Machine Learning Test

☐ Linear Regression

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

Simple Linear Regression: Simple linear regression involves only **one independent variable** and **one dependent variable**. It models the relationship between the independent variable (predictor) and the dependent variable (response) using a straight line.

The equation for simple linear regression can be represented as:

$$Y = B0 + B1*X + E$$

Where:

- Y is the dependent variable.
- X is the independent variable.
- B0 is the intercept.
- B1 is the slope of the line.
- E represents the error term.

Multiple Linear Regression: Multiple linear regression involves **two or more independent variables** and **one dependent variable**.

The equation for multiple linear regression can be represented as:

$$Y = B0 + B1X1 + B2X2 + ... + Bn*Xn + E$$

Where:

- Y is the dependent variable.
- X1, X2, ..., Xn are the independent variables.
- B0 is the intercept (constant term).
- B1, B2, ..., Bn are the coefficients of the independent variables.
- E represents the error term.

2.Explain the concept of the cost function in linear regression.

The cost function in linear regression quantifies the difference between predicted and actual values, reflecting model accuracy. common cost functions include:

Mean Squared Error (MSE)
Mean Absolute Error (MAE)
Root Mean Squared Error (RMSE)

3.How do you interpret the coefficients in a linear regression model?

In a linear regression model, coefficients signify the change in the dependent variable for a unit change in the corresponding independent variable, holding other variables constant.

4.What are the assumptions of linear regression?

There are four assumptions of linear regression model:

Linearity: The relationship between X and the mean of Y is linear.
Homoscedasticity: The variance of residual is the same for any value of X.
Independence: Observations are independent of each other.
Normality: For any fixed value of X, Y is normally distributed.

☐ Logistic Regression

1.How does logistic regression differ from linear regression?

- Linear regression predicts continuous outcomes, while logistic regression predicts probabilities for binary outcomes.

- Linear regression models a linear relationship, while logistic regression models the log-odds of the outcome.

2.Explain the sigmoid function and its role in logistic regression.

- In logistic regression, the sigmoid function is used to model the probability that an observation belongs to a certain class or category, given its input features.

- The sigmoid function, also known as the logistic function, is a mathematical function that maps any real-valued number to a value between 0 and 1.

3.What are the key performance metrics used to evaluate a logistic regression model?

- **Accuracy**
- **Precision**
- **Recall (Sensitivity)**
- **F1 Score**
- **ROC Curve**
- **Confusion Matrix**

4.How do you handle multicollinearity in logistic regression?

In logistic regression, multicollinearity can distort the relationships between predictors and the outcome variable, affecting the accuracy and reliability of the model.

To handle multicollinearity in logistic regression:

Feature Selection
Principal Component Analysis (PCA)
Ridge Regression

☐ Naive Bayes

1.What is the Naive Bayes algorithm based on?

Naive Bayes is a probabilistic machine learning algorithm based on Bayes' theorem with the assumption of independence between the features. It's called "naive" because it assumes that the presence of a particular feature in a class is independent of the presence of any other feature.

2.Explain the concept of conditional probability in the context of Naive Bayes.

Conditional probability in Naive Bayes refers to the likelihood of an event occurring given that another event has already occurred. In the context of classification, Naive Bayes calculates the probability of a class given the values of input features using conditional probabilities.

3.What are the advantages and disadvantages of Naive Bayes?

Advantages:

- Simple and easy to implement.
- Requires a small amount of training data to estimate parameters.
- Fast and efficient for both training and prediction.
- Performs well in practice, especially in text classification and spam filtering tasks.

Disadvantages:

- Strong assumption of feature independence, which may not hold true in some real-world scenarios.
- Tends to perform poorly when the independence assumption is violated or when features are highly correlated.
- Cannot learn interactions between features.

4.How does Naive Bayes handle missing values and categorical features?

- Naive Bayes can handle missing values by ignoring them during training and prediction.
- Naive Bayes can handle categorical features by treating them as discrete variables.

☐ Decision Trees

1.How does a decision tree make decisions?

A decision tree makes decisions by recursively splitting the dataset into subsets based on the values of input features. At each step, it selects the feature that best separates the data into pure or nearly pure subsets according to some criterion.

2.What are the main criteria for splitting nodes in a decision tree?

The main criteria for splitting nodes in a decision tree include:

- **Gini impurity:** Measures the impurity of a node by calculating the probability of misclassifying a randomly chosen sample.
- **Entropy:** Measures the level of disorder or randomness in a node's samples.
- **Information gain:** Measures the reduction in entropy or impurity achieved by splitting a node based on a particular feature. It selects the feature that maximizes information gain.

3.How do decision trees handle categorical variables?

Decision trees can handle categorical variables by considering each category as a separate branch in the tree. They split the data based on the categories of the categorical variable, creating separate branches for each category. Alternatively, decision trees can use binary encoding or one-hot encoding to convert categorical variables into numerical values, making them suitable for splitting nodes based on numerical criteria.

4.What are some common techniques to prevent overfitting in decision trees?

Pruning: Pruning techniques such as cost-complexity pruning (or reduced-error pruning) remove branches of the tree that do not significantly improve the model's performance on a validation dataset. This prevents the tree from becoming overly complex and captures the underlying patterns more effectively.

Maximum depth: Limiting the maximum depth of the tree prevents it from growing too deep and overfitting to the training data.

Minimum information gain: Setting a minimum threshold for information gain ensures that only splits that significantly reduce impurity are considered, preventing the tree from capturing noise.

☐ Support Vector Machines (SVM)

1.What is the basic idea behind SVM?

The basic idea behind Support Vector Machines (SVM) is to find the **hyperplane that best separates the data points into different classes in a high-dimensional space**. SVM aims to maximize the margin between the hyperplane and the nearest data points of each class, while minimizing the classification error.

2.Explain the concepts of margin and support vectors in SVM.

Margin: The distance between the decision boundary and the nearest data points; SVM aims to maximize this margin.
Support vectors: Data points closest to the decision boundary; they determine the position and orientation of the boundary.

3.What are the different kernel functions used in SVM, and when would you use each?

- **Linear:** For linearly separable data.
- **Polynomial:** For curved separation boundaries.
- **RBF (Radial Basis Function):** For non-linear separation; versatile for various data distributions.

4.How does SVM handle outliers?

SVM is robust to outliers due to margin focus. Adjusting regularization parameters like C or using kernel functions can mitigate outlier effects. Outliers lying far from the boundary have minimal impact.