# CSC 411, Fall 2006 — Assignment #3

*Due at **start** of lecture on November 20. Note that this assignment is to be done by each student individually. You may discuss it in general terms with other students, but the work you hand in should be your own. Handing in work that is not your own is a serious academic offense. Fabricating results, such handing in fake output that was not actually produced by your program, is also an academic offense.*

For this assignment, you will implement a Bayesian mixture model using a simple Monte Carlo method and apply it to some artificial two-dimensional data that is provided. The implementation will use sampling from the prior to obtain a sample of parameter values. You will see how this works when there are 15 training cases and when there are 50 training cases.

As discussed during lecture, one approach to unsupervised learning, and in particularly to clustering, is to model the data as coming from a mixture of simple distributions. For real-valued data, Gaussian distributions are often used as mixture components, and often the variables in a training case are considered to be independent given that the case comes from a particular mixture component (or in other words, the Gaussian component distributions have diagonal covariance matrices).

The parameters of such a model, with $K$ mixture components, are the mixing proportions, $\rho_1, \ldots, \rho_K$, a vectors of means for each component, $\mu_1, \ldots, \mu_K$, with $\mu_{k,j}$ being the mean of variable $j$ if the case it is part of comes from component $k$, and a vector of standard deviations for each component, $\sigma_1, \ldots, \sigma_K$, with $\sigma_{k,j}$ being the standard deviation of variable $j$ within component $k$. These parameters define the probability density for a case, $y = (y_1, \ldots, y_p)$, as follows:

$$P(y \mid \rho, \mu, \sigma) \;=\; \sum_{k=1}^{K} \rho_k \prod_{j=1}^{p} (2\pi\sigma_{k,j}^2)^{-1/2} \exp\Big( -(y_j - \mu_{k,j})^2 / 2\sigma_{k,j}^2 \Big)$$

For a Bayesian model of this sort, we need to specify prior distributions for all the parameters. In a real problem, these priors would need to be chosen to be appropriate given what we know of the source of the data. The data in this assignment is artificial, so you will just use a simple prior that doesn't say anything very specific, but which does limit the range of the parameters. In particular, you should assume that the prior for $\mu_{k,j}$ is uniform over the interval $(-1, +1)$, and that the prior for $\sigma_{k,j}$ is uniform over the interval $(0.1, 0.5)$. Assume that all the $\mu_{k,j}$ and $\sigma_{k,j}$ are independent of each other, and of $\rho$. However, $\rho_1, \ldots, \rho_K$ can't be independent of each other, since they have to sum to one. We'll assume that $\rho$ is uniform over the region satisfying this constraint (as well as the constraint that the $\rho_k$ be non-negative).

To obtain a sample of $N$ parameter vectors from the posterior distribution given a set of $n$ training cases, you should sample $M$ parameter vectors independently from the prior distribution, and for each compute the likelihood. You should then sample $N$ parameter vectors (with replacement) from this set of $M$ parameter vectors, with the probability of selecting a parameter vector being proportional to its likelihood. These $N$ parameter vectors will, approximately, be a sample from the posterior distribution. You can look at these $N$ parameter vectors to get an idea of what the posterior distribution is like.

A two-dimension data set with 50 training cases is provided on the web page. You should model this data using $K = 3$ mixture components. You should first try to sample from the posterior distribution in the way described above using just the first 15 cases, and then try to sample from the

posterior distribution using all 50 training cases, using $M = 300000$ and $N = 12$. (This may take ten or twenty minutes; reduce $M$ as necessary if it's taking longer than you can tolerate.)

For each sampling run (for 15 and for 50 training cases), you should display the $N$ sampled parameter vectors, preferably in some informative graphical way. Note that some of these parameter vectors may be identical, since you are supposed to sample from the $M$ parameter vectors drawn from the prior with replacement.

You should discuss how well the sampling works in each case, based on your examination of the $N$ sampled parameter vectors, and also on the "effective sample size" that was obtained. If the weights for each of the $M$ points sampled from the prior (obtained by normalizing their likelihoods) are $w_1, \ldots, w_M$, the effective sample size is $1/\sum_{h=1}^{M} w_h^2$. This is equal to the reciprocal of the probability that two of the $N$ points that you sample using these weights will be identical. When the effective sample size is very low (close to one), the sample will not have points from all (or even any) of the regions that have high posterior probability.

You should also discuss whether the results appear to be reasonable for this dataset, after viewing the training cases yourself, and what they may say about the degree of uncertainty in clustering a small number of training cases.

You should hand in a paper listing of your functions, written in R, Matlab, Octave, or some other suitable language (if you check with me first). Your program should be written in a readable style, and be reasonably efficient (though there is no need to be fanatical in this regard). You should also hand in your discussion, accompanied by printed or plotted results to support your conclusions.

**Note:** Some hints on technical details of doing some of the operations needed for this assignment will be put on the web page soon.