SAVEETHA SCHOOL OF ENGINEERING

SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES

COMPUTER SCIENCE AND ENGINEERING

Subject Code: DSA0420
Subject Name: Fundamentals of data science for healthcare informatics
Faculty Name: Mangaiyarkarasi K

VIVA QUESTIONS

1. What is the driving force behind the surge in demand for data science?

Answer: The exponential growth of data generation across various fields has created a need to extract meaningful insights and make data-driven decisions. Traditional methods struggle to handle the volume, variety, and velocity of Big Data. Data science offers a comprehensive approach to harnessing this data for valuable discoveries.

2. Elaborate on some of the key benefits of data science?

Answer: Data science unlocks a multitude of benefits:

- **Improved Decision-Making:** Data-driven insights guide informed choices across business operations, marketing, finance, and other domains.
- **Enhanced Customer Experience:** By analyzing customer behavior, data science helps personalize interactions and improve customer satisfaction.
- **Reduced Costs:** Data science can identify operational inefficiencies and predict future trends, leading to cost optimization.
- **Innovation:** Discovering hidden patterns and trends in data fuels innovation and the development of new products and services.
- **Risk Management:** Data analysis allows for proactive risk identification and mitigation strategies.

3. What are the different facets of data that data scientists work with?

Answer: Data scientists deal with several data facets:

- **Structured Data:** Organized and well-defined data stored in relational databases (e.g., customer information, financial transactions).
- **Unstructured Data:** Text, images, video, social media posts, etc., requiring specialized techniques for extraction and analysis.
- **Semi-structured Data:** Data with some inherent structure but not strictly adhering to a fixed format (e.g., emails, logs).

4. How does the Big Data ecosystem function?

Answer: The Big Data ecosystem comprises various components working in concert:

- **Data Sources:** Generate vast amounts of data (e.g., social media, sensors, IoT devices).
- **Data Acquisition:** Techniques like data scraping, APIs, and data warehousing collect the data.
- **Data Storage:** Distributed storage systems (e.g., Hadoop) handle massive datasets efficiently.
- **Data Processing:** Cleaning, transforming, and preparing data for analysis.
- **Data Analysis Tools:** Specialized software (R, Python) enables exploration and statistical calculations.
- **Data Visualization:** Tools (Tableau, Power BI) create clear and compelling insights.

5. Explain the steps involved in the data science process.

Answer: The data science process follows a structured approach:

1. **Data Retrieval:** Gathering data from diverse sources, ensuring its relevance to the problem.
2. **Data Cleaning, Integration, and Transformation:** Addressing missing values, inconsistencies, and transforming the data into a usable format.
3. **Exploratory Data Analysis (EDA):** Gaining initial understanding and identifying patterns through visualization and summary statistics.
4. **Model Building:** Choosing appropriate algorithms (e.g., regression, classification) and training models on the data.
5. **Model Evaluation:** Assessing model performance with metrics like accuracy, precision, and recall.
6. **Model Deployment:** Putting the finalized model into production for real-world use.
7. **Communication and Insight Presentation:** Communicating findings effectively to decision-makers.

6. How do data scientists typically retrieve data?

Answer: Data retrieval methods vary depending on the data source:

- **Databases:** SQL queries for structured data in relational databases.
- **APIs:** Application Programming Interfaces for real-time or programmatic access to data.
- **Web Scraping:** Extracting data from websites using tools or code.
- **Data Warehousing:** Accessing consolidated data from multiple sources.
- **Sensors and IoT Devices:** Using specific protocols or software to collect sensor data.

7. Why is data cleaning such a crucial step in the data science process?

Answer: Data cleaning is essential because:

- **Incomplete or inaccurate data** leads to unreliable results and misleading insights.
- **Cleaning ensures data** is consistent and adheres to the chosen format for analysis.
- **It optimizes processing times** and facilitates smoother model training.

