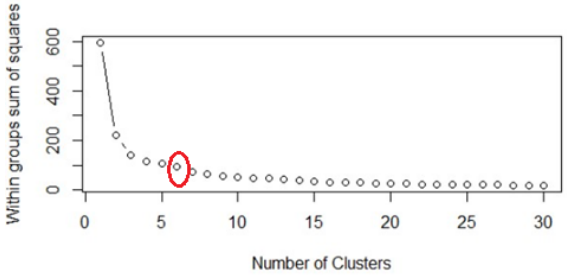
20. How will you define the number of clusters in a clustering algorithm?

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



different groups.

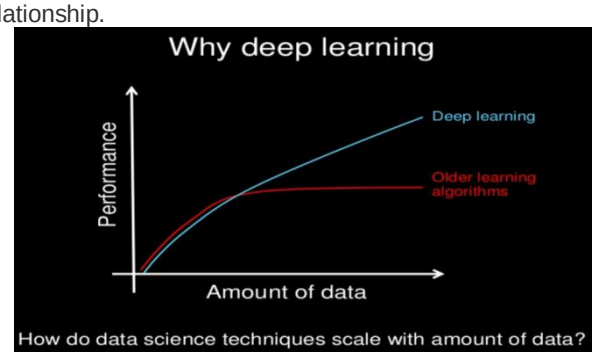
Within Sum of squares is generally used to explain the homogeneity within a cluster. If you plot WSS for a range of number of clusters, you will get the plot shown below. The Graph is generally known as Elbow Curve.



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

21. What is deep learning?

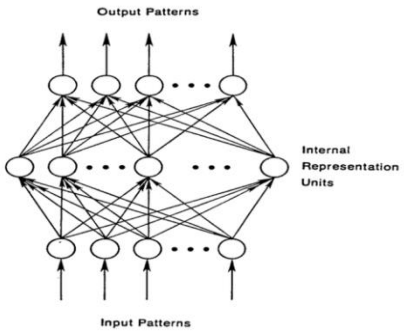
Deep learning is subfield of machine learning inspired by structure and function of brain called artificial neural network. We have a lot numbers of algorithms under machine learning like Linear regression, SVM, Neural network etc and deep learning is just an extention of Neural networks. In neural nets 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.



22. 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.

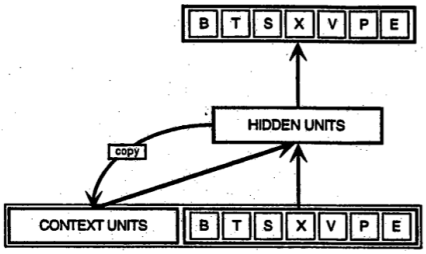


Recurrent networks on the other hand, take as their input not just the current input example they see, but also the what they have percieved previously in time. The BTSXPE at the bottom of the drawing represents the input example in the current moment, and CONTEXT UNIT represents the output of the previous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



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

Machine learning:

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

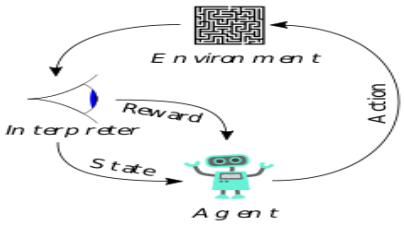
1. Supervised machine learning,
2. Unsupervised machine learning,
3. Reinforcement learning

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.

24. What is reinforcement learning ?

Reinforcement learning



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

25. 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.

26. 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.

27. What is TF/IDF vectorization ?

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

28. What are Recommender Systems?

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

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

Both Regression and classification machine learning techniques come under **Supervised machine learning algorithms**. In Supervised machine learning algorithm, we have to train the model using 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.

30. 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?

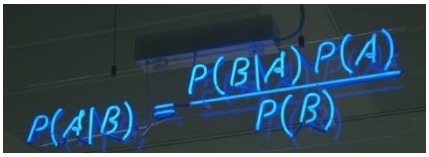
When you perform a hypothesis test in statistics, a p-value can help you determine the strength of your results. p-value is a number between 0 and 1. Based on the value it will denote the strength of the results. The claim which is on trial is called Null Hypothesis.

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.



What is Naive ?
The Algorithm is 'naive' because it makes assumptions that may or may not turn out to be correct.
Company 1 (20-25 minutes) – Global service based company
As in your CV, what have you done in building document exploration system ? I have implemented document clustering through topic modeling with LDA technique?
Let's say, you are given a scenario where you have terabytes of data files consisting of pdfs, text files, images, scanned pdfs etc. What approach will you take in understanding or classifying them ?
How will you read the content of scanned pdfs or written documents in image formats ?
Why is naive bayes called "naive" ?
Tell me about naive bayes classifier ?
What is deep learning ? Difference between machine learning & Deep learning ?
Experience: There were few more questions like this but it went terrible for me. I really could not get what his expectation was as he was not going technical at all where as I was trying to dig deep into technical stuff. I tried discussing about training a tesseract or language models but he does not seem to be convinced. May be he wanted to hear something out of box or may be just good explanations or may be some better approach. I did not get the difference between getting interviewed as fresher and experienced professional as I was taking it for the first time in last 5 years.
Company 2 (40-45 minutes) – Global service based company
How can you cluster documents in unsupervised way ?
How do you find the similar documents related to some query sentence/search ?
Explain TF-IDF ?
So As per my experience, Tf-Idf fails in document classification/clustering ? How can you improve further ?
What is LSTM neural network ? Explain me !! How does it work ?
What are word2vec vectors?
What do you mean by mutable and immutable objects in python ?
What are the data structures you have used in python ?
Experience: Though there was lot of discussion around document similarity problem, I was through the interview process. Again, there was not much interest in going to technical depth. Mostly the company had short automation projects in text analytics. I was offered the ML architect job position.
Company 3 – (40 minutes) Global product & service based
How do you handle multi-class classification with unbalanced dataset ?
How did you perform language identification from text sentence ? (As in my CV)
How does you represent the symbolic chinese or japanese alphabets here ?
How can I design a chatbot ? (I had little idea but I tried answering it with intent and response tf-idf based similarity)
Can I develop a chatbot with RNN providing a intent and response pair in input ?
Suppose I developed a chatbot with RNN/LSTMs on Reddit dataset. It gives me 10 probable responses ? How can I choose the best reply Or how can I eliminate others replies ?
How does SVM learns non-linear boundaries ? Explain.
Experience: There were few more questions which I do not remember now. This was the first technical interview where I had to go on explaining technical details which made me happy. I was offered job position in the company.
Company 4 – (50 minutes) – 1 year old Heath care start-up
What is precision and recall ? Which one of this do you think is important in medical diagnosis ?
Define precision and recall ?
How do you draw ROC curve ? What does area under ROC curve signify ?
How will you draw ROC for multi class classification problem ?
Tell me any other metric to measure multi-class classification result ?
What is sensitivity and specificity ?
What is random about Random Forest ?
How do you perform text classification ?
How can you make sure to learn a context !! Well its not possible with TF-IDF ? (I told him about taking n-grams say n = 1, 2, 3,

4 and concatenating tf-idf of them to make a long count vector ?
Okay that is the baseline people start with ? What can you do more with machine learning ? (I tried suggesting LSTM with word2vec or 1D-CNN with word2vec for classification but he wanted improvements in machine learning based methods :-|)
How does neural networks learns non-linear shapes when it is made of linear nodes ? What makes it learn non-linear boundaries ?
Experience: There were few more good questions which I do not remember now. Though the interview discussion was fantastic, in some of the questions we were not on same platform. Also during the interview I found out that being a startup there were only 2-3 people working in ML/DL/DS and showed the concern to the interviewer. I was not selected for the position here.
Company 5 – (50-55 minutes) – Amazon Inc.
What are the parameters in training a decision tree ?
What are the criteria for splitting at a node in decision trees ?
What is the formula of Gini index criteria?
What is the formula for Entropy criteria?
How is it decided that on which features it has to split ?
How do you calculate information gain mathematically ? Are you sure of formula!!
What is the advantage with random forest ?
Tell me about boosting algorithms ?
Okay !! So how does gradient boosting works ?
Do you know about Adaboost algorithm ? How and why does it work ?
What are the kernels used in SVM ? What is the optimization technique of SVM ?
How does SVM learns the hyperplane ? Talk more about mathematical details ?
Talking about unsupervised learning ? What are the algorithms ?
How do you decide K in K-Means clustering algorithm ?
Tell me at least 3 ways of deciding K in clustering ?
What other clustering algorithms do you know ?
Okay !! Can you tell DB-SCAN algorithm ?
How does HAC (Hierarchical Agglomerative clustering) work ?
Explain PCA ? Tell me the mathematical steps to implement PCA ?
What is disadvantage of using PCA ?
How does CNN work ? Explain the implementation details ?
Explain the back propagation steps of Convolution Neural Network ?
How do you deploy Machine Learning models ?
Lot of times, we may have to write ML models from scratch in C++ ? Will you be able to do that ?
Experience: This was Amazon interview for level 6 position**.** All I can say that it was mainly focused on algorithms and their mathematics behind them. Unfortunately, I was taking all interviews extempore and was not in a position to discuss all the mathematics in detailed way. Though I really enjoyed the interview discussing at least all the methodologies and mathematical explanations (which i remembered), the interviewer does not find me suitable for Level 6 position. I suppose one can easily crack initial technical round if you brush up the common machine learning algorithms in detail.
Company 6 – (50-55 minutes) – Global service based giant
What is the range of sigmoid function ?
Name the package of scikit-learn that implements logistic regression ?
What is mean and variance of standard normal distribution ?
What are the data structures you have used in python ?
Text classification method. How will you do it ?
Explain Tf-Idf ? What is the drawback of Tf-Idf ? How do you overcome it ?
What are bigrams & Tri-grams ? Explain with example of Tf-Idf of bi-grams & trigrams with a text sentence.
What is an application of word2vec ? Example.
How will you design a neural network ? How about making it very deep ? Very basic questions on neural network.?
How does LSTM work ? How can it remember the context ?
What is naive bayes classifier ?
What is the probability of heads coming 4 times when a coin is tossed 10 number of times ?
How do you get an index of an element of a list in python ?
How do you merge two data-set with pandas ?
From user behavior, you need to model fraudulent activity. How are you going to solve this ? May be anomaly detection problem

or a classification problem !!
What will you prefer a decision tree or a random forest ?
How is using a logistic regression different from using a random forest ?
Will you use decision tree or random forest for a classification problem ? What is advantage of using random forest ?
Experience: Well I was offered the data scientist job position. In-fact I got a very good feedback. It was a good technical discussion which I enjoyed. Also, you may feel that these questions are basic in ML/DS field. I do not want to demean but somewhere I felt that interviewer may not have worked much in the domain or may not be aware of current stuff going in the field.
Company 7 (25-30 minutes) – Global business process management company
Which model would you use in case of unbalanced dataset: Random Forest or Boosting ? Why ?
What are the boosting techniques you know ?
Which model would you like to choose if you have many classes in a supervised learning problem ? Say 40-50 classes !!
How do you perform ensemble technique?
How does SVM work ?
What is Kernel ? Explain a few.
How do you perform non-linear regression ?
What are Lasso and Ridge regression ?
Experience: Frankly speaking, the technical interview was little vague (personal opinion) as I do not find the conversation serious enough. But questions were good. The job position was for a senior position where I would be leading 15-16 people team. It was followed by a client interview and HR interview. I was offered the consultant job position with good compensation.
Company 8 (60 minutes) – 4 years old product & services company
As per CV, you have done speaker identification in speech. What approach have you followed ?
What are MFCCs ?
What is Gaussian Mixture model ? How does it perform clustering ?
How is Expectation Maximization performed ? Explain both the steps ?
How is likelihood calculated in GMM ?
How did you perform MAP adaptation for GMM-UBM technique of speaker identification ?
Tell me about I-vector technique you implemented ?
Okay !! What is factor analysis in this context ?
What was the difference between JFA and I-vector ? Why choose I-vectors over JFA
Have you implemented PLDA I-vector technique ?
Have you read Deep Speaker paper from Baidu ?
How do you select between 2 models (Model Selection techniques)?
Okay great BIC & AIC !! How does it work mathematically ?
Explain the intuition behind BIC or AIC ?
How do you handle missing data or NANs in your MFCC feature vector matrix ?
How did you perform language identification ? What were the feature ?
How did you model classifiers like speech vs music and speech vs non-speech ?
How can deep neural network be applied in these speech analytics applications ?
Experience: Yeah

