Data Science Assorted Interview Questions

We believe that you have learned both theoritical and practical knowledge on various algorithms through your assignment.

So let's have some assorted interview questions. This will help you to be prepared for interviews too!

Best with Quest

1. What is the Significance of correlation in data? What do you mean by Pearson correlation? what value of Pearson correlation is significant according to you?

```
Correlation explains how one or more variables are related to
each other. A +ve relation means that if the value of X increases then
the value of Y also increases. Correlation is imp as it'll help us
reduce our features.
If 2 columns are highly correlated, it is better to remove 1 of them as
both will provide same info to model. Just finding a correlation that
tells you whether the relation is + or -ve is not enough. We need a
metric, Pearson correlation is just that, It is a measure of the
strength of a linear association between two variables. The value lies
between -1 to 1, 1 being strong +ve relation 0 means no relation and -1
means strong -ve relation. values above 0.9 and below -0.9 are generally significant.
values close to 0 are discarded in some
cases.
```

2. Difference between correlation and covariance? When should we prefer one over the other?

```
Correlation is a function of the covariance. Covariance is used
to determine how much two random variables vary together, whereas
correlation is used to determine when a change in one variable can
result in a change in another. Covariance gives the extent to which
two random variables change. Correlation represents how strongly two
random variables are related.
Tho not a rule, but is preferred to use the covariance matrix when the
variable are on similar scales and the correlation matrix when the
scales of the variables differ.
```

3. Difference between vectors and scalers?

```
Scalars are quantities that are fully described by a magnitude
(or numerical value) alone. Vectors are quantities that are fully
described by both a magnitude and a direction.
```

4. If a dataset has 1000 columns, how will you deal with them? Can we perform Dimentionality reduction with Deep Learning?

```
When we have huge datasets with many features it is hard to visualize
and group the data, and also we need to see if 2 features are highly
```

correlated.
Thus, we use some Techniques to map multiple dimensions into small space, such as 2 dimensions for plottings. This is called Dimensionality Reduction.
Some Common Techniques are:
- Linear Discriminant Analysis (LDA)
- Principal component analysis (PCA)
- A high correlation between two columns
- Backward/Forward Elimination
- t-distributed Stochastic Neighbor Embedding (t-SNE)

Yes, we can use techniques such as AutoEncoders, Self Organizing Maps(SOM) to perform DR using Deep Learning.

5. How will you find if a feature is significant to your model?

There is a saying in DS: "Garbage in Garbage out". When we have a large no. of features, it is a good idea to see if 2 or more features have 1:1 mapping btw them or are some features too correlated or are they even significant to predict our output?
In order to see their significance to predict output, we use 2 common Elimination Techniques.
Forward and Backward elimination.
Backward elimination is a feature selection technique while building a machine learning model.

It is used to remove those features that do not have a significant effect on the dependent variable or prediction of output.
- We fit the model with all features first and find the p-value for all columns.
- Then we remove the column with the highest p-value, provided it is more than our significance level, typically 0.05.
- Then we fit the model again and do the same until only columns with p-value<SL are left.

Forward Elimination is just the opposite of that.
There are many other ways as well to find imp. Features as we will discuss in further questions.

6. How can you avoid the overfitting your model?

Overfitting is when the model learns the training data too much and doesn't generalize well.
In order to prevent overfitting we can take these steps:
1. Reduce the complexity of the model.
2. Add Dropout in case of ANN
3. Use cross-validation techniques, such as k folds cross-validation
4. Use Regularization techniques such as L1,L2 regularization.

7 What are some assumptions or pre-requisite for Pearson correlation?

For the Pearson r correlation, both variables should be normally distributed.
1.There should be no significant outliers.
2.Each variable should be continuous

```
3.The two variables have a linear relationship.
4.The observations are paired observations.
5.Homoscedascity- Homoscedascity simply refers to ‘equal variances’.
```

8. How do you see if your model's inputs are explaining the outcome properly? Can you use r^2 and adjusted r^2?

```
Yes, These metrics can help us judge the goodness of the model.
R-squared explains the degree to which your input variables explain the
variation of your output / predicted variable. So, if R-square is 0.8,
it means 80% of the variation in the output variable is explained by
the input variables.
The problem with R-squared is that it will either stay the same or
increase with the addition of more variables, even if they do not have
any relationship with the output variables. This is where “Adjusted R
square” comes to help. Adjusted R-square penalizes you for adding
variables that do not improve your existing model.
```

