

Advanced Data Science Concepts

Data Science · Answer Key · 20 Questions

1. Which statistical measure is most resilient to the influence of outliers in a skewed dataset?

- A) Arithmetic Mean
- B) Median**
- C) Mode
- D) Range

2. In the context of supervised learning, what does the 'Bias-Variance Tradeoff' describe?

- A) The relationship between model complexity and error**
- B) The speed of CPU vs GPU processing
- C) The difference between training data and validation data
- D) The cost of data storage vs retrieval speed

3. What is the primary objective of applying 'Normalization' to numerical input features in a machine learning model?

- A) To remove missing values
- B) To increase the number of dimensions
- C) To scale values to a standard range, usually 0 to 1**
- D) To convert categorical labels into integers

4. Which term describes the phenomenon where a model performs exceptionally well on training data but fails to generalize to new, unseen data?

- A) Underfitting
- B) Gradient Descent
- C) Overfitting**
- D) Feature Selection

5. What is the 'Curse of Dimensionality' in data science?

- A) The difficulty of visualizing data in 3D
- B) The exponential increase in data volume as dimensions grow
- C) The requirement for more data as the number of features increases to maintain statistical significance**
- D) The inability of computers to process high-resolution images

6. In a confusion matrix for a binary classifier, what does the 'False Negative' rate represent?

- A) Type I error
- B) Type II error**
- C) Accuracy
- D) Precision

7. Which algorithm is fundamentally based on the concept of maximizing the margin between classes in a vector space?

- A) K-Means Clustering
- B) Support Vector Machine**
- C) Linear Regression
- D) Decision Tree

8. What does the 'p-value' indicate in the context of hypothesis testing?

- A) The probability that the null hypothesis is true**
- B) The effect size of the experiment
- C) The significance level of the independent variable
- D) The correlation coefficient of the data

9. Which data preprocessing technique is used to handle categorical variables by creating binary columns for each category?

- A) Standardization
- B) One-Hot Encoding**
- C) Logarithmic Transformation
- D) Feature Aggregation

10. In SQL, which clause is used to filter aggregated data after a GROUP BY operation has been performed?

- A) WHERE
- B) HAVING**
- C) ORDER BY
- D) LIMIT

11. What is the primary purpose of cross-validation in model evaluation?

- A) To speed up training time
- B) To reduce the amount of training data needed
- C) To provide a more robust estimate of model performance on unseen data**
- D) To optimize the hyperparameters automatically

12. Which of the following is a non-parametric machine learning algorithm that classifies data points based on their proximity to neighbors?

- A) Logistic Regression
- B) K-Nearest Neighbors**
- C) Naive Bayes
- D) Random Forest

13. In the context of information theory, what metric is commonly used to measure the impurity of a node in a decision tree?

- A) Gini Impurity**
- B) Euclidean Distance
- C) Standard Deviation
- D) Manhattan Distance

14. Which mathematical operation is at the core of the backpropagation algorithm in neural networks?

- A) Integration
- B) Chain Rule of Calculus**
- C) Long Division
- D) Matrix Transposition

15. What does the 'F1-Score' represent in classification evaluation?

- A) The sum of precision and recall
- B) The difference between true positives and false positives
- C) The harmonic mean of precision and recall**
- D) The ratio of accuracy to error rate

16. Which type of regularization adds a penalty equal to the absolute value of the magnitude of coefficients, often leading to sparse models?

- A) L1 Regularization (Lasso)**
- B) L2 Regularization (Ridge)
- C) Elastic Net
- D) Dropout

17. What is the main advantage of using a 'Random Forest' over a single 'Decision Tree'?

- A) Reduced computational cost
- B) Improved interpretability
- C) Reduction of variance via bagging**
- D) Increased bias

18. Which statistical term refers to the correlation between two variables where the influence of a third variable is controlled?

A) Partial Correlation

B) Spurious Correlation

C) Covariance

D) Multicollinearity

19. In time-series analysis, what does 'Stationarity' imply about a dataset?

A) The mean and variance remain constant over time

B) The data exhibits a strong seasonal trend

C) The data has no missing values

D) The data is linearly distributed

20. Which encoding technique is best suited for ordinal categorical data where the order of categories matters?

A) Label Encoding

B) One-Hot Encoding

C) Dummy Encoding

D) Frequency Encoding