Lab 3
Experimenting with the 1R Algorithm

Overview: This week we see the 1R algorithm in action, and conduct some experiments to see how it works. Additionally, we review some of our skills from previous activities – including the construction of ARFF files and decision trees – and take a first look at creating association rules.

1. Data preparation review

In one of our problem sets we learned how to construct an ARFF file out of `.data` and `.names` files – one of the most common formats for datasets on the Internet. Since this is an important skill to have when we start running experiments on datasets of our own choosing, we'll practice it again today. This time we'll use another common format for datasets – Microsoft Excel worksheets. You have been e-mailed two such files, entitled `CreditCardPromotion.xlsx` and `Titanic.xlsx`. Save these to your USB drive and open them to inspect the contents.

a. Using your knowledge of the ARFF format from the text, the Weka sample files, and your previous experimentation, you could easily convert both of these files into new files `CreditCardPromotion.arff` and `Titanic.arff`. You should try this for practice, but there is a nice shortcut offered by Weka:

   If you choose FILE > Save As in Excel and select the file type `CSV (Comma delimited)`, the result will be a file with the extension `.csv`. If you inspect the contents of this file using a text editor (not Excel), you will find it to be pleasantly similar to the format that you already need for ARFF. Even more pleasantly, Weka can open `.csv` files directly. Give it a try with both files.

b. Be sure both files load successfully into Weka. Then save them each to your USB drive in the `.arff` format.

c. You want `Life Ins Promo` to be the class attribute for the Credit Card Promotion file, and `Survived` to be the class attribute for the Titanic file. You don't want to keep re-setting this over and over, so edit the files as necessary so that these will be the default class attributes. You should remember the details of the `.arff` format (from problem set #2) that will allow you to do this.

2. The 1R Algorithm

a. Load the credit card dataset into Weka. Make sure that the life insurance promotion attribute is the class attribute. Under the "Preprocess" tab click **Visualize All**. Based on the histogram information displayed, which attribute would be you guess will be the most influential in predicting the class attribute? Justify your answer. It's okay to be wrong, as long as you give an adequate justification.

b. Go to the "Classify" tab. Choose the *OneR* algorithm under *weka.classifiers.rules*. Keep the default settings, and use the training set under "Test options." Again, make sure that `LifeInsPromo` (or whatever you named it) is still the class attribute. (If this isn't the last attribute listed in your ARFF file, Weka frequently changes the class attribute.)

c. Run *OneR* on the dataset. What rule was generated? Based on what we've learned in lecture so far, why was this attribute selected above all the others? What is the accuracy percentage when tested on the training data?

d. Now click on the name of the 1R filter and change `MinBucketSize` to 1. What rule was generated? How does the accuracy change on the training data? Can you explain what happened here?

e. Return `MinBucketSize` back to 6. This time use a 66% percentage split as the training option. What happens to the accuracy percentage?

f. Re-set the Test options to "Use training set." Try setting each of the other six attributes as the class attribute in turn. Record the attribute/rule that was selected in each case, along with the test accuracy. Which input attribute is most predictive of which output (class) attribute? Can you give a logical, real world explanation for this?

g. What happens when `age` is chosen as the output/class attribute? Why?

h. Now choose the J48 classifier algorithm, using a 66% split to test. Be sure to re-set `LifeInsPromo` as the class attribute. Run the algorithm and then right-click on the latest model in the Result List and visualize the decision tree. Which attributes are used in the tree? The answer might seem odd to you. Think about why I might say that.

i. What is the accuracy of this model compared to the 66% split test of 1R? What do you conclude from this?

3. Association Learning: a first look

As we have learned, unsupervised association learning has the objective of generating rules which suggest strong associations between different attributes of the input dataset. These rules are differentiated from classification rules by the absence of a designated class attribute. The *Apriori* algorithm is the primary means for mining such rules. We'll learn all about this algorithm later, but now let's give it a test drive.

a. Switch to the "Associate" tab. The **Start** button is grayed out for the *Apriori* algorithm. This is because the algorithm requires that all attributes be nominal. Can you think of a reason for this?

b. Go back to the "Preprocess" tab. From *weka.filters.unsupervised.attribute* select the *Discretize* filter. Click **Apply** and then look at the `Age` attribute. What changed?

c. Return to the "Associate" tab. You should understand why the **Start** button is now an option again. Run the *Apriori* algorithm with the default settings, and look over the rules generated. Record three that you find interesting, and give a good real world explanation for why each might be true.

4. "I'm the king of the world!"

a. Now load the `Titanic.arff` file that you have prepared. Edit the dataset to make sure that no errors were introduced in your file conversion process. Make sure the attribute `Survived` is the class attribute.

b. Run *OneR* on the dataset with 66% split. Which single attribute had the most impact on who survived? Explain. How does the rule perform?

c. Run *J48* on the dataset. How does the tree perform with 66% split? What can you conclude based on the comparison of this figure to the performance of *1R*?

d. Visualize the tree. Would you predict that a male adult in second class would survive or not?

e. Run *Apriori* on this dataset and once again record three rules that you might consider to be interesting, along with a plausible real world explanation for why each rule is true.

Lab 3: The 1R Algorithm

Name: ____________________________________

Exercise 1.c: What steps did you take to set the default class attribute in the credit card promotion dataset?

Exercise 2.a: Which attribute do you predict will be most influential, and why?

Exercise 2.c: What's the rule to predict the class attribute? What criteria was used to select it? What is the accuracy percentage on training data?

Exercise 2.d: In one word, what happened here?

Exercise 2.e: What is the accuracy percentage with 66% split?

Exercise 2.f:

Output (Class) Attribute	Rule Generated by 1R	Accuracy
IncomeRange		
MagazinePromo		
WatchPromo		
CreditCardIns		
Sex		
Age		
LifeInsPromo		

Which input attribute is most predictive of which output attribute? What is your explanation for this?

Exercise 2.g: Explain what happens when `Age` is chosen as the class attribute.

Exercise 2.h: What attributes does J48 use in the decision tree for this problem?

Exercise 2.i: How does the accuracy compare to 1R? What do you conclude from this?

Exercise 3.a: Why does the *Apriori* algorithm require nominal values? (Think about the rules that might be generated for numeric values.)

Exercise 3.b: What did the *Discretize* filter do to the `Age` attribute?

Exercise 3.c: Identify the three most interesting association rules from the credit card promotion dataset, and explain each in words.

Exercise 4.b: Which rule does 1R generate to predict survival? Give a good explanation for this. What is the 66% split accuracy of 1R?

Exercise 4.c: How does J48 perform with 66% split accuracy? What can you conclude from the comparison of this figure to the accuracy of 1R?

Exercise 4.d: What does the tree predict would happen to an adult male in second class?

Exercise 4.e: Identify the three association rules from the Titanic dataset that you find most interesting, and explain in words the possible correlation behind each.