9. What if your dataset has 1 column related to multiple columns of your dataset, how will you handle your data?

```
In this case, if we draw a correlation heat map, we will see a
lot of collinearity and won't be able to decide the columns we want to
remove.
This problem is known as the MultiCollinearity Problem. One good way to
deal with Multicollinearity is the Variance Inflation Factor (VIF).
Variance inflation factors (VIF) measure how much one column is related
to all others in terms of their dependency. It is obtained by
regressing each independent variable, say X on the remaining
independent variables (say Y and Z) and checking how much of it (of X)
is explained by these variables.
Intuitively, we just perform linear regression with all the columns as
target variable one by one, in eat iteration we keep 1 column as target
and the rest as features and see how it is explained by other
variables. We do this with all the columns.
Hence
In general, the Higher the VIF, the higher the R2 which means the
variable X is collinear with Y and Z variables. Columns with values of
VIF above 2 or 5 and in some cases 10, are removed.
```

10. How to handle skewness?]What are the mathematical properties of skewed data? How can you fix it?

```
Skewness is when the distribution of data which is concentrated
in 1 side(left or right) more, or it is sort of collected on one side.
data can be left-skewed or right-skewed.
Right skewed is when data is concentrated on the left, and the tail is
on right. This is also known as positive skewed. Left is -ve skewed
when the tail is on left.
Skewness is caused by the presence of Outliers in the data.
In mathematical terms, the right skewness is when the mean is greater
than the median.
We can remove Skewness by omitting Outliers or taking Log of the
features.
```

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

```
● Univariate data contains only one variable. The purpose of
the univariate analysis is to describe the data and find
patterns that exist within it.
● Bivariate data involves two different variables. The
analysis of this type of data deals with causes and
relationships and the analysis is done to determine the
relationship between the two variables.
● Multivariate data involves three or more variables, it is
categorized under multivariate. It is similar to a bivariate
but contains more than one dependent variable.
```

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

```
Recommendation systems are built by 2 methods. Content-based
filtering and collaborative filtering. Collaborative filtering explains
the behavior of other users and their purchase history in terms of
ratings, selection, etc.
It takes into account all the choices of other users and features
in order to recommend it. The engine makes predictions on what might
interest a person based on the preferences of other users. In this
algorithm, item features are unknown.
```

13. What is Selection Bias? Any Examples?

```
The Selection bias is an experimental error that occurs when the
population being studied does not provide the data that we require to
make conclusions. Basically, our Sample is not representative of the
target population
For eg if you want to study college student's behavior and you only
selected students from 1 college, then that sample of students will not
be a good representative of all the college students in India. Thus
introducing Sample/Selection Bias.


Selection bias is a kind of error that occurs when the researcher decides what has to be
studied. It is
associated with research where the selection of participants is not random. Therefore,
some conclusions
of the study may not be accurate.
The types of selection bias include:
• Sampling bias: It is a systematic error due to a non-random sample of a population
causing some
members of the population to be less likely to be included than others resulting in a
biased
sample.
• Time interval: A trial may be terminated early at an extreme value (often for ethical
reasons), but
the extreme value is likely to be reached by the variable with the largest variance, even
if all
variables have a similar mean.
• Data: When specific subsets of data are chosen to support a conclusion or rejection of
bad data
on arbitrary grounds, instead of according to previously stated or generally agreed
criteria.
```

• Attrition: Attrition bias is a kind of selection bias caused by attrition (loss of participants)
discounting trial subjects/tests that did not run to completion.

14. Explain the difference between supervised and unsupervised machine learning?

Although this isn't one of the most common data scientist interview questions and has more to
do with machine learning than with anything else, it still falls under the umbrella of data science,
so it's worth knowing.
During supervised learning, you would infer a function from a labeled portion of data that's
designed for training. Basically, the machine would learn from objective and concrete examples
that you provide.
Unsupervised learning refers to a machine training method which uses no labeled responses -
the machine learns by descriptions of the input data

15. What are the parametric models? Give an example.

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.

