Data Science

List of Common Interview Questions

– Compiled by Ankita Sinha (2019)

PART I (Theory)

What is Data Science?

Data Science is a blend of Statistics, technical skills and business vision which is used to analyze the available data and predict the future trend.

How is it different from Big Data and Data Analytics?

Big Data	Data Science	Data Analytics
Huge volumes of data-structured, unstructured and semi-structured	Deals with slicing and dicing the data	Contributing operational insights into complex business scenarios
Requires a basic knowledge of statistics and mathematics	Requires in-depth knowledge of statistics and mathematics	Requires moderate amount of statistics and mathematics

What is a Recommender System?

A recommender system is today widely deployed in multiple fields like movie recommendations, music preferences, social tags, research articles, search queries and so on.

The recommender systems work as per collaborative and content-based filtering or by deploying a personality-based approach.

This type of system works based on a person’s past behavior in order to build a model for the future.

This will predict the future product buying, movie viewing or book reading by people. It also creates a filtering approach using the discrete characteristics of items while recommending additional items.

How do Data Scientists use Statistics?

Statistics helps Data Scientists to look into the data for patterns, hidden insights and convert data into profitable insights.

It helps to get a better idea of what the customers are expecting.

Data Scientists can learn about consumer behavior, interest, engagement, retention and finally conversion all through the power of insightful statistics.

It helps them to build powerful data models in order to validate certain inferences and predictions.

All this can be converted into a powerful business proposition by giving users what they want at precisely when they want it.

What is logistic regression?

It is a statistical technique or a model in order to analyze a dataset and predict the binary outcome.

The outcome has to be a binary outcome that is either zero or one or a yes or no. Random Forest is an important technique which is used to do classification, regression and other tasks on data.

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.

Why data cleansing is important in data analysis?

With data coming in from multiple sources it is important to ensure that data is good enough for analysis.

This is where data cleansing becomes extremely vital.

Data cleansing extensively deals with the process of detecting and correcting of data records, ensuring that data is complete and accurate and the components of data that are irrelevant are deleted or modified as per the needs.

This process can be deployed in concurrence with data wrangling aka batch processing.

Once the data is cleaned it confirms with the rules of the data sets in the system.

Data cleansing is an essential part of the data science because the data can be prone to error due to human negligence, corruption during transmission or storage among other things.

Data cleansing/Data Wrangling takes a huge chunk of time and effort of a Data Scientist because of the multiple sources from which data emanates and the speed at which it comes.

Describe univariate, bivariate and multivariate analysis.

These are methodologies for analysis having single, double or multiple variables.

1. A univariate analysis will have one variable and due to this there are no relationships, causes.
2. The major aspect of the univariate analysis is to summarize the data and find the patterns within it to make actionable decisions.
3. A Bivariate analysis deals with the relationship between two sets of data.
4. These sets of paired data come from related sources, or samples.
5. There are various tools to analyze such data including chi-squared tests and t-tests when the data are having a correlation.
6. If the data can be quantified then it can analyzed using a graph plot or a scatterplot.
7. The strength of the correlation between the two data sets will be tested in a Bivariate analysis.

How machine learning is deployed in real world scenarios?

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

- Ecommerce: Understanding the customer churn (the steady-state of customers a business supports), deploying targeted advertising, remarketing etc.
- Search engine: Ranking pages depending on the personal preferences of the searcher.
- Finance: Evaluating investment opportunities & risks, detecting fraudulent transactions.
- Medicare: Designing drugs depending on the patient's history and needs
- Robotics: Machine learning for handling situations that are out of the ordinary.
- Social media: Understanding relationships and recommending connections.

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

What are the various aspects of a Machine Learning process?

