Pattern Recognition: Homework 7

Due date: 2023.4.11

Problem 1 (20 pt)

Suppose we have 8 one-dimensional samples as

$$-5.5, -4.1, -3.0, -2.6, 10.1, 11.9, 12.3, 13.6$$

Please use hierarchical clustering to cluster the samples. The distance between two classes D_i, D_j is defined as

$$d_{\min}(D_i, D_j) = \min_{x \in D_i, x' \in D_j} \|x - x'\|.$$

Draw the cluster tree and specify the value of two axes. How many classes do you tend to cluster them? Why? (It is a simple question, so just calculate it by hand and paper is enough)

Problem 2 (40 pt)

In this problem, you need to use EM algorithm to estimate the distribution in homework 5. Namely $\pi_1 = \frac{2}{3}, \pi_2 = \frac{1}{3}$ and

$$P(x, y|\omega = \omega_1) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}$$

$$P(x, y|\omega = \omega_2) = \frac{1}{2\pi} e^{-\frac{(x-2)^2+(y-2)^2}{2}}.$$

Set the cluster number to be exactly 2, how good is the estimation for these parameters $\pi_1, \pi_2, \mu_1 = (0, 0), \mu_2 = (2, 2), \Sigma_1 = \Sigma_2 = I$? Draw a diagram for the prediction error $\|\mu_1^* - \mu_1^{(t)}\|, \|\mu_2^* - \mu_2^{(t)}\|$ with respect to steps t and total number of samples N. It is also recommended (not required) to compare its density error with the best window method in homework 5. Which is better?

Problem 3 (40 pt)

We provide you with a tiny MNIST dataset, which consists of 1000 training digit images and 200 test images. You can use `data=numpy.load('toy_mnist.npz')` to load it, and then use `data['X_train']` or `data['X_test']` to load the image data, `data['X_train']` or `data['X_test']` to load the one-hot label. Please remember to convert them into float type array and divide by 255 to normalize it.

- Do the clustering on **training** data points using package `sklearn.KMeans`. Plot the diagram of J_e with respect to number of clusters. Where is the elbow point?

- Visualize the learned means of each cluster as 28×28 images. What do they look like? Try to make a predictor which predicts a sample's label to the nearest cluster's center. What is the prediction accuracy on MNIST test dataset?
- Try `sklearn.mixture.GaussianMixture` to use EM on MNIST. Visualize the learned means vector as 28×28 images. Compare its cluster centers' image and accuracy with KMeans.