16. What is the difference between classification and regression?

Classification is used to produce discrete results, classification is used to classify data into some specific categories. For example, classifying emails into spam and non-spam categories.
Whereas, We use regression analysis when we are dealing with continuous data, for example predicting stock prices at a certain point in time.

17 How Do You Handle Missing or Corrupted Data in a Dataset?

One of the easiest ways to handle missing or corrupted data is to drop those rows or columns or replace them entirely with some other value.
There are two useful methods in Pandas:
- IsNull() and dropna() will help to find the columns/rows with missing data and drop them
- Fillna() will replace the wrong values with a placeholder value

18. Explain the Bias-Variance Tradeoff.

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 overfitting (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.

Bias: Bias is an error introduced in the model due to the oversimplification of the algorithm used (does
not fit the data properly). It can lead to under-fitting.
Low bias machine learning algorithms – Decision Trees, k-NN and SVM
High bias machine learning algorithms – Linear Regression, Logistic Regression
Variance: Variance is error introduced in the model due to a too complex algorithm, it performs very well
in the training set but poorly in the test set. It can lead to high sensitivity and overfitting.
Possible high variance – polynomial regression

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 until a particular point. As you continue to make your model
more complex, you end up over-fitting your model and hence your model will start suffering from high
variance

Bias-Variance trade-off: The goal of any supervised machine learning algorithm is to have low bias and
low variance to achieve good prediction performance.
1. The k-nearest neighbor 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.
3. The decision tree has low bias and high variance, you can decrease the depth of the tree or use
fewer attributes.
4. The linear regression has low variance and high bias, you can increase the number of features or
use another regression that better fits the data.
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 bias.

19. What is the difference between stochastic gradient descent (SGD) and gradient descent (GD)?

Both algorithms are methods for finding a set of parameters that minimize a loss function by evaluating parameters against data and then making adjustments.
In standard gradient descent, you'll evaluate all training samples for each set of parameters. This is akin to taking big, slow steps toward the solution.

In stochastic gradient descent, you'll evaluate only 1 training sample for the set of parameters before updating them. This is akin to taking small, quick steps toward the solution.

20. How Can You Choose a Classifier Based on a Training Set Data Size?

When the training set is small, a model that has a right bias and low variance seems to work better because they are less likely to overfit.
For example, Naive Bayes works best when the training set is large. Models with low bias and high variance tend to perform better as they work fine with complex relationships.

21. What are 3 data preprocessing techniques to handle outliers?

1. Winsorize (cap at threshold).
2. Transform to reduce skew (using Box-Cox or similar).
3. Remove outliers if you're certain they are anomalies or measurement errors.

22. How much data should you allocate for your training, validation, and test sets?

You have to find a balance, and there's no right answer for every problem.
If your test set is too small, you'll have an unreliable estimation of model performance (performance statistic will have high variance). If your training set is too small, your actual model parameters will have a high variance.
A good rule of thumb is to use an 80/20 train/test split. Then, your train set can be further split into train/validation or into partitions for cross-validation.

23. What Is a False Positive and False Negative and How Are They Significant?

False positives are those cases which wrongly get classified as True but are False. False negatives are those cases which wrongly get classified as False but are True.
In the term ‘False Positive,’ the word ‘Positive’ refers to the ‘Yes’ row of the predicted value in the confusion matrix. The complete term indicates that the system has predicted it as a positive, but the actual value is negative.

24. What Are the Applications of Supervised Machine Learning in Modern Businesses?

Applications of supervised machine learning include:
- Email Spam Detection

Here we train the model using historical data that consists of emails categorized as spam or not spam. This labeled information is fed as input to the model.
- Healthcare Diagnosis

By providing images regarding a disease, a model can be trained to detect if a person is suffering from the disease or not.
- Sentiment Analysis

This refers to the process of using algorithms to mine documents and determine whether they’re positive, neutral, or negative in sentiment.
- Fraud Detection

Training the model to identify suspicious patterns, we can detect instances of possible fraud.

