Machine Learning Theoretical Questions

Linear Regression

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

Simple linear regression and multiple linear regression are both techniques used in statistical modeling to understand the relationship between independent variables (features) and a dependent variable (target). The main difference between them lies in the number of independent variables they consider.

Simple Linear Regression: Simple linear regression involves only two variables: one independent variable (predictor) and one dependent variable (response).
The relationship between the independent and dependent variables is modeled using a straight line.
The equation of the regression line is expressed as: $y = m.x + c + \varepsilon$ where y is the dependent variable, x is the independent variable, c is the intercept, m is the slope coefficient, and ε is the error term.
Simple linear regression is suitable when there is a linear relationship between the two variables.

Multiple Linear Regression: Multiple linear regression involves two or more independent variables (predictors) and one dependent variable (response).
The relationship between the independent variables and the dependent variable is modeled using a linear equation involving multiple predictors.
The equation of the regression line is expressed as: $y = c + m_1.x_1 + m_2.x_2 + m_3.x_3 ..... + m_n.x_n + \varepsilon$ where y is the dependent variable, $x_1, x_2,..., x_n$ are the independent variables, c is the intercept, $m_1, m_2,..., m_n$ are the coefficients of the independent variables, and ε is the error term.
Multiple linear regression allows for modeling more complex relationships between the dependent variable and multiple predictors.

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

In linear regression, the cost function (also known as the loss function or objective function) measures the difference between the predicted values and the actual values of the dependent variable. The goal of linear regression is to minimize this cost function, which quantifies the overall error of the model.

Mean Squared Error (MSE) Cost Function:
The most commonly used cost function in linear regression is the Mean Squared Error (MSE), which calculates the average squared difference between the predicted values

$J(\theta) = 1/2m\ \Sigma_{i=1}^{m}\ (\hat{}y_i - y_i)^2$

where:
- $J(\theta)$ is the cost function.
- m is the number of observations.
- (θ) represents the model parameters (intercept and coefficients).
- $\hat{}y_i$ is the predicted value of the dependent variable for observation i.
- y_i is the actual value of the dependent variable for observation i.

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

In a linear regression model, the coefficients (also known as weights or parameters) represent the change in the dependent variable for a one-unit change in the corresponding independent variable, while holding all other variables constant.

- **Intercept c :** Represents the value of the dependent variable when all independent variables are zero. It is the predicted value of the dependent variable when all predictors have no effect.
- **Coefficients m_1, m_2,.., m_n:** Represents 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 that an increase in the independent variable is associated with an increase in the dependent variable, while a negative coefficient indicates the opposite.

Q.4.What are the assumptions of linear regression?

- **Linearity:** There exists a linear relationship between the independent variables and the dependent variable.
- **Independence:** The errors (residuals) of the model are independent of each other.
- **Less Variance**: The variance of the errors is constant across all levels of the independent variables (i.e., the spread of residuals is consistent).
- **No Multicollinearity:** The independent variables are not highly correlated with each other.

Logistic Regression

Q.1 How does logistic regression differ from linear regression?

Linear Regression: Linear regression is used for predicting a continuous dependent variable based on one or more independent variables. It models the relationship between the independent variables and the dependent variable using a straight line.

Logistic Regression: Logistic regression is used for predicting the probability of a categorical outcome (binary or multinomial) based on one or more independent variables. It models the relationship between the independent variables and the probability of a specific outcome using a logistic function.

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

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. It is defined as:

$\sigma(z) = 1/1 + e^{-z}$

where z is the input to the function.

The sigmoid function is the core of logistic regression. It transforms the linear combination of independent variables and their coefficients into a probability value between 0 and 1. The output of the sigmoid function represents the probability that a given observation belongs to a specific category.

The sigmoid function has an S-shaped curve, which allows logistic regression to model non-linear relationships between the independent variables and the probability of the outcome.

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

- **Accuracy:** The proportion of correctly classified instances out of the total instances.

- **Precision:** The proportion of true positive predictions out of all positive predictions. It measures the accuracy of positive predictions.

- **Recall (Sensitivity):** The proportion of true positive predictions out of all actual positive instances. It measures the ability of the model to identify positive instances.

- **F1 Score:** The harmonic mean of precision and recall. It provides a balance between precision and recall.

- **ROC Curve (Receiver Operating Characteristic Curve):** A graphical representation of the trade-off between true positive rate (sensitivity) and false positive rate (1 - specificity) at various threshold settings.

- **AUC (Area Under the ROC Curve):** A measure of the model's ability to distinguish between positive and negative instances. A higher AUC indicates better performance.

Q4.How do you handle multicollinearity in logistic regression?

- **Feature Selection:** Remove highly correlated independent variables from the model.

- **Dimensionality Reduction:** Use techniques like principal component analysis (PCA) to reduce the dimensionality of the feature space and remove multicollinearity.

- **Regularization:** Apply regularization techniques such as L1 (Lasso) or L2 (Ridge) regularization to penalize large coefficients and reduce the impact of multicollinearity.

- **VIF (Variance Inflation Factor):** Check for multicollinearity using VIF and drop variables with high VIF values (> 5).

Q.1 What is the Naive Bayes algorithm based on?

The Naive Bayes algorithm is based on Bayes' theorem, which describes the probability of an event based on prior knowledge of conditions that might be related to the event.

It is a probabilistic classifier that calculates the probability of each class given a set of features and selects the class with the highest probability as the prediction.

