GOOGLE DATA SCIENCE INTERVIEW QUESTINOS

WHAT ARE THE ASSUMPTIONS OF ERROR IN LINEAR REGRESSION

Independence of Errors - The error terms should be independent of each other. This means that there should be no correlation between consecutive errors (no autocorrelation). This assumption is often tested using the Durbin-Watson test in time series data.

Homoscedasticity - The variance of the error terms should remain constant across all levels of the independent variables. If the variance of the errors increases or decreases (heteroscedasticity), it can lead to inefficiencies in the estimation of coefficients.

Normality of Errors - The error terms should be normally distributed, especially for hypothesis testing (i.e., t-tests for coefficients). This assumption is crucial when constructing confidence intervals and p-values.

WHAT IS THE FUNCTION OF P-VALUES IN HIGH DIMENSIONAL LINEAR REGRESSION?

P-values are used to test the null hypothesis that a specific regression coefficient (for a predictor) is zero. A low p-value suggests that the predictor is statistically significant, meaning it likely has an effect on the response variable.

In high-dimensional models, testing many predictors increases the chance of false positives (Type I errors), meaning some predictors might appear significant purely by chance. Traditional p-values need to be adjusted (e.g., Bonferroni correction, FDR methods) to account for this.

High-dimensional data often has strong multicollinearity, meaning many predictors are highly correlated. This can cause unstable estimates of regression coefficients, leading to unreliable p-values. So make sure to remove correlated features

@karunt

LET'S SAY YOU HAVE A CATEGORICAL VARIABLE WITH THOUSANDS OF DISTINCT VALUES, HOW WOULD YOU ENCODE IT?

Leave-One-Out Encoding A variation of target encoding, leave-one-out encoding, computes the target mean for each category, but excludes the current observation to avoid target leakage.
Pros: Reduces target leakage, works well with high-cardinality features.
Cons: Computationally more expensive than simple target encoding.

Embedding-Based Encoding - For extremely high cardinality categorical features, embedding-based approaches are often effective. This technique involves learning a dense vector representation of each category, and you typically use a NN to get the embedding.
Pros: Captures latent structure
Cons: More complex to implement

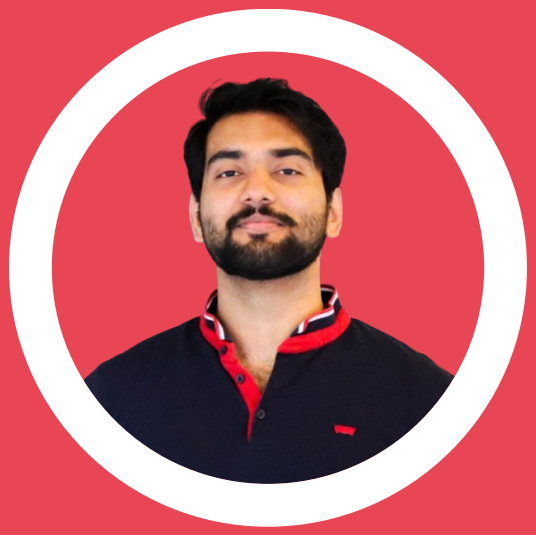
@karunt

DESCRIBE TO ME HOW PCA WORKS

PCA is a dimensionality reduction technique used if you think you have correlated features, noisy data, or to visualize data in fewer dimensions.

To perform PCA you normalize features, calculate covariance matrix (to indicate if variable increase/decrease when another variables does), find eigenvectors (directions where data is most spread out) or eigenvalues (amount of variance/spread)

PCA does assume variables are linearly related, so cant be used for non-linear relationships. Also, new dimension are linear combination of older dimension so interpretation does become harder.



WAS THIS HELPFUL?

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

@karunt