8. Differentiate between supervised and unsupervised learning in data science.

Answer: Data science utilizes two main learning paradigms:

- **Supervised Learning:** Models are trained on labeled data (inputs with known outputs) to learn relationships and make predictions for unseen data. (e.g., Customer churn prediction using labeled data on customer characteristics).
- **Unsupervised Learning:** Used for unlabeled data where the objective is to discover hidden patterns or clusters. (e.g., Customer segmentation based on purchase behavior).

9. What are common techniques used for presenting data science findings?

Answer: Effective communication is key:

- Interactive Dashboards (Tableau, Power BI): Provide dynamic visualizations for exploration
- Static Reports and Presentations: Communicate insights with data visualizations and explanations
- Interactive Storytelling: Use data-driven narratives to engage audiences

10. How do data scientists collaborate with other professionals?

Answer: Data scientists work with various teams:

- Domain Experts: Provide subject-matter

11. Explain the process of reading data from a CSV file using Python.

```
import pandas as pd

data = pd.read_csv("your_data.csv")  # Replace "your_data.csv" with the actual file path
print(data.head())  # Display the first few rows of the data
```

12. How do you select specific columns from a pandas DataFrame?

Answer: bracket notation to select columns by name or position:

```
selected_columns = data[["column1", "column3"]]  # Select columns by name
selected_columns = data.iloc[:, [0, 2]]  # Select columns by position (0-based indexing)
```

13. Describe methods for filtering data in a pandas DataFrame based on conditions.

Answer:boolean indexing with logical operators (&, |, ~) to filter rows:

```
filtered_data = data[data["column1"] > 10]  # Filter rows where "column1" is greater than 10
filtered_data = data[(data["column2"] == "category A") & (data["column3"] < 5)]  # Combine conditions
```

14. How can you handle missing data in a pandas DataFrame?

Answer: There are several ways to address missing data:

- .fillna(value): Replace missing values with a specified value (e.g., 0, mean)
- .dropna(): Drop rows or columns with missing values (use with caution)

- .interpolate(): Fill missing values with estimated values (e.g., linear interpolation)

15. How do you sort data in a pandas DataFrame?

Answer: The .sort_values() method, specifying the column and sort order (ascending or descending):

Python

sorted_data = data.sort_values(by="column1", ascending=False) # Sort by "column1" in descending order

16. Explain how to group data in a pandas DataFrame by a specific column.

Answer: The .groupby() method to create groups based on a column, then perform operations on each group (e.g., calculate means):

Python

grouped_data = data.groupby("column2")["column1"].mean() # Calculate mean of "column1" for each group in "column2"

17. How can you rank data in a pandas DataFrame?:.rank(): Assigns a rank to each data point within a group (requires grouping)

- .nlargest() or .nsmallest(): Select the N largest or smallest values

18. What is the purpose of the NumPy library in data science?

- Creating and manipulating multidimensional arrays
- Performing numerical computations (vectorized operations)
- Linear algebra operations

19. Briefly describe the functionalities of SciPy for scientific computing.

- Optimization algorithms
- Integration and differentiation
- Statistical functions (beyond NumPy's basics)

20. How does scikit-learn support machine learning tasks in Python?

- Data preprocessing (scaling, feature selection)
- Supervised learning models (classification, regression)
- Unsupervised learning m**10 Viva Questions and Answers on Data Preparation and Exploratory Data Analysis (EDA)**

21. Describe the different stages of data preparation in a data science project.

Answer: Data preparation involves several steps:

- **Data Acquisition:** Gathering data from relevant sources (databases, APIs, etc.).

- **Cleaning:** Addressing missing values, inconsistencies, and formatting errors.
- **Integration:** Combining data from different sources into a unified format.
- **Transformation:** Feature engineering (creating new features) and data scaling/normalization if needed.

22. What are the key objectives of Exploratory Data Analysis (EDA)?

Answer: EDA aims to:

- **Gain initial understanding** of the data and its characteristics.
- **Identify patterns, trends, and relationships** between variables.
- **Discover potential outliers** or anomalies.
- **Inform decisions** about further data cleaning, feature selection, and model building.

