Introduction to ML (CS771), Autumn 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Anshu Kumar Rajkumar
Roll Number: 210155
Date: November 17, 2023

QUESTION
1

Solution to problem 1

Step 1: Determine the x_n best cluster

$$z_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j ||x_n - \mu_j||^2 \\ 0 & \text{otherwise} \end{cases}$$

z_{nk} is set to 0 if k is not a minimal space group.

Step 2: Updating Cluster Means by SGD

The objective K-means function L can be written as:

$$L = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} ||x_n - \mu_k||^2$$

Using Stochastic Gradient Descent (SGD) to update the ensemble means that for each data point x_n and μ_k we must find the gradient of μ_k respectively.) each other. The gradient for a single data point x_n:

$$\frac{\partial l_n}{\partial \mu_k} = -2 z_{nk} (x_n - \mu_k)$$

Update equation for μ_k using SGD is then:

$$\mu_k^{(t+1)} = \mu_k^{(t)} + 2\eta z_{nk} (x_n - \mu_k^{(t)})$$

where t is the number of iteration, η is the step size, and x_n is randomly selected for the SGD update.

Step Size η

A suitable step size (η) for a convergent algorithm can be:

$$\eta_t = \frac{1}{t}$$

This option provides a step size that allows the algorithm to converge better, since SGD is a random type. As we approach the optimum, we will find that we are not oscillating, knowing in theory that our choice of step size can go anywhere.

Introduction to ML (CS771), Autumn 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Anshu Kumar Rajkumar
Roll Number: 210155
Date: November 17, 2023

QUESTION
2

Solution to problem 2

An Ideal Projection

The training data labeled $\{(x_n, y_n)\}$ R^D is given by $x_n and y_n$ $\{-1 + 1\}$. Using the vector R^D, project inputs in one dimension, project the distance between input means from two classes as close as possible, and inputs in each class as close as possible.

Objective Function:

The objective/loss function to achieve this is given by:

$$L(w) = \frac{\sum_{i=1}^{N_+} \|w^T(x_i^+ - \mu_+)\|^2 + \sum_{i=1}^{N_-} \|w^T(x_i^- - \mu_-)\|^2}{\|w^T\mu_+ - w^T\mu_-\|^2}$$

subject to the constraint:

$$\|w\|^2 = 1$$

where:

- μ_+ is the mean of the inputs with label $+1$,
- μ_- is the mean of the inputs with label -1,
- x_i^+ is the i-th sample in the positive class,
- x_i^- is the i-th sample in the negative class,
- $\|\cdot\|^2$ denotes the squared Euclidean norm,
- The constraint $\|w\|_2^2 = 1$ ensures that the projection vector w has a unit Euclidean norm.

Justification:

1. **Maximizing Distance Between Class Means:** The denominator $\|w^T\mu_+ - w^T\mu_-\|^2$. This is to ensure that there are as many elements of the positive and negative classes as possible (the class difference is quite high).
2. **Minimizing Within-Class Variance:** The index contains the sum of the squared distances from each sample to the class after projection.

Introduction to ML (CS771), Autumn 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Anshu Kumar Rajkumar
Roll Number: 210155
Date: November 17, 2023

QUESTION
3

Solution to problem 3

Eigenchangers

Given:

- X is an $N \times D$ matrix.
- Assume $D > N$.
- Conventional PCA involves computing the eigenvectors of the covariance matrix $S = \frac{1}{N}X^TX$ (assuming the data are centered).

Problem Statement:

For others to show that the matrix $\frac{1}{N}XX^T$ gives the eigenvector $v \in \mathbb{R}^N$, you can use it to get the eigenvector $u \in \mathbb{R}^D S = \frac{1}{N}X^TX$.

Solution:

Let $v \in \mathbb{R}^N$ be the eigenvector of the matrix $\frac{1}{N}XX^T$ with corresponding eigenvalue λ. We want to find an eigenvector $u \in \mathbb{R}^D$ of the covariance matrix $S = \frac{1}{N}X^TX$. Claim: $u = X^Tv$ is an eigenvector of the covariance matrix $S = \frac{1}{N}X^TX$ with same eigen-value λ.

Poof:

$$Su = \frac{1}{N}X^TXu$$

Substitute $u = X^Tv$:

$$Su = \frac{1}{N}X^TX(X^Tv)$$

$$\implies Su = X^T\frac{1}{N}(XX^T)v$$

$$\implies Su = X^T(\lambda v)$$

$$\implies Su = \lambda X^Tv$$

$$\implies Su = \lambda u$$

Advantage:

Calculating the eigenvectors of $\frac{1}{N}XX^T$ can be cheaper than calculating the eigenvectors of S, especially if D is larger than N. This approach makes it possible to reduce the dimensionality of the calculation.

Introduction to ML (CS771), Autumn 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Anshu Kumar Rajkumar
Roll Number: 210155
Date: November 17, 2023

QUESTION 4

Solution to problem 4

Ordinary probability linear models can be used when we need to fit a linear model to data that is truly linear. This model will not work well if the data is more than a single line curve (or as line segments in separate clusters). The latent variable model models the data as a combination of K different linear curves, that is, divides the data into K different groups and then estimates y based on the group's linear curve.

Show Latent Hidden model as follows:

$$\mathbb{P}(z_n = k \mid y_n, \theta) = \frac{\mathbb{P}(z_n = k) \cdot p(y_n \mid z_n = k, \theta)}{\sum_{j=1}^{K} \mathbb{P}(z_n = j) \cdot p(y_n \mid z_n = j, \theta)}$$

where

$$\mathbb{P}(z_n = k) = \pi_k$$
$$p(y_n \mid z_n = k, \theta) = \mathcal{N}(w_k^T x_n, \beta^{-1})$$

ALT-OPT Algorithm

Step 1 : Finding optimal z_n for x_n based on current marginals

$$z_n = \arg\max_k \frac{\pi_k \mathcal{N}(w_k^T x_n, \beta^{-1})}{\sum_{m=1}^{K} \pi_m \mathcal{N}(w_m^T x_n, \beta^{-1})}$$

$$\implies z_n = \arg\max_k \frac{\pi_k \exp\left(-\frac{\beta}{2}(y_n - w_k^T x_n)^2\right)}{\sum_{m=1}^{K} \pi_m \exp\left(-\frac{\beta}{2}(y_n - w_m^T x_n)^2\right)}$$

Step 2 : Re-estimating marginals and class weight w_k after clustering update

$$N_k = \sum_{n=1}^{N} z_{nk}$$
$$w_k = (X_k^T X_k)^{-1} X_k^T y_k$$
$$\pi_k = \frac{N_k}{N}$$

We first determine the cluster based on Bayes' rule using the current class and state, then recalculate the class limit and repeat the process until convergence. **If $\pi_k = \frac{1}{K}$:**

$$z_n = \arg\max_k \frac{\exp\left(-\frac{\beta}{2}(y_n - w_k^T x_n)^2\right)}{\sum_{m=1}^{K} \pi_m \exp\left(-\frac{\beta}{2}(y_n - w_m^T x_n)^2\right)}$$

Equivalent to multi class logistic regression.

Introduction to ML (CS771), Autumn 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Anshu Kumar Rajkumar
Roll Number: 210155
Date: November 17, 2023

QUESTION 5

My solution to problem 5

5.1.1 : Figure 1
Root Mean Squared Error (RMSE) for lambda = 0.1 : 0.03257767029357561
Root Mean Squared Error (RMSE) for lambda = 1 : 0.17030390344202548
Root Mean Squared Error (RMSE) for lambda = 10 : 0.6092671596540066
Root Mean Squared Error (RMSE) for lambda = 100 : 0.9110858052767243

Overfitting with increasing lambda values leads to overfitting of the model by fully exploiting the train data features and increasing the RMSE.

5.1.2 : Figure 2
Landmarks: 2, RMSE: 0.9747234557672884
Landmarks: 5, RMSE: 0.9171537975413196
Landmarks: 20, RMSE: 0.13611014771281066
Landmarks: 50, RMSE: 0.08852463717613687
Landmarks: 100, RMSE: 0.0582412809194834

As the number of markers increases, the feature representation becomes more dimensional and richer, leading to a more complex model capable of identifying nonlinearities in the data. 50 marks seems good enough.

