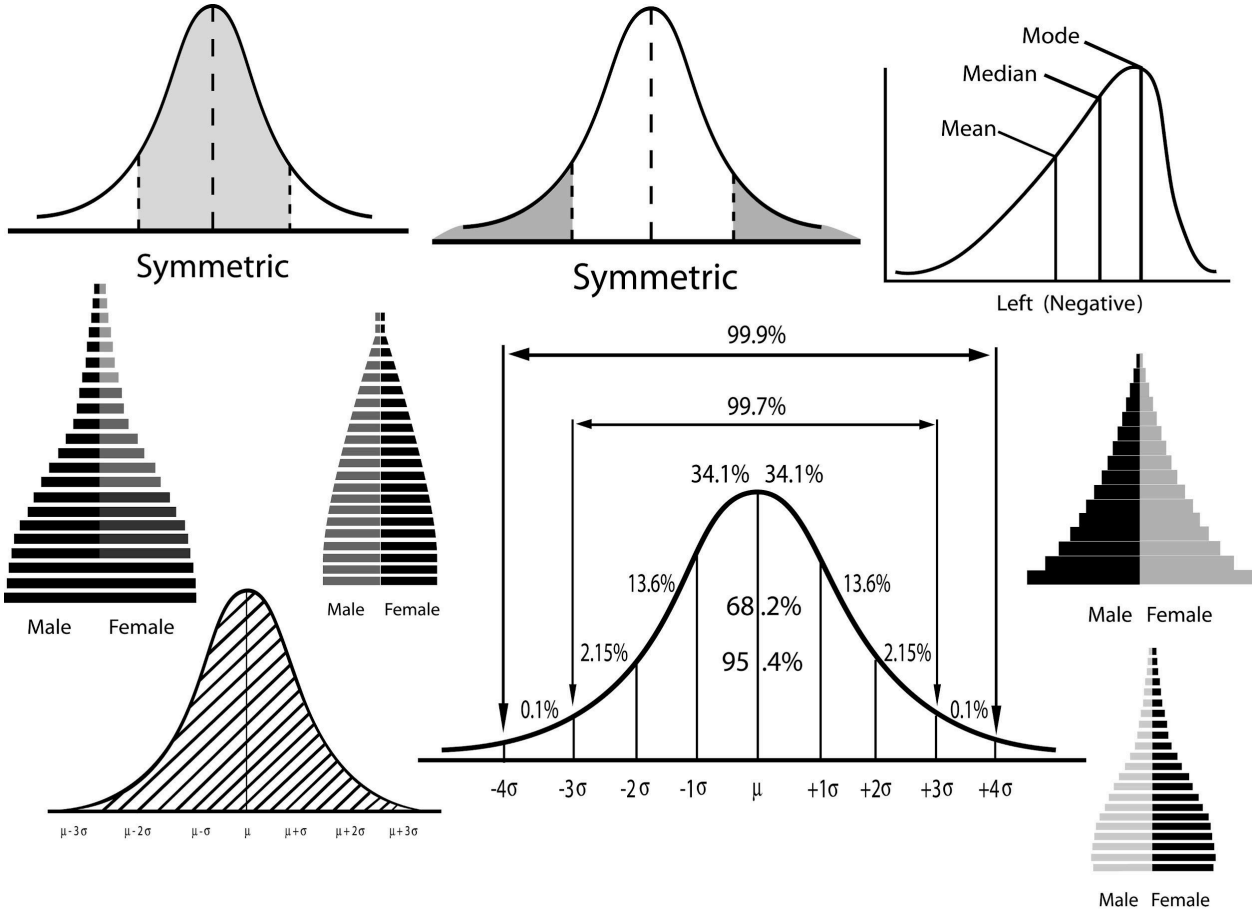
Comprehensive Data Science & Machine Learning Interview Question Bank

Part 1: Statistics, Probability, and Core Data Concepts

Section 1: Statistics & Probability

1. What is the Central Limit Theorem (CLT) and why is it important? Answer: The CLT states that the sampling distribution of the sample mean approaches a normal distribution as the sample size gets larger, regardless of the shape of the population distribution. This is crucial because it allows us to apply statistical techniques that assume normality (like hypothesis testing and confidence intervals) even if the original data is not normally distributed, provided the sample size is sufficient (usually n>30).



POPULATION PYRAMID SHAPES

Shutterstock

2. Explain the p-value to a non-technical stakeholder. Answer: The p-value is a measure of evidence against a null hypothesis. Imagine we want to test if a new medicine works. The "null hypothesis" assumes it doesn't work. The p-value tells us the probability of seeing results as strong as ours if the medicine didn't actually work. A p-value of 0.05 means there is only a 5% chance the results happened by random luck. If it's lower than that (e.g., < 0.05), we conclude the medicine likely works.

3. What is the difference between Type I and Type II errors? Answer:

- **Type I Error (False Positive):** Rejecting a true Null Hypothesis. *Example: Convicting an innocent person.*
- **Type II Error (False Negative):** Failing to reject a false Null Hypothesis. *Example: Letting a guilty person go free.*

4. Explain the Bias-Variance Tradeoff. Answer: It describes the balance between two sources of error in models.

- **Bias:** Error from erroneous assumptions in the learning algorithm (e.g., using a linear model for complex data). High bias leads to underfitting.
- **Variance:** Error from sensitivity to small fluctuations in the training set. High variance leads to overfitting.
- **Goal:** To find the optimal balance where total error is minimized.

5. What is the difference between Correlation and Covariance? Answer:

- **Covariance:** Indicates the direction of the linear relationship between variables (Positive means they move together, negative means inversely). The magnitude is not standardized.
- **Correlation:** Measures both the strength and direction of the linear relationship. It is a standardized version of covariance, ranging from -1 to 1.

6. What is A/B Testing? Answer: A/B testing is a randomized experiment with two variants, A (control) and B (treatment). It is used to compare two versions of a single variable (like a web page or marketing email) to determine which one performs better on a specific metric (like click-through rate).

7. What is the Law of Large Numbers? Answer: It states that as the size of a sample increases, the sample mean gets closer to the population mean. It ensures that larger datasets yield more reliable statistics.