23. Explain the concept of data summarization in EDA.

Answer: Data summarization involves calculating key statistical measures to describe the data:

- **Central tendency:** Mean, median, mode (indicate "average" values).
- **Spread:** Variance, standard deviation (measure how data points are distributed around the mean).
- **Shape:** Skewness, kurtosis (describe the symmetry or peakedness of the distribution).

24. How do you assess the distribution of data in EDA?

Answer: Techniques for analyzing data distribution include:

- **Visualizations:** Histograms, box plots, and density plots provide a visual representation of the data spread.
- **Descriptive statistics:** Measures like skewness and kurtosis quantify the distribution's shape (symmetrical, skewed, etc.).

25. Explain outlier treatment methods in data science.

Answer: When dealing with outliers (data points significantly different from the rest), you can:

- **Investigate the cause:** Understand if it's an error or a valid data point.
- **Winsorization:** Replace extreme values with values closer to the distribution's tail.
- **Capping:** Set a maximum or minimum value for outliers.
- **Removal:** Remove outliers if they're determined to be errors or irrelevant to the analysis (use with caution).

26. How can you measure asymmetry (skewness) in a data distribution?

Answer: Skewness measures the asymmetry of a distribution:

- **Positive skew:** The right tail is longer (distribution leans to the left).
- **Negative skew:** The left tail is longer (distribution leans to the right).
- **Zero skew:** The distribution is symmetrical.

27. What are continuous distributions in data science?

Answer: Continuous distributions describe data that can take on any value within a specific range. They are often represented by smooth curves. Examples include normal distribution, uniform distribution, etc.

28. Explain the concept of mean and variance in estimation statistics.

Answer: Mean and variance are used to estimate population parameters from sample data:

- **Mean:** The average value of a dataset (central tendency).
- **Variance:** The average squared deviation from the mean (measures spread).

29. What is the difference between sampling and the entire population in statistics?

Answer: Population refers to the entire collection of data points you're interested in. Sampling involves drawing a representative subset of the population for analysis.

30. Describe the concepts of covariance and correlation in data science.

Answer: Covariance and correlation assess the relationship between two variables:

- **Covariance:** Measures the direction and magnitude of the linear relationship (positive or negative association).
- **Correlation:** Represents the strength of the linear relationship between two variables (values range from -1 to 1). Correlation doesn't imply causation.
- odels (clustering, dimensionality reduction)
- Model evaluation and selection

31. What are point estimates in statistics?

Answer: Point estimates are single values used to estimate population parameters from sample data. Examples include:

- **Mean:** Average value of the data (central tendency).
- **Median:** Middle value when data is ordered.
- **Mode:** Most frequent value.

32. Explain the concept of confidence intervals in statistics.

Answer: Confidence intervals provide a range of values within which you are confident the true population parameter is likely to lie. They are constructed with a specific confidence level (e.g., 95%).

33. How are confidence intervals used in the frequentist approach?

Answer: Confidence intervals provide an uncertainty measure around a point estimate. They account for sampling variability and allow you to express the range of plausible values for the population parameter.

34. Describe the relationship between confidence level and confidence interval width.

Answer: As the confidence level increases, the confidence interval typically becomes wider. This is because you're capturing a larger range of possible values with higher confidence.

35. How can you use confidence intervals for hypothesis testing?

Answer: In some cases, you can use confidence intervals to draw conclusions about hypothesis testing. If the hypothesized value falls within the confidence interval, you fail to reject the null hypothesis (there's no significant difference).

36. Explain the concept of hypothesis testing in frequentist statistics.

Answer: Hypothesis testing is a statistical method to assess claims about a population parameter. It involves:

- **Null hypothesis (H0):** The default hypothesis, assuming no significant difference.
- **Alternative hypothesis (H1):** The hypothesis you aim to support, proposing a difference.

