Introduction to ML (CS771), Autumn 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 2

QUESTION 1

Student Name: Havi Bohra
Roll Number: 210429
Date: November 16, 2023

Standard K-means has objective function: $\mathcal{L} = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk}||\mathbf{x}_n - \boldsymbol{\mu_k}||^2$.

(1) In this step, we need to assign x_n greedily to the best cluster ,
we can do it using **ALT-OPT** method:
Fix $\boldsymbol{\mu}$ as $\hat{\boldsymbol{\mu}}$ and find the optimal $\mathbf{Z}$ as

$$\hat{\mathbf{Z}} = \arg\min_{\mathbf{z}} \sum_{k=1}^{K} z_{nk}||\mathbf{x}_n - \hat{\boldsymbol{\mu_k}}||^2$$

(2) In this step, we have to update cluster mean using SGD:
Solving for $\boldsymbol{\mu}$ using with $\mathbf{Z}$ fixed at $\hat{\mathbf{Z}}$

$$\hat{\boldsymbol{\mu}} = \arg\min_{\mu} \mathcal{L}(\mathbf{X}, \hat{\mathbf{Z}}, \boldsymbol{\mu}) = \arg\min_{\mu} \sum_{k=1}^{K} \sum_{n:\hat{z_n}=k} ||\mathbf{x}_n - \boldsymbol{\mu_k}||^2$$

$$\hat{\boldsymbol{\mu_k}} = \arg\min_{\mu_k} \sum_{n:\hat{z_n}=k} ||\mathbf{x}_n - \boldsymbol{\mu_k}||^2$$

Each μ_k can be optimized independently.
At any iteration t, pick an index i$\in \{1, 2, ..., N\}$ uniformly randomly and approximate $\mathbf{g}$ as

$$\mathbf{g} \approx \mathbf{g}_i = \frac{\partial}{\partial \boldsymbol{\mu_k}}(||\mathbf{x}_n - \boldsymbol{\mu_k}||^2) = -2(\mathbf{x}_n - \boldsymbol{\mu_k})$$

Now update mean as: $\boldsymbol{\mu_k}^{(t+1)} = \boldsymbol{\mu_k}^{(t)} - \eta \mathbf{g}^{(t)} \implies \boldsymbol{\mu_k}^{(t+1)} = \boldsymbol{\mu_k}^{(t)} + 2\eta(\mathbf{x}_n - \boldsymbol{\mu_k}^{(t)})$
Intuitvely, we are moving means such that the overall loss (sum of intra-class variances) reduces in each step.

(3) Step size $\eta \propto \frac{1}{N_k}$, where N_k is number of data points in k-th cluster.
As it would make updated mean in the ratio of sum of features of every data point in the cluster to N_k.

Introduction to ML (CS771), Autumn 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Havi Bohra
Roll Number: 210429
Date: November 16, 2023

QUESTION 2

The objective function could be a combination of maximizing the distance between the means of the inputs from the two classes and minimizing the dispersion of inputs within each class.
Fisher's Linear Discriminant Analysis (LDA) Objective Function:

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

Where:

- w is the projection vector.
- S_B is the between-class scatter matrix, measuring the variance between class means.
- $S_W$is the within-class scatter matrix, measuring the variance within each class.

Justification:

- Maximizing the between-class variance S_B helps in achieving a clear separation between the means of the two classes, facilitating better class discrimination. And can be done by involving maximization of the difference between the means of the projected inputs for the two classes in the objective function.
- Minimizing the within-class variance $S_W$helps in reducing the dispersion of data points within each class, making the data points in the same class closer to each other in the projected space. And can be done by involving minimization of the variance or distance of the projected inputs within each class in the objective function.

