CSCI 6600/7350 Homework 4 (Due midnight, Thursday April 25th)
Submit PDF file to D2L Dropbox "Homework 4"

1. Apply the following feature selection methods to pick the top 2 features for the iris data (https://archive.ics.uci.edu/ml/datasets/iris):
    a. Signal to noise ratio
    b. Two tailed t-test
    c. Relief, assuming the two randomly selected objects are:
        i. 4.3,3.0,1.1,0.1,Iris-setosa. (line 14)
        ii. 6.2,2.2,4.5,1.5,Iris-versicolor (line 69)

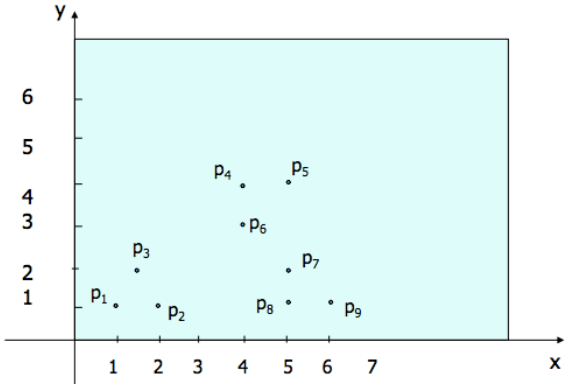
2. Apply the PCA and LDA dimensionality reduction methods on the following two data sets:
    a. biodegradation data (https://archive.ics.uci.edu/ml/datasets/QSAR+biodegradation#)
    b. glass data (https://archive.ics.uci.edu/ml/datasets/glass+identification)

    For each data set, split the data with 80% training and 20% testing, and use scikit learn programs (program templates can be downloaded from the course web site) to answer the following questions. Justify your answer with experimental results:

    - For PCA, what is the optimal number of principle components to use?
    - Do PCA and LDA improve the classification accuracy obtained from the original data?

3. Perform K-means clustering on the example data discussed in class.
    a. Let K=2, and the initial seed objects for the two classes be: class 1 (object 3) and class 2 (object 6). Show the clustering results.
    b. Compute the mean squared errors of the clustering result with K=2. Determine which clustering partition size, e.g., k=2 or K=3, is more suitable for this data by comparing the mean squared error results from K=2 clustering partition and K=3 clustering partition (use the results given in the class notes).

P1(1, 1)
P2(2,1)
P3(1.5,2)
P4(4, 4)
P5(5,4)
P6(4,3)
P7(5,2)
P8(5,1)
P9(6,1)



4. Given the distance(dissimilarity)table for 6 data objects as shown below. Assuming O2 and O4 are the medoids of the 2 clusters initially. After the objects are distributed to the two clusters, we randomly selected O1 to replace O4 as the new medoid for cluster 2, what's the total cost for this replacement? Should this replacement be carried out?

	O1	**O2**	**O3**	**O4**	**O5**	**O6**
O1	--					
O2	0.94	--				
O3	0.36	0.91	--			
O4	0.19	0.75	0.39	--		
O5	1.15	1.38	0.99	1.3	--	
O6	1.38	2.16	1.22	1.5	1.2	--

5. Perform UPGMA average link hierarchical clustering on 6 objects. The distance/dissimilarity values between pairwise objects are shown in the table below. Show the intermediate distance tables computed as well as the intermediate clustering hierarchies constructed. If we are to recommend K clusters for the data, what will be a reasonable distance threshold to use to obtain the clusters from the dendrogram?

	Obj1	Obj2	Obj3	Obj4	Obj5	Obj6
Obj1	0					
Obj2	0.5	0				
Obj3	4	3.5	0			
Obj4	1.5	2	2.5	0		
Obj5	4.5	1	3	5	0	
Obj6	6	2.5	6.5	4	3.5	0