CS 335 & CS 337

Adarsh Raj - 190050004

Assignment-4

October 2021

Question 1 - Clustering

1.1 - KMeans Implementation

(i) Refer to `assignment_4.ipynb` for `fit()` and `predict()` function of `Kmeans` class.

(ii) Clusters Formed using Kmeans Algorithm for three different seeds are as follows. The seeds used as **123, 200, 400**, for each of the three data sets.

- Kmeans on Dataset 1 and seed 123

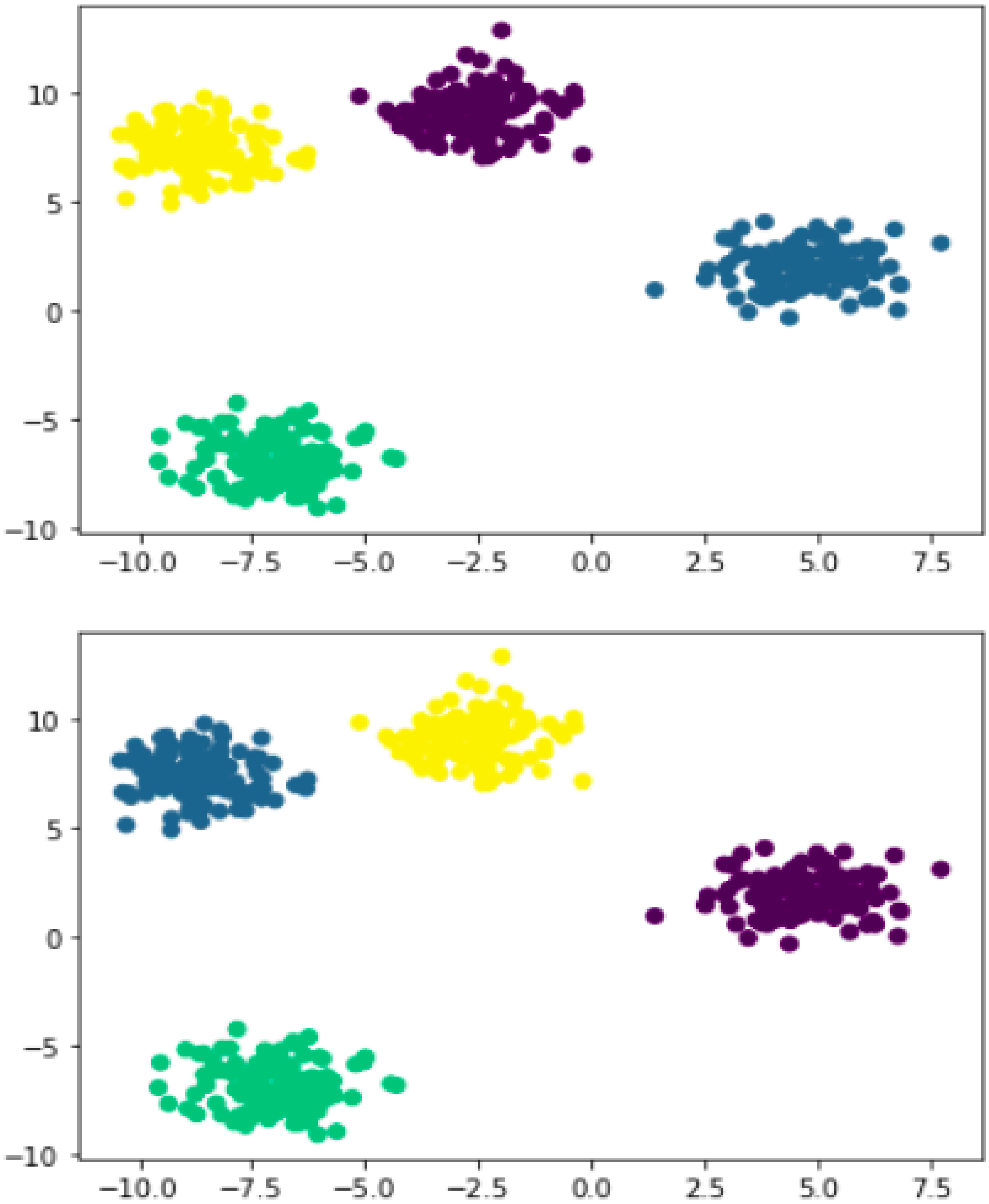

Figure 1: Seed - 123 and Dataset - 1

- Kmeans on Dataset 2 and seed 123

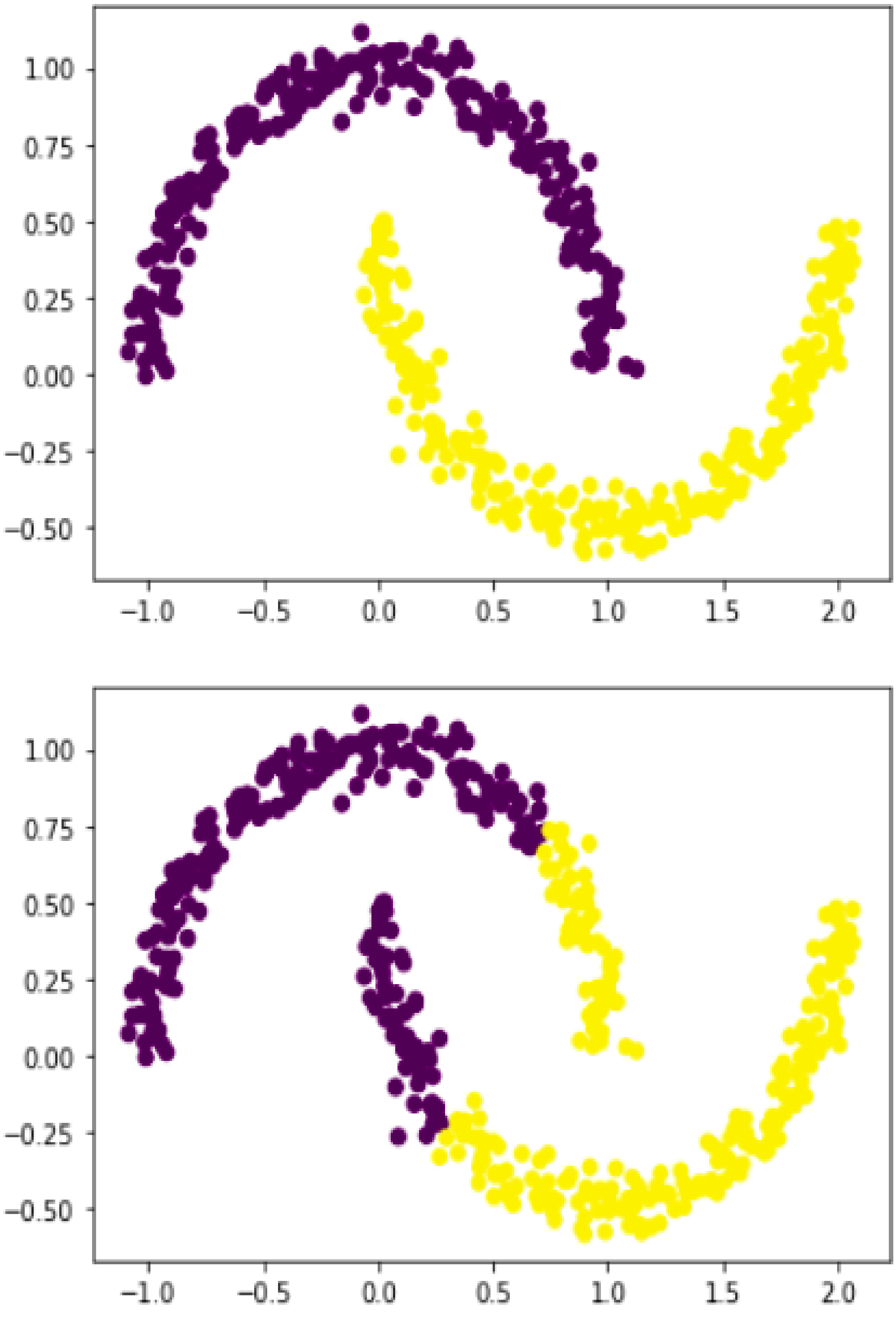

