Thomas Charuel - Kyota Lannelongue - Hugo Cambier

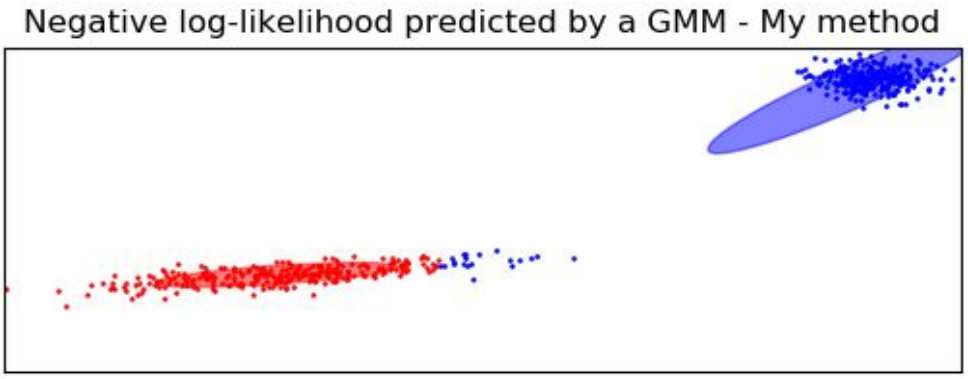
# Session 3 Report

In this homework we study the Gaussian Mixture Model. This model allows us to study classification problems through the density function estimation. In a classification problem like such, we are dealing with unsupervised learning where we do not know the outputs of the data. The objective is to create efficient clusters in order to classify efficiently our inputs.

Here we have a dataset of dots generated. We want to classify it knowing that there is 2 cluster that we want to create. GMM, much like K-Means, allows us to answer such answer. The difference being that GMM takes variance into consideration when it calculate the measurement, whereas K-Means only calculates conventional Euclidean distances.

We write down a specific bootstrap procedure, with the idea to guess the value of z (i) using a model. Afterwards, we use the estimated z (i) to fit the parameters of the model. Then, we iterate these two steps until we have a better guess of z (i) , while optimizing the model parameters. This bootstrap procedure is known as the Expectation-Maximization (EM) algorithm. The E-step guess the value of z (i) and the M-step update the parameters according to the new z (i).

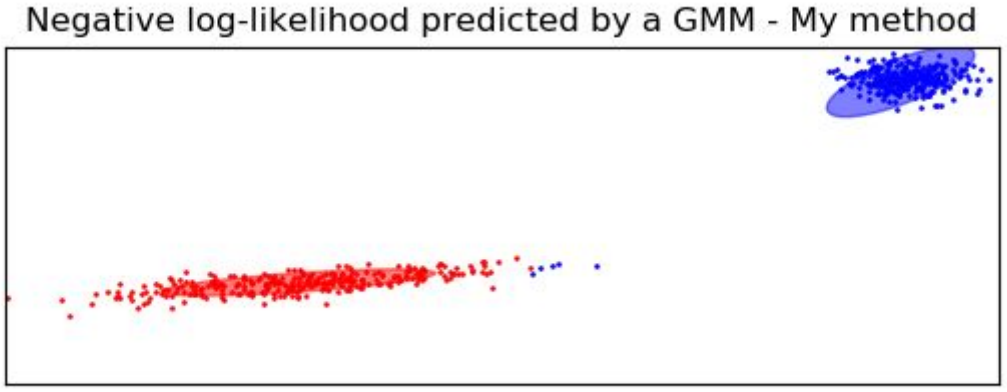
Here is a plot of the implemented Expectation-Maximization algorithm running. On 1 iteration we can see that some data points clearly fail to get in the right cluster.




```
Log-likelihood of the best GMM = -72121.98908082508
Phi of the best GMM = 0.466769369444
Mu of the best GMM = [[ -0.61070372  -0.16645449]
 [ 19.08046897  18.8102089 ]]
Sigma of the best GMM = [[[  8.64564078   1.71143026]
  [  1.71143026   0.775077  ]]

 [[ 11.51674249  14.64075055]
  [ 14.64075055  21.3852369 ]]]
```

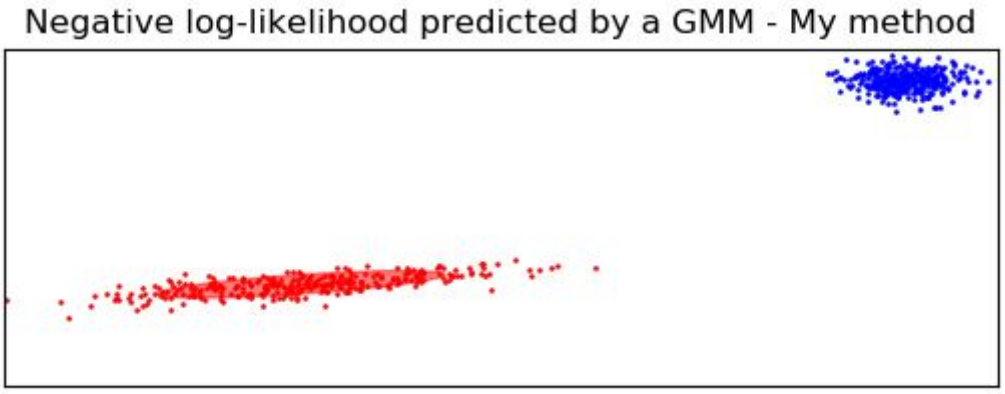
As the iterations progresses, we can see that the classification is doing better and the data points are joining their meant-to-be clusters. Phi, Mu and Sigma all seem to evolve in the right direction.




```
Log-likelihood of the best GMM = -3046.54952456787
Phi of the best GMM = 0.492387168667
Mu of the best GMM = [[ -0.26651599  -0.09362649]
 [ 19.74036342  19.69726445]]
Sigma of the best GMM = [[[ 10.38035907   2.08505833]
  [  2.08505833   0.85366503]]

 [[  3.00865745   3.18961792]
  [  3.18961792   6.06435518]]]
```

We can see after 3 iterations that all the data points seems to have joined the right clusters.




An error occurs after 3 iterations, this must be because of a problem in the implementations and the evolutions of the variables.

Those are the final values given by the GMM after 3 iterations. The results seem to correspond well with what we were looking for. Phi stabilized at 0,5.

```
Log-likelihood of the best GMM = -2672.419251753084
Phi of the best GMM = 0.5
Mu of the best GMM = [[ -0.13607006  -0.07059606]
 [ 19.91453549  19.97556345]]
Sigma of the best GMM = [[[  1.13328030e+01   2.25048269e+00]
  [  2.25048269e+00   8.77008968e-01]]

 [[  1.02179864e+00   3.28158686e-03]
  [  3.28158686e-03   9.90374215e-01]]]
```