CMPE544-HW2

Sadullah Gültekin

2020 - Fall

1 Implementation Details

In this project expectation maximization(EM) algorithm is implemented. For the sake of readability and usability, the algorithm is implemented in a class structure. To use the algorithm, one only needs to create an EM object with a number of Gaussian distributions and call the fit function using the data. After the convergence, all the resulting parameters will be stored in the class variables.

Fit function takes the data and runs the EM algorithm n times using the given threshold value in the stopping condition. The algorithm is run n times, because, like other clustering algorithms, in some cases the algorithm doesn't converges to the most optimal solution. This situation depends on the initialization. To avoid sub-optimal solutions, the algorithm is run n times (10 is the default value), and the parameters of the run with the highest likelihood value is selected as the final parameters.

In the main loop of the algorithm, there are e-steps and m-steps that are executed one after another until the model converges. In the e-step the gamma values are calculated. In the m-step, the new parameters are calculated for each cluster using the gamma values calculated in the e-step, and the new parameters are stored in a class variable. After reaching to a convergence in a single run, all calculated parameters are added to a list. After n runs, the parameters of the run with the highest likelihood value is selected as the final parameters. If the visualization parameter is set to true, the final clusters are plotted and saved as a PNG file.

Used libraries:

- For visualization **matplotlib** is used.
- To generate semi positive definite matrix in the initialization of sigma **sklearn.dataset** is used.
- In the initialization of pi values, the softmax function of **scipy.special** is used.
- At the calculations of the mutivariate normal distribution, **scipy.stats** is used.
- Other parts are written using pure **numpy** functions

2 Results

	x	y
cluster 1	4.37901524	4.35181941
cluster 2	9.60515914	9.16835945
cluster 3	0.70207122	0.66133847

Table 1: Mean values of clusters after the convergence

	Covariance Matrix
cluster 1	[2.74795674, -0.11917498] [-0.11917498, 0.61809318]
cluster 2	[2.01245136, -0.64166751] [-0.64166751, 0.82171149]
cluster 3	[2.11797871, -0.09975965] [-0.09975965, 0.6407947]

Table 2: Covariance Matrix of clusters after the convergence

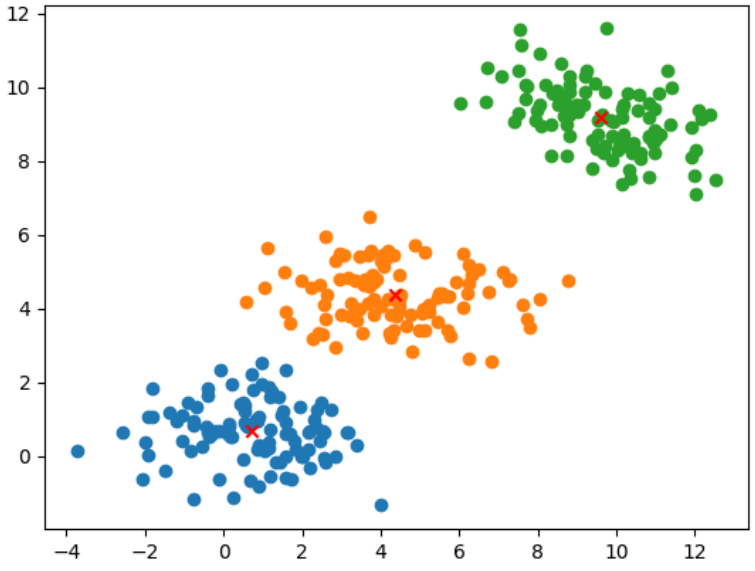

Figure 1: Final clusters (3 clusters)

3 Bonus

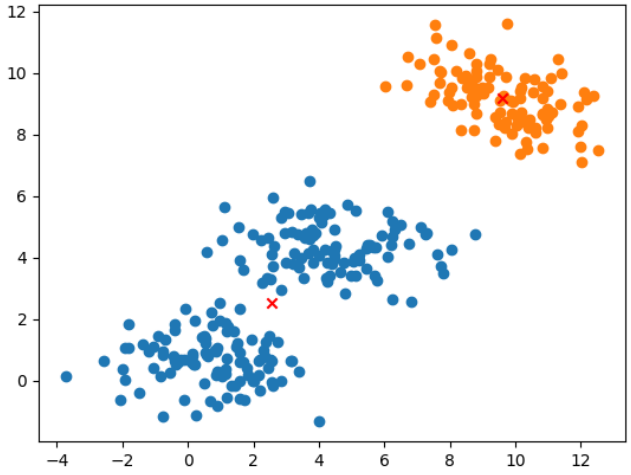

Figure 2: Final clusters (2 clusters)

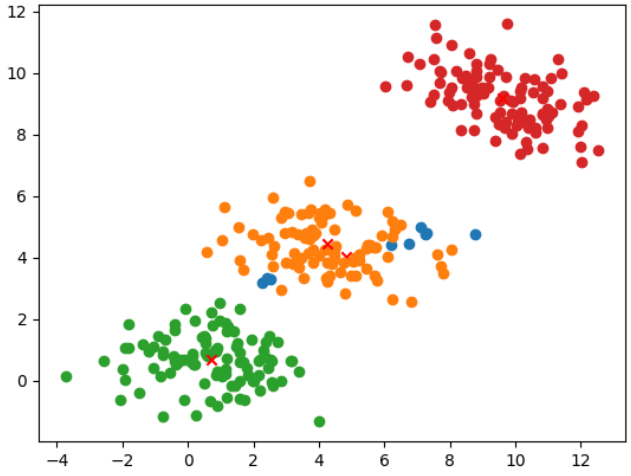

Figure 3: Final clusters (4 clusters)

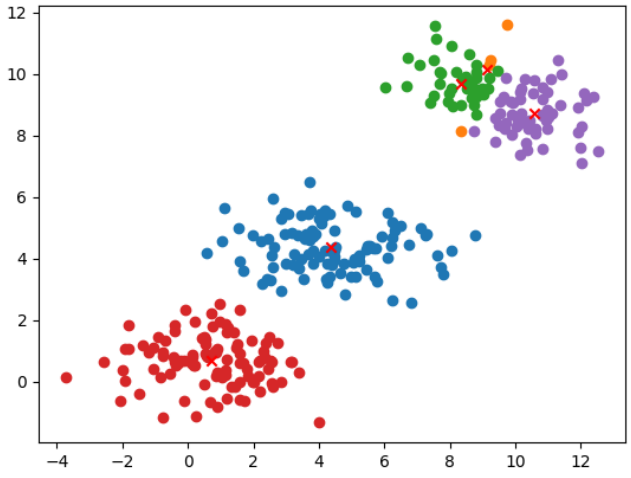

Figure 4: Final clusters (5 clusters)