Figure 2: Seed - 123 and Dataset - 2

- Kmeans on Dataset 3 and seed 123

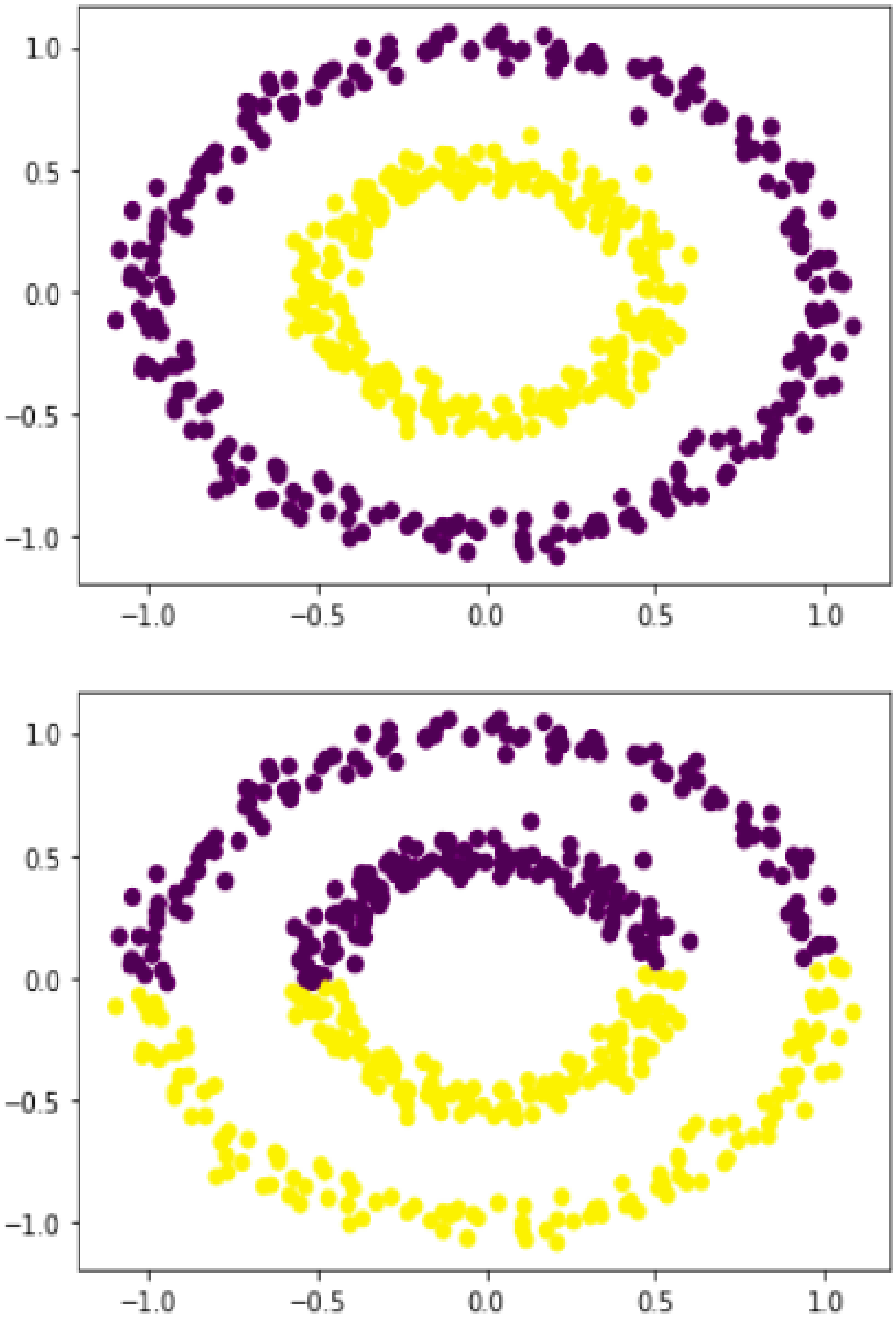

Figure 3: Seed - 123 and Dataset - 3

- Kmeans on Dataset 1 and seed 200

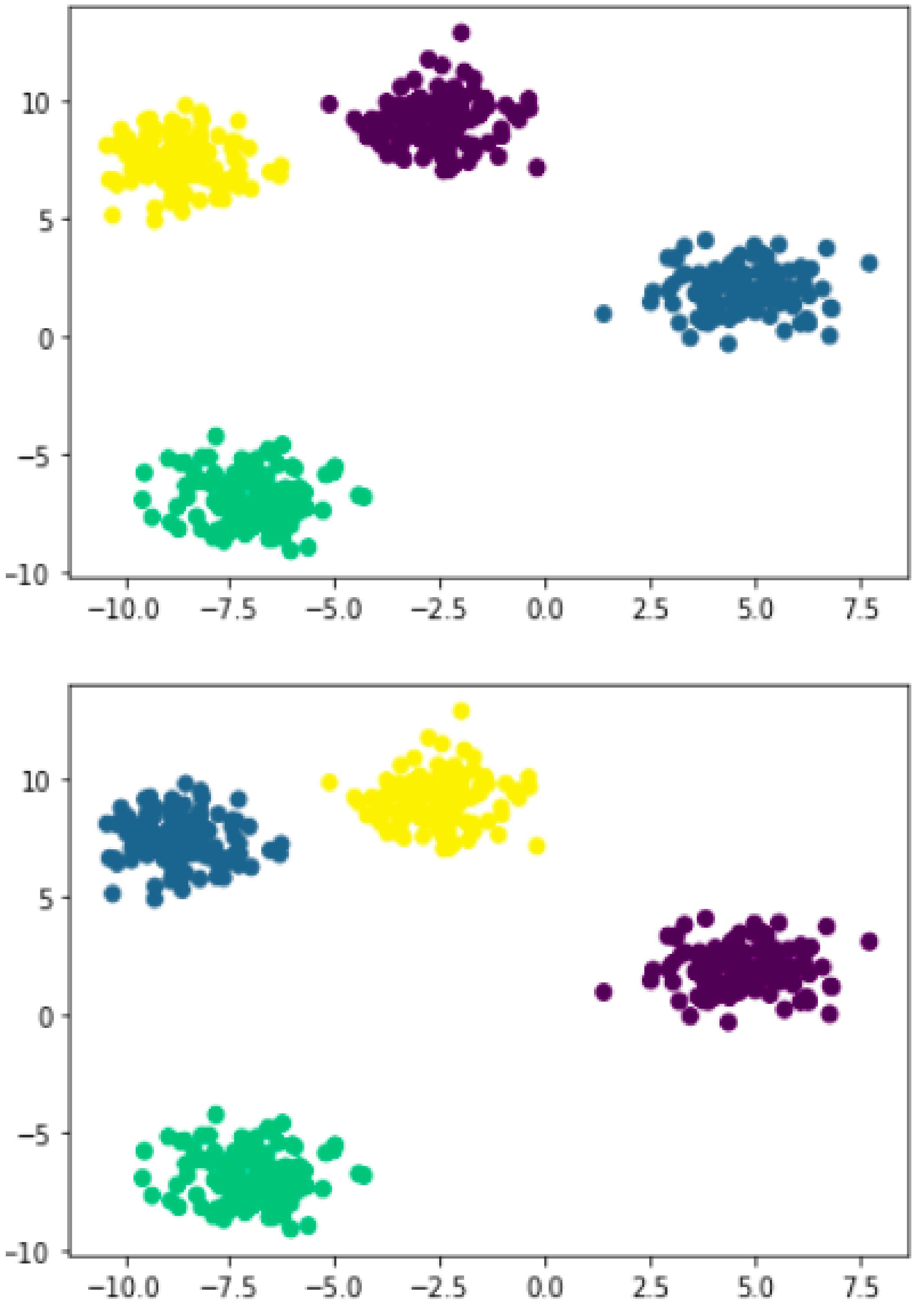

Figure 4: Seed - 200 and Dataset - 1

