Company Name - JPMorgan
Role - Data Analyst

Q. What is a sensitivity analysis in the decision-making process?

A. Sensitivity analysis is a method for predicting the outcome of a decision if a situation turns out to be different compared to the key predictions. It helps in assessing the riskiness of a strategy. Helps in identifying how dependent the output is on a particular input value.

Q. How do you interpret the data using statistical techniques?

A. Most Important Methods For Statistical Data Analysis Mean.
1. Standard Deviation.
2. Regression.
3. Sample Size.
4. Determination.
5. Hypothesis Testing.

Q. Describe Map Reduce?

A. MapReduce facilitates concurrent processing by splitting petabytes of data into smaller chunks, and processing them in parallel on Hadoop commodity servers. In the end, it aggregates all the data from multiple servers to return a consolidated output back to the application.

Q. What is a Pivot Table?

A. A pivot table is a table of grouped values that aggregates the individual items of a more extensive table within one or more discrete categories.

Q. Difference between 1-Sample T-test, and 2-Sample T-test?

A. The 2-sample t-test takes your sample data from two groups and boils it down to the t-value. The process is very similar to the 1-sample t-test, and you can still use the analogy of the signal-to-noise ratio. Unlike the paired t-test, the 2-sample t-test requires independent groups for each sample.

Q. Variance and covariance difference?

A. Variance and covariance are mathematical terms frequently used in statistics and probability theory. Variance refers to the spread of a data set around its mean value, while a covariance refers to the measure of the directional relationship between two random variables.

Q. types of Joins in SQL? Difference?

A. INNER JOIN: The INNER JOIN keyword selects all rows from both the tables as long as the condition satisfies. This keyword will create the result-set by combining all rows from both the tables where the condition satisfies i.e value of the common field will be same. LEFT JOIN: This join returns all the rows of the table on the left side of the join and matching rows for the table on the right side of join. The rows for which there is no matching row on right side, the result-set will contain null. LEFT JOIN is also known as LEFT OUTER JOIN. RIGHT JOIN: RIGHT JOIN is similar to LEFT JOIN. This join returns all the rows of the table on the right side of the join and matching rows for the table on the left side of join. The rows for which there is no matching row on left side, the result-set will contain null. RIGHT JOIN is also known as RIGHT OUTER JOIN.
FULL JOIN: FULL JOIN creates the result-set by combining result of both LEFT JOIN and RIGHT JOIN. The result-set will contain all the rows from both the tables. The rows for which there is no matching, the result-set will contain NULL values.

Q. Define Degree of Freedom with example.

A. Degrees of Freedom Definition Degrees of freedom (df) refers to the number of independent values (variable) in a data sample used to find the missing piece of information (fixed) without violating any constraints imposed in a dynamic system. Exp: degree of freedom for given sequence: x = 2, 8, 3, 6, 4, 2, 9, 5.

Given n=" 8

Therefore,

DF = n-1

DF = 8-1

DF = 7.

Q. Explain the difference between Variance and R squared error.

A. Considering this aspect in regression analysis, the variance is the mean squared error that measures the squared and thus, the summed difference between the actual values and the values predicted through the formed regression equation. R-squared error is completely different in concept as compared to variance.

Q. What is an example of a data set with a non-Gaussian distribution?

A. Any distribution of money or value will be non-Gaussian. For example: distributions of income; distributions of house prices; distributions of bets placed on a sporting event. These distributions cannot have negative values and will usually have extended right hand tails.

Q. When is median better measure than the mean?

A. If your data contains outliers, then you would typically rather use the median because otherwise the value of the mean would be dominated by the outliers rather than the typical values. In conclusion, if you are considering the mean, check your data for outliers, if any then better choose median.

Q. Correlation and what is its range?

A. “Correlation” on the other hand measures both the strength and direction of the linear relationship between two variables. It is scaled between the range, -1 and +1.

Q. You have two dices to play with. You win Rs10 every time you roll a 5. If you play till you win and then stop, what is the expected winning money?

