Company Name - Nokia
Role - Data Scientist

Q. Difference between LSTM and Simple RNN?

A. The main difference between RNN and LSTM is in terms of which one maintain information in the memory for the long period of time. Here LSTM has advantage over RNN as LSTM can handle the information in memory for the long period of time as compare to RNN.

Q. How do you know if you have enough data for your model?

A. Right kind of data is much more important than amount of data for your model, and should be prioritize first. If your model has diverse features, then surely, you'll require more data as possible in order to understand it better. Having as much data possible is always preferable though one can always check with plotting learning curve in order to determine enough data.

Q. P-value and confidence interval?

A. The P-value is known as the level of marginal significance within the hypothesis testing that represents the probability of occurrence of the given event. A confidence interval is a kind of interval calculation, obtained from the observed data that holds the actual value of the unknown parameter. It displays the probability that a parameter will fall between a pair of values around the mean.

Q. ANOVA vs Chi-Square test

A. Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples. 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. How model behaves when you have multicollinearity?

A. Multicollinearity causes the following two basic types of problems: The coefficient estimates can swing wildly based on which other independent variables are in the model. The coefficients become very sensitive to small changes in the model.

Q. Name algorithms which do not need scaling?

A. CART, Random Forests, Gradient Boosted Decision Trees. Algorithms that do not require normalization/scaling are the ones that rely on rules. They would not be affected by any monotonic transformations of the variables. Scaling is a monotonic transformation.

Q. How to deal with imbalance data in classification modelling?

A. Follow these techniques:

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

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. Pickling and Unpickling?

A. Pickling in python refers to the process of serializing objects into binary streams, while unpickling is the inverse of that. It’s called that because of the pickle module in Python which implements the methods to do this.

Q. What is Dimension reduction?

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

Q. What is Perceptron?

A. Perceptron is a single layer neural network and a multi-layer perceptron is called Neural Networks. Perceptron is a linear classifier (binary). Also, it is used in supervised learning. It helps to classify the given input data. Also it is usually used to classify data into two parts

Therefore, it is also known as Linear Binary Classifier.

Q. What is Incremental learning algorithm in ensemble?

A. An Incremental Learning Algorithm of Ensemble Classifier Systems is basically a method of machine learning in which input data is continuously used to extend the existing model's knowledge i.e., to further train the model. Simply the algorithm that can facilitate the incremental learning are known as incremental machine learning algorithms.

Q. What should you do when your model is suffering from low bias and high variance?

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. How to Standardize data?

A. Standardization comes into picture when features of input data set have large differences between their ranges, or simply when they are measured in different measurement units (e.g., Pounds, Meters, Miles ... etc).

These differences in the ranges of initial features causes trouble to many machine learning models. For example, for the models that are based on distance computation, if one of the features has a broad range of values, the distance will be governed by this particular feature.

Q. What is Z-score?

A. A z-score (also called a *standard score*) gives you an idea of how far from the mean a data point is. But more technically it's a measure of how many standard deviations below or above the population mean a raw score is.

A z-score can be placed on a normal distibution curve. Z-scores range from -3 standard deviations (which would fall to the far left of the normal distribution curve) up to +3 standard deviations (which would fall to the far right of the normal distribution curve). In order to use a z-score, you need to know the mean μ and also the population standard deviation σ.

Z-scores are a way to compare results to a “normal” population. Results from tests or surveys have thousands of possible results and units; those results can often seem meaningless. For example, knowing that someone's weight is 150 pounds might be good information, but if you want to compare it to the "average" person's weight, looking at a vast table of data can be overwhelming (especially if some weights are recorded in kilograms). A z-score can tell you where that person's weight is compared to the average population's mean weight.