5.2.1 : Figure 4

The proposition must be changed from XY coordinates to R - θ because the group cannot be divided linearly, but resembles a radical group.

5.2.2 : Figures 5 to 14 : Landmark Shown by blue cross

The selected point will be the starting point for clustering, since the behavior of the RBF kernel depends only on the distance between two points, a radial function. Therefore, a randomly selected Landmark will only work well if the randomly selected key data points have coordinates close to the center (ie 0,0).

5.3 : Figures 15 and 16

t-SNE works well because if two local points are neighbors in the original space, they should be local neighbors in the predicted space, maintaining the interclass space and class separation. In linear PCA, the separation between classes can be lost, so the PCA kernel will reduce the loss of separation, but not as much as t-SNE.

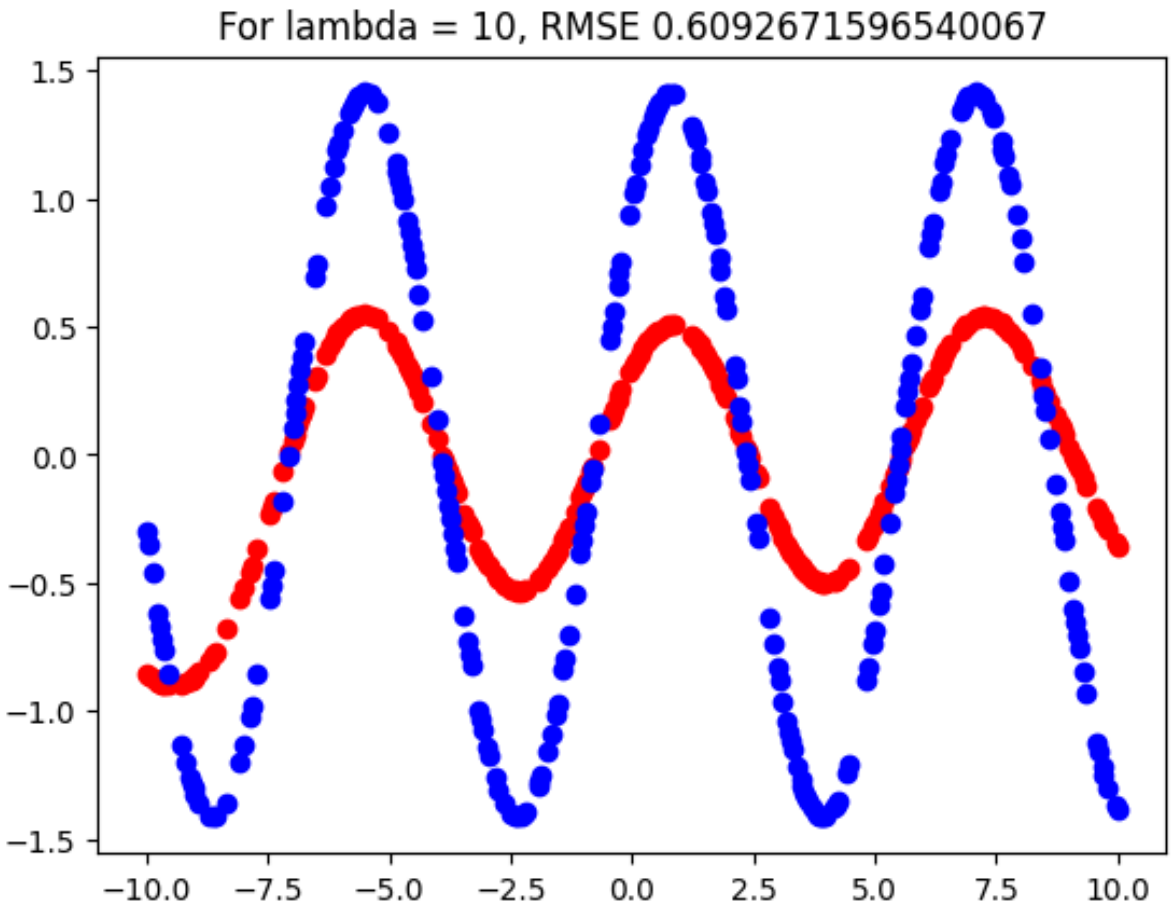

Figure 1: Kernel Regression with different values of λ

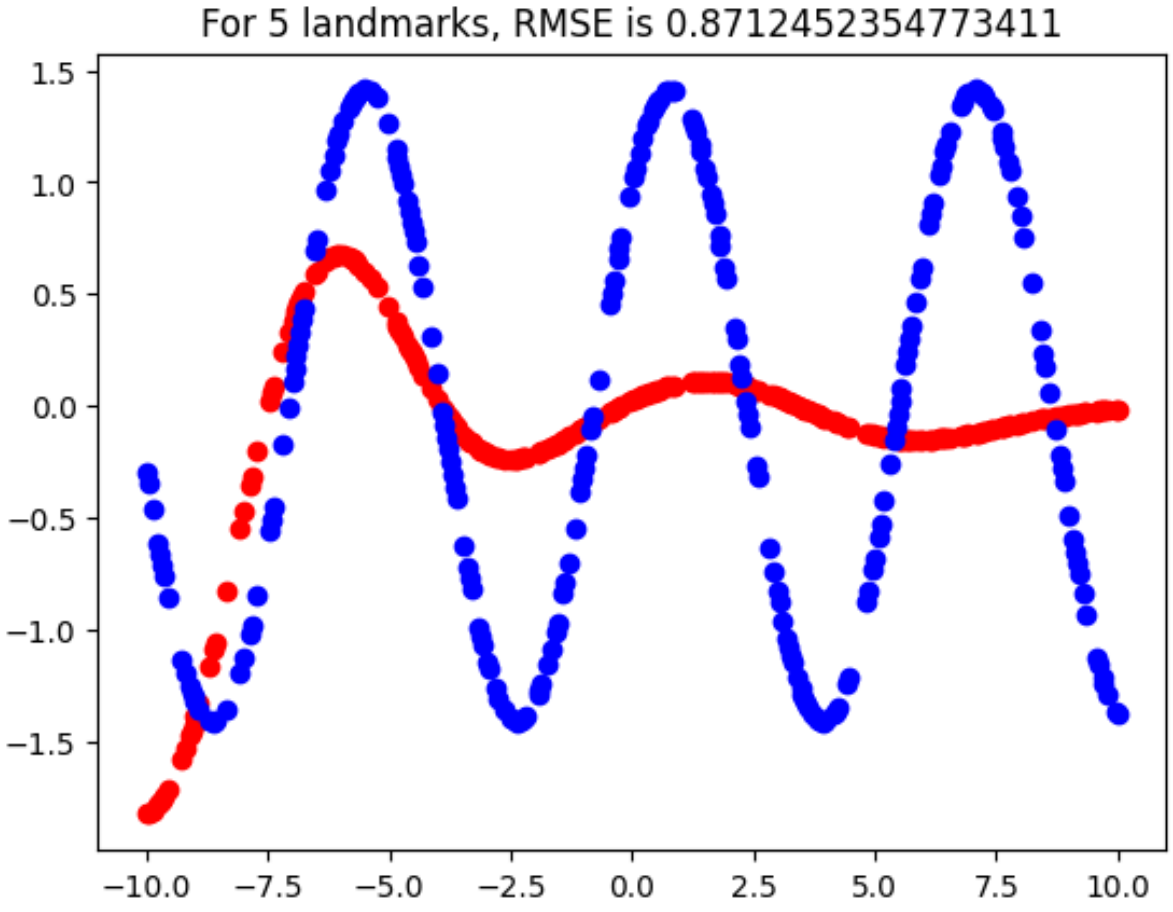

Figure 2: Landmark Regression with different number of landmarks

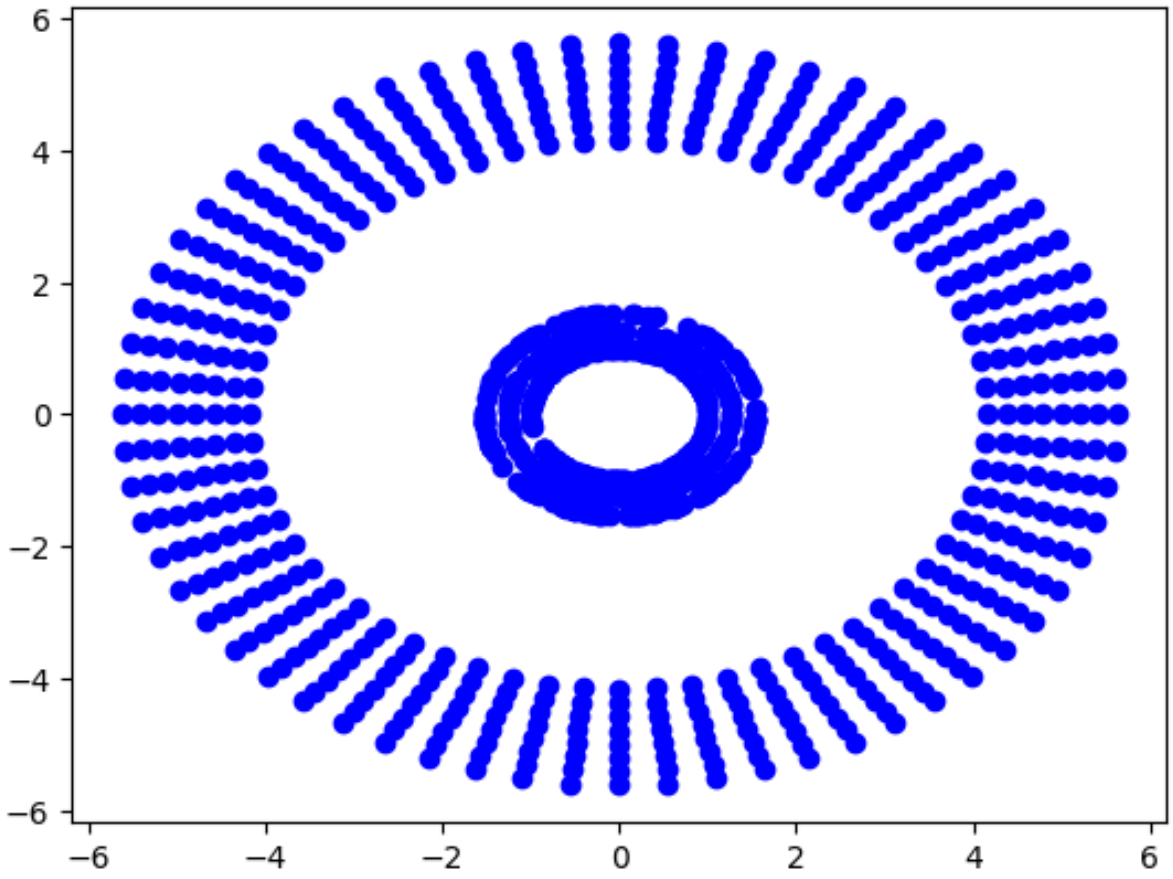

Figure 3: Original Data for Part 2

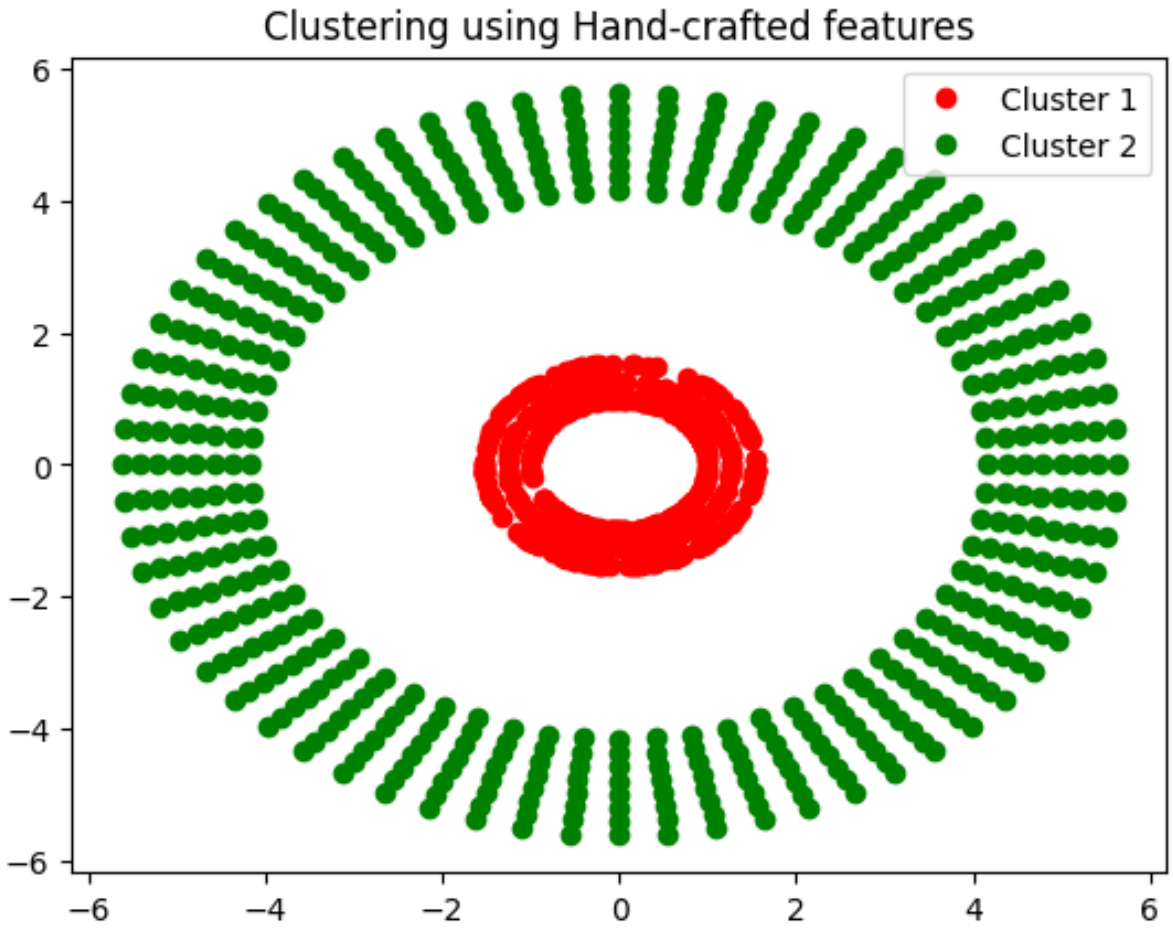

Figure 4: Handcrafted Clustering on the basis of radius(domain transform to polar)

