ENGR 421 / Homework 5: Expectation Maximization Clustering

Umur Berkay Karakaş

May 24, 2021

In homework 5, we are given 300 2-dimensional data points and 5 initial centroids. We are asked to implement EM algorithm for 100 iterations to find the final centroids and memberships.

I manually wrote the mean and covariance matrix values that are used to generate our data points to initial_means and initial_covariances arrays. I found memberships for each data point by assigning each data point to the closest centroid using spa.distance_matrix method. Then, I created a one-hat-encoding matrix for the membership values.

For the initial covariance matrices, I used np.cov for each cluster. For the priors, I simply divided the count of each cluster to number of total data points.

For E-step, I created gaussians array to store gaussian densities of the data points in each cluster and for the H (success) matrix, I used the corresponding formula from the lecture 22.

For M-step, I used the corresponding formula to update centroids, covariances and priors. Final centroids were:

```
In [30]: centroids
Out[30]:
array([[-2.04419197, -2.69776844],
       [ 2.48874351,  2.67687075],
       [ 2.6622246 , -2.3091108 ],
       [ 0.15535175,  0.05773829],
       [-2.67591954,  2.44658904]])
```

Figure 1: Final centroids

After 100 iterations are done, since we have probability of being in cluster k for each data point i in the success matrix, I used argmax function with success matrix as an input to find the memberships. After coloring each cluster with a different color and plotting the densities with initial and finals means, I had the following final plot:

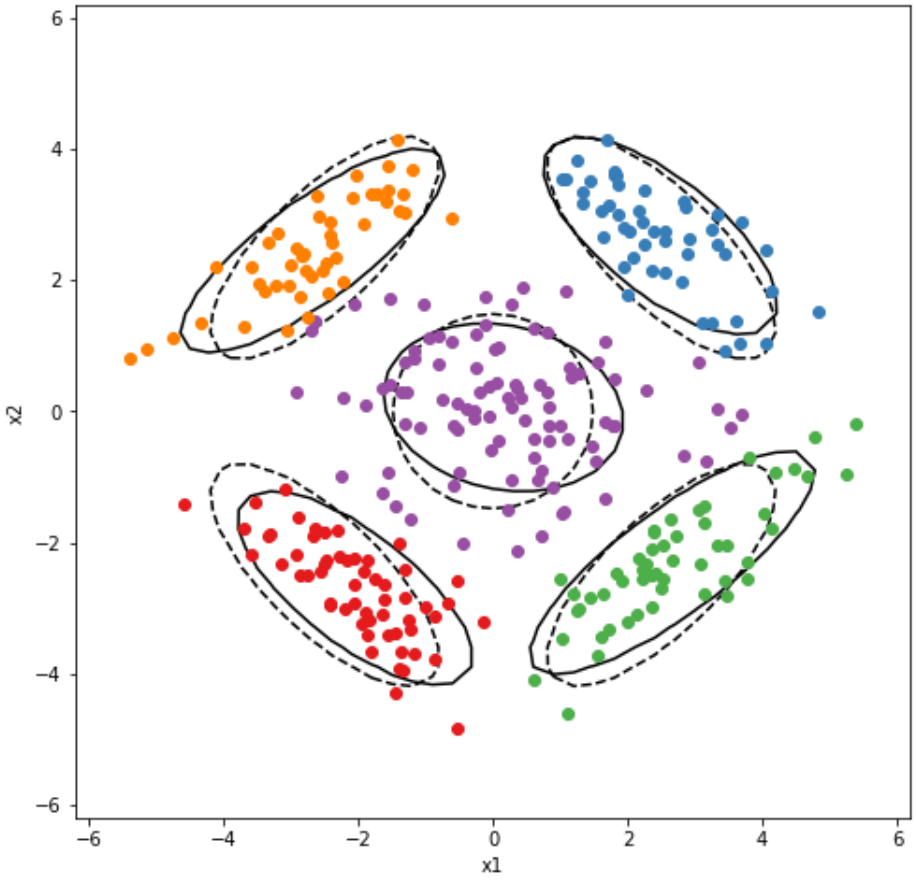

Figure 2: Final plot