- Kmeans on Dataset 2 and seed 200

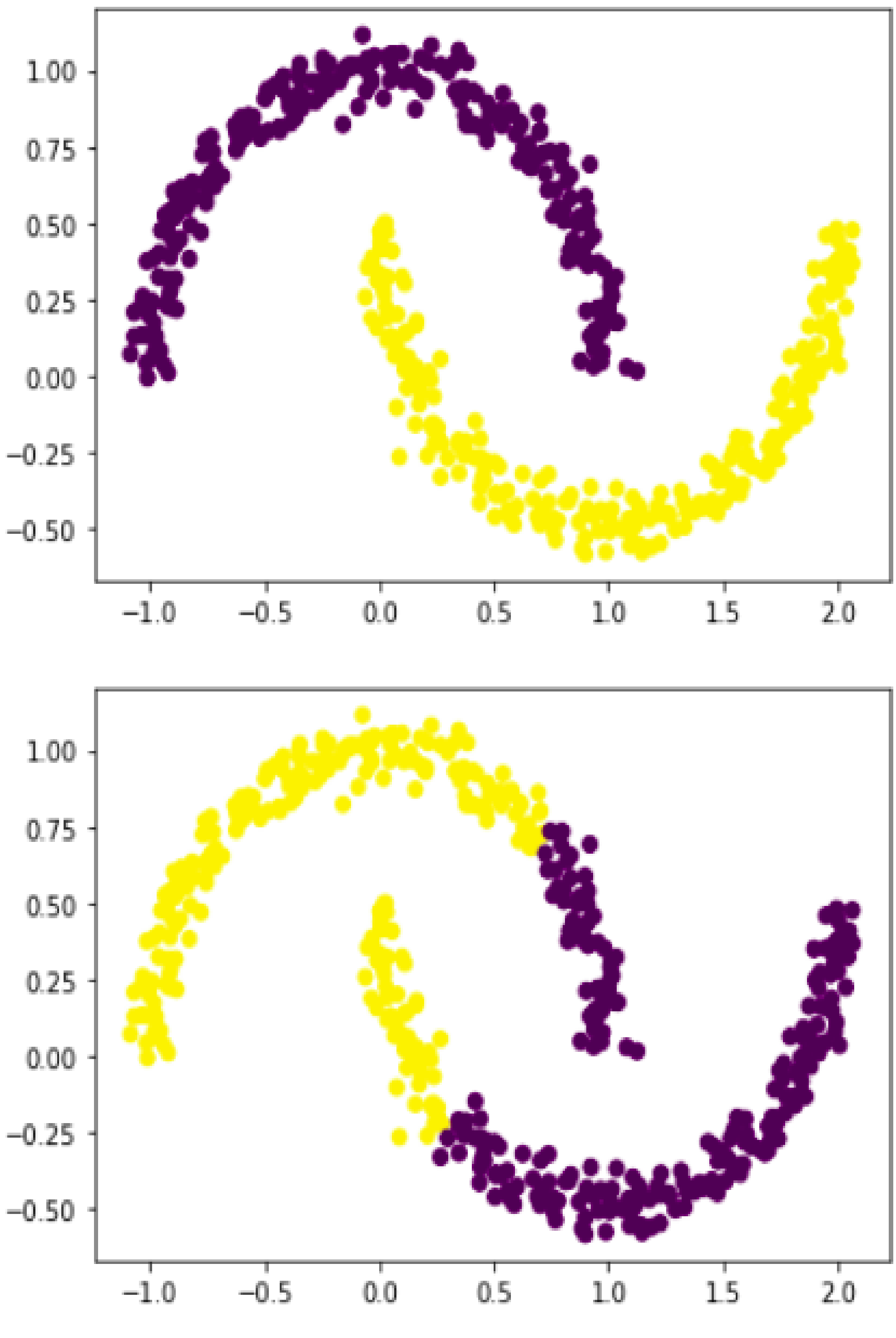

Figure 5: Seed - 200 and Dataset - 2

- Kmeans on Dataset 3 and seed 200

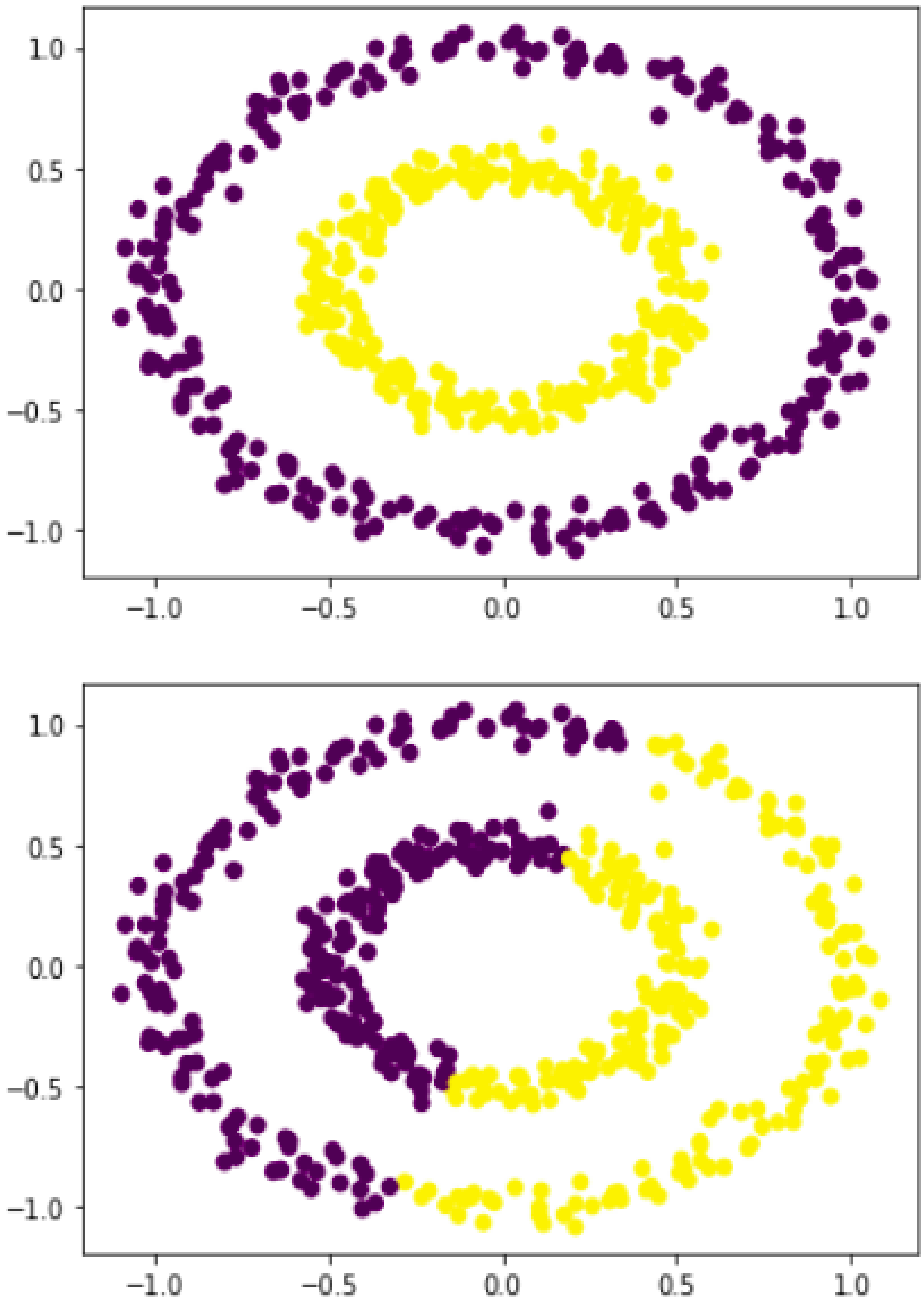

Figure 6: Seed - 200 and Dataset - 3

- Kmeans on Dataset 1 and seed 400