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

Conditional probability is the probability of an event occurring given that another event has already occurred. In the context of Naive Bayes, conditional probability is used to calculate the probability of a class (target variable) given the values of the features (predictor variables).

Bayes' theorem states that the conditional probability of a class C given a set of features X can be calculated as:

$P(C \mid X) = P(X \mid C). P(C)/P(X)$

where:
- $P(C \mid X)$ is the posterior probability of class C given features X.
- $P(X \mid C)$ is the likelihood of observing features X given class C.
- P(C) is the prior probability of class C.
- P(X) is the probability of observing features X (the evidence).

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

Advantages

- Simple and easy to implement.
- Efficient and scalable, particularly with high-dimensional data.
- Performs well with categorical features and can handle a large number of features.
- Works well with small datasets and is less prone to overfitting.
- Can handle missing values gracefully.

Disadvantages

- Assumes independence between features, which is not always true in real-world data.
- Sensitivity to the quality of input data and the presence of irrelevant features.
- Unable to capture complex relationships between features.
- Prone to the "zero probability" problem if a categorical variable and class combination never occur together in the training data.

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

Missing Values: Naive Bayes can handle missing values by ignoring them during model training and prediction. It calculates probabilities based only on available data for features relevant to the prediction task.

Categorical Features: Naive Bayes naturally handles categorical features by estimating the probability of each class given the observed values of the features. It calculates the likelihood of observing a specific feature value within each class and combines these probabilities using Bayes' theorem to make predictions.

Decision Trees

Q.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 features.
At each node of the tree, it selects the feature that best separates the data into distinct classes or reduces impurity the most.
This process continues until a stopping criterion is met, such as reaching a maximum depth, having a minimum number of samples in a node, or when further splitting does not significantly improve the model's performance.
Once the tree is built, it follows the path from the root node to the leaf nodes based on the feature values of the input data, and the final decision is made at the leaf node.

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

- **Gini Impurity:** Measures the probability of incorrectly classifying a randomly chosen element in the dataset if it were randomly labeled. Nodes are split to minimize the Gini impurity, resulting in homogeneous subsets.

- **Entropy:** Measures the amount of uncertainty or disorder in a set of data. Nodes are split to maximize the information gain, which is the reduction in entropy after the split.

- **Classification Error:** Measures the probability of misclassification at a node. Nodes are split to minimize the classification error, resulting in the majority class being assigned to a subset.

Q.3.How do decision trees handle categorical variables?

Decision trees handle categorical variables by splitting the data based on the categories of the variables. For binary categorical variables, the tree branches into two based on the presence or absence of the category. For multi-level categorical variables, the tree creates branches corresponding to each category. The splitting criterion evaluates the purity or impurity of subsets created by splitting categorical variables.

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

- **Pruning:** Pruning involves removing parts of the tree that do not provide significant predictive power on the validation dataset. This helps prevent the tree from becoming too complex and overfitting the training data.

- **Limiting Tree Depth:** Restricting the maximum depth of the tree prevents it from growing too deep and capturing noise or irrelevant patterns in the data.

- **Minimum Sample Split:** Setting a minimum number of samples required to split a node ensures that only nodes with sufficient data are split, reducing the chance of overfitting.

- **Minimum Leaf Samples:** Similar to minimum sample split, setting a minimum number of samples required to be at a leaf node prevents the tree from creating leaf nodes with too few samples.

SVM

Q.1.What is the basic idea behind SVM?

The basic idea behind Support Vector Machine (SVM) is to find the optimal hyperplane that best separates the data points into different classes in a high-dimensional space.

SVM aims to maximize the margin, which is the distance between the hyperplane and the nearest data points from each class, known as support vectors.

By maximizing the margin, SVM aims to find the decision boundary that generalizes well to unseen data and has better classification performance.

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

Margin: The margin is the distance between the decision boundary (hyperplane) and the nearest data points from each class. The goal of SVM is to find the hyperplane that maximizes this margin.

Support Vectors: Support vectors are the data points that lie closest to the decision boundary and have a non-zero contribution to defining the hyperplane. They are the critical data points that determine the position and orientation of the decision boundary.

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

- **Linear Kernel:** The linear kernel is the simplest kernel function, which represents the dot product between the feature vectors in the original feature space. It is suitable for linearly separable data.

- **Polynomial Kernel:** The polynomial kernel computes the similarity between two feature vectors using a polynomial function. It is useful for capturing non-linear relationships in the data.

- **RBF (Radial Basis Function) Kernel:** The RBF kernel measures the similarity between two feature vectors based on the Gaussian (radial basis) function. It is suitable for non-linearly separable data and provides more flexibility in capturing complex patterns.

- **Sigmoid Kernel:** The sigmoid kernel computes the similarity between feature vectors using a sigmoid function. It can be useful for binary classification problems.

Q.4.How does SVM handle outliers?

SVM is robust to outliers to some extent due to its focus on maximizing the margin. Outliers that lie far away from the decision boundary may have little influence on the position of the hyperplane. However, if outliers significantly affect the margin or decision boundary, SVM may produce suboptimal results.

To handle outliers in SVM, preprocessing techniques such as outlier detection and removal, data normalization, or feature scaling can be applied to make the algorithm more robust to outliers.

Additionally, tuning the regularization parameter C in SVM can help control the impact of outliers on the decision boundary. Increasing C allows for more flexibility in fitting the data, while decreasing C penalizes misclassifications more heavily, potentially reducing the influence of outliers.