Assignment 5
Due July 21, 2006 by 9:00 AM. Drop in box 89 at the general office (attention Reza Shahidi).

PROBLEM

The purpose of this assignment is to demonstrate your understanding of K-means clustering by implementing the algorithm in MATLAB. Your first task is to create a function or program to randomly generate several hundred samples for several classes. The features should follow a uncorrelated Gaussian distribution in two dimensions. Then you need to create two algorithms to separate the points into \( k \) classes: the basic K-means clustering algorithm using Euclidean distance, and a modification that minimizes the global representation error. Finally, you will need to plot the results, showing the actual points labeled according their assigned classes as well as the cluster means.

SUBMISSION

Include the following in your submission:

1. Function listings of your code files. These should be well commented and easy to follow. Make sure to include the following code:
   a. Generating the random distributions
   b. Basic K-means using Euclidean distance
   c. K-means using the global minimization extension. You may use the same basic program structure from part b if you wish.
   d. Generating the plots

2. Plots of four cases: 3 generated classes with results from basic K-means where \( k=3 \) and \( k=5 \); and 3 generated classes with results from extended K-means where \( k=3 \) and \( k=5 \). You should pick examples that demonstrate your algorithms working well. The plots should show the means of the actual and found clusters, and should show how the samples are clustered. Use at least 50 samples per class.

3. Include a brief discussion about how you could modify your code to deal with the cases where the wrong value of \( k \) was chosen. How would the program adjust \( k \)?