Short Questions

1. What is data mining? What are the various applications of data mining.
2. What is an outlier? Explain with one example.
3. How to handle missing values in a dataset?
4. What do you mean by normalization? Explain Z-score normalization with example.
5. What is 68-95-99.7 rule?
6. What is the difference between Business Intelligence and data science? Explain with one example?
7. What is data warehouse? Why do we need data warehouse?
8. What are the applications of data warehouse?
9. What is the difference between OLTP and OLAP?
10. Explain OLTP and OLAP with one example from each.
11. What is ETL process?

Long Questions

1. What is data mining? Explain KDD process with suitable diagram?
2. What is data preprocessing? Explain various steps (6 steps) of data preprocessing.
3. What are the various types of attributes in data mining concept. Explain each one with example.
4. What is sampling? Explain any 2 probabilisic and 2 non probabilistic sampling methods with examples.
5. What do you mean by normalization? Explain decimal scaling, z-score and min-max normalization with example.
6. Find the Eucledian, Minkowski distance and cosine similarity distance of the following dataset.

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

7. What is data warehouse? Explain the three tier data warehouse architecture with diagram.
8. What is the difference between database and data warehouse?
9. What is the difference between OLTP and OLAP?
10. What is the difference between transactional processing and analytical processing in DMDW?
11. What is data mart? Explain various types of data marts.
12. The 'database' below has four transactions. What association rules can be found in this set, if the minimum support (i.e coverage) is 60% and the minimum confidence (i.e. accuracy) is 80% ?

Transaction ID	Item List
T1	K, A, D, B
T2	D, A, C, E, B
T3	C, A, B, E
T4	B, A, D

Questions on Association Rule Mining

Q1. You are given the transaction data shown in the Table below from a fast food restaurant. There are 9 distinct transactions (order:1 – order:9) and each transaction involves between 2 and 4 meal items. There are a total of 5 meal items that are involved in the transactions. For simplicity we assign the meal items short names (M1 – M5) rather than the full descriptive names (e.g., Big Mac).

Meal Item	List of Item IDs	Meal Item	List of Item IDs
Order:1	M1, M2, M5	Order:6	M2, M3
Order:2	M2, M4	Order:7	M1, M3
Order:3	M2, M3	Order:8	M1, M2, M3, M5
Order:4	M1, M2, M4	Order:9	M1, M2, M3
Order:5	M1, M3		

For all of the parts below the minimum support is 2/9 (.222) and the minimum confidence is 7/9 (.777). Note that you only need to achieve this level, not exceeds it. Show your work for full credit (this mainly applies to part a).

a. Apply the Apriori algorithm to the dataset of transactions and identify all frequent kitemsets. Show all of your work. You must show candidates but can cross them off to show the ones that pass the minimum support threshold. This question is a bit longer than the homework questions due to the number of transactions and items, so proceed carefully and neatly.

   [Note: if a candidate itemset is pruned because it violates the Apriori property, you must indicate that it fails for this reason and not just because it does not achieve the necessary support count (i.e., in these cases there is no need to actually compute the support count). So, explicitly tag the itemsets that are pruned due to violation of the Apriori property. This really did not come up on the homework because those problems were quite short. (If you do not know what the Apriori property is, do not panic. You will ultimately get the exact same answer but will just lose a few points).

b. Find all strong association rules of the form: X ^ Y -> Z and note their confidence values. Hint: the answer is not 0 so you should have at least one frequent 3-frequent itemset.

Q2. The 'database' below has four transactions. What association rules can be found in this set, if the minimum support (i.e coverage) is 60% and the minimum confidence (i.e. accuracy) is 80% ?

Transaction ID	Item List
T1	K, A, D, B
T2	D, A, C, E, B
T3	C, A, B, E
T4	B, A, D

Q3. Consider a transactional database where 1, 2, 3, 4, 5, 6, 7 are items.

Transaction ID	Items
T1	1, 2, 3, 5
T2	1,2, 3, 4, 5
T3	1, 2, 3, 7
T4	1, 3, 6
T5	1, 2, 4, 5, 6

Suppose the minimum support is 60%. Find all frequent itemsets. Indicate each candidate set Ck, k = 1, 2… the candidates that are pruned by each pruning step, and the resulting frequent itemsets Lk.

-Best of Luck-