Machine Learning

Supervised ML:

# Regression

## Linear Regression

Explanation for the practice questions:

1. **What is linear regression?**

Linear regression is a supervised machine learning algorithm used to predict a continuous target variable based on one or more input features.

1. **What are the assumptions of linear regression?**

The assumptions of linear regression include linearity, independence, homoscedasticity, and normality of residuals.

1. **How do you interpret the coefficients in linear regression?**

The coefficients in linear regression represent the change in the target variable for a one-unit change in the corresponding input feature, assuming all other variables are held constant.

1. **What is the difference between simple linear regression and multiple linear regression?**

Simple linear regression involves predicting a target variable using only one input feature, while multiple linear regression involves predicting the target variable using multiple input features.

1. **How do you evaluate the performance of a linear regression model?**

The performance of a linear regression model can be evaluated using metrics such as mean squared error (MSE), root mean squared error (RMSE), and R-squared.

1. **What is the purpose of the cost function in linear regression?**

The cost function in linear regression measures the difference between the predicted values and the actual values, and it is used to optimize the model parameters during the training process.

1. **How can you handle categorical variables in linear regression?**

Categorical variables can be handled in linear regression by using techniques such as one-hot encoding or ordinal encoding to convert them into numerical values.

1. **What is the difference between correlation and regression?**

Correlation measures the strength and direction of the linear relationship between two variables, while regression predicts the value of a dependent variable based on the values of independent variables.

1. **What is the purpose of regression analysis in machine learning?**
   * Regression analysis is used to model the relationship between a dependent variable and one or more independent variables. It helps in predicting continuous numerical values and understanding the impact of independent variables on the dependent variable.
2. **Explain the concept of linear regression and its assumptions.**
   * Linear regression is a statistical approach to modeling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables and that the errors are normally distributed and have constant variance.
3. **What is the difference between simple linear regression and multiple linear regression?**
   * Simple linear regression involves only one independent variable, while multiple linear regression involves two or more independent variables. Simple linear regression aims to find a linear relationship between the dependent variable and a single independent variable, whereas multiple linear regression considers the combined effect of multiple independent variables on the dependent variable.
4. **How do you interpret the coefficients in a linear regression model?**
   * The coefficients in a linear regression model represent the change in the dependent variable for a one-unit change in the corresponding independent variable, holding all other variables constant. A positive coefficient indicates a positive relationship, while a negative coefficient indicates a negative relationship.
5. **What is the purpose of data splitting in regression analysis?**
   * Data splitting is done to evaluate the performance of a regression model. The dataset is divided into training and testing sets. The training set is used to train the model, while the testing set is used to assess how well the model generalizes to unseen data.
6. **Describe the process of cross-validation and its advantages.**
   * Cross-validation is a technique used to assess the performance of a model by splitting the data into multiple subsets. It involves training the model on a subset of the data and evaluating it on the remaining subset. This process is repeated multiple times, and the average performance is calculated. Cross-validation helps in estimating the model's performance on unseen data and reduces the risk of overfitting.
7. **What is the bias-variance trade-off in machine learning?**
   * The bias-variance trade-off refers to the trade-off between the model's ability to fit the training data (low bias) and its ability to generalize to unseen data (low variance). A model with high bias may underfit the data, while a model with high variance may overfit the data. Finding the right balance is crucial for building a good predictive model.
8. **Explain the concept of regularization and its role in regression models.**
   * Regularization is a technique used to prevent overfitting in regression models. It adds a penalty term to the loss function, which discourages the model from fitting the training data too closely. Regularization helps in reducing the model's complexity and improving its generalization performance.
9. **What are the differences between Ridge, LASSO, and Elastic Net regularization techniques?**
   * Ridge regularization adds the sum of squared coefficients multiplied by a regularization parameter to the loss function. It shrinks the coefficients towards zero but does not set them exactly to zero.
   * LASSO regularization adds the sum of absolute values of coefficients multiplied by a regularization parameter to the loss function. It can set some coefficients exactly to zero, effectively performing feature selection.
   * Elastic Net regularization is a combination of Ridge and LASSO regularization. It adds both the sum of squared coefficients and the sum of absolute values of coefficients multiplied by regularization parameters to the loss function. It provides a balance between Ridge and LASSO regularization.
10. **How can you evaluate the performance of a regression model using R-squared?**
    * R-squared, also known as the coefficient of determination, is a metric used to evaluate the goodness of fit of a regression model. It measures the proportion of the variance in the dependent variable that is predictable from the independent variables. R-squared ranges from 0 to 1, where 1 indicates a perfect fit. Higher R-squared values indicate better model performance.