IS 467

Assignment 3

Total number of points: 55 points (+5 bonus available)

Submission Instructions

1. Save your solutions with **clearly marked problem numbers**, clear and succinct writing, all software **output** and any **code** if applicable into a single PDF file (you can write in Word and save to PDF).
2. Submit your file online at the course website at **http://d2l.depaul.edu** and double-check it.
3. Keep a copy of all your submissions!
4. If you have questions about the homework, email me BEFORE the deadline.
5. The assignment will lose 20%, if submitted after the due date.
6. Assignments submitted after the class session following the due date will not be accepted.
7. Ask me ahead of time if you need an exception.

Problem 1 (20 points): This problem illustrates the classification approach by using decision trees and the Lupus data (download the data file "sledata" from D2L). The data consists of 300 patient records. Each record contains 12 elements. The first 11 elements stand for different symptoms and the final element of each record indicates the diagnosis. Build a decision tree and answer these questions.

1) Build the best decision tree you can and explain what makes it the best. Show what criteria you used including the number of cases allowed in parents and children and the depth and stopping condition.
2) How many nodes does the final tree have and how many of them are terminal nodes?
3) What are the most important three Lupus data features in building the tree? Explain your answer.
4) Increase the parameters that let you set the number of cases allowed in parent and child nodes. What do you notice with the complexity (number of nodes) of the tree? Does it increase? Explain your answer.

Problem 2 (30 points): This problem illustrates the effect of the class imbalance of the accuracy of the decision trees. Download the red wine quality data from the UCI machine learning repository at:
http://archive.ics.uci.edu/ml/datasets/Wine+Quality
(for a reminder of how to get this kind of data ready for SPSS, see Assignment 1)

1. Consider each quality level of wine to be a different class. Report how many classes there are and what is the distribution of these classes for the red wine data (how many cases of each class are there).
2. Repeat **Problem 1** on the red wine data.
3. Now bin the class variable in such a way that data is not so imbalanced with respect to the class variable. Repeat **Problem 1** but on the data you have processed with this smoothing.
4. How does the performance of the best classification model on the original class variable compare with the accuracy of the best classification model on the binned classification variable?
5. Do you have any other ideas on how you can improve the results further?
   Showing that your idea will actually work will be graded with five **extra credit** points.

Problem 3 (5 points): Differentiate between the following terms:

a. feature selection and feature extraction
b. training and testing data
c. parametric reduction techniques and non-parametric reduction techniques
d. uniform binning and non-uniform binning
e. covariance matrix and correlation matrix