國立成功大學

114學年度碩士班招生考試試題

編　號：189

系　所：數據科學研究所

科　目：計算機概論

日　期：0210

節　次：第 2 節

注　意：1.不可使用計算機
2.請於答案卷(卡)作答，於試題上作答，不予計分。

1. (32%) In the following statements, please specify if the statement is **True** or **False**. If the statement is True, explain why it is True. If it is False, give correct answer or explain why. (Each 2%)

   (a) In a binary search tree (BST), every node's left child holds a value less than the node's value, and every node's right child holds a value greater than the node's value.
   (b) In object-oriented programming, polymorphism allows multiple methods with the same name but different parameter lists or implementations.
   (c) In open addressing for hash tables, using double hashing completely eliminates collisions.
   (d) In multi-threading, if there is only a single shared variable, race conditions cannot occur.
   (e) Under Transmission Control Protocol (TCP), packet loss can be completely avoided by proper flow control settings.
   (f) A Turing Machine is a theoretical device that manipulates symbols on a tape according to a set of rules, and it forms a fundamental model for defining computability.
   (g) Binary search on a sorted singly linked list guarantees an $O(\log n)$ time complexity.
   (h) If problem A is NP-hard and problem B is reducible to A in polynomial time, then B must also be NP-hard.
   (i) In operating systems, a process in the "blocked" state is waiting for some event (such as I/O completion) to proceed.
   (j) In a relational database, the primary key of a table can be NULL for exactly one record.
   (k) In an undirected, weighted graph, Dijkstra's algorithm correctly handles negative edge weights.
   (l) In operating systems, "thrashing" occurs when excessive paging/swapping dominates CPU usage, leaving little time for real work.
   (m) In 2's complement representation for a 32-bit signed integer, the range is $-2^{31}+1$ to $2^{31}-1$.
   (n) An operating system can resolve deadlock automatically by letting one of the processes wait forever.
   (o) In a priority queue implemented via a binary min-heap, the element with largest priority is always at the root.
   (p) The IP protocol guarantees packets arrive in-order at the destination.

2. You are given a **directed, weighted graph**. The graph may contain **positive, zero, or negative edge weights, but no negative cycles** exist. Your tasks are described as follows:
   (a) **Sparse vs. Dense:** Propose an appropriate data structure to store a sparse version of the graph (i.e., relatively few edges compared to the number of vertices). Then explain how you would store the graph if it were dense (i.e., the number of edges is close to the maximum possible). Compare the space complexity of these two approaches, and justify why each is more suitable under different circumstances. (4%)

(b) **Shortest Path with Negative Edges:** Which single-source shortest path algorithm(s) can handle negative edges but still avoid issues with negative cycles? State the time complexity of the chosen algorithm(s) and briefly explain why a typical algorithm for non-negative edges (e.g., Dijkstra's) is not immediately applicable when edges can be negative. (4%)

(c) **Scalability:** Suppose the graph is very large (e.g., millions of vertices) but still mostly sparse. Discuss strategies to store and process this graph efficiently (consider factors such as memory usage, parallelization, or distributed computing). (4%)

3. Many data science workloads involve concurrent tasks (e.g., data preprocessing, model training). Compare the **process-based** and **thread-based** models of parallelism in an operating system:
   (a) What are the **main differences** in terms of memory sharing, context switching, and communication costs? (3%)
   (b) In a high-throughput data pipeline, which model might be **more efficient**, and under what circumstances? (3%)

4. When multiple machine-learning processes compete for limited resources (e.g., GPUs, large memory pools), deadlocks can occur.
   (a) Describe four conditions necessary for deadlock. (3%)
   (b) Propose a practical **deadlock prevention** or **avoidance** strategy in a GPU-based cluster running multiple training jobs simultaneously. (3%)

5. Describe the differences between the following pairs of terms.
   (a) Stacks vs. Queues (3%)
   (b) Pass-by-Value vs. Pass-by-Reference (3%)
   (c) Relational Databases vs. NoSQL Databases (3%)
   (d) Greedy Algorithms vs. Dynamic Programming (3%)

6. Insert the following 10 distinct integers into an initially empty Binary Search Tree (BST) in the given order:

   `[15, 9, 20, 6, 12, 17, 25, 11, 14, 19]`.

   (a) Draw the **resulting BST** after all insertions. Show intermediate steps (or at least the final tree structure). (3%)
   (b) Write down the **preorder**, **inorder**, and **postorder** traversals of the resulting BST. (6%)
   (c) Delete the key **9** and redraw or describe the updated BST. Why did you rearrange the nodes the way you did? (3%)

7. Answer the following questions on data science.
   (a) Compare **supervised** and **unsupervised** learning in terms of data requirements and typical algorithms. Give an example use case for each. (3%)
   (b) Describe the differences between **overfitting** and **underfitting** in predictive modeling. How can each phenomenon be detected, and what practical steps can be taken to reduce them? (3%)
   (c) Discuss the purpose of **train, validation,** and **test data splits**. How can an improper split (e.g., data leakage) distort a model's reported performance? (3%)
   (d) Differentiate a **parameter** (e.g., model coefficients in linear regression) from a **hyperparameter** (e.g., learning rate). How does each affect the model's behavior? (3%)

8. Given a dataset containing 7 data points with x1 and x2 as two features, shown as below.

Points	x1	x2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

   (a) If we set $K = 2$, the **initial centroids are Point 1 and Point 4**, and we are using **Euclidean distance**, what are the final outputs of **K-means clustering**? (3%)
   (b) When using Euclidean distance, the cost over iterations always decreases? Explain your answer. (2%)
   (c) Explain two different types of 2-dimensional data distributions (you can plot the data points) where K-means might fail to produce accurate clusters. (3%)