SEMINAR 3
PATTERN RECOGNITION

1. Use the k-means algorithm and Euclidean distance to cluster the following 8 examples into 3 clusters: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).
The distance matrix based on the Euclidean distance is given below:

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

Suppose that the initial seeds (centers of each cluster) are A1, A4 and A7. Run the k-means algorithm for 1 epoch only. At the end of this epoch show:

a) The new clusters (i.e. the examples belonging to each cluster).
b) The centers of the new clusters.
c) Make a plot with the 8 points and show the clusters after the first epoch and the new centroids.
d) How many more iterations are needed to converge? Draw the result for each epoch.

2. Consider a (two-dimensional) dataset composed of two points.

1) Build a similarity matrix using a threshold function on Euclidean (norm-2) distance. The metric outputs 1 if the points are close enough according to a threshold and zero otherwise. Consider two cases: when the two datapoints are close or far.
2) Build the Laplacian in each case and discuss the eigenvalues and eigenvectors.

Now consider a dataset composed of four points with two pairs of points that are close to each other, one pair being far from the other.

More formally, assume that the similarity matrix looks as follows:

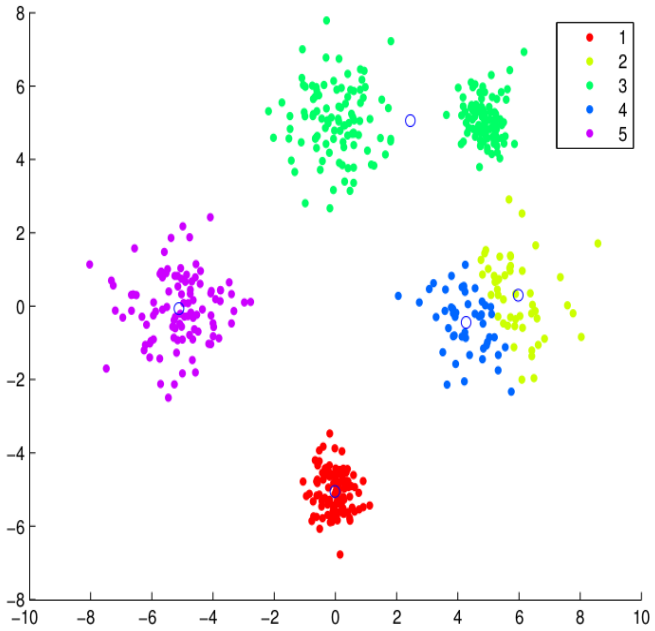
$$S = \begin{bmatrix} 1 & 0.8 & 0 & 0 \\ 0.8 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.5 \\ 0 & 0 & 0.5 & 1 \end{bmatrix}$$

1) What are the eigenvalues and eigenvectors of L = D – S? How many connected components do you obtain?

Hint: Look at the ratio of the eigenvalues

3. Load the file ex1P3.mat and do:
   a) Plot the 2D Points in a figure.

   b) Apply K-Means clustering over this data using K equal to 2,4,5,10 and 20. For each value, use the function gscatter to draw the samples with different colours depending on its corresponding assigned cluster. (See Fig. 1). Comment the results.



4. If you execute different times the function over the same data and the same K, you will observe that the final results differs between executions. This is because the random initialization of the centroids. Propose a way to solve this problem and that you think that is better than the random initialization. Defend why. You do not have to code it.

5. Load the image imgP3.jpg with the function imread and do the following:
   a) Use the function reshape in order to obtain a matrix $X \in R^{N \times 3}$ containing the values R,G,B of the N image pixels.
   b) Use k-Means to group the pixel colours into 10 clusters.
   c) Create a new image with the same number of pixels than the original. The R,G,B values of a given pixel won't be the original ones but the value corresponding to the centroid which have been assigned with k-Means. See the Fig. 2 as an example.
   d) Do the same with 2,5,10,50 and 100 clusters. Discuss the results.

Imatge Original

10 Clusters