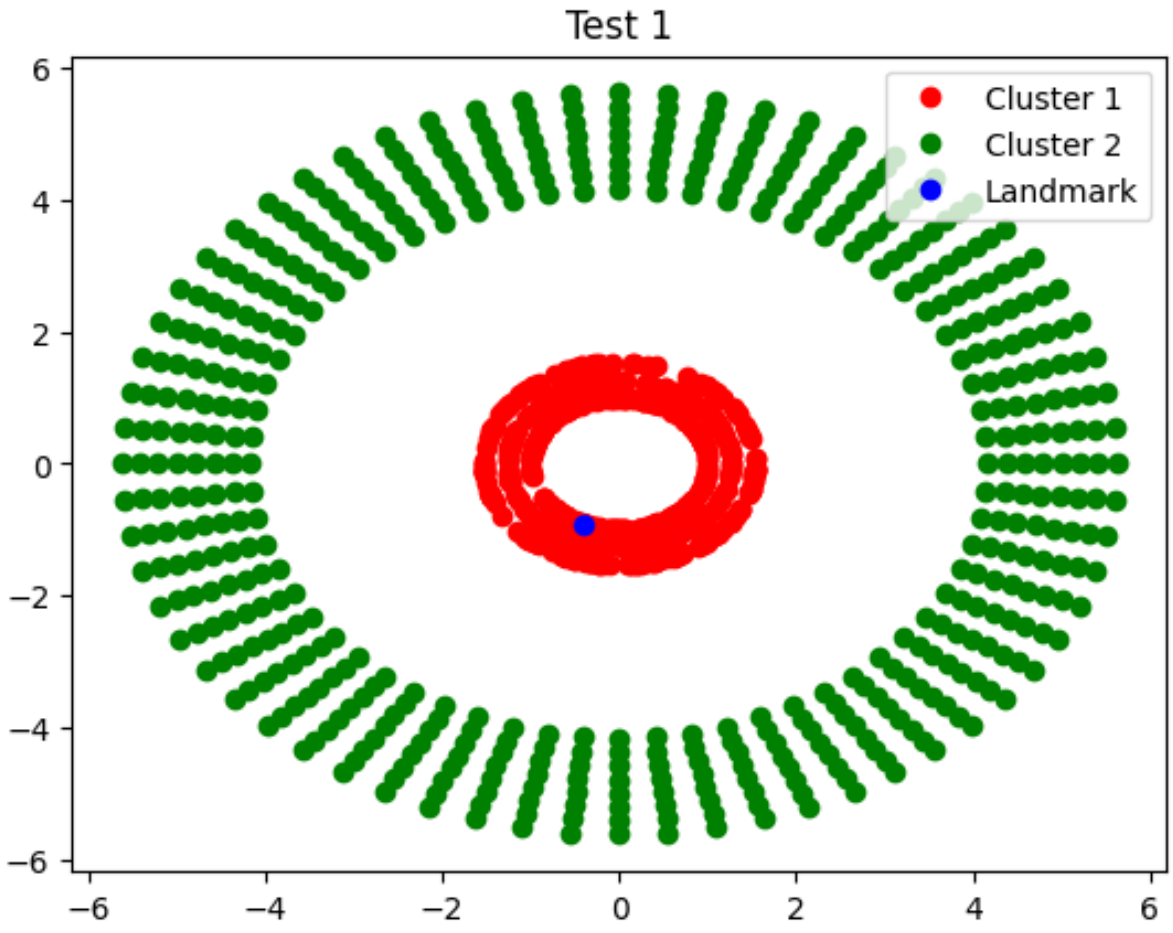

Figure 5: Single Landmark : Run 1

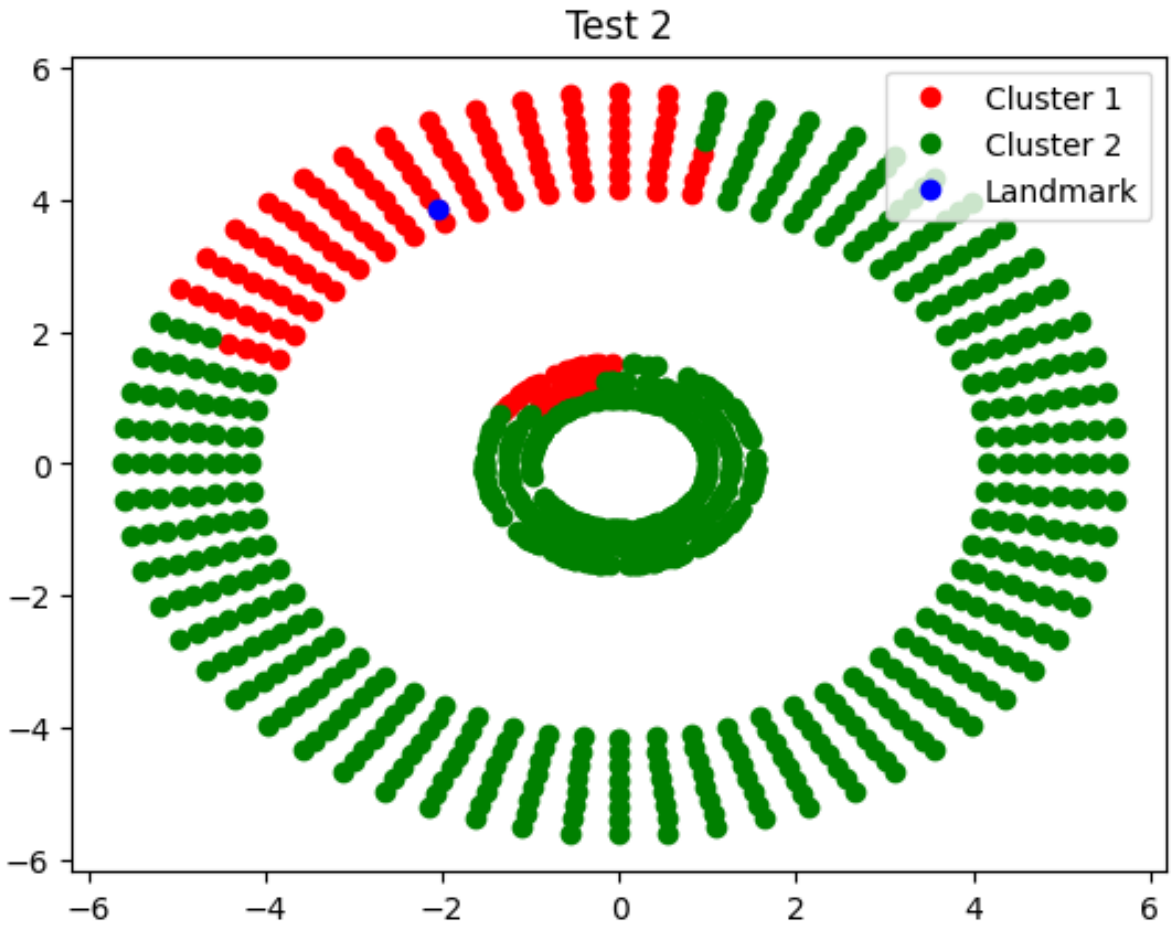

Figure 6: Single Landmark : Run 2

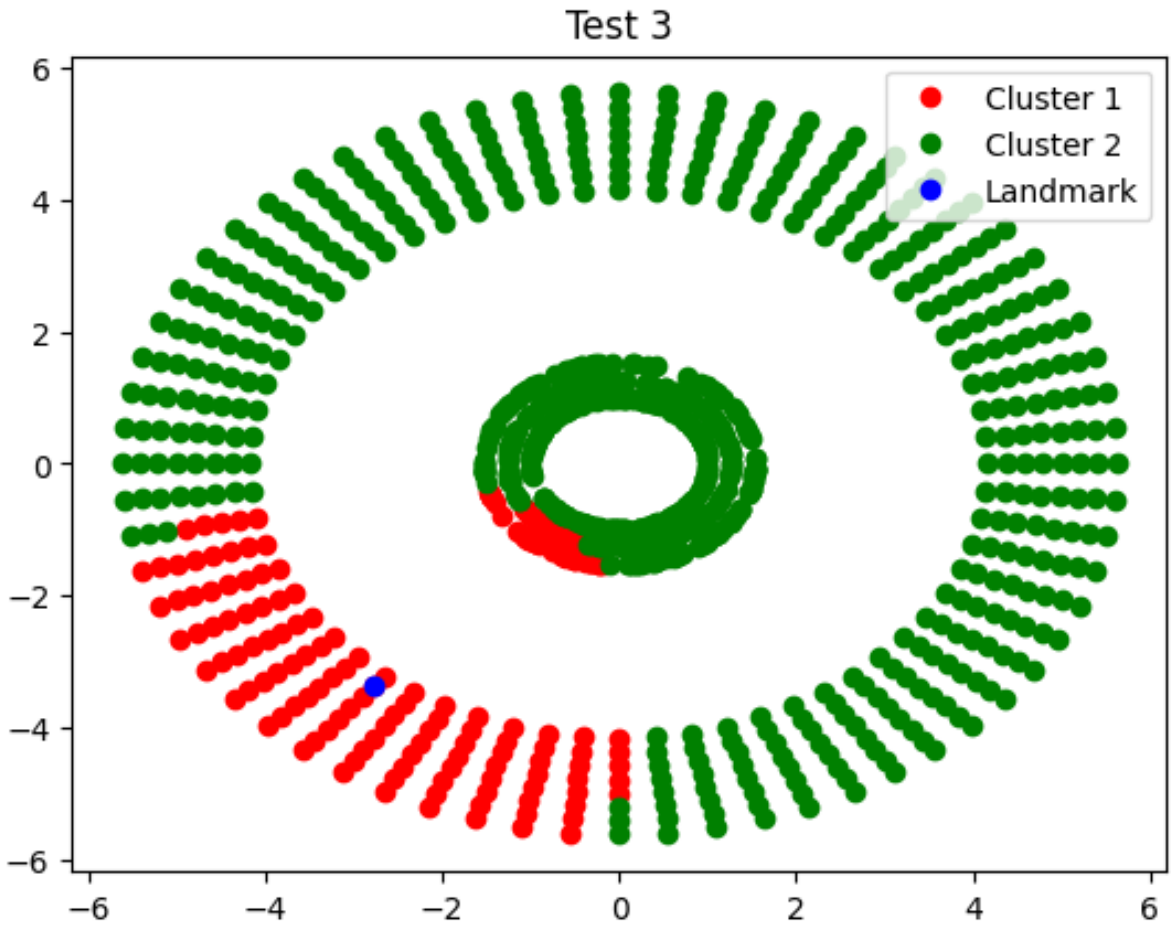