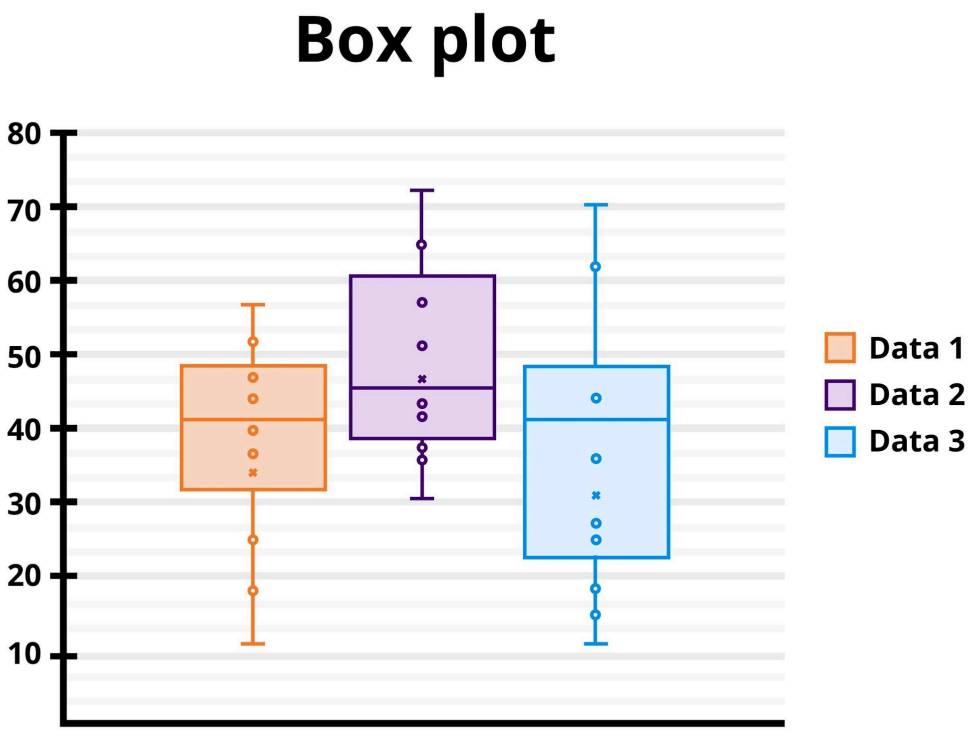
8. What is Sampling Bias? Give examples. Answer: Sampling bias occurs when the sample collected is not representative of the population intended to be analyzed.

- **Selection Bias:** Selecting specific individuals more often.

- **Survivorship Bias:** Focusing only on subjects that "survived" a process (e.g., analyzing only successful startups).
- **Under-coverage Bias:** Some members of the population are inadequately represented.

9. What is the difference between a Histogram and a Boxplot? Answer:

- **Histogram:** Shows the frequency distribution of a continuous variable (shape of data).
- **Boxplot:** Summarizes data through quartiles. It clearly shows the median, interquartile range (IQR), and explicitly identifies outliers.



Shutterstock

10. Explain Conditional Probability. Answer: It is the probability of an event occurring given that another event has already occurred. Denoted as $P(A|B)$, which is the probability of A given B. It is calculated as $P(A \cap B)/P(B)$.

Section 2: Data Preprocessing & Feature Engineering

11. How do you handle missing values in a dataset? Answer: Techniques depend on the nature of the data:

- **Deletion:** Remove rows/columns (if missing data is <5% and MCAR - Missing Completely At Random).
- **Imputation:** Fill with Mean/Median (for numerical), Mode (for categorical).
- **Advanced Imputation:** Using KNN or prediction models to estimate missing values.
- **Flagging:** Create a new column "Is_Missing" to capture the pattern of missingness.

12. How do you detect and treat outliers? Answer:

- **Detection:** Boxplots (points outside 1.5×IQR), Z-Score (points beyond 3 standard deviations).
- **Treatment:** Trimming (removing), Capping (setting to a threshold like the 99th percentile), or transforming (Log transformation) to reduce impact.

13. What is the difference between Normalization and Standardization? Answer:

- **Normalization (Min-Max Scaling):** Scales data to a range of [0, 1]. Useful for algorithms like Neural Networks and KNN that calculate distances.
- **Standardization (Z-Score Scaling):** Centers data around mean 0 with a standard deviation of 1. Better for algorithms that assume a Gaussian distribution (e.g., Logistic Regression, SVM).

14. What is One-Hot Encoding and when should you use it? Answer: One-Hot Encoding converts categorical variables into binary vectors (0s and 1s). For a feature "Color" with Red, Green, Blue, it creates 3 columns. It is used when there is no ordinal relationship between categories.

15. What is Multicollinearity and why is it a problem? Answer: Multicollinearity occurs when independent variables in a regression model are highly correlated. It makes it difficult to determine the individual effect of each independent variable on the dependent variable and inflates the variance of coefficients, making the model unstable.

16. How do you handle imbalanced datasets? Answer:

- **Resampling:** Undersampling the majority class or Oversampling the minority class (e.g., SMOTE - Synthetic Minority Over-sampling Technique).
- **Algorithmic:** Using class weights (penalizing mistakes on minority class).
- **Metrics:** Using F1-Score, Precision-Recall AUC instead of Accuracy.

17. What is PCA (Principal Component Analysis)? Answer: PCA is a dimensionality reduction technique. It projects high-dimensional data into a lower-dimensional space by finding new axes (principal components) that maximize the variance (information) in the data, while keeping them orthogonal (uncorrelated).

18. What is the Curse of Dimensionality? Answer: As the number of features (dimensions) increases, the amount of data needed to generalize accurately increases exponentially. In high dimensions, data becomes sparse, and distance metrics (like Euclidean distance) lose meaning, making algorithms like KNN ineffective.

19. What is Data Leakage? Answer: Data leakage happens when information from outside the training dataset is used to create the model. This typically occurs when the test data accidentally "leaks" into the training process (e.g., normalizing the whole dataset before splitting), resulting in overly optimistic performance scores that fail in production.

20. Feature Selection vs. Feature Engineering: What's the difference? Answer:

- **Feature Engineering:** Creating new features from raw data (e.g., extracting "Day of Week" from a "Date" column) to help the model learn better.
- **Feature Selection:** Selecting the most relevant subset of existing features to reduce dimensionality and overfitting (e.g., using Lasso regression or Recursive Feature Elimination).

Section 3: Core Machine Learning Concepts

21. Differentiate between Supervised and Unsupervised Learning. Answer:

- **Supervised:** Training on labeled data where the target outcome is known (e.g., Regression, Classification).
- **Unsupervised:** Training on unlabeled data to find hidden structures or patterns (e.g., Clustering, Association rules).

22. What is Overfitting and how can you avoid it? Answer: Overfitting happens when a model learns the training data too well, capturing noise instead of the signal. It performs great on training data but fails on test data.

- **Prevention:** Cross-validation, Regularization (L1/L2), Pruning (Trees), Dropout (Neural Networks), Early Stopping, and gathering more data.

23. What is Regularization (L1 vs L2)? Answer: Regularization adds a penalty term to the loss function to prevent overfitting by discouraging complex models (large coefficients).

- **L1 (Lasso):** Adds absolute value of magnitude of coefficients. Can shrink coefficients to zero (feature selection).
- **L2 (Ridge):** Adds squared magnitude of coefficients. Shrinks coefficients but not to zero.

24. Explain Cross-Validation. Answer: A technique to evaluate model stability. In k-fold cross-validation, the data is split into 'k' subsets. The model is trained on k-1 subsets and tested

on the remaining one. This process is repeated 'k' times, and the results are averaged. It ensures the model's performance is not dependent on just one specific train-test split.

25. Why do we need a Validation Set (distinct from Test Set)? Answer:

- **Training Set:** To fit the parameters (weights).
- **Validation Set:** To tune hyperparameters (e.g., depth of tree, K in KNN) and prevent overfitting during development.
- **Test Set:** To provide an unbiased evaluation of the final model.

