**CS405 Machine Learning**

**Pre Lab #6 EM**

**Pre-Lab (25 points)**:

In this pre-lab, firstly you should use the K-means method to implement the image (Input image) segmentation



Exercise 1:

1. Implement the image segmentation based K-means method;
2. You should use the different color to represent these classifications, as shown in the following.



Hint:

* Transform the color space “RGB” to “Lab”;

The following parameters are as used to generate the original data (ground truth). Please the give the plots of GMM(Gaussian Mixture Model) of these two sets of data respectively. Then you are asked to merge these two sets of data as the test data, denoted by *A*. Assumed there are only two classifications, please estimate the distributions of the data *A* based EM.

N=[50;50]; %两类样本分别包含点数

mu=[2；10]; %对应样本均值

std= [1；1];

You should give the original distribution in a figure. As the following figure shows:



Then you should use the EM method to estimate the optimal parameters of GMM, including mean value *mu* and variance value *std*. Firstly, you can use a small number of samples (like 7 samples) to initialize the parameter *mu* and *std*. The EM method is based on the initialization parameters and the iteration number is 20. After that you should give the estimation results in a table and the figures.

|  |  |  |
| --- | --- | --- |
|  | Ground truth | Estimation (final) |
| mu | [2;10] | xxx |
| std | [1;1] | xxx |

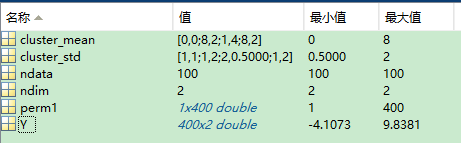
Estimation results:

Final distributions with the optimal estimation parameters:



**Lab (75 points):**

In this lab, a dataset named “data\_c3\_2.mat” is obtained by Matlab. You can use the python or Matlab to finish this lab. If you use the python to finish this lab, you can use this package “scipy.io” to load the data. The following figure shows the content of this dataset.



|  |  |
| --- | --- |
| Cluster\_mean | 每类样本在对应维度上的均值 |
| Cluster\_std | 每类样本对应的协方差矩阵 |
| ndata | 样本总数 |
| ndim | 每个样本维度 |
| Perm1 | 随机样本顺序，用于初始化参数 |
| Y | 样本集 |

Exercise 2:

1. Implement data classification with EM method;
2. Plot the figure of classification results，as shown in the following figure. Note: the results shown below are just one example, and you may have different results due to different initialization parameters.



The x-axis in the following figure denotes the candidate clustering number, and the y-axis denotes the the successive values of the cost function.

