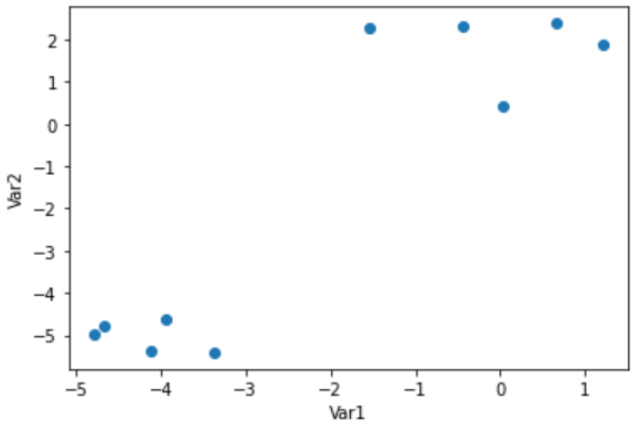
CLL 788 Assignment 3 Report

Aditi Singh
2017CH10188

Q1. Manually perform K Means clustering on Manual_Data.xlsx. There are 10 data points given and you have to separate them into 2 clusters.

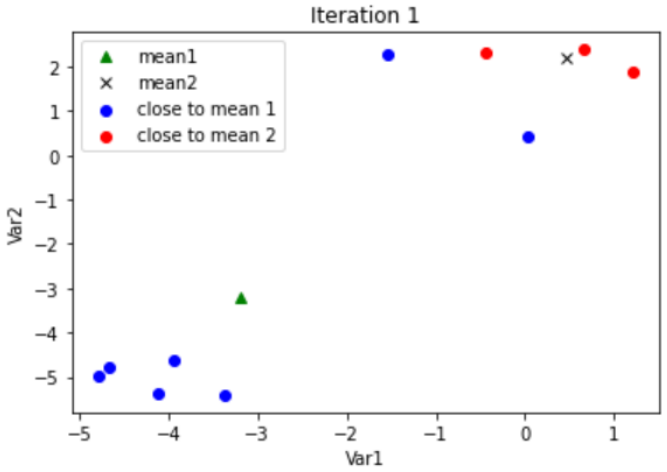
To manually perform K Means, I used excel sheet to do calculations which are provided in the same directory, it is named as 'q1_calculations.xlsx', the plotting has been done in python jupyter notebook.
Now let us plot the data points



From the above figure we can see that one can divide the dataset into 2 clusters.
Now to start the iterations I randomly selected the two means as (0,0) and (1,1)

		1st Iteration		
Var1	**Var2**	**Distance From 1st Mean**	**Distance From 2nd Mean**	**Class**
-1.54	2.29	2.759655776	2.848806768	1
-0.44	2.34	2.38100819	1.967028215	2
0.03	0.41	0.411096096	1.135341358	1
1.2	1.87	2.22191359	0.892692556	2
0.65	2.39	2.476812468	1.433387596	2
-4.67	-4.8	6.696932133	8.111035692	1
-3.37	-5.41	6.373774392	7.7578992	1
-3.93	-4.64	6.080666082	7.49096122	1
-4.78	-4.96	6.88839604	8.302409289	1
-4.12	-5.36	6.760473356	8.164802508	1