26. Explain Gradient Descent. Answer: It is an optimization algorithm used to minimize the cost function. It iteratively adjusts the model parameters (weights) in the direction of the steepest descent (negative gradient) of the cost function until it converges to a minimum.

27. Stochastic vs. Batch vs. Mini-batch Gradient Descent. Answer:

- **Batch:** Updates weights after calculating gradients for the entire dataset (Slow, stable).
- **Stochastic (SGD):** Updates weights after every single record (Fast, noisy/unstable).
- **Mini-batch:** Updates weights after a small batch of records (Balance between speed and stability).

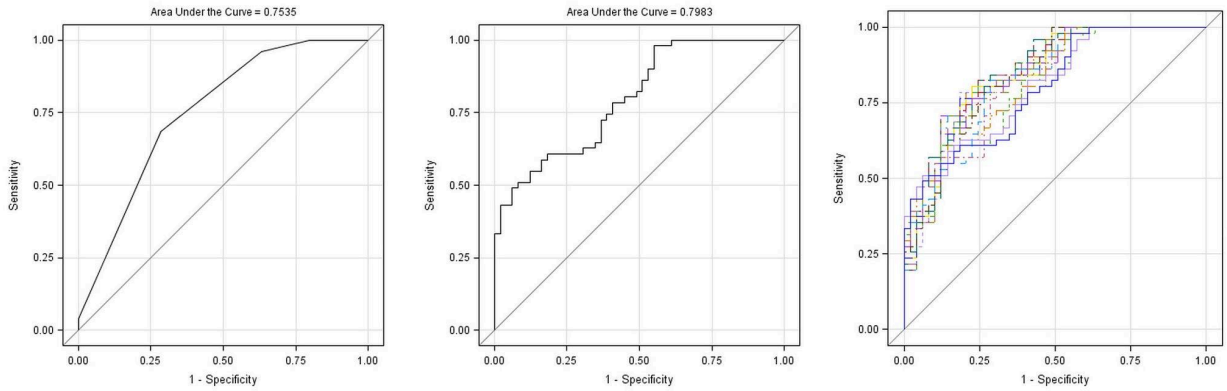
28. What is a Confusion Matrix? Answer: A table used to evaluate the performance of a classification model. It displays True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), allowing calculation of Accuracy, Precision, Recall, and F1.

29. What is the difference between Precision and Recall? Answer:

- **Precision:** Out of all predicted positives, how many are actually positive? (Use when False Positives are costly, e.g., Spam detection).
- **Recall (Sensitivity):** Out of all actual positives, how many did we capture? (Use when False Negatives are costly, e.g., Cancer detection).

30. What is the ROC Curve and AUC? Answer:

- **ROC (Receiver Operating Characteristic):** A plot of True Positive Rate (Recall) vs. False Positive Rate at various threshold settings.
- **AUC (Area Under Curve):** Represents the degree or measure of separability. Higher AUC means the model is better at distinguishing between classes.



Getty Images

31. What is the difference between Bagging and Boosting? Answer:

- **Bagging (Bootstrap Aggregating):** Trains multiple models (e.g., Decision Trees) independently in parallel on random subsets of data and averages the results to reduce variance (e.g., Random Forest).
- **Boosting:** Trains models sequentially, where each new model focuses on correcting the errors of the previous ones to reduce bias (e.g., Gradient Boosting, AdaBoost).

32. What is Pruning in Decision Trees? Answer: Pruning is the process of removing sections of the tree that provide little power to classify instances. It reduces the size of the tree to prevent overfitting. It can be pre-pruning (stop growing early) or post-pruning (cut branches after full growth).

33. Explain Entropy and Information Gain. Answer:

- **Entropy:** A measure of impurity or randomness in a dataset. (0 = pure sample, 1 = impure).
- **Information Gain:** The reduction in entropy achieved by splitting a dataset on a specific attribute. Decision Trees construct branches by choosing the split that maximizes Information Gain.

34. What is the Gini Index? Answer: Another metric for splitting in Decision Trees (specifically CART). It measures the probability of a random element being misclassified. A Gini index of 0 means pure. It is computationally faster than Entropy as it doesn't involve logarithms.

35. Parametric vs. Non-Parametric Models. Answer:

- **Parametric:** Assumes a fixed number of parameters and a specific functional form (e.g., Linear Regression). Good for smaller data, faster.
- **Non-Parametric:** Does not assume a functional form; the number of parameters grows with data (e.g., KNN, Decision Trees). More flexible, requires more data.

36. Generative vs. Discriminative Models. Answer:

- **Discriminative:** Learns the decision boundary between classes directly ($P(Y|X)$). Example: Logistic Regression, SVM.
- **Generative:** Learns the distribution of each class ($P(X|Y)$ and $P(Y)$) to generate new data or classify. Example: Naive Bayes.

37. What is Hyperparameter Tuning? Answer: The process of optimizing the configuration variables (hyperparameters) that govern the training process (e.g., learning rate, number of trees). Techniques include Grid Search (exhaustive) and Random Search (random sampling).

38. Explain Ensemble Learning. Answer: A technique that combines predictions from multiple machine learning models to reduce variance (Bagging), bias (Boosting), or improve predictions (Stacking), producing a model more accurate than any single individual model.