1. What is the difference between inductive machine learning and deductive machine learning ? (machine learning interview questions and answers)
Answer:
In inductive machine learning, the model learns by examples from a set of observed instances to draw a generalized conclusion whereas in deductive learning the model first draws the conclusion and then the conclusion is drawn. Let's understand this with an example, for instance, if you have to explain to a kid that playing with fire can cause burns. There are two ways you can explain this to kids, you can show them training examples of various fire accidents or images with burnt people and label them as "Hazardous". In this case the kid will learn with the help of examples and not play with fire. This is referred to as Inductive machine learning. The other way is to let your kid play with fire and wait to see what happens. If the kid gets a burn they will learn not to play with fire and whenever they come across fire, they will avoid going near it. This is referred to as deductive learning.
(machine learning interview questions and answers)
2. Why is Harmonic mean used to calculate F1 score and not arithmetic mean ?

Answer:
Because the Harmonic mean gives more weightage to lower values. Thus, we will only get high F1 score if both Precision and Recall are higher.
3. Does 100% precision mean that our model predicts all the values correctly ?
Answer:
No. We can get perfect precision in many ways, but it doesn't mean that our model predicts every value accurately. For ex. if we make one single positive prediction and make sure it is correct, our precision reaches 100%. Generally, precision is used with other metrics (recall) to measure the performance.
4. What are the similarities & difference between machine learning and human learning ?
Answer:
Machine learning and human learning actually quite similar. Machine learning is about an algorithm or computer. Actually engaging with its environment with data and adopting a coding too on the things that it learns. Let us suppose a program fails to make the right predictions & it will balance itself in some sense. In order to make better predictions next time. Now, That is very similar the way human learns. Human is actually engaging with its environment & learning from it. So, Machine learning has an aspect of kind of an evolutionary aspect to it. Which I think quite new to the area of this artificial intelligence.
5. How to decide one problem is a machine learning problem or not ?
Answer:
When you are analyzing a problem. If that problem consisting patterns and that pattern we can't extract from mathematical equations. If you found such kind of problem then we need to use machine learning to extract those pattern by using lots of data. These above key features are helpful to predict whether the problem is a machine learning problem or not.
Example:– We need to find whether the number is even or odd. This example seems very simple. Yes, this problem is very simple because we know the logic to find whether a number is odd or even and we also know about the mathematics behind this problem. Let us suppose the number is divided by 2. Then the remainder is 1 then we call that number is odd. Whether the remainder is 0 then we call that number as an even. So, this problem has some pattern but we can solve through mathematical equations. do not need a lot of data also. So, definitely, this is not a machine learning problem. Till now this is not a machine learning problem but if you want to make it as a machine learning problem. We can do one thing. We can feed lots of data as an individual number by telling the number is odd the number is even. The machine will classify whether the number is even or odd. But as we know logic and mathematics behind this problem this problem can't come under as a machine learning problem.
Example1:- Let's say you have a lot of photos or images. We need to find whether particular photo contains human face or not. Here there is a pattern that we need to find human across all the photos. Can we solve this problem through mathematical equations? So, it's very difficult. SImply we can take this lot of data and we can feed this data to the algorithm as a training data. Which means training the machine using this data. After this training, we will get some mathematical equations based on the patterns that we got from training. But humans can't write this logic as their own. So Definitely, this is a machine learning problem. Here machine will automatically form a rule based on training data. That rule is nothing but to detect whether the photo contains human face or not.
6. What is AdaGrad algorithm in machine learning ?
Answer:
A sophisticated gradient descent algorithm that rescales the gradients of each parameter, effectively giving each parameter an independent learning rate.
7. What is a binary classification in machine learning ?
Answer:
A type of classification task that outputs one of two mutually exclusive classes. For example, a machine learning model that evaluates email messages and outputs either "spam" or "not spam" is a binary classifier.
8. What is Rectified Linear Unit (ReLU) in Machine learning ?
Answer:
An activation function with the following rules:
If input is negative or zero, output is 0.
If input is positive, output is equal to input.
machine learning interview questions and answers
9. Explain what is the function of 'Unsupervised Learning ?
Answer:
Find clusters of the data
Find low-dimensional representations of the data
Find interesting directions in data
Interesting coordinates and correlations
Find novel observations/ database cleaning
10. Why over fitting happens ?

Answer:
The possibility of overfitting happens as the criteria used for training the model is not the same as the criteria used to judge the efficiency of a model.
11. What is a recommendation system?
Answer:
Anyone who has used Spotify or shopped at Amazon will recognize a recommendation system: It's an information filtering system that predicts what a user might want to hear or see based on choice patterns provided by the user.
12. Explain what is precision and Recall?
Answer:
Recall:
It is known as a true positive rate. The number of positives that your model has claimed compared to the actual defined number of positives available throughout the data.
Precision:
It is also known as a positive predicted value. This is more based on the prediction. It is a measure of a number of accurate positives that the model claims when compared to the number of positives it actually claims.
13. Pick an algorithm and write a Pseudo code for the same ?
Answer:
This question depicts your understanding of the algorithm. This is something that one has to be very creative and also should have in-depth knowledge about the algorithms and first and foremost the individual should have a good understanding of the algorithms. Best way to answer this question would be start off with Web Sequence Diagrams.
14. List out some important methods of reducing dimensionality?
Answer:
Combine features with feature engineering.
Use some form of algorithmic dimensionality reduction like ICA or PCA.
Remove collinear features to reduce dimensionality.
15. Explain the Bias-Variance Tradeoff ?
Answer:
Predictive models have a tradeoff between bias (how well the model fits the data) and variance (how much the model changes based on changes in the inputs).
Simpler models are stable (low variance) but they don't get close to the truth (high bias).
More complex models are more prone to being overfit (high variance) but they are expressive enough to get close to the truth (low bias).
The best model for a given problem usually lies somewhere in the middle.
16. When is Ridge regression favorable over Lasso regression ?
Answer:
You can quote ISLR's authors Hastie, Tibshirani who asserted that, in presence of few variables with medium / large sized effect, use lasso regression. In presence of many variables with small / medium sized effect, use ridge regression.
Conceptually, we can say, lasso regression (L1) does both variable selection and parameter shrinkage, whereas Ridge regression only does parameter shrinkage and end up including all the coefficients in the model. In presence of correlated variables, ridge regression might be the preferred choice. Also, ridge regression works best in situations where the least square estimates have higher variance. Therefore, it depends on our model objective.
machine learning interview questions and answers
17. What is convex hull ?
Answer:
In case of linearly separable data, convex hull represents the outer boundaries of the two group of data points. Once convex hull is created, we get maximum margin hyperplane (MMH) as a perpendicular bisector between two convex hulls. MMH is the line which attempts to create greatest separation between two groups.
18. What's the difference between a generative and discriminative model ?
Answer:
A generative model will learn categories of data while a discriminative model will simply learn the distinction between different categories of data. Discriminative models will generally outperform generative models on classification tasks.
19. What is your training in machine learning and what types of hands-on experience do you have ?
Answer:
Your answer to this question will depend on your training in machine learning. Be sure to emphasize any direct projects you've completed as part of your education. Don't fail to mention any additional experience that you have including certifications and how they have prepared you for your role in the machine learning field.
20. How do bias and variance play out in machine learning ?
Answer:
Both bias and variance are errors. Bias is an error due to flawed assumptions in the learning algorithm. Variance is an error resulting from too much complexity in the learning algorithm.
21. What is supervised versus unsupervised learning ?

Answer:
Supervised learning is a process of machine learning in which outputs are fed back into a computer for the software to learn from for more accurate results the next time. With supervised learning, the "machine" receives initial training to start. In contrast, unsupervised learning means a computer will learn without initial training.
22. How will you know which machine learning algorithm to choose for your classification problem ?
Answer:
If accuracy is a major concern for you when deciding on a machine learning algorithm then the best way to go about it is test a couple of different ones (by trying different parameters within each algorithm) and choose the best one by cross-validation. A general rule of thumb to choose a good enough machine learning algorithm for your classification problem is based on how large your training set is. If the training set is small then using low variance/high bias classifiers like Naïve Bayes is advantageous over high variance/low bias classifiers like k-nearest neighbour algorithms as it might overfit the model. High variance/low bias classifiers tend to win when the training set grows in size.
23. Logistic regression gives probabilities as result then how do we use it to predict a binary outcome ?
Answer:
A logistic model outputs a value between 0 and 1. To convert these probabilities into classes we use decision boundaries. We can set equal or unequal boundaries depending upon the requirement.
24. What are some common unsupervised tasks other than clustering ?
Answer:
Visualization, Dimensionality reduction, association rule learning.
machine learning interview questions and answers
25. What is the difference between A.I. and machine learning, and has A.I. been oversold for decades because of sci-fi ?
Answer:
More People thought of that A.I. that mean artificial intelligence may be than machine learning. Artificial intelligence actually it's like which we go to see that alan turing aim was to somehow make a machine have the sort of intelligence that human might have. In particularly a program actually to convince you that it is human if you chat it with it. But i think artificial intelligence has evolved since then to make a unique sort of intelligence that machine might have. Machine learning has slightly a different quality. It is like a more specific part of artificial intelligence. Which is the idea of the program is going to change a world through a coding that makes interactions as same like humans. By the end, the program might not know actually how the program is written. Because it's been changing as it's been interacting. So, might be when we look at the program & see the actual program we do not know why necessarily it decided to write these decisions in a particular way. Because there are a lot of connections between artificial intelligence have with machine learning.
26. Differentiate between inductive and deductive machine learning ?
Answer:
In inductive machine learning, the model learns through examples obtained from a set of observed instances to draw generalized conclusions, whereas in deductive machine learning certain statements are combined in a logical order as per some predefined rules to obtain new statements. Basically, inductive learning is instruction based and deductive learning is experience based. (company)
27. What is the "Curse of Dimensionality ?
Answer:
The term, "Curse of Dimensionality" refers to the difficulty of searching through a space with multiple dimensions; more the dimensions, more the difficulty. If talk of this term, particularly in context of machine learning, it has to do with the difficulty associated with non-intuitive properties of data observed when working in a high-dimensional space.
28. Can you name some popular machine learning algorithms ?
Answer:
Yes, they are:
Nearest Neighbour
Neural Networks
Decision Trees
Support vector machines
29. What do you understand by decision tree classification ?
Answer:
Decision tree classification in machine learning refers to a tree-like classification model where the data is continuously split as per certain parameters. There are two primary entities in this model, namely decision nodes and leaves. The leaves denote the final outcome of decisions, while the nodes signify the point where the data is split. A decision tree classification greatly facilitates a visual and explicit representation of the decisions and decision making process.
30. What is the difference between supervised and unsupervised learning ?
Answer:
In the supervised learning process, outputs are fed back into a computer system so that the software can learn from it

and produce more accurate results in the successive occurrences; it is a kind of initial training for a system. On the other hand, unsupervised learning is a machine learning algorithm that draws inferences on its own from the unlabeled data set, without any external aid or input.

