CSECE 8735 Fall 2015
Unsupervised Learning

Assignment 4

due Thursday 11/12/2015

Problem 1 Carry out spectral clustering on the dataset Circle.dat, with the parameters $\sigma^2 = 2$, and $\varepsilon = 1.5$. Use the median value of the vector y's components as the threshold for partitioning data into two clusters: assign x_i to C_1 if $y_i < median(y)$, otherwise assign x_i to C_2.

(a) Show the smallest 5 eigenvalues.
(b) Make a plot of the data samples with their assigned clusters shown by different colors or different symbols.
(c) Include your MATLAB code in the report.

Problem 2 Carry out latent sematic analysis on the following four commends:
(c.f. class notes of 10/27/15, 10/29/15, and the papers of LSA-LM and SemanticClassification)

d-1: what is the time
d-2: what is the day
d-3: what time is the meeting
d-4: cancel the meeting

(a) Build the matrix W (word by document), with the words arranges as
w-1: what
w-2: is
w-3: the
w-4: time
w-5: day
w-6: meeting
w-7: cancel
w-8: when

Please use log with base 2 for computing the normalized entropy ε_i, and set $\varepsilon_i = 0$ if the word w-*i* does not exist in the training documents (e.g., the word when) .

(b) Perform SVD decomposition on W (svd() in MATLAB) and report the values of your U, S, V matrices.
(c) Keep the eigenvectors corresponding to the largest two eigenvalues, and compute the scaled document vectors for the four documents. Report their values.
(d) For the following new test commend
d-5: when is the meeting
use the fold-in method to compute its scaled document vector.

(e) Compute the Euclidean distance between the test commend d-5 and the training commends d-1, d-2, d-3, d-4 by using their scaled document vector. Determine the 1st, 2nd, 3rd, and 4th nearest neighbors of d-5.
(f) Include your MATLAB code in the report.

Problem 3 A Euclidean distance matrix of five data samples is given below:

$$D = \begin{bmatrix} 0 & 1 & \sqrt{5} & \sqrt{5} & 2 \\ 1 & 0 & \sqrt{2} & 2 & \sqrt{5} \\ \sqrt{5} & \sqrt{2} & 0 & \sqrt{2} & \sqrt{5} \\ \sqrt{5} & 2 & \sqrt{2} & 0 & 1 \\ 2 & \sqrt{5} & \sqrt{5} & 1 & 0 \end{bmatrix}$$

(a) Use the classical multidimensional scaling algorithm (MDS) to estimate the 2-D coordinates of the five data samples (include in your report the eigenvalue matrix, the eigenvector matrix, the estimated data coordinates, and the MATLAB code).
(b) Verify that the estimated data coordinates are centered, i.e., $\frac{1}{5}\sum_{i=1}^{5} x_i = 0$, and the distance matrix D is completely recovered by the Euclidean distance between the estimated data samples.

Problem 4 Three different clusterings (a,b, and c) have been generated for 12 data samples, shown in the table below:

Data samples	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
Class labels	1	1	1	1	1	2	2	2	2	2	2	2
Cluster labels (a)	2	2	2	2	2	1	1	1	1	1	1	1
Cluster labels (b)	2	2	1	1	1	2	2	2	2	2	2	2
Cluster labels (c)	1	1	1	2	2	2	2	2	2	2	1	1

Evaluate the quality of the three clusterings by using the measure of normalized mutual information (NMI) as defined on page 5 of the random projection paper (and discussed in class on 10/15/2015). Specifically, do the following three parts:

(a) For each clustering, compute the probability distributions Pij, Pi, and Pj, where i represents the class labels, and j represents the cluster labels.
(b) For each clustering, compute the mutual information and normalized mutual information.
(c) Comment on your results.

This assignment is complete.