

Advanced Data Science Fundamentals

Data Science · Answer Key · 20 Questions

1. Which statistical method is specifically designed to handle multicollinearity by adding a penalty proportional to the square of the magnitude of coefficients?

- A) Lasso Regression
- B) Ridge Regression**
- C) Elastic Net
- D) Principal Component Regression

2. In the context of the Bias-Variance tradeoff, what does the term 'irreducible error' represent?

- A) The error remaining after model training
- B) The variance of the training dataset
- C) The noise inherent in the system**
- D) The bias of the chosen algorithm

3. Who is credited with inventing the term 'exploratory data analysis' in 1977?

- A) John Tukey**
- B) Ronald Fisher
- C) George Box
- D) C.R. Rao

4. Which algorithm is defined as a non-parametric method used for classification and regression that relies on the 'majority vote' of its constituent decision trees?

- A) Gradient Boosting Machines
- B) Random Forest**
- C) Support Vector Machines
- D) Naive Bayes

5. What is the primary mathematical objective of the Expectation-Maximization (EM) algorithm?

- A) Finding maximum likelihood estimates of parameters in latent variable models**
- B) Minimizing the residual sum of squares in linear models
- C) Reducing the dimensionality of high-rank matrices
- D) Calculating the exact global minimum of non-convex functions

6. In reinforcement learning, what does the Bellman equation describe?

- A) The probability distribution of hidden states
- B) The recursive relationship between the value of a state and its successor states**
- C) The gradient descent update rule for deep Q-networks
- D) The rate of decay for an exploration-exploitation strategy

7. What is the 'Curse of Dimensionality' primarily characterized by in high-dimensional feature spaces?

- A) Exponential increase in computational requirements for matrix inversion
- B) The volume of the space increases so rapidly that available data becomes sparse**
- C) A breakdown in the central limit theorem
- D) The inability to calculate eigenvectors for sparse matrices

8. Which distribution is the conjugate prior to the Bernoulli distribution in Bayesian statistics?

- A) Normal distribution
- B) Poisson distribution
- C) Beta distribution**
- D) Gamma distribution

9. What is the definition of a 'stationary process' in time series analysis?

- A) A process where the mean, variance, and autocorrelation structure are constant over time**
- B) A process that has undergone de-trending and seasonal adjustment
- C) A process with a unit root
- D) A process that cannot be forecasted

10. The Vapnik-Chervonenkis (VC) dimension is a measure of what property of a statistical classification algorithm?

- A) Training time complexity
- B) Model capacity or complexity**
- C) Convergence rate
- D) Number of hyper-parameters

11. Which specific technique is used to prevent overfitting by penalizing large weights in neural networks?

- A) Dropout
- B) Batch Normalization
- C) Weight Decay**
- D) Gradient Clipping

12. In the context of information theory, what does the Kullback-Leibler (KL) divergence measure?

- A) The mutual information between two random variables
- B) The difference between two probability distributions**
- C) The entropy of a single distribution
- D) The variance of a sample population

13. What is the primary advantage of using a 'Leaky ReLU' activation function over a standard ReLU?

- A) It eliminates the vanishing gradient problem entirely
- B) It prevents the 'dying ReLU' problem by allowing a small gradient when the input is negative**
- C) It is computationally cheaper to compute the derivative
- D) It ensures the output is normalized between 0 and 1

14. Which algorithm is most commonly associated with the solution to the 'Traveling Salesperson Problem' using biological inspiration?

- A) Ant Colony Optimization**
- B) K-Nearest Neighbors
- C) Apriori Algorithm
- D) Hidden Markov Models

15. In SQL, what is the fundamental difference between a 'LEFT JOIN' and an 'INNER JOIN' when the right table has no matching rows?

- A) The LEFT JOIN returns NULL for the right side, while INNER JOIN excludes the row**
- B) They are functionally identical in modern databases
- C) LEFT JOIN duplicates the row, INNER JOIN deletes it
- D) INNER JOIN returns NULL, LEFT JOIN returns 0

16. Which theorem states that the sum of a large number of independent and identically distributed random variables tends toward a normal distribution?

- A) Law of Large Numbers
- B) Central Limit Theorem**
- C) Bayes' Theorem
- D) Slutsky's Theorem

17. What is the 'Manhattan Distance' between two points (x1, y1) and (x2, y2) mathematically equivalent to?

- A) The Euclidean distance squared
- B) The L1 norm of the difference vector**
- C) The L2 norm of the difference vector
- D) The cosine similarity

18. What does the 'Support' metric represent in Association Rule Learning (e.g., in the Apriori algorithm)?

- A) The confidence of an implication
- B) The frequency of an itemset in the dataset**
- C) The lift ratio relative to independent occurrence
- D) The error rate of the classification

19. Which dimensionality reduction technique is based on identifying orthogonal axes that maximize the variance of the projected data?

A) Independent Component Analysis

B) Principal Component Analysis

C) Factor Analysis

D) Multidimensional Scaling

20. In Python's Scikit-learn, what is the default solver used for Logistic Regression for small datasets?

A) liblinear

B) sag

C) lbfgs

D) newton-cg