31. What is activation function in Machine Learning ?

Answer:

A function (for example, ReLU or sigmoid) that takes in the weighted sum of all of the inputs from the previous layer and then generates and passes an output value (typically nonlinear) to the next layer.

32. What is baseline in machine learning ?

Answer:

A simple model or heuristic used as reference point for comparing how well a model is performing. A baseline helps model developers quantify the minimal, expected performance on a particular problem.

machine learning interview questions and answers

33. What is batch in machine learning ?

Answer:

The set of examples used in one iteration (that is, one gradient update) of model training.

34. What is bias in machine learning ?

Answer:

An intercept or offset from an origin. Bias (also known as the bias term) is referred to as b or w0 in machine learning models.

35. What is calibration layer in machine learning ?

Answer:

A post-prediction adjustment, typically to account for prediction bias. The adjusted predictions and probabilities should match the distribution of an observed set of labels.

36. What is class-imbalanced data set in machine learning ?

Answer:

A binary classification problem in which the labels for the two classes have significantly different frequencies. For example, a disease data set in which 0.0001 of examples have positive labels and 0.9999 have negative labels is a class-imbalanced problem, but a football game predictor in which 0.51 of examples label one team winning and 0.49 label the other team winning is not a class-imbalanced problem. (Machine Learning Training)

37. What is confusion matrix in machine learning ?

Answer:

An NxN table that summarizes how successful a classification model's predictions were; that is, the correlation between the label and the model's classification. One axis of a confusion matrix is the label that the model predicted, and the other axis is the actual label. N represents the number of classes.

38. How do you choose an algorithm for a classification problem ?

Answer:

The answer depends on the degree of accuracy needed and the size of the training set. If you have a small training set, you can use a low variance/high bias classifier. If your training set is large, you will want to choose a high variance/low bias classifier.

39. What is 'Overfitting' in Machine learning ?

Answer:

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

40. What is the difference between Type 1 and Type 2 errors ?

Answer:

Type 1 error is classified as a false positive. I.e. This error claims that something has happened but the fact is nothing has happened. It is like a false fire alarm. The alarm rings but there is no fire.

Type 2 error is classified as a false negative. I.e. This error claims that nothing has happened but the fact is that actually, something happened at the instance.

The best way to differentiate a type 1 vs type 2 error is:

Calling a man to be pregnant- This is Type 1 example

Calling pregnant women and telling that she isn't carrying any baby- This is type 2 example

machine learning interview questions and answers

41. What are parametric models? Give an example ?

Answer:

Parametric models are those with a finite number of parameters. To predict new data, you only need to know the parameters of the model. Examples include linear regression, logistic regression, and linear SVMs.

Non-parametric models are those with an unbounded number of parameters, allowing for more flexibility. To predict new data, you need to know the parameters of the model and the state of the data that has been observed. Examples include decision trees, k-nearest neighbors, and topic models using latent dirichlet analysis.

42. What is the difference between covariance and correlation ?
Answer:
Correlation is the standardized form of covariance.
Covariances are difficult to compare. For example: if we calculate the covariances of salary ($) and age (years), we'll get different covariances which can't be compared because of having unequal scales. To combat such situation, we calculate correlation to get a value between -1 and 1, irrespective of their respective scale.
43. How would you evaluate a logistic regression model ?
Answer:
A subsection of the question above. You have to demonstrate an understanding of what the typical goals of a logistic regression are (classification, prediction etc.) and bring up a few examples and use cases.
44. How will you explain machine learning in to a layperson ?
Answer:
Machine learning is all about making decisions based on previous experience with a task with the intent of improving its performance. There are multiple examples that can be given to explain machine learning to a layperson –
Imagine a curious kid who sticks his palm
You have observed from your connections that obese people often tend to get heart diseases thus you make the decision that you will try to remain thin otherwise you might suffer from a heart disease. You have observed a ton of data and come up with a general rule of classification.
You are playing blackjack and based on the sequence of cards you see, you decide whether to hit or to stay. In this case based on the previous information you have and by looking at what happens, you make a decision quickly.
45. What is decision tree classification ?
Answer:
A decision tree builds classification (or regression) models as a tree structure, with datasets broken up into ever smaller subsets while developing the decision tree, literally in a tree-like way with branches and nodes. Decision trees can handle both categorical and numerical data.
46. What are some methods of reducing dimensionality ?
Answer:
Deductive machine learning starts with a conclusion, then learns by deducing what is right or wrong about that conclusion. Inductive machine learning starts with examples from which to draw conclusions.
48. What is kernel SVM ?
Answer:
Kernel SVM is the abbreviated version of kernel support vector machine. Kernel methods are a class of algorithms for pattern analysis and the most common one is the kernel SVM.
49. How do we separate one dimensional, two dimensional and three-dimensional data ?
Answer:
One dimensional can be separated using a point, two dimensional using a line and three dimensional can be separated by a hyperplane.
50. What kind problems are solved by regularization ?
Answer:
In machine learning, regularization is basically a process of introducing additional information with the purpose of solving an ill-posed problem or to avoid over fitting. It is basically a form of regression, which regularizes or constrains the coefficient estimates to zero. The technique of regularization prevents learning a more complex or flexible model in order to avoid the over fitting risk.

What is deep learning?

Deep learning is an area of machine learning focus on using deep (containing more than one hidden layer) artificial neural networks, which are loosely inspired by the brain. The idea dates back to the mid 1960s, Alexey Grigorevich Ivakhnenko published the first general, working deep learning network. Deep learning is applicable over a range of fields such as computer vision, speech recognition, natural language processing.

Question 2

Why are deep networks better than shallow ones?

Both shallow and deep networks are capable of approximating any function. For the same level of accuracy, deeper networks can be much more efficient in terms of computation and number of parameters. Deeper networks are able to create deep representations, at every layer, the network learns a new, more abstract representation of the input.

Question 3

What is a cost function?

Cost function tells us how well the neural network is performing. Our goal during training is to find parameters that minimize the cost function. For an example of a cost function, consider Mean Squared Error function:
$$MSE=\frac{1}{n} \sum_{i=0}^n(\hat Y_i – Y_i)^2$$
The mean of square differences between our prediction \(\hat Y\) and desired value \(Y\) is what we want to minimize.

Question 4

What is a gradient descent?

Gradient descent is an optimization algorithm used in machine learning to learn values of parameters that minimize the cost function. It's an iterative algorithm, in every iteration, we compute the gradient of the cost function with respect to each parameter and update the parameters of the function via the following.

$$\Theta := \Theta\,\, – \alpha\frac{d}{\partial\Theta}J(\Theta)$$

\(\Theta\) – is the parameter vector, \(\alpha\) – learning rate, \(J(\Theta)\) – is a cost function

Take a look at the linear regression simulator to see how gradient descent is used to perform linear regression.

Question 5

What is a backpropagation?

Backpropagation is a training algorithm used for a multilayer neural networks. It moves the error information from the end of the network to all the weights inside the network and thus allows for efficient computation of the gradient.

The backpropagation algorithm can be divided into several steps:

1. Forward propagation of training data through the network in order to generate output.
2. Use target value and output value to compute error derivative with respect to output activations.
3. Backpropagate to compute the derivative of the error with respect to output activations in the previous layer and continue for all hidden layers.
4. Use the previously calculated derivatives for output and all hidden layers to calculate the error derivative with respect to weights.
5. Update the weights.

The neural network simulator will guide you step-by-step through backpropagation.

Question 6

Explain the following three variants of gradient descent: batch, stochastic and mini-batch?

Stochastic Gradient Descent

Uses only single training example to calculate the gradient and update parameters.

Batch Gradient Descent

Calculate the gradients for the whole dataset and perform just one update at each iteration.

Mini-batch Gradient Descent

Mini-batch gradient is a variation of stochastic gradient descent where instead of single training example, mini-batch of samples is used. It's one of the most popular optimization algorithms.

Question 7

What are the benefits of mini-batch gradient descent?

- Computationally efficient compared to stochastic gradient descent.
- Improve generalization by finding flat minima.
- Improving convergence, by using mini-batches we approximating the gradient of the entire training set, which might help to avoid local minima.

Question 12

What is data normalization and why do we need it?

Data normalization is very important preprocessing step, used to rescale values to fit in a specific range to assure better convergence during backpropagation. In general, it boils down to subtracting the mean of each data point and dividing by its standard deviation.

Question 13

Weight initialization in neural networks?

Weight initialization is a very important step. Bad weight initialization can prevent a network from learning. Good initialization can lead to quicker convergence and better overall error. Biases can be generally initialized to zero. The general rule for setting the weights is to be close to zero without being too small.

Question 14

Why is zero initialization not a recommended weight initialization technique?

As a result of setting weights in the network to zero, all the neurons at each layer are producing the same output and the same gradients during backpropagation.

The network can't learn at all because there is no source of asymmetry between neurons. That is why we need to add randomness to weight initialization process.

Question 15

What is the role of the activation function?

The goal of an activation function is to introduce non-linearity into the neural network so that it can learn more complex function. Without it, the neural network would be only able to learn function which is a linear combination of its input data.

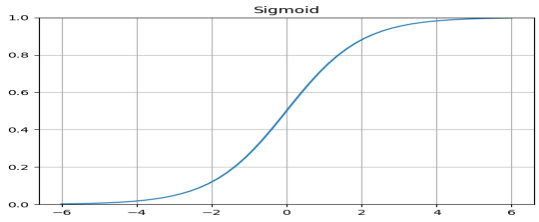
Question 16

Provide some examples of activation functions?