25. What Are Unsupervised Machine Learning Techniques?

There are two techniques used in unsupervised learning: clustering and association.
Clustering
- Clustering problems involve data to be divided into subsets. These subsets, also called clusters, contain data that are similar to each other. Different clusters reveal different details about the objects, unlike classification or regression.

Association
- In an association problem, we identify patterns of associations between different variables or items.
- For example, an eCommerce website can suggest other items for you to buy, based on the prior purchases that you have made, spending habits, items in your wishlist, other customers' purchase habits, and so on.

25 What's the F1 score? How would you use it?

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.

26 How Will You Know Which Machine Learning Algorithm to Choose for Your Classification Problem?

While there is no fixed rule to choose an algorithm for a classification problem, you can follow these guidelines:
- If accuracy is a concern, test different algorithms and cross-validate them
- If the training dataset is small, use models that have low variance and high bias
- If the training dataset is large, use models that have high variance and little bias

30 What evaluation approaches would you work to gauge the effectiveness of a machine learning model?

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.

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

This question has enough hints for you to start thinking! Since the data is spread across the 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.

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

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 the 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).

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

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

34. OLS is to linear regression. The maximum likelihood is logistic regression. Explain the statement.

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 values of parameters which maximizes the likelihood that the parameters are most likely to produce observed data.

35. When does regularization becomes necessary in Machine Learning?

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 the cost term. This helps to reduce model complexity so that the model can become better at predicting (generalizing).

36. What is the Variance Inflation Factor?

Variance Inflation Factor (VIF) is the estimate of the volume of multicollinearity in a collection of many regression variables.
VIF = Variance of the model / Variance of the model with a single independent variable
We have to calculate this ratio for every independent variable. If VIF is high, then it shows the high collinearity of the independent variables.

37. We know that one hot encoding increases the dimensionality of a dataset, but label encoding doesn't. How?

Don't get baffled at this question. It's a simple question asking the difference between the two.

Using 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.

38. What is the Binarizing of data? How to Binarize?

In most of the Machine Learning Interviews, apart from theoretical questions, interviewers focus on the implementation part. So, this ML Interview Questions focused on the implementation of the theoretical concepts.
Converting data into binary values on the basis of threshold values is known as the binarizing of data. The values that are less than the threshold are set to 0 and the values that are greater than the threshold are set to 1. This process is useful when we have to perform feature engineering, and we can also use it for adding unique features.

39 What is cross-validation?

Cross-validation is essentially a technique used to assess how well a model performs on a new independent dataset. The simplest example of cross-validation is when you split your data into two groups: training data and testing data, where you use the training data to build the model and the testing data to test the model.

40 When would you use random forests Vs SVM and why?

There are a couple of reasons why a random forest is a better choice of the model than a support vector machine:
- Random forests allow you to determine the feature importance. SVM's can't do this.
- Random forests are much quicker and simpler to build than an SVM.
- For multi-class classification problems, SVMs require a one-vs-rest method, which is less scalable and more memory intensive.

41. What is the difference between regularization and normalisation?

Normalisation adjusts the data; regularisation adjusts the prediction function. If your data is on very different scales (especially low to high), you would want to normalise the data. Alter each column to have compatible basic statistics. This can be helpful to make sure there is no loss of accuracy. One of the goals of model training is to identify the signal and ignore the noise if the model is given free rein to minimize error, there is a possibility of suffering from overfitting. Regularization imposes some control on this by providing simpler fitting functions over complex ones.

42. Mention some of the EDA Techniques?

Exploratory Data Analysis (EDA) helps analysts to understand the data better and forms the foundation of better models.
Visualization
- Univariate visualization

- Bivariate visualization
- Multivariate visualization

Missing Value Treatment – Replace missing values with Either Mean/Median
Outlier Detection – Use Boxplot to identify the distribution of Outliers, then Apply IQR to set the boundary for IQR

43. What's the Relationship between True Positive Rate and Recall?

The True positive rate in machine learning is the percentage of the positives that have been properly acknowledged, and recall is just the count of the results that have been correctly identified and are relevant. Therefore, they are the same things, just having different names. It is also known as sensitivity.

