

ENGR 421 / DASC 521: Introduction to Machine Learning

Homework 07: Expectation-Maximization Clustering

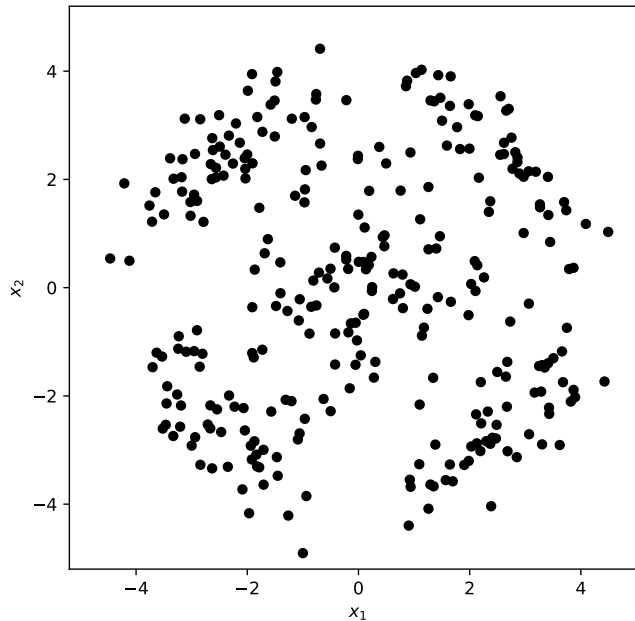
Deadline: December 27, 2021, 11:59 PM

In this homework, you will implement an expectation-maximization (EM) clustering algorithm in Python. Here are the steps you need to follow:

1. You are given a two-dimensional data set in the file named `hw07_data_set.csv`, which contains 300 data points generated randomly from five bivariate Gaussian densities with the following parameters.

$$\begin{aligned}\mu_1 &= \begin{bmatrix} +2.5 \\ +2.5 \end{bmatrix}, & \Sigma_1 &= \begin{bmatrix} +0.8 & -0.6 \\ -0.6 & +0.8 \end{bmatrix}, & N_1 &= 50 \\ \mu_2 &= \begin{bmatrix} -2.5 \\ +2.5 \end{bmatrix}, & \Sigma_2 &= \begin{bmatrix} +0.8 & +0.6 \\ +0.6 & +0.8 \end{bmatrix}, & N_2 &= 50 \\ \mu_3 &= \begin{bmatrix} -2.5 \\ -2.5 \end{bmatrix}, & \Sigma_3 &= \begin{bmatrix} +0.8 & -0.6 \\ -0.6 & +0.8 \end{bmatrix}, & N_3 &= 50 \\ \mu_4 &= \begin{bmatrix} +2.5 \\ -2.5 \end{bmatrix}, & \Sigma_4 &= \begin{bmatrix} +0.8 & +0.6 \\ +0.6 & +0.8 \end{bmatrix}, & N_4 &= 50 \\ \mu_5 &= \begin{bmatrix} +0.0 \\ +0.0 \end{bmatrix}, & \Sigma_5 &= \begin{bmatrix} +1.6 & +0.0 \\ +0.0 & +1.6 \end{bmatrix}, & N_5 &= 100\end{aligned}$$

The given data points are shown in the following figure.

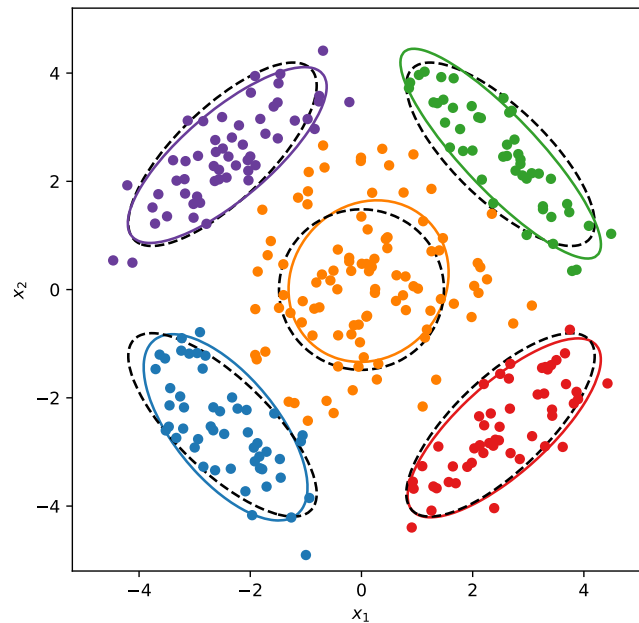


2. To initialize your EM algorithm, you should take the centroids given in the file named `hw07_initial_centroids.csv` as the initial values for the mean vectors. By assigning the data points to the nearest center, estimate the initial covariance matrices and prior probabilities in your EM algorithm.

3. After the initialization step, run your EM algorithm for 100 iterations. Report the mean vectors your EM algorithm finds. Your results should be similar to the following matrix.

```
[ [-2.44390052 -2.5453942 ]  
 [ 2.50354332  2.51134859]  
 [ 2.56726404 -2.55477253]  
 [ 0.12794677  0.15595776]  
 [-2.41465305  2.4855615  ]]
```

4. Draw the clustering result obtained by your EM algorithm by coloring each cluster with a different color. You should also draw the original Gaussian densities you use to generate data points and the Gaussian densities your EM algorithm finds with dashed and solid lines, respectively. Draw these Gaussian densities where their values are equal to 0.05. Your figure should be similar to the following figure.



What to submit: You need to submit your source code in a single file (.py file) and a short report explaining your approach (.doc, .docx, or .pdf file).

How to submit: Submit the two files (source code and short report) you created to Blackboard. Submissions that do not follow these guidelines will not be graded.

Late submission policy: Late submissions will not be graded.

Cheating policy: Very similar submissions will not be graded.