EECS 349 Machine Learning

Project 6

Ding Xiang

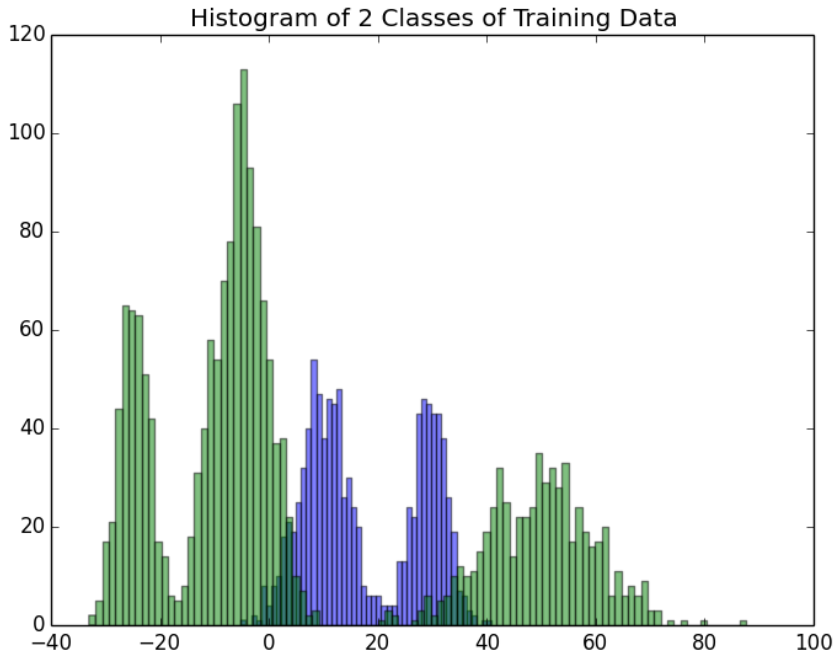
Nov. 5, 2017

Problem 1

Please see *gmm_est.py.*

Problem 2

To choose an appropriate number of Gaussian components K, we first plot the histogram of two class of data in the training set as below.



It's clear that for class 1 there are 2 peaks and for class 2 there are 3 peaks. So it's good to choose K=2 for class 1 and

K=3 for class 2. Also since for normal distribution about 99.7% of data are within 3 standard deviation of the mean i.e. $\mu \pm 3\sigma$. So by observation we can roughly assign initial values as the table below.

Table 1. Number of Gaussian Components and Initializations of Parameters

Parameters	**Class 1**	**Class 2**
K	2	3
μ	(10, 30)	(-25, -5, 50)
σ^2	(25, 11)	(11, 25, 64)
w	(0.6, 0.4)	(0.3, 0.4, 0.3)

For the first 20 iterations, the plot of data log-likelihood values are shown below.

