ISYE 6740 Midterm 1

Prof. Yao Xie

Released Sept. 15. Due Sept. 30, 11:55pm
Total point: 100

1 K-means (15 points)

Given $m = 5$ data points configuration in Figure 1. Assume $K = 2$ and use Euclidean distance. Assuming the initialization of centroid as shown, after one iteration of k-means algorithm, answer the following questions.

(a) Show the cluster assignment;

(b) Show the location of the new center;

(c) Will it terminate in one step?

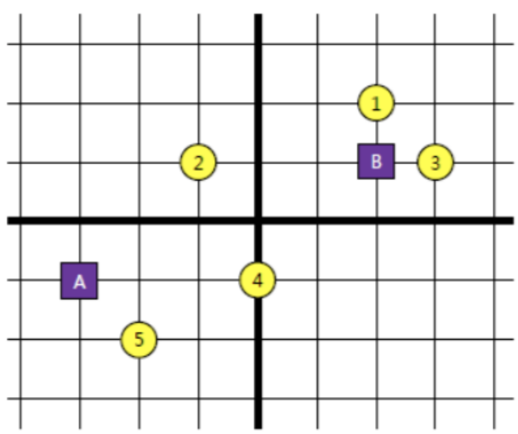


Figure 1: Question 1.

Now answer the above questions using Manhattan distance.

2 Spectral clustering (15 points)

Consider the data point setting in Figure 2. We will use spectral clustering to divide these points into two clusters. Our version of spectral clustering uses a neighborhood graph obtained by connecting each point to its two nearest neighbors (breaking ties randomly), and by weighting the resulting edges between points x_i and x_j by $W_{ij} = \exp(-\|x_i - x_j\|)$.

Indicate on Figure 2b the clusters that we will obtain from spectral clustering. Please provide an argument for your answer. Any reasonable answer will be given credits.

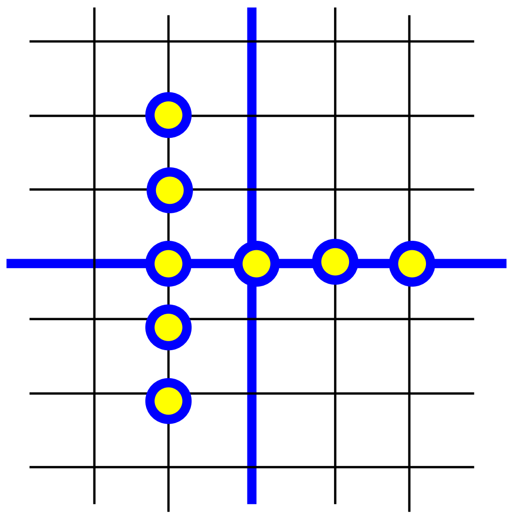


Figure 2: Question 2.

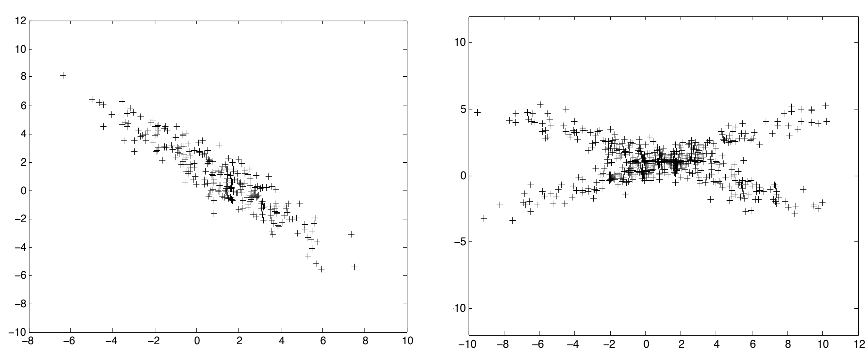
3 Principal Component Analysis (20 pts)

Suppose we have 4 points in 3-dimensional Euclidean space, namely $(4, -2, 4)$, $(5, -3, 5)$, $(2, 0, 2)$, and $(3, -1, 3)$.

(a) Find the first principal direction.

(b) When we reduce the dimensionality from 3 to 1 based on the principal direction you found in (a), what is the reconstruction error in terms of variance?

(c) You are given the following 2-D datasets, approximately draw the first and second principal directional on each plot.



(a) (b)

4 PCA for face recognition (25 points)

This question is a simplified illustration of using PCA for face recognition using a subset of data from the famous Yale Face dataset.

Remark: you have to perform downsampling of the image by a factor of 4 to turn them into a lower resolution image before we do anything.

1. First, given a set of images for each person, we generate the so-called eigenface using these images. The procedure to obtain eigenface is explained as follows. Given n images of the *same person* denoted by $x_1, \ldots, x_n$. Each image originally is a matrix. We vectorize each image to form the vector $x_i \in \mathbb{R}^p$. Now form a matrix

   $$X = [x_1, \ldots, x_n] \in \mathbb{R}^{p \times n}.$$

   The eigenfaces correspond to the largest k eigenvector of the data matrix $XX^\top$.

   Perform analysis on the Yale face dataset for subject 14 and subject 01, respectively, using all the images EXCEPT for the two images named `subject01-test.gif` and `subject14-test.gif`. Plot the top 6 eigenfaces for each subject. When visualizing the eigenvalues, you have to reshape the eigenvectors into images with the same dimension as the original images.

2. Now we will perform a face recognition task.

   For doing face recognition through PCA we proceed as follows. Given the test image `subject01-test.gif` and `subject14-test.gif`, we vectorize each image. Take the top eigenfaces of Subject 1 and Subject 14, respectively, project the 2 vectorized test images using the vectorized eigenfaces to obtain scores, respectively. Report four scores: (1) projecting test image of Subject 1 using eigenface of Subject 1; (2) projecting test image of Subject 1 using eigenface of Subject 14; (3) projecting test image of Subject 14 using eigenface of Subject 1; (4) projecting test image of Subject 14 using eigenface of Subject 14.

   Explain whether or not (and how) can you recognize the faces of the test images using these scores.

5 Order of faces using ISOMAP (25 points)

The objective of this question is to reproduce the ISOMAP algorithm results that we have seen discussed in class. The file isomap.mat (or isomap.dat) contains 698 images, corresponding to different poses of the same face. Each image is given as a 64×64 luminosity map, hence represented as a vector in $\mathbb{R}^{4096}$. This vector is stored as a row in the file. [This is one of the datasets used in J.B. Tenenbaum, V. de Silva, and J.C. Langford, Science 290 (2000) 2319-2323]

(a) Choose the Euclidean distance between images (i.e. in this case a distance in $\mathbb{R}^{4096}$). Construct a similarity graph with vertices corresponding to the images, and connecting each image to the k nearest neighbors in the dataset, for $k = 100$. (Notice that as a result, each vertex is in general connected to more than k neighbors.) Visualize the similarity graph (e.g., plot the adjacency matrix where weights are shown using intensity).

(b) Implement the ISOMAP algorithm and apply it to this graph to obtain a $d = 2$-dimensional embedding. Present a plot of this embedding. Find three points that are close to each other and show what they look like. Do you see any similarity among them?