36-350: Data Mining

Homework 4
Date: September 21, 2001 **Due: start of class September 28, 2001**

---

1. For each abstraction scenario below, explain whether k-means or `bhist.merge` would be more appropriate.

   (a) An online music seller wants to recommend music to a customer based on their previous purchases. Some people like to buy music from a specific time period, like the 1970's or 1980's. The seller wants to use the time period of the music as a basis for the recommendation. This requires dividing the time axis into bins which reflect a particular customer's preferences.

   (b) 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. Given a set of pages relevant to a search, they need to abstract the popularity measure into "popular" vs. "unpopular".

   (c) In each customer profile, a grocery store keeps track of the average amount spent per visit. The store wants to determine what causes some customers to spend more or less per visit. They decide to abstract this number into "high", "medium", and "low" categories.

2. 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?

3. We want to abstract the data in `hw4.dat`.

   (a) Use `break.hclust` to divide the data into 3 clusters. Show how the data is divided.

   (b) Why is three an interesting number of clusters for this data?

   (c) Now use `break.kmeans` to divide the data into 3 clusters. Which method achieves the lower sum of squares? Why?

   (d) Find another interesting number of clusters and show the result.

4. A data miner runs k-means to break 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.

5. In class we showed how to use Ward's method to find change-points in a time series. A simpler idea for finding change-points is to look for a large difference in the level between consecutive years. Does this idea work on the Lake Huron data? The data is plotted below, along with the result of Ward's method. The individual measurements are shown as circles.