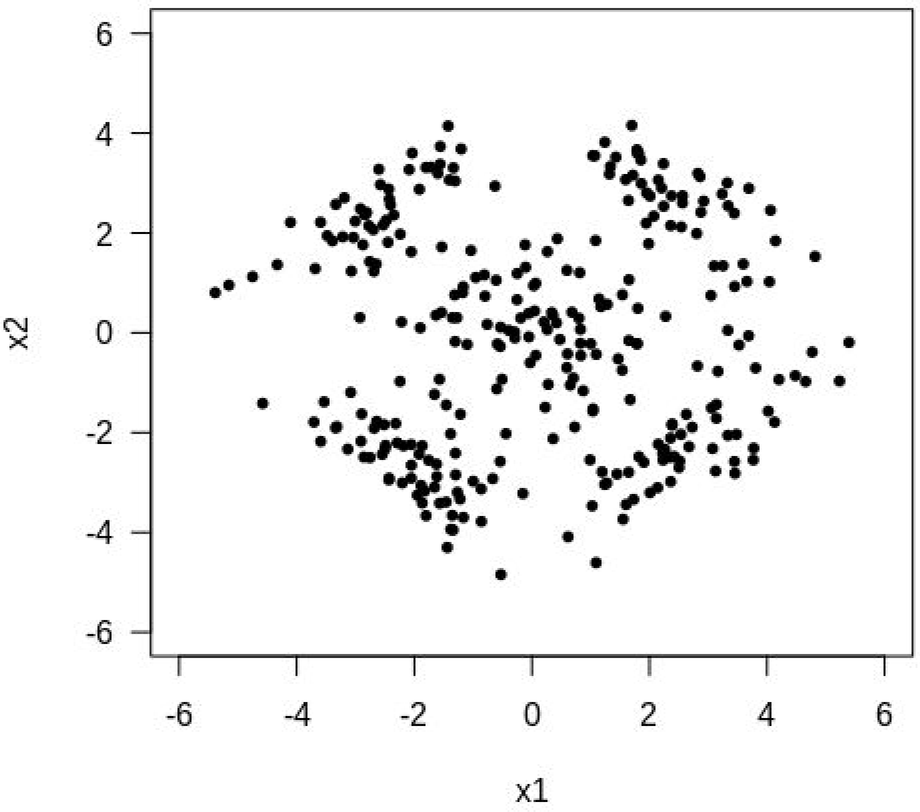
Aykut Aykut December 14th, 2018
0034039
Comp421 Homework06

In this homework, I was asked to implement an expectation-maximization (EM) clustering algorithm. In order to do that , I followed these 10 steps:

1) I generated random data points from the given 5 bivariate Gaussian density parameters and plotted the data.



2) I set *k* parameter to 5 and then randomly picked *k* centroids from *data.*

3) I set *k_means_max_iteration* parameter to 2 as it is stated in the homework description.

4) I ran the k-means clustering algorithm for *k_means_max_iteration* times. In each iteration, I first applied E-step and then M-step. At E-step, I calculated b_{ik} values for each data point and cluster using the following formula:

$$b_{ik} = 1 \; if \; \|x_i - \widehat{\mu}_k\| = min_c \|x_i - \widehat{\mu}_c\| \; ; \; 0 \; otherwise$$

At M-step, I updated centroids using the following formula:

$$\widehat{\mu}_k = \frac{\sum_{i=1}^{N} b_{ik} * x_i}{\sum_{i=1}^{N} b_{ik}} where \; N \; is \; the \; sample \; size$$

5) I took last centroids of k-means algorithm as the initial mean vectors of EM algorithm. I estimated prior probability and covariance matrix of each cluster using the following formulas:

$$\widehat{p}_k = \frac{\sum_{i=1}^{N} y_{ik}}{N} \; where \; y_{ik} = 1 \; if \; x_i \; is \; in \; cluster \; k; \; 0 \; otherwise \; and \; N \; is \; the \; sample \; mean$$

$$\overset{\wedge}{\underset{k}{\sum}} = \frac{\sum_{i=1}^{N} y_{ik} * (x_i - \widehat{\mu}_k) * (x_i - \widehat{\mu}_k)^T}{\sum_{i=1}^{N} y_{ik}} \; where \; y_{ik} = 1 \; if \; x_i \; is \; in \; cluster \; k; \; 0 \; otherwise \; and \; N \; is \; the \; sample \; mean$$

6) I set *EM_max_iteration* parameter to 100 as it is stated in the homework description.