Sigmoid

The sigmoid function also known as the logistic function is used for binary classification, it's continuous and it has an easily calculated derivative. It squashes real numbers to range between [0,*1*].

$$ f(x)= \frac{1}{(1 + e^{-x})}$$



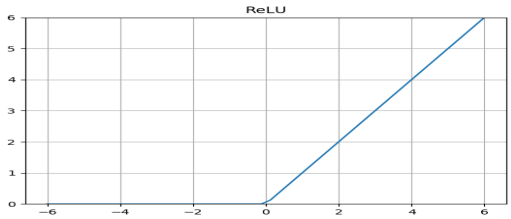
Softmax

Softmax is a generalization of the sigmoid function to the case where we want to handle multiple classes. All output values are in the range (0, 1) and sum up to 1.0 and therefore can be interpreted as probabilities that our input belongs to one of a set of output classes.

Rectified linear units – ReLU

ReLU outputs 0 if the input is less or equal to 0 and raw output otherwise, we can think of them as switches. Biologically inspired, enable training much deeper nets by backpropagation. It does not suffer from the vanishing gradient problem and it's very fast. It has been used in convolutional networks more effectively than the widely used logistic function.

$$ f(x)= max(0, x)$$



Question 17

What are hyperparameters, provide some examples?

Hyperparameters as opposed to model parameters can't be learn from the data, they are set before training phase.

Learning rate

It determines how fast we want to update the weights during optimization, if learning rate is too small, gradient descent can be slow to find the minimum and if it's too large gradient descent may not converge(it can overshoot the minima). It's considered to be the most important hyperparameter.

Number of epochs

Epoch is defined as one forward pass and one backward pass of all training data.

Batch size

The number of training examples in one forward/backward pass.

Question 18

What is a model capacity?

Ability to approximate any given function. The higher model capacity is the larger amount of information that can be stored in the network.

Question 19

What is a convolutional neural network?

Convolutional neural networks, also known as CNN, are a type of feedforward neural networks that use convolution in at least one of their layers. The convolutional layer consists of a set of filter (kernels). This filter is sliding across the entire input image, computing dot product between the weights of the filter and the input image. As a result of training, the network learns filters that can detect specific features.

Question 20

What is an autoencoder?

Autoencoder is artificial neural networks able to learn representation for a set of data (encoding), without any supervision. The network learns by copying its input to the output, typically internal representation has smaller dimensions than input vector so that they can learn efficient ways of representing data. Autoencoder consist of two parts, an encoder tries to fit the inputs to an internal representation and decoder converts internal state to the outputs.

Question 21

What is a dropout?

Dropout is a regularization technique for reducing overfitting in neural networks. At each training step we randomly drop out (set to zero) set of nodes, thus we create a different model for each training case, all of these models share weights. It's a form of model averaging.

Question 22

How we define the cross-entropy cost function?

Cross-entropy cost function is used for classification, it's a natural choice if there is a sigmoid or softmax nonlinearity in the output layer.

$$C=- \frac{1}{n} \sum_{i=1}^n(y_i\ln a_i + (1-y_i)\ln (1-a_i))$$

The a – represents the output of the neural network, y – target value, n – is the total number of training examples.

Question 23

What are the differences between feedforward neural network and a recurrent neural network?

Feedforward network allows signals to travel one way only, from input to output. A recurrent neural network is a special network, which has unlike feedforward networks, recurrent connections. The RNN can be described using this recurrent formula:

$$s_t = f(s_{t-1}, x_t)$$

The state \(s_{t}\) at a time 't' is a function of previous state \(s_{t-1}\) and the input \(x_{t}\) at the current time step, recurrent neural network maintains an internal state \(s_{t}\), by using their own output as a part of the input for next time step. This state vector summarizes the history of the sequence it has seen so far. Recurrent neural networks are Turing complete, can simulate arbitrary programs. Whereas feedforward network can just compute one fixed-size input to one fixed-size output, the RNN can handle sequential data of arbitrary length.

Question 24

What are some limitations of deep learning?

- Deep learning usually requires large amounts of training data.
- Deep neural networks are easily fooled.
- Successes of deep learning are purely empirical, deep learning algorithms have been criticized as uninterpretable "black-boxes".
- Deep learning thus far has not been well integrated with prior knowledge.

Please leave your comments, suggestion, feedback.

Share this:

Statistics

The foundation of machine learning and analytics, is statistics. So having a basic understanding of the same, is very important. While you're not expected to be a statistician, knowing the fundamentals will ensure the interviewers that you're not just plugging data into models and actually understand what you are doing.

Question 1: What is linear regression?

A linear approach to modelling the relationship between a scalar response and one or more variables, is known as linear regression. When it's the case of one explanatory variable in the equation, it is called simple linear regression.

- LASSO and Ridge Regression by Experfy

Question 2: What is interpolation and extrapolation?

Extrapolation refers to an estimation of a value, based on extending a known sequence of facts beyond the area for which they are known. Interpolation, is the estimation of a value within two known values in a sequence of values.

- What is interpolation and extrapolation by Mario's Math Tutoring

Question 3: What is the difference between univariate, bivariate, and multivariate analysis?

Bivariate and multivariate, are terms that are used to describe how many variables are being analysed at the moment. While multivariate means more than two variables, bivariate refers to two and univariate means that only one variable is under examination.

- Classification Models by Experfy
- The difference between bivariate and multivariate analysis by Sciencing

Question 4: What does p-value signify about the statistical data?

A small p-value signifies evidence against the null hypothesis and is typically less than or equal to 0.05. A large p-value is greater than 0.05 and fails to reject the null hypothesis during data analysis.

- What a p-value tells you about statistical data by dummies
- Probability and statistics for data science with R by Experfy

Question 5: What is the difference between Type I and Type II error?

A type I error is also known as false positive finding and refers to the rejection of a true null hypothesis. The type II error is also called a false negative finding, retaining a false null hypothesis during an analysis.

- The difference between Type I and Type II errors by ThoughtCo

Question 6: How do you deal with outliers?
An outlier refers to any data point that is distinctly different from the rest of your data points. In case you have outlier records in your data set, you can choose to either cap your data, assign a new value to it, try a transformation or completely remove it from your data set.

- Data on the edge: Handling outliers by Veera
- Data pre-processing by Experfy

Question 7: How do you handle missing data?
There are several ways to handle missing data. The easiest way(not necessarily the best way) is to just plainly get rid of all missing way. Another way is to replace the missing value with the mean(average) of the time series. A more sophisticated way is to impute the missing values using various statistical and machine learning techniques.

- Working with missing data in machine learning
- Data pre-processing course by Experfy

Question 8: What is nonparametric testing?
Nonparametric tests are sometimes called distribution-free tests because they are based on fewer assumptions. They don't assume the underlying distribution is normal. You use nonparametric tests when your data is not normal.

Question 9: Describe the central limit theorem.
The central limit theorem just says that with a large sample size, sample means are normally distributed. A sample mean is the average of a random subset of a larger group. So if you randomly picked 10 people out of 100 and recorded their weights, the average of those 10 weights would be the sample mean. You could do this many times and, since it is a random selection, the sample mean would be different each time.
The CLT make no assumptions about the distribution of your underlying data. The distribution of people's weights does not need to be normally distributed in order to know that the sample means of the weights are normally distributed.

- Ingredients in the making of a data scientist

Question 10: Alice has 2 kids and one of them is a girl. What is the probability that the other child is also a girl?
You can assume that there is an equal number of males and females in the world.
The outcomes for two kids can be: {BB, BG, GB, GG}
Since it is mentioned that one of them is a girl, we can remove the BB option from the sample space. Therefore the sample space has 3 options while only one fits the second condition. Therefore the probability the second child will be a girl too is 1/3.

Machine Learning & Theory
In this category, employers want to make sure you can explain the basic concepts behind popular machine learning algorithms and models.
Nowadays, machine learning algorithms are simply just library calls from scikit-learn(if you are using Python) or various packages (if you are using R). So using the machine learning algorithm is just several short lines of code.
However, do you understand the library functions you are calling?
These are just a sample of the various questions that might be asked.

Question 1: What's the difference between Supervised and Unsupervised Learning?
In supervised learning, the machine learning algorithm learns a function that maps an input to an output based on examples of input-output pairs. Examples of supervised learning include regression, neural networks, random forest, deep learning, etc.
In unsupervised learning, we give the machine learning algorithms data and it infers structure from the data. Examples of unsupervised learning are the various classification algorithms where it finds groups in unlabeled data.

- Machine learning foundations: Supervised Learning by Experfy

Question 2: Describe PCA(Principal Component Analysis).
PCA is a dimensionality reduction technique. Let's say we have a data set with a higher number of dimensions (n dimensions). We select k features (also called variables/factors) among a larger set of n features, with k much smaller than n.
This smaller set of k features created using PCA is the best subset of k features(in that it minimizes the variance of the residual noise when fitting data to a linear model). Note that PCA transforms the initial features into new ones, that are linear combinations of the original features.

- Unsupervised learning: Dimensionality reduction and representation by Experfy

Question 3: Why is naive Bayes so 'naive' ?
Naive Bayes is called 'naive' because it assumes that all of the features in a dataset are statistically independent. This is rarely the case in real life since there's always some kind of correlation between the features. Having said that, Naives Bayesian algorithm is a surprisingly effective machine learning algorithm for a number of use cases.

Question 4: Discuss bias and variance tradeoffs. Give examples of ML algorithms that have low/high bias and low/high variance.
The bias-variance tradeoff is the central problem in supervised machine learning. We want to choose a model that both accurately captures the regularities in the training data but also be able to generalize to unseen data. High bias models underfit the data by missing relevant relationships between features and the target outputs. High variance model overfits the training data by being too sensitive to all the minutiae and fluctuations in the training data.
The best model has low bias and low variance. But there's usually a tradeoff.

- Parametric or linear machine learning algorithms often have a high bias but a low variance.

- Non-parametric or non-linear machine learning algorithms often have a low bias but a high variance.
- Examples of low-variance machine learning algorithms include Linear Regression, Linear Discriminant Analysis, and Logistic Regression.
- Examples of high-variance machine learning algorithms include Decision Trees, k-Nearest Neighbors and Support Vector Machines.

Question 5: How do you handle imbalanced dataset?

Certain data science problems have highly imbalanced data. For example, in the case of fraud detection, the fraud cases might be significantly less than 1% of the sample. Any machine learning models will overlearn the non-fraudulent cases and not be able to pick up the fraudulent cases. So, how do we fix this problem?

There are several approaches:

1. Oversample: Oversample the minority class(fraud cases) by increasing the quantity of the rare class so that it becomes more representative. There are several statistical methods used such as bootstrapping or SMOTE(Synthetic Minority Oversampling Technique).
2. Undersample: Undersample the majority class so that it can be balanced with the minority class.
3. Both oversample/undersample: In some cases, we likely do both. Oversample the minority class and undersample the majority class.

Question 6: Name several clustering algorithms.

- K- mean clustering
- Agglomerative Hierarchical Clustering
- Gaussian Mixture Models using Expectation Maximization
- Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

Question 7: Describe one method to select the optimal number of clusters in k-means?

One method to pick the optimal number of clusters is the "elbow" method. The basic idea is to run k-means clustering through the dataset for different values of k(e.g. 1 to 10). For each value of k, calculate the sum of squared errors(SSE).We then plot a line chart of the SSE for each value of *k*. If the line chart looks like an arm, then the "elbow" on the arm is the value of *k* that is optimal. The basic idea is that we want a small SSE, but that the SSE tends to decrease toward 0 as we increase *k*. The objective is to choose a small value of *k* that still has a low SSE, and the elbow usually represents where we start to have diminishing returns by increasing *k*.

