EECE5644 Homework #4

Yuxin Lin

December 12, 2022

Question 1

We chose the following parameters to satisfy the requirements that the probability of error should fall within the range of 10% and 20% when classifying the generated data using the MAP method:

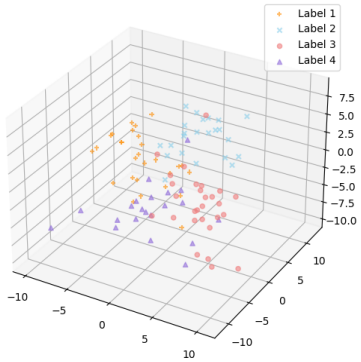
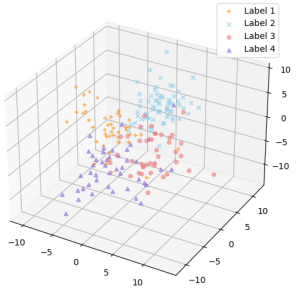
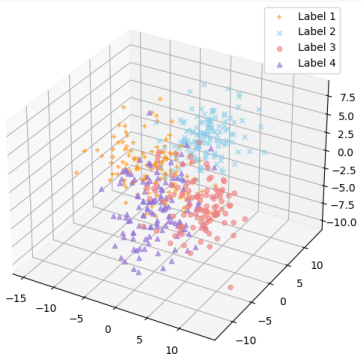
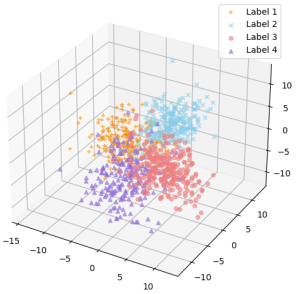
$$P_1 = \frac{1}{4} \quad Mean_1 = \begin{bmatrix} -5 \\ 3 \\ -2 \end{bmatrix} \quad Covariance_1 = \begin{bmatrix} -10 & 3 & -2 \\ 3 & 5 & 0 \\ -2 & 0 & 10 \end{bmatrix}$$

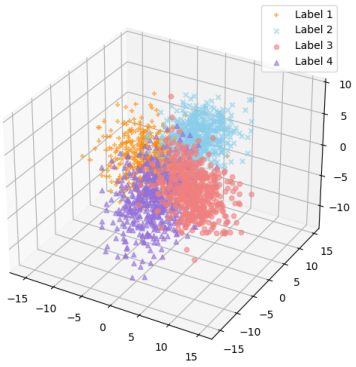
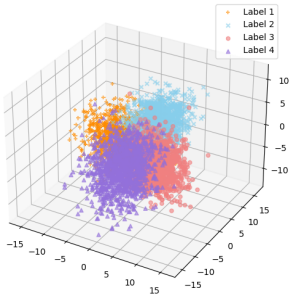
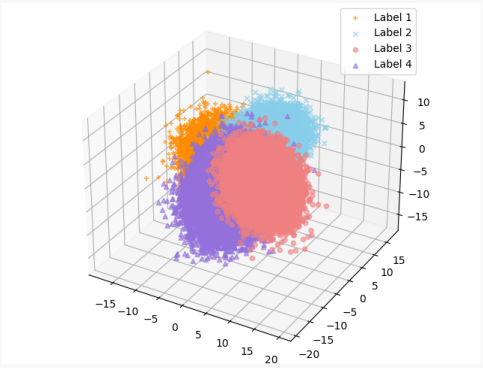
$$P_2 = \frac{1}{4} \quad Mean_2 = \begin{bmatrix} 2 \\ 5 \\ 3 \end{bmatrix} \quad Covariance_2 = \begin{bmatrix} 5 & 1 & 0 \\ 1 & 10 & 0 \\ 0 & 0 & 5 \end{bmatrix}$$

$$P_3 = \frac{1}{4} \quad Mean_3 = \begin{bmatrix} 5 \\ -2 \\ -2 \end{bmatrix} \quad Covariance_3 = \begin{bmatrix} 10 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 10 \end{bmatrix}$$

$$P_4 = \frac{1}{4} \quad Mean_4 = \begin{bmatrix} 0 \\ -5 \\ 3 \end{bmatrix} \quad Covariance_4 = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 10 & 5 \\ 0 & 5 & 10 \end{bmatrix}$$

The data distributions of the generated datasets with numbers of samples N=100, 200, 500, 1000, 2000, 5000 and 100000 are shown below:

N	Figure	N	Figure
N=100		N=200	
N=500		N=1000	

N	Figure	N	Figure
N=2000		N=5000	
N=100000			

Based on the minimum-probability-of-error classification rule, we estimated the optimal probability of error on the test dataset, which contains 100000 samples. Then we got:

$$P(error) = 0.128$$

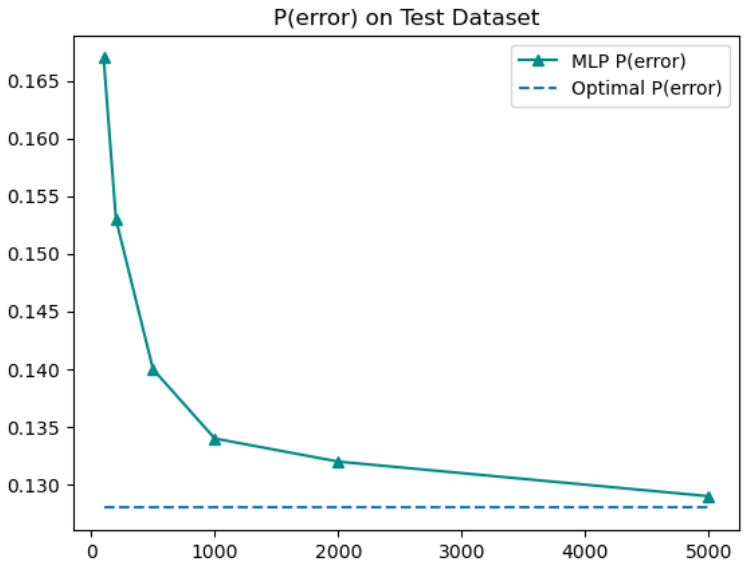
To implement the 2-layer MLP as the problem describes, we used the `sklearn.neural_network.MLPClassifier`, which contains a hidden layer with P perceptrons and a smooth-ramp style activation function 'tanh'. It takes cross-entropy loss to measure the performance of the model, and supports multi-classification by applying the softmax function as the output function.

For each training dataset, we performed 10-fold cross-validation to select the best number of perceptrons that could minimize P(error). The relationship between the number of perceptrons and the calculated P(error) is visualized as below:

According to the running results, for each training dataset, the ideal number of perceptrons should be:

Training Dataset	N=100	N=200	N=500	N=1000	N=2000	N=5000
Number of Perceptrons	10	19	21	7	9	30

With the best number of perceptrons, we trained several MLP classifiers on each training dataset and selected those who could provide the best classification accuracy on its training data. Then, we used these MLP classifiers to classify the test dataset. The estimated probabilities of error for each classifier are visualized as below:



where the optimal P(error) is 0.128, and:

Training Dataset	N=100	N=200	N=500	N=1000	N=2000	N=5000
P(error) on Test Dataset	0.167	0.153	0.140	0.134	0.132	0.129

We can see that the MLP classifier trained with a larger training dataset could achieve a higher classification accuracy as well as a lower P(error) on the test dataset.

Question 2

For data generation, we determined the parameters for the 4 components of our GMM as below:

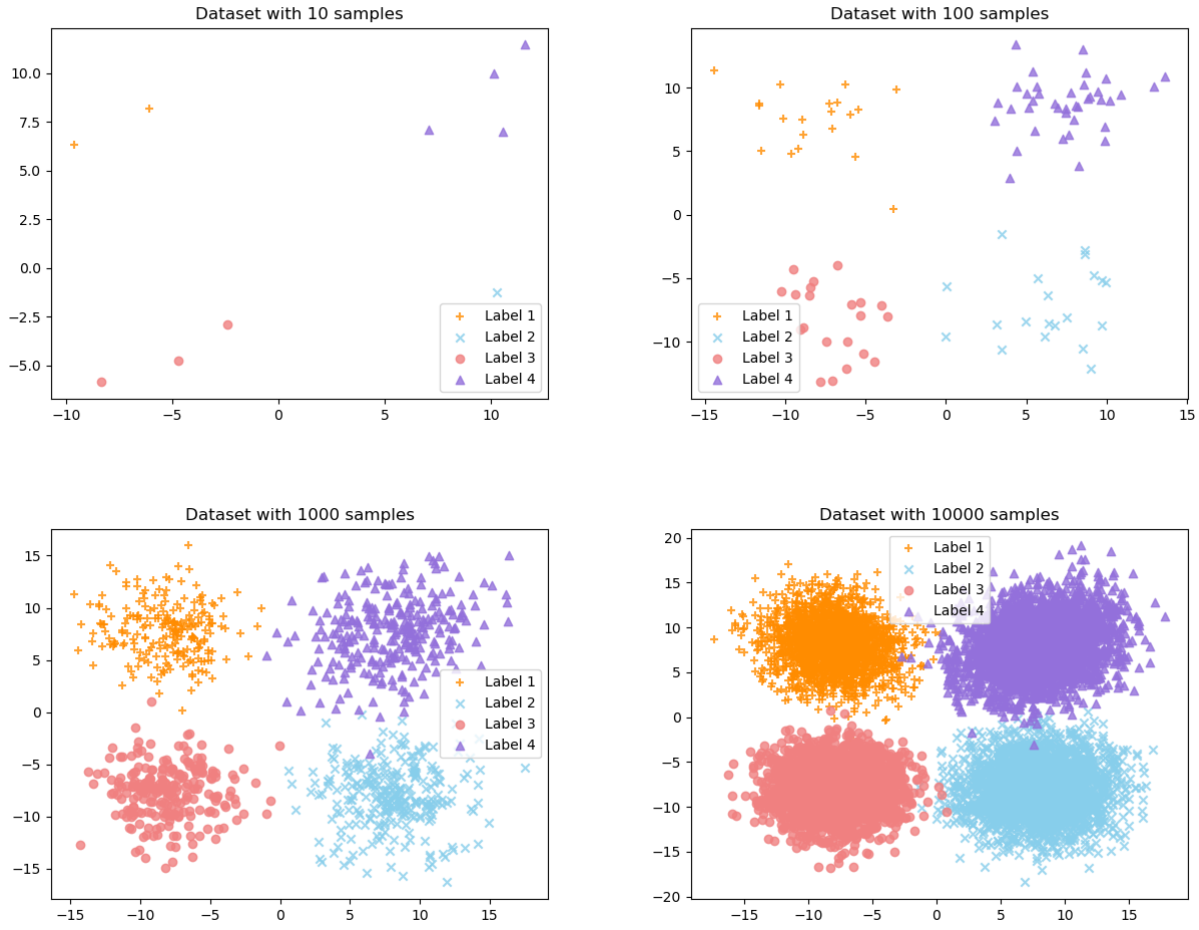
$$P_1 = 0.22 \quad Mean_1 = \begin{bmatrix} -8 \\ 8 \end{bmatrix} \quad Covariance_1 = \begin{bmatrix} 7 & -1 \\ -1 & 7 \end{bmatrix}$$

$$P_2 = 0.28 \quad Mean_2 = \begin{bmatrix} 8 \\ -8 \end{bmatrix} \quad Covariance_2 = \begin{bmatrix} 8 & 0 \\ 0 & 8 \end{bmatrix}$$

$$P_3 = 0.24 \quad Mean_3 = \begin{bmatrix} -8 \\ -8 \end{bmatrix} \quad Covariance_3 = \begin{bmatrix} 6 & 0 \\ 0 & 6 \end{bmatrix}$$

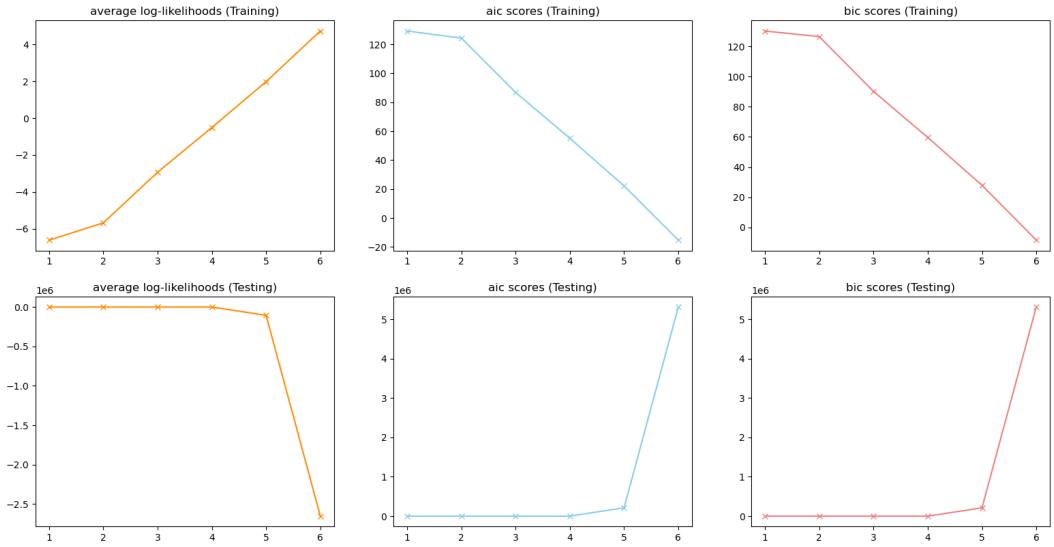
$$P_4 = 0.26 \quad Mean_4 = \begin{bmatrix} 8 \\ 8 \end{bmatrix} \quad Covariance_4 = \begin{bmatrix} 9 & 2 \\ 2 & 9 \end{bmatrix}$$