39. Name distinct Distance Metrics. Answer:

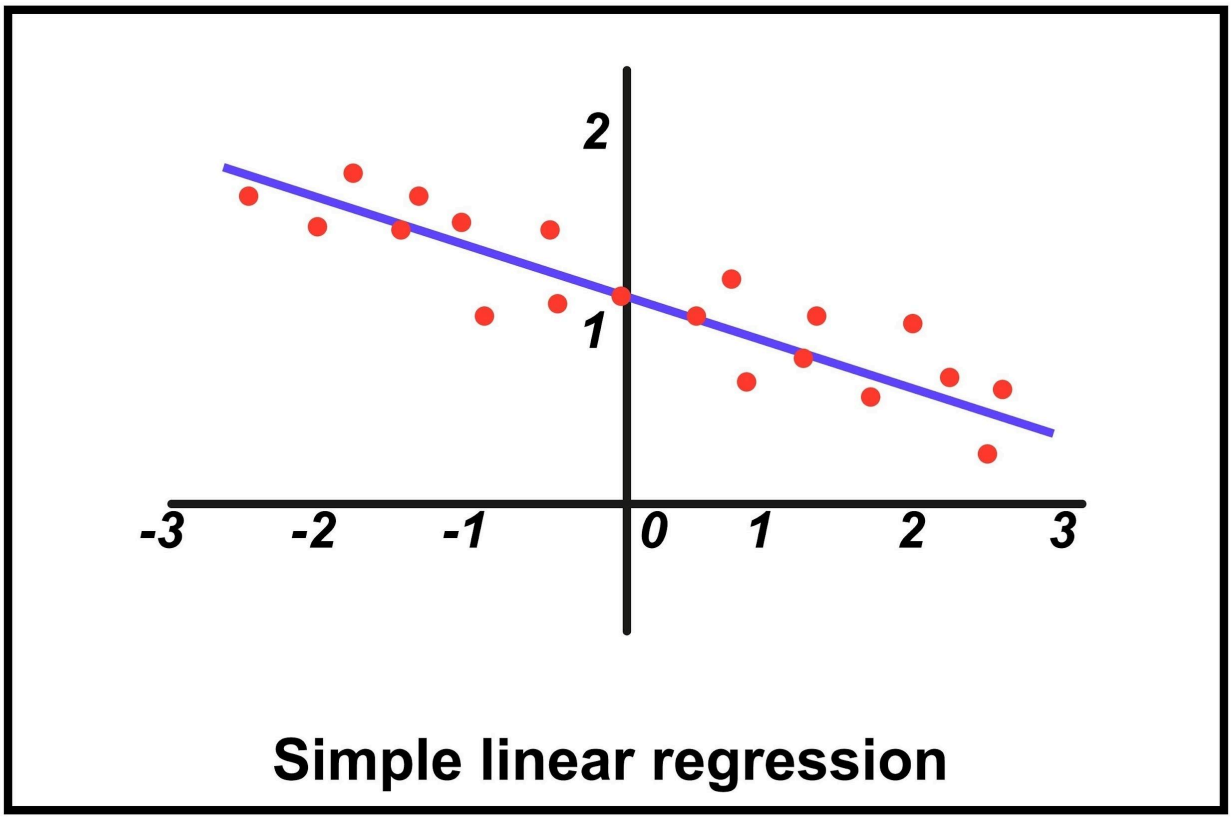
- **Euclidean:** Straight-line distance (L2 norm).
- **Manhattan:** Sum of absolute differences (L1 norm), grid-like.
- **Cosine:** Measures the angle between vectors (used in text analysis).
- **Minkowski:** A generalization of Euclidean and Manhattan.

40. What is the bias term in a linear equation (y=mx+c)? Answer: The bias term (c, or intercept) allows the regression line to shift up or down, ensuring it doesn't have to pass through the origin (0,0). It represents the value of y when all input features x are 0.

Section 4: Machine Learning Algorithms (Foundational)

41. Explain Linear Regression Assumptions. Answer:

1. **Linearity:** Relationship between X and Y is linear.
2. **Homoscedasticity:** Constant variance of errors.
3. **Independence:** Residuals are independent.
4. **Normality:** Residuals are normally distributed.



Shutterstock

42. What is the Sigmoid Function in Logistic Regression? Answer: The sigmoid function maps any real-valued number into a value between 0 and 1. S(x)=1/(1+e−x). This turns the output of a linear equation into a probability for binary classification.

43. How does the Random Forest algorithm work? Answer: It builds multiple decision trees during training. It uses Bootstrapping (random sampling with replacement) for data and Feature Randomness (selecting a subset of features) for splitting. The final prediction is the mode (classification) or mean (regression) of the individual trees.

44. What is the "Kernel Trick" in SVM? Answer: SVMs work well for linearly separable data. The Kernel Trick allows SVM to solve non-linear problems by implicitly mapping input data into a higher-dimensional space where it becomes linearly separable, without calculating the coordinates in that high-dimensional space.

45. Why is Naive Bayes called "Naive"? Answer: It assumes that all features are independent of each other given the class label. This is a "naive" assumption because, in real-world data, features are often correlated (e.g., age and salary), yet the algorithm still performs surprisingly well.

46. How does K-Nearest Neighbors (KNN) work? Answer: KNN is a lazy learner. It doesn't "learn" a model. When predicting a new point, it calculates the distance to all training points,

picks the 'K' nearest neighbors, and assigns the class based on a majority vote (classification) or average (regression).

47. How do you choose 'K' in K-Means Clustering? Answer: Using the Elbow Method. You plot the WCSS (Within-Cluster Sum of Squares) against the number of clusters. The point where the plot bends (like an elbow), indicating that adding more clusters yields diminishing returns in variance reduction, is the optimal K.

48. Difference between K-Means and Hierarchical Clustering. Answer:

- **K-Means:** Partitional clustering. You define 'K' beforehand. Faster for large data.
- **Hierarchical:** Builds a tree of clusters (dendrogram). No need to define 'K' initially. Better for small datasets and visualizing hierarchy.

49. What is Gradient Boosting? Answer: An ensemble technique where new models (usually decision trees) are added sequentially to correct the errors made by previous models. It uses gradient descent to minimize the loss function when adding new models.

50. What is XGBoost? Answer: eXtreme Gradient Boosting is an optimized implementation of gradient boosting. It is designed for speed and performance, featuring parallel processing, tree-pruning, handling missing values internally, and regularization to avoid overfitting.

Part 2: Deep Learning & Advanced Algorithms

Section 5: Neural Networks Fundamentals

51. What is a Perceptron? Answer: A Perceptron is the fundamental building block of neural networks. It is a binary linear classifier that takes multiple inputs, multiplies them by weights, sums them up, adds a bias, and passes the result through an activation function to produce an output.

52. Explain the Backpropagation algorithm. Answer: Backpropagation is the learning mechanism for neural networks. It calculates the gradient of the loss function with respect to each weight by applying the chain rule, moving backward from the output layer to the input layer. These gradients are then used to update the weights (via Gradient Descent) to minimize error.

53. What is an Activation Function and why is it necessary? Answer: An activation function determines the output of a neural network node given an input. Crucially, it introduces non-linearity into the network. Without them, a neural network would behave just like a single-layer linear regression model, unable to learn complex patterns.

54. Compare Sigmoid, Tanh, and ReLU activation functions. Answer:

- **Sigmoid:** Squashes output between 0 and 1. Good for binary classification output, but suffers from the vanishing gradient problem.
- **Tanh:** Squashes output between -1 and 1. Zero-centered, generally better than sigmoid, but still has vanishing gradients.
- **ReLU (Rectified Linear Unit):** f(x)=max(0,x). Solves vanishing gradient for positive values, computationally efficient, but can suffer from "dying ReLU".

55. What is the "Vanishing Gradient Problem"? Answer: In deep networks with certain activation functions (like Sigmoid), gradients become increasingly small as they backpropagate to earlier layers. This means the weights in the initial layers barely update, preventing the network from learning effectively.

56. What is the "Exploding Gradient Problem"? Answer: The opposite of vanishing gradients, where error gradients accumulate and become very large during training. This causes large updates to weights, making the model unstable. It is often handled using Gradient Clipping.

57. What is Dropout? Answer: Dropout is a regularization technique where randomly selected neurons are ignored (dropped out) during training. This forces the network to learn more robust features that are redundant across many neurons, preventing overfitting.

58. What is Batch Normalization? Answer: It is a technique to improve the speed, performance, and stability of neural networks. It normalizes the inputs of each layer to have a mean of 0 and variance of 1.