Question 8: What are the advantages and disadvantages of neural networks?

Advantages: Neural networks have led to performance breakthroughs for unstructured data sets such as video, audio, and images(especially deep learning neural networks). They have been able to do things that no other ML algorithms have achieved.

Disadvantages: However, they require a large amount of training data. It's also difficult to pick the right deep learning architecture, and the internal "hidden" layers are incomprehensible. Very difficult to explain. Blackbox.

Question 9: What is the ROC Curve and what is AUC (a.k.a. AUROC)?

The ROC (receiver operating characteristic) curve is a plot of the performance of binary classifiers of True Positive Rate (y-axis) vs. False Positive Rate (x-axis).

AUC is an area under the ROC curve, and it's a common performance metric for evaluating binary classification models.

The higher the AUC the better the classifier.

Question 10: What is cross validation?

Cross-validation is a technique to evaluate predictive models and machine learning algorithms by partitioning the data set into test and training sets. The way to go about doing this is to use something called k-fold cross-validation. The original sample is randomly partitioned into k equal size subsets. Of the k subsets, a single subset is kept as the validation set for testing the model, and the remaining k-1 subsets are used as training data.

The cross-validation process is then repeated k times (k folds), with each of the k subsets used exactly once as the validation data. The k results from the folds can then be averaged (or otherwise combined) to produce a single estimate.

The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.

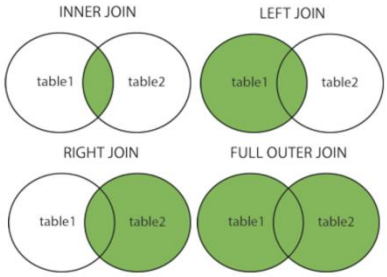
Data Munging, SQL, Visualization, Programming

This is a really broad topic. Depending on the role, the questions can be quite technical. Whereas the other topics are more conceptual. Questions in this category are more tactical and often times you are asked to even write code on the spot.

Question 1: Describe all the joins in SQL.

Here are the different types of the JOINs in SQL:

- (INNER) JOIN: Returns records that have matching values in both tables
- LEFT (OUTER) JOIN: Return all records from the left table, and the matched records from the right table
- RIGHT (OUTER) JOIN: Return all records from the right table, and the matched records from the left table
- FULL (OUTER) JOIN: Return all records when there is a match in either the left or right table



Question 2: What is the SQL query to find the second highest salary employee? Assume we have a table called employee with salary as a field.
Select max(salary) from employee
Where salary not in (select max(salary) from employee);
Question 3: What is the difference between Union and Union All?
UNION command is used to combine the result set of two or more select statements. However, union eliminates the duplicates. UNION ALL includes duplicates.
Question 4: How would you visualize a dataset with height, weight, and eye color in a 2-D graph. (Or ask to visualize any dataset on the whiteboard with more than 2 dimensions)
The basic idea is to encode dimensions beyond 2 as shape or colors or symbols. Anyone who has used Tableau will know this right away. I've personally interviewed doctoral candidates(people in a Ph.D. program) in a quantitative field who struggle with this question.
So, it's not really a measure of intelligence. Just understanding how to visualize higher dimensional data using different types of encoding.
Question 5: What is features engineering? Describe some feature engineering you have done in the past to improve the results of your data science project.
Features engineering is the process of using knowledge about the problem space to create features(factors/variables) that make machine learning algorithms work better.
As the famous Andrew Ng said, "Coming up with features is difficult, time-consuming, requires expert knowledge. "Applied machine learning" is basically feature engineering."
Candidate should explain how they transform existing features into new features that are meaningful to their data science projects.
Question 6: Explain MapReduce conceptually.
The map is a function which "transforms" items in some kind of list to another kind of item and put them back in the same kind of list.
Reduce is a function which "collects" the items in lists and perform some computation on all of them, thus reducing them to a single value.
Question 7: Write a regular expression to parse dates of the YYYY-MM-DD format.
/^\d{4}[\/\-](0?[1-9]|1[012])[\/\-](0?[1-9]|[12][0-9]|3[01])$/
Question 8: Describe the data cleaning process.
How can we cleanse:

- Standardization - making data follow the same rules, notifications, codes
- Enrichment - filling in missing data based on some reference value (eg. City name)
- De-duplication - finding and removing seemingly same but actually duplicate data
- Validations - commonly used for making sure data follows business rules

Question 9: Why is Apache Spark faster than Hadoop?
Basic idea is that it uses in-memory computation instead of writing everything to files like Hadoop.
Also read: A big data or a big data developer? What do you want to become?
Question 10: What is normalization/denormalization? Tradeoffs of each. What is the use case for denormalization?
Most relational databases are denormalized. This means the data is reorganized so that it contains no redundant data and related data separated into different tables. Normalizing reduces disk storage. The normalized the database the more complex the queries are because a query has to join many tables.
The data in a data warehouse, on the other hand, are organized to be read-only and for analytics purposes. Therefore, it does not need to be organized for a normalized fashion. A denormalized data warehouse uses fewer tables and includes many redundancies which are used for reporting and analytical purposes.

Projects & Fit Questions
These are "soft" open-ended questions to understand how you tackle a data science project. How well do you work in teams? What does your prior role look like?
Sometimes, certain behavioral questions are asked during the interview and you need to be prepared for it.

- Question 1: Describe a recent project you've worked on.
- Question 2: What are some of your favorite machine learning algorithms or statistical models and why?
- Question 3: How big was your current team and the team structure? What was your role? Data scientist, data engineer, analyst, architect?
- Question 4: Give an example of a team project you worked on and your contribution to the team.
- Question 5: You are assigned a new data analytics/data science project. How will you begin with and what are the steps you will follow?
- Question 6: What types of management styles do you work best under?
- Question 7: What are your best technical skills? What areas are you weakest in?

1. What Is Ai ? (Artificial Intelligence Questions)
Answer :
Artificial intelligence ("AI") can mean many things to many people. Much confusion arises that the word 'intelligence' is ill-defined. The phrase is so broad that people have found it useful to divide AI into two classes: strong AI and weak AI.
Artificial Intelligence Questions
2. What is TensorFlow and what is it used for ?
Answer :
TensorFlow is an open-source software library originally developed by the Google Brain Team for use in machine learning and neural networks research. It is used for data-flow programming. TensorFlow makes it much easier to build certain AI features into applications, including natural language processing and speech recognition.
3. What are neural networks and how do they relate to AI ?
Answer :
Neural networks are a class of machine learning algorithms. The neuron part of neural is the computational component and the network part is how the neurons are connected. Neural networks pass data among themselves, gathering more and more meaning as the data moves along. Because the networks are interconnected, more complex data can be processed more easily.
4. Which algorithm is used for solving temporal probabilistic reasoning ?
Answer :
To solve temporal probabilistic reasoning, HMM (Hidden Markov Model) is used, independent of transition and sensor model.
5. In Hidden Markov Model, how does the state of the process is described ?
Answer :
The state of the process in HMM's model is described by a 'Single Discrete Random Variable'.
6. In HMM, where does the additional variable is added ?
Answer :
While staying within the HMM network, the additional state variables can be added to a temporal model.
7. Which Is Not Commonly Used Programming Language For Ai ?
Answer :
Perl language is not commonly used programming language for AI
8. What are disadvantages Uniform Cost Search Algorithm ?
Answer :
There can be multiple long paths with the cost ≤ C*.
Uniform Cost search must explore them all.
Artificial Intelligence Questions
9. What Are The Two Different Kinds Of Steps That We Can Take In Constructing A Plan ?
Answer :
a) Add an operator (action)
b) Add an ordering constraint between operators
10. In 'artificial Intelligence' Where You Can Use The Bayes Rule ?
Answer :
In Artificial Intelligence to answer the probabilistic queries conditioned on one piece of evidence, Bayes rule can be used.
2018 Latest Machine learning Interview Questions And Answers
11. To Answer Any Query How The Bayesian Network Can Be Used ?
Answer :
If a Bayesian Network is a representative of the joint distribution, then by summing all the relevant joint entries, it can solve any query.
12. Give some best books to A.I ?
Answer :
a. Artificial Intelligence (3rd edition) –

Patrick Henry Winston has written this book. As it is an introduction to AI. Also, this book is best for the non-programmers. As they can easily understand the explanations and concepts. Moreover, advanced AI topics are covered but haven't been explained in depth. Further, it teaches to build intelligent systems using various real-life examples.

b. Artificial Intelligence: A Modern Approach –

If you have opted a course from Norvig to understand his style of teaching. Hence, you will long for it! Stuart Russell and Peter Norvig have written this book. This is the best book for newcomers for A.I. Also, covers subjects from search algorithm, working with logic to more advanced topics. Moreover, make this book your first choice for A.I.

c. Artificial Intelligence For Humans-

Jeff Heaton has written this book. This books will help to understand the basic artificial intelligence algorithms. Such as dimensionality, distance metrics, clustering, and linear regression. Interesting examples and cases were used to explain these algorithms. Although, to understand this book you need a good command of math. Otherwise, you'll require more time to learn the equations.

d. Paradigm of Artificial Intelligence Programming-

Another one by Peter Norvig! This book will help you to understand the advanced common lisp techniques to build major A.I systems. It is all about practical aspects. Also, it teaches readers the method to build and debug robust practical programs. Moreover, It gives better understanding superior programming style and essential AI concepts. Further, if you are serious about a career, this book is best for you.

13. What is Uniform Cost Search Algorithm ?

Answer :

Basically, it performs sorting in increasing cost of the path to a node. Also, always expands the least cost node. Although, it is identical to Breadth-First search if each transition has the same cost. It explores paths in the increasing order of cost.

14. Which search algorithm will use a limited amount of memory in online search ?

Answer :

RBFE and SMA* will solve any kind of problem that A* can't by using a limited amount of memory.

15. For building a Bayes model how many terms are required ?

Answer :

For building a Bayes model in AI, three terms are required; they are one conditional probability and two unconditional probability.

16. In top-down inductive learning methods how many literals are available? What are they ?

Answer :

There are three literals available in top-down inductive learning methods they are

a) Predicates

b) Equality and Inequality

c) Arithmetic Literals

Artificial Intelligence Questions

17. What combines inductive methods with the power of first order representations ?

Answer :

Inductive logic programming combines inductive methods with the power of first order representations. (Artificial Intelligence training)

18. 2 Batsman Are On 94 Notout,need To Win 7 Runs Off 2 Balls,both Hit A Century? How It Is Possible ?

Answer :

First batsman hit 4 on no ball and then took a single on next ball. Thus completed his century. Second batsman hit 6 on last ball and completed his century too.

19. In Speech Recognition What Kind Of Signal Is Used ?

Answer :

In speech recognition, Acoustic signal is used to identify a sequence of words.