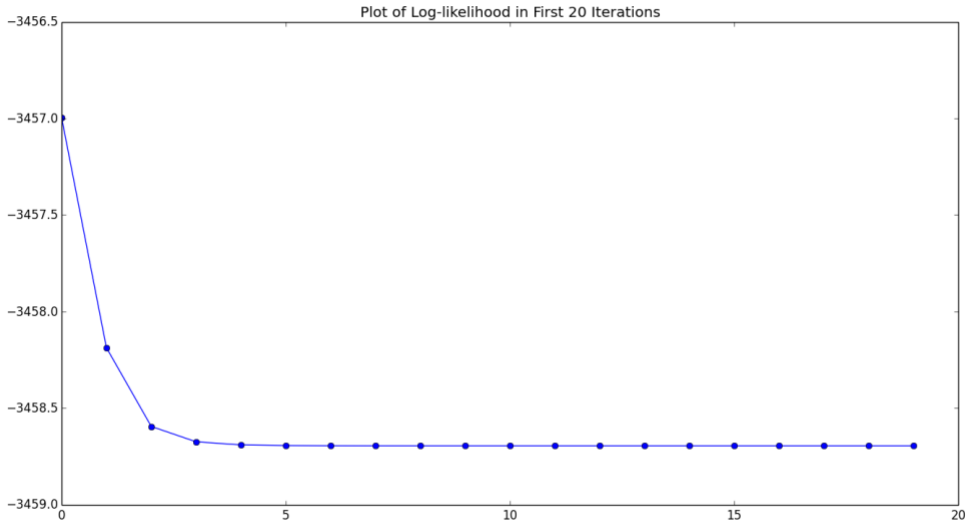


Figure 1. Plot of Log-likelihood in First 20 Iterations for Class 1

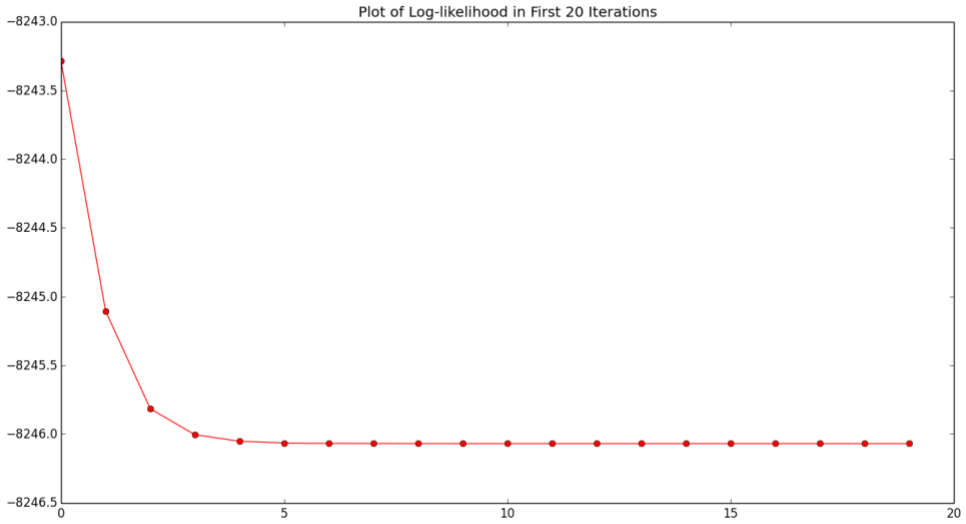


Figure 2. Plot of Log-likelihood in First 20 Iterations for Class 2

From the above picture we can see that the iteration has converged.

The final values of the GMM parameters for class 1

and class 2 are shown in the following picture and given in the table 2.

```
Class 1
mu = [9.7748859236208396, 29.582587182945414]
sigma^2 = [21.922804563227473, 9.7837696129445995]
w = [0.59765463038641087, 0.40234536961368628]

Class 2
mu = [-24.822751728696254, -5.0601582832343865, 49.624444719527624]
sigma^2 = [7.9473354077562393, 23.322661814350976, 100.02433750441195]
w = [0.20364945852723454, 0.49884302379593665, 0.29750751767685862]

Process finished with exit code 0
```

Table 2. Final Values of the GMM Parameters

Parameters	**Class 1**	**Class 2**
μ	(9.77, 29.58)	(-24.82, -5.06, 49.62)
σ^2	(21.92, 9.78)	(7.95, 23.32, 100.02)
w	(0.60, 0.40)	(0.20, 0.50, 0.30)

Problem 3

Please see "*gmm_classify.py*".

Problem 4

The prior probability of class 1 from training data is 1000/3000 = 0.333. The accuracy rate of the classifier is 0.935 (= 93.5%).

The plots of histograms of true classes and predicted classes for testing data are shown below.

