CSL407 Machine Learning
Homework 6

Due on 11/14/2014, 11.55pm

Instructions: Upload to your moodle account one zip file containing the following. Please do not submit hardcopy of your solutions. In case moodle is not accessible email the zip file to the instructor at ckn@iitrpr.ac.in. Late submission is not allowed without prior approval of the instructor. You are expected to follow the honor code of the course while doing this homework.

1. A neatly formatted PDF document with your answers for each of the questions in the homework. You can use latex, MS word or any other software to create the PDF.
2. Include a separate folder named as 'code' containing the scripts for the homework along with the necessary data files. Name the scripts using the problem number.
3. Include a README file explaining how to execute the scripts.
4. Name the ZIP file using the following convention **rollnumber_hwnumber.zip**

In this homework you will be experimenting with clustering and dimensionality reduction techniques

1. Suppose that we have four observations, for which we compute a dissimilarity matrix, given by

$$\begin{bmatrix} 0 & 0.3 & 0.4 & 0.7 \\ 0.3 & 0 & 0.5 & 0.8 \\ 0.4 & 0.5 & 0 & 0.45 \\ 0.7 & 0.8 & 0.45 & 0 \end{bmatrix}$$

   For instance, the dissimilarity between the first and second observations is 0.3, and the dissimilarity between the second and fourth observations is 0.8.

   (a) On the basis of this dissimilarity matrix, sketch the dendrogram that results from hierarchically clustering these four observations using complete linkage. Be sure to indicate on the plot the height at which each fusion occurs, as well as the observations corresponding to each leaf in the dendrogram.

   (b) Repeat (a), this time using single linkage clustering.

   (c) Suppose that we cut the dendrogram obtained in (a) such that two clusters result. Which observations are in each cluster?

   (d) Suppose that we cut the dendrogram obtained in (b) such that two clusters result. Which observations are in each cluster?

2. In this problem, you will perform K-means clustering manually, with K = 2, on a small example with n = 6 observations and p = 2 features. The observations are provided in Table 1.

   (a) Plot the observations.

   (b) Randomly assign a cluster label to each observation. Report the cluster labels for each observation.

Table 1: Observations for question 2

Obs.	$\mathbf{X}_1$	$\mathbf{X}_2$
1	1	4
2	1	3
3	0	4
4	6	2
5	5	1
6	4	0

(c) Compute the centroid for each cluster.

(d) Assign each observation to the centroid to which it is closest, in terms of Euclidean distance. Report the cluster labels for each observation.

(e) Repeat (c) and (d) until the answers obtained stop changing.

(f) In your plot from (a), color the observations according to the cluster labels obtained.

3. In this problem, you will generate simulated data, and then perform PCA and K-means clustering on the data.

(a) Generate a simulated data set with 20 observations in each of three classes (i.e. 60 observations total), and 50 features.

(b) Perform PCA on the 60 observations and plot the first two principal component score vectors. Use a different color to indicate the observations in each of the three classes. If the three classes appear separated in this plot, then continue on to part (c). If not, then return to part (a) and modify the simulation so that there is greater separation between the three classes. Do not continue to part (c) until the three classes show at least some separation in the first two principal component score vectors. Once you reach the situation where the three classes appear separated, save the data samples. This will help the TA's to reproduce the results of your code.

(c) Perform K-means clustering of the observations with K = 3. How well do the clusters that you obtained in K-means clustering compare to the true class labels?

(d) Perform K-means clustering with K = 2. Describe your results.

(e) Now perform K-means clustering with K = 4, and describe your results.

(f) Now perform K-means clustering with K = 3 on the first two principal component vectors, rather than on the raw data. That is, perform K-means clustering on the 60 2 matrix of which the first column is the first principal component score vector, and the second column is the second principal component score vector. Comment on the results.

(g) Perform K-means clustering with K = 3 on the data after scaling each feature to have standard deviation one. How do these results compare to those obtained in (b)? Explain.

4. In this question you will experiment with ISOMAP - nonlinear dimensionality reduction technique as discussed in class. The accompanying `hw6data.txt` contains 698 images, corresponding to different poses of a face. Each image is given as a 64x64 intensity map and is represented as a vector in $\mathbb{R}^{4096}$. Each image vector is stored as a row in the file.

(a) Choose a metric to measure the distance between images. Construct a proximity graph with vertices corresponding to the images, and connecting each image to the k nearest neighbors in the dataset, for a suitable k.

(b) Implement ISOMAP algorithm and apply it to this graph to obtain a two dimensional embedding. Present a plot of this embedding.

(c) Repeat (a) and (b) for some other distance metric. Report your observations on the embedding.