Company Name - HP
Role - Data Scientist

Q. If through training all the features in the dataset, an accuracy of 100% is obtained but with the validation set, the accuracy score is 75%. What should be looked out for?

A. Training accuracy is much higher than validation accuracy, proving that it's the case of overfitting, so in this case, try regularization or making less complex model or any other method to avoid overfitting.

Q. How is skewness different from kurtosis?

A. Skewness basically measures the asymmetry in data. Kurtosis on the other hand, measures the bulge / peak of a distribution curve.

Q. How to calculate the accuracy of a binary classification algorithm using its confusion matrix?

A. Accuracy is calculated as the total number of two correct predictions (TP + TN) divided by the total number of a dataset (P + N).

Q. How will you measure the Euclidean distance between the two arrays in numpy?

A. eucl_distance = np. linalg. norm(point_a - point_b) where np stands for numpy.

Q. In a survey conducted, the average height was 164cm with a standard deviation of 15cm. If Alex had a z-score of 1.30, what will be his height?

A. Alex's height = 164 + 1.30*15 = 183.5 cm.

Q. Unsupervised algorithms?

A. Unsupervised learning aims to discover the dataset's underlying pattern, assemble that data according to similarities, and express that dataset in a precise format. Example: K-means clustering, anomaly detection, PCA etc.

Q. How can you make data normal using Box-Cox transformation?

A. A Box Cox transformation turns non-normal dependent variables into normal shapes. The Lambda value specifies the level of data that should be raised to. The Box-Cox power transformation does this by searching from Lambda = -5 to Lambda = +5 until the best value is discovered.

Q. How do you handle categorical data?

A. One-Hot Encoding is the most common, correct way to deal with non-ordinal categorical data. It consists of creating an additional feature for each group of the categorical feature and mark each observation belonging (Value=1) or not (Value=0) to that group.

Q. Ways to avoid overfitting?

A. Some steps that we can take to avoid it:
1. Data augmentation
2. L1/L2 Regularization
3. Remove layers / number of units per layer
4. Cross-validation

Q. Image classification algorithms?

A. Image Classification algorithms are the algorithms which are used to classify labels for images using their characteristics. Example: Convolutional Neural Networks.

Q. How do you handle null values and which Imputation method is more favourable?

A. Ways to handle missing values in the dataset:

1. Deleting Rows with missing values.
2. Impute missing values for continuous variable.
3. Impute missing values for categorical variable.
4. Other Imputation Methods.
5. Using Algorithms that support missing values.
6. Prediction of missing values.

Multiple imputation is more advantageous than the single imputation because it uses several complete data sets and provides both the within-imputation and between-imputation variability.

Q. args will return?

A. The special syntax *args in function is used to pass a variable number of arguments to a function. It is used to pass a non-key worded, variable-length argument list. The syntax is to use the symbol * to take in a variable number of arguments; by convention, it is often used with the word args.

Q. Difference between having and where clause in SQL.

A. WHERE Clause is used to filter the records from the table based on the specified condition. HAVING Clause is used to filter record from the groups based on the specified condition.

Q. How do you handle categorical data?

A. One-Hot Encoding is the most common, correct way to deal with non-ordinal categorical data. It consists of creating an additional feature for each group of the categorical feature and mark each observation belonging (Value=1) or not (Value=0) to that group.

Q. SQL joins and Groups

A. The SQL Joins clause is used to combine records from two or more tables in a database. The GROUP BY statement groups rows that have the same values into summary rows, like "find the number of customers in each country".

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