60-473/574 Pattern Recognition - Fall 2017
Assignment 3

Deadline: **November 17, 2017, at 11:59pm**

This assignment must be done **individually**. The main goal of this assignment is that you obtain hands-on experience in using tools for clustering and apply them to real problems. For this assignment, use your favorite software package or programming language. Note that you **do not** have to implement the clustering algorithms. Implementations of these algorithms are available in Matlab, Octave, R, Python, C/C++, Java, Weka, Scikit and others – you can use any of these implementations.

You will work with the 4 datasets given in Resources (clusterincluster, halfkernel, twogaussians and twosprials).

1. For each of the 4 datasets do the following:
    a. Remove the class labels.
    b. Run *k*-means (with Euclidean distance) and EM, where $k = 2$ for both algorithms.
    c. Add the original labels to the existing clusters and plot the points as follows: points in the new clusters should have the same shape; points in the original classes should have the same color.
2. Pick an index of clustering validity discussed in class, and find the best value of *k* for each dataset, for both *k*-means and EM.
3. Comment on the results obtained for **all** datasets **and** clustering algorithms. How good is a particular clustering algorithm on a particular dataset? Why?

Submit:

1) A report (in PDF or Word) that includes all the times as required:
    a) In one paragraph, explain how you used the tools to perform the clustering.
    b) All plots resulting from the clusterings.
    c) Comments and comparison for each clustering scheme, and for each dataset.
2) One or two screenshots that show how you ran the algorithms in Weka, Scikit or the tool you used.

Note: Missing explanations about your implementations will imply marks deducted!

Upload *a single* Zip file that includes all the files to the Blackboard system no later than the date/time specified as the deadline. Late submissions will be penalized with 10% per day for up to 3 days – after the 3rd day, the mark will be **zero**.