COMPUTER SCIENCE 4373
Assignment #3

Points: 100 **Weight**: 3%
Due: Friday, March 2, 2018 at 2:00 pm in class as well as on BlackBoard
Note: Late assignment will not be accepted without instructor's pre–approval.

Instruction: This assignment should be completed individually. Please make sure your answer is legible (and preferably formatted using MS Words or LaTeX/LyX). If a question requires you to follow an algorithm, show a clear trace of the algorithm. If the algorithm is iterative, show the details in the first two iterations. For each of the remaining iterations, show the status of the algorithm at the end of the iteration. Please also submit to the BlackBoard a single your_name_hwk03.zip file that contains a PDF or a Word version of your solutions (**no scanned image please**), and the source code, the input data, and the output of your program.
The following questions are based on this dataset (table) from an employee database.

department	status	age	salary	count
sales	senior	31..35	46k..50k	30
sales	junior	26..30	26k..30k	40
sales	junior	31..35	31k..35k	40
systems	junior	21..25	46k..50k	20
systems	senior	31..35	66k..70k	5
systems	junior	26..30	46k..50k	3
systems	senior	41..45	66k..70k	3
marketing	senior	36..40	46k..50k	10
marketing	junior	31..35	41k..45k	4
secretary	senior	46..50	36k..40k	4
secretary	junior	26..30	26k..30k	6

The data is a summary of the original data table. For example, the first row indicates that 30 employees in the sales department has an age between 31 and 35 inclusive and a salary between 46K and 50K inclusive. The attribute status is the class label.

1. [**30**] Explain how the basic decision tree algorithm can be extended to incorporate the ranges (for Age and Salary) and the counts (in the count column) into the calculation of the impurity measures. Then, use the extended algorithm by hand to find the best split of the given data using the following impurity measures. You only need to show how to find the best split at the root node of the decision tree. You need to show the details of the calculations for at least one attribute, and show the results for the rest of attributes. You may want to write a program to perform the calculations. If you do so, also hand in your program source code.

   (a) information gain

   (b) gain ratio

(c) gini index

2. [**30**] Extend the Naive Bayes classifier algorithm so that it can also incorporate the ranges and counts in calculation of the probabilities.

   (a) Show how the extended algorithm would calculate the prior probabilities and the conditional probabilities $P(A_k \mid C)$ using the data table as the training data

   (b) Show how the extended algorithm would determine the status of the following data tuple

   $$t =< department : systems, status :?, age : 28, salary : 50K >$$

   Again, you need to show the details of the calculation for some of the probabilities, and for tuple t.

3. [**20**] Use this dataset to create a suitable new data file, either hwk03.arff or hwk03.csv, by replicating each row with the number of copies as indicated in the count column. For example, you should make the first row in the given table appear 30 times in the new table. Then, remove the count column.

   Write a program that trains a decision tree using the new data file as the training data and use the decision tree to predict the status of a user provided unseen data, for example,

   $$t =< department : systems, status :?, age : 28, salary : 50K >$$

   Specifically, you either write a Java program that uses Weka's J48 or a Python Jupyter notebook that uses SciKit-Learn's DecisionTreeClassifier to learn the decision tree. Notice that SciKit-Learn requires to encode categorical attributes as integer attributes.

   You may have to convert the actual age and salary into the corresponding ranges for the decision tree to work on the unseen data.

4. [**20**] Make another new dataset (named hwk03-02.arff or hwk03-02.csv) from the data file obtained in the previous exercise by converting the values in the age and salary columns to random values drawn from the specific range for each row. For example, suppose the age of a row is "31..35", replace it by a random integer between 31 and 35 inclusively.

   Write a program that uses either Weka or SciKit-Learn to learn a Naive Bayes classifier and use it to find the status of a user provided unseen data, for example,

   $$t =< department : systems, status :?, age : 28, salary : 50K >$$