43. How is machine learning used in day-to-day life?

Most of the people are already using machine learning in their everyday life. Assume that you are engaging with the internet, you are actually expressing your preferences, likes, dislikes through your searches. All these things are picked up by cookies coming on your computer, from this, the behavior of a user is evaluated. It helps to increase the progress of a user through the internet and provide similar suggestions.
The navigation system can also be considered as one of the examples where we are using machine learning to calculate a distance between two places using optimization techniques. Surely, people are going to more engage with machine learning in the near future

44. What is feature engineering? How do you apply it in the process of modelling?

Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

45. How can learning curves help create a better model?

Learning curves give the indication of the presence of overfitting or underfitting.
In a learning curve, the training error and cross-validating error are plotted against the number of training data points.

46. What's the difference between a 'data scientist' and a 'data analyst'?

Even though this is also one of the basic data science interview questions, the terms still often
tend to get mixed up.
Data scientists mine, process and analyze data. They are concerned with providing predictions
for businesses on what problems they might come across.
Data analysts solve the unavoided business problems instead of predicting them beforehand. They identify issues, perform analysis of statistical information and document everything.

47. What's A/B Testing

47. What's A/B testing

While A/B testing can be applied in various different niches, it is also one of the more prominent data science interview questions. So what is it?
A/B testing is a form of tests conducted to find out which version of the same thing is more
worth using to achieve the desired result.
Say, for example, that you want to sell apples. You're not sure what type of apples - red or green ones - your customers will prefer. So you try both - first you try to sell the red
apples, then the green ones. After you're done, you simply calculate which were the more profitable ones and that's it - that's A/B testing!

48. Which is better - good data or good models?

The answer to this question is truly very subjective and case-by-case dependant. Bigger companies might prefer good data, for it is the core of any successful business. On the other
hand, good models couldn't really be created without having good data.
You should probably pick according to your own personal preference - there really isn't any
right or wrong answer (unless the company is specifically searching for either one of them).
So, do your research about the company. Try to see if they're testing your knowledge of their
product or is it a 'trick question'

49. What is 'Expected Value' Vs. 'Mean Value'

When it comes to functionality, there's no difference between the two. However, they are both
used in different situations.
Expected values usually reflect random variables, while mean values reflect the sample population

50. Name a reason why Python is better to use in data science instead of most other programming languages

Naturally, Python is very rich in data science libraries, it's amazingly fast and easy to read or
learn. Python's suite of specialized deep learning and other machine learning libraries includes
popular tools like scikit-learn, Keras, and TensorFlow, which enable data scientists to develop
sophisticated data models that plug directly into a production system.
To unearth insights from the data, you'll have to use Pandas, the data analysis library for Python.
It can hold large amounts of data without any of the lag that comes from Excel. You can do numerical modeling analysis with Numpy. You can do scientific computing and calculation with
SciPy. You can access a lot of powerful machine learning algorithms with the scikit-learn code
library. With Python API and the IPython Notebook that comes with Anaconda, you will get powerful options to visualize your data.
Naturally, Python is very rich in data science libraries, it's amazingly fast and easy to read or

learn. Python's suite of specialized deep learning and other machine learning libraries includes
popular tools like scikit-learn, Keras, and TensorFlow, which enable data scientists to develop
sophisticated data models that plug directly into a production system

51. What do the terms p-value, coefficient, and r squared value mean? What is the significance of each of these components

Imagine you want to predict the price of a house. That will depend on some factors, called independent
variables, such as location, size, year of construction… if we assume there is a linear relationship between
these variables and the price (our dependent variable), then our price is predicted by the following
function:
$Y = a + b X$
The p-value in the table is the minimum α (the significance level) at which the coefficient is relevant. The
lower the p-value, the more important is the variable in predicting the price. Usually we set a 5% level, so
that we have a 95% confidentiality that our variable is relevant.
The p-value is used as an alternative to rejection points to provide the smallest level of significance at
which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in
favor of the alternative hypothesis.
The coefficient value signifies how much the mean of the dependent variable changes given a one-unit
shift in the independent variable while holding other variables in the model constant. This property of
holding the other variables constant is crucial because it allows you to assess the effect of each variable
in isolation from the others.
R squared (R2
) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

52. What is a confusion matrix?

The confusion matrix is a 2X2 table that contains 4 outputs provided by the binary classifier.

