Homework #5 (due Nov. 21, 2005)

14th November 2005

1. Implement a four cluster, 2D Gaussian, expectation-maximization (EM) algorithm with isotropic variances. You do not need to solve for the prior probabilities which can all be set to $P(i) = \frac{1}{4}$. Generate more than 100 data points,uniformly sampled from the four clusters. Assuming that you are generating the data in a $[0,1]^2$ grid, set the means to
$\{[0.25, 0.25], [0.25, 0.75], [0.75, 0.25], [0.75, 0.75]\}$. For testing purposes, generate the data multiple times using different standard deviations ranging from 0.1 to 1. Assess the performance of the algorithm by reporting the difference between the true means and variances and the estimated means and variances respectively. Assume that all clusters have the same variance. Compare with the K-means clustering algorithm for different choices of initial conditions.
2. Show that

$$\min_{\{z_{ia}\}} \sum_{i=1}^{N} \sum_{a=1}^{K} z_{ia} ||x_i - \mu_a||^2 + \sum_{i=1}^{N} \sum_{a=1}^{K} z_{ia} \log\, z_{ia} = -\sum_{i=1}^{N} \log \sum_{a=1}^{K} \exp(-||x_i - \mu_a||^2)$$

when z_{ia} has the constraints $z_{ia} > 0$ and $\sum_{a=1}^{K} z_{ia} = 1$. Assuming a mixture model $p(x|\mu) = \sum_{a=1}^{K} \frac{1}{(2\pi)^{D/2}} \exp(-||x_i - \mu_a||^2/2)$, comment on the significance of the above result.