CS 699
Assignment 5

Note: R is recommended since that is what is taught in class, but you may use any software you choose. As explained in submission guidelines on the next page, please provide your "final answers" in a document and provide a separate file containing your code or other detail that explains how you obtained your answers.

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). Use the dataset *basketanalysis.csv*. This dataset has 999 observations and 17 features.

(1). Load the data, discard the transaction id feature, and convert the data frame from strings containing "True" and "False" to a matrix of logical values (TRUE and FALSE). Then convert this matrix to a transactions object. Then use the an apriori rule miner with a minimum support of 15% and a minimum confidence of 50%. How many rules are mined?

(2). Determine which rule has the highest confidence. What is this rule? Capture the portion of the output that states the rule, as well as the support, confidence, coverage, and lift.

(Assignment continues on the next page)

(3). Return to the matrix of logical values. For the rule you identified in (2), find the number of transactions with both the antecedent and consequent itemsets, the number of transactions with the antecedent itemset, and the number of transactions with the consequent itemset. Use these values to compute the coverage (LHS-support), support, confidence, and lift for the rule. Provide any code you use to do these calculations. They should match the results displayed in step (2).

Problem 3 (20 points). Use the dataset *bookratings-small.csv*. This dataset has 139984 observations and 3 features: a user ID code, a book code, and the rating.

(1). Load the data and force the data.frame to a realRatingMatrix object. Fit a user-based collaborative filtering model to all but 3 of the users (you can choose which three to hold out).
(2). Use the model to make predictions for the three users you held back.

Problem 4 (10 points). Consider the following A/B testing scenario for a manufacturing plant which is having a problem with a customer refusing to buy their product because too many of the items produced have a manufacturing defect. The plant makes some changes to the production process and uses A/B testing to determine whether they have made a significant reduction of the defect rate.

Before the changes, the plant manufactured 200 products for the customer, but 20 of them had defects. After making their process improvements, they made another 250 products, and only 10 of them had defects. It appears the defect rate has decreased. Use a statistical test to determine whether the reduction is statistically significant.

Problem 5 (20 points). Use the dataset *uplift-small.csv.* This dataset has 28123 observations and 9 features. The class feature is named conversion and the treatment variable is named offer.

(1). Generate training and holdout partitions on the data set. Use 1/3 of the data in the holdout.
(2). Fit a random forest model, applying a training grid to tune the parameters. Use the random forest model to make probability-metric predictions on two versions of the holdout data: one with the offer (treatment) feature set to Discount and the other with the offer feature set to No Offer. Then compute the uplift for the treatment. Obtain Q1, the median, and Q3 for the predicted uplift. Make conclusions.

Submission Guidelines:

- **Submit the solutions in a single Word or PDF document and upload it to Blackboard. You should separately upload the code or another document that explains how you arrived at the answers. Make sure each filename includes your name.**

    - **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.**
- **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 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.**