CS 599 – Information Retrieval, Homework 4

- Download and install the latest version of WEKA to your computer
- For this assignment you will be working with "Explorer" interface and "KnowledgeFlow" interface, just like we did during the Weka Workshop.
- You will work with the Adult (http://archive.ics.uci.edu/ml/datasets/Adult) dataset from the US Census Bureau. This dataset, as well as other datasets frequently used for machine learning experiments can be downloaded from the Univ. of California Irvine (UCI) Data Repository. The adult.data file contains labeled examples: one example per line. The adult.names file contains information on the attributes used in the data. Note that these files are used by C4.5, not Weka, and hence you will need to generate an arff file before you can use this data. But the names file should be understandable just by reading it. Download the adult data and convert it to arff format, using information in the .names file, and save it on your computer.

  If you get completely stuck, you can search for a version of adult.arff on the web. But I strongly encourage you to try it by yourself first and spend time on the conversion process. The reason is that for your future work you may not have an .arff file ready and you may be forced to do the conversion by yourself. Either way, make sure that you understand the file. If you do the conversion simply by copying the file from the web, you still need to write down the steps for the conversion, or you will lose points for Step 1 on the homework.

  Required Steps for the Homework Assignment

  Note: Handwritten submissions will not be accepted. Please submit a hardcopy of your homework on or before the due date. Use the numbering below when answering the questions. In some cases it may make sense to include some of the figures generated by Weka into your submission. You can make a copy of a screenshot or individual graphic in WEKA by using Alt+Shift+Left-Click. You can then save that into a file, you can copy the image and then paste it into your submission.

1. (10 points) Briefly describe how you generated the arff file for the adult data set.
2. (10 points) After reading in the adult data via the arff file to "Explorer", while in the preprocess tab, look at how each feature correlates with the class variable (you can do this all at once by clicking on the "Visualize All" button. List 3 features that you think will be useful for predicting the class variable and, in one or two sentences, justify why you think these features will be useful.
3. (15 points) Run Weka's J48 classifier on the initial data with the test option set to 66% so that 66% of the data is used for training and the rest is used for test. Answer the following questions
    a. What is the accuracy of the classifier on the test data?
    b. How many leaves were there in the tree?
    c. How long did it take to build the model?
    d. Copy and paste the confusion matrix

4. (15 points) On the initial unaltered data set, run the ZeroR classifier, which can be found under the rules classifiers in WEKA. Answer parts a, c, and d from Question 3 for this classifier and characterize the differences between these results and those for Question 3.
5. (20 points) On the unaltered data set, use J48 to build a classifier to predict “sex” rather than the default class (named “class” but represents income level). You can do this by using the “Classify” tab to select J48, as usual, and then you can go to the left column to the button under the “More Option …” button and manually change the class/target variable from “class” to “(Nom) sex”. As before, use 66% for training data. Answer the questions from parts a-d for Question 3. Then answer the following additional question
    a. Based on the results in the confusion matrix, specify the number of females and males in the test set as counts (whole numbers) and as percentages.
6. (30 points) Now open the Weka “KnowledgeFlow” interface. Design a KnowledgeFlow to load the adult.arff file and train/test with three classifiers J48, ZeroR and NaiveBayes using Cross Validation. Configure your ClassAssigner with “class” and ClassValuePicker with class value “>50”. Your design should include ModelPerformanceChart to create all 3 classifier ROC curves visualized in a single chart (See related paper from our discussion on 2/15 for a sample ROC curve). Save and print the image from the ModelPerformanceChart. Include both the screenshot of the KnowledgeFlow design and the image of the ROC curve chart with your submission.