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

Student Name: Manish Agrawal
Roll Number: 231110028
Date: November 16, 2023

QUESTION 1

As we already know and given as well The standard k-means loss function is as:

$$L(X, Z, \mu) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \|x_n - \mu_k\|^2$$

Now Apply SGD K means

STEP 1: To post it online by assigning x_n greedily to the best cluster after picking a random example x_n at a time.

Fix $\mu = \hat{\mu}$ and solve for z_n.

$$\hat{z}_n = \arg\min_{z_n} \textstyle\sum_{k=1}^{K} z_{nk} \|x_n - \hat{\mu}_k\|^2$$

For the assigned cluster, set $z_{nk} = 1$, indicating the cluster of data point x_n, and set all other z_{nk} values to zero. so no need of summation.

$$\hat{z}_n = \arg\min_{z_n} \; z_{nk} \|x_n - \hat{\mu}_{zn}\|^2$$

for each example x_n in the set $\{x_n\}_{n=1}^{N}$we must use the above equation to assign a cluster to x_n in order to complete Step 1.

STEP 2: Now we will update the cluster means
Now we will fix $z = \hat{z}$ then Solving for μ using SGD on the objective function

$$\hat{\mu} = \arg\min_{\mu} L(X, \hat{Z}, \mu)$$

$$\hat{\mu} = \arg\min_{\mu} \sum_{n=1}^{N} \sum_{n:\hat{z}_n = k} \|x_n - \mu_k\|^2$$

Now we will focus on kth cluster mean

$$\hat{\mu}_k = \arg\min_{\mu_k} \sum_{n:\hat{z}_n = k} ||x_n - \mu_k||^2$$

we will chose uniformly randomly x_n and approximate gradient g as we do in SGD at any iteration t

$$g \approx g_n$$

$$g_n = \frac{\partial}{\partial \mu_k} \left(\|x_n - \mu_k\|^2\right)$$

$$g_n = -2(x_n - \mu_k)$$

Now the update equation of mean in SGD is following:

$$\mu_k^{(t+1)} = \mu_k^{(t)} - \eta g^{(t)}$$

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

$\mu_k^{(t)}$ is the current estimate of the cluster mean for cluster k at iteration t. and $x_n^{(t)}$ is the randomly chosen data point at iteration t. η is the learning rate.

The step size can be $\eta \propto \frac{1}{N_k}$, where N_k is the number of data points in the k-th cluster, so that the updated mean is in the ratio of the sum of features of every data point to the total number of data points in that cluster.

Lets take $\eta = \frac{1}{N_k+1}$

$$\mu_k^{(t+1)} = \mu_k^{(t)} + 2\frac{1}{N_k+1}(x_n^{(t)} - \mu_k^{(t)})$$

This step size($\eta \propto \frac{1}{N_k}$) is a good choice because The learning rate is inversely proportional to the number of data points N_k in the cluster. This ensures that larger clusters contribute less to the update, preventing them from dominating the adjustment of the cluster mean. It helps in achieving a balanced influence of all data points on the cluster mean updates.This step size selection strategy helps balance the influence of individual data points on the mean during the SGD process, promoting stability in cluster updates.the step size ensures controlled mean movement, even for starting clusters with no data

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

Student Name: Manish Agrawal
Roll Number: 231110028
Date: November 16, 2023

QUESTION 2

The objective/loss function for an ideal projection can be formulated as follows:

$$J(w) = \frac{\|w^T\mu_+ - w^T\mu_-\|^2}{w^T S_w w}$$

Here:
μ_+ and μ_- are the means of the positive and negative classes, respectively.

S_w is the within-class covariance matrix. The numerator $(\mathrm{w}^T\mu_+ - \mathbf{w}^T\mu_-)^2$ maximizes the separation between class means, and the denominator $\mathbf{w}^T S_w \mathbf{w}$ minimizes the spread of data within each class.

The justification is as follows:

1. Maximizing the numerator increases the distance between the means of the two classes, promoting better separation.
2. Minimizing the denominator ensures that the data within each class is more compact, reducing overlap and improving class discrimination.

