**PubH 7475/8475/Stat 8931 Homework 4** (Spring 2016)

*Due on March 24, 2016*

1. Apply 1) K-means or K-medoids or kernel K-means; 2) Normal mixture model-based clustering; 3) spectral clustering to one of the following two data sets, and use a method to select the number of clusters (and possibly other parametrs too): (20-20-20=60 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, it is allowed to consider only a smaller subset of the data (e.g. $n = 200$ to 1000). But these are optional.

   You need to assess 1) the number of clusters (and other possible parameters) and 2) how well the clusters predict the classes.

   **Please attach your computer program and relevant output.**

2. (EM for mixture model) We have iid observations (1-dim) $x_1, ..., x_n$ from the distribution

$$f(x; \Theta) = \sum_{k=1}^{K} \pi_k \phi(x; \mu_k, \sigma_k^2),$$

   where $\Theta$ represents all unknown parameters, $0 \le \pi_k \le 1$ for and $1 \le k \le K$ and $\sum_{k=1}^{K} \pi_k = 1$, and $\phi(x; \mu_k, \sigma_k^2)$ is the density function for a Normal distribution $N(\mu_k, \sigma_k^2)$. Derive the EM algorithm to estimate $\Theta$. (20 pts)

3. **(Optional)** 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 in addition to $x_1, ..., x_n$, we have another $m$ 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 $\hat{p}$. (15 pts)

4. **(8000)** Choose two papers from the lists given under Weeks 7-8 on the course Updates page: summarize the main points of and comment on each paper. (20 pts)