Company Name - Tiger Analytics
Role - Data Scientist

Q. Explain the difference between an array and a linked list?

A. An array is a collection of elements of a similar data type. A linked list is a collection of objects known as a node where node consists of two parts, i.e., data and address. Array elements store in a contiguous memory location. Linked list elements can be stored anywhere in the memory or randomly stored.

Q. How do you ensure you are not overfitting a model?

A. Keep your model simple. Use regularization technique. Use cross-validation.

Q. How do you fix high variance in a model?

A. You can reduce High variance, by reducing the number of features in the model. There are several methods available to check which features don't add much value to the model and which are of importance. Increasing the size of the training set can also help the model generalize.

Q. What are hyperparameters? How do they differ from model parameters?

A. Model Parameters: These are the parameters in the model that must be determined using the training data set. These are the fitted parameters. Hyperparameters: These are adjustable parameters that must be tuned in order to obtain a model with optimal performance.

Q. You are told that your regression model is suffering from multicollinearity. How do verify this is true and build a better model?

A. A simple method to detect multicollinearity in a model is by using something called the variance inflation factor or the VIF for each predicting variable. We can follow these steps in order to build a better model:

1. Remove some of the highly correlated independent variables.
2. Linearly combine the independent variables, such as adding them together.
3. Perform an analysis designed for highly correlated variables, such as principal components analysis or partial least squares regression.

Q. You build a random forest model with 10,000 trees. Training error as at 0.00, but validation error is 34.23. Explain what went wrong.

A. It means that the model has mimiced the training pattern perfectly that it will cause overfitting problem in test samples. To avoid this overfitting, use techniques like less complex model or cross validation etc.

Q. What is the recall, specificity and precision of the confusion matrix?

A. The recall is the ratio of the relevant results returned by the search engine to the total number of the relevant results that could have been returned. Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. The precision is the proportion of relevant results in the list of all returned search results.

Q. Filter and wrapper feature selection methods?

A. Wrapper methods measure the "usefulness" of features based on the classifier performance. In contrast, the filter methods pick up the intrinsic properties of the features (i.e., the "relevance" of the features) measured via univariate statistics instead of cross-validation performance. So, wrapper methods are essentially solving the "real" problem (optimizing the classifier performance), but they are also computationally more expensive compared to filter methods due to the repeated learning steps and cross-validation.
Filter methods: information gain, chi-square test, fisher score, correlation coefficient, variance threshold
Wrapper methods: recursive feature elimination, sequential feature selection algorithms, genetic algorithms

Q. C and degree in SVM?

A. The C parameter tells the SVM optimization how much you want to avoid misclassifying each training example. For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. Degree: It is the degree of the polynomial kernel function ('poly') and is ignored by all other kernels. The default value is 3.

Q. What is Dbscan vs kmeans?

A. In Data Science and Machine Learning, KMeans and DBScan are two of the most popular clustering(unsupervised) algorithms. Density clustering algorithms use the concept of reachability i.e. how many neighbors has a point within a radius. DBScan is more lovely because it doesn't need parameter, k, which is the number of clusters we are trying to find, which KMeans needs. When you don't know the number of clusters hidden in the dataset and there's no way to visualize your dataset, it's a good decision to use DBScan. DBSCAN produces a varying number of clusters, based on the input data.

Here's a list of advantages of KMeans and DBScan:

- KMeans is much faster than DBScan
- DBScan doesn't need number of clusters

Here's a list of disadvantages of KMeans and DBScan:

- K-means need the number of clusters hidden in the dataset
- DBScan doesn't work well over clusters with different densities

- DBScan needs a careful selection of its parameters

Q. adjusted R1 and R2?

A. R2 is statistical measurement used to explain the dependent and independent variables. Adjusted R Squared is a measurement that predicts the regression variables. This model will take additional input variable that predicts to solve the problems. Adjusted R2 is the better model when you compare models that have a different number of variables.

Q. How to decide imputation method if there are null values in a dataset?

A. **Mean/Median Imputation: -** In a mean or median substitution, the mean or a median value of a variable is used in place of the missing data value for that same variable. Median over mean when the data column has any outliers.
Mode substitution: - In mode substitution, the highest occurring value for categorical value is used in place of the missing data value of the same variable. Deleting Column: If the number of null values is more than 70%, then prefer deleting the column as it doesn't contribute to the model. If it does, try other methods such as regression or k-means imputation.

Q. What is Autoencoder methods?

A. Autoencoder is a type of neural network where the output layer has the same dimensionality as the input layer. In simpler words, the number of output units in the output layer is equal to the number of input units in the input layer. Various techniques exist to prevent autoencoders from learning the identity function and to improve their ability to capture important ‘ information and learn richer representations. 1.Sparse autoencoder (SAE) 2. Denoising autoencoder (DAE) 3. Contractive autoencoder (CAE) 4. Principal component analysis.

Q. L1 and L2 regularization?

A. L1 regularization gives output in binary weights from 0 to 1 for the model's features and is adopted for decreasing the number of features in a huge dimensional dataset. L2 regularization disperse the error terms in all the weights that leads to more accurate customized final models.

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

A. Euclidean distance is defined in mathematics as the magnitude or length of the line segment between two points. There are multiple methods for measuring the Euclidean methods.

Method 1. In this method, we first initialize two numpy arrays. Then, we use linalg.norm() of numpy basically to compute the euclidean distance directly.
Method 2. In this method, we first initialize two numpy arrays. Then, we take the difference of the two arrays, compute the dot product of the result, and transpose of the result. Then we take the square root of the answer. This is another way to implement Euclidean distance.
Method 3. In this method, we first initialize two numpy arrays. Then, we compute the difference of these arrays and take their square. We take the sum of the squared elements, and after that, we take the square root in the end. This is another way to implement Euclidean distance.

Q. What are the support vectors in SVM?

A. Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyperplane. These are the points that help us build our SVM.

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. What is correlation?

A. Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect.

Q. What is covariance?

A. Covariance is nothing but a measure of correlation. Covariance is a measure of how much two random variables vary together. It’s similar to variance, but where

variance tells you how a single variable varies, co variance tells you how two variables vary together

---

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