A data set used for performance evaluation is called a test data set. It should contain the correct labels
and predicted labels. The predicted labels will exactly the same if the performance of a binary classifier 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 data set as either positive or negative. This produces
four outcomes: TP, FP, TN, FN. Basic measures derived from the confusion matrix:

1. $Error\ Rate$
2. $Accuracy$
3. $Sensitivity\ (Recall\ or\ True\ positive\ rate)$
4. $Specificity\ (True\ negative\ rate)$
5. $Precision\ (Positive\ predicted\ value)$

6. $F - Score$ $(Harmonic\ mean\ of\ precision\ and\ recall)$

53. What is the difference between “long” and “wide” format data?

```
In the wide-format, a subject’s repeated responses will be in a single row, and each
response is in a
separate column. In the long-format, each row is a one-time point per subject. You can
recognize data in
wide format by the fact that columns generally represent groups (variables).
```

54. What are the differences between over-fitting and under-fitting?

```
In statistics and machine learning, one of the most common tasks is to fit a model to a
set of training data,
so as to be able to make reliable predictions on general untrained data.
In overfitting, a statistical model describes random error or noise instead of the
underlying relationship.
Overfitting occurs when a model is excessively complex, such as having too many parameters
relative to
the number of observations. A model that has been overfitted, has poor predictive
performance, as it
overreacts to minor fluctuations in the training data.
Underfitting occurs when a statistical model or machine learning algorithm cannot capture
the underlying
trend of the data. Underfitting would occur, for example, when fitting a linear model to
non-linear data.
Such a model too would have poor predictive performance.
```

55. How to combat Overfitting and Underfitting?

```
To combat overfitting:
1. Add noise
2. Feature selection
3. Increase training set
4. L2 (ridge) or L1 (lasso) regularization; L1 drops weights, L2 no
5. Use cross-validation techniques, such as k folds cross-validation
6. Boosting and bagging
7. Dropout technique
8. Perform early stopping
9. Remove inner layers

To combat underfitting:
1. Add features
2. Increase time of trainin
```

56. What Is the Law of Large Numbers?

```
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 means, the
sample variance
and the sample standard deviation converge to what they are trying to estimate. According
to the law,
```

the average of the results obtained from a large number of trials should be close to the expected value
and will tend to become closer to the expected value as more trials are performed.

57. How does data cleaning play a vital role in the analysis?

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

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

Let us first understand what false positives and false negatives are
• False Positives are the cases where you wrongly classified a non-event as an event a.k.a Type I
error.
• False Negatives are the cases where you wrongly classify events as non-events, a.k.a Type II error.
Example 1: In the medical field, assume you have to give chemotherapy to patients. Assume a patient
comes to that hospital and he is tested positive for cancer, based on the lab prediction but he actually
doesn’t have cancer. This is a case of false positive. Here it is of utmost danger to start chemotherapy on
this patient when he actually does not have cancer. In the absence of cancerous cell, chemotherapy will
do certain damage to his normal healthy cells and might lead to severe diseases, even cancer.
Example 2: Let’s say an e-commerce company decided to give $1000 Gift voucher to the customers whom
they assume to purchase at least $10,000 worth of items. They send free voucher mail directly to 100
customers without any minimum purchase condition because they assume to make at least 20% profit on
sold items above $10,000. Now the issue is if we send the $1000 gift vouchers to customers who have not
actually purchased anything but are marked as having made $10,000 worth of purchase.

59. Can you cite some examples where a false negative important than a false positive? And vice versa?

Example 1 FN: What if Jury or judge decides to make a criminal go free?
Example 2 FN: Fraud detection.
Example 3 FP: customer voucher use promo evaluation: if many used it and actually if was not true,

```
promo sucks
```

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

```
In the Banking industry giving loans is the primary source of making money but at the same time if your
repayment rate is not good you will not make any profit, rather you will risk huge losses.
Banks don’t want to lose good customers and at the same point in time, they don’t want to acquire bad
customers. In this scenario, both the false positives and false negatives become very important to measure.
```

