Instructor:
Prof. Nakul Verma
COMS W4771–Spring, 2022

Student:
Vishweshwar Tyagi
vt2353 @columbia.edu

Homework 4

Due: 11:59pm, April 25, 2022

Homework Problems

Solution-1:

(i) Fix $d \geq 1$. For $y \in \mathbb{R}^d$, y^t will denote the t'th component of y.

We know that for $e_i \in \mathbb{R}^d$, we have, $e_i = I[:, i]$ $\forall 1 \leq i \leq d$ where $I_{d\times d}$ is the identity matrix. Stated in words, e_i is the i'th column of the identity matrix $I_{d\times d}$.

Consider the set of d (distinct) points $S = \{x_j = j * e_j \mid 1 \leq j \leq d\}$. Again, stated in words, $x_j = j * e_j$, is the same vector as e_j except with the element 1 replaced by j.

We claim that $\mathcal{F}$ can shatter S.

Consider any arbitrary labeling of the points.

Let $\alpha_j = j$ if x_j is a positive sample, otherwise let $\alpha_j = 0$. This ensures that $\alpha_j \geq 0$ $\forall 1 \leq j \leq d$. Also note that, each distinct x_j fixes only the j'th coordinate of α, and has no say in fixing any other coordinate of α.

Then, f_α achieves the labeling of these points. To see this, if x_j is a positive sample $\to \alpha_j = j$ and so for $i \neq j$, $x_j^i = 0 \leq \alpha_i$ and $x_j^j = j \leq \alpha_j$. Hence, x_j gets classified as positive. If x_j is negative sample $\to \alpha_j = 0$ and so $x_j^j = j > \alpha_j$. Hence, x_j gets classified as negative.

Thus, we have shown that VC dimension is atleast d.

We now claim that $\mathcal{F}$ cannot shatter any set of $(d+1)$ points.

Consider any set of $(d+1)$ points in $\mathbb{R}^d$. For each of the d axes in $\mathbb{R}^d$, select the point that has the largest value for that axis and label it positive. This way, you can label atmost d points. Label the remaining points as negative.

Suppose $\exists$ $f_\alpha \in \mathcal{F}$ that achieves this labeling.

Consider a negative point y. Because of how we labelled the points, we have, for every $j \in \{1, 2 \ldots d\}$, $\exists$ x a positive point such that $x^j \geq y^j$. Since f_α will classify x as positive, we must have $y^j \leq x^j \leq \alpha_j$. Hence, f_α will also classify y as positive, which is a contradiction.

Hence, no such f_α can exist.

(ii) Axes aligned rectangles in $\mathbb{R}^2$ that label samples in their interior (and boundary) as negative and others as positive have VC dimension 4. Therefore, VC dimension of $\mathcal{F}$ is atleast 4.

For $k > 4$, select any k distinct points on a unit circle centered at $(0, 0)$ and consider any possible labeling of these points. Note that it is possible to select these points on a unit circle such that none of the points overlap/coincide because a circle is made of infinitely many distinct points.

If there is only one negative sample, we can enclose it with an axes aligned rectangle such that all the remaining points lie outside this rectangle. Clearly, this labeling is achieved.

If there are only two negative samples u, v, draw an edge between them $e(u, v)$. If these points are diametrically opposite, i.e, $e(u, v)$ forms a diameter of the circle, then consider two cases: 1. if $(0, \frac{1}{2})$ does not lie on $e(u, v)$, we draw the edges $e(u, (0, \frac{1}{2}))$ and $e(v, (0, \frac{1}{2}))$ 2. if $(0, \frac{1}{2})$ does lie on $e(u, v)$, then we draw edges $e(u, (\frac{1}{2}, 0))$ and $e(v, (\frac{1}{2}, 0))$.

If these points are not diametrically opposite, we draw edges $e(u, (0, 0))$ and $e(v, (0, 0))$.

In all the cases above, we obtain a convex triangle such that negative samples are two of its vertices and the remaining points lie outside this triangle. Clearly, this triangle achieves the labeling.

If there are more than two negative samples, moving in anti-clockwise direction, join the consecutive negative samples by an edge. This will result in a convex polygon having the same number of vertices as there are negative samples as well as all the negative samples are present on the vertices. The remaining samples will lie in the exterior of this convex polygon. This convex polygon will achieve the labeling.

