**Crash Course in Statistical Learning Quiz Questions**

Manimanya Reddy - 002763582

1. What is cross-validation in statistical learning?

a) The process of dividing the data into training and test sets.

b) The process of evaluating the performance of a model using different subsets of the data.

c) The process of selecting the optimal hyperparameters for a model.

d) The process of fitting the model to the data using an optimization algorithm.

Answer: b) The process of evaluating the performance of a model using different subsets of the data.

Explanation: Cross-validation is a statistical method used to evaluate the performance of a model by using different subsets of the data. In this method, the available data is divided into two sets, training and testing sets. The model is then trained on the training set and tested on the testing set. This process is repeated multiple times, with different subsets of the data being used for training and testing each time. The average performance of the model over all the different subsets of data is used to evaluate its performance. Cross-validation is an important technique in statistical learning as it helps to reduce the risk of overfitting the model to the training data.

1. What is the difference between L1 and L2 regularization?

a) L1 regularization adds a penalty proportional to the absolute value of the coefficients, while L2 regularization adds a penalty proportional to the squared value of the coefficients.

b) L1 regularization adds a penalty proportional to the squared value of the coefficients, while L2 regularization adds a penalty proportional to the absolute value of the coefficients.

c) L1 regularization is only used for linear models, while L2 regularization can be used for both linear and non-linear models.

d) L1 regularization is more computationally efficient than L2 regularization.

Answer: a) L1 regularization adds a penalty proportional to the absolute value of the coefficients, while L2 regularization adds a penalty proportional to the squared value of the coefficients.

Explanation: L1 and L2 regularization are techniques used to prevent overfitting in linear regression models. L1 regularization adds a penalty term to the loss function that is proportional to the absolute value of the coefficients, while L2 regularization adds a penalty term that is proportional to the squared value of the coefficients. The effect of this penalty term is to shrink the coefficients towards zero, which can help to reduce the variance of the model and prevent overfitting.

1. What is the difference between a confusion matrix and a ROC curve?

a) A confusion matrix is a plot of the true positive rate versus the false positive rate, while a ROC curve is a table that summarizes the performance of a classifier.

b) A confusion matrix is a table that summarizes the performance of a classifier, while a ROC curve is a plot of the true positive rate versus the false positive rate.

c) A confusion matrix and a ROC curve are the same thing.

d) A confusion matrix and a ROC curve are both used for clustering tasks.

Answer: b) A confusion matrix is a table that summarizes the performance of a classifier, while a ROC curve is a plot of the true positive rate versus the false positive rate.

Explanation: A confusion matrix and a ROC curve are both tools used to evaluate the performance of a classifier. A confusion matrix is a table that summarizes the performance of a classifier by showing the number of true positives, true negatives, false positives, and false negatives. A ROC (Receiver Operating Characteristic) curve, on the other hand, is a plot of the true positive rate (TPR) versus the false positive rate (FPR) at different classification thresholds. A ROC curve can be used to evaluate the performance of a classifier by comparing the tradeoff between TPR and FPR at different thresholds.

1. What is the difference between a decision tree and a random forest?

a) A decision tree is a single tree that makes predictions based on a set of rules, while a random forest is an ensemble of decision trees that makes predictions based on the average prediction of the individual trees.

b) A decision tree is an ensemble of trees that makes predictions based on the average prediction of the individual trees, while a random forest is a single tree that makes predictions based on a set of rules.

c) A decision tree and a random forest are the same thing.

d) A decision tree and a random forest are both unsupervised learning techniques.

Answer: a) A decision tree is a single tree that makes predictions based on a set of rules, while a random forest is an ensemble of decision trees that makes predictions based on the average prediction of the individual trees.

Explanation: A decision tree is a type of model that is used to make predictions by recursively splitting the data into smaller subsets based on a set of rules. Each internal node of the tree represents a decision based on a feature, while each leaf node represents a prediction. A random forest, on the other hand, is an ensemble of decision trees that makes predictions based on the average prediction of the individual trees. In a random forest, each tree is trained on a subset of the data and a random subset of the features. The final prediction is then made by averaging the predictions of all the trees in the forest.

1. What is the difference between a support vector machine and a neural network?

a) A support vector machine is a linear model that separates the classes using a hyperplane, while a neural network is a non-linear model that uses multiple layers of neurons to learn complex relationships in the data.

b) A neural network is a linear model that separates the classes using a hyperplane, while a support vector machine is a non-linear model that uses multiple layers of neurons to learn complex relationships in the data.

c) A support vector machine and a neural network are the same thing.

d) A support vector machine and a neural network are both unsupervised learning techniques.

Answer: a) A support vector machine is a linear model that separates the classes using a hyperplane, while a neural network is a non-linear

Explanation: A decision tree is a type of model that is used to make predictions by recursively splitting the data into smaller subsets based on a set of rules. Each internal node of the tree represents a decision based on a feature, while each leaf node represents a prediction. A random forest, on the other hand, is an ensemble of decision trees that makes predictions based on the average prediction of the individual trees. In a random forest, each tree is trained on a subset of the data and a random subset of the features. The final prediction is then made by averaging the predictions of all the trees in the forest.

