NUIM/CS401 **Fall 2008**

Assignment 2: Digital k-Means Fun!

We will make use of the MNIST handwritten digits dataset as cleaned by Yann LeCun and Corinna Cortes, `http://yann.lecun.com/exdb/mnist/`.

You may work in small teams. You are also welcome to share I/O code for reading the dataset and displaying images even between teams.

Each team should turn in one report describing the results, with the computer code as an Appendix. Please describe difficulties encountered and how they were overcome. Also, please show anything interesting you noticed about the data or algorithm, such as interesting outliers in the data set, better or worse ways of initializing the cluster centers, etc.

1. To make sure you can read them correctly, print 8 randomly chosen samples of each digit (eight 1's, eight 2's, etc) from the training set.
2. Use the k-means algorithm to cluster the training set into 10 clusters. Initialize the cluster means randomly. Show the cluster means as bitmaps, along with the numbers of digits of various sorts associated with each cluster as a table of counts. (In such a table each row is one cluster, each column is one kind of digit, and the intersections are counts: the number of digit images of the given identity assigned to the given cluster.)
3. Show the same table using the test data instead of the training data.
4. Repeat 2 & 3 with 40 clusters.
5. Calculate the digit classification accuracy given by the above k-means clustering, assuming each point is assigned to the nearest cluster and then given the class corresponding to the most prevalent digit in that cluster. Calculate this for both the training set and the test set, for both the 10-cluster and 40-cluster versions. (This amounts to some arithmetic on the table described above.)
6. How much do you trust the above recognition accuracy figures? Suggest measures that could be taken to improve their reliability.
7. Make some changes to the system in order to improve recognition accuracy on the test set; describe them; and test to see how well they actually work.

Due by email to `barak+cs401-1@cs.nuim.ie` before 23:59 Sun 26-Oct-2008.