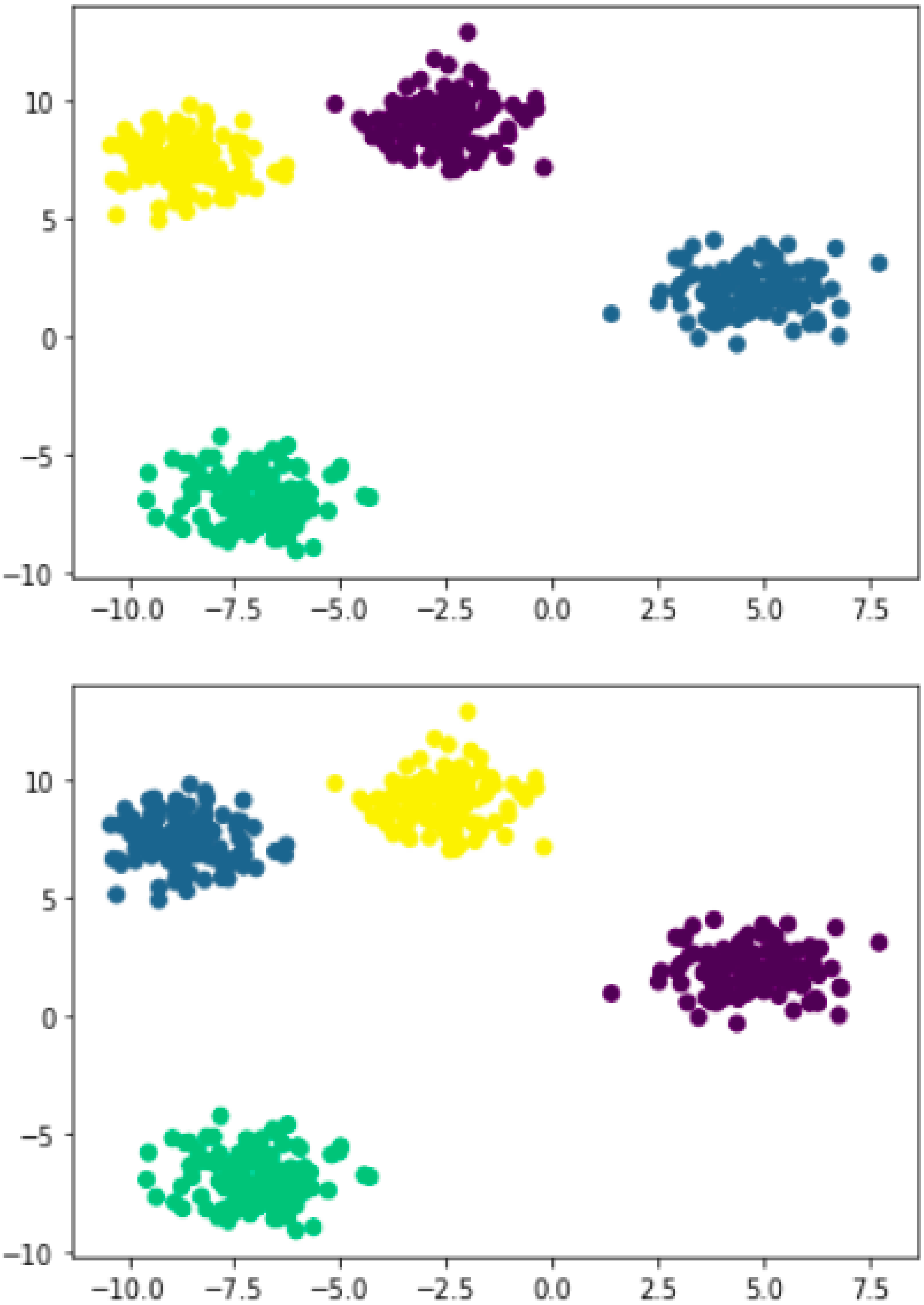

Figure 7: Seed - 400 and Dataset - 1

- Kmeans on Dataset 2 and seed 400

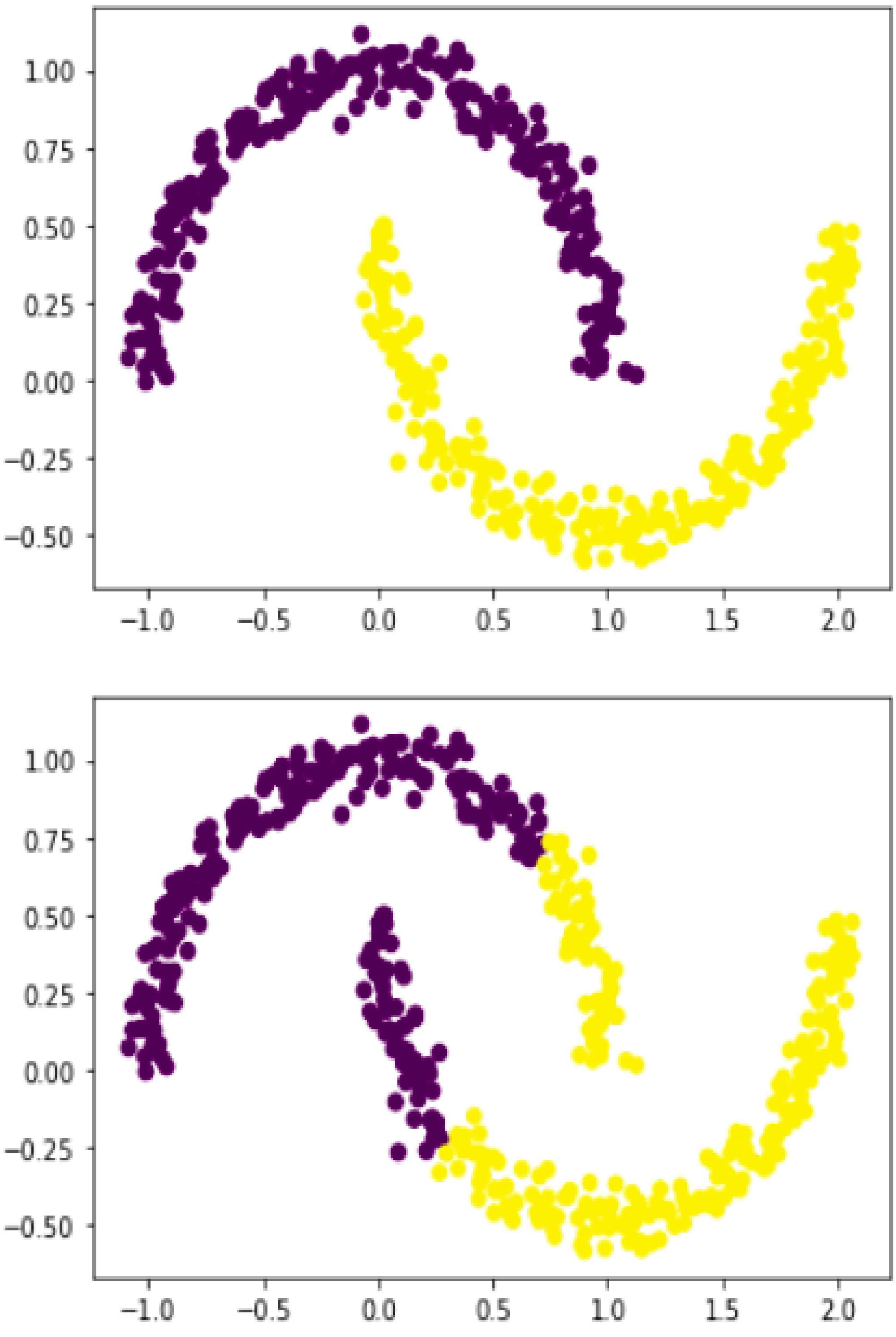

Figure 8: Seed - 400 and Dataset - 2

- Kmeans on Dataset 3 and seed 400

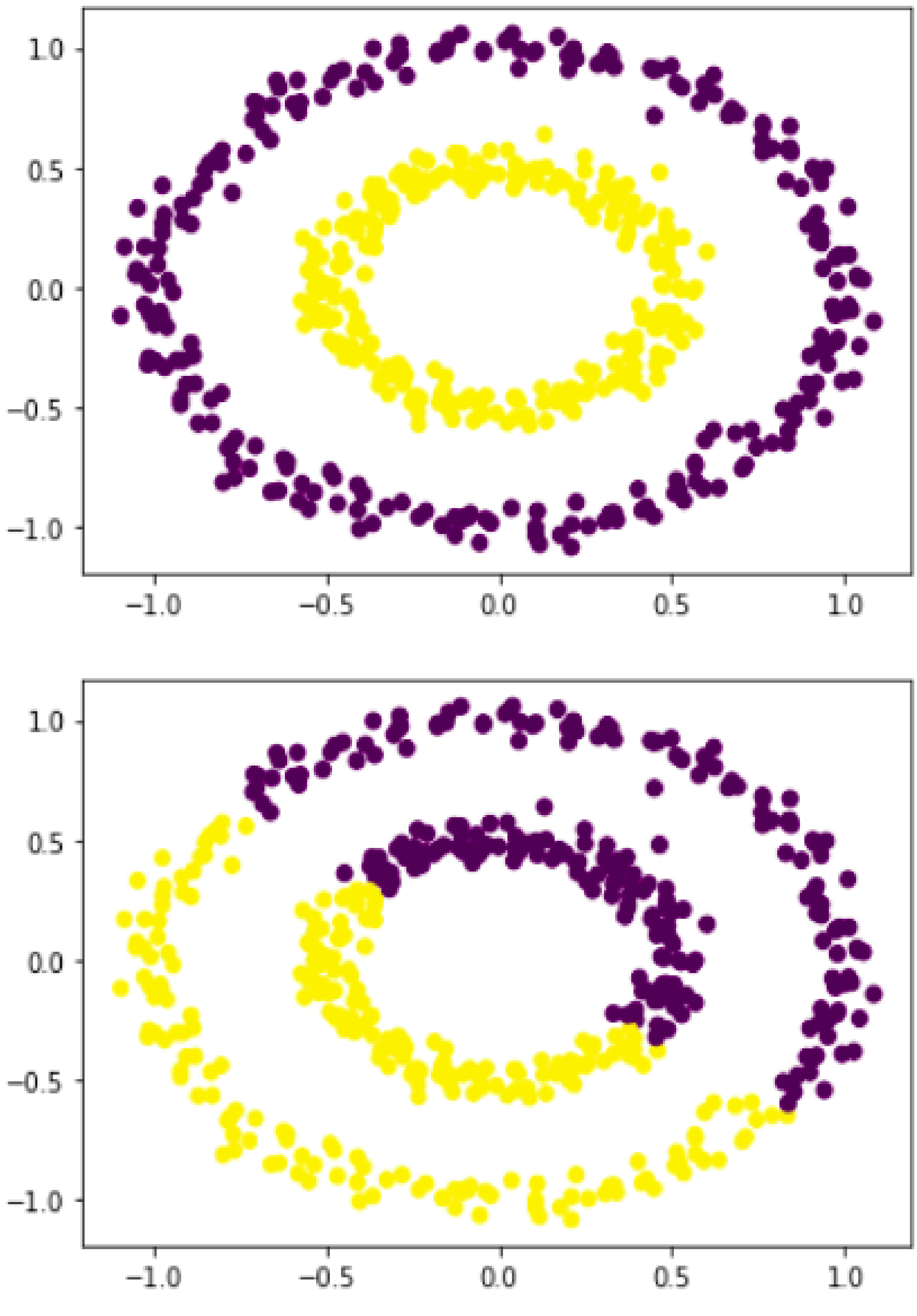

Figure 9: Seed - 400 and Dataset - 3

Comments:

- For the first dataset, since all the clusters are linearly separable, normal kmeans works fine. But, with random initialisation of initial centroids, the result will not be good for some seeds due to bad initialisation of centroids.

