APPLE DATA SCIENCE INTERVIEW QUESTIONS

EXPLAIN AUTO-REGRESSION.

Auto-regression (AR) is a time series modeling technique where the value of a variable at a particular time step is regressed on its previous values. Some key aspects are -

Lagged Observations: In an AR model, predictions are based on past observations of the same variable, known as "lagged values." For example, an AR model of order 1 (AR(1)) might predict the current value based solely on the previous time step, The coefficients on the model determine how much weight each past observation carries in predicting the current value.

Order of the Model: The order of an AR model, often denoted as p in AR(p), represents the number of past values (lags) used in the model. For example, in an AR(3) model, the current value depends on the previous three values.

Stationary: For an AR model to be valid, the time series data should generally be stationary, meaning its statistical properties (mean, variance) do not change over time. Non-stationary data often need to be transformed (e.g., using differencing) to achieve stationary.

Error Term: The error term ϵt\epsilon_tϵt represents the random fluctuations or noise that the past values can't explain. It's assumed to have a mean of zero and a constant variance.

@karunt

WHAT IS THE MEANING AND CALCULATION OF ACF AND PACF?

The Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) are tools used in time series analysis to understand the relationships and dependencies between observations at different time lags. These functions are essential for identifying the order of Auto-Regressive (AR) and Moving Average (MA) models, especially in models like ARIMA.

The Autocorrelation Function (ACF) measures the correlation between observations in a time series separated by various lags. Essentially, it tells us how strongly current values in the time series are related to past values. In an ACF plot, a spike at lag k indicates significant autocorrelation at that lag. For example, a slow decay in the ACF plot suggests that an MA component may be present in the time series.

The Partial Autocorrelation Function (PACF) measures the correlation between observations separated by a lag, controlling for the influence of intermediate lags. In a PACF plot, spikes at specific lags indicate that those lags are directly influencing the time series. For instance, a single spike at lag k with no significant spikes afterward suggests an AR model of order k.

HOW DOES XGBOOST HANDLE THE BIAS-VARIANCE TRADE-OFF?

XGBoost addresses the bias-variance trade-off through its core methodology of gradient boosting, which balances the trade-off by combining the strengths of multiple weak learners (typically decision trees) to build a strong model

In gradient boosting, each weak learner (decision tree) is trained to correct the errors (residuals) of the previous learners. This approach allows XGBoost to incrementally reduce the bias of the model by correcting mistakes and adding more detail with each additional tree.

XGBoost includes regularization terms (λ and α) in its objective function to penalize the complexity of individual trees, which helps to control variance and prevent overfitting.

XGBoost allows tuning of tree depth and minimum child weight, which controls the maximum depth of the trees and the minimum number of samples required to form a new split. Shallow trees prevent the model from capturing too much detail (lower variance), while deeper trees allow capturing more complexity (lower bias). Tuning these parameters can help find a balance between bias and variance.

There are other parameters like learning rate, subsampling, early stopping as well that can help control bias-variance.

@karunt

SUPPOSE YOU HAVE CREATED 1 DECISION TREE , THE MODEL HAS BIAS (B) AND VARIANCE (V). WHAT WILL HAPPEN TO THE BIAS AND VARIANCE IF I ADD ONE MORE DECISION TREE TO THE MODEL - (1) PARALLELLY LIKE BAGGING (2) SEQUENTIALLY LIKE BOOSTING

Adding a Tree in Parallel (Bagging)

Bagging (Bootstrap Aggregating) involves training multiple decision trees independently in parallel on different random samples of the data and averaging their predictions (or taking the majority vote for classification). This approach does not specifically target the bias of the individual models because each tree is trained independently. Therefore, adding an additional tree in bagging generally does not reduce bias significantly; the model retains a similar bias level to the individual trees. But, bagging reduces variance by averaging the predictions of multiple trees. Each tree's errors (random fluctuations) are averaged out, leading to a more stable overall prediction.

Adding a Tree Sequentially (Boosting)

Boosting is a sequential process where each new tree attempts to correct the errors (or residuals) made by the previous trees. This approach reduces the overall bias of the ensemble because each new tree incrementally improves the fit to the data. Since boosting iteratively fits the model to residuals and each tree depends on the previous ones, the model becomes increasingly complex and specific to the training data. This process can increase variance.

@karunt

WAS THIS HELPFUL?

Be sure to save it so you
can come back to it later!

@karunt