By optimizing the objective function *J(w)*, one can find the optimal projection direction $\mathbf{w}$, achieving better class separability in the one-dimensional projected space

Introduction to ML (CS771), Autumn 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 2

Student Name: Havi Bohra
Roll Number: 210429
Date: November 16, 2023

QUESTION 3

Let $\mathbf{v} \in \mathbb{R}^n$ be eigenvector of matrix $\frac{1}{N}\mathbf{X}\mathbf{X}^T$, then
$\frac{1}{N}\mathbf{X}\mathbf{X}^T\mathbf{v} = \lambda\mathbf{v}$, for some $\lambda \in \mathbb{R}$
Pre-multiply both side $\mathbf{X}^T$,
$\implies \frac{1}{N}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{v}) = \lambda(\mathbf{X}^T\mathbf{v})$
Substitute $\mathbf{u} = \mathbf{X}^T\mathbf{v}$
$\implies \frac{1}{N}\mathbf{X}^T\mathbf{X}\mathbf{u} = \lambda\mathbf{u}$
$\implies \mathbf{u} = \mathbf{X}^T\mathbf{v}$ is eigenvector of matrix $\mathbf{S} = \frac{1}{N}\mathbf{X}^T\mathbf{X}$,
as $\mathbf{X}$ is N×D and $\mathbf{v}$ is N×1 hence $\mathbf{u}$ is D×1 i.e. $\mathbf{u} \in \mathbb{R}^D$.

This way of obtaining eigenvectors is advantageous as,
if we directly decompose matrix $\frac{1}{N}\mathbf{X}^T\mathbf{X}$ to obtain k eigenvectors it would take $\mathcal{O}(kD^2)$ as the matrix $\frac{1}{N}\mathbf{X}^T\mathbf{X} \in \mathbf{M}_{D^2}$
But by above way it will take $\mathcal{O}(kN^2) + \mathcal{O}(kND)$, {decomposing matrix $\frac{1}{N}\mathbf{X}\mathbf{X}^T$ for k eigenvectors + matrix multiplication $\mathbf{X}^T\mathbf{v}$}
Overall $\mathcal{O}(kND)$ as N<D, which is less than $\mathcal{O}(kD^2)$.

Introduction to ML (CS771), Autumn 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 2

QUESTION 4

Student Name: Havi Bohra
Roll Number: 210429
Date: November 16, 2023

Part 1:
A standard linear model is limited to linear regression, while this alternative model accommodates a combination of K diverse linear curves. Instead of solely fitting one linear curve, this model clusters the data into K distinct linear segments before predicting y values. Furthermore, this approach aids in identifying and addressing outliers within these linear segments, as the clustering process potentially isolates outlier data points from the main cluster.

Part 2:
Our LVM will be:

$$p(z_n = k) = \pi_k$$
$$p(y_n|z_n, \theta) = \mathcal{N}(w_{z_n}^T x_n, \beta^{-1})$$
$$p(y_n, z_n|\theta) = p(y_n|z_n, \theta)p(z_n|\theta)$$
$$p(z_n = k|y_n, \theta) = \frac{p(z_n = k)p(y_n|z_n = k, \theta)}{\sum_{l=1}^{K} p(z_n = l)p(y_n|z_n = l, \theta)} = \frac{\pi_k \mathcal{N}(w_{z_n}^T x_n, \beta^{-1})}{\sum_{l=1}^{K} \pi_l \mathcal{N}(w_l^T x_n, \beta^{-1})}$$

ALT-OPT algorithm, step 1 (find the best z_n:)

$$z_n = \arg\max_{z_n} p(z_n = k|y_n, \theta) = \arg\max_{z_n} \frac{\pi_k exp(-\frac{\beta}{2}(y_n - w_{z_n}^T x_n)^2)}{\sum_{l=1}^{K} \pi_l exp(-\frac{\beta}{2}(y_n - w_l^T x_n)^2)}$$

Step 2 re-estimate the parameters:
N_k denotes no. of elements in cluster k. X_k are $N_K \times D$ matrix containing training points and y_k are $N_k \times 1$ vectors containing training point labels, which are clustered in class k.

$$w_k = (X_k^T X_k)^{-1} X_k^T y_k$$
$$\pi_k = N_k/N$$

If $\pi_k = \frac{1}{K}$ then:

$$z_n = \arg\max_{z_n} \frac{exp(-\frac{\beta}{2}(y_n - w_{z_n}^T x_n)^2)}{\sum_{l=1}^{K} exp(-\frac{\beta}{2}(y_n - w_l^T x_n)^2)}$$

This update is similar to that of multi-class logistic regression.

