CSE471: Statistical Methods in AI -- Spring 2016
Assignment 5: Dimensionality Reduction and Clustering

DUE: Before 12:00 midnight on 30 Mar 2016

INSTRUCTIONS:

1. You may do the assignment in Matlab/Octave, R, Python, C/C++ or Java.
2. You need to upload pdf files in the Course Portal. One file should contain your answers, results and analysis. A separate file should contain code you have written and its sample output.
3. At the top-right of the first page of your submission, include the assignment number, your name and roll number.
4. *IMPORTANT:* Make sure that the assignment that you submit is your own work. *Do not copy any part from any source* including your friends, seniors or the internet. Any breach of this rule could result in serious actions including an **F grade** in the course.
5. Your grade will depend on the correctness of answers and output. In addition, due consideration will be given to the clarity and details of your answers and the legibility and structure of your code.

Preamble:

The aim of this assignment is to experiment with *dimensionality reduction* and *clustering* techniques we learned in the class on real world problems.

A. Dimensionality Reduction

(A.1) Implement dimensionality reduction techniques such as Principal Component Analysis (PCA) and Fisher's Linear Discriminant Analysis (LDA) using your own code.

(A.2) Apply PCA and LDA on IRIS dataset. Show projection of the original data in PCA space (PC1 versus PC2; PC1 versus PC3 and PC2 versus PC3) and 1-dimensional LDA space. Please label each point with their class labels for better visualization. Now, comment on the similarities and differences of results obtained in these two approaches.

(A.3) Apply PCA to a high dimensional dataset such as UCI Arcene cancer classification data. Generate the Scree Plot. Comment on how many components to choose for explaining 85%, 90%, 95% and 99% of the variance of the data. Project the data into first two PCs as well as 1-dimensional LDA space (To avoid clutter plot only 10% of the data but equal number from each class. Use class labels on the data points to visualize better) and comment on the "representation" of the data in the lower dimensional subspace.

B. Clustering

Apply K-means clustering on the following two datasets.

(B.1) Write your own code for K-means clustering.

(B.2) IRIS dataset (omit the class labels and perform clustering).

(B.3) One dataset of your choice from the UCI Machine Learning Repository

(B.4) There are two kinds of Cluster validation measures – Internal Measures and External Measures. Describe any two cluster validation measures of each kind. Apply these measures on your datasets as well as compute the Confusion Matrix. Present an analysis and discussion of your results.

Questions to be Answered :

(Don't directly copy answers from online resources, but consult them to address these questions in your own words).

1. Compare and contrast K-Means, K-median and K-medoid approaches – when do you use each method, what are the differences in the objective function that is optimized in each case.
2. What is the difference between using Covariance matrix versus Correlation matrix for doing PCA on data? When do you recommend using one over the other? Demonstrate this on a small dataset to support your argument.
3. Show the process of kernelizing PCA. What kind of dimensionality reduction does kernel-PCA accomplish?