Figure 7: Single Landmark : Run 3

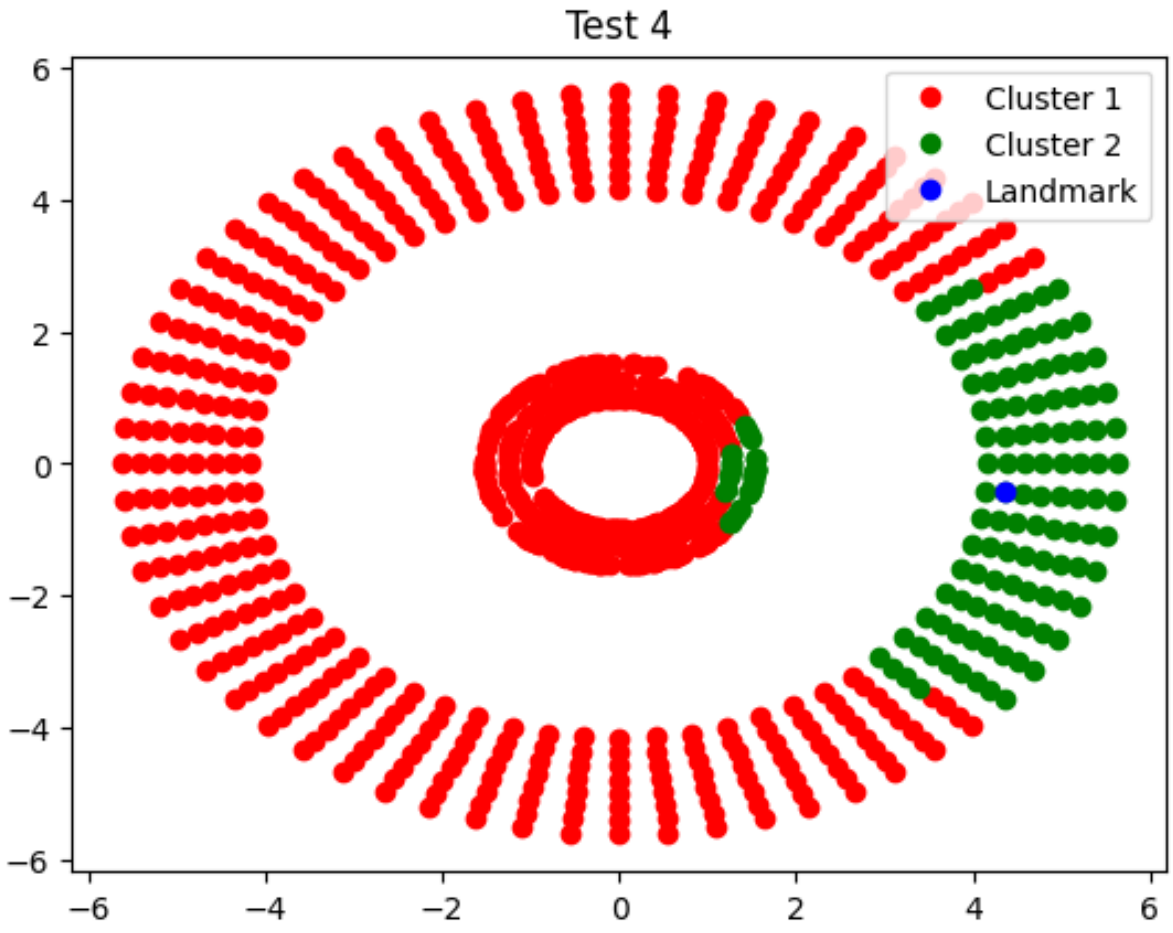

Figure 8: Single Landmark : Run 4

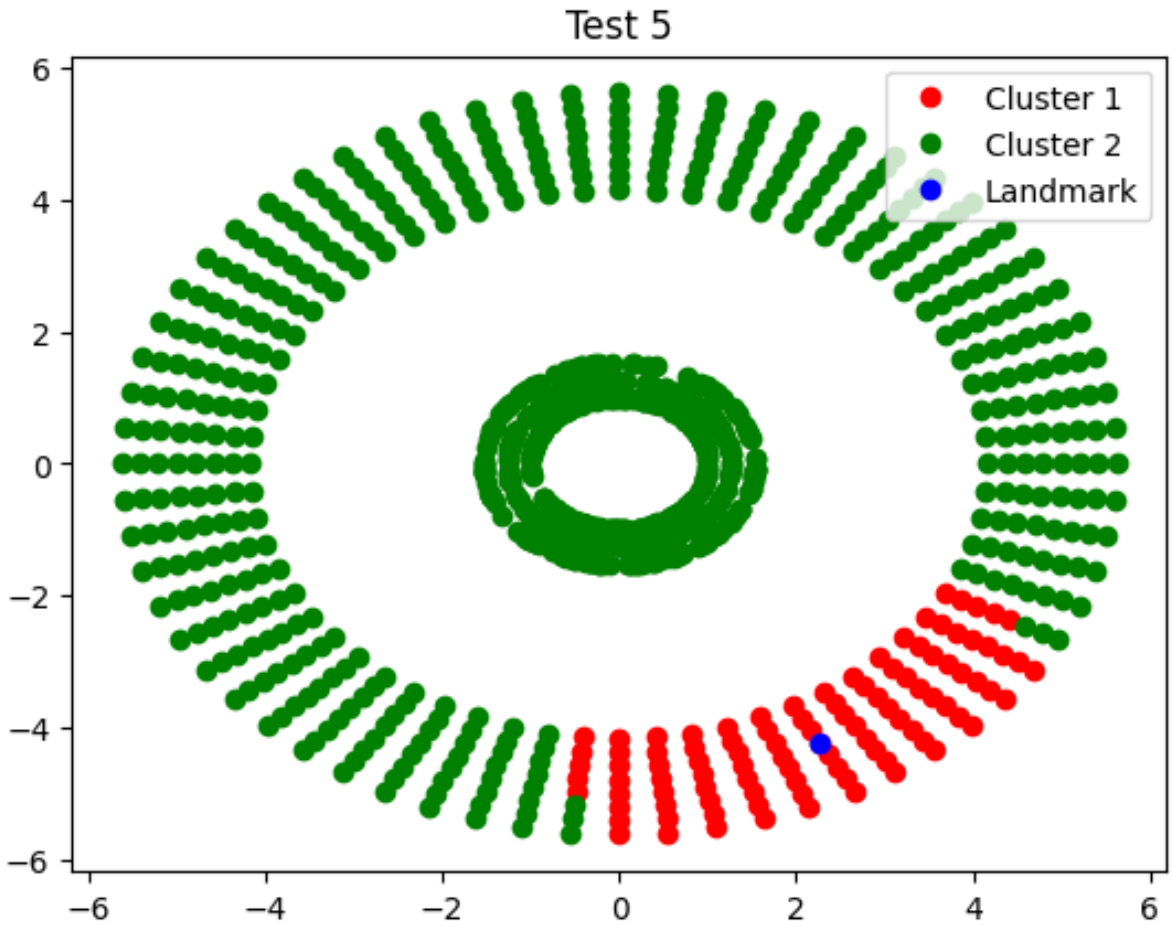

