David Bozarth
CES 514
Fall 2005

Homework 4, Problem 3

I built a Java program to perform k-means clustering for any choice of k, on an arbitrary data set of real numbers formatted as a comma-separated value text file. The first field of each row must contain a positive integer, which the program interprets as a "Point identifier".

There was little error-checking built in; for example, the behavior when trying to form more clusters than there are points, was not considered.

I exercised the program thoroughly for k = 2, 3, 4; one result with k = 16 was obtained. The table (next page) shows results organized on the left side by cluster id#. On the right side are mean values for clusters sorted by number of points. A glance at the data reveals strongly correlated clusters for all 3 values of k. Only the first of 9 data runs yielded an anomalous distribution of points among the clusters. The number of iterations required to achieve stability was directly related to k.

Intracluster distance is reported as mean and as root mean square. The overall mean-of-means, and mean-of-rms for the values of k are:

k = 2:	20259 avg,	23201 rms
k = 3:	17710 avg,	19197 rms
k = 4:	15528 avg,	16725 rms

For this range of k-values, a tradeoff appears to be processing time vs. clustering precision. For the data set under test (about 6400 rows by 32 columns) on my desktop PC, with k = 4, the program runs to completion in under 5 minutes. For k = 16, completion occurs in under 10 minutes.

k-means clustering for k = 2, 3, 4

					mean (small to large by #points)	
points 0	5726	5712	672		points	667.3333
avg 0	8338.247	8241.885	32220.52		avg	32243.73
rms 0	10674.75	10542.17	35813.08		rms	35815.47
points 1	658	672	5712		points	5716.667
avg 1	32290.15	32220.52	8241.885		avg	8274.006
rms 1	35820.25	35813.08	10542.17		rms	10586.36
iterations	5	14	14		iterations	11
points 0	906	5216	262		points	262
avg 0	17214.14	5689.737	30226.48		avg	30226.48
rms 0	18338.59	7045.779	32205.31		rms	32205.31
points 1	262	262	5216		points	906
avg 1	30226.48	30226.48	5689.737		avg	17214.14
rms 1	32205.31	32205.31	7045.779		rms	18338.59
points 2	5216	906	906		points	5216
avg 2	5689.737	17214.14	17214.14		avg	5689.737
rms 2	7045.779	18338.59	18338.59		rms	7045.779
iterations	18	19	19		iterations	18.66667
points 0	1005	511	1005		points	228
avg 0	11466.91	16912.72	11466.91		avg	29676.11
rms 0	12387.57	18058.69	12387.57		rms	31507.55
points 1	4640	1005	4640		points	511
avg 1	4055.851	11466.91	4055.851		avg	16912.72
rms 1	4944.805	12387.57	4944.805		rms	18058.69
points 2	511	4640	228		points	1005
avg 2	16912.72	4055.851	29676.11		avg	11466.91
rms 2	18058.69	4944.805	31507.55		rms	12387.57
points 3	228	228	511		points	4640
avg 3	29676.11	29676.11	16912.72		avg	4055.851
rms 3	31507.55	31507.55	18058.69		rms	4944.805
iterations	26	23	24		iterations	24.33333