21. Explain artificial intelligence examples and applications ?

Answer :

a. Virtual Personal Assistants

Basically, it is processed in which we have to collect a huge amount data. That is collected from a variety of sources to learn about users. Also, one needs to be more effective in helping them organize and track their information. For Example There are various platforms like iOS, Android, and Window mobile. We use intelligent digital personal assistants are like Siri, Google Now, and Cortana. AI plays an important role in this apps. If you demand they use to collect the information. And this information is used to recognize your request and serves your result.

b. Smart Cars

There are two examples: That are featured Google's self-driving car project and Tesla's "autopilot". Also. the artificial intelligence is been used since the invention of the first video game.

c. Prediction

We call it as the use of predictive analytics. Its main purpose is potential privacy. Also, we can use in many ways. As its also sending you coupons, offering you discounts. That is close to your home with products that you will like to buy. Further, we can call it as the controversial use of artificial intelligence.

d. Fraud Detection

We use AI to detects fraud. As many frauds always happen in banks. Also, computers have a large sample of fraudulent and non-fraudulent purchases. As they asked to look for signs that a transaction falls into one category or another.

22. What are artificial intelligence career domains ?

Answer :

A career in this can be realized within a variety of settings including : private companies public organizations education the arts healthcare facilities government agencies and the military.

23. In Artificial Intelligence, What Do Semantic Analysis Used For ?

Answer :

In Artificial Intelligence, to extract the meaning from the group of sentences semantic analysis is used.

Artificial Intelligence Questions

24. What is the philosophy behind Artificial Intelligence ?

Answer :

As if we see the powers that are exploiting the power of computer system, the curiosity of human lead him to wonder, "Can a machine think and behave like humans do?" Thus, AI was started with the intention of creating similar intelligence in machines. Also, that we find and regard high in humans. company

25. What is AI technique ?

Answer :

Basically, its volume is huge, next to unimaginable. Although, it keeps changing constantly. As AI Technique is a manner to organize. Also, we use it efficiently in such a way that − Basically, it should be perceivable by the people who provide it. As it should be easily modifiable to correct errors. Moreover, it should be useful in many situations. Though it is incomplete or inaccurate.

27. Name search algorithm technology ?

Answer :

a. Problem Space Basically, it is the environment in which the search takes place. (A set of states and set of operators to change those states)

b. Problem Instance It is a result of Initial state + Goal state.

c. Problem Space Graph We use it to represent problem state. Also, we use nodes to show states.

d. The depth of a problem We can define a length of the shortest path.

28. What is the function of the third component of the planning system ?

Answer :

In a planning system, the function of the third component is to detect when a solution to problem has been found.

29. What is "Generality" in AI ?

Answer :

Generality is the measure of ease with which the method can be adapted to different domains of application.

30. Suppose I Have Gmail Account, I Want To Delete All The Mails In My Inbox Having The Same Name(for Eg., Orkut). I Have Thousands Of Mails Like That. So, How Can I Delete All The Mails Having Single Name. Is There Any Option Provided In Gmail ?

Answer :

Yes, its very easy …just do one thing ..in the top of the Inbox page there is a search box just search whatever you want to delete then click .. after few sec all the mail with concerned name get displayed .. just select them and delete them .. as you delete your spam or other mails..

31. What is FOPL stands for and explain its role in Artificial Intelligence ?

Answer :

FOPL stands for First Order Predicate Logic, Predicate Logic provides

a) A language to express assertions about certain "World"

b) An inference system to deductive apparatus whereby we may draw conclusions from such assertion

c) A semantic based on set theory

Predictive Modeling

Q1. (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.

Q2. 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

Q3. 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

Q4. 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

Q5. 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.

Q6. 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

Q7. 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.

Q8. 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.

Q9. 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

Q10. 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.

Q11. 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).

Q12. 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

Q13. 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.

Q14. 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

Q15. 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

Q16. 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.

Q17. 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

Q18. 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.

Q19. 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

Q21. 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

Q22. 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?

1-(0.8)^4. Or, we can use Poisson processes

Q24. 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.

Q25. 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

Q26. 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

Q27. 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

Q28. 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

Q29. 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

Q30. 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

Q31. 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

Q32. 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

Q33. 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

Q34. 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

Q35. 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

Q36. 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.
Q37. 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
Q38. 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
Q39. Let's say you're building the recommended music engine at Spotify to recommend peoplemusic based on past listening history. How would you approach this problem?
collaborative filtering
Q40. 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
Q41. 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
Q42. 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.
Q43. 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.
Q44. 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.
Q45. 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.

Q46. 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

Q47. 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.

Q48. 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

Q49. 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

Q50. 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.

Q51. 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.

Q52. 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.

Q53. 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.

Q54. 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.

Q55. 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}.

Q56. 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).

Q57. 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

Q58. 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

Q59. 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

Q60. 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

Q61. 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.

Q62. 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.

Q63. 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.
Q64. How would you conduct an A/B test on an opt-in feature?
Ask someone for more details.
Q65. 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).
Q66. 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
Q67. 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
Q68. 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.
Q69. 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.
Q70. 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
Q71. 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.
Q72. 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.
Q73. 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

Q74. 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.

Q75. 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.

Q76. 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

Q77. 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

Q78. 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

Q79. 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

Q80. 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

Q81. 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.

Q82. 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.

Q83. QWhat are the different types of missing value treatment?

- Deletion of values
- Guess the value
- Average Substitution
- Regression based substitution
- Multiple Imputation

Q84. 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

Q85. 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

Q86. 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.

Q87. 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

Q88. How to solve overfitting?

- Introduce Regularization
- Perform Cross Validation
- Reduce the number of features
- Increase the number of entries
- Ensembling

Q89. 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

Q90. 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

Q91. 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

Q92. 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

Q93. What are the different ways of performing aggregation in python using pandas?

Group by function

Pivot function

Aggregate function

Q94. 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))

Q95. 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

Q97. 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

Q98. Name few libraries that is used in python for data analysis?

Numpy

Scipy

Pandas

Scikit learn

Matplotlib\ seaborn

Q99. 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

Q100. 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

Q101. What are the different sampling methods?

- Random Sampling
- Systematic Sampling
- Stratified Sampling
- Quota Sampling

Q102. Common Data Quality Issues

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

Q103. 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.

Q104. 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.

Q105. 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).

Q106. Examples of Supervised learning algorithm?

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

Q107. 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.

Q108. 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

Q109. 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

Q110. 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.

Q111. 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.

Q112. 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

Q113. 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.

Q114. 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.

Q115. 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.

Q116. 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

Q117. 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"

Q118. 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.

Q119. 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

Q120. 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.

Q121. 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

Q122. 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

Q123. 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

How the root node is predicted in Decision Tree Algorithm?

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

Q125. What are the different Backend Process available in Keras?

- TensorFlow
- Theano
- CNTK

Q126. Name Few Deep Learning Algorithm

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

Q127. How to split the data with equal set of classes in both training and testing data?

Using Stratified Shuffle package

Q128. What do you mean by giving "epoch = 1" in neural network?

It means that "traversing the data set one time

Q129. 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.

Q130. 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

Q131. Applications of Machine Learning?

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

Q132. If you are given with a use case – 'Predict whether the transaction is fraud (or) not fraud", which algorithm would you choose

Logistic Regression

Q133. If you are given with a use case – 'Predict the house price range in the coming years", which algorithm would you choose

Linear Regression

Q134. What is the underlying mathematical knowledge behind Naïve Bayes?

Bayes Theorem

Q135. 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.

Q136. 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).

Q137. 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.

Q138. 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

Q139. 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

Q140. 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

Q141. What are the different method to split the tree in decision tree?

Information gain and gini index

Q142. What is the weakness for Decision Tree Algorithm?

Not suitable for continuous/Discrete variable

Performs poorly on small data

Q143. Why do we use PCA(Principal Components Analysis) ?

These are important feature extraction techniques used for dimensionality reduction.

Q144. During Imbalanced Data Set, will you

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

We need to calculate precision, Recall separately

Q145.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-out in case of neural network

Q146. 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.

Q147. 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).

Q148. Explain Bigram with an Example?

Eg: I Love Data Science

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

Q149. What are the different activation functions in neural network?

Relu, Leaky Relu , Softmax, Sigmoid

Q150. Which Algorithm Suits for Text Classification Problem?

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

Q151. 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.

Q152. 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.

Q153. 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.

Q154. 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.

Q155. 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.

Q156. 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

Q157. 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

Q158. 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)

Q159. 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.

Q160. 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).

Q161. 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.

Q162. 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).

Q163. 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.

Q164. 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.

Q165. Which Algorithms are used to do a Binary classification?

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

Q166. Which Algorithms are used to do a Multinomial classification?

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

Q167. 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.

Q168. What are all the pre-processing steps that are highly recommended?

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

Q169. What is Normal Distribution?

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

Q170. 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

Q171. What is Regression problem statement?

With the help of Independent variables(X), we predict target variable(Y), if your target variable having infinite possibilities, then the problem will fall under Regression problem statement.

Q172. What are all the Error metrics for Regression problem statement?

Standard error metrics are RMSE & MAPE.

RMSE: Root Mean Squared Error (where we use least square values).

MAPE: Mean Absolute Percent Error (Here, we use absolute values).

Q173. What is R value in Linear regression?

R is the correlation coefficient. Which will be in the range of 0 to 1. If value is closer to 1, it means that Independent variables are highly correlated to your target variable.

Can be given by the formula: (slope*standard deviation(X))/ standard deviation(Y)

Q174. What is an Outlier?

An outlier is an observation that lies in an abnormal distance from other values. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal.

Example:data – (2,1,1,3,4,2,1,4,5,6,2,6,8,9,64,1,7,9)

Only one data point is not in the distribution. You could see all data points are within the range of 1-9. But one data point has a value of 64. Which can be considered as an Influential data point.

Q175. What are all the mechanisms which can identify Outliers?

Box plot is the standard mechanism which can be used in the univariate Analysis.

Scatter plot can be used for Bi-variate Analysis.

Q176. How can we treat Outliers?

Outliers should be to investigated first. Investigation should be in a way that, what is the reason behind that outlier value? Is it possible to change those values by our investigations manually? If can't be treated manually, need to remove the observation if the values are highly deviated. If the deviation is low, can keep the outliers as such and we can proceed.

Q177. What are all the standard imputations that can be carried for missing value treatments?

Mean, Median & Mode can be always the better replacements.

- Central Imputations
- KNN Imputations

Q178. What is the formula for calculating Upper whisker & Lower whisker value in Box plot?

Upper Whisker: Q3 + 1.5(IQR)

Lower Whisker: Q1 – 1.5(IQR)

IQR: Inter-Quartile Range. Which is given by Q3 – Q1.

Q179. What is the skewed Distribution & uniform distribution?

Uniform Distribution is identified when the data spread is equal in the range. Right/Left skewed data is something if data is distributed on any of one side of the plot.

Q180. What is the key assumption for Naive Bayes?

Naïve Bayes assumption tells that all independent variables are equally important as well independent of each other. The reality doesn't support this idea much. But surprisingly Naïve Bayes model sometimes works efficient for classification problem.