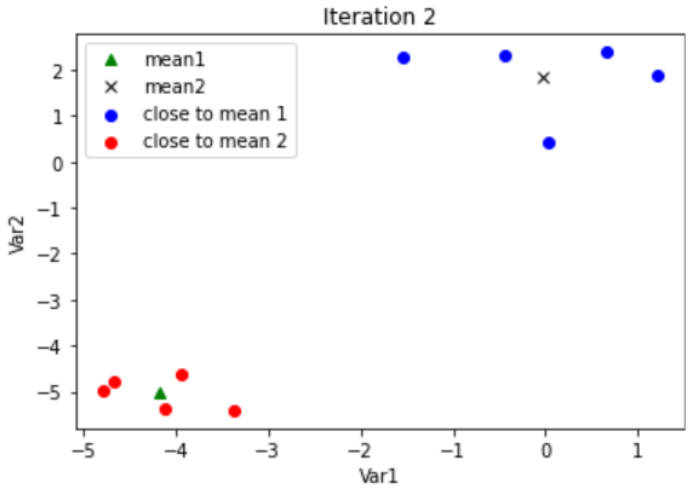
From this iteration, one can see that 3 data points belong to class 2 while 7 belong to class 1, now for these new classes we get the means as **(-3.197,-3.21)** and **(0.47,2.2)**



Now, using these means as new centers of clusters we get the next set of iterations as given:

		2nd Iteration		
Var1	**Var2**	**Distance From 1st Mean**	**Distance From 2nd Mean**	**Class**
-1.54	2.29	5.744183928	2.012013916	2
-0.44	2.34	6.197059706	0.920706251	2
0.03	0.41	4.84952874	1.843285111	2
1.2	1.87	6.718631483	0.80112421	2
0.65	2.39	6.794071607	0.261725047	2
-4.67	-4.8	2.167447577	8.684445866	1
-3.37	-5.41	2.206791562	8.523948616	1
-3.93	-4.64	1.606919102	8.132994528	1
-4.78	-4.96	2.359743418	8.878519021	1
-4.12	-5.36	2.339749773	8.844303251	1

In this iteration, one can see that now 5 points are in cluster 1 and 5 in cluster 2 which is similar to the cluster manually estimated in the first figure, we get the final means as **(-4.174,-5.034)** and **(-0.02,1.86)**

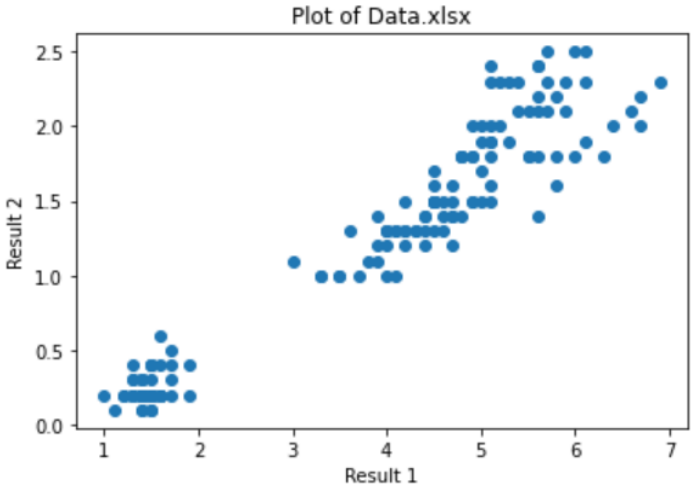


Now after doing another iteration we get the same means thus verifying that 2 iterations were enough to get 2 separate classes.

		3rd Iteration		
Var1	Var2	Distance From 1st Mean	Distance From 2nd Mean	Class
-1.54	2.29	7.78324688	1.57965186	2
-0.44	2.34	8.265508575	0.637808749	2
0.03	0.41	6.878281181	1.450861813	2
1.2	1.87	8.749005201	1.220040983	2
0.65	2.39	8.853629312	0.854283325	2
-4.67	-4.8	0.548426841	8.122690441	1
-3.37	-5.41	0.887576476	8.004711113	1
-3.93	-4.64	0.463435001	7.585387268	1
-4.78	-4.96	0.610501433	8.316850365	1
-4.12	-5.36	0.330442128	8.302915151	1

Q2. Carbon and nitrogen emission tests of 2 different types of vehicles were done. Test results are provided in Excel sheets. Your task is to identify the two groups of vehicles from the data.

a) Plot the data (Data.xlsx) to get an idea of the data distribution.



