Homework 3

Due by: 4/3/2019, 10:00AM

Instructions

1. Put all your answers and results in a **single Microsoft Word document** with your name, and save all the R codes used for this assignment in a **single R Script file** with your name. **Submit your Word file and R file into eLearning under "Assignments"**. Grading will be based on the answers and results provided in your **Word document**. We check you R codes only when we see any potential problems (e.g., something suspicious). Correct R codes without good answers in the Word file receive no credits.
2. **Late submissions are not acceptable** and will be rejected by eLearning.
3. A professional quality report is expected–messy or hard-to-read reports will be penalized.
4. Explain your answers. **Be as clear as possible**. Vague answers–even if they are long–will not receive full credit. Information in excess of what the question warrants is acceptable as long as it is relevant and correct. Incorrect information, even if unwarranted, will be penalized. Therefore, **proofread your report to tidy it up before submission**.

Questions

1. In Fall 2018, UTD opened a new buffet where there are many food selections for faculty and students. For simplicity, suppose five types of foods are offered daily: salad, hamburger, taco, soup and pasta. Suppose you are the manager and you decide to use associate rules **(manually)** to figure out what foods customers tend to purchase together. You recorded selections by five customers as shown in the Table below. You also decide to use the following cutoffs: minimum support 40% and minimum confidence 80%. What valid rules will you generate? Provide detailed steps with your relevant calculations. Also report support, confidence and lift for the final rules you generate. **(2.5 Points)**

Customer ID	Food
1	Salad, Hamburger, Taco
2	Soup, Hamburger, Pasta
3	Salad, Soup, Hamburger, Pasta
4	Soup, Pasta
5	Taco, Pasta, Soup

2. The following contingency table summarizes supermarket transaction data. (It is similar in format to the table you see on slide 20 in our lecture notes on association rules. ∑ means sum by row or column.)

	Hot dog	No hot dog	∑
Hamburger	1800	700	2500
No hamburger	1200	1300	2500
∑	3000	2000	5000

(a) Suppose that the association rule "hot dogs ⇒ hamburgers" is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, is this association rule valid? **(0.5 Point)**

(b) Based on the given data, is the purchase of hot dogs independent of the purchase of hamburgers? If not, what kind of correlation relationship exists between the two (i.e., if a customer purchases hot dogs, will that increase or decrease her chance of purchasing hamburgers)? (**1 Point)**

3. Conducting an Association Analysis Using R: A store is interested in determining the associations between items purchased from the Health and Beauty Aids department and the Stationery Department. The store chose to conduct a market basket analysis of specific items purchased from these two departments. "transactions" contains information about over 400,000 transactions made over the past three months. The following 17 products are represented in the data set: bar soap, bows, candy bars, deodorant, greeting cards, magazines, markers, pain relievers, pencils, pens, perfume, photo processing, prescription medications, shampoo, toothbrushes, toothpaste, and wrapping paper. **(4 Points)**

There are four variables in the data set:

Name	Model Role	Data Type	Description
Store	Ignore	Numeric	Identification number of the store
Customer	ID	Numeric	Customer identification number
Product	Target	Categorical	Product purchased
Quantity	Ignore	Numeric	Quantity of this product purchased

Use R to generate Association Rules (based on the code shown in class):

(a) Import the data to R. Copy the R code used below. (Tip: use read.transactions)

(b) Set Support to 0.01, Confidence to 0.10, and Min Length to 2. Run apriori to obtain the rules. Sort the rules according to "Lift" with descending order. Copy the R code used below.

(c) Show the top ten Association Rules. Copy the code used and the result below

(d) What is the highest lift value for the resulting rules? Which rule has this value? Show how this lift value was calculated.

(e) Interpret the first five rules in the output in words.

(f) Reviewing the top 10 rules, based on their lift ratios, comment on their redundancy and how you would assess their utility as a decision maker.