CS 345: Data Mining   Names:_____

In-Class Activity 3   _____

Work through the Tutorial Exercises for the Weka Explorer. Put your solutions for each exercise below.

Exercise 17.1.1. What are the values that the attribute *temperature* can have?

Exercise 17.1.2. Load a new dataset. Click the *Open file* button and select the file *iris.arff*, which corresponds to the iris dataset in Table 1.4. How many instances does this dataset have? How many attributes? What is the range of possible values of the attribute *petallength*?

Exercise 17.1.3. What is the function of the first column in the Viewer window?

Exercise 17.1.4. What is the class value of instance number 8 in the weather data?

Exercise 17.1.5. Load the iris data and open it in the editor. How many numeric and how many nominal attributes does this dataset have?

Exercise 17.1.8. How would this instance be classified using the decision tree?
outlook = sunny, temperature = cool, humidity = high, windy = TRUE

Exercise 17.1.9. Load the iris data using the Preprocess panel. Evaluate C4.5 on this data using (a) the training set and (b) cross-validation. What is the estimated percentage of correct classifications for (a) and (b)? Which estimate is more realistic?

Exercise 17.1.10. Use the *Visualize classifier errors* function to find the wrongly classified test instances for the cross-validation performed in Exercise 17.1.9. What can you say about the location of the errors?

Exercise 17.6.1. Based on the output, what is the support for this item set?
outlook = rainy humidity = normal windy = FALSE play = yes

Exercise 17.6.2. Suppose you want to generate all rules with a certain confidence and minimum support. This can be done by choosing appropriate values for *minMetric*, *lowerBoundMinSupport*, and *numRules*. What is the total number of possible rules for the weather data for each combination of values in Table 17.6?

Exercise 17.6.3. Run *Apriori* on the weather data with each of the four rule-ranking metrics, and default settings otherwise. What is the top-ranked rule that is output for each metric?

Exercise 17.6.4. Run *Apriori* on this data with default settings. Comment on the rules that are generated. Several of them are quite similar. How are their support and confidence values related?

Exercise 17.6.5. It is interesting to see that none of the rules in the default output involve *Class = republican*. Why do you think that is?

Exercise 17.6.6. Experiment with *Apriori* and investigate the effect of the various parameters described before. Write a brief report on the main findings of your investigation.