Hence, we have shown that $\mathcal{F}$ can shatter a set of k points arranged on the unit circle for arbitrary $k \geq 1$. Thus, the VC dimension of this class is ∞

Solution-2:

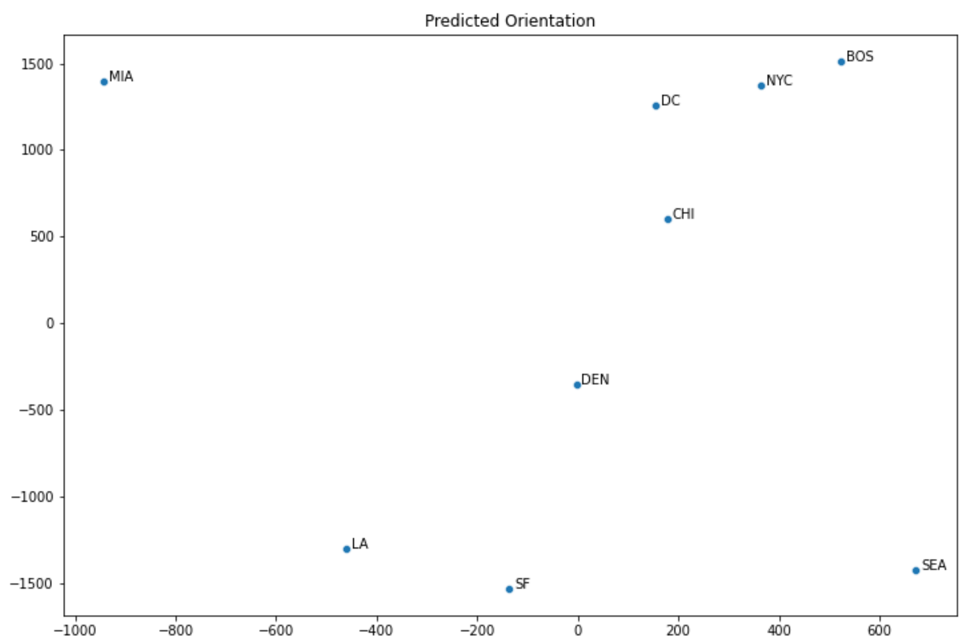
(i) Let $i \in \{1, 2, \dots 9\}$

$$
\begin{aligned}
F(x_i) &= \sum_{u,v} (\|x_u - x_v\| - D_{uv})^2 \\
&= \sum_{v \neq i} (\|x_i - x_v\| - D_{iv})^2 + \sum_{u \neq i} (\|x_u - x_i\| - D_{ui})^2 + \sum_{u \neq i, v \neq i} (\|x_u - x_v\| - D_{uv})^2 \\
&= 2 \sum_{v \neq i} (\|x_i - x_v\| - D_{iv})^2 + \sum_{u \neq i, v \neq i} (\|x_u - x_v\| - D_{uv})^2 \\
\rightarrow \nabla_{x_i} F(x_i) &= 4 \sum_{v \neq i} \frac{\left[\|x_i - x_v\| - D_{iv}\right](x_i - x_v)}{\|x_i - x_v\|}
\end{aligned}
$$

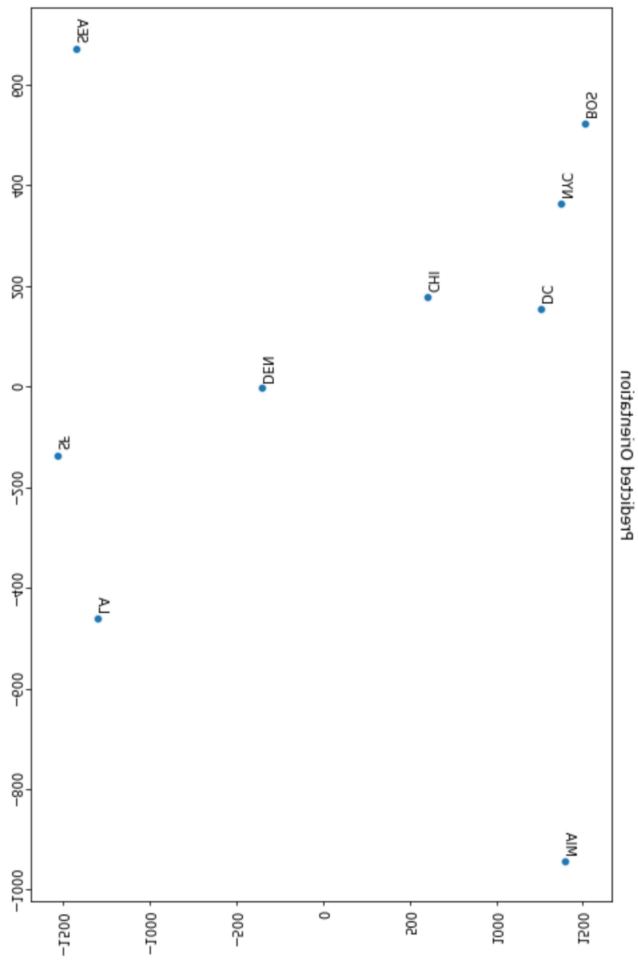
which is well defined because $\|x_i - x_v\| = 0 \iff v = i$

(ii) Submitted.

(iii) The following orientation was obtained using gradient descent -



A 90° anti-clockwise rotation followed by horizontal flip results in orientation similar to the actual orientation of the cities.



Hence, I'm assuming that the result I have obtained can be mapped back to the actual geographical locations using a rotation, flip, and translation.