  Hence, I used spaced out initialisation for initial centroids, which uses distances from selected centroids and choose initial centroids as far as possible from each other, which is optimal. Hence, the clusters are converged optimally for all seeds for first dataset.

- For second and third datasets, since the datasets are not linearly separable, and Kmeans results in clusters for linearly separable datasets, we get a line trying to separate two clusters for these datasets, but that was not desired for, as the separation is higher dimensional for second and third datasets. Therefore kernel kmeans can be used for these datasets, with appropriate kernels.

(iii) **Good Initialisation of Kmeans:**

A good initialisation can be done by selecting cluster centers far from each other, so that the data space can be uniformly clustered by normal KMeans. Folowing steps can be taken:

- Select the first cluster center randomly from the datapoints.
- After that, we can ensure that the next cluster center is spaced far from the chosen cluster center. This can be done by calculating distance of all datapoints from the selected cluster center and choosing the datapoint with maximum distance value, as the next cluster center.
- Repeat the above step till all cluster centers are initialised, by ensuring far distance from all previously selected centers.

This is what I have implemented in my code. This initialisation ensures that the algorithm has an idea of the size of the dataset which result in faster convergence for individual clusters and thus, generating optimal results, for different seeds.

Question 2 - Kernel Design and Kernelized Clustering

2.1 - Proving Kernel Validity

To Prove: The function $K_\sigma : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined as $K_\sigma(\mathbf{x}, \mathbf{y}) = \exp\left(\frac{-||x-y||^2}{2\sigma^2}\right)$ is a valid Kernel.

Proof:

We have,

$$K(x, y) = \exp\left(-\frac{||x-y||^2}{2\sigma^2}\right) = \exp\left(-\frac{||x||^2}{2\sigma^2} - \frac{||y||^2}{2\sigma^2} + \frac{x^T y}{2\sigma^2}\right)$$

$$K(x, y) = \exp\left(-\frac{||x||^2}{2\sigma^2}\right)\exp\left(-\frac{||y||^2}{2\sigma^2}\right)\exp\left(\frac{x^T y}{2\sigma^2}\right)$$

Now, using Taylor's expansion for third term, we have:

$$\exp\left(\frac{x^T y}{2\sigma^2}\right) = 1 + \frac{x^T y}{2\sigma^2} + \frac{1}{2!}.\left(\frac{x^T y}{2\sigma^2}\right)^2 + \frac{1}{3!}.\left(\frac{x^T y}{2\sigma^2}\right)^3 + ...\infty$$

$$K(x, y) = \exp\left(-\frac{||x||^2}{2\sigma^2}\right)\exp\left(-\frac{||y||^2}{2\sigma^2}\right).\sum_{i=0}^{\infty}\frac{1}{i!}.\left(\frac{x^T y}{2\sigma^2}\right)^i$$

Since, x and y are from $\mathbb{R}^n$ (vectors), we can use multinomial theorem to expand the third term that we got from taylor's expansion. Following is the expression after expansion:

$$\sum_{i=0}^{\infty}\frac{1}{i!}.\left(\frac{x^T y}{2\sigma^2}\right)^i = \sum_{i=0}^{\infty}\frac{1}{i!(2\sigma^2)^i}\sum_{\sum n_j = i}\left(\frac{x_1^{n_1}....x_k^{n_k}}{\sqrt{n_1!...n_k!}}\right)\left(\frac{y_1^{n_1}....y_k^{n_k}}{\sqrt{n_1!...n_k!}}\right)$$

Combining terms , we have the Kernel expression as follows:

$$K(x, y) = \sum_{i=0}^{\infty}\frac{1}{i!(2\sigma^2)^i}\sum_{\sum n_j = i}\exp\left(-\frac{||x||^2}{2\sigma^2}\right)\frac{x_1^{n_1}....x_k^{n_k}}{\sqrt{n_1!...n_k!}}\exp\left(-\frac{||y||^2}{2\sigma^2}\right)\frac{y_1^{n_1}....y_k^{n_k}}{\sqrt{n_1!...n_k!}}$$

Now, we can prove that the given kernel is a mercer kernel, to prove that we can show the integration of kernel with a arbitary square integrable function g is positive. Let us denote that value by I.

$$I = \int_x \int_y K(x, y) g(x) g(y) dx dy$$

$$I = \int_x \int_y \sum_{i=0}^{\infty} \frac{1}{i!(2\sigma^2)^i} \sum_{\sum n_j = i} \exp\left(-\frac{||x||^2}{2\sigma^2}\right) \frac{x_1^{n_1}....x_k^{n_k}}{\sqrt{n_1!...n_k!}} \exp\left(-\frac{||y||^2}{2\sigma^2}\right) \frac{y_1^{n_1}....y_k^{n_k}}{\sqrt{n_1!...n_k!}} g(x) g(y) dx dy$$

Now, we have,

$$I = \sum_{i=0}^{\infty} \frac{1}{i!(2\sigma^2)^i} \int_x \int_y \sum_{\sum n_j = i} \exp\left(-\frac{||x||^2}{2\sigma^2}\right) \frac{x_1^{n_1}....x_k^{n_k}}{\sqrt{n_1!...n_k!}} \exp\left(-\frac{||y||^2}{2\sigma^2}\right) \frac{y_1^{n_1}....y_k^{n_k}}{\sqrt{n_1!...n_k!}} g(x) g(y) dx dy$$

$$I = \sum_{i=0}^{\infty} \frac{1}{i!(2\sigma^2)^i} \sum_{\sum n_j = i} \left(\int_x \exp\left(-\frac{||x||^2}{2\sigma^2}\right) \frac{x_1^{n_1}....x_k^{n_k}}{\sqrt{n_1!...n_k!}} g(x) dx \right) \left(\int_y \exp\left(-\frac{||y||^2}{2\sigma^2}\right) \frac{y_1^{n_1}....y_k^{n_k}}{\sqrt{n_1!...n_k!}} g(y) dy \right)$$

Now, for some parameter z, we have:

$$I_i = \sum_{i=0}^{\infty} \frac{1}{i!(2\sigma^2)^i} \sum_{\sum n_j = i} \left(\int_z \exp\left(-\frac{||z||^2}{2\sigma^2}\right) \frac{z_1^{n_1}....z_k^{n_k}}{\sqrt{n_1!...n_k!}} g(z) \right)^2$$

Since, sum of all integrations for all i are positive, we have overall I greater than equal to 0.

$$I \geq 0$$

Therefore, we have shown that the Kernel $K(x, y)$ is a mercer kernel and now, by Mercer Theorem, $(K(x, y))$ is a valid kernel. Hence proved.