61. What is a Box-Cox Transformation?

```
The dependent variable for a regression analysis might not satisfy one or more assumptions of an ordinary
least squares regression. The residuals could either curve as the prediction increases or follow the skewed
distribution. In such scenarios, it is necessary to transform the response variable so that the data meets
the required assumptions. A Box-Cox transformation is a statistical technique to transform non-normal
dependent variables into a normal shape. If the given data is not normal then most of the statistical
techniques assume normality. Applying a Box-Cox transformation means that you can run a broader
number of tests.
A Box-Cox transformation is a way to transform non-normal dependent variables into a normal shape.
Normality is an important assumption for many statistical techniques, if your data isn’t normal, applying
a Box-Cox means that you are able to run a broader number of tests. The Box-Cox transformation is named
after statisticians George Box and Sir David Roxbee Cox who collaborated on a 1964 paper and developed
the technique.
```

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

```
First of all, you have to ask which ML model you want to train.
For Neural networks: Batch size with Numpy array will work. Steps:
1. Load the whole data in the Numpy array. Numpy array has a property to create a mapping of the
complete data set, it doesn’t load complete data set in memory.
2. You can pass an index to Numpy array to get required data.
3. Use this data to pass to the Neural network.
4. Have a small batch size.
For SVM: Partial fit will work. Steps:
1. Divide one big data set in small size data sets.
2. Use a partial fit method of SVM, it requires a subset of the complete data set.
3. Repeat step 2 for other subsets.
However, you could actually face such an issue in reality. So, you could check out the best laptop for
```

Machine Learning to prevent that. Having said that, let's move on to some questions on deep learning

63. 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?

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.

64. 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?

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.

65.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?

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.

66. 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?

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.

67. Explain machine learning to me like a 5 year old.

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.

68. You have a data science project assignment where you have to deal with 1000 columns and around 1 million rows. The objective of the problem is to carry out classification. You are required to reduce the dimensions of this data in order to reduce the model computation time. Furthermore, your machine suffers from memory constraints. What will you do in this situation?

Considering memory constraints, developing a machine learning model would prove to be a laborious task. However, one can carry this out with the following steps:

Since we are low on our RAM, we can preserve the memory by closing the other miscellaneous applications that we do not require.
We will then perform sampling on our data randomly. This sample will be a much smaller version of the bigger dataset.
We will then reduce the dimensionality by removing the correlated variables. Furthermore, using PCA, we will select those features that can explain maximum variance in our data.
We will further create a linear model using stochastic gradient descent.
Using domain knowledge, we will further drop the predictor variables that do not have much effect on the response variable. This will further lead to a reduction in the number of dimensions.

69. Assume that you are working with categorical features wherein you do not know about the distribution of the categorical variable present in the validation set. Now, you wish to apply one hot encoding on the categorical features. What are the various challenges that you can encounter once you have applied one hot encoding on the categorical variable belonging to the train set?

Applying One Hot Encoding to encode the categories present in the test set but not in the train set, will not involve all the categories of the categorical variable present in the dataset. Secondly, there could be a possible mismatch between the frequency distribution of the categories present in the training set and the validation set.

70. For tuning hyperparameters of your machine learning model, what will be the ideal seed?

There is no fixed value for the seed and no ideal value. The seed is initialized randomly in order to tune the hyperparameters of the machine learning model.

71. Is it true that Pearson captures the monotonic behavior of the relation between the two variables whereas Spearman captures how linearly dependent the two variables are?

No. It is actually the opposite. Pearson evaluates the linear relationship between the two variables whereas Spearman evaluates the monotonic behavior that the two variables share in a relationship.

72. What is the difference between an error and a residual error?

An error occurs in values while the prediction gives us the difference between the observed values and the true values of a dataset. Whereas, the residual error is the difference between the observed values and the predicted values. The reason we use the residual error to evaluate the performance of an algorithm is that the true values are never known. Hence, we use the observed values to measure the error using residuals. It helps us get an accurate estimate of the error.

73. What is the cost function?

Cost functions are a tool to evaluate how good the model performance has been made. It takes into consideration the errors and losses that are made in the output layer during the backpropagation process. In such a case, the errors are moved backward in the neural network, and various other training functions are applied.

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

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

Now Rest with this Quest :)