Introduction to ML (CS771), Autumn 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 2

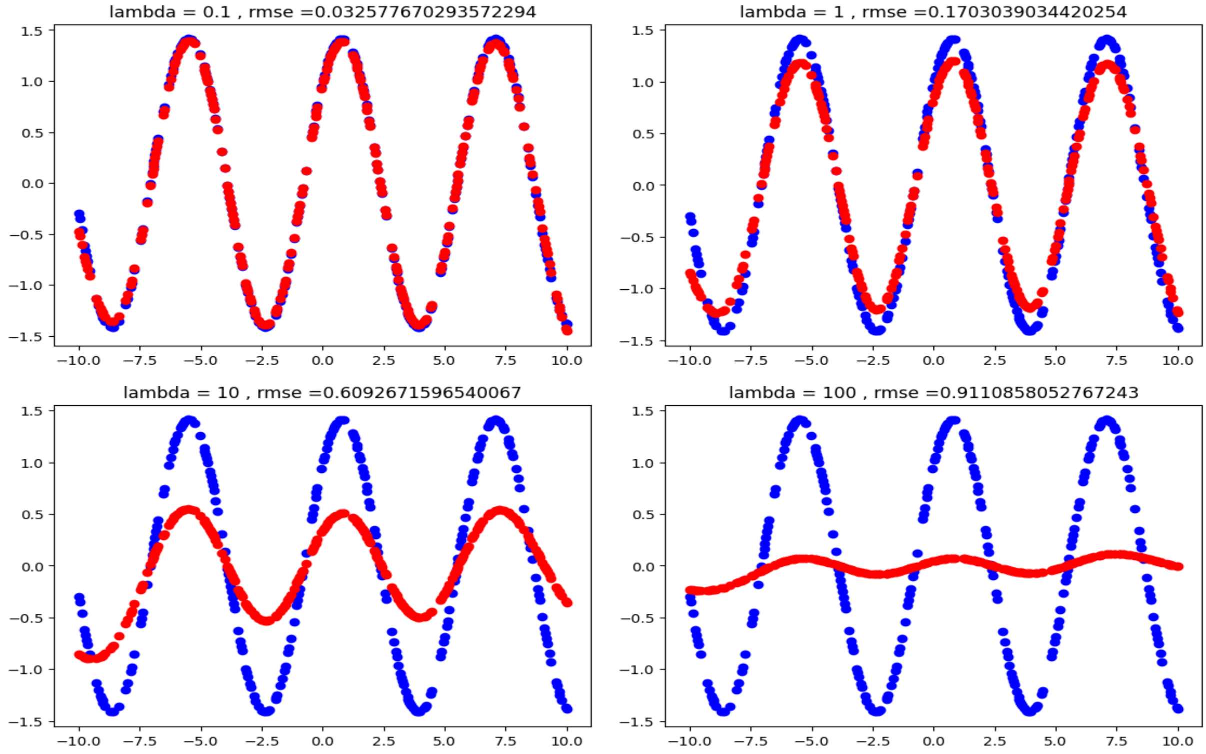
QUESTION 5

Student Name: Havi Bohra
Roll Number: 210429
Date: November 16, 2023

Part 1:

(1) **Kernel Ridge:**

λ	RMSE
0.1	0.032577670293572294
1	0.1703039034420254
10	0.6092671596540067
100	0.9110858052767243


