CS753 Assignment 4: Classification

Due on Nov 13, 2018

Notes:

- This is an independent, individual assignment. Group work is prohibited.
- Copying others' work is considered cheating.
- The homework is worth *5 points* toward the final grade.
- The homework will be due at the beginning of the class on the due date. No late submission is accepted.
- Submit the printout of your homework. Do NOT submit it via email.
- Make sure you have your name on the printout.

Question 1:

A data mining routine has been applied to a transaction dataset and has classified 88 records as fraudulent (30 correctly so) and 952 as nonfraudulent (920 correctly so). Construct the classification matrix and calculate the following metrics:

- Accuracy and error rate
- Sensitivity and specificity
- False positive and false negative rates

Question 2:

The file eBayAuctions.xls on the Blackboard contains information on 1972 auctions transacted on eBay.com during May-June 2004. The goal is to use these data to build a model that will classify competitive auctions from noncompetitive ones. A competitive one is defined as an auction with at least two bids placed on the item auctioned. The data include variables that describe the item (auction category), the seller (his/her rating), and the auction terms that the seller selected (auction duration, opening price, currency, day-of-week of auction close). In addition, we have the price at which the auction closed. The goal is to predict whether or not the auction will be competitive.

- Data preprocessing
    a. Add the source data from the spreadsheet to Local Repository > Data.
    b. Drag the ebay data from the Data folder into your Process/Design canvas.
    c. Use the *Numerical to Binominal* operator to change "Competitive?" variable to binominal (true/false). Set both min and max to be 0.0.
    d. Set the "Competitive?" variable's role to be label (use the *Set Role* operator).

e. Use the *Split Data* operator to partition the dataset into training (60%) and validation (40%) sets.

- Classifier construction
    a. Fit a classification tree using all predictors, setting the maximum depth to be 7. Copy and paste the tree to your homework printout.
    b. Derive and write up all the rules from this tree, and make sure to report the smallest set of rules required for classification. For example, if a rule is like "IF A > 3 and A > 5 THEN …" you should consolidate it to "IF A > 5 THEN …"
    c. Add the *Apply Model* operator and connect the mod port from *Decision Tree* to its model port and the second par port to its unl port. Add a *Performance* operator at the end. Run the process and report the confusion matrix.
    d. Is this model practical for predicting the outcome of a new auction? Which variables should not be used as predictors?
    e. Fit another classification tree this time only with predictors that can be used for predicting the outcome of a new auction. Set the maximum depth of the tree to be 7. Copy and paste this tree. Derive all the rules from this tree, and make sure to report the smallest set of rules required for classification.
    f. Add the *Lift Chart* (simple) operator. Copy and paste this chart. Draw the baseline (random prediction). What can you say about the predictive performance of this model?
    g. Based on this tree, what can you conclude from these data about the chances of an auction obtaining at least two bids and its relationship to the auction settings set by the seller? What would you recommend for a seller as the strategy that will most likely lead to a competitive auction?