Company Name - Philips
Role - Data Scientist

Q. Time Series (ARIMA)?

A. ARIMA, short for 'AutoRegressive Integrated Moving Average', is a forecasting algorithm based on the idea that the information in the past values of the time series can alone be used to predict the future values.

Q. How to reduce overfitting?

A. Techniques to reduce overfitting:

1. Increase training data.
2. Reduce model complexity.
3. Early stopping during the training phase.
4. Ridge Regularization and Lasso Regularization.

Q. What is precision/recall ratio?

A. When it comes to precision we're talking about the true positives over the true positives plus the false positives. As opposed to recall which is the number of true positives over the true positives and the false negatives.

Q. Dimensionality reduction?

A. Dimensionality Reduction is used to reduce the feature space with consideration by a set of principal features.

Q. Bias and variance?

A. Bias is one type of error which occurs due to wrong assumptions about data such as assuming data is linear when in reality, data follows a complex function. On the other hand, variance gets introduced with high sensitivity to variations in training data.

Q. Difference between classification and clustering?

A. In classification data are grouped by analysing data objects whose class label is known. Clustering analyses data objects without knowing class label. There is some prior knowledge of attributes of each classification. There is no prior knowledge of attributes of data to form clusters.

Q. Deal unbalanced classification?

A. Techniques to Handle Imbalanced Data

1. Use the right evaluation metrics
2. Use K-fold Cross-Validation in the right way
3. Ensemble different resampled datasets
4. Resample with different ratios
5. Cluster the abundant class
6. Design your own models

Q. working of bagging and boosting?

A. To understand the working of Bagging, assume we have an N number of models and a Dataset D. Where m is the number of data and n is the number of features in each data. And we are supposed to do binary classification. First, we will split the dataset. For now, we will split this dataset into training and test set only. Let's call the training dataset, where is the total number of training examples.

Take a sample of records from the training set and use it to train the first model, say m1. For the next model, m2 resample the training set and take another sample from the training set. We will do this same thing for the N number of models. Since we are resampling the training dataset and taking the samples from it without removing anything from the dataset, it might be possible that we have two or more training data record common in multiple samples. This technique of resampling the training dataset and providing the sample to the model is termed Row Sampling with Replacement. Suppose we have trained each model, and now we want to see the prediction on test data. Since we are working on binary classification, the output can be either 0 or 1. The test dataset is passed to each model, and we get a prediction from each model. Let's say out of N models more than N/2 models predicted it to be 1; hence, using the model averaging technique like maximum vote, we can say that the predicted output for the test data is 1.

In boosting, we take records from the dataset and pass it to base learners sequentially; here, base learners can be any model. Suppose we have m number of records in the dataset. Then we pass a few records to base learner BL1 and train it. Once the BL1 gets trained, then we pass all the records from the dataset and see how the Base learner works. For all the records classified incorrectly by the base learner, we only take them and pass it to other base learners say BL2 and simultaneously pass the incorrect records classified by BL2 to train BL3. This will go on unless and until we specify some specific number of base learner models we need. Finally, we combine the output from these base learners and create a strong learner; thus, the

model's prediction power gets improved. Ok. So now we know how the Bagging and Boosting work.

Q. Why do we do Regularization and how?

A. Regularization is the process which regularizes or shrinks the coefficients towards zero. In simple words, regularization discourages learning a more complex or flexible model, to prevent overfitting. Regularization is a technique used for tuning the function by adding an additional penalty term in the error function. Regularization, significantly reduces the variance of the model, without substantial increase in its bias.

Q. C and gamma in SVM.. how these improve model?

A. C: It is a hypermeter in SVM to control error. 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.

gamma: Gamma is used when we use the Gaussian RBF kernel. if you use linear or polynomial kernel then you do not need gamma only you need C hypermeter. Somewhere it is also used as sigma. Gamma decides that how much curvature we want in a decision boundary.

Q. Is outlier always need to be removed? What if.. it is relevant to business?

A. Depending on the problem and dataset, we decide whether outliers are important or not. Thus, it is not necessary that Outliers need to be removed all the time because sometimes they provide important information, especially when it is relevant to business.

Q. How do we use confusion metrics?

A. Confusion matrix. This is an NXN matrix where N is called the number of classes being predicted. This metric is called an error matrix and it portrays a dominant role for prediction mainly in the issues of statistical categorization. A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class.

Q. lambda function in python with example?

A. An anonymous function is known as a lambda function. This function can have any number of parameters but, can have just one statement.

Example: a = lambda x,y : x+y
print(a(5, 6))

Output: 11.

Q. What is Chi-Square test?

A. A chi-square test is a statistical test used to compare observed results with expected results. The purpose of this test is to determine if a difference between observed data and expected data is due to chance, or if it is due to a relationship between the variables you are studying.

Q. What is ensemble learning?

A. Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model.

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