CAP 6638 **Pattern Recognition** **Spring 2015**

MW 05:00 PM – 06:15 PM in NES 013

Project #4

In this project we explore unsupervised learning. We develop or tune a clustering algorithm on Dataset A, and test it on the other datasets. Our objective is to assign characters with the same label to the same class, and characters with different labels to different classes.

.

Write a clustering algorithm (or use existing software) for our sample data set. Do not use any information from the previous assignments other than the values of the pattern vectors *without* the class information. Specifically, do not use information about the sequence of the patterns for the choice of initial cluster seeds, the relative frequencies of the classes, or the values of the scores from classification.

Cluster the 100 samples of Dataset A into 9, 10, and 11 classes using the algorithm of your choice.
You may use either the moment or the binary data. You are encouraged to find any available software instead of writing your own: most statistical or image processing packages have some clustering routines. You can try several experiments, methods or parameter settings on Dataset A.

Repeat the method, ***without alteration*****,** on the other three sets, but with only 10 classes.

When you have completed the clustering process, report your results in a table (analogous to the confusion tables in earlier assignments) showing for each cluster, the number of characters of each class. The tables for Dataset A will be 9 x 10, 10 x 10, and 11 x 10 rows by columns. The other datasets will each have a 10 x 10 table.

Assign a letter identity to each cluster according to which letter occurs with the *highest frequency* in that cluster. So you will have the largest numbers on the diagonal of the table. Decide ties by lexical order. Once you have assigned a letter identity to each cluster, consider characters in that cluster with a different identity as "errors" and report the number of errors for each run in a summary table.

In reporting the results you are, of course, entitled to use the letter labels. Use the alphabetic labels for the clusters whenever possible, assigned according to the majority class (and lexical order in case of ties). Note that it is possible to have two clusters for the same letter: if so, use a1, a2.

Please write a *short* (c. 1-2 page) report explaining what you did and what you observed. If you have any arbitrary parameters, explain how you set them (For instance, initialization, ordering of samples). Include in the report a summary table with the number of errors for each data set. Since I expect some diversity in the methods selected, I may distribute these reports to the class.

Reference:

A.K. Jain, R.C. Dubes: ALGORITHMS FOR CLUSTERING DATA, Prentice Hall, 1988.
S. Theodoridis' and K. Koutroumbas' PATTERN RECOGNITION (Academic 1999) has six chapters (over 200 pages) on clustering.