1. What is the curse of dimensionality in statistical learning?

a) The curse of dimensionality refers to the difficulty of visualizing high-dimensional data.

b) The curse of dimensionality refers to the difficulty of finding a good approximation of the true model in high-dimensional data.

c) The curse of dimensionality refers to the difficulty of finding a good balance between bias and variance in high-dimensional data.

d) The curse of dimensionality refers to the phenomenon that the number of training samples required to learn a good model increases exponentially with the number of dimensions.

Answer: d) The curse of dimensionality refers to the phenomenon that the number of training samples required to learn a good model increases exponentially with the number of dimensions.

Explanation- The curse of dimensionality in statistical learning refers to the phenomenon that the number of training samples required to learn a good model increases exponentially with the number of dimensions. This is due to the fact that the volume of the input space increases exponentially with the number of dimensions, making it harder to find relevant patterns and relationships in the data. As the number of dimensions increases, the amount of available training data becomes less and less representative of the input space, leading to overfitting and poor generalization performance.

1. What is the difference between a parametric model and a non-parametric model?

a) A parametric model has a fixed number of parameters, while a non-parametric model has a flexible number of parameters.

b) A parametric model has a flexible number of parameters, while a non-parametric model has a fixed number of parameters.

c) A parametric model is a linear model, while a non-parametric model is a non-linear model.

d) A parametric model is used for classification tasks, while a non-parametric model is used for regression tasks.

Answer: a) A parametric model has a fixed number of parameters, while a non-parametric model has a flexible number of parameters.

Explanation- The difference between a parametric model and a non-parametric model is that a parametric model has a fixed number of parameters, while a non-parametric model has a flexible number of parameters. A parametric model makes assumptions about the functional form of the relationship between inputs and outputs, and the number of parameters is fixed by the choice of model architecture. In contrast, a non-parametric model does not make any assumptions about the functional form of the relationship, and the number of parameters is determined by the complexity of the data. Non-parametric models are often more flexible and can capture more complex relationships in the data, but they require more data to avoid overfitting.

1. What is the difference between unsupervised learning and supervised learning?

a) In unsupervised learning, the data is not labeled, while in supervised learning, the data is labeled.

b) In unsupervised learning, the goal is to find patterns in the data, while in supervised learning, the goal is to make predictions based on the data.

c) Unsupervised learning is only used for regression tasks, while supervised learning is only used for classification tasks.

d) Unsupervised learning and supervised learning are the same thing.

Answer: b) In unsupervised learning, the goal is to find patterns in the data, while in supervised learning, the goal is to make predictions based on the data.

Explanation- The difference between unsupervised learning and supervised learning is that in unsupervised learning, the goal is to find patterns in the data, while in supervised learning, the goal is to make predictions based on the data. In unsupervised learning, the data is not labeled, so the algorithm must find patterns and relationships in the data without any knowledge of what the patterns represent. In supervised learning, the data is labeled with the correct output values, and the algorithm learns to predict these values from the inputs. Supervised learning is used for both classification and regression tasks, while unsupervised learning is typically used for clustering, dimensionality reduction, and other data exploration tasks.

1. What is the difference between a linear model and a non-linear model?

a) A linear model uses a linear equation to make predictions, while a non-linear model uses a non-linear equation to make predictions.

b) A linear model is only used for classification tasks, while a non-linear model is only used for regression tasks.

c) A linear model has a fixed number of parameters, while a non-linear model has a flexible number of parameters.

d) A linear model is more complex than a non-linear model.

Answer: a) A linear model uses a linear equation to make predictions, while a non-linear model uses a non-linear equation to make predictions.

Explanation - A linear model is a model that assumes a linear relationship between the independent variables and the dependent variable, where the output is a linear combination of the input features. This means that the model uses a linear equation to make predictions. In contrast, a non-linear model does not make this assumption and can capture more complex relationships between the input features and output. Non-linear models can use more complex functions such as polynomials or neural networks to make predictions.

10. What is the difference between a generative model and a discriminative model?

a) A generative model models the joint probability distribution of the features and the target variable, while a discriminative model models the conditional probability distribution of the target variable given the features.

b) A discriminative model models the joint probability distribution of the features and the target variable, while a generative model models the conditional probability distribution of the target variable given the features.

c) A generative model is only used for regression tasks, while a discriminative model is only used for classification tasks.

d) A discriminative model is more complex than a generative model.

Answer: a) A generative model models the joint probability distribution of the features and the target variable, while a discriminative model models the conditional probability distribution of the target variable given the features.

Explanation - A generative model is a model that models the joint probability distribution of the input features and the target variable, meaning it learns the underlying probability distribution of the data. In contrast, a discriminative model only models the conditional probability distribution of the target variable given the input features, meaning it learns the boundary between the classes. Generative models can be used for both classification and regression tasks, while discriminative models are typically used for classification tasks. Generative models can also be used to generate new data samples that are similar to the training data, while discriminative models cannot.