Dashboard Courses PW Skills Lab Job Portal Experience Portal Become an affiliate Hall of Fame KHUSHI

Quiz -5

19 out of 20 correct

1. In the Naive Approach, feature independence is assumed. What does this mean?

● Features are unrelated to each other

○ Features are dependent on each other

○ Features are normally distributed

○ Features are linearly related

Explanation: In the Naive Approach, it is assumed that the features are independent of each other. This assumption simplifies the model and allows the probability of each feature to be estimated separately

2. Which of the following is NOT an application of KNN?

○ Classification

○ Regression

○ Anomaly detection

● Dimensionality reduction

Explanation: KNN is primarily used for classification and regression tasks. It is not typically used for dimensionality reduction.

3. Which clustering algorithm is sensitive to the initial choice of cluster centers?

● K-means clustering

○ Hierarchical clustering

○ DBSCAN

○ Mean Shift clustering

Explanation: K-means clustering is sensitive to the initial placement of cluster centers. Different initializations can lead to different clustering results.

4. Which algorithm is commonly used for anomaly detection?

○ K-means clustering

● K-nearest neighbors (KNN)

○ Support Vector Machines (SVM)

○ Principal Component Analysis (PCA)

Explanation: KNN can be used for anomaly detection by measuring the distance of a data point to its nearest neighbors. Unusually distant points can be identified as anomalies.

5. Which technique is used for reducing the dimensionality of a dataset?

● Principal Component Analysis (PCA)

○ K-means clustering

○ Random Forests

○ Support Vector Machines (SVM)

Explanation: PCA is a widely used technique for dimensionality reduction. It transforms the original features into a new set of orthogonal features called principal components.

6. Which technique is used to select the most important features in a dataset?

○ PCA

○ K-means clustering

● Feature selection

○ Anomaly detection

Explanation: Feature selection is the process of selecting the most relevant features from a dataset. It aims to reduce dimensionality and improve model performance by focusing on the most informative features.

7. What is the purpose of data drift detection?

○ To identify anomalies in the dataset

○ To prevent data leakage

● To monitor changes in the data distribution over time

○ To measure the impact of feature selection

Explanation: Data drift detection helps to identify changes in the underlying data distribution, which can affect the performance of machine learning models. It is important to monitor and adapt models to changing data conditions.

8. What is data leakage in machine learning?

○ Unintentional disclosure of sensitive data

○ Unreliable data sources

○ Inconsistent labeling of data samples

● Incorporating information from the future into the training process

Explanation: Data leakage refers to the situation when information from the future or outside the training set is inadvertently used during model training, leading to overly optimistic performance estimates

9. Which technique is used for preventing data leakage in machine learning?

○ Feature selection

● Cross-validation

○ Dimensionality reduction

○ Anomaly detection

Explanation: Cross-validation is a technique used to evaluate the performance of a machine learning model on unseen data. It helps prevent overfitting and provides a more reliable estimate of model performance.

10. Which evaluation technique is used to assess the performance of a machine learning model on unseen data?

● Cross-validation

○ Feature selection

○ Anomaly detection

○ Data leakage detection

Explanation: Cross-validation is used to assess the performance of a machine learning model on unseen data. It involves splitting the data into multiple subsets, training and evaluating the model on different subsets, and averaging the performance metrics.

11. Which of the following is an unsupervised learning algorithm used for anomaly detection?

○ Decision Tree

○ Random Forest

● Isolation Forest

○ Gradient Boosting

Explanation: Isolation Forest is an unsupervised learning algorithm that identifies anomalies by isolating them into separate regions of a random partition tree.

12. Which dimensionality reduction technique aims to preserve the pairwise distances between data points?

○ Principal Component Analysis (PCA)

○ Linear Discriminant Analysis (LDA)

● t-SNE

○ Singular Value Decomposition (SVD)

Explanation: t-SNE (t-Distributed Stochastic Neighbor Embedding) is a dimensionality reduction technique that aims to preserve the pairwise distances between data points in the lower-dimensional space.

13. Which feature selection technique uses statistical tests to evaluate the significance of each feature?

○ Recursive Feature Elimination (RFE)

○ Mutual Information

● Chi-square test

○ Lasso regularization

Explanation: The Chi-square test is a feature selection technique that uses statistical tests to evaluate the significance of each feature's association with the target variable.

14. What is the purpose of the Data Drift Detection technique?

● To detect changes in the data distribution over time

○ To identify outliers in the dataset

○ To reduce the dimensionality of the dataset

○ To select the most important features in the dataset

Explanation: Data Drift Detection is a technique used to monitor and detect changes in the statistical properties of the data over time, such as shifts in means, variances, or distributions.

15. Which technique can be used to prevent Data Leakage in machine learning?

○ Proper data cleaning and preprocessing

○ Implementing robust feature selection methods

● Applying strict privacy and security measures

○ Following strict model deployment and monitoring protocols

Explanation: Proper model deployment and monitoring protocols help prevent unintentional data leakage by ensuring that the model only uses information available at the time of prediction and does not rely on future or external data.

16. Which evaluation technique is used to estimate the performance of a machine learning model on unseen data?

○ Data Leakage Detection

○ Feature Selection

● Cross-Validation

○ Anomaly Detection

Explanation: Cross-Validation is an evaluation technique where the data is divided into subsets, and the model is trained and evaluated on different subsets to estimate its performance on unseen data.

17. Which technique is used to detect potential data leakage in machine learning pipelines?

● Data Leakage Detection

○ Cross-Validation

○ Anomaly Detection

○ Feature Selection

Explanation: Data Leakage Detection involves carefully examining the data and the steps in the machine learning pipeline to identify potential sources of data leakage, such as incorporating information from the future or using information that should not be available at prediction time.

18. What is the purpose of Cross-Validation in machine learning?

● To prevent overfitting and assess model generalization

○ To identify outliers in the dataset

○ To reduce the dimensionality of the dataset

○ To select the most important features in the dataset

Explanation: Cross-Validation is used to assess the performance and generalization ability of a machine learning model by training and evaluating it on multiple subsets of the data. It helps detect overfitting and provides a more reliable estimate of model performance.

19. Which of the following techniques can be used to handle imbalanced datasets?

○ Oversampling the minority class

○ Undersampling the majority class

○ Using appropriate evaluation metrics (e.g., F1-score)

● All of the above

Explanation: To handle imbalanced datasets, various techniques can be employed, such as oversampling the minority class, undersampling the majority class, and using appropriate evaluation metrics that consider both precision and recall, such as the F1-score

20. Which of the following clustering algorithms is density-based and capable of discovering clusters of arbitrary shapes?

○ K-means clustering

○ Hierarchical clustering

● DBSCAN

○ Mean Shift clustering

Explanation: DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that is capable of discovering clusters of arbitrary shapes. It groups together data points that are close to each other in density-connected regions.

Submit