Company Name - Genpact
Role - Data Scientist

Q. Inter quartile ranges?

A. The most effective way to find all of your outliers is by using the interquartile range (IQR). The IQR contains the middle bulk of your data, so outliers can be easily found once you know the IQR. Quartiles divide the entire set into four equal parts. So, there are three quartiles, first, second and third represented by Q 1 , Q 2 and Q 3 , respectively. Q 2 is nothing but the median.

Q. Imputation methods?

A. They are:

- List wise or case deletion
- Pairwise deletion
- Mean substitution
- Regression imputation
- Maximum likelihood.

Q. Deal 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. Grid search vs Random search?

A. Random search differs from grid search in that we no longer provide an explicit set of possible values for each hyperparameter; rather, we provide a statistical distribution for each hyperparameter from which values are sampled. Essentially, we define a sampling distribution for each hyperparameter to carry out a randomized search.

Q. Hyperparameters in SVM?

A. kernel: It maps the observations into some feature space. Ideally the observations are more easily (linearly) separable after this transformation. There are multiple standard kernels for these transformations, e.g. the linear kernel, the polynomial kernel and the radial kernel.
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.

CLOUDYML

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. Given a user's history of purchases, how do you predict their next purchase?

A. We can try to compare the products that a user has purchased so far to the purchases of other users (user based recommendation). Based on the goods purchased by another client with comparable tastes, the next purchase can be anticipated. Alternatively, we can try mining association rules among the products that the user has purchased as well as other items (item based recommendation). You may also use clustering algorithms to locate groups of objects or users that are similar. Another strategy is to try to forecast a user's upcoming purchase if you know the information about that user's previous purchase. A Markov Model can be used in this method. The most recent state in a Markov model is anticipated based on a given number of prior states, and this fixed number of previous states is referred to as the Markov model's order. Each state may be a distinct buy in our situation.

Q. Does discarding correlated variables have any effect on PCA?

A. There is no need to remove highly correlated variables because PCA offers a means to deal with them. If N variables are highly correlated, they will all load on the SAME Principal Component (Eigenvector), rather than separate ones. This is how you can tell if they're substantially linked.
If you want to undertake more research, you can either:
1) Use the PCA and interpret it based on the variables that are loaded on it.
2) Pick one of the strongly correlated variables (those that all load onto the same variable) and focus your analysis on it.

CLOUDYML

Q. How would you build a forecasting system that predicts product sales?

A. Sales predictions can be used to set benchmarks, assess the incremental effects of new efforts, allocate resources in response to anticipated demand, and forecast future budgets. This can be accomplished by using different machine learning models like XGBoost, ARIMA, LSTM, etc. The first step would be loading and

cleaning the data (processing outliers and null values), scaling the data. Next would be splitting the data into training and testing data. Further, its required to use the models to predict the results and then score the models based on different metrics. The model giving the best performance should be chosen.

Q. Is standard deviation robust to outliers?

A. Outliers have an effect on standard deviation. When we look at the standard deviation formula, we can see that an extremely high or very low value will raise standard deviation because it is so far from the mean. As a result, outliers have an impact on standard deviation. A data value that is distinct from the rest of the data can boost the statistics' value by an arbitrary amount.

Q. Explain how the map, reduce, and filter functions work.

A. The map() method takes two parameters: another function and a list of 'iterables,' and returns results after applying the function to each iterable in the list. When the filter() method is used, it generates an output list of values that all return true. The reduce() function takes 'iterables' and applies a given function to them, returning a single value.

Q. Are dictionaries or lists faster for lookups?

A. In dictionaries, lookups are faster because Python uses hash tables to implement them. To acquire what you want, we'll have to go through the full list. A dictionary, on the other hand, will return the value you're looking for without going through all of the keys. Because a dictionary may very instantaneously jump to the key it is asked for, the two times above for 100 and 10000000 are nearly identical for a dictionary. If we use Big O concepts to illustrate the distinction, dictionaries have constant time complexity, O(1), whereas lists have linear time complexity, O(n).

Q. Ridge vs lasso?

A. Ridge and Lasso regression are some of the simple techniques to reduce model complexity and prevent over-fitting which may result from simple linear regression . Ridge Regression : In ridge regression, the cost function is altered by adding a

penalty equivalent to square of the magnitude of the coefficients. lasso (least absolute shrinkage and selection operator; also Lasso or LASSO) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.

Q. Goal of A/B Testing?

A. A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

Q. Feature Vectors?

A. A feature vector is a vector containing multiple elements about an object. Putting feature vectors for objects together can make up a feature space.

Q. Write the formula to calculate R-square?

A. R^2 = 1 - (RSS/TSS) where RSS = sum of squares of residual and TSS = Total sum of squares

---

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

CLOUDYML