Solution-3:

(i) Note that

$$\sum_{i=1}^{n} \min_{j \in \{1,2,\dots k\}} \|x_i - c_j\|^2 \geq 0$$

So, the least possible value is 0. For $k = n$ and $c_i = x_i \ \forall i = 1, 2, \dots n$, we have

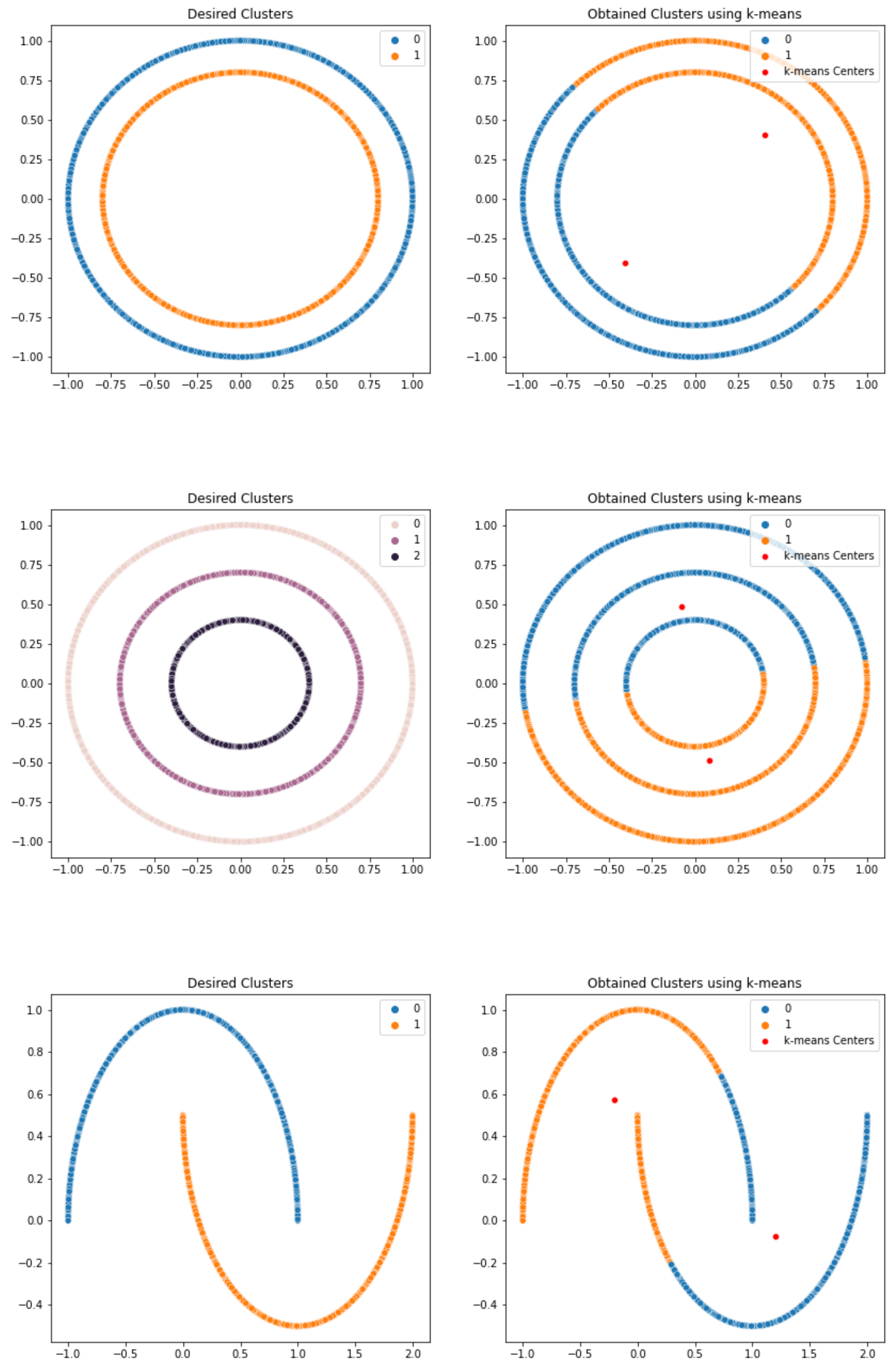
$$\begin{aligned} &\sum_{i=1}^{n} \min_{j \in \{1,2,\dots k\}} \|x_i - c_j\|^2 \\ = &\sum_{i=1}^{n} \|x_i - x_i\|^2 \\ = &\ 0 \end{aligned}$$

Hence, if we allow k to vary, then, the minimum value will be 0 which is achieved when $k = n$ and $c_i = x_i \ \forall i = 1, 2, \dots n$.

This is a bad idea because in this case each data point will be a separate cluster, which defeats the purpose of k-means clustering as it doesn't reveal any pattern in the data.

