ISYE 6740 Homework 3

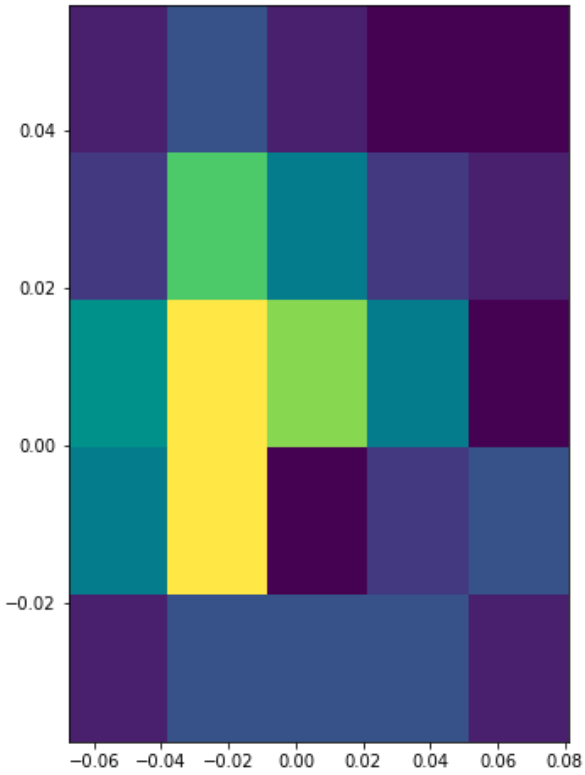
100 points total.

1. **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) 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.

Figure 1: Example of a 2 dimensional histogram with 5 bins. Amygdala is on the x-axis and acc is on the y-axis.



(b) 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

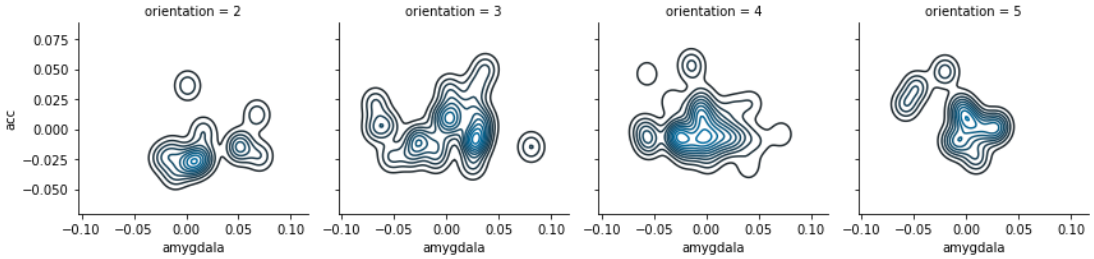
$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_1^2 + x_2^2)}{2}}.$$

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.

Figure 2: Example of a 2-dimensional KDE plot using a simple Gaussian kernel with bandwidth=.0085.



(c) 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.

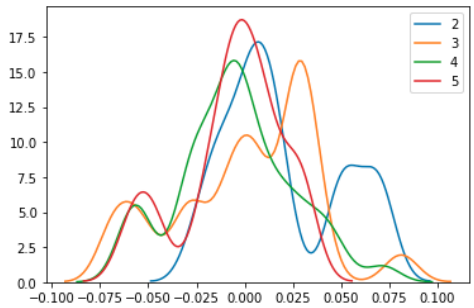


Figure 3: 1-dimensional KDE plot using a simple Gaussian kernel with bandwidth=.0085 for the Amygdalla sample.

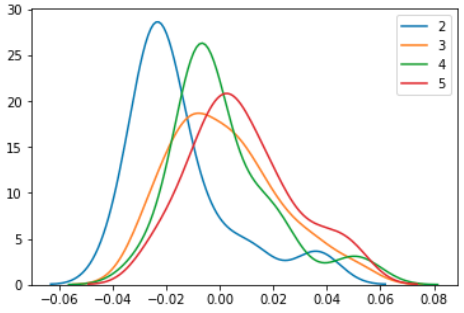


Figure 4: 1-dimensional KDE plot using a simple Gaussian kernel with bandwidth=.0085 for the Acc sample.

