IE 583 Homework 2

This homework assignment is due on Tuesday, March 6th, by the end of the day (midnight).

Vote Data

For this assignment you will use association rule mining to discover interesting relationships in the Vote data. Recall that this dataset includes the votes of every U.S. House of Representatives Congressmen for 16 votes cast in the 2nd session of 1984. This dataset has a class attribute, which identifies party affiliation (Republican versus Democrat) but since this set of attributes (votes) will never repeat, predicting future data is not the key issue, but rather obtaining insights from the voting record.

Association Rule Mining in R

You should use R for this analysis, and the analysis should include the following:

1. Identify the "best" association rules involving any attributes
2. Identify the "best" association rules that describe a Democrat, that is, rules that include the item "Class=Democrat" in the consequent.
3. Identify the "best" association rules that describe a Republican, that is, rules that include the item "Class=Republican" in the consequent.

The determination of "best" should include some combination of our standard measures of association rule quality. The exact combination cannot be determined beforehand.

Report

You are expected summarize your results in a report that uses direct output from R sparingly. (E.g., use figures etc. from the software, but I'm more interested in your interpretation and analysis than screenshots.) There is no fixed page length for the report.

Types of Analysis

In addition to simply reporting rules, you should do analysis and interpretation of the rules that you obtain. In other words, focus on what the rules mean and how useful they are in providing insights.

This analysis should include (but is not limited to) the following:

- *Comparison to decision trees*. Since there is a class attribute (Republican/Democrat), you can also learn a decision tree from this data. Each leaf node in the decision tree represents a rule. Are these rules also learned through your association rule analysis? Does association rule mining find new and/or different rules? What is the advantage and drawbacks of each approach?
- *Evaluation of rule quality*. For every association rule that you include in your report, you should include the *support*, *confidence* and *lift* of the rule. [It is acceptable to replace lift with another measure of correlation if you prefer.]
- *Visualization of association rules*. You should use one or more graphical representation of rules to help the reader identify the quality of the rules that you report and make comparison between the rules that you report.
- *Comparison of association rules*. Are any of the rules generated redundant, that is, they are simply implied by other rules? If so, they shouldn't be included in your final output of useful rules. What relationships exist between the rules that you have chosen to report?