CISC-483-683: Assignment 3: Decision Trees

50 points

Due Monday, Sept. 23, 2019

1. Consider the following values for the numeric attribute SCORE and the following class values for the response attribute DECISION:

SCORE	DECISION
105	Reject
110	Admit
115	Admit
120	Admit
140	Admit
146	Wait
150	Admit
156	Admit
160	Admit
165	Admit
171	Reject
171	Wait
172	Wait
172	Admit
173	Wait
175	Wait
180	Reject
184	Admit
185	Wait

   (a) The possible values for the attribute SCORE range from 101 to 200. For equal interval binning with 4 bins, which instances go in each bin?

   (b) For equal frequency binning with 4 bins, which instances go in each bin?

   (c) Suppose that you are using top-down, supervised, entropy-based discretization. Which is a better first split point, 148 or 168? Show your work.

   (d) Suppose that you are using bottom-up, supervised, chi-square based discretization. Suppose that you have already constructed the following intervals: (-∞,104), (105,148), (148,168), and (168,∞).

      i. Using a confidence level of .90 (ie., want 90% confidence that we can reject the null hypothesis that the intervals are similar in terms of their class values), do you merge the interval (148,168) with the interval (168,∞)? Show your work.

      ii. Using a confidence level of .90, do you merge the interval (105,148) and the interval (148,168)? Show your work.

2. Use the file crx-partial-19 that is available from the class web site. The class attribute is CREDIT.

   (a) Using 10-fold cross-validation, use Weka to develop and evaluate a model built using the J48 decision tree classifier.

      i. What is the accuracy of the model?

      ii. (CISC-683 only) Use the Weka information and output to show that numeric attributes are being handled by top-down local discretization. (You will need to examine what Weka produces and make an informed judgement — no help will be given on this.)

(b) Now we want to do global discretization of numeric attributes. But we need to make certain that the discretization is done on the training data only, and then these same discrete classes are used on the test data. (We don't want to re-discretize on the test data.) So do the following:

- Click on `Choose` and double-click on `meta` under that, then select `FilteredClassifier`.
- Now when you get back to the main Explorer window, you will see "FilteredClassifier" to the right of "Choose". This filtered classifier is Weka's means of enforcing discretization based solely on the training data. Click on `FilteredClassifier`.
- The window that appears has a box labelled `classifier` and another box labelled `filter`; the former lets you choose a classification method and the latter lets you choose a global discretization method. Use 10-fold cross validation, J48 as the classification method, and the following discretization methods: supervised Discretize, unsupervised Discretize, and unsupervised PKIDiscretize. (Once you select *Supervised* or *Unsupervised*, you will need to click on *Attribute* to get a list of the possible discretization methods.)
  i. For each discretization method, give what kind of discretization the method does (such as equal frequency binning), the accuracy rate achieved using this discretization method, the number of nodes in the decision tree, and the depth of the tree.
  ii. Experiment with different numbers of bins for the unsupervised Discretize filter. How many bins give the best result?