A. 1/6 = Probability of getting 5 on dice. It will be (1/6 + 5/6*1/6 + 5/6*5/6*1/6 +....)*10 = 1/6*1/(1-(5/6))*10 = 1*10. Expected money to win is 10Rs. As eventually, we're not stopping till win, so prize money is 10Rs.

Q. Explain How a System Can Play a Game of Chess Using Reinforcement Learning?

A. Reinforcement learning has an environment and an agent. The agent performs some actions to achieve a specific goal. Every time the agent performs a task that is taking it towards the goal, it is rewarded. And, every time it takes a step that goes against that goal or in the reverse direction, it is penalized.

Earlier, chess programs had to determine the best moves after much research on numerous factors. Building a machine designed to play such games would require many rules to be specified.

With reinforced learning, we don’t have to deal with this problem as the learning agent learns by playing the game. It will make a move (decision), check if it’s the right move (feedback), and keep the outcomes in memory for the next step it takes (learning). There is a reward for every correct decision the system takes and punishment for the wrong one.

Q. Difference between Gini Impurity and Entropy in a Decision Tree?

A. The gini impurity measures the frequency at which any element of the dataset will be mislabelled when it is randomly labeled.

The minimum value of the Gini Index is 0. This happens when the node is **pure**, this means that all the contained elements in the node are of one unique class. Therefore, this node will not be split again. Thus, the optimum split is chosen by the features with less Gini Index. Moreover, it gets the maximum value when the probability of the two classes is the same.

Whereas entropy is a measure of information that indicates the disorder of the features with the target. Similar to the Gini Index, the optimum split is chosen by the feature with less entropy. It gets its maximum value when the probability of the two classes is the same and a node is pure when the entropy has its minimum value, which is 0.

The Gini Index and the Entropy have two main differences:

- Gini Index has values inside the interval [0, 0.5] whereas the interval of the Entropy is [0, 1]. In the following figure, both of them are represented. The gini index has also been represented multiplied by two to see concretely the differences between them, which are not very significant.
- Computationally, entropy is more complex since it makes use of **logarithms** and consequently, the calculation of the Gini Index will be faster

Q. Suppose you found that your model is suffering from low bias and high variance. How would you tackle it?

A. 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

Q. Difference between stochastic gradient descent (SGD) and gradient descent (GD)?

A. In both gradient descent (GD) and stochastic gradient descent (SGD), you update a set of parameters in an iterative manner to minimize an error function.
While in GD, you have to run through ALL the samples in your training set to do a single update for a parameter in a particular iteration, in SGD, on the other hand, you use ONLY ONE or SUBSET of training sample from your training set to do the update for a parameter in a particular iteration. If you use SUBSET, it is called Minibatch Stochastic gradient Descent.
Thus, if the number of training samples are large, in fact very large, then using gradient descent may take too long because in every iteration when you are updating the values of the parameters, you are running through the complete training set. On the other hand, using SGD will be faster because you use only one training sample and it starts improving itself right away from the first sample.
SGD often converges much faster compared to GD but the error function is not as well minimized as in the case of GD. Often in most cases, the close approximation that you get in SGD for the parameter values are enough because they reach the optimal values and keep oscillating there.

Q. What are crosstab in python?

A. Crosstab function builds a cross-tabulation table that can show the frequency with which certain groups of data appear.

Q. How complex would SVM become if you have many features (eg 30)?

A. SVM does not overfit when using a lot of features, provided that you regularize correctly. With less features also, sometimes SVM model can have less accuracy. So, if you have many features SVM wouldn't be much complex model.

Q. What is Bimodel, unimodel and skewed data?

A. Unimodal Distribution, as the name suggests, is a single peaked distribution which means one value occurs with the greatest frequency than the other values. Distributions often have a clear peak to their shape. If a distribution has two fairly equal high points, it is called a bimodal distribution. It is a distribution where two values occur with the greatest frequency. The graph resembles two humps on a camel's back. A data is called as skewed when curve appears distorted or skewed either to the left or to the right, in a statistical distribution. In a normal distribution, the graph appears symmetry but here, the data is unsymmetrical.

NOTE:
If you want to learn Data Science from scratch and become interview ready, check out our popular Data science course.

https://ai.cloudyml.com/Learn-Data-Science-from-Scratch