7) I ran EM algorithm for *EM_max_iteration* times. In each iteration, I first applied E-step and then M-step. At E-step, I calculated h_{ik} values for each data point and cluster using the following formula:

$$h_{ik} = \frac{p(x_i | C_k, \phi^t) * \widehat{p}_k}{\sum_{c=1}^{K} p(x_i | C_c, \phi^t) * \widehat{p}_c} \; where \; K \; is \; the \; number \; of \; clusters$$

At M-step, I updated mean vectors, prior probabilities and covariance matrices using the following formulas:

$$\widehat{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^{N} h_{ik} * x_i}{\sum_{i=1}^{N} h_{ik}} where \; N \; is \; the \; sample \; size$$

$$\widehat{p}_k^{(t+1)} = \frac{\sum_{i=1}^{N} h_{ik}}{N} \; where \; N \; is \; the \; sample \; mean$$

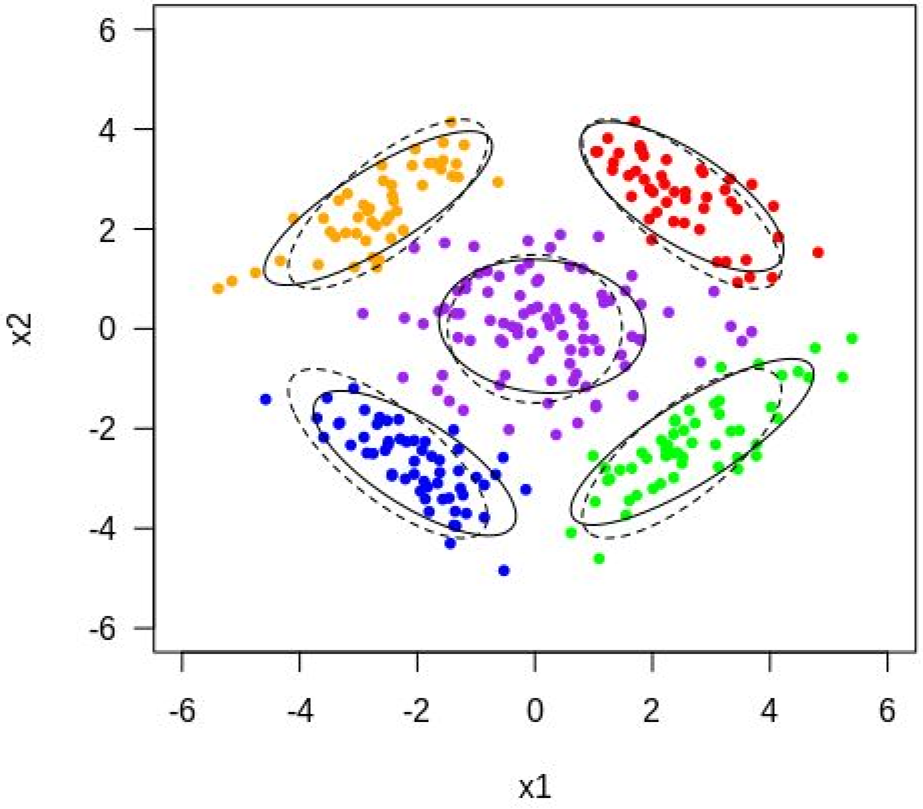
$$\overset{\wedge}{\underset{k}{\sum}}^{(t+1)} = \frac{\sum_{i=1}^{N} h_{ik} * (x_i - \widehat{\mu}_k^{(t+1)}) * (x_i - \widehat{\mu}_k^{(t+1)})^T}{\sum_{i=1}^{N} h_{ik}} \; where \; N \; is \; the \; sample \; mean$$

Finally, I assigned each data point to a cluster whose probability is the highest for that data point.

8) I printed mean vectors of EM algorithm after *EM_max_iteration* iterations.

```
> print(t(mean_vectors))
            x1          x2
[1,] -2.6659954  2.41685609
[2,]  2.5005444  2.64800061
[3,]  2.6772251 -2.26589452
[4,] -2.0426009 -2.69489423
[5,]  0.1213246  0.05096945
```

9) I drew the clustering result obtained by my EM algorithm by coloring each cluster with a different color. I also drew the original Gaussian densities which are used to generate data points and the Gaussian densities my EM algorithm found where densities are equal to 0.05 with dashed and solid lines, respectively. The figure is as follows:



10) I obtained the same/very close results, data point figure, mean vectors of EM algorithm and clustering results with the results given in the homework description.