ASSIGNMENT 1

Dimensionality reduction and classification.

1) Using "MixSampling.m" function, generate 300 observations by a bivariate Gaussian mixture with a structure of three groups.

2) Apply the Well Structured Partition (WSP) and Well Structured Perfect Partition (WSPP) models on data generated in the step 1) fixing the number of clusters equal to 3. Compare the membership matrices obtained by the two models with the "real" memberships values. Comment the results.

3) Generate 4 noise variables of 300 observations by a Gaussian distribution with $\mu = 0$ and $\sigma = 2$. Add these 4 variables to the data generated in the step 1). Now, you have a new 300×6 data matrix.

4) Apply the *k*-means model on this new data set fixing the number of clusters equal to 3. Compare the membership matrix obtained by *k*-means with the "real" memberships values. Comment the results.

5) "*Using a dimensionality reduction technique, such as principal components analysis (PCA), to create new variables and then using cluster analysis, such as k-means, to form groups using these new variables*". This technique is called as "*Tandem Analysis*". Use this approach on the new data set generated in the step 3) fixing the number of clusters equal to 3. Compare the membership matrix obtained by *k*-means with the "real" memberships values. Comment the results.

6) Apply Reduced *k*-means (REDKM) on data generated in the step 3) fixing the number of clusters equal to 3. Compare the membership matrix obtained by the REDKM with the "real" memberships values. Comment the results underling the simultaneous approach advantages with respect to the sequential approach.

7) Load MatLab workspace named "data02.mat". Repeat the steps 5) and 6) on the ECSI data set. Define the partition obtained by both approaches and comment the results.