36-350: Data Mining

Homework 4
Date: September 15, 2002 **Due: start of class September 22, 2002**

---

1. A dataset is currently divided into three clusters. The first cluster has 100 measurements and mean 0. The second cluster has 5 measurements and mean 2. The third cluster has 3 measurements and mean 5. We want to merge two clusters, while minimizing the sum of squares. Which two should we merge?

2. A web search engine keeps track of the popularity of each web site it indexes. Popularity is measured by the number of hits per day. The organizers want to improve the result of a search by displaying "popular" results separately from "unpopular" results, in two different columns. The problem is that the definition of "popular" versus "unpopular" has to be defined relative to the set of results, e.g. 100 hits per day might be "popular" when most of the other results have 10 hits per day, but would be considered "unpopular" when the other results have 1000 hits per day. They need a solution to this which is automatic—no human intervention. Describe how the organizers can do this using the methods in class.

3. Troubled by the fact that k-means gives a different answer each time it is run, Louis Reasoner suggests removing the randomness from k-means. Instead of starting with random prototypes, he suggests always starting with the first k data points as the prototypes. Explain why this is a bad idea.

4. Three applications of clustering were described in class: browsing images hierarchically, segmenting color images, and organizing video segments. In each case, k-means or Ward's method was used. These algorithms try to minimize the sum of squares, which means they prefer the clusters to be balanced in size. Is balance desirable in each of these applications? If not, why not?

5. A data miner runs k-means to divide a dataset into 3 clusters. The cluster means turn out to be evenly spaced across the range of the data. Can the miner conclude that the data consists of three separate subgroups? Explain.

6. In the computer lab, you will have computed a cluster hierarchy by Ward's method. How many clusters does the tree suggest are contained in the data? Why?