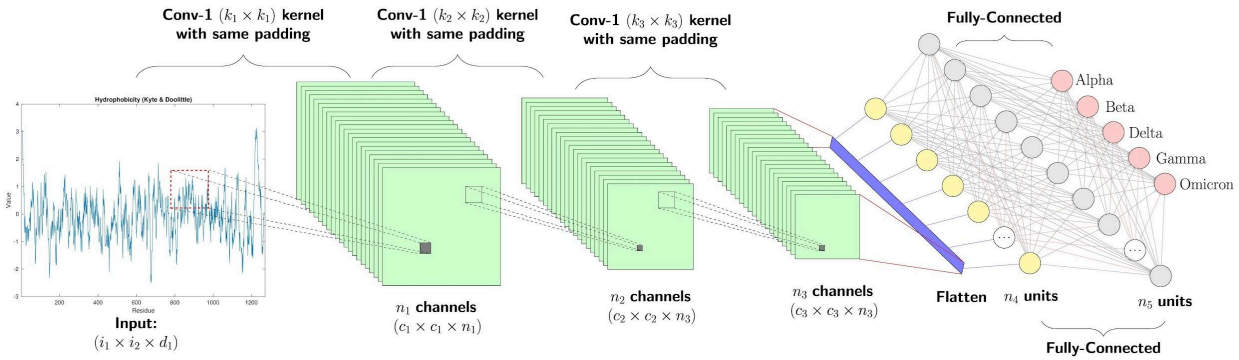
59. What is an Epoch, Batch, and Iteration? Answer:

- **Epoch:** One complete pass of the entire training dataset through the algorithm.
- **Batch:** A subset of the training dataset used for one gradient update.
- **Iteration:** The number of batches needed to complete one epoch.

60. What are Optimizers? Name a few. Answer: Optimizers are algorithms used to change the attributes of the neural network (weights and learning rate) to reduce losses. Examples: SGD, RMSprop, and Adam (Adaptive Moment Estimation).

Section 6: Convolutional Neural Networks (CNNs)

61. What is a CNN and where is it used? Answer: A Convolutional Neural Network is a deep learning algorithm designed to process structured grid data, such as images. It uses convolution operations to automatically and adaptively learn spatial hierarchies of features.



Shutterstock

62. Explain the Convolution Operation. Answer: It involves sliding a filter (kernel) over the input image to perform element-wise multiplication and summation. This produces a feature map that highlights specific features (like vertical edges) detected by that kernel.

63. What is Pooling (Max Pooling vs. Average Pooling)? Answer: Pooling reduces the spatial dimensions of the input volume.

- **Max Pooling:** Takes the maximum value from a feature map region. Preserves dominant features.
- **Average Pooling:** Takes the average value. Smooths the feature map.

64. What is Padding in CNNs? Answer: Padding involves adding layers of zeros to the input image borders. It allows the kernel to cover the edges of the image and controls the spatial size of the output volume.

65. What is Stride? Answer: Stride is the number of pixels the kernel moves over the input matrix. A stride of 2 downsamples the output feature map by a factor of 2.

66. What is a Fully Connected Layer in a CNN? Answer: After convolutional and pooling layers, the high-level reasoning is done via fully connected layers. Neurons here have connections to all activations in the previous layer, similar to a regular Neural Network.

67. Explain Data Augmentation. Answer: A technique to artificially increase the size of a training dataset by creating modified versions of images (rotations, flips, crops). It helps prevent overfitting.

68. What is Transfer Learning? Answer: Taking a pre-trained model (e.g., trained on ImageNet) and repurposing it for a related task. You typically freeze the early layers and fine-tune the later layers for your specific problem.

69. Name some popular CNN architectures. Answer: LeNet-5, AlexNet, VGG-16, ResNet (Residual Networks), Inception (GoogLeNet), and EfficientNet.

70. What is the difference between R-CNN, Fast R-CNN, and Faster R-CNN? Answer:

- **R-CNN:** Uses selective search to propose regions, then runs CNN on each (Slow).
- **Fast R-CNN:** Runs CNN on the whole image once, then maps proposals to feature map (Faster).
- **Faster R-CNN:** Replaces selective search with a Region Proposal Network (RPN) learned by the neural network (Fastest).

Section 7: Recurrent Neural Networks (RNNs) & NLP

71. What is an RNN and how is it different from a standard NN? Answer: RNNs are designed for sequential data (time series, text). Unlike standard NNs, RNNs have a "memory" (hidden state) that captures information about what has been calculated so far.

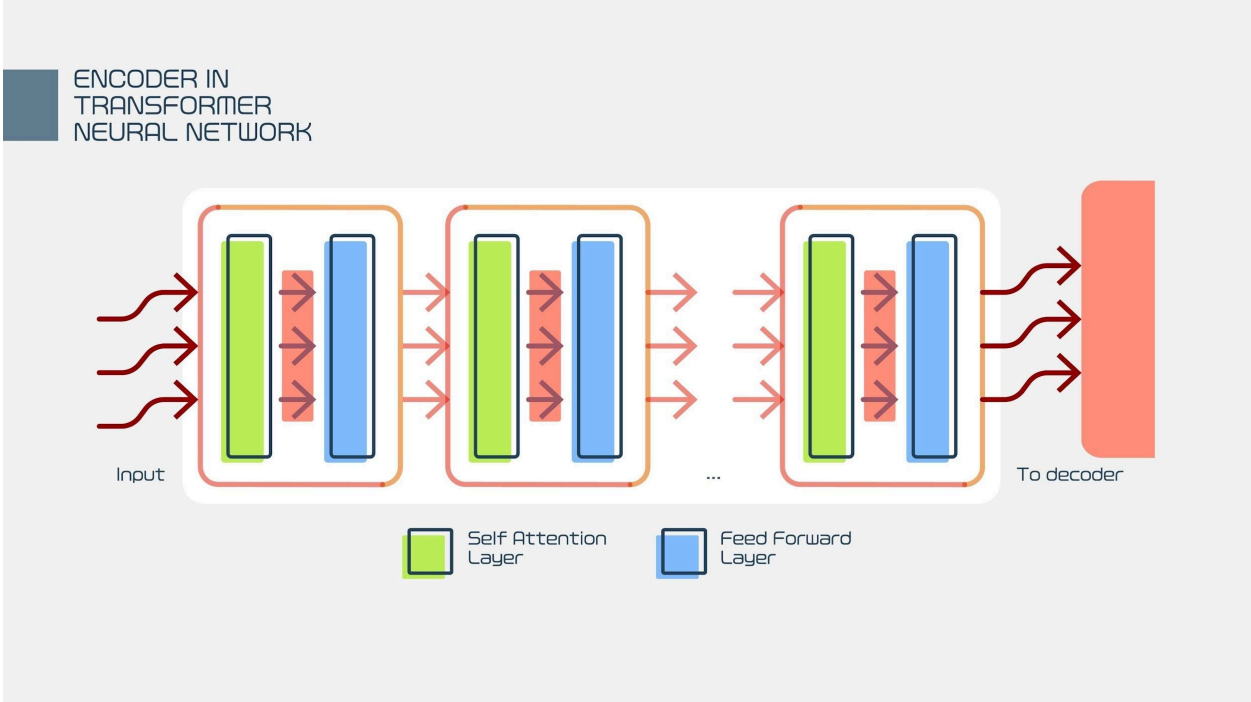
72. What is the major issue with standard RNNs? Answer: They suffer severely from the Vanishing Gradient Problem over long sequences, making it difficult to carry information from the beginning of the sequence to the end (forgetting long-term dependencies).