2. **Implementing EM algorithm for MNIST dataset. (50 points)**

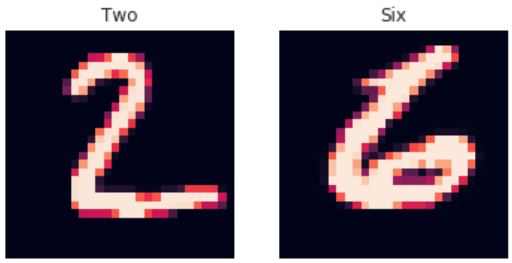
Implement the EM algorithm for fitting a Gaussian mixture model for the MNIST dataset. We reduce the dataset to be only two cases, of digits "2" and "6" only. Thus, you will fit GMM with $C = 2$. Use

the data file data.mat or data.dat on Canvas. True label of the data are also provided in label.mat and label.dat

The matrix images is of size 784-by-1990, i.e., there are totally 1990 images, and each column of the matrix corresponds to one image of size 28-by-28 pixels (the image is vectorized; the original image can be recovered, e.g., using MATLAB code, reshape(images(:,1),28, 28).

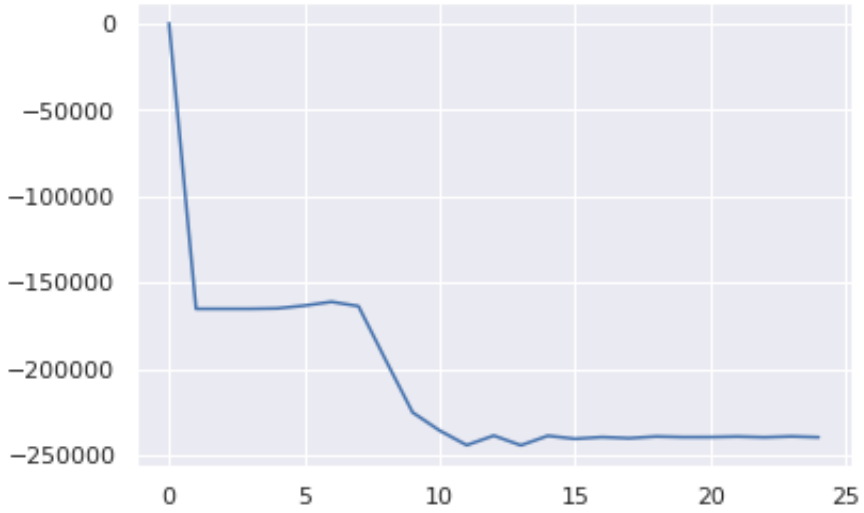
(a) Select from data one raw image of "2" and "6" and visualize them, respectively.

Figure 5: Example of a 2 and 6 hand written digit from the MNIST data set.



(b) Use random Gaussian vector with zero mean as initial means, and identity matrix as initial covariance matrix for the clusters. Please plot the log-likelihood function versus the number of iterations to show your algorithm is converging.

Figure 6: Log-likelihood per iteration of the EM algorithm where K=2 using a subset of the MNIST dataset



I am not sure what I have done to invert my log-likelihood. It is obviously decreasing instead of increasing. I have checked my calculations and the appear correct, but I have obviously missed something.

(c) Report the finally fitting GMM model when EM terminates: the weights for each component, the mean vectors (please reformat the vectors into 28-by-28 images and show these images in your submission). Ideally, you should be able to see these means corresponds to "average" images. No need to report the covariance matrices.

My ending π_k values were 7.76e-09 and 2.74e-08 respectively, average images are seen below.

It is interesting to compare the images to the plot of log-likelihood. There is little to no change between iteration 2 and 7, then just a slight improvement in the log-likelihood corresponds to finally being able to distinguish the separate numbers in iteration 8. Then there is not much visual difference between iteration 9 and the final iteration, although there was significant improvement in the log-likelihood.

Figure 7: Average of the 2 clusters for chosen iterations.