36-350: Data Mining

Homework 2
Date: September 7, 2001 **Due: start of class September 14, 2001**

---

1. (Data-driven abstraction) A company wants to profile its customers by location. However, the set of possible locations is too large, so we want to abstract it. Consider the following hierarchy of Canadian provinces, together with sales count for each:

   - Western
     - British Columbia (68)
     - Prairies
       - Alberta (40)
       - Middle Prairies
         - Manitoba (8)
         - Saskatchewan (3)
   - Central
     - Ontario (212)
     - Quebec (97)
   - Maritime
     - Nova Scotia (21)
     - New Brunswick (15)
     - Newfoundland (13)

   In class we discussed two automatic algorithms for merging categories according to a conceptual hierarchy. Algorithm A repeatedly merges the categories which, taken together, have smallest count. Algorithm B merges the categories which, taken together, most reduce the quadratic imbalance function. That is, it merges the categories with smallest value of

   $$\Delta = (\sum_i n_i)^2 - \sum_i n_i^2 \qquad (1)$$

   where the n_i are the counts for the categories being merged. When only two categories are merged, $\Delta = 2n_1n_2$ where n_1 and n_2 are the counts for the categories being merged. Note that category merges must respect the hierarchy.

   (a) Use algorithm A to abstract the Canadian provinces into 7 balanced bins. Report which groups you merged.

   (b) Use algorithm B to do the same. Report which groups you merged.

2. (Classification and anomaly detection) In this problem, you will perform the image classification problem described in lecture (day 3 slide 18). Treat each image as a random independent sample of pixels from a pixel population. The two populations in this case are “flower” and “tiger”. There are two images labeled “flower” from which you can estimate the flower pixel population, and two images labeled “tiger” from which you can estimate the tiger pixel population.

   The set of possible colors has been reduced to 64. The file `flower.dat` contains a vector reporting the total count of each color across the labeled flower images. Similarly, the file `tiger.dat` has counts for the labeled tiger images. The file `test1.dat` contains a vector of color counts, in the same order, for a new image. Similarly for `test2.dat` and `test3.dat`. You can read these into S or R using `scan`.

   (a) For the image described by `test1.dat`, what is the log-likelihood for each class (“flower” and “tiger”)? Which is the most likely class of the image? As always, please submit your code. Hint: if you have more than 15 lines of code, you’re probably doing something wrong.

   (b) Do the same for `test2.dat` and `test3.dat`.

   (c) One of the test images doesn’t belong in either category. Which one? How can you tell?