36-350: Data Mining

Homework 6
Date: October 5, 2001 **Due: start of class October 12, 2001**

1. A retailer surveys 500 customers on their breakfast habits. The number of people found to eat any yogurt is 300 and the number eating any cereal is 375 (these are marginal totals). Of these, 200 were found to eat both yogurt and cereal. We want to describe the association between yogurt and cereal.

   (a) Using the support/confidence framework, what is the support of the prediction rule "eating yogurt implies eating cereal"? What is its confidence?

   (b) What is the marginal probability of eating cereal? Is "eating yogurt implies eating cereal" a good rule?

   (c) What is the support and confidence of "eating yogurt implies not eating cereal"?

   (d) Which of these rules does the support/confidence framework prefer?

   (e) Compute the lift of (yogurt,cereal). Which rule does it prefer?

2. A phone company wants to know which of its customers are likely to switch ("churn") to another carrier. You can help them by summarizing the relationship between a customer's history and the probability of churn.

   (a) The file `churn-cat.dat` is a four-way contingency table describing customers. The dimensions are "intl" (international plan, yes or no), "vmail" (voice mail plan, yes or no), "state" (US state), and "churn" (yes or no). Source the `crosstab2` package and read the table using `read.crosstab`. Use `mine.associations` to find the major associations between "churn" and the other three variables. What are the top three associations?

   (b) The file `churn-svc.dat` is a table listing the number of customer service calls made by customers who stayed versus churned. What is the relationship between the number of service calls and the probability of churn?

   (c) It seems that the relationship can be summarized by abstracting `service.calls` into two bins. Make an appropriate mosaic plot and suggest what the bins should be.

   (d) Use `merge.table` to abstract `service.calls` into two bins. Remember that `service.calls` is ordered. Does it agree with your answer in part (c)?

   (e) The file `churn-day.dat` is a table listing the monthly charge for daytime minutes of customers who stayed versus churned. Use `merge.table` to abstract `day.charge` into 5 bins. Why is five an interesting number of bins?

   (f) Make a mosaic plot of the merged table, showing how churn probability varies with `day.charge`. Describe the relationship.

3. The file `adult-marital.dat` has census data on adults, relating age to marital status.

   (a) Use `merge.table` to simultaneously abstract the age dimension into four bins and marital status into four bins. Remember that age is ordered but marital status is not.

   (b) Make a mosaic plot of the merged table. For each marital status group, briefly explain the trend with age.

   (c) What are the two largest associations found by `mine.associations`? Summarize what they say about age and marital status.

4. A biology researcher is trying to study patterns in gene sequences. The file `gene1.dat` is a sequence of 369 amino acids which come from part of the human genome. There are 20 possible acids and each is coded by a different letter. You are going to perform market basket analysis to find unusually common acid pairs. The file `gene1-pair.dat` is a contingency table of adjacent acids. That is, the count for (A,G) is the number of times A was immediately followed by G in the sequence.

   (a) List the five acid pairs with highest lower bound on lift. You may use `mine.associations`. Describe, as if to the researcher, what the highest lift value means.

   (b) For the pair with highest lower bound on lift, what is the actual number of times that pair occurred? How do you explain its lift value?

   (c) Use `merge.table` to merge the two most similar rows. Are they the same as the acid pair with highest lift? Explain any discrepancy.