This formulation aligns with the intuition of pushing means of different classes apart while pulling the data within each class closer together.

Student Name: Manish Agrawal
Roll Number: 231110028
Date: November 16, 2023

QUESTION 3

given covariance matrix is $S = \frac{1}{N}(XX^T)$.

Consider an eigenvector v associated with the matrix
so it will satisfy the following equation where eigen value associated with eigenvector v is λ.

$$\boxed{Sv = \lambda v.}$$

So, Now the equation is :: $\frac{1}{N}(XX^T)v = \lambda v$.

Now we will multiply X^T both side of equation then we get ::

$$\frac{1}{N}(X^T X)(X^T v) = \lambda(X^T v).$$

Now substitute $u = X^T v$

Then equation will become $\frac{1}{N}(X^T X)u = \lambda u$.

$$\boxed{Su = \lambda u.}$$

u is nothing else but eigen vector of matrix S.
In Normal scenario to compute K eigen vectors which has maximum variance of matrix S takes $O(KD^2)$
But in this approach time complexity is $O(KN^2) + O(KND)$

$$\boxed{D > N}$$

so overall time complexity is $O(KND)$ which is lesser then $O(KD^2)$
so this approach reduces the time complexity

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

Student Name: Manish Agrawal
Roll Number: 231110028
Date: November 16, 2023

QUESTION
4

(1)

A conventional linear model is limited to scenarios requiring regression on a single linear curve but this approach involves blending various linear curves, amounting to a model with K distinct linear components.Essentially, before producing predictions for y, the model intiates by groups the data along K different linear curves. This approach is robust against outliers,as each point is assigned to its cluster using latent variables.Due to the possibility of outlier separation from the clustering operation, this procedure is benificial for reducing outliers inside a linear curve.

(2)

given ::

$$p(y_n|z_n,\theta) = \mathcal{N}(w_{z_n}^T x_n, \beta^{-1})$$

$$p(z_n = k) = \pi_k$$

Our latent variable model ::

$$p(z_n = k|y_n,\theta) = \frac{p(z_n = k)\cdot p(y_n|z_n = k,\theta)}{\sum_{l=1}^K p(z_n = l)\cdot p(y_n|z_n = l,\theta)}$$

$$p(y_n, z_n|\theta) = p(y_n|z_n,\theta)\cdot p(z_n|\theta)$$

Now will apply ALT-OPT algorithm

Step 1:
first we will find optimum z$_n$ ::

$$z_n = \arg\max_{z_n} \frac{\pi_k \cdot \mathcal{N}(w_{z_n}^T x_n, \beta^{-1})}{\sum_{l=1}^K \pi_l \cdot \mathcal{N}(w_l^T x_n, \beta^{-1})}$$

$$\boxed{\therefore \mathcal{N}(w_{z_n}^T x_n, \beta^{-1}) = \exp\left(-\frac{\beta}{2}(y_n - w^T z_n x_n)^2\right)}$$

$$\implies z_n = \arg\max_{z_n} \frac{\pi_k \cdot \exp\left(-\frac{\beta}{2}(y_n - w^T z_n x_n)^2\right)}{\sum_{l=1}^K \pi_l \cdot \exp\left(-\frac{\beta}{2}(y_n - w_l^T x_n)^2\right)}$$

Step 2:
Now we will re-estimate the parameters

$$w_k = (X_k^T X_k)^{-1} X_k^T y_k$$

$$N_k = \sum_{n=1}^{N} z_{nk}$$

$$\pi_k = \frac{N_k}{N}$$

In this case, training sets are clustered in class k and are represented by a matrix X_k which have dimension $N_k \times D$, whereas training set labels are clustered in class k and are represented by $N_k \times 1$ vectors y_k.
If $\pi_k = \frac{1}{K}$, then:

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

This update is equivalent to multi-output logistic regression.