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

2. 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

3. How can you generate a random number between 1 - 7 with only a die?

- Launch it 3 times: each throw sets the nth bit of the result.
- For each launch, if the value is 1-3, record a 0, else 1. The result is between 0 (000) and 7 (111), evenly spread (3 independent throw). Repeat the throws if 0 was obtained: the process stops on evenly spread values.

4. 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.

5. 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

6. 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

7. 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

8. You have a group of couples that decide to have children until they have their first girl, after which 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

9. 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

10. 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

11. You call 2 UberX's and 3 Lyfts. If the time that each takes to reach you is IID, what is the probability 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

12. 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

13. On a dating site, users can select 5 out of 24 adjectives to describe themselves. A match is declared 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

14. 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

15. 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

16. What's the expected number of coin flips until you get two heads in a row? What's the expected number of coin flips until you get two tails in a row?

17. 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

18. 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

19. You have a 0.1% chance of picking up a coin with both heads, and a 99.9% chance that you pick up a fair coin. You flip your coin and it comes up heads 10 times. What's the chance that you picked up the fair coin, given the information that you observed?

- Events: F = "picked a fair coin", T = "10 heads in a row"
- (1) P(F|T) = P(T|F)P(F)/P(T) (Bayes formula)
- (2) P(T) = P(T|F)P(F) + P(T|¬F)P(¬F) (total probabilities formula)
- Injecting (2) in (1): P(F|T) = P(T|F)P(F)/(P(T|F)P(F) + P(T|¬F)P(¬F)) = 1 / (1 + P(T|¬F)P(¬F)/(P(T|F)P(F)))
- Numerically: 1/(1 + 0.001 * 2^10 /0.999).
- With 2^10 ≈ 1000 and 0.999 ≈ 1 this simplifies to 1/2

20. What is a P-Value ?

- The probability to obtain a similar or more extreme result than observed when the null hypothesis is assumed.

- ⇒ If the p-value is small, the null hypothesis is unlikely

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

2. What might be the benefits of running an A/A test, where you have two buckets who are exposed to the exact same product?

- Verify the sampling algorithm is random.

3. 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.

4. 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.

5. How would you conduct an A/B test on an opt-in feature?

- Ask someone for more details.

6. 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).

7. 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

8. 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

9. What is a p-value? What is the di erence between type-1 and type-2 error?

-
- type-1 error: rejecting Ho when Ho is true
- type-2 error: not rejecting Ho when Ha is true

10. 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.

11. How would you design an experiment to determine the impact of latency on user engagement?

- 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.

12. 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

13. What's the di erence between a MAP, MOM, MLE estima- tor? 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 is 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.

14. 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.

15. 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 1. (Given a Dataset) Analyze this dataset and tell me what you can learn from it.

2. 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

3. 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

4. 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.

5. What are advantages of plotting your data before per- forming analysis?

- i. 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.
2. Variables can have skewness, outliers etc. Then the arithmetic mean might not be useful. Which means the standard deviation isn't useful.
3. Variables can be multimodal! If a variable is multimodal then anything based on its mean or median is going to be suspect.

6. 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.

7. What is the role of trial and error in data analysis? What is the the role of making a hypothesis before 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.

8. How can you determine which features are the most im- portant 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

9. 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.

10. You have several variables that are positively correlated with your response, and you think combining 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

11. 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

12. 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.

13. 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.

14. You run your regression on different subsets of your data, and find that in each subset, the beta value 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.

15. 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.

16. Given that you have wi data in your o ce, how would you determine which rooms and areas are 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.

17. How could you use GPS data from a car to determine the quality of a driver?

18. Given accelerometer, altitude, and fuel usage data from a car, how would you determine the optimum acceleration pattern to drive over hills?

19. Given position data of NBA players in a season's games, how would you evaluate a basketball player's defensive ability?

20. How would you quantify the influence of a Twitter user?

- like page rank with each user corresponding to the webpages and linking to the page equivalent to following.

21. Given location data of golf balls in games, how would construct a model that can advise golfers where to aim?

22. 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 di culty?

- 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}.

23. You have 5000 people that rank 10 sushis in terms of saltiness. 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).

24. Given data on congressional bills and which congressional 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

25. 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 words) then compare those with other texts by calculating the similarity

26. 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 to include?

- KNN
- choose a small value of k that still has a low SSE (elbow method)
- https://bl.ocks.org/rpgove/0060ff3b656618e9136b

27. Let's say you're building the recommended music engine at Spotify to recommend people music based on past listening history. How would you approach this problem?

- collaborative filtering

Bias1. (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.

2. 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

3. 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

4. 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

5. 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.

6. 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

7. 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.

8. 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.

9. Given training data on tweets and their retweets, how would you predict the number of retweets 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

10. 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.

11. How would you construct a feed to show relevant content for a site that involves user interactions 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).

12. 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

13. 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.

14. 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

15. 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

16. 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.

17. 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

18. 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.

19. 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.
- See Q3.

1. What would be good metrics of success for an advertising-driven consumer product? (Buzzfeed, YouTube, Google Search, etc.) A service-driven consumer product? (Uber, Flickr, Venmo, etc.)

- advertising-driven: Pageviews and daily actives, CTR, CPC (cost per click)
    - click-ads
    - display-ads
- service-driven: number of purchases, conversion rate

2. What would be good metrics of success for a productiv- ity tool? (Evernote, Asana, Google Docs, etc.) A MOOC? (edX, Coursera, Udacity, etc.)

- productivity tool: same as premium subscriptions
- MOOC: same as premium subscriptions, completion rate

3. What would be good metrics of success for an e-commerce product? (Etsy, Groupon, Birchbox, etc.) A subscrip- tion product? (Net ix, Birchbox, Hulu, etc.) Premium subscriptions? (OKCupid, LinkedIn, Spotify, etc.)

- e-commerce: number of purchases, conversion rate, Hourly, daily, weekly, monthly, quarterly, and annual sales, Cost of goods sold, Inventory levels, Site traffic, Unique visitors versus returning visitors, Customer service phone call count, Average resolution time
- subscription

  - churn, CoCA, ARPU, MRR, LTV
- premium subscriptions:

4. What would be good metrics of success for a consumer product that relies heavily on engagement and interac- tion? (Snapchat, Pinterest, Facebook, etc.) A messaging product? (GroupMe, Hangouts, Snapchat, etc.)

- heavily on engagement and interaction: uses AU ratios, email summary by type, and push notification summary by type, resurrection ratio
- messaging product:

5. What would be good metrics of success for a product that o ered in-app purchases? (Zynga, Angry Birds, other gaming apps)

- Average Revenue Per Paid User
- Average Revenue Per User

6. A certain metric is violating your expectations by going down or up more than you expect. How would you try to identify the cause of the change?

- breakdown the KPI's into what consists them and find where the change is
- then further breakdown that basic KPI by channel, user cluster, etc. and relate them with any campaigns, changes in user behaviors in that segment

7. Growth for total number of tweets sent has been slow this month. What data would you look at to determine the cause of the problem?

8. You're a restaurant and are approached by Groupon to run a deal. What data would you ask from them in order to determine whether or not to do the deal?

- for similar restaurants (they should define similarity), average increase in revenue gain per coupon, average increase in customers per coupon

9. You are tasked with improving the e ciency of a subway system. Where would you start?

- define efficiency

10. Say you are working on Facebook News Feed. What would be some metrics that you think are important? How would you make the news each person gets more relevant?

- rate for each action, duration users stay, CTR for sponsor feed posts
- 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

11. How would you measure the impact that sponsored stories on Facebook News Feed have on user engagement? How would you determine the optimum balance between sponsored stories and organic content on a user's News Feed?

- AB test on different balance ratio and see

12. You are on the data science team at Uber and you are asked to start thinking about surge pricing. What would be the objectives of such a product and how would you start looking into this?

- there is a gradual step-function type scaling mechanism until that imbalance of requests-to-drivers is alleviated and then vice versa as too many drivers come online enticed by the surge pricing structure.
- I would bet the algorithm is custom tailored and calibrated to each location as price elasticities almost certainly vary across different cities depending on a huge multitude of variables: income, distance/sprawl, traffic patterns, car ownership, etc. With the massive troves of user data that Uber probably has collected, they most likely have tweaked the algos for each city to adjust for these varying sensitivities to surge pricing. Throw in some machine learning and incredibly rich data and you've got yourself an incredible, constantly-evolving algorithm.

13. Say that you are Net ix. How would you determine what original series you should invest in and create?

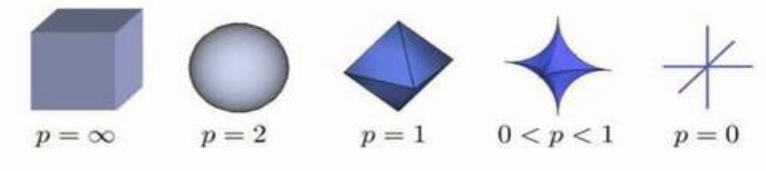
- Netflix uses data to estimate the potential market size for an original series before giving it the go-ahead.

14. What kind of services would nd churn (metric that tracks how many customers leave the service) helpful? How would you calculate churn?

- subscription based services

Q1. Explain what regularization is and why it is useful.

Answer by Matthew Mayo. Regularization is the process of adding a tuning parameter to a model to induce smoothness in order to prevent overfitting. (see also KDnuggets posts on Overfitting) This is most often done by adding a constant multiple to an existing weight vector. This constant is often either the L1 (Lasso) or L2 (ridge), but can in actuality can be any norm. The model predictions should then minimize the mean of the loss function calculated on the regularized training set.



Q2. Which data scientists do you admire most? which startups?

Answer by Gregory Piatetsky: This question does not have a correct answer, but here is my personal list of 12 Data Scientists I most admire, not in any particular order.

Geoff Hinton, Yann LeCun, and Yoshua Bengio - for persevering with Neural Nets when and starting the current Deep Learning revolution.

Demis Hassabis, for his amazing work on DeepMind, which achieved human or superhuman performance on Atari games and recently Go.

Jake Porway from DataKind and Rayid Ghani from U. Chicago/DSSG, for enabling data science contributions to social good.

DJ Patil, First US Chief Data Scientist, for using Data Science to make US government work better.

Kirk D. Borne for his influence and leadership on social media.

Claudia Perlich for brilliant work on ad ecosystem and serving as a great KDD-2014 chair.

Hilary Mason for great work at Bitly and inspiring others as a Big Data Rock Star.

Usama Fayyad, for showing leadership and setting high goals for KDD and Data Science, which helped inspire me and many thousands of others to do their best.

Hadley Wickham, for his fantastic work on Data Science and Data Visualization in R, including dplyr, ggplot2, and Rstudio.

There are too many excellent startups in Data Science area, but I will not list them here to avoid a conflict of interest. Here is some of our previous coverage of startups.

Q3. How would you validate a model you created to generate a predictive model of a quantitative outcome variable using multiple regression.

Answer by Matthew Mayo. Proposed methods for model validation:

- If the values predicted by the model are far outside of the response variable range, this would immediately indicate poor estimation or model inaccuracy.
- If the values seem to be reasonable, examine the parameters; any of the following would indicate poor estimation or multi-collinearity: opposite signs of expectations, unusually large or small values, or observed inconsistency when the model is fed new data.
- Use the model for prediction by feeding it new data, and use the coefficient of determination (R squared) as a model validity measure.
- Use data splitting to form a separate dataset for estimating model parameters, and another for validating predictions.
- Use jackknife resampling if the dataset contains a small number of instances, and measure validity with R squared and mean squared error (MSE).