We can see that the data provided should be divided into two clusters following are the properties of the dataset.

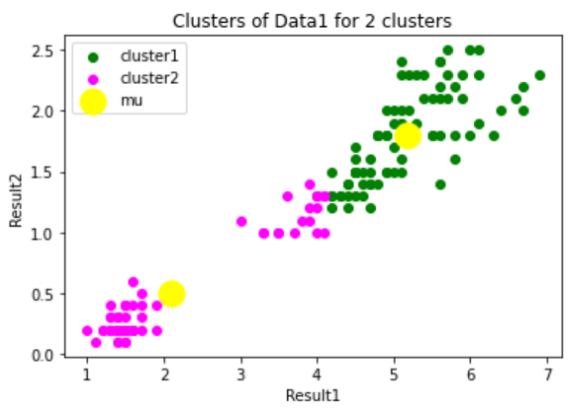
	Result 1	Result 2
count	150.000000	150.000000
mean	3.758000	1.199333
std	1.765298	0.762238
min	1.000000	0.100000
25%	1.600000	0.300000
50%	4.350000	1.300000
75%	5.100000	1.800000
max	6.900000	2.500000

b) Apply K-Means clustering on the data to find out the 2 clusters.

K means clustering algorithm's code was developed to get the clusters, to start the clustering the means selected were chosen to be random data points from the dataset. After approximately **6 iterations** means got converged below are the iterations of one of the K means algorithm
(Since starting mean is selected at random thus at every running of code one will get different set of iterations photos but in all cases at max it takes 6 iterations to converge)

```
Iteration  1
```


```
Iteration  2
```

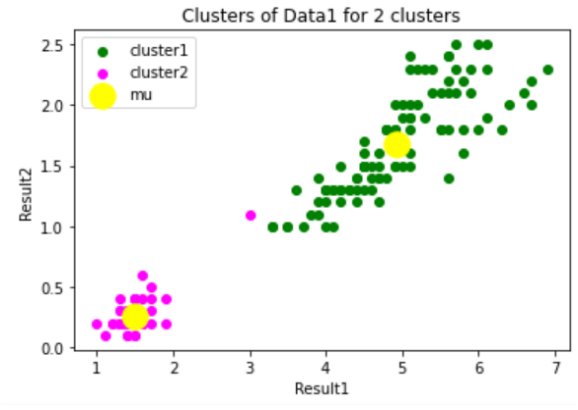



```
Iteration  3
```


```
Iteration  4
```

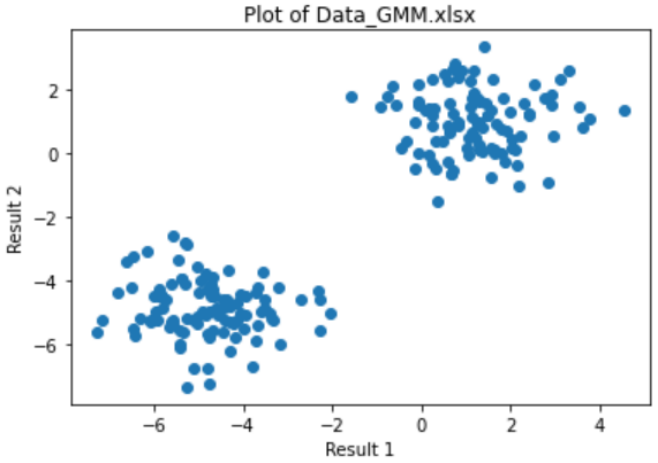

```
Iteration  5
```


```
Iteration  6
```



The converged means obtained were:
cluster 1 (99 points): **[4.92525253, 1.68181818],**
cluster 2 (51 points): **[1.49215686, 0.2627451]**
From the above case one can also observe that point (3.1,1) is mis classified it should have been classified with cluster 1