- Domain Knowledge
  This is the first step where we need to understand how to extract the various features from the data and learn more about the data that we are dealing with. It has got more to do with the type of domain that we are dealing with and familiarizing the system to learn more about it.
- Feature Selection
  This step has got more to do with the feature that we are selecting from the set of features that we have. Sometimes it happens that there are a lot of features and we must make an intelligent decision regarding the type of feature that we want to select to go ahead with our machine learning task.
- Algorithm
  This is a vital step since the algorithms that we choose will have a very major impact on the entire process of machine learning. You can choose between the linear and nonlinear algorithm. Some of the algorithms used are Support Vector Machines, Decision Trees, Naïve Bayes, K-Means Clustering, etc.
- Training
  This is the most important part of the machine learning technique and this is where it differs from the traditional programming. The training is done based on the data that we have and providing more real-world experiences. With each consequent training step, the machine gets better and smarter and able to take improved decisions.
- Evaluation
  In this step we evaluate the decisions taken by the machine in order to decide whether it is up to the mark or not. There are various metrics that are involved in

this process and we must closely monitor each of these to decide on the efficacy of the whole machine learning task.

- Optimization

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

- Testing

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

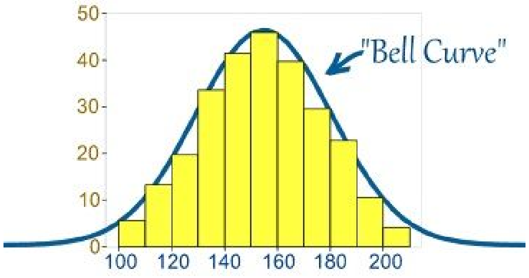
What do you understand by the term Normal Distribution?

It is a set of continuous variables spread across a normal curve or in the shape of a bell curve. It can be considered as a continuous probability distribution and finds much use in statistics. It is the most common distribution curve and it becomes very useful when we must analyze the variables and their relationships with normal distribution curve.

The normal distribution curve is symmetrical.

The non-normal distribution approaches the normal distribution as the size of the samples increases aka law of large numbers.

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



What is Linear Regression?

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

The Linear Regression consists of the following three methods:

- Determining and analyzing the correlation and direction of the data
- Providing Estimation of the model performance
- Ensuring the usefulness and validity of the model

- It is extensively used in scenarios where the cause effect model comes into play.

For example, you want to know the effect of a certain action in order to determine the various outcomes and extent of effect the cause has in determining the final outcome.

What is Interpolation and Extrapolation?

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

Interpolation on the other hand is the method of determining a certain value which falls between a certain set of values or the sequence of values. This is especially useful when you have data at the two extremities of a certain region, but you don't have enough data points at the specific point. This is when you deploy interpolation to determine the value that you need.

What is Power Analysis?

The power analysis is a vital part of the experimental design. It is involved with the process of determining the sample size needed for detecting an effect of a given size with a certain degree of assurance.

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

Power analysis lets you understand the sample size estimate so that they are neither high nor low.

A low sample size means there will not be enough authentication to provide reliable answers and if it is too large then there will be unnecessary wastage of resources.

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

K-means clustering can be termed as the basic unsupervised learning algorithm. It is the method of classifying data using a certain set of clusters called as K clusters. It is deployed for grouping data in order to find similarity in the data.
It includes defining K centers. The clusters are defined into K groups with K being predefined. The K points are selected at random as cluster centers. The data points are assigned to their nearest cluster center. The objects within a cluster are as closely related as possible and differ as much as possible to the objects in other clusters. K-means clustering works very well for large sets of data.

Explain what regularization is and why it is useful.

Regularization is the process of adding a tuning parameter to a model to induce smoothness in order to prevent overfitting.

This is most often done by adding a constant multiple to an existing weight vector. This constant is often either the L1 (Lasso) or L2 (ridge).

The model predictions should then minimize the mean of the loss function calculated on the regularized training set.

What is data normalization and why do we need it?

I felt this one would be important to highlight. 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. If we don't do this then some of the features (those with high magnitude) will be weighted more in the cost function (if a higher-magnitude feature changes by 1%, then that change is pretty big, but for smaller features it's quite insignificant). The data normalization makes all features weighted equally.

Explain dimensionality reduction, where it's used, and it's benefits?