73. What is an LSTM (Long Short-Term Memory)? Answer: LSTMs are a special kind of RNN capable of learning long-term dependencies. They use a gating mechanism (Input, Output, and Forget gates) to regulate the flow of information.

74. What is a GRU (Gated Recurrent Unit)? Answer: A simplified version of the LSTM. It combines the forget and input gates into a single "update gate." It is computationally more efficient than LSTM.

75. What is Word Embedding (Word2Vec)? Answer: A technique to represent words as dense vectors of real numbers. Words with similar meanings are mapped to nearby points in the vector space.

76. Explain the Transformer Architecture. Answer: Transformers abandoned recurrence entirely and rely on the Attention Mechanism to draw global dependencies between input and output. They allow for massive parallelization and are the foundation of modern NLP (BERT, GPT).



Shutterstock

77. What is the Attention Mechanism? Answer: It allows the model to "focus" on different parts of the input sequence when producing a specific part of the output sequence.

78. What is BERT (Bidirectional Encoder Representations from Transformers)? Answer: BERT is a transformer-based model designed to pre-train deep bidirectional representations from unlabeled text. It is trained on Masked Language Modeling and Next Sentence Prediction.

79. What is the difference between Tokenization, Stemming, and Lemmatization? Answer:

- **Tokenization:** Splitting text into individual words or subwords.
- **Stemming:** Cutting off prefixes/suffixes to find the root (crude).
- **Lemmatization:** Reducing a word to its base dictionary form (lemma) using morphological analysis.

80. What are Stop Words? Answer: Common words (like "the", "is", "at") that appear frequently but often carry little meaningful information for classification. They are often removed during preprocessing.

Section 8: Advanced ML & Model Evaluation

81. What is Cross-Entropy Loss? Answer: Also known as Log Loss, it measures the performance of a classification model whose output is a probability value between 0 and 1.

82. What is the difference between L1 and L2 loss? Answer:

- **L1 Loss (MAE):** Mean Absolute Error. Robust to outliers.
- **L2 Loss (MSE):** Mean Squared Error. Penalizes larger errors more heavily. Not robust to outliers.

83. Difference between K-Means and KNN? Answer:

- **K-Means:** Unsupervised clustering. 'K' is the number of clusters.
- **KNN:** Supervised classification/regression. 'K' is the number of neighbors.

84. What is Dimensionality Reduction with t-SNE? Answer: t-Distributed Stochastic Neighbor Embedding (t-SNE) is a non-linear technique for dimensionality reduction, particularly well suited for the visualization of high-dimensional datasets.

85. What is Time Series Analysis? Answer: Analyzing data points collected at specific time intervals to identify trends, cycles, and seasonal variations. Models: ARIMA, SARIMA, LSTMs.

86. What is Stationarity in Time Series? Answer: A time series is stationary if its statistical properties (mean, variance) do not change over time. Most forecasting models assume stationarity.

87. What is Recommendation System (Collaborative vs. Content-based)? Answer:

- **Collaborative Filtering:** Based on user-item interactions ("Users who liked X also liked Y").
- **Content-Based:** Based on item attributes ("You liked X, so you might like Y because it has similar features").

88. What is Reinforcement Learning? Answer: A type of ML where an agent learns to make decisions by performing actions in an environment and receiving rewards or penalties. (Example: AlphaGo).

89. What is Heteroscedasticity? Answer: In regression, when the variability of a variable is unequal across the range of values of a second variable that predicts it (The "cone shape" in residual plots).

90. What is the F1 Score? Answer: The harmonic mean of Precision and Recall. F1=2×Precision+RecallPrecision×Recall. Useful for imbalanced classes.

91. What is SMOTE? Answer: Synthetic Minority Over-sampling Technique. Addresses imbalanced datasets by generating synthetic examples for the minority class.

92. How does SVM work? Answer: SVM finds the hyperplane that best separates the classes with the maximum margin between the hyperplane and the nearest data points (support vectors).

93. What is an Autoencoder? Answer: An unsupervised neural network used to learn efficient codings. It has an Encoder (compresses input) and a Decoder (reconstructs input).

94. What is a GAN (Generative Adversarial Network)? Answer: A framework where two neural networks contest with each other. The Generator creates fake data, and the Discriminator tries to distinguish fake data from real data.

95. What is Mean Squared Error (MSE)? Answer: The average of the squares of the errors. Common loss function for regression.

96. What is R-Squared? Answer: A statistical measure representing the proportion of the variance for a dependent variable that's explained by an independent variable.

97. What is Ensemble Learning - Stacking? Answer: Stacking involves training a new model (meta-learner) to combine the predictions of several base models.

98. What is Transfer Learning in NLP? Answer: Pre-training a language model on a large corpus (like Wikipedia) to understand language structure, then fine-tuning it on a specific task (e.g., ULMFiT).

99. Explain "explaining" a model (SHAP/LIME). Answer:

- **SHAP:** Assigns each feature an importance value for a particular prediction based on game theory.
- **LIME:** Approximates the complex model locally with a simple interpretable model.

100. Difference between Parameters and Hyperparameters? Answer:

- **Parameters:** Learned internal variables (Weights, Biases).
- **Hyperparameters:** External configurations set before training (Learning Rate, K in KNN).

Part 3: SQL, Big Data & MLOps

Section 9: SQL & Database Concepts

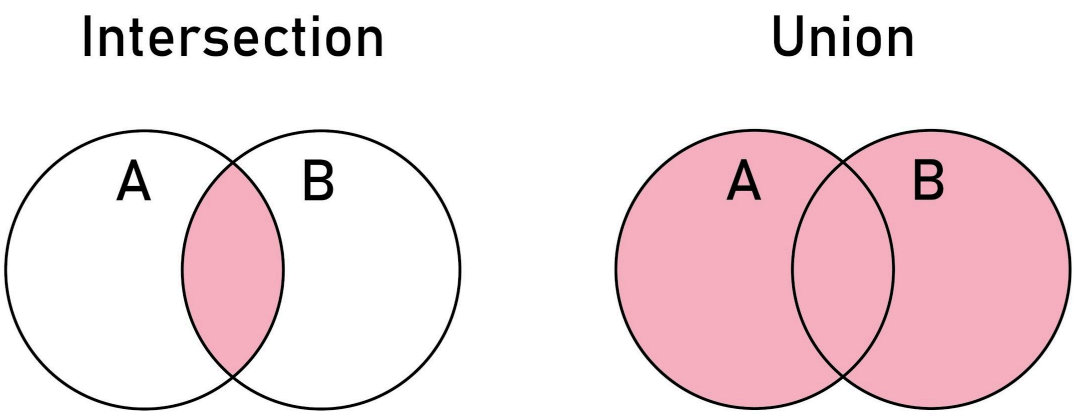
101. What is the difference between WHERE and HAVING clauses? Answer:

- **WHERE:** Filters rows before grouping. Cannot be used with aggregate functions.
- **HAVING:** Filters groups after grouping. Used with aggregate functions.