lambda = 0.1 , rmse =0.032577670293572294

lambda = 1 , rmse =0.1703039034420254

lambda = 10 , rmse =0.6092671596540067

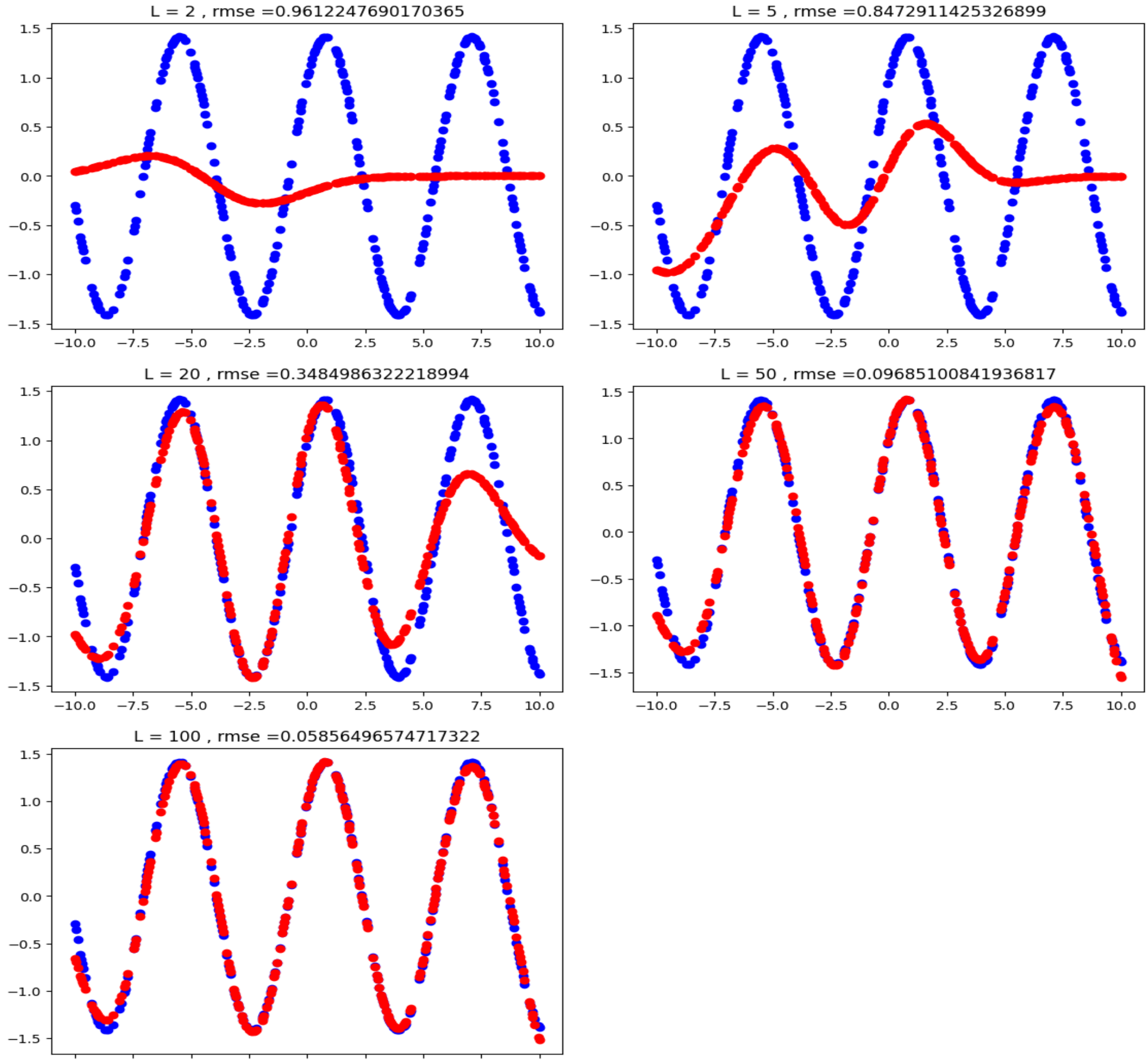
lambda = 100 , rmse =0.9110858052767243

Actual- Blue, Predicted- Red.

Data sample looks like sampled from a sine wave, hence kernel function is required. Also from figures, we can see smaller λ performs better .

