PubH 7475/8400 Homework 5 (Spring 2007)

Due by 4pm on May 10, 2007 in Biostat frontdesk or my mailbox

1. Apply 1) hierachical clustering; 2) K-means or K-medoids; 3) Normal mixture model-based clustering to one of the following two data sets: (10-10-20=40 pts)

   - NCI (NCI60) microarray data: there are $p = 6830$ predictors (i.e. genes).
   - Spam data: there are $p = 57$ variables (in the Data file) to distinguish two classes, spam (coded as 1) and email (coded as 0).

   For the NCI60 data, you may want to do some gene selection (**without** using the class label information). For the Spam data, you can consider only a small subset of the data (e.g. $n = 100$ to 1000).

   The data and some information on the data are available from the Data link on our course homepage.

   You need to assess 1) the number of clusters and 2) whether the clusters separate the classes. If you like, you can use training data to do clustering and then use test data or CV to evaluate the predictive performance of the clustering model.

   Please attach your computer program and relevant output.

2. `(EM for mixture model)` We have iid observations $x_1, ..., x_n$ from the distribution

$$f(x;\Theta) = \sum_{r=1}^{K} \pi_r \phi(x; \mu_r, \sigma_r^2),$$

   where Θ represents all unknown parameters, $0 \leq \pi_r \leq 1$ for and $1 \leq r \leq K$ and $\sum_{r=1}^{K} \pi_r = 1$, and $\phi(x; \mu_r, \sigma_r^2)$ is the density function for a Normal distribution $N(\mu_r, \sigma_r^2)$. Derive the EM algorithm to estimate Θ. (40 pts)

3. **(PubH 8400)** Suppose we observe $x_1, ..., x_n$ iid from $Bin(1, p)$.

   (a) Derive the MLE $\hat{p}$ for p. (5 pts)

   (b) Suppose that we have m iid observations $y_1, ..., y_m$ from $Bin(1, p)$ that are randomly missing. Derive the EM algorithm for estimating p, and show whether it is the same as or better than $\hat{p}$. (25 pts)

4. **(PubH 8400)** Choose two papers from the lists given under Week 13 and Week 14 on the course Updates page: summarize the main points of and comment on each paper. (20 pts)