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

Student Name: Manish Agrawal
Roll Number: 231110028
Date: November 16, 2023

QUESTION 5

Programming Problem: Part 1

1.Kernal ridge regression:

Our observed RMSE Values are as follows:
RMSE for $\lambda = 0.1$ is =0.032577670293572177
RMSE for $\lambda = 1$ is =0.17030390344202517
RMSE for $\lambda = 10$ is =0.6092671596540067
RMSE for $\lambda = 100$ is =0.9110858052767243
so we are observing that as we increase λ (regularization parameter) value our RMSE value also increase means if we incraese λ value then our model prediction accuracy decrease.Less regularization means a better curve toward train data and, indirectly, a better curve toward test data. This could be the case because the train and test sets appear to have been taken from the same distribution set with few outliers.
Plots are following:

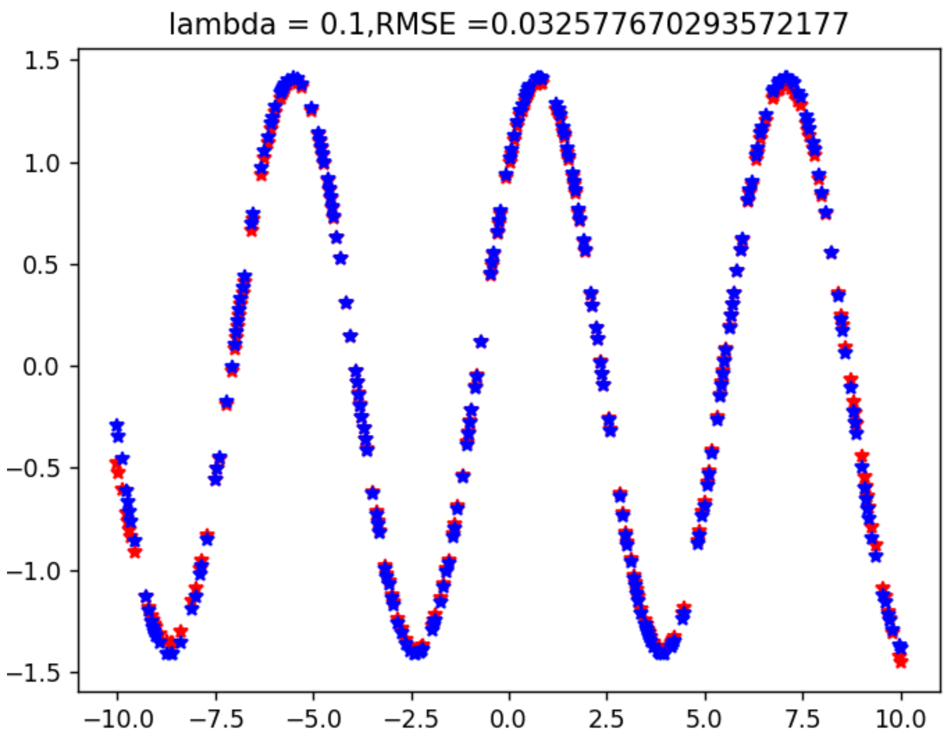

Figure 1: for λ=0.1

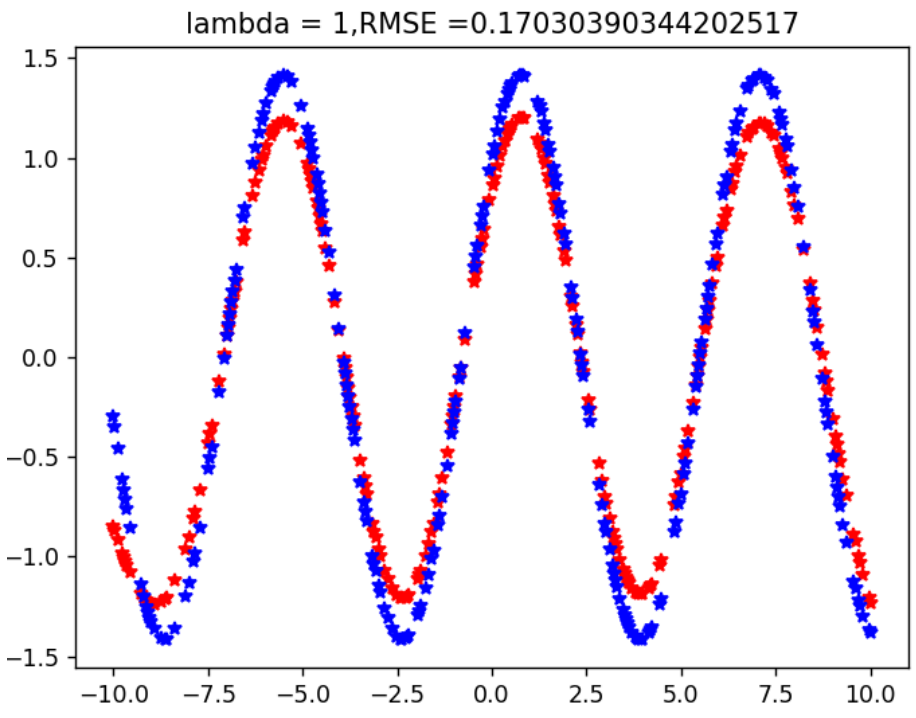

Figure 2: for λ=1

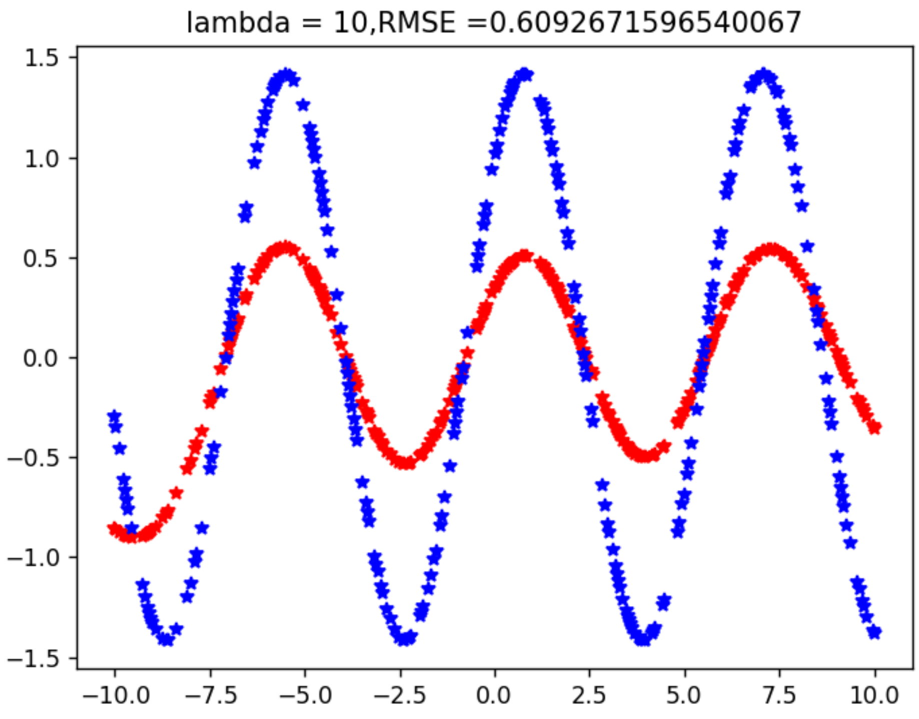

Figure 3: for λ=10

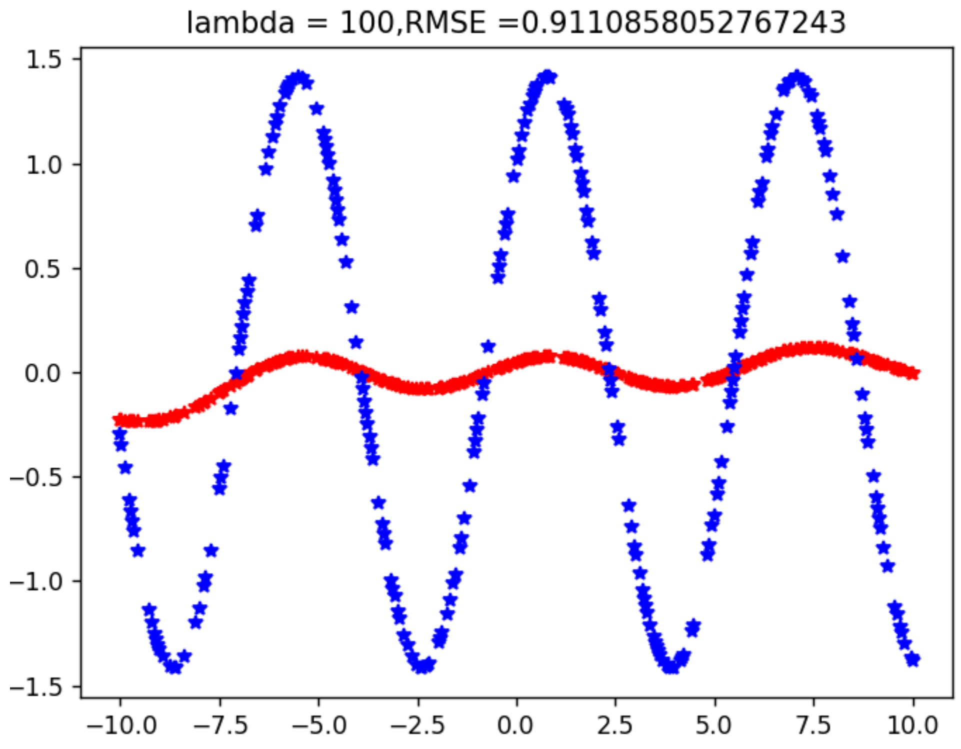

Figure 4: for λ=100

2.Landmark ridge regression:

here λ is same but L value is changed We are choosing landmarks uniformly randomly from the data set so every run of the program may give different answer.
Our observation are follows:

λ value is : 0.1
RMSE for L = 2 is =0.9745378656924272
λ value is : 0.1
RMSE for L = 5 is =0.9092751376410605
λ value is : 0.1
RMSE for L = 20 is =0.25841967010040406
λ value is : 0.1
RMSE for L = 50 is =0.12372811665710566
λ value is : 0.1
RMSE for L = 100 is =0.0662316467507647

We are observing that if we increase L value then RMSE value decrease so our model prediction accuracy increase.Because fewer feature points are taken, the prediction error increases with decreasing L value.
L=100 seems good enough because it is giving minimum error.L=50 is also seems to be good enough because there is very less difference in RMSE values.so both are good enough.
Plots are following:

