Company Name - IBM
Role - Data Scientist

Q. How do you prevent 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. How do you deal with outliers in your data?

A. Removing outliers is legitimate only for specific reasons. Outliers can be very informative about the subject-area and data collection process. If the outlier does not change the results but does affect assumptions, you may drop the outlier. Or just trim the data set, but replace outliers with the nearest "good" data, as opposed to truncating them completely.

Q. How do you analyse the performance of the predictions generated by regression models versus classification models?

A. For regression, R-square or average error. While for classification, evaluation metrics can be Precision, Recall or F1-score.

Q. What's the name of the matrix used to evaluate predictive models?

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.

Q. What's the relationship between Principal Component Analysis (PCA) and Linear & Quadratic Discriminant Analysis (LDA & QDA)?

A. LDA focuses on finding a feature subspace that maximizes the separability between the groups. While Principal component analysis is an unsupervised Dimensionality reduction technique, it ignores the class label. PCA focuses on capturing the direction of maximum variation in the data set.

Q. What's the difference between logistic and linear regression? How do you avoid local minima?

A. In linear regression, the outcome (dependent variable) is continuous. It can have any one of an infinite number of possible values. In logistic regression, the outcome (dependent variable) has only a limited number of possible values. Logistic regression is used when the response variable is categorical in nature. How do you avoid local minima, use stochastic gradient descent?

Q. What is P value?

A. The p-value is the probability that the null hypothesis is true. (1 – the p-value) is the probability that the alternative hypothesis is true. A low p-value shows that the results are replicable. A low p-value shows that the effect is large or that the result is of major theoretical, clinical or practical importance.

Q. Forward and Backward selection? It's working?

A. Forward Selection chooses a subset of the predictor variables for the final model. Unlike forward stepwise selection, it begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time. forward selection starts with a null model (with no predictors) and proceeds to add variables one at a time, and so unlike backward selection, it DOES NOT have to consider the full model (which includes all the predictors).

Q. How to Fix multicollinearity?

A. To remove multicollinearities, we can do two things.

1. We can create new features
2. remove those features from our data.

Q. Difference between Gradient boosting and random forest?

A. The two main differences are: How trees are built: random forests builds each tree independently while gradient boosting builds one tree at a time. Combining results: random forests combine results at the end of the process (by averaging or "majority rules") while gradient boosting combines results along the way.

Q. Precision and Recall? How they are related to ROC curve?

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. The precision is the proportion of relevant results in the list of all returned search results. When dealing with highly skewed datasets, Precision-Recall (PR) curves give a more informative picture of an algorithm's performance. We show that a deep connection exists between ROC space and PR space, such that a curve dominates in ROC space if and only if it dominates in PR space.

Q. Overfitting and Underfitting? It's relation with bias and variance?

A. Overfitting is a modeling error which occurs when a function is too closely fit to a limited set of data points. Underfitting refers to a model that can neither model the training data nor generalize to new data.

When the bias is more the prediction of the model is very far from the actual value. It means that the model is not having capacity to generalize the distribution of the data. This is underfitting. You will have to increase the complexity of the model so that it can better generalize the data distribution.

On the other hand when the variance of the model is more, then the values predicted by the model are highly spread from the expected value predicted by the model(not the actual value). This is overfitting. The model is highly complicated and needs to be made simple. Otherwise, noise and outliers can take a great toll on the model.

Q. Precision-Recall Trade off?

A. The Idea behind the precision-recall trade-off is that when a person changes the threshold for determining if a class is positive or negative it will tilt the scales. It means that it will cause precision to increase and recall to decrease, or vice versa.

Q. Explain about optimizers?

A. Optimizers are algorithms or methods used to change the attributes of the neural network such as weights and learning rate to reduce the losses. Optimizers are used to solve optimization problems by minimizing the function.

Q. Bias-variance trade-off?

A. The bias–variance trade-off is the property of a model that the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters.

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