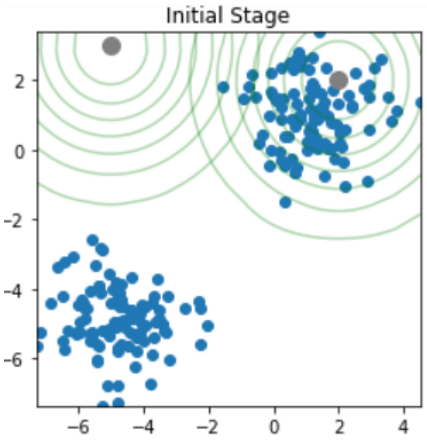
c) Plot the data (Data_GMM.xlsx) to get an idea of the data distribution.

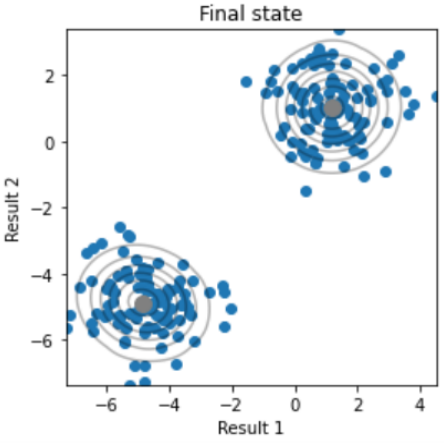


We can see that the data provided should be divided into two clusters following are the properties of the dataset.

	Result 1	Result 2
count	200.000000	200.000000
mean	-1.810469	-1.931323
std	3.205968	3.094894
min	-7.275102	-7.347239
25%	-4.799860	-4.949301
50%	-1.807678	-2.031475
75%	1.151394	1.041118
max	4.526678	3.365225

d) Apply Gaussian Mixture Model on the Data_GMM.xlsx to find out the 2 clusters.
GMM algorithm's code was developed to get the distributions, to start with the means were randomly chosen. After approximately **8 iterations** means got converged below are the initial randomised and final state of the GMM algorithms
(Since starting mean is selected at random thus at every running of code one will get different set of photos for the initial and final state but in all cases at max it takes 8 iterations to final distributions)





The following gaussian model characteristics were obtained:

Cluster 1:

Mean= (-4.818151 , -4.87158382),

Covariance Matrix = [[1.1696876 , -0.10528557],[-0.10528557, 0.80018897]]

Cluster 2:

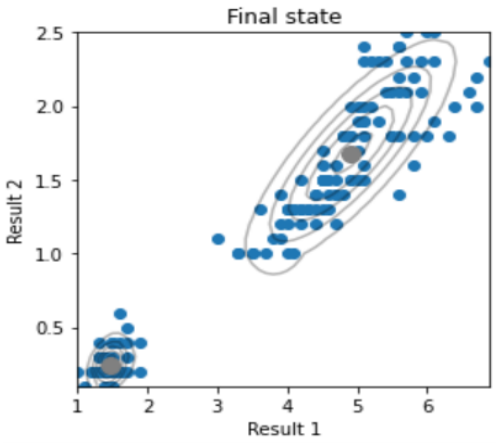
Mean= (1.197213 , 1.00893773),

Covariance Matrix = [[1.19169011, 0.00910688], [0.00910688, 0.97049769]]

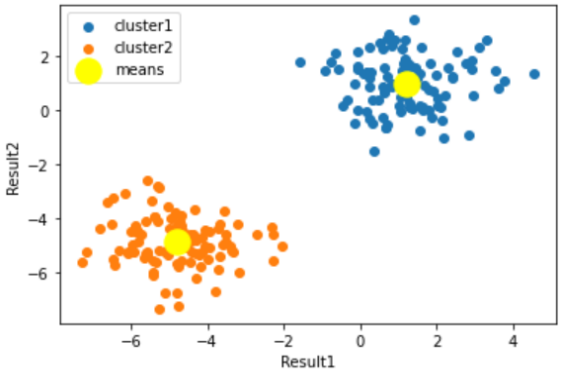
e) Compare the two methods used.