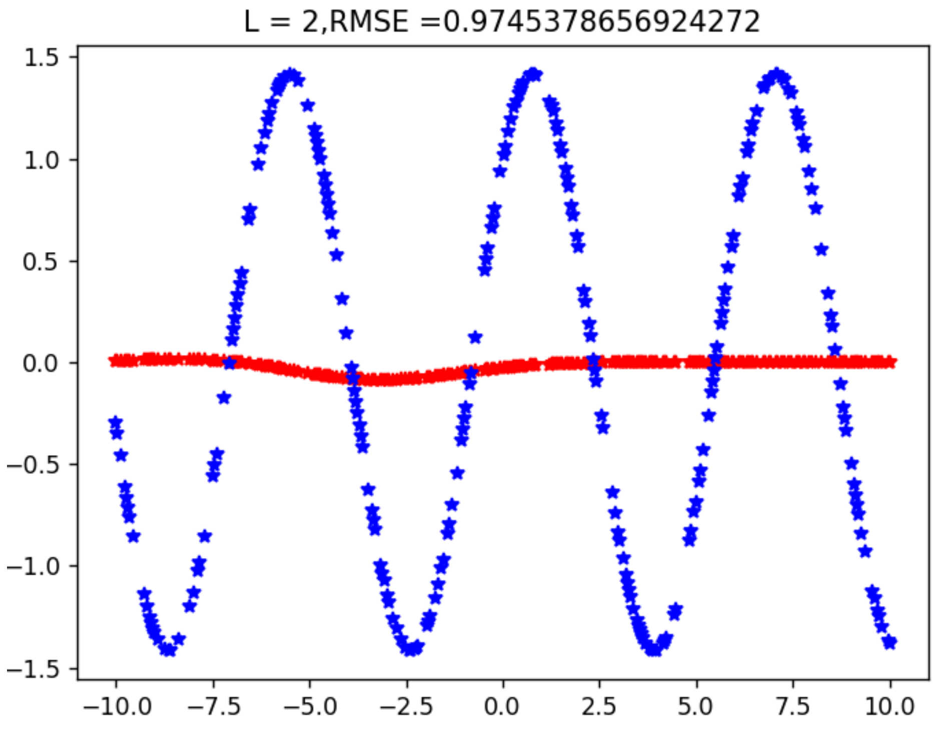

Figure 5: for L=2

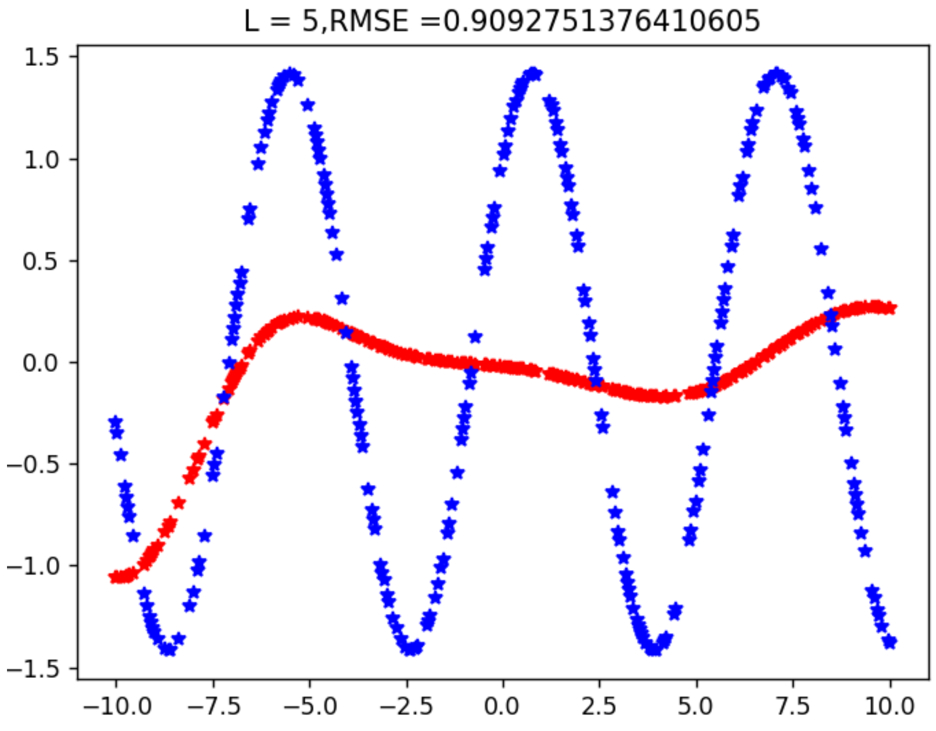

Figure 6: for L=5