37. How do p-values play a role in hypothesis testing?

Answer: P-value represents the probability of observing a test statistic as extreme as the one obtained, assuming the null hypothesis is true.

- **Low p-value (e.g., < 0.05):** Evidence against the null hypothesis (reject H0).
- **High p-value (e.g., > 0.05):** Fail to reject the null hypothesis (insufficient evidence for H1).

38. What are some limitations of relying solely on p-values for hypothesis testing?

Answer: While p-values are helpful, they have limitations:

- **P-value depends on sample size:** Larger samples can lead to smaller p-values, even with weak effects.
- **Focuses on statistical significance:** A statistically significant result may not be practically important.

39. What are some best practices for interpreting hypothesis testing results?

Answer: Consider both p-value and effect size when drawing conclusions:

- **Effect size:** Quantifies the magnitude of the observed effect (e.g., difference between groups).
- **Consider context:** A statistically significant result may not be practically meaningful in all situations.

40. How does the frequentist approach compare to the Bayesian approach in statistics?

Answer: Frequentist statistics focuses on population parameters and p-values for hypothesis testing. Bayesian statistics incorporates prior knowledge and updates beliefs based on data, providing probability distributions for parameters.

41. Briefly describe supervised learning and its key applications.

Answer: Supervised learning involves training models on labeled data (inputs with known outputs) to make predictions for unseen data. It's used for tasks like:

- **Classification:** Categorizing data points (e.g., spam detection, image recognition).
- **Regression:** Predicting continuous values (e.g., stock price forecasting, sales prediction).

42. Explain the k-Nearest Neighbors (kNN) classifier algorithm.

Answer: kNN classifies data points based on the majority vote of their k nearest neighbors in the training data. It's simple to implement but can be sensitive to high dimensionality and noisy data.

43. How do decision trees work in supervised learning?

Answer: Decision trees are tree-like models that make predictions by asking a series of questions about the data. They are interpretable and robust to outliers but can be prone to overfitting.

44. What is the CART algorithm used for in decision trees?

Answer: CART (Classification and Regression Trees) is a specific algorithm for building decision trees. It uses a greedy approach to select the best splitting features at each node based on a purity measure (e.g., Gini impurity for classification).

45. Define the concept of regression analysis in supervised learning.

Answer: Regression analysis aims to model the relationship between a dependent variable (to be predicted) and one or more independent variables (predictors).

46. Differentiate between linear regression and logistic regression.

- **Linear regression:** Models the relationship between a continuous dependent variable and independent variables using a linear equation.
- **Logistic regression:** Used for classification tasks where the dependent variable is binary (e.g., 0 or 1). It predicts the probability of an event occurring.

47. What is the core idea behind unsupervised learning?

Answer: Unsupervised learning deals with unlabeled data, where the objective is to discover hidden patterns or structures within the data. Common applications include:

- **Clustering:** Grouping data points based on similarities.
- **Dimensionality reduction:** Reducing the number of variables in a dataset while retaining essential information.

48. Explain the k-means clustering algorithm.

Answer: k-means is a popular clustering algorithm that partitions data points into a predefined number of clusters (k). It iteratively minimizes the within-cluster sum of squared distances between points and their assigned cluster centers.

49. What are some common evaluation metrics used to assess the performance of machine learning models?

Answer: Evaluation metrics vary depending on the task (classification vs. regression):

- **Classification:** Accuracy, precision, recall, F1-score, confusion matrix.
- **Regression:** Mean squared error (MSE), R-squared (coefficient of determination).

50. How can you address the issue of overfitting in supervised learning models?

Answer: Overfitting occurs when a model memorizes the training data too well, leading to poor performance on unseen data. Here are some techniques to mitigate it:

- **Regularization:** Penalizing complex models to discourage overfitting (e.g., L1/L2 regularization).
- **Dropout:** Randomly dropping out neurons during training to prevent co-adaptation between features.
- **Cross-validation:** Evaluating model performance on separate validation sets to avoid overfitting to the training data.