Homework 11 Possible Answers

Abra Brisbin

2021-08-05

Important:

This document contains R code solutions and example answers for problems posed in the DS740: Data Mining homework. We are intentionally sharing this document with learners *who have already completed the associated homework.* We want you to be able to review and troubleshoot your code, as well as ask questions to prepare you for later assessments. *By using this document, you are accepting responsibility* **not** *to share it with anyone else, including other students in the course or the program who might not have completed the homework yet.* By upholding this agreement, you are helping us use this tool with learners in future terms.

From Problem 1:

Question

Make a barplot of the relative frequencies of each item, including all items that were purchased by at least 5% of the customers.

Possible Answer:

```
itemFrequencyPlot(Groceries, support = .05)
```

Question

Mine all simple association rules with at least .001 support and at least 0.5 confidence. (Recall that a “simple” association rule is one with a single item in the consequent, which is the type of rule that `apriori` returns.)

Possible Answer:

```
rules = apriori(Groceries, parameter = list(support = .001, confidence =
0.5))
```

Question 6

Print the 10 rules with the highest lift values.

Possible Answer:

```
sub_rules = head(rules, n = 10, by = "lift")
inspect(sub_rules)
```

Question

How many customers in the data set purchased soda, popcorn, and a salty snack?

Possible Answer: (Support of the rule {soda, popcorn} -> {salty snack}) * (Number of customers in data set)

Question

Return to working with the original set of rules from the third question in this problem. Filter out any rules that have lower confidence than more general versions of the same rules. How many rules remain?

Numeric Answer (AUTOGRADED on Canvas):

```
non_redundant = !is.redundant(rules)
rules_non_red = rules[non_redundant]
summary( rules_non_red )

# Could also use
#non_redundant = (interestMeasure(rules, measure = "improvement",
transactions = NULL, reuse = TRUE, quality_measure = "confidence") > 0)
#rules_non_red = rules[non_redundant]
```

Question

Suppose that you work for a baking company, and you want to offer a coupon to customers who are likely to buy pastry. Using your filtered rules from the previous question, identify combination(s) of items that are associated with an increased probability of buying "pastry".

Possible Answer:

```
pastry_rules = subset( rules_non_red, subset = rhs %in% c("pastry") )
inspect(pastry_rules)
```

From Problem 2:

Question

A value of **ChestPain** equal to 4 indicates that the patient is asymptomatic (does not have any kind of chest pain). Create a new variable, **hasCP**, which equals 1 for all individuals with chest pain, and which equals 0 otherwise. Make it a factor variable. Remove the original **ChestPain** variable.

Possible Answer:

```
hd = read_csv("../HeartDisease.csv")
hd <- hd %>%
  mutate(hasCP = factor(ifelse(ChestPain == 4, 0, 1))) %>%
  select(-ChestPain)
```

Question

Discretize **Age** into 3 ordered categories, using equal interval lengths.

Possible Answer:

```
hd$Age = discretize(hd$Age, breaks=3, ordered=T, method="interval")
```

Question

Discretize **BloodPressure** into 3 ordered categories with fixed boundaries determined by the first and third quartiles of the data.

Possible Answer:

```
bp_cutoffs = quantile(hd$BloodPressure, probs = c(0, .25, .75, 1)) #94, 120, 140, 200
hd$BloodPressure = discretize(hd$BloodPressure, "fixed", breaks=bp_cutoffs, ordered=T)
```

Question

Tell R to treat the other variables (which you didn't create using `discretize()`) in **HeartDisease.csv** as discrete factors. Create a data frame containing the discrete versions of all of the variables (it should have 14 columns). Then convert the data frame to a format suitable for association rule mining.

Possible Answer:

```
hd <- hd %>%
  mutate(across(c(-Age, -BloodPressure), factor))

hd_trans = as(hd, "transactions")
```

Question

Which **two** of the following risk factors, along with being female, are associated with the greatest elevation in probability of heart disease?

- Reversible defects on a thallium heart scan (**Thal** equal to 7)
- No chest pain

Notes: The elevation in probability of heart disease is the lift, so this can be determined using the code provided below.

```
sub_rules = head(female_rules, n = 1, by = "lift")
inspect(sub_rules)
```

Note that this gives risk factors that give a "lift" above the baseline of the entire sample, not just compared to the women in the sample. To find which variables are risk factors specifically for women, we could repeat the association rule mining on the subset of data including only women.

Question

There are 2340 rules in which being male is an antecedent. Why are there fewer rules in which being female is an antecedent? Give **two** reasons.

Possible Answer:

1.Being male is a risk factor for heart disease. This has been reported in various studies, and it can be seen in this data set: The rule “Sex=1” -> “hasHD=1” has a lift of 1.2.

```
hd_rules = apriori(hd_trans, parameter = list(support = .03, confidence =
0.5, maxlen = 2),
                   appearance = list(rhs = c("hasHD=1"), default = "lhs"))
inspect(hd_rules)
```

2.There are more men than women in the sample, so there are fewer combinations of “female + other variables” that have sufficient support to be considered as possible rules.