Machine Learning

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

   Supervised learning involves training a model on labeled data, where the input data is paired with the correct output. The model learns to map inputs to outputs and can be used for tasks such as classification and regression. Unsupervised learning involves training a model on unlabeled data to identify patterns or structures within the data. It is often used for clustering and dimensionality reduction.

2. **Explain the purpose of training and testing datasets in the context of building a machine learning model.**

   The training dataset is used to train the machine learning model by allowing it to learn the patterns and relationships within the data. The testing dataset is used to evaluate the model's performance on unseen data to ensure it generalizes well and does not overfit to the training data.

3. **What is overfitting, and how can it be prevented?**

   Overfitting occurs when a model learns the noise or random fluctuations in the training data rather than the underlying patterns, resulting in poor performance on new, unseen data. It can be prevented by using techniques such as cross-validation, regularization, pruning, and ensuring the model has sufficient training data.

4. **What is cross-validation?**

   Cross-validation is a technique used to evaluate the performance of a machine learning model by dividing the data into multiple subsets or folds. The model is trained on some of these folds and tested on the remaining ones. This process is repeated multiple times to ensure that the model's performance is consistent and reliable.

5. **Explain the bias-variance tradeoff.**

   The bias-variance tradeoff refers to the balance between the bias (error due to overly simplistic models) and variance (error due to overly complex models) of a machine learning model. High bias can lead to underfitting, while high variance can lead to overfitting. The goal is to find a model that minimizes both bias and variance.

6. **What is a confusion matrix?**

   A confusion matrix is a table used to evaluate the performance of a classification model. It shows the number of true positives, false positives, true negatives, and false negatives, allowing for the calculation of various performance metrics such as accuracy, precision, recall, and F1 score.

7. **What is precision and recall?**

   Precision measures the proportion of true positive predictions out of all positive predictions made by the model (i.e., TP / (TP + FP)). Recall measures the proportion of true positive predictions out of all actual positives in the data (i.e., TP / (TP + FN)).

8. **What is the purpose of feature scaling?**

   Feature scaling is used to standardize the range of features in the data, so they contribute equally to the model's performance. This is particularly important for algorithms that rely on distance calculations, such as k-nearest neighbors and gradient descent-based methods.

9. **Explain the difference between classification and regression.**

   Classification is a task where the goal is to assign input data into discrete categories or classes (e.g., spam vs. non-spam emails). Regression is a task where the goal is to predict a continuous numerical value (e.g., predicting house prices).

10. **What is a decision tree?**

    A decision tree is a supervised learning algorithm that models decisions and their possible consequences in a tree-like structure. Each internal node represents a decision based on a feature, each branch represents an outcome, and each leaf node represents a class label or value.

11. **What is an ensemble method?**

    Ensemble methods combine the predictions of multiple models to improve performance. Examples include bagging (e.g., Random Forests) and boosting (e.g., Gradient Boosting), which aggregate the results from multiple models to enhance accuracy and reduce overfitting.

12. **What is a support vector machine (SVM)?**

    A support vector machine is a supervised learning algorithm used for classification and regression tasks. It works by finding the hyperplane that best separates the data into different classes, with the maximum margin between the classes.

13. **What is the role of the kernel function in SVM?**

    The kernel function in SVM transforms the input data into a higher-dimensional space where a linear separator may be more effective. Common kernel functions include the linear, polynomial, and radial basis function (RBF) kernels.

14. **What is a neural network?**

    A neural network is a computational model inspired by the human brain, consisting of layers of interconnected nodes (neurons). It is used for a variety of tasks including classification, regression, and feature extraction.

15. **What is deep learning?**

    Deep learning is a subset of machine learning that involves training neural networks with many layers (deep neural networks). It is used for complex tasks such as image recognition, natural language processing, and autonomous driving.

16. **What is the difference between bagging and boosting?**

    Bagging (Bootstrap Aggregating) combines the predictions of multiple models trained on different subsets of the training data to reduce variance. Boosting sequentially trains models, where each new model attempts to correct the errors of the previous ones, to reduce bias.

17. **What is a random forest?**

    A random forest is an ensemble learning method that combines multiple decision trees to improve accuracy and robustness. It works by averaging the predictions of multiple trees to reduce overfitting and improve generalization.

18. **What is the k-nearest neighbors (k-NN) algorithm?**

    The k-nearest neighbors algorithm is a non-parametric supervised learning method used for classification and regression. It classifies data points based on the majority class or average value of their k-nearest neighbors in the feature space.

19. **What is principal component analysis (PCA)?**

    Principal component analysis is a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional form while retaining as much variance as possible. It identifies the principal components that capture the most variance in the data.

20. **What is the purpose of regularization in machine learning?**

    Regularization is used to prevent overfitting by adding a penalty to the model's complexity. Common regularization techniques include L1 (Lasso) and L2 (Ridge) regularization, which add penalties based on the magnitude of the model's coefficients.

21. **What is the difference between L1 and L2 regularization?**

    L1 regularization (Lasso) adds a penalty proportional to the absolute values of the coefficients, encouraging sparsity (some coefficients become zero). L2 regularization (Ridge) adds a penalty proportional to the squared values of the coefficients, discouraging large coefficients but not necessarily forcing them to zero.

22. **What is the ROC curve?**

    The ROC (Receiver Operating Characteristic) curve is a graphical representation of a classifier's performance across different threshold values. It plots the true positive rate (sensitivity) against the false positive rate (1-specificity), helping to evaluate the trade-offs between sensitivity and specificity.

23. **What is the AUC-ROC score?**

    The AUC-ROC score (Area Under the ROC Curve) measures the overall performance of a classification model. It represents the probability that the model will correctly rank a randomly chosen positive instance higher than a randomly chosen negative instance.

24. **What is gradient descent?**

    Gradient descent is an optimization algorithm used to minimize the cost function of a machine learning model by iteratively adjusting the model's parameters in the direction of the steepest decrease in the cost function.

25. **What is a loss function?**

    A loss function is a mathematical function that measures the difference between the predicted values and the actual values in a model. It quantifies the model's prediction error and guides the optimization process.

26. **What is a hyperparameter in machine learning?**

    A hyperparameter is a parameter whose value is set before the learning process begins and controls the training process of a machine learning model. Examples include the learning rate, number of layers in a neural network, and regularization strength.

27. **What is the difference between a parameter and a hyperparameter?**

    Parameters are the internal variables of a machine learning model that are learned from the training data (e.g., weights in a neural network). Hyperparameters are external variables set before training begins that control the training process and model architecture.

28. **What is the purpose of feature selection?**

    Feature selection involves choosing a subset of relevant features from the original feature set to improve the model's performance, reduce overfitting, and decrease computational complexity. It helps to focus on the most informative features.

29. **What is feature engineering?**

    Feature engineering is the process of creating new features or modifying existing ones to improve the performance of a machine learning model. It involves transforming raw data into meaningful features that can better represent the underlying patterns.

30. **What is the difference between parametric and non-parametric models?**

    Parametric models assume a specific form for the underlying data distribution and have a fixed number of parameters (e.g., linear regression). Non-parametric models do not assume a fixed form and can adapt to the data's complexity (e.g., k-nearest neighbors).