James Mortensen

CIS 600-Assignment 3

Show your work for all answers.

1. 50 pts. From *Tan et al. text Exercise 5.2. Association rules*
   Consider the table below.

	Customer ID	Transaction ID	Items Bought
1	1	0001	{a, d, c}
2	1	0024	{a, b, c, e}
3	2	0012	{a, b, d, e} ✓ ✓
4	2	0031	{a, c, d, e}
5	3	0015	{b, c, e}
6	3	0022	{b, d, e} ✓ ✓
7	4	0029	{c, d}
8	4	0040	{a, b, c}
9	5	0033	{a, d, e}
10	5	0038	{a, b, e}

$$\text{support(itemset)} = \frac{\text{\# of transactions containing set}}{\text{total \# of transactions}}$$

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

$\{c\} \rightarrow 6 \rightarrow$ $\boxed{6/10 = .6 \rightarrow 60\%}$

$\{b,d\} \rightarrow 2 \rightarrow$ $\boxed{2/10 = .2 \rightarrow 20\%}$

$\{b,d,e\} \rightarrow 10 \rightarrow$ $\boxed{2/10 = .2 \rightarrow 20\%}$

(b) Use the results in (a) to compute the confidence for the association rules {b, d} -> {e} and {c} -> {b,d}. Is confidence a symmetric measure?

- Confidence$(\{A\} \rightarrow \{B\})$ = Support$(\{A,B\})$ / Support$(\{A\})$
- Support$(\{b,d,e\}) = .2$, Support$(\{b,d\}) = .2 \rightarrow$ Confidence$(\{b,d\} \rightarrow \{e\}) = .2/.2 = 1$
- Support$(\{b,c,d\}) = 0$, Support$(\{C\}) = .6 \rightarrow$ Confidence$(\{c\} \rightarrow \{b,d\}) = 0/.6 = 0$

(c) Repeat (a) by treating each customer ID as a market basket. Each item should be treated as a binary variable (1 if an item appears in at least one transaction bought by the customer, and 0 otherwise).

Market Basket: $\text{support(Itemset)} = \frac{\text{\# of Customers buying all items in itemset}}{\text{Total \# of customers}}$

Support$(\{c\}) = \frac{4}{5} = .8 \rightarrow 80\%$

support$(\{b,d\}) = \frac{2}{5} = .4 \rightarrow 40\%$

support$(\{b,d,e\}) = \frac{2}{5} = .4 \rightarrow 40\%$

(d) Use the results of (c) to compute the confidence of the association rules {b, d}->{e} and {e}->{b,d}

{b,d} → {e}: .4/.4 → 1

{e} → {b,d}: .4/.8 = .5

(e) Suppose s_1 and c_1 are the support and confidence values of an association rule r when treating each transaction ID as a market basket. Also, s_2 and c_2 are the support and confidence values of an association rule *r* when treating each customer ID as a market basket. Discuss whether there are any relationships between s_1 and s_2 or c_1 and c_2.

There is generally no relationship or correlation between s_1 and s_2 or c_1 and c_2, the relationship depends on distribution of items among transactions and customers. Comparison of these values may yield insights into customer buying patterns. High s_1 compared to s_2 suggests customers buy items in an itemset across multiple transactions.

2. From *Tan et al. text Exercise 5.6. Association rules*

Consider the table below.

Transaction ID	Items Bought
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{~~Milk, Diapers~~, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{~~Milk, Diapers~~, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

(a) What is the maximum number of associations rules that can be extracted from this data (including rules that have zero support)?

$n = 6$ unique items: Milk, beer, diapers, bread, butter, cookies

$3^n - 2^n \rightarrow 3^6 - 2^6 = 665$

(b) What is the maximum size of frequent itemsets that can be extracted (assuming minsup > 0)?

- size of largest transaction

→ 4

(c) Write an expression for the maximum number of size itemsets that be dervived from this data set.

n = unique items

k = size of itemset

n choose k → ${}_nC_k$

${}_6C_k$

(d) Find an itemset (of size 2 or larger) that has the largest support.

- Bread and Butter appear in more transactions together than any other size 2 or larger item set.

(e) Fine a pair of items, a and b, such that the rules {a} -> {b} and {b} -> {a} have the same confidence.

{Milk, diapers}, {Bread, butter}, {Beer, diapers} → 4

{Milk, Bread}, {Milk, butter}, {Diapers, Bread} {Diapers, Butter} → 3

Diapers → 7

Milk → 5

Bread → 5

Butter → 4

Beer → 3

Cookies → 3

→ None have the same confidence

{A} → {B} vs {B} → {A}

2. 50 pts. *From tan et al text Exercise 3.12. Learning objective is to show understanding of classifier performance analysis.*

Consider a labeled data set containing 100 data instances, which is randomly partitioned into two sets A and B, each containing 50 instances. We use A as the training set to learn two decision trees, T_{10} with 10 leaf nodes and T_{100} with 100 leaf nodes. The accuracies of the two decision trees on data sets A and B are show in the table below.

	Accuracy	
Data Sets	T_{10}	T_{100}
A	0.86	0.97
B	0.84	0.77

Based on the accuracies shown above, which classification model you expect to have better performance on unseen instances?

T_{10} will do better against unseen instances, given it performs approximately the same with the B data set as with A, the training data. T_{100}'s accuracy drops too much with the testing data, suggesting overfitting to the training set

Now, you have tested T_{10} and T_{100} on the entire data set (A + B) and found the classification accuracy of T_{10} on the entire set (A + B) is 0.85, whereas the classification accuracy of T_{100} on the data set (A + B) is 0.87. Based this new information and your observations from the table, which classification model would you finally choose for classification?

T_{10} is still the better choice; with the combined data it is clear that the accuracies are comparable. It is important to note that A+B is half seen by both models, T_{10} still performs much better with unseen data.