Clustering Methods

Exercises 5/7, 10.4.2017

1. Implement one clustering algorithm using one of the tools: **C**, **Java**, **Matlab**, **R, Python, Excel**. Algorithms available are in the 2nd page. The same algorithm with same language can be implemented only by one student. The ones already existing are not available.

2. Upload your program to Sami. (samisi@cs.uef.fi) by 10.4. by 10.00 latest. In your email, have title "Clustering project work".

3. Make your program (a) easy-to-read, (b) modular, (c) useful for others.

4. Find and remove all bugs found from your implementation. Upload revised version by 31.5.

5. Validate that your method by calculating the following measures:

   (a) SSE = sum-of-squared error (=TSE)
   (b) nMSE = SSE/*nd*
   (c) CI = Centroid index (CI) value
   (d) Success = Repeated 100 times. How many times it finds solution with CI=0?
   (e) ARI = Adjusted Rand Index (only if you have software for it)
   (f) ε = (SSE$_x$-SSE$_{opt}$)/SSE$_{opt}$ (optional; x=your algorithm, opt=best known)

Example values for K-means: (scaled by 10^{p}!!!)

Dataset	Clustering quality			Objective function			**Best**
	Success	**CI**	**ARI**	**SSE**	**nMSE**	**ε**	
A1	1%	2.5	0.82	1.98	3.31	0.64	1.22
A2	0%	4.5	0.82	3.39	3.23	0.67	2.03
A3	0%	6.6	0.82	4.90	3.27	0.69	2.89
S1	3%	1.8	0.85	18.84	18.84	1.11	8.91
S2	11%	1.4	0.86	19.79	19.79	0.48	13.28
S3	12%	1.3	0.84	19.51	19.51	0.14	16.89
S4	26%	0.9	0.84	17.00	17.00	0.07	15.70
Unbalance	0%	3.9	0.64	2.10	1.61	8.81	0.21
Birch1	0%	6.6	0.85	10.95	5.47	0.18	
Birch2	0%	16.6	0.81	15.75	7.87	2.45	
Dim32	0%	3.6	0.76	16.5	504	68.34	
Average:	**5%**	4.5	0.81	---	---	7.60	

If implemented by C using modules package, all points **double**!

Algorithms

	C	Java	Matlab	Python	Excel
Random Swap					
Genetic Algorithm[1]					
Density peaks[2]					
Mean shift[3]					
Affinity propagation[4]					
Stochastic relaxation[5]					
(more to appear)					

[1] P. Fränti, "Genetic algorithm with deterministic crossover for vector quantization", *Pattern Recognition Letters*, 21 (1), 61-68, 2000

[2] A. Rodriquez and A. Laio, Clustering by fast search and find of density peaks, *Science*, 344 (6191), 1492-1496, 2014.

[3] Y. Cheng, "Mean shift, mode seeking, and clustering", *IEEE Trans. on Pattern analysis and Machine Intelligence*, 17 (8), 790-799, 1995.

[4] B.J. Frey, D. Dueck, Clustering by passing messages between data points, *Science*, 315 (2007), pp. 972–976.

[5] K. Zeger and A. Gersho, Stochastic Relaxation Algorithm for Improved Vector Quantiser Design. *Electronics Letters*, Vol. 25 (14), pp. 896-898, July 1989.