ISYE 6740, Homework 3

Prof. Yao Xie

1. Order of faces using ISOMAP (50 points)

The objective of this question is to reproduce the ISOMAP algorithm results that we have seen discussed in lecture as an excercise. 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 the original paper for ISOMAP, J.B. Tenenbaum, V. de Silva, and J.C. Langford, Science 290 (2000) 2319-2323.]

(a) (20 points) 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 tune the threshold ϵ so that each node has at least 100 neighbors. Visualize the similarity graph (e.g., plot the adjacency matrix, or visualize the graph and illustrate a few images corresponds to nodes at different parts of the graph; you can be a bit creative here).

(b) (20 points) 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 in the embedding space, and show what they look like. Do you see any visual similarity among them?

(c) (10 points) Now choose ℓ_1 distance (or Manhattan distance) between images (recall the definition from "Clustering" lecture)). Repeat the steps above. Again construct a similarity graph with vertices corresponding to the images, and tune the threshold ϵ so that each node has at least 100 neighbors. Implement the ISOMAP algorithm and apply it to this graph to obtain a $d = 2$-dimensional embedding. Present a plot of this embedding. Do you see any difference by choosing a different similarity measure?

2. Density estimation: Psychological experiments. (50 points)

The data set n90pol.csv contains information on 90 university students who participated in a psychological experiment designed to look for relationships between the size of different regions of the brain and political views. The variables amygdala and acc indicate the volume of two particular brain regions known to be involved in emotions and decision-making, the amygdala and the anterior cingulate cortex; more exactly, these are residuals from the predicted volume, after adjusting for height, sex, and similar body-type variables. The variable orientation gives the students' locations on a five-point scale from 1 (very conservative) to 5 (very liberal).

(a) (20 points) Form 2-dimensional histogram for the pairs of variables (amygdala, acc). Decide on a suitable number of bins so you can see the shape of the distribution clearly.

(b) (20 points) Now implement kernel-density-estimation (KDE) to estimate the 2-dimensional with a two-dimensional density function of (amygdala, acc). Use a simple multi-dimensional Gaussian kernel, for

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2,$$

where x_1 and x_2 are the two dimensions respectively

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_1^2 + x_2^2)}{2}}.$$

Recall in this case, the kernel density estimator (KDE) for a density is given by

$$p(x) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{h} K\left(\frac{x^i - x}{h}\right),$$

where x^i are two-dimensional vectors, $h > 0$ is the kernel bandwidth. Set an appropriate h so you can see the shape of the distribution clearly. Plot of contour plot (like the ones in slides) for your estimated density.

(c) (10 points) Plot the condition distribution of the volume of the amygdala as a function of political orientation: $p(\text{amygdala}|\text{orientation} = a)$, $a = 1, \ldots, 5$. Do the same for the volume of the acc. Plot $p(\text{acc}|\text{orientation} = a)$, $a = 1, \ldots, 5$. You may either use histogram or KDE to achieve the goal.