2.2 - Simple Kernel Design

(i) Given, blue points at a distance r_1 from origin and red point at a distance from r_2 from origin, ignoring the noise for convenience. Since, vanilla Kmeans decision boundary is based on the distance between the two cluster centers, it will always result in a linear boundary trying to separate the red and blue data points.

Now, the data points are not linearly separable for any r_1 and r_2 since the cluster separation is possible only in higher dimensions (> 1). Hence, for any condition on r_1 and r_2 or any value, it is not possible to get correct clustering using vanilla KMeans.

(ii) For the given configurations of r_1 and r_2 that are not clusterable, we have to modify the function phi such that it increases the dimension of the separation. One way to do this to consider the distance of points from origin, which will result in a 2-dimensional separation, if there exists one. Hence, we can use the distance of origin or L2 norm of each data point in $\mathbb{R}^2$. Hence, we can define the ϕ function as follows:

$$\phi(x) = ||x||_2$$

Depending on the definition of ϕ, we have our Kernel for the same as:

$$K(x, x') = \phi(x)\phi(x') = ||x||_2||x'||_2$$

Also, for two points, we have a distance,

$$Distance(x, x') = (\phi(x) - \phi(x'))^2 = (||x||_2 - ||x'||_2)^2$$

Now, we can prove K is a valid kernel by showing that it is a Mercer Kernel. For any arbitary square integrable function g we have, the integration following assuming similar naming convention as previous part:

$$I = \int_x \int_{x'} K(x, x')g(x)g(x')dxdx'$$
$$I = \int_x \int_{x'} ||x||_2||x'||_2 g(x)g(x')dxdx'$$

$$I = \left(\int_x ||x||_2 g(x)dx \right) \left(\int_{x'} ||x'||_2 g(x')dx' \right)$$

For a parameter z, we have:

$$I = \left(\int_z ||z||_2 g(z) dz \right)^2 \geq 0$$

$$I \geq 0$$

Since, the integration is greater than equal to 0, the Kernel defined above is a Mercer Kernel and by virtue if Mercer's Theorem, $K(x, x')$ is a valid kernel.

2.2 - Kernel Kmeans Lab

Refer to `assignment_4.ipynb` kernel kmeans section, for the implementation of Kernel Kmeans Class and `make_zero_centered()` function.

Kernel used: $K(x, y) = ||x||_2||y||_2$

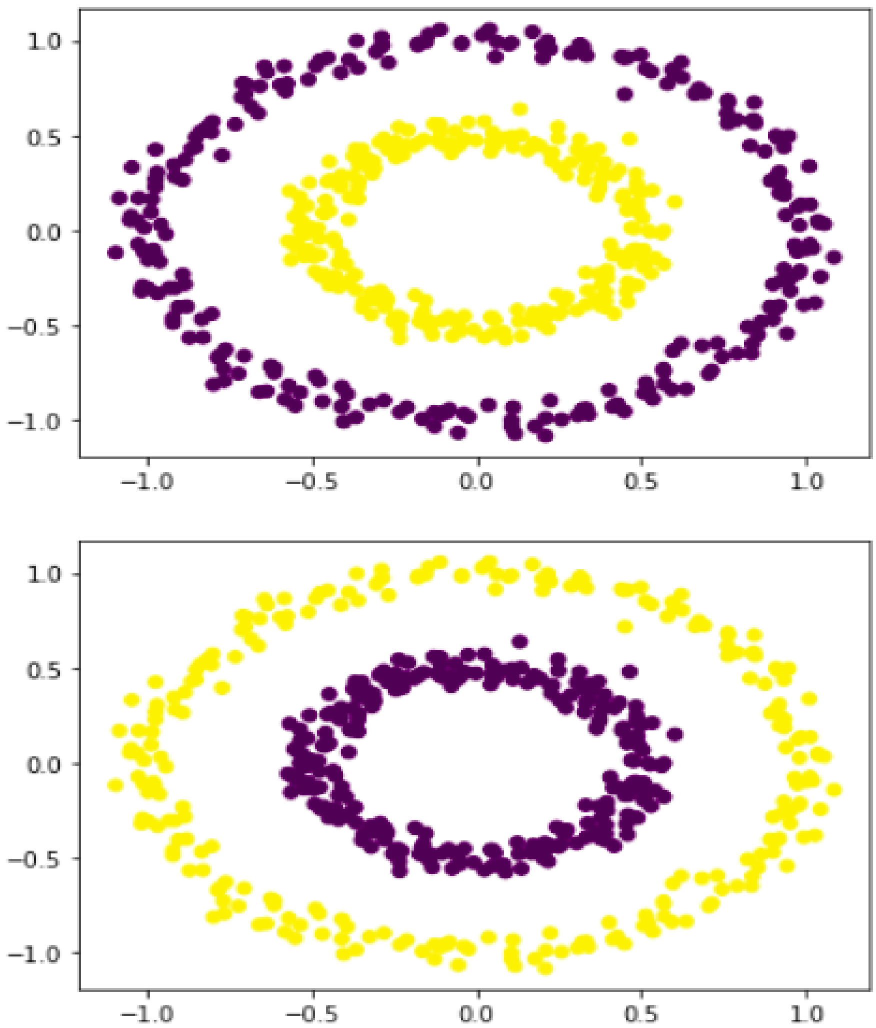

Figure 10: Seed - 123 and Dataset - 3