102. Explain the different types of SQL Joins. Answer:

- **INNER JOIN:** Matching values in both tables.

- **LEFT JOIN:** All from left, matched from right.
- **RIGHT JOIN:** All from right, matched from left.
- **FULL JOIN:** Match in either table.
- **CROSS JOIN:** Cartesian product.



Shutterstock

103. What is a Window Function in SQL? Answer: Performs calculations across a set of table rows related to the current row without collapsing them (unlike GROUP BY). Examples: `RANK()`, `LEAD()`, `LAG()`.

104. Difference between RANK(), DENSE_RANK(), and ROW_NUMBER(). Answer:

- **ROW_NUMBER():** 1, 2, 3, 4 (Unique sequential).
- **RANK():** 1, 2, 2, 4 (Skips numbers for ties).
- **DENSE_RANK():** 1, 2, 2, 3 (No skipping).

105. What are ACID properties? Answer: Atomicity, Consistency, Isolation, Durability. They ensure reliable database transactions.

106. What is Database Normalization (1NF, 2NF, 3NF)? Answer:

- **1NF:** Atomic columns.
- **2NF:** 1NF + no partial dependency.
- **3NF:** 2NF + no transitive dependency.

107. Primary Key vs. Foreign Key? Answer:

- **Primary Key:** Uniquely identifies a row. Cannot be NULL.

- **Foreign Key:** Links to the Primary Key of another table.

108. UNION vs. UNION ALL? Answer:

- **UNION:** Removes duplicates (slower).
- **UNION ALL:** Keeps duplicates (faster).

109. How would you find the second highest salary? Answer: `SELECT MAX(Salary) FROM Employee WHERE Salary < (SELECT MAX(Salary) FROM Employee)`

110. What is a CTE? Answer: Common Table Expression. A temporary result set defined with `WITH`.

111. SQL vs. NoSQL? Answer:

- **SQL:** Relational, structured, scales vertically (MySQL).
- **NoSQL:** Non-relational, unstructured, scales horizontally (MongoDB).

112. What is an Index? Answer: A data structure that improves data retrieval speed but slows down writes (INSERT/UPDATE).

113. What is a "View"? Answer: A virtual table based on the result-set of an SQL statement.

114. How do you handle duplicates in SQL? Answer: Use `DISTINCT`. To delete, use CTE with `ROW_NUMBER()`.

115. DELETE vs. TRUNCATE vs. DROP? Answer:

- **DELETE:** Specific rows, slower, rollback possible.
- **TRUNCATE:** All rows, fast, usually no rollback.
- **DROP:** Removes table and data entirely.

Section 10: Big Data Technologies

116. What is Hadoop and its main components? Answer:

- **HDFS:** Storage (distributed).
- **MapReduce:** Processing model.
- **YARN:** Resource management.

117. Explain MapReduce. Answer:

- **Map:** Converts data into key/value pairs.
- **Reduce:** Aggregates those pairs into a smaller set.

118. Apache Spark vs. Hadoop MapReduce? Answer: Spark processes data in-memory (RAM), making it 100x faster than MapReduce which writes intermediate results to disk.

119. What is an RDD? Answer: Resilient Distributed Dataset. Spark's fundamental data structure (immutable, distributed).

120. What is "Lazy Evaluation" in Spark? Answer: Spark records the lineage of operations but only executes them when an Action (like count or collect) is triggered.

121. Explain the CAP Theorem. Answer: In a distributed system, you can only guarantee two of three: Consistency, Availability, Partition Tolerance.

122. Types of NoSQL databases? Answer: Key-Value, Document, Column-oriented, Graph.

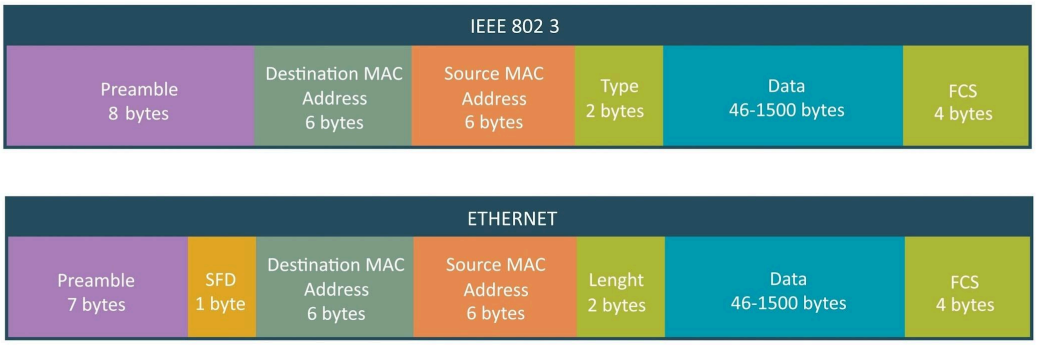
123. What is Hive? Answer: Data warehouse software on Hadoop that provides an SQL-like interface (HQL).

124. What is Sharding? Answer: Breaking a large database into smaller pieces (shards) spread across multiple servers.

125. What is the Parquet file format? Answer: A columnar storage format efficient for compression and analytics queries.

Section 11: Deployment & MLOps

126. What is Docker? Answer: A platform for creating "containers" that package code with dependencies, ensuring it runs the same everywhere.



Shutterstock

127. What is Kubernetes? Answer: A system for automating deployment, scaling, and management of containerized applications.

128. What is a REST API? Answer: An architectural style for APIs using HTTP requests. Used to serve models to other applications.

129. What is "Pickling"? Answer: Serializing a Python object (like a model) into a byte stream to save to disk.

130. Explain Model Drift. Answer:

- **Data Drift:** Input distribution changes.
- **Concept Drift:** Relationship between input and output changes.

131. What is CI/CD in ML? Answer:

- **CI:** Automated testing of code/data.
- **CD:** Automated deployment and retraining pipelines.

132. What is A/B Testing in Deployment? Answer: Comparing a new model against an old one in production on a subset of users.

133. Latency vs. Throughput? Answer:

- **Latency:** Time for one prediction (Real-time).
- **Throughput:** Predictions per second (Batch).

134. How do you monitor a model? Answer: Monitor Service Metrics (Latency, Error rates) and Model Metrics (Drift, Accuracy).

135. What is "Shadow Mode"? Answer: Deploying a new model alongside the live one but logging predictions instead of showing them to users.

136. How would you serve a Deep Learning model? Answer: TensorFlow Serving or TorchServe, wrapped in Docker/Kubernetes.

137. What is ONNX? Answer: Open Neural Network Exchange. A format to represent DL models, allowing transfer between frameworks.

138. Batch vs. Online Prediction? Answer:

- **Batch:** Large set at once (Nightly).
- **Online:** Immediate (On request).