(ii) Suppose $k = 2, n = 4$ and $x_i = i \ \forall i = 1, 2, 3, 4$. Suppose Lloyd's method initializes cluster centers $c_1 = 1$ and $c_2 = 3$ and finds valid clusters C_i with respect to center c_i as $C_1 = \{1\}$ and $C_2 = \{2, 3, 4\}$. Then, $\sum_{i=1}^{n} \min_{j \in \{1,2,\dots k\}} \|x_i - c_j\|^2 = 2$. Next, the centers are stable because the average of points in each cluster coincides with the cluster center. Hence, the algorithm will halt. However, this is clearly sub-optimal since we have for new centers $c_1' = 1.5$ and $c_2' = 3.5$, and clusters $C_1' = \{1, 2\}$ and $C_2' = \{3, 4\}$, the objective $\sum_{i=1}^{n} \min_{j \in \{1,2,\dots k\}} \|x_i - c_j'\|^2 = 1 < 2$. Thus, Lloyd's method can produce sub-optimal configuration.

(iii) (a) After running k-means, the following results were obtained. The left images display desired clusters, while the right images display the results of k-means -



(b) A simple and intuitive way to see why this is true is to observe that kmeans clusters are always convex sets, hence, they must be linearly separable. More rigorously, for two clusters ($k = 2$), the points on the decision boundary will be equi-distance from both the cluster centers c_1 and c_2, that is,

$$\begin{aligned} \|x - c_1\|_2^2 &= \|x - c_2\|_2^2 \\ \rightarrow \|x\|_2^2 + \|c_1\|_2^2 - 2\langle x, c_1\rangle &= \|x\|_2^2 + \|c_2\|_2^2 - 2\langle x, c_2\rangle \\ \rightarrow 2\langle x, c_1 - c_2\rangle &= \|c_1\|_2^2 - \|c_2\|_2^2 \\ \rightarrow x^T w &= b \end{aligned}$$

where $w = c_1 - c_2$ and $b = \frac{\|c_1\|_2^2 - \|c_2\|_2^2}{2}$, which lie on a hyperplane (linear) in $\mathbb{R}^d$ if $x \in \mathbb{R}^d$. Hence, the cluster boundary is linear.

(c)

$$D - W = \begin{bmatrix} \sum_j w_{1j} - w_{11} & -w_{12} & \dots & -w_{1n} \\ -w_{21} & \sum_j w_{2j} - w_{22} & \dots & -w_{2n} \\ \vdots & \vdots & \dots & \vdots \\ -w_{n1} & -w_{n2} & \dots & \sum_j w_{nj} - w_{nn} \end{bmatrix} \tag{1}$$

$$\rightarrow f^T(D - W) = \left[f_1 \sum_j w_{1j} - \sum_i f_i w_{i1}, \quad f_2 \sum_j w_{2j} - \sum_i f_i w_{i2}, \quad \dots \quad f_n \sum_j w_{nj} - \sum_i f_i w_{in}\right]$$

$$\begin{aligned} \rightarrow f^T(D - W)f^T &= \sum_i f_i^2 \sum_j w_{ij} - \sum_j f_j \sum_i f_i w_{ij} \\ \rightarrow f^T L f^T &= \sum_{ij} f_i^2 w_{ij} - \sum_{ij} f_i f_j w_{ij} \\ &= \frac{1}{2} \cdot \Big(\sum_{ij} f_i^2 w_{ij} + \sum_{ij} f_j^2 w_{ij} - \sum_{ij} 2 f_i f_j w_{ij}\Big) \\ &= \frac{1}{2} \sum_{ij} w_{ij}(f_i^2 + f_j^2 - 2f_{ij}) \\ &= \frac{1}{2} \sum_{ij} w_{ij}(f_i - f_j)^2 \end{aligned} \tag{2}$$

where we used that $\sum_{ij} f_i^2 w_{ij} = \sum_{ij} f_j^2 w_{ij}$

(d) We know that $w_{ij} \in \{0, 1\}\ \forall i, j$. Hence, $w_{ij}(f_i - f_j)^2 \geq 0\ \forall i, j$. Therefore, for arbitrary $f \in \mathbb{R}^m$, we have,

$$f^T L f^T = \frac{1}{2} \sum_{ij} w_{ij}(f_i - f_j)^2 \geq 0$$

Hence, L is positive semi-definite.

Also, $\forall i \neq j$, we have, $L_{ij} = -w_{ij} = -w_{ji} = L_{ji}$, where we used $w_{ij} = w_{ji}$ which follows from definiton of W. Hence, L is also symmetric.