(2) **Landmark Ridge:**(λ =0.1)

L	RMSE
2	0.9612247690170365
5	0.8472911425326899
20	0.3484986322218994
50	0.09685100841936817
100	0.05856496574717322

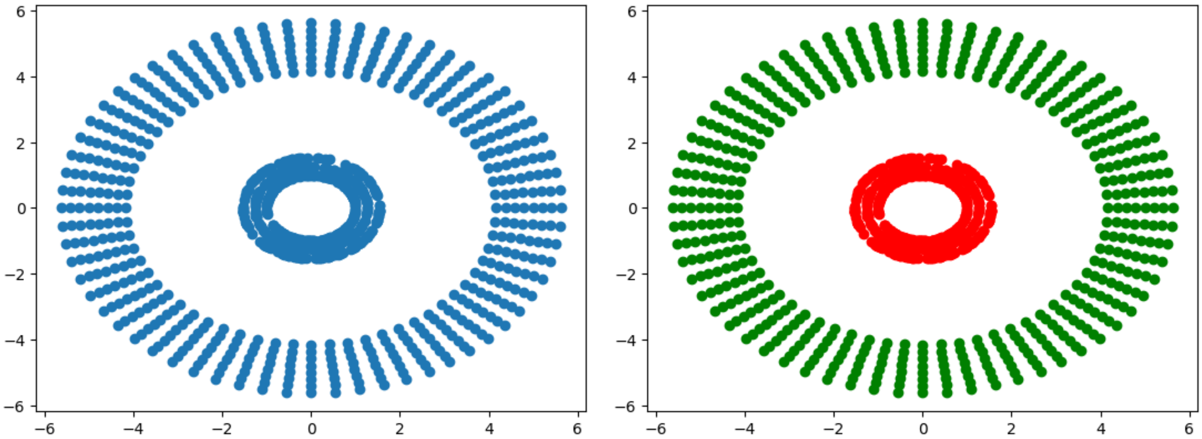


Actual- Blue, Predicted- Red.

As no. of landmarks **L** increases, prediction improves.

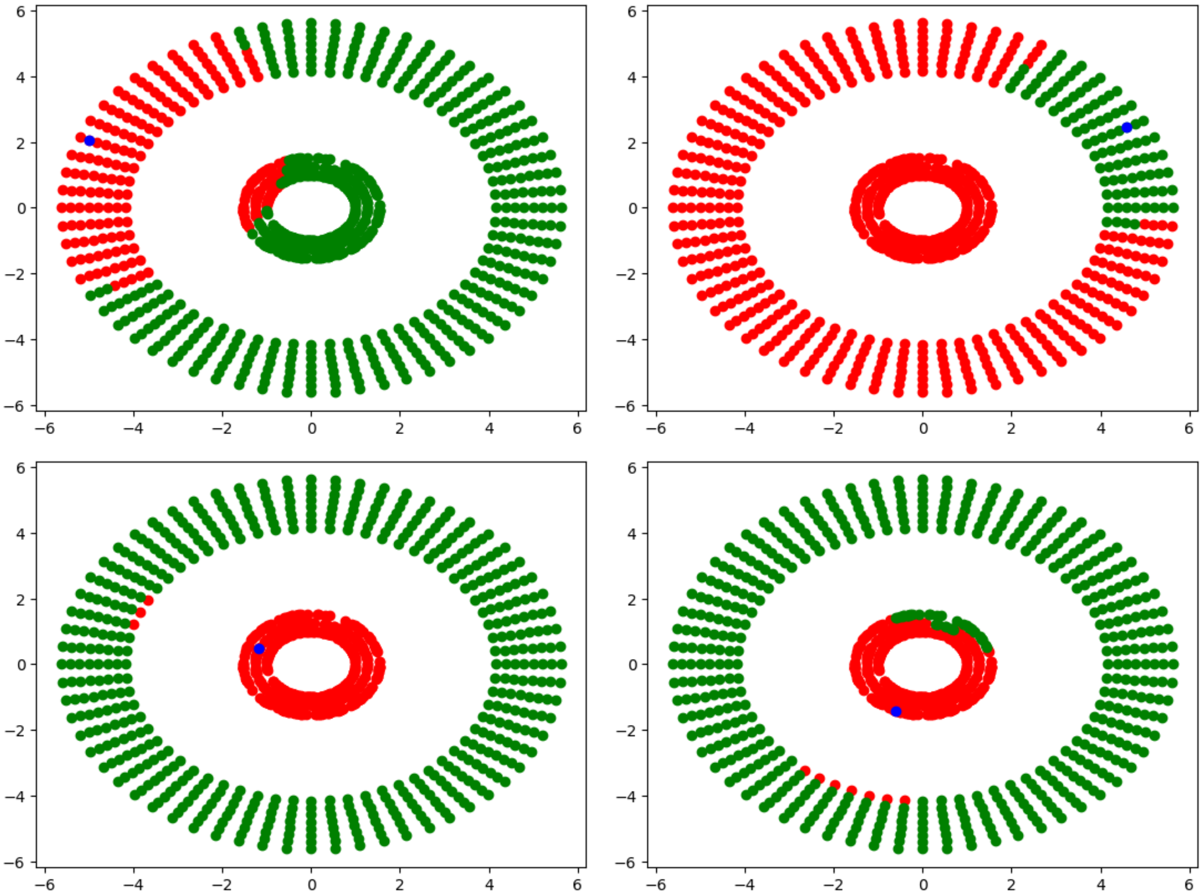
Part 2:

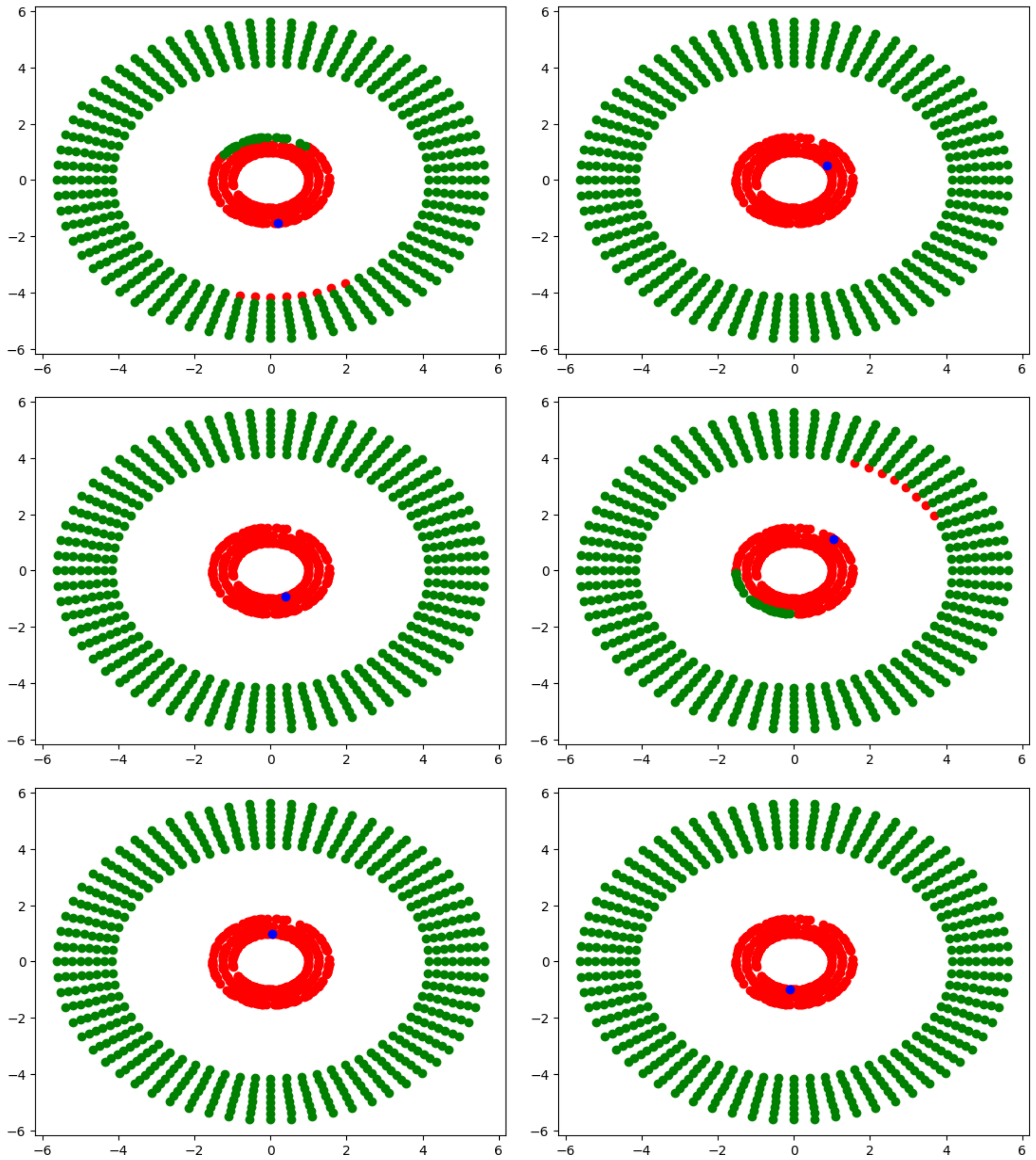
(1) **Original data + Using Hand-crafted Features :**



