STAT 535 Homework 8
Out November 29, 2011
Due December 5, 2011

mmp@stat.washington.edu

Problem 1 – K-mean clustering with Power Initialization

a. Implement the Power Initialization algorithm as a generic function. Inputs: sample $\mathcal{D}$ of size n, consisting of real valued vectors in d dimensions (it is OK to take $d = 2$), number of clusters K, a constant $c \geq 1$.

Set the number of initial centers to $K' = cK \ln K$.

b. Implement the K-means algorithm proper. Inputs: sample $\mathcal{D}$ of size n, consisting of real valued vectors in d dimensions (it is OK to take $d = 2$), number of clusters K, a set of initial centers $\mu^0_{1:K}$, a maximum number of iterations T. The algorithm should run no more than T iterations, but it should stop earlier if convergence is reached.

c. Run the algorithm on the data set `hw8-cluster7-data1000.dat` with $K = 7$ clusters and $T = 100$ iterations. The data file contains $n = 1000$ 2 dimensional real vectors, one per line.

Use the c constant of your choice. Plot the data as points in the plane, and superimposed on them the trajectories of the K centers for the T iterations. Please make as clear a plot as possible.

d. Make also a second plot showing the data and the final positions of the centers. Recommended but optional: mark the data points by their cluster assignments (e.g color the points in different colors, or mark the separation lines between clusters; the latter is OK by hand as long as it's neat enough).

[**e. Optional – Extra credit**] Plot on a graph the cost $\mathcal{L}(\mu_{1:K}) = \sum_k \sum_{i \in C_k} ||x_i - \mu_k||^2$ versus the iteration $t = 1 : T$.

f. Did your algorithm converge? Do you think the clustering achieved is a good clustering of these data?

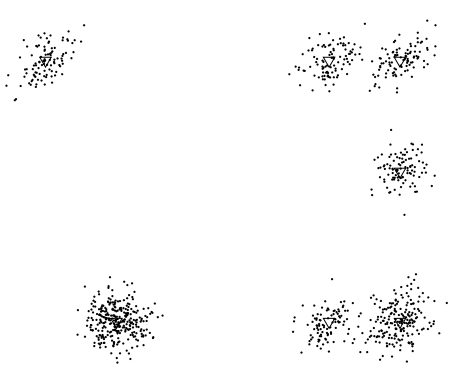
[**g. Optional – Extra credit**] Perform **c, d, e** again for the data set `hw8-cluster5-data1000.dat` with $d = 2$, $K = 4$, $T = 100$ (or $K = 3$) and compare the results on the two data sets?

The data set `hw8-cluster3-data100-debug.dat` with $K = 3$, $d = 2$, $n = 100$ is meant to help you test your code. The optimal cluster labels for this data set are in `hw8-cluster3-data100-debug-labels.dat`, given as the integers 1,2,3, one per line.

What you need to submit: the code through the web site;the answers and plots from c, d, [e], f, [g] on paper.

`hw8-cluster3-data100-debug.dat` $K = 3$

`hw8-cluster7-data1000.dat` $K = 7$



`hw8-cluster5-data1000.dat` $K = 3, 4$