To compare the two methods, I fitted the data1 with Gaussian Mixture Model and data2 with K means clustering algorithm

<u>GMM for first data</u>



In this we can now see that the misclassified point in means fit is correctly classified with GMM while this algo takes approximately 12 iterations to get the gaussian models, which is accurate enough as it correctly classifies the misclassified point as in K means

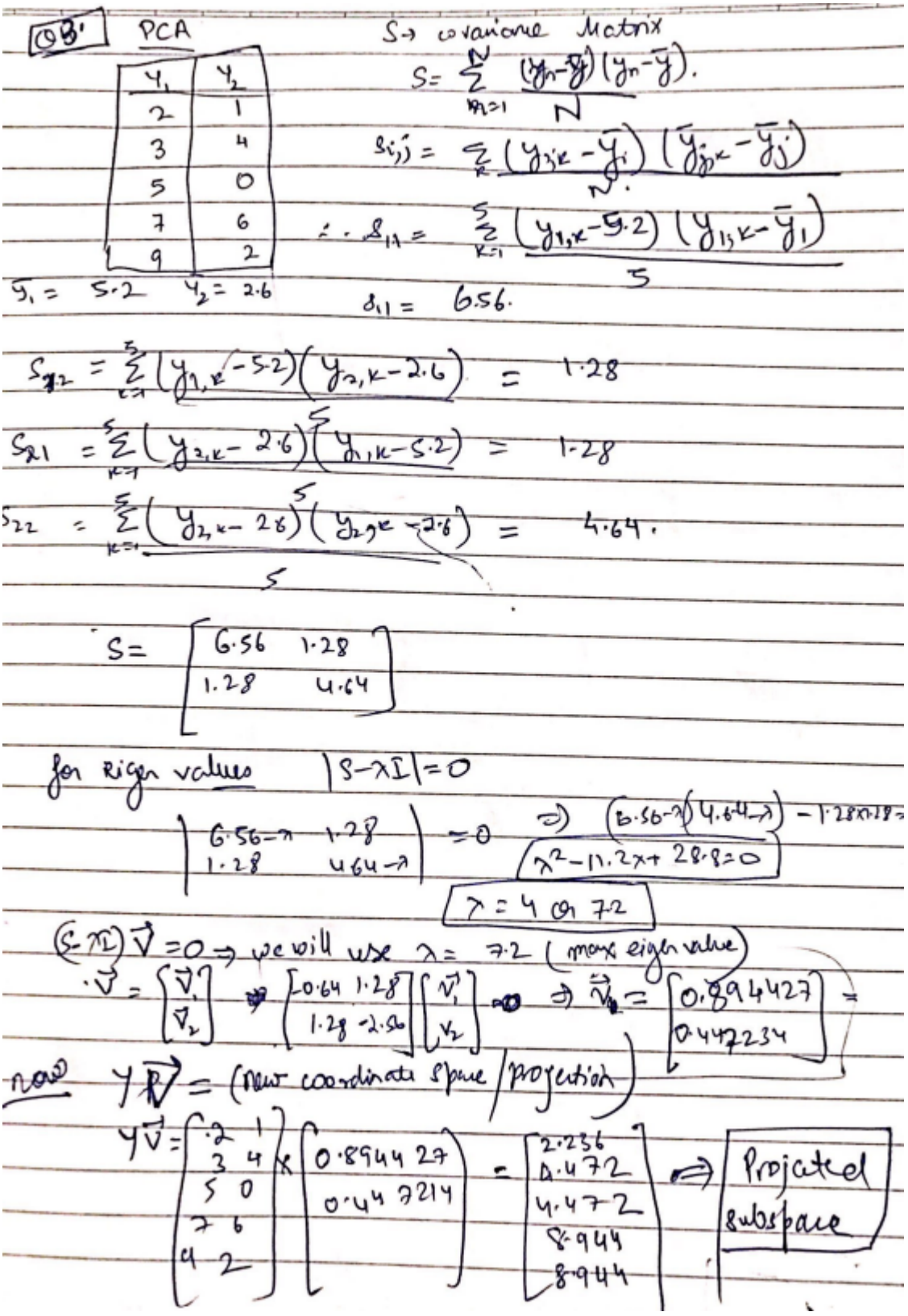


<u>K Means Clustering for 2nd Data</u>

We get the exactly same clusters as with GMM in about 6 iterations.

Out of the 2 GMM is more effective in our case as it correctly classifies even that one point which was misclassified for K means for 1st data, But in K means takes less iterations to converge but overall both the unsupervised algorithm works fine.

Q3. Manually(perform the PCA analysis of the data vectors Y1, Y2 presented. Calculate the eigen vectors corresponding to two principle directions and transform the data into the new coordinate space.



Q3. PCA

Y_1	Y_2
2	1
3	4
5	0
7	6
9	2

$\bar{Y}_1 = 5.2 \quad \bar{Y}_2 = 2.6$

S → covariance Matrix

$$S = \sum_{n=1}^{N} \frac{(y_n - \bar{y})(y_n - \bar{y})}{N}$$

$$s_{ij} = \sum_{k} \frac{(y_{ik} - \bar{y}_i)(\bar{y}_{jk} - \bar{y}_j)}{N}$$

