This page has been left blank intentionally.
Q1.

(a) 

- Explain the role of an online transaction processing (OLTP) database in a business.

- Explain clearly two example business scenarios that are not possible to analyze using an OLTP database.

(b) 

- Explain the role of a data warehouse and an online analytical processing (OLAP) database in a business.

- Explain clearly two example business scenarios that are not possible to analyze using a data warehouse.

(c) Design a data warehouse schema for a regional weather bureau. The weather bureau has about 1000 probes, which are scattered throughout various land and ocean locations in the region to collect basic weather data, including air pressure, temperature, and precipitation (rain fall) at each hour. All data are sent to the central station, which has collected such data for the last 10 years.

- Explain clearly the different dimensions and measures you have used in your schema.

- Assume that a data cube has been implemented based on your schema. Explain four interesting queries that can be answered using your data cube.
Q2.

(a)  
• Explain the two measures *support* and *confidence* for an association rule.
• Explain why the measure *support* alone is not sufficient to determine the importance of an association rule.

(3)

(b)  
• Explain the *Apriori* algorithm for finding frequent itemsets.
• Explain clearly how the *Apriori* algorithm reduces computation cost using the *apriori* principle.

(7)

(c)  
• Explain clearly the major steps in *decision tree classification*.
• Why is *tree pruning* useful in decision tree induction?

(5)
Q3.

(a) Explain clearly with examples when *supervised* and *unsupervised* learning algorithms are used.

- The *support vector machine* (SVM) is a highly accurate classification method. However, SVM classifiers suffer from slow processing when training with large datasets.

Explain the SVM algorithm intuitively and discuss why it is inefficient.

(b) Explain the k-Medoids algorithm for clustering.

- How does the k-Medoids algorithm improve upon the k-Means algorithm?

(c) Explain the DBSCAN algorithm for density-based clustering of high-dimensional data sets.

- What are the main advantages of this algorithm?

(d) Explain intuitively the architecture of a *neural network* and the steps required to design a classifier based on a neural network.

- What are the main advantages and disadvantages of a neural network as a classifier?