Assignment 5

Note: Show all your work as you have on previous assignments.

Problem 1 (25 points). Consider the following transactional database.

TID	Items
100	2, 3, 4, 5, 6, 8
200	1, 2, 3, 5, 6
300	1, 4, 5, 7, 8
400	2, 3, 4, 5, 6
500	1, 2, 3, 4, 5, 7
600	1, 3, 8

(1) Mine all frequent itemsets using the Apriori algorithm, which we discussed in the class, with the minimum support = 50% (or 3 or more transactions). Show all candidate itemsets and frequent itemsets. You should follow the process described in the book and lecture (i.e., C1 → L1 → C2 → L2 → …). You don't need to show pruning steps. To save your time, L1 is given below:

L1:

Itemset	1	2	3	4	5	6	8
Count	4	4	5	4	5	3	3

(2) Sort all frequent 4-itemsets by their item number. Then, select the first frequent 4-itemset form the sorted list of frequent 4-itemsets and mine all strong rules from this itemset that have the format {W, X} => {Y, Z}, where W, X, Y, and Z are individual items. Assume that minimum confidence = 80%.

Problem 2 (25 points). Consider the following training dataset, which is used for classification:

A1	A2	A3	Class
High	On	True	Positive
High	On	False	Positive
Low	Off	True	Negative
High	Off	True	Negative
Low	On	False	Positive
High	Off	True	Positive
High	On	False	Negative

You can generate classification rules from the above dataset using the Apriori algorithm, which we discussed in the class.
(1). Execute the Apriori algorithm on the above dataset with the minimum support = 40% or 3 transactions. You need to proceed as we discussed in the class, i.e., C1 -> L1 -> C2 -> L2 -> . . . You need to show all candidate itemsets, frequent itemsets, and all rules mined from the dataset.
(2). Show only the rules that can be used for classification and calculate their confidences.

You must run the Apriori algorithm yourself as we discussed in the class (i.e., you should not use Weka, JMP Pro or any other software to run an Apriori algorithm on the given dataset). You need to show all intermediate steps.

Problem 3 (25 points). Consider the following contingency table.

	C (buys coffee = Yes)	$\overline{C}$ (buys coffee = No)
T (buys tea = Yes)	142	862
$\overline{T}$ (buys tea = No)	186	1859

Compute the *lift*, *all-confidence*, *cosine*, *Kulczynski* and *imbalance ratio* measures, and determine whether buying coffee and buying tea are positively correlated, negatively correlated, or not correlated. You must show all calculations.

Problem 4 (25 points). You will perform association analysis using JMP Pro. There is a section in *Predictive and Specialized Modeling.pdf* documentation that shows how to perform association analysis. You may want to read this section before starting the assignment. Follow the instructions in *JMP-association-analysis-assignment.pdf* file.

Submission Guidelines:

- **Submit the solutions in a single Word or PDF document and upload it to Blackboard. Make sure the filename includes your name.**
- **Your submission should be organized well and easy to follow. Clearly list which question you are answering and provide the answer requested (one or more numbers, a screenshot, a plot, an explanation, whatever the question asks you to do). If a question contains multiple parts, you should clearly provide your answer for each part.**
- **If a problem requires you to work out formulas or calculations, you should provide this detailed work either in an Appendix in your Word/PDF or as a separate file. All separate files should include your name in the filename.**
- **If you need more than 3-4 separate files, you might ZIP/RAR these files. Please still upload the Word/PDF separately. The separate Word/PDF greatly simplifies the facilitator's ability to comment on your assignment.**
- **If your facilitator tells you to submit the files differently than the above guidelines, you are expected to respect your facilitator's wishes starting on the next assignment.**
- **Facilitators can deduct up to 20% if you fail to follow these requirements (more if the questions are not actually answered).**
- **Facilitators can deduct 5% for each day the assignment is late.**
- **Unless your facilitator or the professor agrees, your assignment will not be graded if it is more than 3 days late (e.g., no credit will be given after Friday at 6 AM Boston time). The professor will usually ask the facilitator to make the decision but in rare cases (<1% of the time) has overridden a facilitator. Do not expect the professor to override in most cases.**