Interview Questions

Scenario on Handling Multicollinearity: Imagine you are working with a dataset intended to predict housing prices based on features like size, location, age of the property, and proximity to amenities. During your analysis, you discover significant multicollinearity between the size of the house and its age. Describe the steps you would take to address this issue. Which specific techniques or metrics would you use to confirm and mitigate multicollinearity to ensure the stability and interpretability of your model?

Step-by-Step Approach

Since the multicollinearity in the dataset will increase the overfitting of the model. So, to reduce the overfitting of the model we need apply the Regularization technique. Regularization is the set of method for reducing the overfitting in machine learning model. By increasing bias and decreasing variance regularization resolve model overfitting.

There are two type of regularization with linear model.
- Lasso Regularization(L1)
- Ridge Regularization(L2)

Since ridge is commonly used to address multicollinearity because it cans shrink the coefficients while still retaining them in the final model, it can be employed in contexts when there are extremely correlated predictors.

Now to apply ridge regularization you need to follow following step:

1. Prepare the data
    a. Data Preprocessing
    b. Splitting of data
2. Compute cost function using following formula

$$\text{Cost} = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

3. Choose the regularization parameter (lambda):
    a. Select lamda using techniques like cross-validation to optimize model performance.
4. Fit the ridge model
    a. Use the computed coefficients beta to train your ridge regression model.
5. Prediction and evaluation

Scenario on Model Evaluation Metrics: You have developed a multiple linear regression model to forecast quarterly sales based on advertising spend, seasonal effects, and economic conditions. The model has an R-squared of 0.85, but your client is concerned about the reliability of predictions. Discuss how you would use MSE and RMSE in this scenario to evaluate model performance further. Explain the implications of these metrics and how they might influence your recommendations for model adjustments or client expectations.

Step by step approach

R-squared is useful when it comes to evaluating regression models which make predictions continuous variables (like sale prices) from training data. R-squared 0.85 means the model explain 85% variance in the sales data.

Mean Squared Error(MSE): It calculates the average squared difference between predicated and actuals sales. Lower MSE indicates better fit.

Root Mean Squared Error(RMSE): It is squared root of Means Squared Error. It is easier to interpret the magnitude of errors.

After calculating MSE and RMSE:

- High MSE and RMSE
    - Large prediction error. Consider
        - Add additional features for refinement of the model
        - Applying data transformation techniques
- Low MSE and RMSE
    - Good average prediction accuracy
    - Recommend the model.

Summary of lesson 3:

Classification is supervised machine learning technique. It gives a class as prediction. For example email spam detection.

Types of classification:

Binary -> two classes

Multi classes -> three or more classes

Classification vs Regression

Classification predict quantities while regression predicts numeric value

Evaluation in classification is more complicated.

Overview of classification Process

- Understand problem
- Data collection
- Data preprocessing
- Splitting the dataset
- Choosing a model
- Training a model
- Model evaluation
- Model Tunning and Optimization -> Hyperparameter tuning, cross validation
- Deployment
- Feedback loop

Data Preprocessing:

Handling missing values -> Imputation and elimination, replace by mean and median

Normalization(Feature scaling)

Encoding categorical variables

One-hot encoding

Ordinal encoding

Logistic Regression-> can be used for classification because it have ability to give probability of class

Gradient Descent