(e) WLOG we can assume that $v_1, v_2 \dots v_k$ belong in disjoint and exhaustive connected components $C_1, C_2 \dots C_k$ respectively. We then have for $1 \leq t \leq k$

$$\begin{aligned}
\mathbf{1}_{C_t} &= \begin{bmatrix} w_{t1} \\ w_{t2} \\ \vdots \\ w_{tn} \end{bmatrix} \\
\rightarrow \mathbf{1}_{C_t}^T L \mathbf{1}_{C_t} &= \frac{1}{2} \sum_{ij} w_{ij} ([\mathbf{1}_{C_t}]_i - [\mathbf{1}_{C_t}]_j)^2 \\
&= \frac{1}{2} \sum_{ij} w_{ij} (w_{ti} - w_{tj})^2 \\
&= \frac{1}{2} \sum_{i \neq j} w_{ij} (w_{ti} - w_{tj})^2 \qquad (3)
\end{aligned}$$

First note that elements of W are either 0 or 1. So, only the following cases are possible:

If $w_{ti} = 0$ and $w_{tj} = 0 \rightarrow w_{ij}(w_{ti} - w_{tj})^2 = 0$
If $w_{ti} = 1$ and $w_{tj} = 1 \rightarrow w_{ij}(w_{ti} - w_{tj})^2 = 0$

Now, consider $w_{ti} = 1$ and $w_{tj} = 0$. If $w_{ij} = 1$, then v_i and v_j are connected, and $w_{ti} = 1 \rightarrow v_t$ and v_i are connected. This implies by transitivity that $w_{tj} = 1$, contradiction. Hence, we must have $w_{ij} = 0 \rightarrow w_{ij}(w_{ti} - w_{tj})^2 = 0$

Again, consider $w_{ti} = 0$ and $w_{tj} = 1$. If $w_{ij} = 1$, then v_i and v_j are connected, and $w_{tj} = 1 \rightarrow v_t$ and v_j are connected. This implies by transitivity that $w_{ti} = 1$, contradiction. Hence, we must have $w_{ij} = 0 \rightarrow w_{ij}(w_{ti} - w_{tj})^2 = 0$

Thus, (3) is a sum of zero terms and using the claim proved on the next page, we have $\mathbf{1}_{C_t}^T L \mathbf{1}_{C_t} = 0 \rightarrow \langle \mathbf{1}_{C_t}, L\mathbf{1}_{C_t} \rangle = 0 \rightarrow L1_{C_t} = 0 \rightarrow \mathbf{1}_{C_t} \in \text{Null}\big(L\big)$ and hence, $\mathbf{1}_{C_t}$ is eigenvector of L corresponding to 0 eigenvalues.

Claim: If $A_{n\times n}(\mathbb{R})$ is symmetric positive semi-definite matrix, then $\langle x, Ax\rangle = 0 \to Ax = 0$

Proof: A is symmetric matrix $\to$ A has n linearly independent and orthonormal eigenvectors $v_i \in \mathbb{R}^n\ \forall i = 1, 2 \ldots n$. Let these eigenvectors correspond to eigenvalues $\lambda_i\ \forall i = 1, 2 \ldots n$ respectively. Since these are n vectors which are linearly independent in n-dimensional space $\mathbb{R}^n$, they form a basis of $\mathbb{R}^n$. Thus, $\exists c_i\ \forall i = 1, 2 \ldots n$ constants such that $x = \sum_i c_i v_i$

Hence,

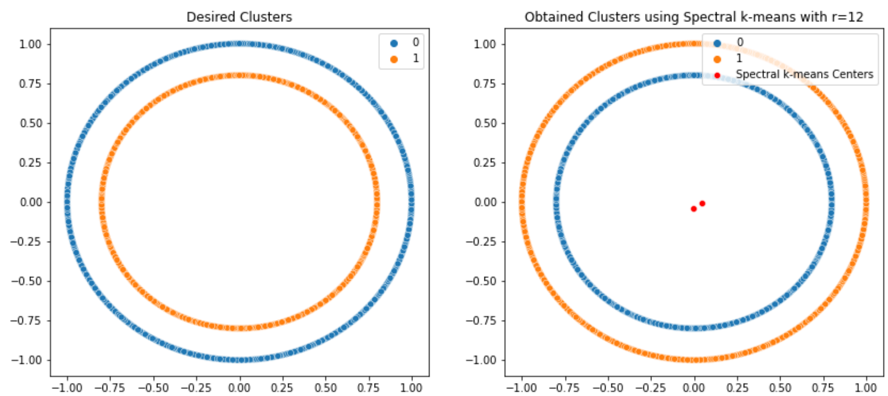
$$
\begin{aligned}
&\langle x, Ax\rangle \\
&= \langle \sum_i c_i v_i, \sum_i c_i A v_i\rangle \\
&= \langle \sum_i c_i v_i, \sum_i \lambda_i c_i v_i\rangle \\
&= \sum_i \sum_j \langle c_i v_i, \lambda_j c_j v_j\rangle \\
&= \sum_i \langle c_i v_i, \lambda_i c_i v_i\rangle \qquad \text{because } \langle v_i, v_j\rangle = 0\ \forall i \neq j \text{ as the eigenvectors are orthogonal} \\
&= \sum_i \lambda_i c_i^2 \|v_i\|^2 \qquad \text{because } v_i \text{ are unit length vectors} \\
&= \sum_i \lambda_i c_i^2
\end{aligned} \tag{4}
$$

