CISC-483-683
Homework 9: Association Rules

Due: Wednesday, Nov. 20, 2019

60 points

1. Suppose that you want to generate association rules with minimum support 25% and minimum accuracy 90%. You have the following dataset:

HAIR	UPKEEP	SHOTS	AGE	HOUSEPET
short	med	some	young	yes
short	high	none	young	no
short	tremendous	all	med	ok
med	tremendous	some	med	ok
horrid	tremendous	some	med	no
med	tremendous	all	old	no
med	low	some	med	ok
long	med	some	young	yes
short	med	none	med	no
long	med	some	old	no
short	low	none	young	no
med	med	all	med	ok

The frequent 1-itemsets are the following:

HAIR=med
HAIR=short
UPKEEP=med
UPKEEP=tremendous
SHOTS=none
SHOTS=some
SHOTS=all
AGE=young
AGE=med
HOUSEPET=ok
HOUSEPET=no

(a) (20 points) Using the Apriori algorithm, generate the frequent itemsets. For each k ($k > 1$) you should list the candidate itemsets and beside each, write one of the following:

pruned-A	itemset is pruned due to Apriori principle
pruned-D	itemset is pruned due to insufficient support from dataset
frequent	itemset is a frequent k-itemset

The elements of your candidate itemsets **MUST** be listed in alphabetical order. For example, a 2-itemset composed of Shots=some and Age=young should be listed as {Age=young,Shots=some}, NOT as {Shots=some,Age=young}.

(b) (20 points) Generate strong association rules from the frequent 3-itemsets. For each frequent 3-itemset, give the candidate rules that are considered along with their accuracy, and then give the final set of strong association rules that are constructed from that itemset.

2. (20 points) Weka: Use the dataset Mushroom-assoc-19.csv that can be found on the class web page at
www.cis.udel.edu/∼carberry/CISC-483-683
Use the Apriori algorithm to extract association rules. Set the Apriori parameter *outputItemSets* to *true* so that the frequent itemsets are output. Leave the other parameters in their default settings.

   (a) Develop the best 5, then best 15, and then best 100 rules. How do the resultant rules differ in terms of range of support for the rules and range of accuracy (confidence) for the rules?

   (b) What does the paraemter *lowerBoundMinSupport* tell you about the minimum coverage. Be specific — you can get this from looking at the description of the parameter and by viewing the output.

   (c) What does the parameter *minMetric* tell you about which rules are generated? Be specific — you can get this from looking at the description of the parameter and by viewing the output.

   (d) We are using *confidence* (which I have referred to as *accuracy*) as our metric. Look at the parameter *metricType* and describe one other metric that could have been used instead of *confidence*.