The data distributions of the generated datasets with numbers of samples N=10, 100, 1000, and 10000 are shown below:

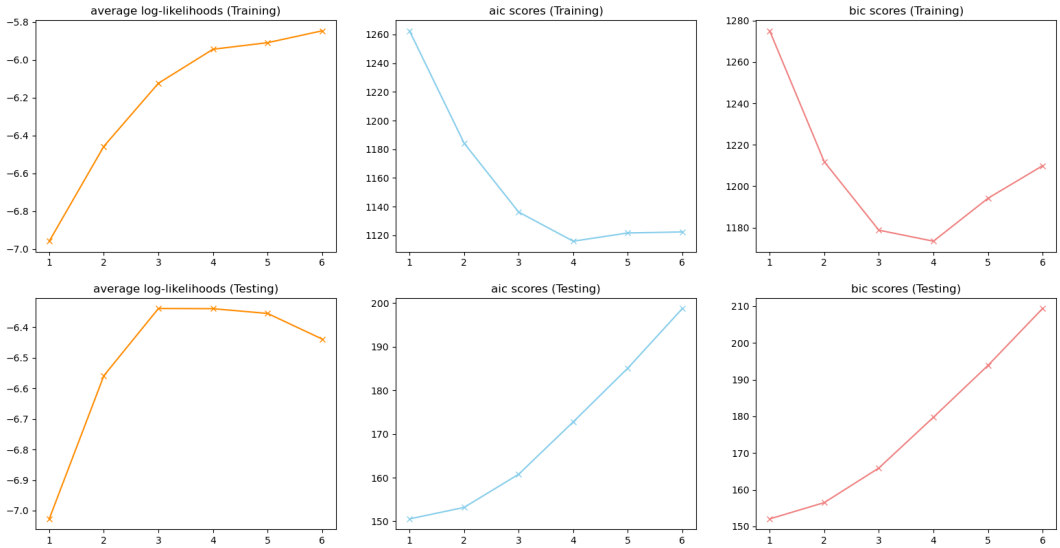


Within the framework of 10-fold cross-validation, we evaluated the generated datasets one by one with 1, 2, 3, 4, 5, 6 components by calculating their average log-likelihoods, aic scores and bic scores. For log-likelihoods, the higher the better. For aic and bic scores, the lower the better. Here we have:

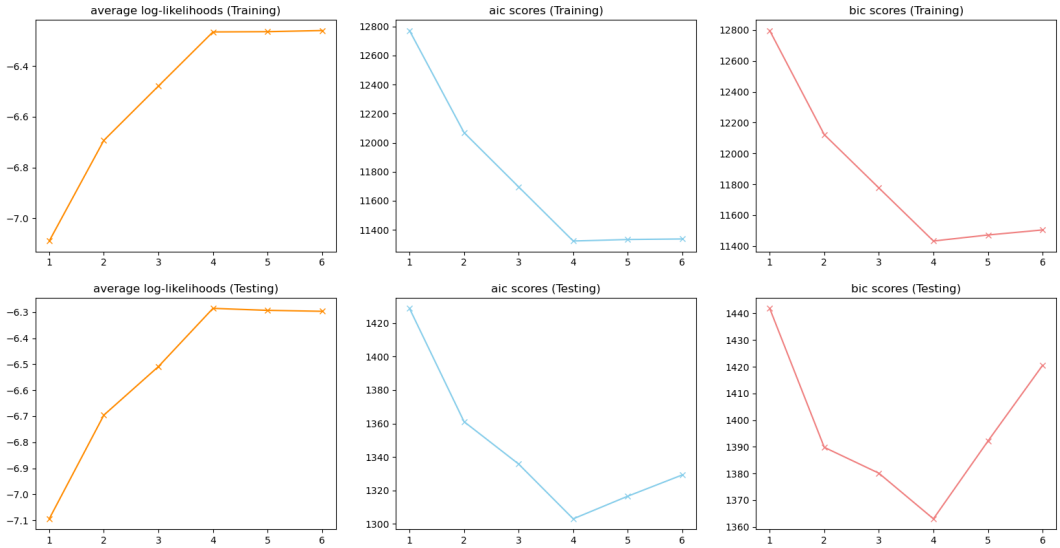
1. N = 10



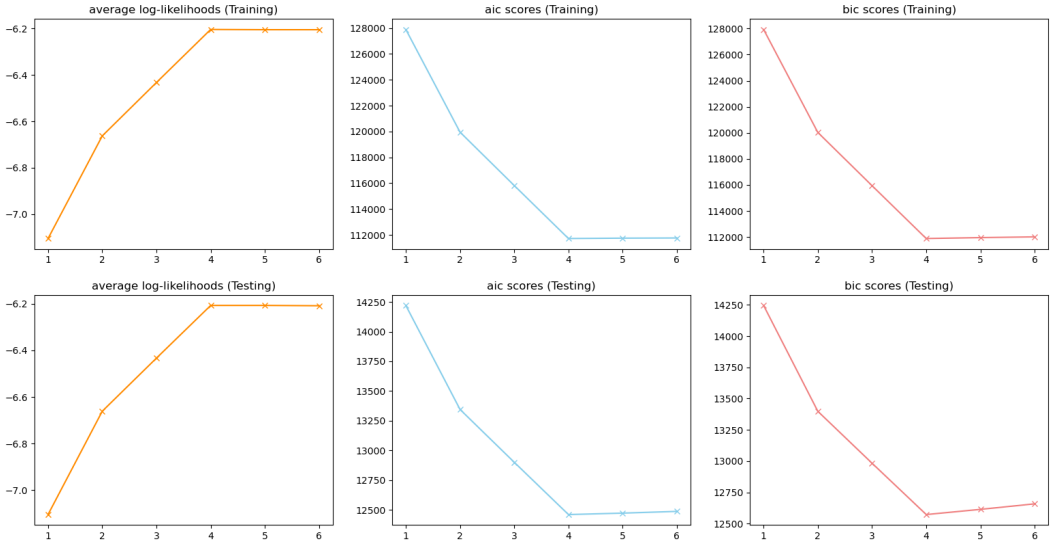
1. N = 100



1. N = 1000



1. N = 10000



The figures above suggest that larger datasets can help us estimate the number of components more accurately, and aic and bic scores are more precise than log-likelihoods in this problem.

Then, we repeated the experiment for 100 times. Each time we determined the number of components on each dataset by selecting the maximum log-likelihoods, minimum aic scores, and minimum bic scores. Calculating the selection rates of the 6 GMM orders on our datasets, we have:

1. N = 10

number of components	likelihood (train)	likelihood (test)	aic (train)	aic (test)	bic (train)	bic (test)
1	0	**1.0**	0	**1.0**	0	**1.0**
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	**1.0**	0	**1.0**	0	**1.0**	0

2. N = 100

number of components	likelihood (train)	likelihood (test)	aic (train)	aic (test)	bic (train)	bic (test)
1	0	0	0	**1.0**	0	**1.0**
2	0	0	0	0	0	0
3	0	0.05	0	0	0	0
4	0	**0.8**	**0.91**	0	**1.0**	0
5	0	0.13	0.05	0	0	0
6	**1.0**	0.02	0.04	0	0	0

3. N = 1000

number of components	likelihood (train)	likelihood (test)	aic (train)	aic (test)	bic (train)	bic (test)
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	**0.81**	**0.99**	**1.0**	**0.99**	**1.0**
5	0.02	0.16	0.01	0	0.01	0
6	**0.98**	0.03	0	0	0	0

4. N = 10000

number of components	likelihood (train)	likelihood (test)	aic (train)	aic (test)	bic (train)	bic (test)
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**
5	0	0	0	0	0	0
6	0	0	0	0	0	0

According to the results above, We find that as the number of samples within the dataset increased, the three methods' accuracy in identifying the number of components increased. When the sample size is small, aic and bic scores give more accurate decisions than log-likelihoods do. When N=10000, all three methods can provide the correct answer that the GMMs have 4 components.