Figure 9: Single Landmark : Run 5

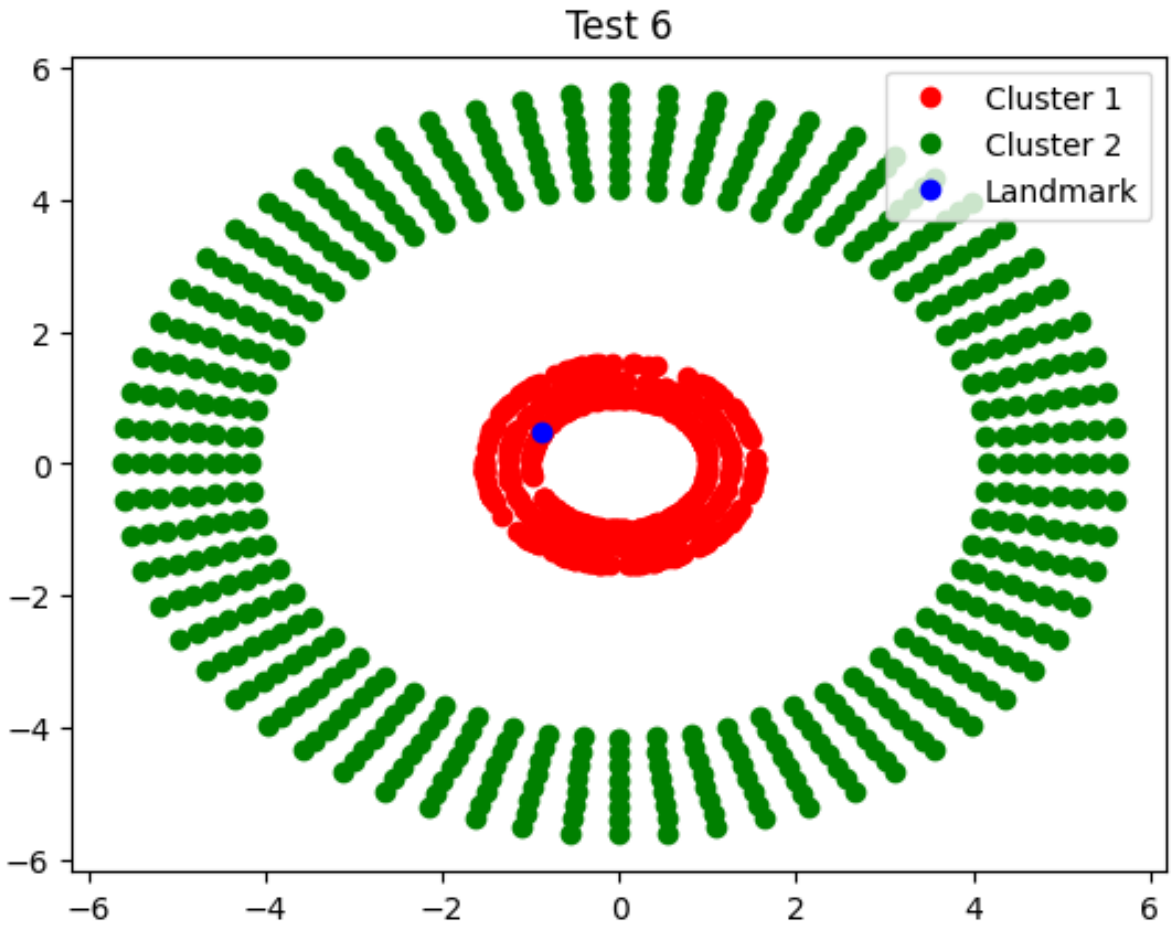

Figure 10: Single Landmark : Run 6

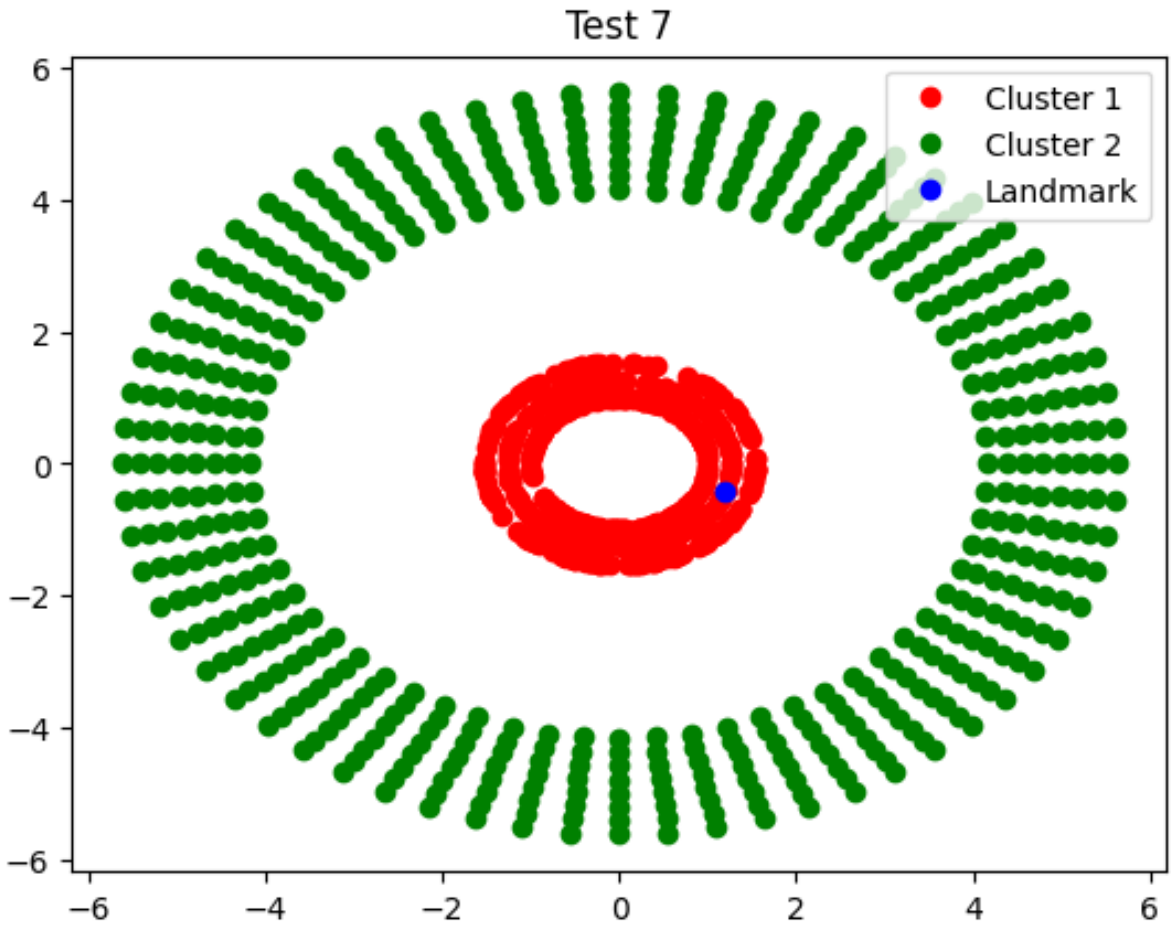

Figure 11: Single Landmark : Run 7