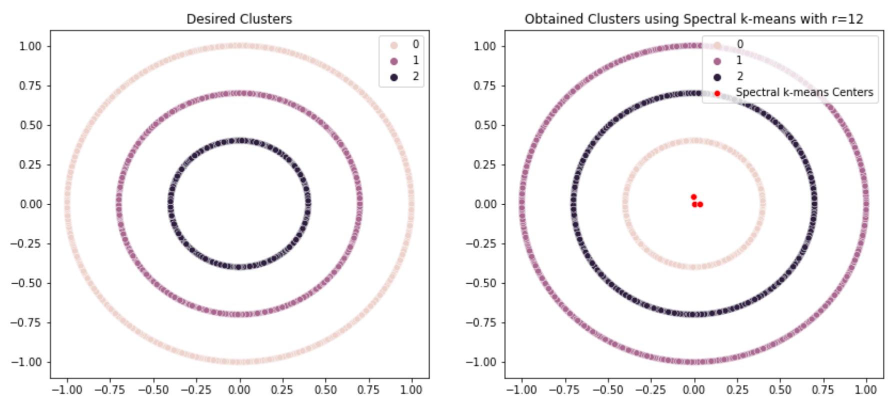
We know that $\lambda_i \geq 0 \to \lambda_i c_i^2 \geq 0$ because L is positive semidefinite.

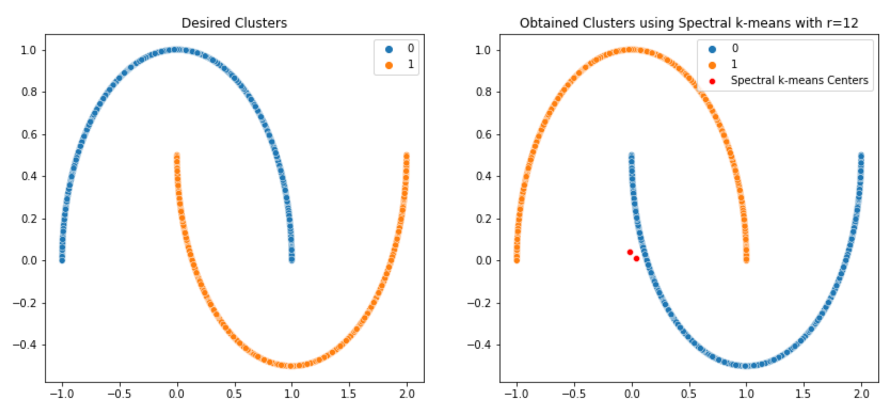
Hence, $\langle x, Ax\rangle = 0 \to \sum_i \lambda_i c_i^2 = 0 \to \lambda_i c_i^2 = 0\ \forall i \to$ if $c_i \neq 0$ then we must have $\lambda_i = 0$

Therefore, $Ax = \sum_i c_i A v_i = \sum_i c_i \lambda_i v_i = \sum_{c_i \neq 0} c_i \lambda_i v_i = 0$

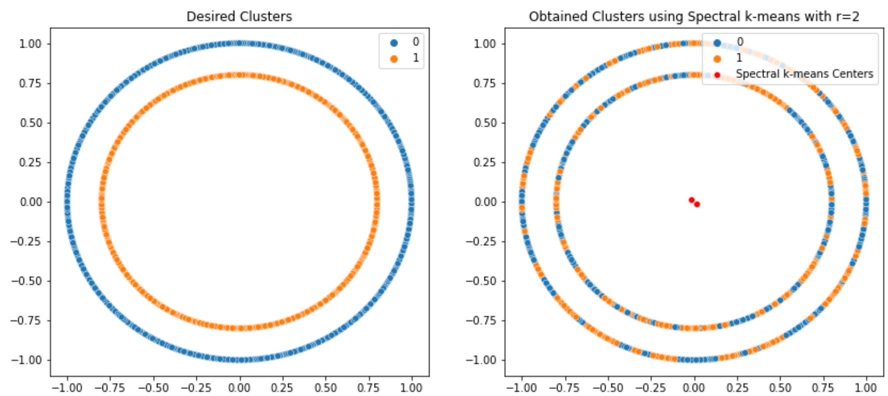
(f) After running the flexible version, I achieved desired clusters with $r = 12$. So, the results improved quite a lot in comparison to vanilla k-means.