Dimensionality reduction is the process of reducing the number of feature variables under consideration by obtaining a set of principal variables which are basically the important features. Importance of a feature depends on how much the feature variable contributes to the information representation of the data and depends on which technique you decide to use. Deciding which technique to use comes down to trial-and-error and preference. It's common to start with a linear technique and move to non-linear techniques when results suggest inadequate fit. Benefits of dimensionality reduction for a data set may be:

(1) Reduce the storage space needed

(2) Speed up computation (for example in machine learning algorithms), less dimensions mean less computing, also less dimensions can allow usage of algorithms unfit for a large number of dimensions

(3) Remove redundant features, for example no point in storing a terrain's size in both sq meters and sq miles (maybe data gathering was flawed)

(4) Reducing a data's dimension to 2D or 3D may allow us to plot and visualize it, maybe observe patterns, give us insights

(5) Too many features or too complex a model can lead to overfitting.

How do you handle missing or corrupted data in a dataset?

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.

How would you go about doing an Exploratory Data Analysis (EDA)?

The goal of an EDA is to gather some insights from the data before applying your predictive model i.e gain some information.

We start by gaining some high-level global insights. Look at mean and variance of each class.

Check out the first few rows to see what it's all about.

Run a pandas df.info() to see which features are continuous, categorical, their type (int, float, string).

Next, drop unnecessary columns that won't be useful in analysis and prediction.

These can simply be columns that look useless, one's where many rows have the same value (i.e., it doesn't give us much information), or it's missing a lot of values.

We can also fill in missing values with the most common value in that column, or the median.

Now we can start making some basic visualizations. Start with high-level stuff.

Do some bar plots for features that are categorical. Look at the most "general features".

Create some visualizations about these individual features to try and gain some basic insights. Now we can start to get more specific. Create visualizations between features, two or three at a time. How are features related to each other?

You can also do a PCA to see which features contain the most information.

Group some features together as well to see their relationships. For example, what happens to the classes when A = 0 and B = 0? How about A = 1 and B = 0?

Compare different features. For example, if feature A can be either “Female” or “Male” then we can plot feature A against which cabin they stayed in to see if Males and Females stay in different cabins.

Beyond bar, scatter, and other basic plots, Look at some statistics like distribution, p-value, etc.

Finally it’s time to build the ML model.

Start with easier stuff like Naive Bayes and Linear Regression.

If you see that those suck or the data is highly non-linear, go with polynomial regression, decision trees, or SVMs.

The features can be selected based on their importance from the EDA. Check ROC curve. Precision, Recall, RMSE etc.

Do you have any other projects that would be related here?

Here you’ll really draw connections between your knowledge and their business. Is there anything you did or any skills you learned that could possibly connect back to their business or the role you are applying for? It doesn’t have to be 100% exact, just somehow related such that you can show that you will be able to directly add lots of value.

PART II (Statistical Aptitude)

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?

A) Yes

B) No

Solution: **(B)**

These two events cannot be disjoint because P(A)+P(B) >1.

P(A⋅B) = P(A)+P(B)–P(A⋅B).

An event is disjoint if P(A⋅B) = 0. If A and B are disjoint P(A⋅B) = 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.

A) 0.5

B) 0.25

C) 0.333

D) 0.75
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 that 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?

A) 1/36

B) 1/18

C) 5/36

D) 1/6

E) 1/3
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) 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).

A) 1/2

B) 3/8

C) 7/16

D) 9/16
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.

5) When an event A independent of itself?

A) Always

B) If and only if P(A)=0

C) If and only if P(A)=1

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.

6) 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=0.05
    - is the marginal value indicating it is possible to go either way.

7) Do gradient descent methods always converge to the 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

8) 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.

9)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.

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

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.

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?

11) Can you cite some examples where both false positive and false negatives are equally important?

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

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

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

What would be more important in case of an interview at a company? Acquiring Wrong candidates or Rejecting Good Candidates?

12) 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.

13) 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.

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

A) Mean and Normal Distribution

B) Mean, Median and Mode

C) Mode, Alpha & Range

D) Standard Deviation, Range and Mean

E) Median, Range and Normal Distribution

Solution: (B)

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

15) What would be the Type I error?

