CS-422 Assignment 3

Association Rule Mining

1. Use Weka
    - Use "Associate" tab

PAGE Limit for this part is 2. Additional page penalty is 2 points per page.

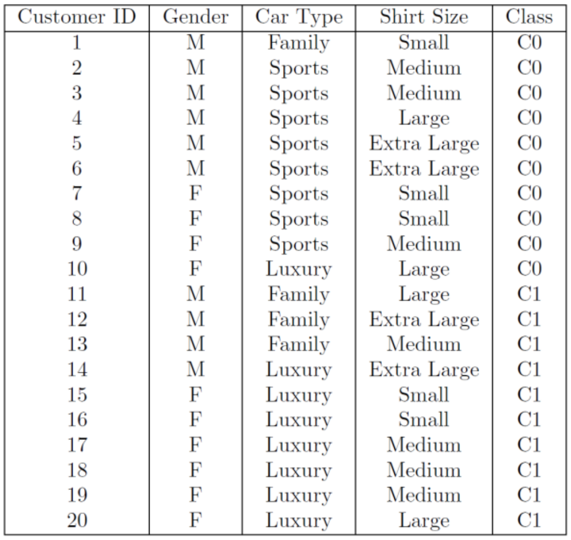
1.1 Supermarket dataset: Show the top 4 association rules with Apriori using the default parameters. Discuss what are the main parameters for Apriori that you can modify. Modify support/confidence 3 times: low, medium high, experiment and report the summary of what your learned - explain briefly how your modifications affected the generation of the rules

1.2 Use the attribute selection tab to remove attributes that appear in most rules. Run the same experiments as above and report what you observe.

1.3 Use vote dataset: Select all attributes in the attribute selection tab. Note that the class label will be used as one of the attributes or "items" and each record is a "market basket". Compute the association rules with the default parameter settings. Look at the right hand size of the top rules, compare the attributes in the association rules to the most important attributes that you computed for the decision tree classifier in HW 2.

2 Chapter 4, Decision Trees. Explain every answer briefly in YOUR OWN WORDS, just a short answer without explanation will be zero point. This is a very good preparation for the midterm and final.

2.1.

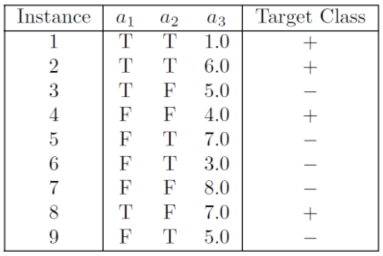


Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

In the table above, you have training examples from the car data set. The last column contains the class label, C0 or C1. Using this table answer the following questions. These questions are based on our discussions in class. In particular in (f) answer based on the analysis we were doing in class for our decision tree examples.

a) Calculate the Gini index for all examples. That is the Gini index at the root of the decision tree.
b) Calculate the Gini index for the attribute CustomerId.
c) Calculate the Gini index for the attribute gender.
d) Calculate the Gini index for the attribute car type.
e) Calculate the Gini index for the attribute shirt size.
f) Make the conclusion. What attribute will you use as the first split attribute of your decision tree? Why? Explain your expectations of the usefulness of each attribute based on the inspection of the table. Consider the distribution of the values of each attribute by the class type and briefly analyze it. Explain how the Gini index reflects what you observe.

2.2.

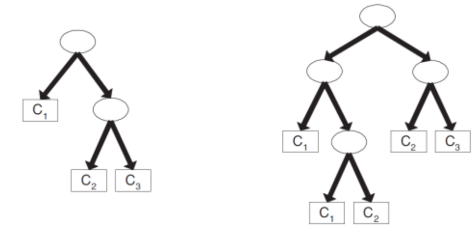


Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

Based on the table above, answer the following:

a) What is the entropy of all examples, with respect to the class "+"? Explain the values – is it large or small? What does it mean for the entropy? Explain in a few words, what does this mean for your analysis of the collection – base it on our discussion how the entropy reflects how expected is the class label, or to say it differently, what is the element of surprise if we see the class label of an example.
b) Calculate the information gain of attribute a1 and of attribute a2. Construct the table with the values of a1 per class label and do the same for a2. Explain how you are using that table in your calculation. How do you calculate the probabilities needed for your IG calculation? Provide the equation that you use in your calculation and explain EACH variable in it.
c) What attribute has a better IG and should be used as the split attribute? Explain briefly but clearly, how the IG reflects the distribution of labels that you observe on the full data set compared to the distribution of labels that you would see after splitting on that attribute. The distribution of labels will be contained in the table you construct in (b).

2.3.



(a) Decision tree with 7 errors

(b) Decision tree with 4 errors

a) Consider these 2 decision trees. They are constructed from data with 3 classes and 16 binary attributes. Calculate the cost of each tree using the MDL principle. What tree should you choose in term of the expected generalization error? Explain briefly but clearly why? Base your answer on the discussions we had in class.

b) Explain how you will use the calculations from (a) in the decision tree post-pruning process. Explain briefly what post-pruning is and why we use it. Outline each step in the decision tree induction process with post-pruning and what will be the resulting tree. You don't have to repeat all steps of the tree induction process, you can mention it as just one step in your outline.