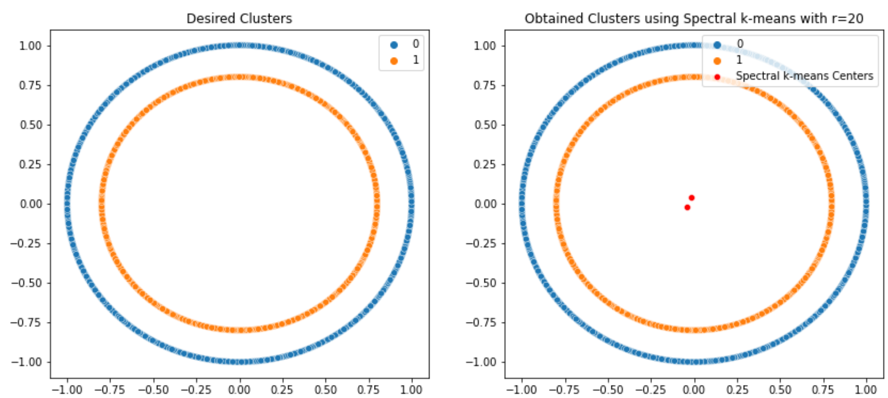


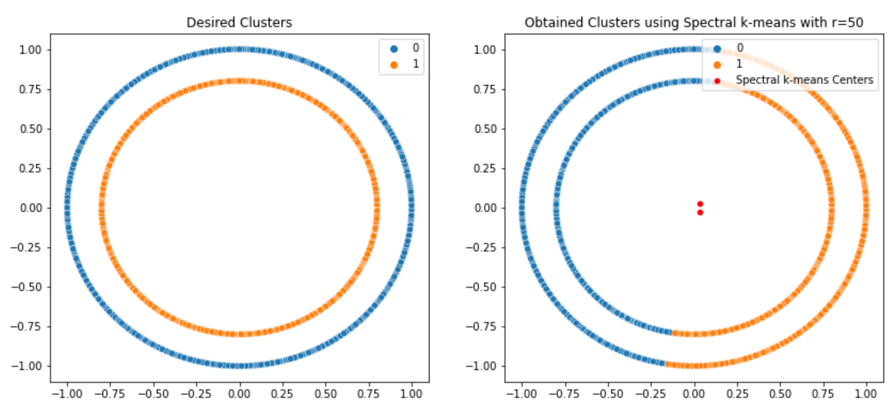


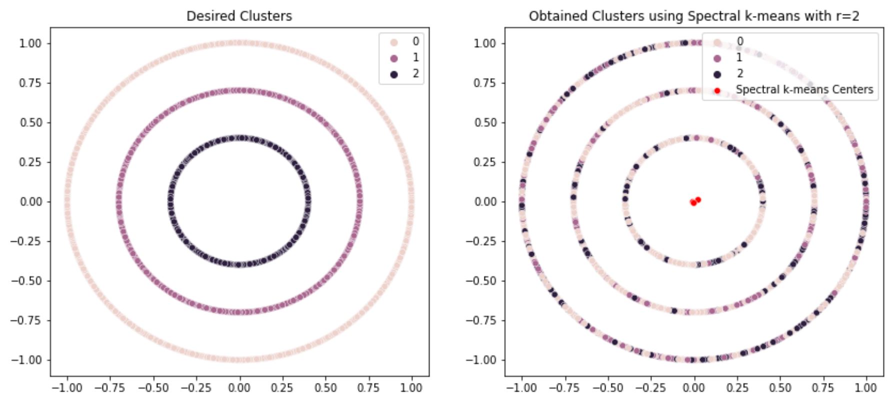


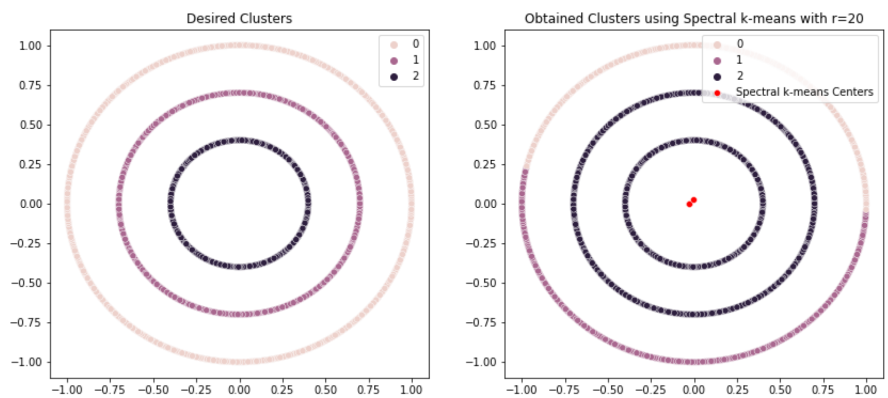
The affect of varying r that I noticed was - Extreme low values gave poor results, then as I increased r, there came a range of values were the results were desirable and after this range, for higher values, the results were similar to vanilla k-means (poor). This is shown below:

