# Estimate Gaussian Mixture Model with EM Algorithm

In this assignment, you will be given ***n*** data points, each of which has ***m*** attributes. The samples are generated from a mixture of a ***k*** number of *unknown* Gaussian distributions. This data is often referred to as Gaussian Mixture Model (GMM). Your task is to estimate the parameters of ***k*** unknown Gaussian distributions. You will be using the EM (Expectation-Maximization) algorithm for this task. Please refer to the class materials for the mathematical backgrounds of this algorithm.

Datasets > [Assignment 3 Materials](https://drive.google.com/drive/folders/1qAhXcGh7b-keuK6d4NRihuE3AsOyOTxa?usp=sharing)

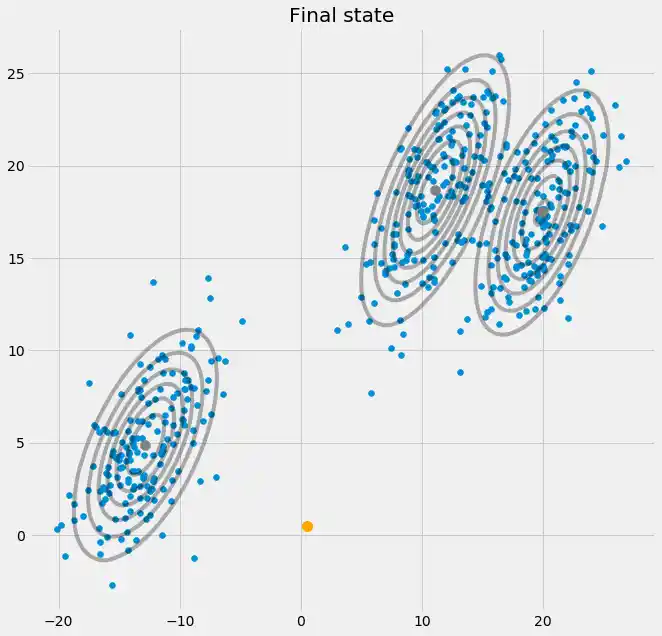
## Task 1: EM Algorithm

* Take a data file as input. The data file contains ***n*** data points, each having ***m*** attributes.
* As the number of components (or, the number of gaussian distributions, ***k***) is usually unknown, you will assume a range for ***k***. For example, from 1 to 10.
* For each value of ***k***,
  + Apply the EM algorithm to estimate the GMM.
  + Keep a note of the converged log-likelihood.
* Show a plot of how converged log-likelihood varies with the number of components (***k***). Choose an appropriate value for ***k*** from this plot. Let’s call it ***k\****.

Now that you have estimated the number of gaussian distributions, the next step is to visualize the EM algorithm for ***k=k\****.

## Task 2: Visualization (m=2)

If the number of attributes is equal to 2, you need to show plots of estimated GMM. After each iteration (an E-step and an M-step), plot the data points and gaussian distributions in a 2D plot. Do not save the plots to a file. After running the EM algorithm, the plot should update as the algorithm advances ([similar to this](https://media.vlpt.us/images/gibonki77/post/32212164-5343-42ee-a2e0-8443ffdb19ec/Gaussian_Mixture_Models__example.gif)).



*Fig. Sample plot after an iteration for 3 components*

## Task 3 (Bonus): Visualization (m>2)

After each iteration (an E-step and an M-step), plot the data points and gaussian distributions in a 2D plot. You can use **PCA, UMAP** or **t-SNE** for dimensionality reduction.

## Additional Information

* You will be given a new dataset during evaluation. Write your program in such a way that a new dataset can be incorporated without any major change.
* Acceptable python libraries for Task 1 and 2: **NumPy, Pandas, Matplotlib, Seaborn**
* You can use any library for Task 3.

## Submission

1705xxx

|-- \*.py

Zip the folder (1705xxx) to 1705xxx.zip

Submit the zipped file.

**Deadline: 11.55 PM. 13th January 2023.**

## Questions to Study

* Explain the terms “hard assignment” and “soft assignment” in light of clustering.
* What are the advantages of GMM clustering compared to k-means clustering?
* What are the intuitions behind the equations of the E-step and M-step of the EM algorithm?
* How to decide ***k***?
* x, y = np.mgrid[np.min(self.data[:, 0]):np.max(self.data[:, 0]):.01,  
   np.min(self.data[:, 1]):np.max(self.data[:, 1]):.01]  
  pos = np.empty(x.shape + (2,))  
  pos[:, :, 0] = x;  
  pos[:, :, 1] = y  
  for i in range(k):  
   rv = multivariate\_normal(mean=means[i], cov=covs[i])  
   ax.contour(x, y, rv.pdf(pos), colors='k')  
  plt.show()  
  plt.pause(0.1) # pause for 0.1 seconds to update the plot  
  ax.clear() # clear the previous plot  
  ax.scatter(self.data[:, 0], self.data[:, 1]) # scatter the points again