139. What is Git? Answer: Version control system to track code changes.

140. What is a Feature Store? Answer: Centralized repository for features ensuring consistency between training and inference.

Section 12: Coding & Practical Scenarios

141. Scenario: High accuracy (98%) but useless model? Answer: Imbalanced dataset. 98% legitimate, 2% fraud. Model predicts "Legitimate" for all. Use Precision/Recall.

142. Scenario: 1000 features, 500 rows? Answer: High variance/Overfitting risk. Use Regularization (Lasso) or PCA.

143. Handling High Cardinality features? Answer: Target Encoding, Frequency Encoding, or Embeddings.

144. Perfectly correlated features in Linear Regression? Answer: Multicollinearity. Matrix becomes singular. Drop one feature.

145. SQL: 3rd highest salary? Answer: `SELECT MIN(Salary) FROM (SELECT DISTINCT Salary FROM Employee ORDER BY Salary DESC LIMIT 3) as T;`

146. Debugging a model that isn't learning? Answer: Check Learning Rate, Data/Labels, Architecture. Try to overfit on a tiny batch.

147. Random Forest vs. XGBoost? Answer: RF for parallel training, less tuning. XGBoost for maximum performance.

148. Scenario: Missing "Age" in production? Answer: Implement robust pipelines (imputation) matching training logic.

149. Optimize slow SQL? Answer: Check Indexes, avoid `SELECT *`, check Explain Plan.

150. What is Stratified Sampling? Answer: Dividing population into subgroups (strata) and sampling from each to maintain class distribution.

Part 4: GenAI, Advanced NLP, Computer Vision & Behavioral Skills

Section 13: Generative AI & LLMs

151. What is a Large Language Model (LLM)? Answer: A deep learning model trained on massive text data using Transformer architecture to understand and generate content (e.g., GPT-4).

152. Explain "Prompt Engineering". Answer: Designing inputs to guide LLMs.

- **Zero-shot:** No examples.
- **Few-shot:** Providing examples.
- **Chain-of-Thought:** Asking for step-by-step reasoning.

153. What is RAG (Retrieval-Augmented Generation)? Answer: Combining an LLM with an external knowledge base. It retrieves relevant docs to provide context for the answer.

154. What is Fine-Tuning? Answer: Training a pre-trained model further on a specific dataset to specialize it.

155. What is PEFT? Answer: Parameter-Efficient Fine-Tuning (e.g., LoRA). Freezes most weights and trains only small adapter layers.

156. What is "Temperature"? Answer: Controls randomness. Low = Deterministic. High = Creative.

157. Hallucination vs. Bias? Answer:

- **Hallucination:** Generating false facts.
- **Bias:** Reflecting stereotypes from training data.

158. What is a Vector Database? Answer: Stores high-dimensional vectors for semantic search (e.g., Pinecone).

159. What are "Embeddings"? Answer: Numerical vector representations of text/images capturing semantic meaning.

160. What is RLHF? Answer: Reinforcement Learning from Human Feedback. Aligning LLMs with human values using a Reward Model.

Section 14: Advanced NLP & Computer Vision

161. What is Named Entity Recognition (NER)? Answer: Classifying named entities (Person, Org, Location) in text.

162. What is Topic Modeling (LDA)? Answer: Unsupervised technique to discover abstract topics in documents.

163. Image Segmentation (Semantic vs. Instance)? Answer:

- **Semantic:** Classifies pixels by category (all cars are red).
- **Instance:** Distinguishes objects (car 1 red, car 2 blue).

164. What is YOLO? Answer: You Only Look Once. Real-time object detection processing the whole image at once.

165. Object Detection vs. Classification? Answer:

- **Classification:** What is it?
- **Detection:** Where is it? (Bounding Box).

166. What is a Vision Transformer (ViT)? Answer: Applying Transformers to images by splitting them into patches.

167. What is OCR? Answer: Optical Character Recognition. Converting images of text into machine-encoded text.

168. "Attention Is All You Need"? Answer: The 2017 paper introducing Transformers.

169. What is Sentiment Analysis? Answer: Identifying polarity (positive/negative) in text.

170. What is Zero-Shot Learning? Answer: Recognizing concepts never seen during training via semantic association.

Section 15: Behavioral & Soft Skills

171. "Explain a complex concept to a non-technical audience." Strategy: Use analogies. (e.g., Random Forest = Average of 100 real estate agents).

172. "Tell me about a time your model failed." Strategy: STAR method. Situation (Failed), Task (Fix), Action (Retrained without leakage), Result (Lower but real accuracy).

173. "Handling disagreement with stakeholder?" Strategy: Discuss trade-offs (Triangle of constraints: Scope, Time, Quality).

174. "What if data is 'dirty'?" Strategy: Investigate source, document cleaning steps, ensure reproducibility.

175. "How do you stay updated?" Strategy: ArXiv, Kaggle, Blogs, Twitter/LinkedIn.

176. "Cross-functional team experience?" Strategy: Highlight communication and enabling others (e.g., building a dashboard for Marketing).

177. "Most important metric?" Strategy: Business impact (Profit/ROI), not just accuracy.

178. "Prioritizing projects?" Strategy: Impact vs. Effort matrix.

179. "Steps from problem to deployment?" Strategy: Define -> Data -> Model -> Evaluate -> Deploy -> Monitor.

180. "Why Data Science?" Strategy: Authentic passion for solving puzzles and creating value.

Section 16: Python & Coding Checks

181. What is a generator? Answer: Function returning an iterator (using `yield`). Memory efficient.

182. List vs. Tuple? Answer: List is Mutable `[]`. Tuple is Immutable `()`.

183. `__init__` method? Answer: Constructor method to initialize objects.

184. What are decorators? Answer: Functions modifying other functions (`@wrapper`).

185. `*args` and `**kwargs`? Answer: Variable positional arguments and keyword arguments.

186. `is` vs `==`? Answer: `is` checks memory reference identity. `==` checks value equality.

187. Python Memory Management? Answer: Private heap, Reference Counting, Garbage Collector.

188. List comprehension? Answer: `[x**2 for x in range(10)]`

189. `loc` vs `iloc`? Answer: `loc` (Label), `iloc` (Index/Position).

190. Merge DataFrames? Answer: `pd.merge(df1, df2, on='key')`

191. GIL? Answer: Global Interpreter Lock. Prevents true multi-threading in Python.

192. Exception Handling? Answer: `try`, `except`, `else`, `finally`.

193. Lambda function? Answer: `lambda a: a + 10`. Anonymous single-line function.

194. `with` statement? Answer: Context manager (auto-closes resources).

195. `copy()` vs `deepcopy()`? Answer: Shallow vs. Independent recursive copy.

196. Plotting libraries? Answer: Matplotlib, Seaborn, Plotly.

197. NumPy broadcasting? Answer: Arithmetic on arrays of different shapes.

198. Reverse string? Answer: `string[::-1]`

199. `if __name__ == "__main__":`? Answer: Runs code only if executed as a script.

200. Sort dictionary by value? Answer: `sorted(d.items(), key=lambda x: x[1])`