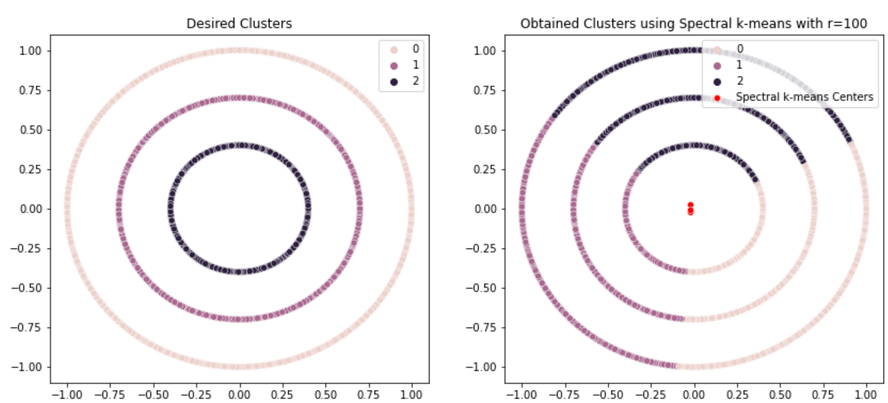


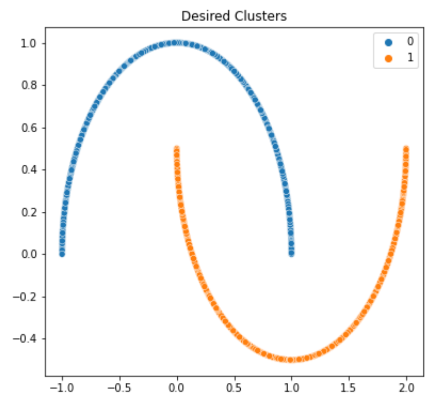


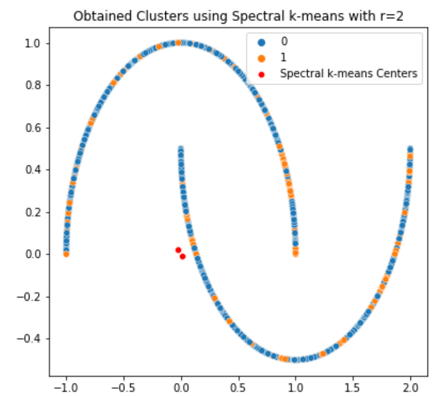


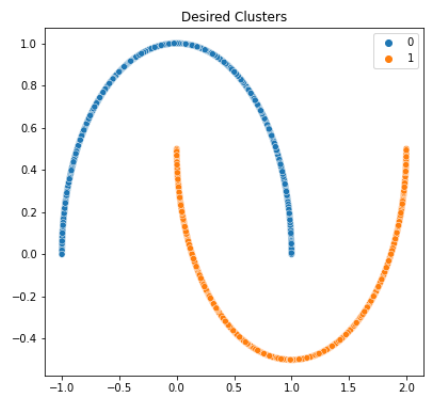

Desired Clusters
0
1
2
Obtained Clusters using Spectral k-means with r=2
0
1
2
Spectral k-means Centers

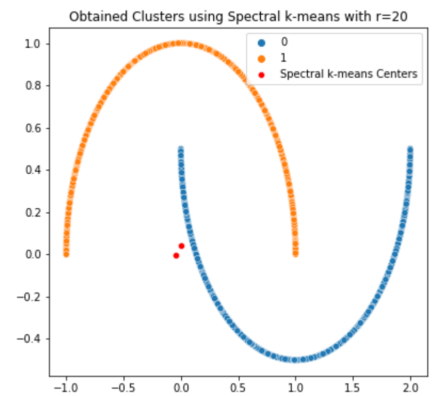

Desired Clusters
0
1
2
Obtained Clusters using Spectral k-means with r=20
0
1
2
Spectral k-means Centers

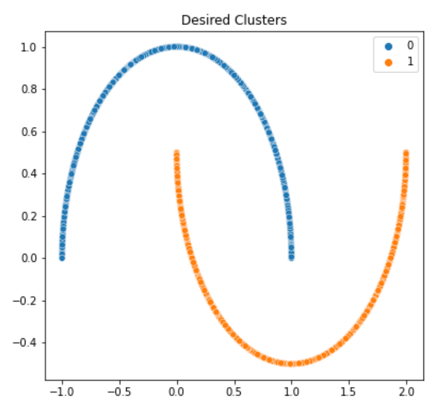

Desired Clusters
0
1
2
Obtained Clusters using Spectral k-means with r=100
0
1
2
Spectral k-means Centers

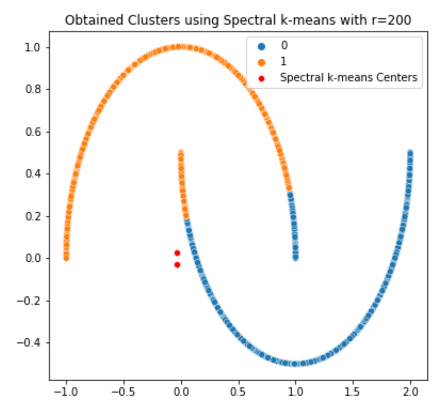

Desired Clusters
0
1


Obtained Clusters using Spectral k-means with r=2
0
1
Spectral k-means Centers


Desired Clusters
0
1


Obtained Clusters using Spectral k-means with r=20
0
1
Spectral k-means Centers


Desired Clusters
0
1


Obtained Clusters using Spectral k-means with r=200
0
1
Spectral k-means Centers