A) Concluding that listening to music while studying improves memory, and it’s right.

B) Concluding that listening to music while studying improves memory when it actually doesn’t.

C) Concluding that listening to music while studying does not improve memory but it does.

Solution: (B)

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

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

A) Type 1 error

B) Type 2 error

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

D) Cannot be determined

Solution: (D)

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

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

A) Confidence interval is robust to outliers

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

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

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

Solution: (B)

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

18) 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.

19) What is the standard error of the mean?

A) 9

B) 6

C) 7.5

D) 18

Solution: (B)

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

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

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

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

A) Only 1

B) 1 and 3

C) 2 and 3

D) All the statements are true

Solution: (C)

Though sometimes causation might be intuitive from a high correlation but 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.

PART III (Machine Learning)

How is kNN different from k-means clustering?

Answer: Don't get misled 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.

k-means 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.

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.

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.

What are the assumptions in data for Linear Regression models?

Assumptions in Regression

Regression is a parametric approach. 'Parametric' means it makes assumptions about data for the purpose of analysis. Due to its parametric side, regression is restrictive in nature. It fails to deliver good results with data sets which doesn't fulfill its assumptions. Therefore, for a successful regression analysis, it's essential to validate these assumptions.

So, how would you check (validate) if a data set follows all regression assumptions? You check it using the regression plots (explained below) along with some statistical test.

Let's look at the important assumptions in regression analysis:

1. There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one unit change in X^1 is constant, regardless of the value of X^1. An additive relationship suggests that the effect of X^1 on Y is independent of other variables.

2. There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Autocorrelation.

3. The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.

4. The error terms must have constant variance. This phenomenon is known as homoskedasticity. The presence of non-constant variance is referred to heteroskedasticity.

5. The error terms must be normally distributed.

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.

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.

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 the 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.

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 these 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.

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).

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 $R^2 = 1 - \Sigma(y - y')^2/\Sigma(y - ymean)^2$ 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, $\Sigma(y - y')^2/\Sigma(y)^2$ equation's value becomes smaller than actual, resulting in higher R^2.

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.

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 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 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 the rise in global average temperature.

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.

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.

‘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.

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).

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.

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.

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.

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 on 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 were 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 etc.

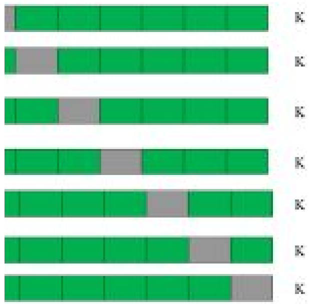
In short, there is no one master algorithm for all situations. We must be scrupulous enough to understand which algorithm to use.

When does regularization becomes necessary in Machine Learning?

Answer: Regularization becomes necessary when the model begins to overfit / 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).

What is k-fold cross validation?

Let's try to visualize how does a k-fold validation work.



This is a 7-fold cross validation.

Here's what goes on behind the scene : we divide the entire population into 7 equal samples. Now we train models on 6 samples (Green boxes) and validate on 1 sample (grey box). Then, at the second iteration we train the model with a different sample held as validation. In 7 iterations, we have basically built model on each sample and held each of them as validation. This is a way to reduce the selection bias and reduce the variance in prediction power. Once we have all the 7 models, we take average of the error terms to find which of the models is best.

What is Confusion Matrix?

A confusion matrix is an N X N matrix, where N is the number of classes being predicted. For the problem in hand, we have N=2, and hence we get a 2 X 2 matrix. Here are a few definitions, you need to remember for a confusion matrix :

- Accuracy : the proportion of the total number of predictions that were correct.
- Positive Predictive Value or Precision : the proportion of positive cases that were correctly identified.
- Negative Predictive Value : the proportion of negative cases that were correctly identified.
- Sensitivity or Recall : the proportion of actual positive cases which are correctly identified.
- Specificity : the proportion of actual negative cases which are correctly identified.

---ENDS HERE--