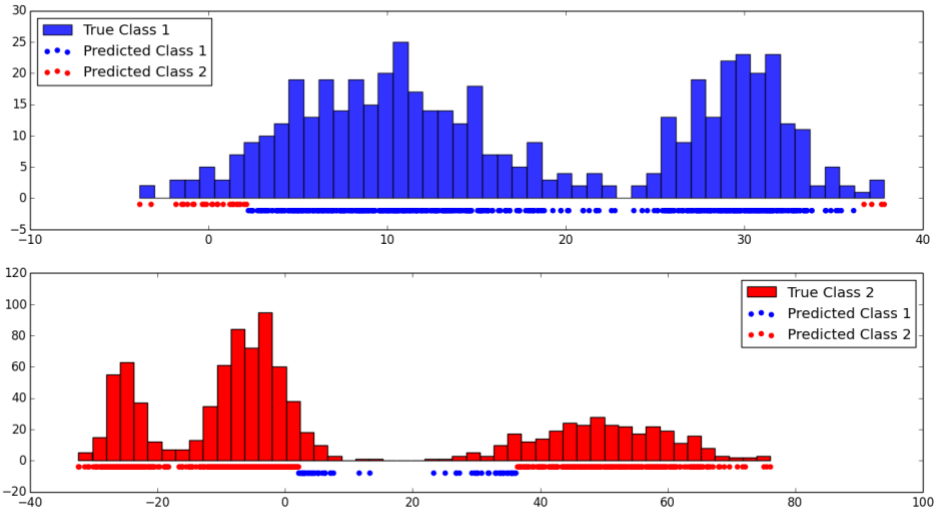


Figure 3. Histogram and Predicted Classes for Testing Data

If we put all of plots in one histogram, then it looks like below.

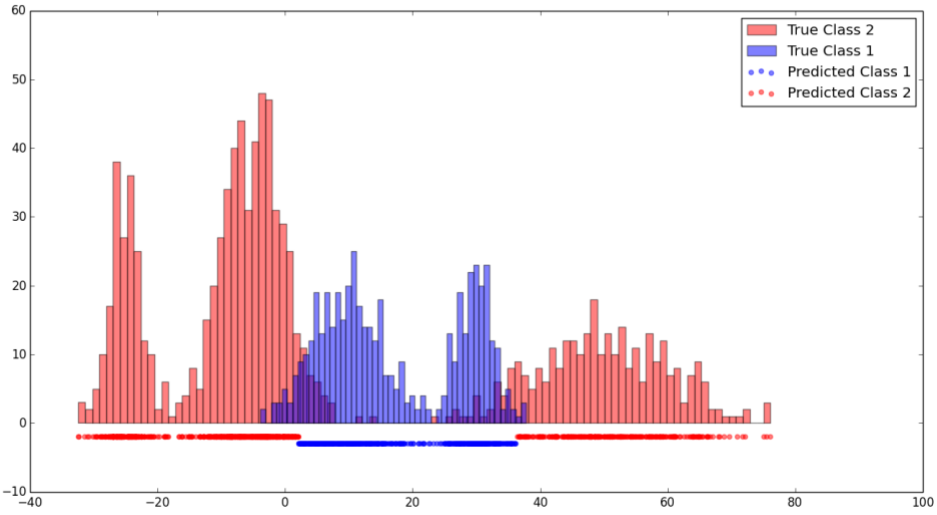


Figure 4. Histogram and Predicted Classes for Testing Data in One Plot

Problem 5

Yes, we can find a closed form.

Since we know exactly one Gaussian and which Gaussian is the one generates the data. We could collect all data that is generated by a same Gaussian $k \in \{1,2,\dots,K\}$. Let's say the data generated by Gaussian k is called D_k and the optimal parameters $\{\mu_k, \sigma_k, w_k\}$ can be calculated as below.

$$\mu_k = \frac{1}{|D_k|}\sum_{x_i \in D_k} x_i$$

$$\sigma_k = \frac{1}{|D_k|}\sum_{x_i \in D_k} (x_i - \mu_k)^2$$

$$w_k = \frac{|D_k|}{N}$$

where, x_i is each data point and there are totally N data points.

Problem 6

Yes, we cannot find a closed form if K is larger than 1.

Because we don't know which Gaussian generates which point, we cannot calculate mean or variance for each Gaussian k. Then we need to use EM algorithm to find a local optimal solutions for the GMM parameters.