Data Mining Homework 4

(60 Points)

This homework covers association rule mining, Naïve Bayes, and clustering.

1. Answer the following questions given the data in the Table below. (10 points)

Customer ID	Transaction ID	Items Bought
1	0001	$\{a, d, e\}$
1	0024	$\{a, b, c, e\}$
2	0012	$\{a, b, d, e\}$
2	0031	$\{a, c, d, e\}$
3	0015	$\{b, c, e\}$
3	0022	$\{b, d, e\}$
4	0029	$\{c, d\}$
4	0040	$\{a, b, c\}$
5	0033	$\{a, d, e\}$
5	0038	$\{a, b, e\}$

   a. Compute the support for itemsets {e}, {b,d}, and {b,d,e} by treating each transaction ID as a market basket.

   b. Use the results in part (a) to compute the confidence for the association rules {b,d} → {e} and {e} → {b,d}.

   c. Repeat part (a) by treating each customer ID as a market basket. Each item should be treated as a binary variable. Note that in this case multiple transactions need to be merged together.

   d. Compute the confidence for the association rules {b,d} → {e} and {e} → {b,d}

2. (20 points) 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 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 exceed 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 k-itemsets. Show all your work. You must show candidates but can cross them off to show the ones that pass the minimum support threshold.

   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 So, explicitly tag the itemsets that are pruned due to violation of the Apriori property.

b. Find all *strong* association rules of the form: $X \wedge Y \rightarrow Z$ and note their confidence values. That is, there should be two items on the left and one on the right. Hint: the answer is not 0 so you should have at least one frequent 3-frequent itemset.

3. (20 points) A competing used car dealership that sells US, European, and Japanese cars is trying to install a machine learning system that will automatically detect whether a car is stolen or not given such parameters as: the car's size, type and the country of origin. Below is the data that the dealership has already collected:

Instance ID	Size	Type	Country	Stolen?
1	Large	Family	USA	**Yes**
2	Large	Family	USA	**No**
3	Large	Family	Europe	**Yes**
4	Medium	Family	Japan	**No**
5	Medium	Family	Europe	**No**
6	Medium	Luxury	Japan	**No**
7	Medium	Luxury	Japan	**Yes**
8	Medium	Luxury	USA	**Yes**
9	Large	Luxury	Japan	**No**
10	Large	Family	USA	**Yes**

a) Fill-in the following table. Leave answers in fractional form (8 points)

Size	
P(Large|No) =	P(Large|Yes) =
P(Medium|No) =	P(Medium |Yes) =
Type	
P(Family|No) =	P(Family|Yes) =
P(Luxury |No) =	P(Luxury |Yes) =
Country	
P(USA|No) =	P(USA|Yes) =
P(Europe|No) =	P(Europe|Yes) =
P(Japan|No) =	P(Japan |Yes) =
Stolen	
P(No) =	P(Yes) =

b) Using the above table estimate whether the car with the following parameters that just arrived at the dealership is stolen or not. **Show your work.** (12 points)

Large, Luxury, USA

4. (10 points: 2 point each) Here are several short questions on clustering.

    a. List one significant commonality between clustering algorithms and instance based learning algorithms like nearest-neighbor.

    b. A decision tree can be used to generate a partitional clustering. How?

    c. Will outliers have a big impact on the K-means algorithm? Why or why not?

    d. Will outliers have a big impact on the DB-scan algorithm?

    e. The outcome of K-means is often determined by the random points that are used to start the algorithm. How can this be addressed?