**HW5 (ECE8133 Spring 2009)**

1. The data set, hw5-gmm.txt, is generated by a Gaussian mixture model. Find a Matlab routine that perform the EM algorithm described in Lecture Note 18 to estimate the unknown parameters, mixture gains, means and variances. You can first plot the histogram of the data. Then you can try to fit 2, 3, 4 and 5 mixture components, and compare the overall likelihoods of the training data in each case. Remember EM is an iterative procedure that will often converge to a local minimum solution that maximize the overall likelihood locally, plot in each case the overall likelihood as a function of the number of iteration. Set a stopping criterion for your iteration process and state why you choose such a criterion. You are required in some cases to choose some initial guess of the parameter estimates in order to start the iteration. State your reason for your choice.

2. The data set, training + testing, contains training Reuters news stories in four fixed categories, as shown in the four folders, and a testing folder containing testing data to be used in Lab6. Download the entire zip file, and for Lab5 assume that you don't have the knowledge of the category membership, your assignment is to perform LSA on all documents to generate the term-document matrix. Then: (a) what are the values of $M$ and $N$? (a) using raw full vectors design a k-means clustering procedure with your favorite vector distance metric to cluster all documents into 2, 4, and 8 categories, and find out how the clustering results agree with the original document membership; (b) find out the rank of the matrix using SVD **(look in http://www.netlib.org/svdpack for software)**; (c) perform dimension reduction on choosing medium and low dimensions and repeat the clustering experiments in Part (a). Do you obtain similar results, why or why not?

**NB. Quiz 1 will cover all the materials we have covered so far with three problems: (1) dynamic programming; (2) MLE and optimization; (3) Entropy.**