Original data (image on the left) look-like two co-centric annular regions, which suggests clustering based on distance from origin, $(x_n, y_n) \to x_n^2 + y_n^2$.
Applying above transformation and using standard k-means, we got clustering (image on the right), which works absolutely perfect.

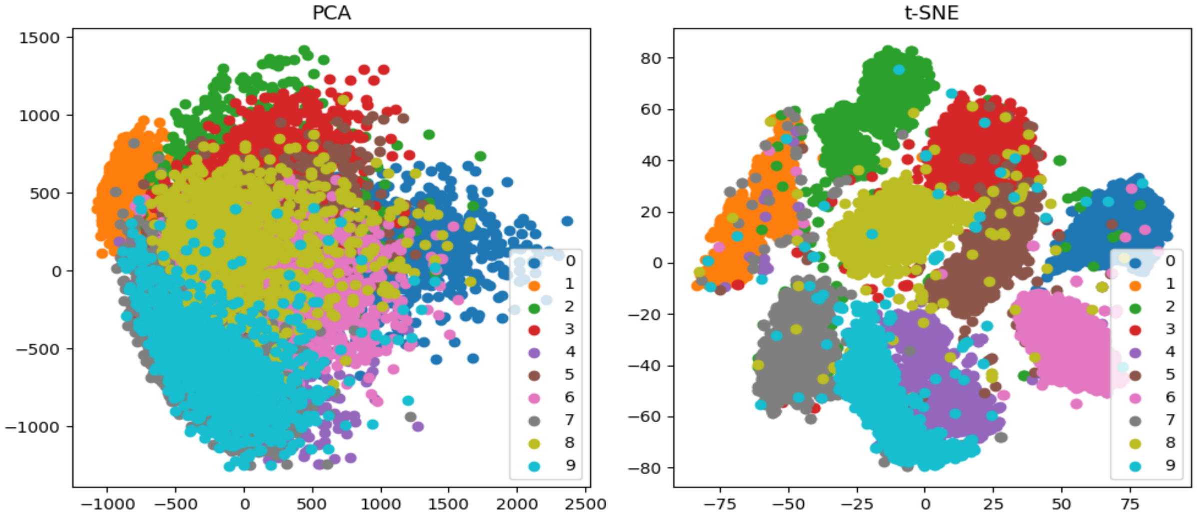
(2) **Using Kernels:**





Easy to see that, when the random landmark is inside the smaller annulus, and the more inside it is, the better clustering we get. This is because the RBF kernel in some sense inverts the distance, so points that are far apart are brought close in the mapping. If the random landmark is outside the inner annulus, points in the inner annulus stay are close to 1 in the mapping space, and the points in the outer annulus are close to zero. However, if the landmark is in the outer annulus, points in both the annulus can come close together.

Part 3:



The t-SNE plot have more well defined clusters and the PCA plot have clusters in overlapping to each other. Hence, t-SNE is more better way to get a low-dim representation of the provided dataset.