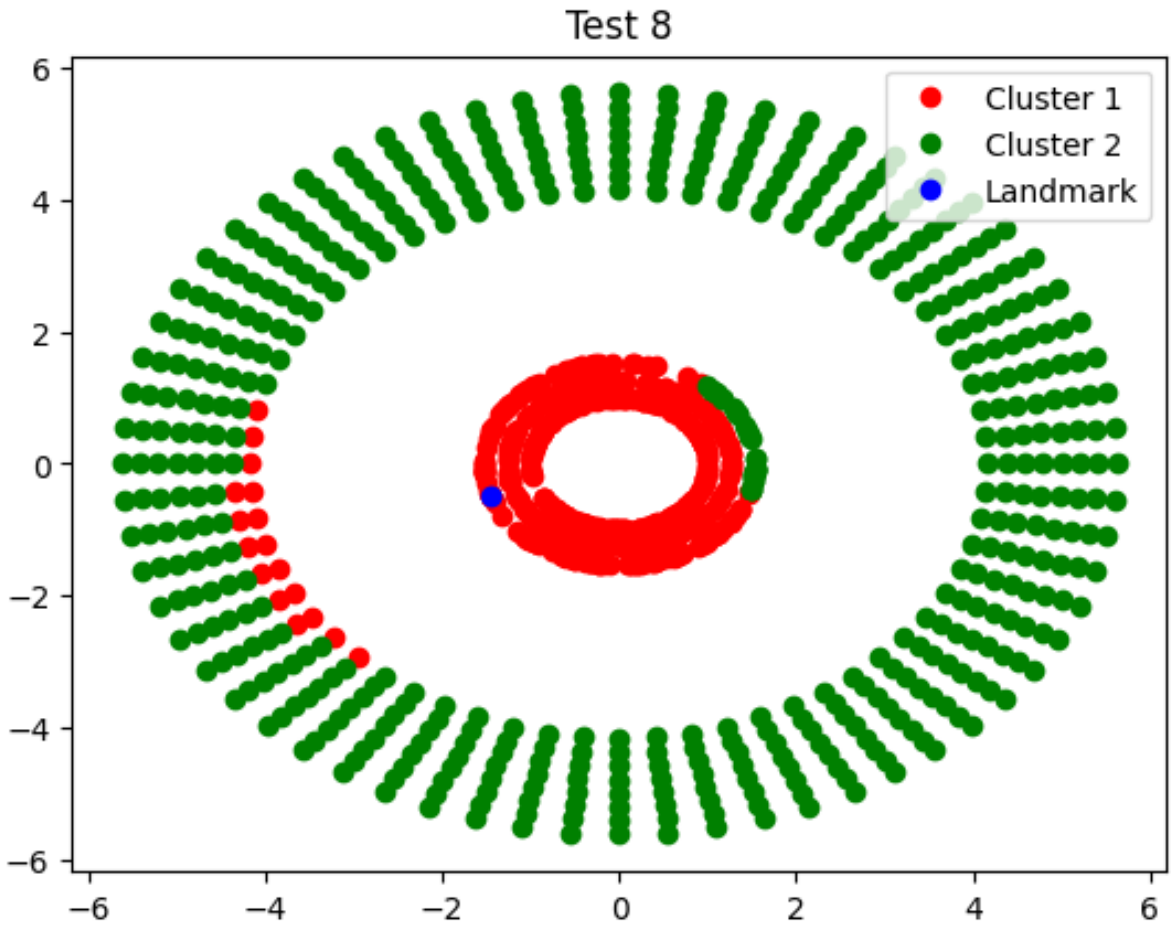

Figure 12: Single Landmark : Run 8

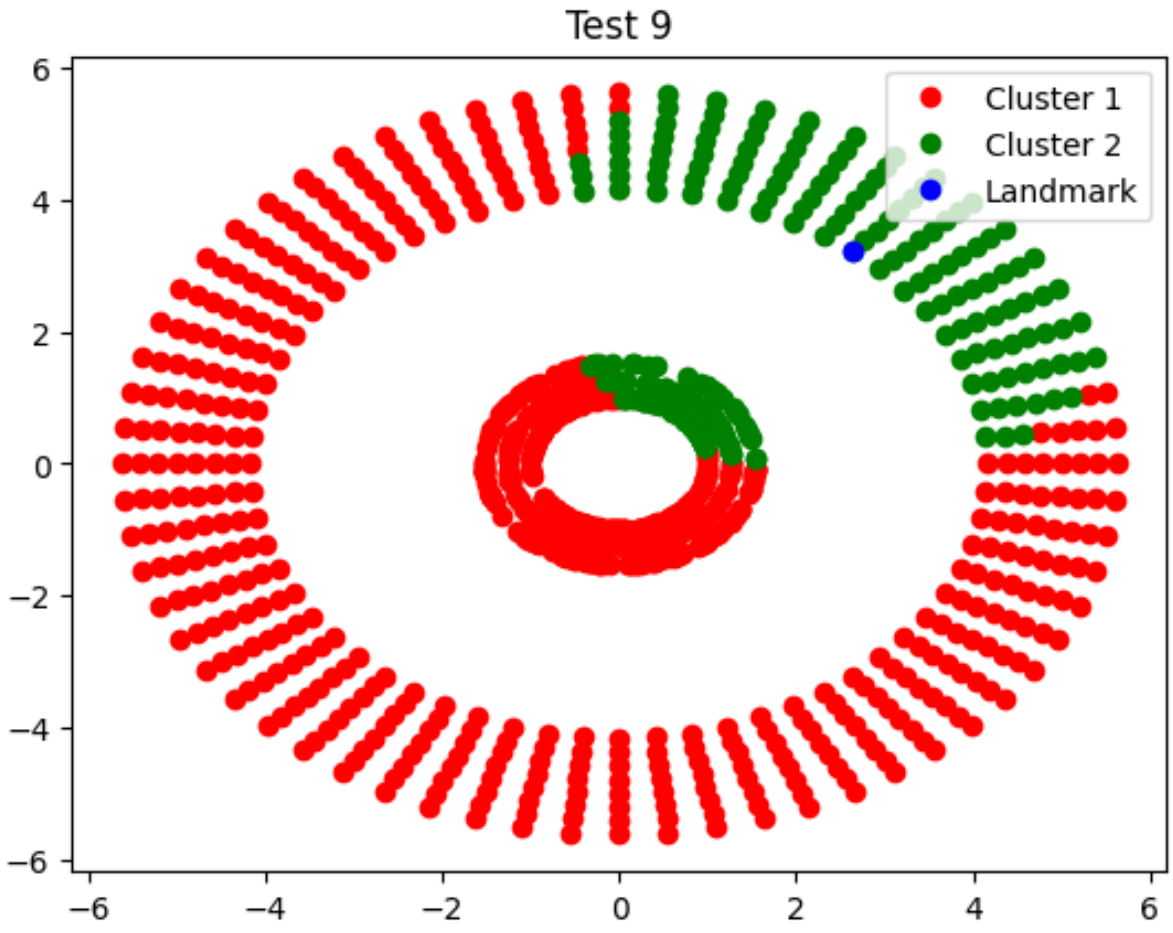

Figure 13: Single Landmark : Run 9

