1. What is the Central Limit Theorem?

Answer: The Central Limit Theorem states that, regardless of the shape of the population distribution, the sampling distribution of the mean tends to be normally distributed with a large enough sample size.

2. What is the difference between Type I and Type II errors?

Answer: Type I error is rejecting a true null hypothesis, while Type II error is failing to reject a false null hypothesis.

3. What is regularization in machine learning?

Answer: Regularization is a technique used to prevent overfitting by adding a penalty term to the model's objective function, encouraging simpler models.

4. Explain the difference between bagging and boosting.

Answer: Bagging involves training multiple models independently and averaging their predictions, while boosting trains models sequentially, with each subsequent model trying to correct the mistakes of the previous models.

5. What is the curse of dimensionality?

Answer: The curse of dimensionality refers to the challenges and issues that arise when working with high-dimensional data, including increased computational complexity and the sparsity of data in high-dimensional space.

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

Answer: Unsupervised learning involves finding patterns or structures in data without labeled outputs, while supervised learning learns from labeled data to make predictions or classifications.

7. Explain precision and recall.

Answer: Precision measures the accuracy of positive predictions, while recall measures the proportion of true positives correctly identified.

8. What is the purpose of cross-validation?

Answer: Cross-validation is used to assess the performance and generalization of a model by partitioning the data into multiple subsets for training and evaluation.

9. What is the difference between variance and bias?

Answer: Variance measures the variability of model predictions, while bias measures the error between the predicted values and the true values.

10. What is the difference between clustering and classification?

Answer: Clustering is an unsupervised learning task that groups similar data points together, while classification is a supervised learning task that assigns labels to data points based on their features.

11. What is feature engineering?

Answer: Feature engineering involves creating or transforming input features to improve the performance and interpretability of machine learning models.

12. What is the purpose of A/B testing?

Answer: A/B testing is used to compare two versions (A and B) of a webpage or application to determine which one performs better based on user behavior or metrics.

13. What is the difference between overfitting and underfitting?

Answer: Overfitting occurs when a model learns too much from the training data and fails to generalize well to unseen data, while underfitting occurs when a model is too simple and fails to capture the underlying patterns in the data.

14. What is feature selection?

Answer: Feature selection is the process of selecting a subset of relevant features from the available input features to improve model performance and reduce complexity.

15. What is the purpose of a validation set?

Answer: A validation set is used to tune model hyperparameters and assess the generalization performance before applying the model to unseen data.

16. Explain the bias-variance tradeoff.

Answer: The bias-variance tradeoff refers to the relationship between the model's flexibility (variance) and its ability to capture the true underlying patterns (bias). Increasing model complexity reduces bias but increases variance, and vice versa.

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

Answer: L1 regularization (Lasso) encourages sparsity by adding the absolute value of the coefficients as a penalty term, while L2 regularization (Ridge) adds the squared value of the coefficients.

18. What is the purpose of the ROC curve?

Answer: The Receiver Operating Characteristic (ROC) curve is used to evaluate and visualize the performance of a binary classification model by plotting the true positive rate against the false positive rate at different classification thresholds.

19. What is the difference between batch gradient descent and stochastic gradient descent?

Answer: Batch gradient descent updates the model parameters using the entire training dataset, while stochastic gradient descent updates the parameters for each individual training example.

20. Explain the term "p-value" in hypothesis testing.

Answer: The p-value is the probability of observing the given test statistic (or more extreme) under the null hypothesis. It is used to determine the statistical significance of the results.

21. What is the purpose of dimensionality reduction techniques like PCA?

Answer: Dimensionality reduction techniques like Principal Component Analysis (PCA) are used to reduce the number of input features while retaining the most important information, capturing the variability in the data.

22. What is the difference between stratified sampling and random sampling?

Answer: Stratified sampling involves dividing the population into homogeneous subgroups and then randomly sampling from each subgroup, while random sampling involves selecting individuals randomly from the entire population.

23. What is the purpose of outlier detection in data analysis?

Answer: Outlier detection is used to identify and handle data points that deviate significantly from the majority of the data, which can affect the model's performance and accuracy.

24. Explain the concept of ensemble learning.

Answer: Ensemble learning combines multiple individual models (e.g., decision trees, neural networks) to make predictions or classifications, often resulting in improved performance and generalization.

25. What is the difference between precision and accuracy?

Answer: Precision measures the correctness of positive predictions relative to all positive predictions, while accuracy measures the correctness of all predictions (both positive and negative) relative to the entire dataset.