36-350: Data Mining

Homework 1
Date: August 31, 2001 **Due: start of class September 7, 2001**

1. (Definition of data mining) How much data mining is involved in the following tasks? How could the tasks be generalized to involve more data mining?

   (a) An electric company needs to predict the amount of energy demand for each day. Energy demand is known to depend on certain factors: the day of the week, holidays, and the season, as well as growing steadily with time.

   (b) It is possible to lower your airfare when taking multiple flights by nesting the trips inside each other. This produces a specific pattern in your itinerary that is easy to spot. An airline company wants to find this pattern automatically in their bookings database.

2. (Visualization and transformation) An experiment was conducted to measure the effectiveness of different insect sprays. Each of the six sprays was tried twelve times. The results are provided in the file `InsectSprays.dat`. You can read it into S or R via `read.table("InsectSprays.dat")`. Each row in this file is an insect count (high count is bad) for a particular trial of a particular spray.

   Find a transformation in the root/reciprocal family to roughly equalize the spread of the spray batches. (You can use `tapply` to compute the `mad` or `IQR` of the batches.) Make a plot of the batches which clearly but fairly reveals the differences between them. Minimize the effort of the viewer to understand the plot. Please submit your code.

3. (Basic probability) In the orchard sprays example (day 2 slide 39), the spread of each batch increases with the median of the batch. This is not a property of the particular sprays but the nature of the experiment. It often happens when the response (loss, in this case) is a sum of positive contributions. Argue, if each bee takes away a random, independent amount of sugar, that a more repelling spray must lead to a lower median as well as a lower spread in the amount of sugar taken away.

4. (Classification) In this problem, you will classify a newsgroup article into "politics" vs. "religion". There are 5,586 different words, listed in the file `words`. The file `politics` contains a vector reporting the total count of each word in the labeled politics articles. For example, the first number is the count for the first word, "abandoned." Similarly, the file `religion` has counts for the religion articles. The file `test` contains a vector of word counts, in the same order, for a new article. You can read these into S or R using `scan`. You may want to consult the manual section on "vector arithmetic."

   What is the log-likelihood for each class? Which is the most likely class of the article? Please submit your code.