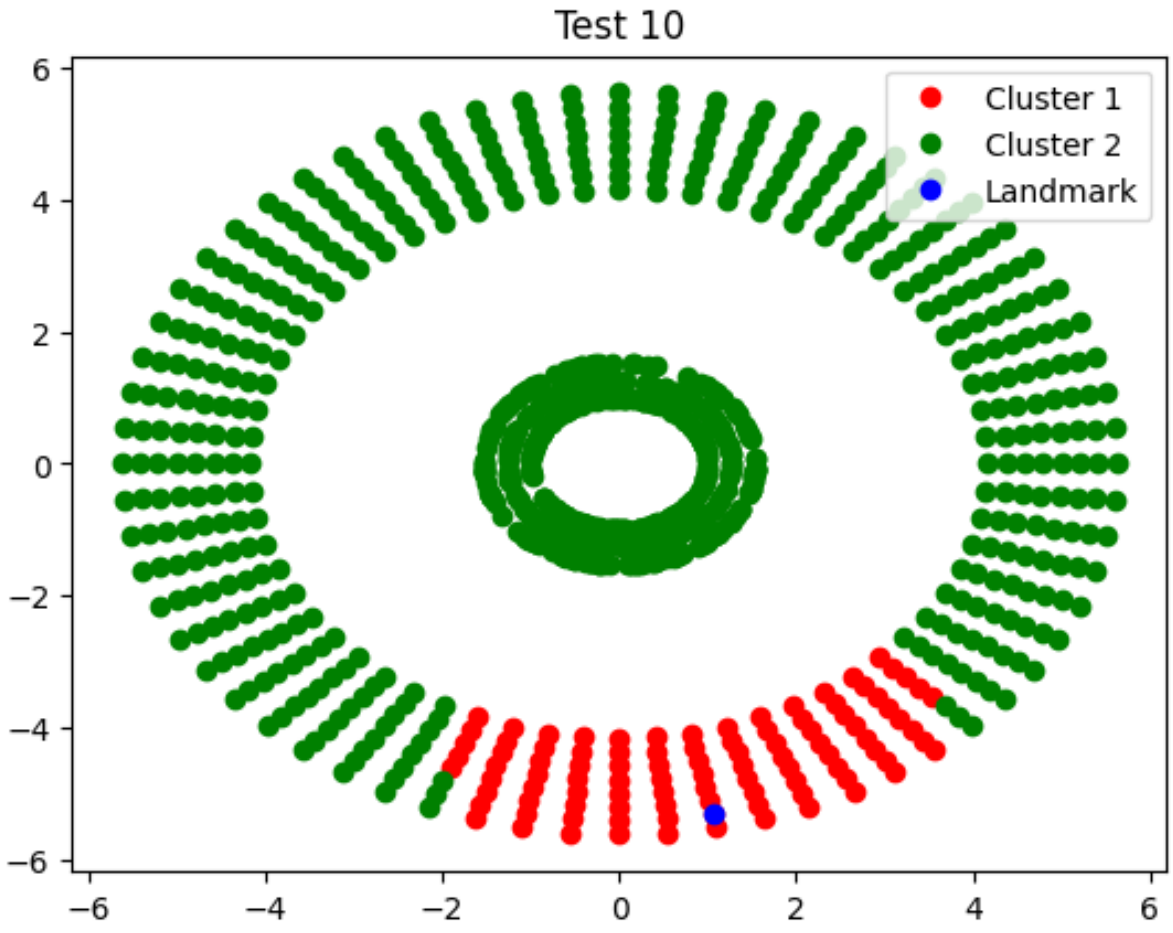

Figure 14: Single Landmark : Run 10

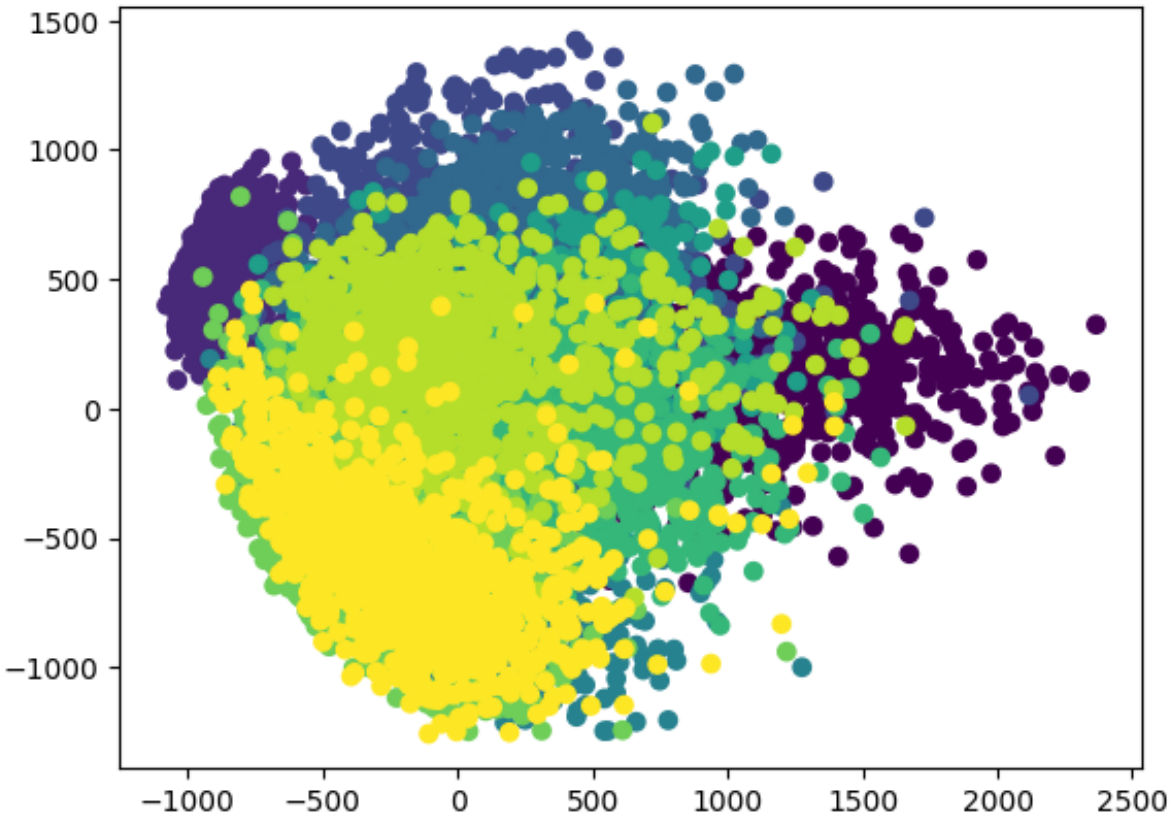

Figure 15: PCA

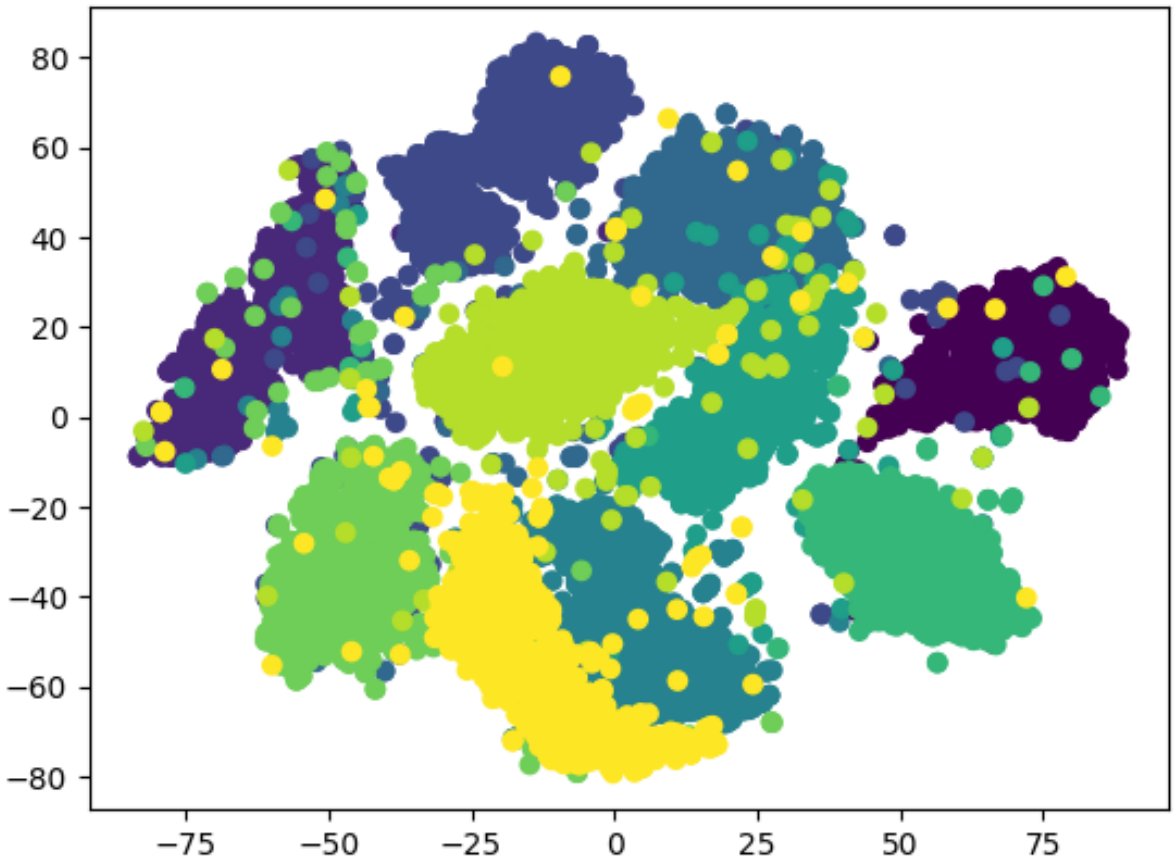

Figure 16: tSNE