Editorial-W7A2: Algorithm Analysis and Machine Learning Fundamentals

Question 1

Question: A software development team is analyzing the performance of their new algorithm. They want to understand how "Big O" notation applies to their analysis. What does "Big O" notation represent in this context?

(MSQ)

A) The best-case time complexity
B) The worst-case time complexity
C) The average-case time complexity
D) The upper bound of time complexity

Correct Answer: B,D

Explanation: Big O notation represents the worst-case or, upper bound of an algorithm's time complexity, indicating its worst-case performance.

Question 2

Question: A company's dataset is already sorted. You are working as an intern there and out of curiosity run bubble sort on this dataset. What would be the time complexity of the algorithm?

A) $O(n)$
B) $O(n \log n)$
C) $O(n^2)$
D) $O(1)$

Correct Answer: A

Explanation: The best-case time complexity of bubble sort is $O(n)$ when the list is already sorted.

Question 3

Question: A team of developers is designing an algorithm that involves a series of operations. They need to analyze its performance over time. What does the term "amortized analysis" refer to in this context?

A) Analyzing the worst-case scenario
B) Analyzing the average-case scenario over multiple operations
C) Analyzing the best-case scenario
D) Analyzing the space complexity

Correct Answer: B

Explanation: Amortized analysis considers the average-case scenario over multiple operations, providing a more realistic view of an algorithm's performance.

Question 4

Question: A data processing company needs to select an algorithm for sorting large datasets consistently. Which algorithm has a time complexity of O(n^2) in all cases (best, average, worst)?

A) Bubble Sort
B) Insertion Sort
C) Selection Sort
D) Binary search

Correct Answer: C

Explanation: Selection sort has a time complexity of O(n^2) in all cases because it always checks each element against every other element.

Question 5

Question: Let's suppose there are 3 algorithms named Blue, Orange, Green which performs a certain task you want to perform. And their respective time complexities are shown with the same colour in the attached image. Which is the best algorithm? (considering only the info given in question).

A) Blue
B) Orange C) Green D) None of the above

Correct Answer: A

Explanation: However large the input becomes, the time complexity for Blue Algorithm is constant which is less than for the other algorithms. Therefore Blue Algorithm is best.

Question 6

Question: Let's suppose there are 3 algorithms named Blue, Orange, Green which performs a certain task you want to perform. And their respective space complexities are shown with the same colour in the attached image. Which is the best algorithm? (considering only the info given in question).

A) Green B) Orange C) Blue
D) They all perform the same

Correct Answer: C

Explanation: However large the input becomes, the space complexity for Blue Algorithm is constant which is less than for the other algorithms. Therefore Blue Algorithm is best.

Question 7

Question: A software development team is using the step counting approach to compare the efficiency of different algorithms. However, they encounter several challenges. Which of the following are flaws in the step counting approach for comparing efficiency in this context?

A) Input Sensitivity
B) Compiler Optimizations
C) Hardware Dependency
D) None of the above

Correct Answer: A, B, C

Explanation:

- **Hardware Dependency**: The time it takes to execute a single line of code can vary greatly depending on the computer's processor, memory, and other hardware components. This makes it difficult to compare algorithms across different systems.

- **Compiler Optimizations**: The compiler (the program that translates your code into machine instructions) can optimize the code in various ways, making step counting inaccurate as the actual execution time may differ from the counted steps.

- **Input Sensitivity**: The number of steps can depend heavily on the specific input to the algorithm. For example, in a linear search, if the key is the first element in the array, the algorithm finds it quickly. If the key is the last element or not in the array at all, it takes much longer. This variability affects the reliability of step counting for efficiency comparison.

Question 8

Question: You are working as an intern at a real estate company and you want to predict house prices based on square footage, number of bedrooms, location, number of years built. What type of machine learning problem is this?

A) Classification
B) K-Means Clustering C) Regression D) PCA

Correct Answer: C

Explanation: Since the task involves predicting a continuous value (house prices), it is a regression task.

Question 9

Question: You are creating an application to predict if the image is of dog or cat (only these 2 images will be inputted). What type of machine learning problem is this?

A) Clustering
B) Classification
C) Linear Regression D) None of the above

Correct Answer: B

Explanation: This is a classification task because it involves categorizing images into "dog" or "cat".

Question 10

Question: You are working as an intern at a digital marketing agency and want to categorize customers based on their purchasing behavior without prior labels. Which machine learning technique should you use?

A) Supervised Learning
B) Unsupervised Learning
C) Reinforcement Learning
D) Neural Networks

Answer: B) Unsupervised Learning

1. Identify if labeled data is available.

2. Determine whether clustering or classification is needed.

3. Choose an appropriate learning method.

Explanation:
Since customer segmentation lacks predefined labels, unsupervised learning (e.g.- Clustering) is the best approach. It groups customers based on patterns without prior categorization.

Question 11

Question: A marketing firm is analyzing customer data to understand purchasing behavior. The dataset includes over 50 variables such as age, income, location, and purchase history. To simplify the analysis, they decide to use PCA. What is the primary goal of using Principal Component Analysis (PCA) in this case?

A) To classify customers into categories
B) To reduce the dimensionality of the dataset while preserving significant information
C) To increase the number of features in the dataset
D) To predict continuous values

Correct Answer: B

Explanation: PCA aims to reduce the dimensionality of a dataset by transforming it into a smaller set of new variables (principal components) that retain most of the original information.

Question 12

Scenario: A data scientist is working on a project involving a high-dimensional dataset with over 100 features. The model performance is suffering due to the complexity of the data.

What is the primary challenge posed by this high-dimensional dataset, and how can dimensionality reduction help?

A) The challenge is data visualization, and dimensionality reduction simplifies it.
B) The challenge is computational inefficiency, and dimensionality reduction increases computational speed.
C) The challenge is the increased risk of overfitting and data sparsity, and dimensionality reduction helps mitigate these risks.
D) The challenge is the lack of features, and dimensionality reduction adds more features.

Correct Answer: C

Explanation: The curse of dimensionality refers to the challenges posed by high-dimensional data, such as increased risk of overfitting and data sparsity. Dimensionality reduction techniques help mitigate these issues by simplifying the data.

Question 13

Scenario: A team of researchers is developing a predictive model for disease diagnosis based on a dataset with over 500 genetic markers. Next, they decide to apply feature selection. What is the primary purpose of feature selection in this scenario?

A) To increase the number of features in the dataset
B) To reduce model complexity and improve interpretability
C) To classify data into categories
D) To predict continuous values

Correct Answer: B

Explanation: Feature selection aims to select the most relevant features from a dataset to reduce model complexity, improve interpretability, and enhance performance.

Question 14

Scenario: A company is analyzing customer purchase data using both PCA and feature selection techniques. How do these two methods differ in their approach to handling the dataset?

A) PCA combines features to create new ones, while feature selection selects existing features.
B) PCA selects existing features, while feature selection combines features.
C) PCA is used for classification, while feature selection is used for regression.
D) PCA is used for regression, while feature selection is used for classification.

Correct Answer: A

Explanation: PCA creates new features (principal components) by combining the original features, whereas feature selection involves selecting a subset of the existing features.

Question 15

Scenario: A researcher is working with a large dataset containing many correlated features to predict customer behavior. In what scenario would PCA be particularly useful for this researcher?

A) When dealing with a small dataset with few features
B) When dealing with a large dataset with many correlated features
C) When the goal is to classify data into categories
D) When the goal is to predict continuous values

Correct Answer: B

Explanation: PCA is particularly useful when dealing with large datasets containing many correlated features, as it helps reduce dimensionality while preserving significant information.