$$\therefore s_{11} = \sum_{k=1}^{5} \frac{(y_{1,k} - 5.2)(y_{1,k} - \bar{y}_1)}{5}$$

$s_{11} = 6.56$

$$s_{12} = \sum_{k=1}^{5} (y_{1,k} - 5.2)(y_{2,k} - 2.6) = 1.28$$

$$s_{21} = \sum_{k=1}^{5} (y_{2,k} - 2.6)(y_{1,k} - 5.2) = 1.28$$

$$s_{22} = \sum_{k=1}^{5} \frac{(y_{2,k} - 2.6)(y_{2,k} - 2.6)}{5} = 4.64.$$

$$S = \begin{bmatrix} 6.56 & 1.28 \\ 1.28 & 4.64 \end{bmatrix}$$

for Eigen values $|S - \lambda I| = 0$

$$\begin{vmatrix} 6.56 - \lambda & 1.28 \\ 1.28 & 4.64 - \lambda \end{vmatrix} = 0 \Rightarrow (6.56 - \lambda)(4.64 - \lambda) - 1.28 \times 1.28 =$$

$$\lambda^2 - 11.2\lambda + 28.8 = 0$$

$\lambda = 4$ or 7.2

$(S - \lambda I)\vec{V} = 0 \Rightarrow$ we will use $\lambda = 7.2$ (max eigen value)

$$\vec{V} = \begin{bmatrix} \vec{V}_1 \\ \vec{V}_2 \end{bmatrix} \Rightarrow \begin{bmatrix} -0.64 & 1.28 \\ 1.28 & -2.56 \end{bmatrix} \begin{bmatrix} \vec{V}_1 \\ V_2 \end{bmatrix} \Rightarrow \vec{V}_1 = \begin{bmatrix} 0.894427 \\ 0.447234 \end{bmatrix}$$

now $Y\vec{V}$ = (new coordinate space / projection)

$$Y\vec{V} = \begin{bmatrix} 2 & 1 \\ 3 & 4 \\ 5 & 0 \\ 7 & 6 \\ 9 & 2 \end{bmatrix} \times \begin{bmatrix} 0.894427 \\ 0.447214 \end{bmatrix} = \begin{bmatrix} 2.236 \\ 4.472 \\ 4.472 \\ 8.944 \\ 8.944 \end{bmatrix} \Rightarrow \text{Projected subspace}$$