Q4. Explain what precision and recall are. How do they relate to the ROC curve?

Answer by Gregory Piatetsky:

Here is the answer from KDnuggets FAQ: Precision and Recall:
Calculating precision and recall is actually quite easy. Imagine there are 100 positive cases among 10,000 cases. You want to predict which ones are positive, and you pick 200 to have a better chance of catching many of the 100 positive cases. You record the IDs of your predictions, and when you get the actual results you sum up how many times you were right or wrong. There are four ways of being right or wrong:

- TN / True Negative: case was negative and predicted negative
- TP / True Positive: case was positive and predicted positive
- FN / False Negative: case was positive but predicted negative
- FP / False Positive: case was negative but predicted positive

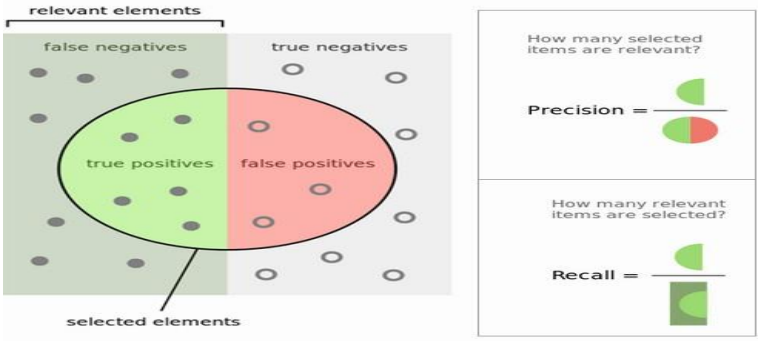
Makes sense so far? Now you count how many of the 10,000 cases fall in each bucket, say:

	Predicted Negative	**Predicted Positive**
Negative Case	TN: 9,760	FP:140
Positive Case	FN: 40	TP: 60

Now, your boss asks you three questions:

1. What percent of your predictions were correct? You answer: the "accuracy" was (9,760+60) out of 10,000 = 98.2%
2. What percent of the positive cases did you catch? You answer: the "recall" was 60 out of 100 = 60%
3. What percent of positive predictions were correct? You answer: the "precision" was 60 out of 200 = 30%

See also a very good explanation of Precision and recall in Wikipedia.



ROC curve represents a relation between sensitivity (RECALL) and specificity(NOT PRECISION) and is commonly used to measure the performance of binary classifiers. However, when dealing with highly skewed datasets, Precision-Recall (PR) curves give a more representative picture of performance.

Q5. How can you prove that one improvement you've brought to an algorithm is really an improvement over not doing anything?

Answer by Anmol Rajpurohit.
Often it is observed that in the pursuit of rapid innovation (aka "quick fame"), the principles of scientific methodology are violated leading to misleading innovations, i.e. appealing insights that are confirmed without rigorous validation. One such scenario is the case that given the task of improving an algorithm to yield better results, you might come with several ideas with potential for improvement.
An obvious human urge is to announce these ideas ASAP and ask for their implementation. When asked for supporting data, often limited results are shared, which are very likely to be impacted by selection bias (known or unknown) or a misleading global minima (due to lack of appropriate variety in test data).

- Data scientists do not let their human emotions overrun their logical reasoning. While the exact approach to prove that one improvement you've brought to an algorithm is really an improvement over not doing anything would depend on the actual case at hand, there are a few common guidelines:
- Ensure that there is no selection bias in test data used for performance comparison
- Ensure that the test data has sufficient variety in order to be symbolic of real-life data (helps avoid overfitting)
- Ensure that "controlled experiment" principles are followed i.e. while comparing performance, the test environment (hardware, etc.) must be exactly the same while running original algorithm and new algorithm
- Ensure that the results are repeatable with near similar results

- Examine whether the results reflect local maxima/minima or global maxima/minima

One common way to achieve the above guidelines is through A/B testing, where both the versions of algorithm are kept running on similar environment for a considerably long time and real-life input data is randomly split between the two. This approach is particularly common in Web Analytics.

Q6. What is root cause analysis?

Answer by Gregory Piatetsky:

According to Wikipedia,

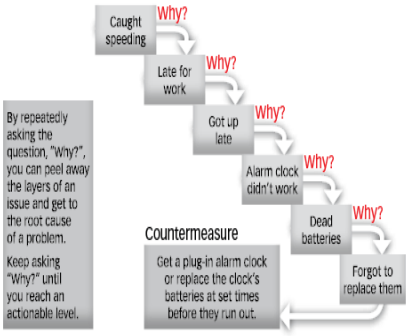
Root cause analysis (RCA) is a method of problem solving used for identifying the root causes of faults or problems. A factor is considered a root cause if removal thereof from the problem-fault-sequence prevents the final undesirable event from recurring; whereas a causal factor is one that affects an event's outcome, but is not a root cause.

Root cause analysis was initially developed to analyze industrial accidents, but is now widely used in other areas, such as healthcare, project management, or software testing.

Here is a useful Root Cause Analysis Toolkit from the state of Minnesota.

Essentially, you can find the root cause of a problem and show the relationship of causes by repeatedly asking the question, "Why?", until you find the root of the problem. This technique is commonly called "5 Whys", although is can be involve more or less than 5 questions.

Five whys analysis example / ONLINE FIGURE 1



Q7. Are you familiar with price optimization, price elasticity, inventory management, competitive intelligence? Give examples.

Answer by Gregory Piatetsky:

Those are economics terms that are not frequently asked of Data Scientists but they are useful to know.

Price optimization is the use of mathematical tools to determine how customers will respond to different prices for its products and services through different channels.

Big Data and data mining enables use of personalization for price optimization. Now companies like Amazon can even take optimization further and show different prices to different visitors, based on their history, although there is a strong debate about whether this is fair.

- Price elasticity in common usage typically refers to Price elasticity of demand, a measure of price sensitivity. It is computed as:

Price Elasticity of Demand = % Change in Quantity Demanded / % Change in Price.

Similarly, Price elasticity of supply is an economics measure that shows how the quantity supplied of a good or service responds to a change in its price.

Inventory management is the overseeing and controlling of the ordering, storage and use of components that a company will use in the production of the items it will sell as well as the overseeing and controlling of quantities of finished products for sale.

Wikipedia defines Competitive intelligence: the action of defining, gathering, analyzing, and distributing intelligence about products, customers, competitors, and any aspect of the environment needed to support executives and managers making strategic decisions for an organization.

Tools like Google Trends, Alexa, Compete, can be used to determine general trends and analyze your competitors on the web.

Q8. What is statistical power?

Answer by Gregory Piatetsky:

Wikipedia defines Statistical power or sensitivity of a binary hypothesis test is the probability that the test correctly rejects the null hypothesis (H0) when the alternative hypothesis (H1) is true.

To put in another way, Statistical power is the likelihood that a study will detect an effect when the effect is present. The higher the statistical power, the less likely you are to make a Type II error (concluding there is no effect when, in fact, there is).
Here are some tools to calculate statistical power.

Q9. Explain what resampling methods are and why they are useful. Also explain their limitations.

Answer by Gregory Piatetsky:

Classical statistical parametric tests compare observed statistics to theoretical sampling distributions. Resampling a data-driven, not theory-driven methodology which is based upon repeated sampling within the same sample.

- Resampling refers to methods for doing one of these
- Estimating the precision of sample statistics (medians, variances, percentiles) by using subsets of available data (jackknifing) or drawing randomly with replacement from a set of data points (bootstrapping)
- Exchanging labels on data points when performing significance tests (permutation tests, also called exact tests, randomization tests, or re-randomization tests) Validating models by using random subsets (bootstrapping, cross validation)

Q10. Is it better to have too many false positives, or too many false negatives? Explain.

Answer by Devendra Desale.

It depends on the question as well as on the domain for which we are trying to solve the question.
In medical testing, false negatives may provide a falsely reassuring message to patients and physicians that disease is absent, when it is actually present. This sometimes leads to inappropriate or inadequate treatment of both the patient and their disease. So, it is desired to have too many false positive.
For spam filtering, a false positive occurs when spam filtering or spam blocking techniques wrongly classify a legitimate email message as spam and, as a result, interferes with its delivery. While most anti-spam tactics can block or filter a high percentage of unwanted emails, doing so without creating significant false-positive results is a much more demanding task. So, we prefer too many false negatives over many false positives.

Q11. What is selection bias, why is it important and how can you avoid it?

Answer by Matthew Mayo.

Selection bias, in general, is a problematic situation in which error is introduced due to a non-random population sample. For example, if a given sample of 100 test cases was made up of a 60/20/15/5 split of 4 classes which actually occurred in relatively equal numbers in the population, then a given model may make the false assumption that probability could be the determining predictive factor. Avoiding non-random samples is the best way to deal with bias; however, when this is impractical, techniques such as resampling, boosting, and weighting are strategies which can be introduced to help deal with the situation.

Q1. What are feature vectors?

Answer: A feature vector is an n-dimensional vector of numerical features that represent some object. In machine learning, feature vectors are used to represent numeric or symbolic characteristics, called features, of an object in a mathematical, easily analyzable way.

Q2. 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 into 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 if you went too far doing splits.

Q3. 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 a problem-solving technique used for isolating the root causes of faults or problems. A factor is called a root cause if its deduction from the problem-fault-sequence averts the final undesirable event from reoccurring.

Q4. What is logistic regression?

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

Q5. What are Recommender Systems?

Answer: 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.

Q6. 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. It is 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 gain insight on how the model will generalize to an independent data set.

Q7. What is Collaborative Filtering?

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

Q8. Do gradient descent methods at all times converge to a similar point?

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

Q9. What is the goal of A/B Testing?

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

Q10. What are the drawbacks of the 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 or binary outcomes

There are overfitting problems that it can't solve

Q11. 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.

Q12. 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.

Q13. Explain star schema?

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

Q14. How regularly must an algorithm be updated?

Answer: You will want to update an algorithm when:

You want the model to evolve as data streams through infrastructure

The underlying data source is changing

There is a case of non-stationarity

Q15. 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. Eigenvalues are the directions along which a particular linear transformation acts by flipping, compressing or stretching.

Q16. Why is resampling done?

Answer: Resampling is done in any of these cases:

Estimating the accuracy of sample statistics by using subsets of accessible data or drawing randomly with replacement from a set of data points

Substituting labels on data points when performing significance tests

Validating models by using random subsets (bootstrapping, cross validation)

Q17. Explain selective bias?

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

Q18. What are the types of biases that can occur during sampling?

Answer:

Selection bias

Under coverage bias

Survivorship bias

Q19. Explain survivorship bias?

Answer: It is the logical error of focusing aspects that support surviving some process and casually overlooking those that did not because of their lack of prominence. This can lead to wrong conclusions in numerous different means.

Q20. How do you work towards a random forest?

Answer: The underlying principle of this technique is that several weak learners combined to 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