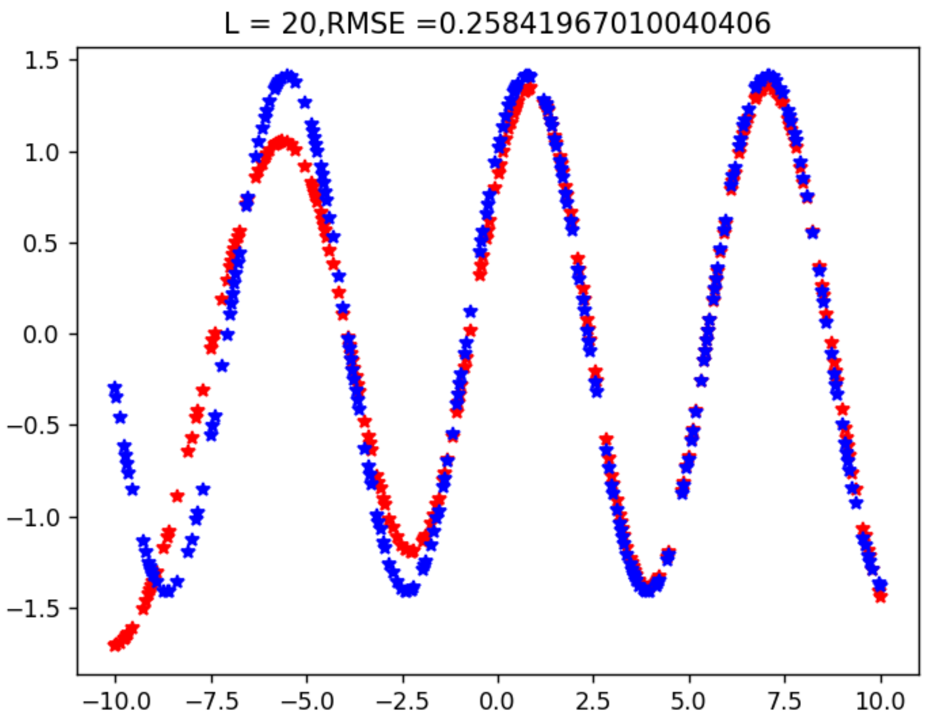

Figure 7: for L=20

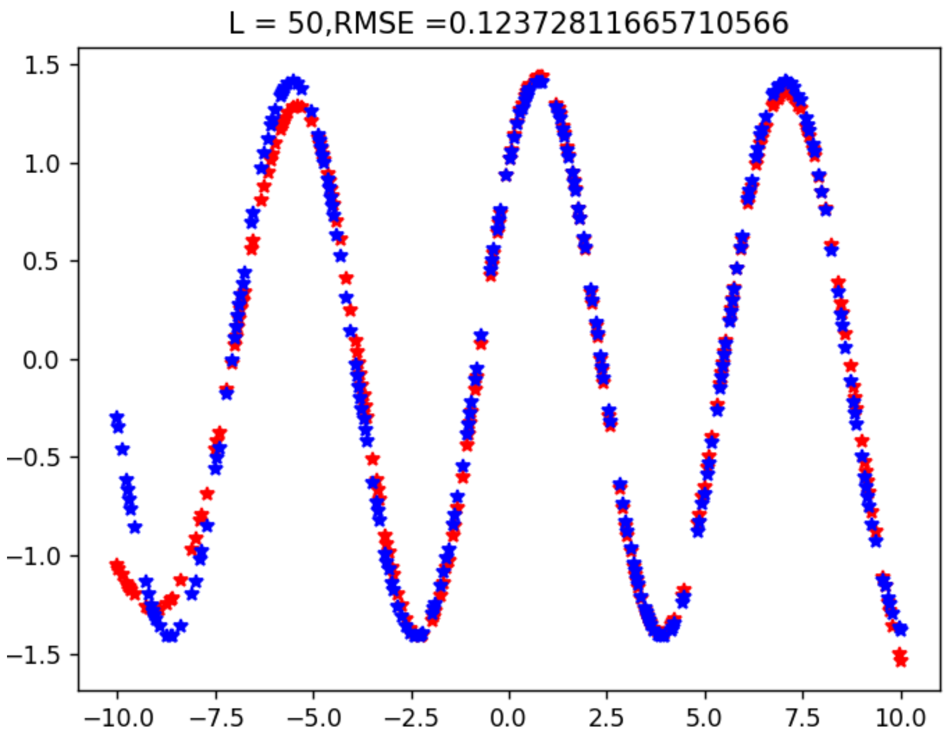

Figure 8: for L=50

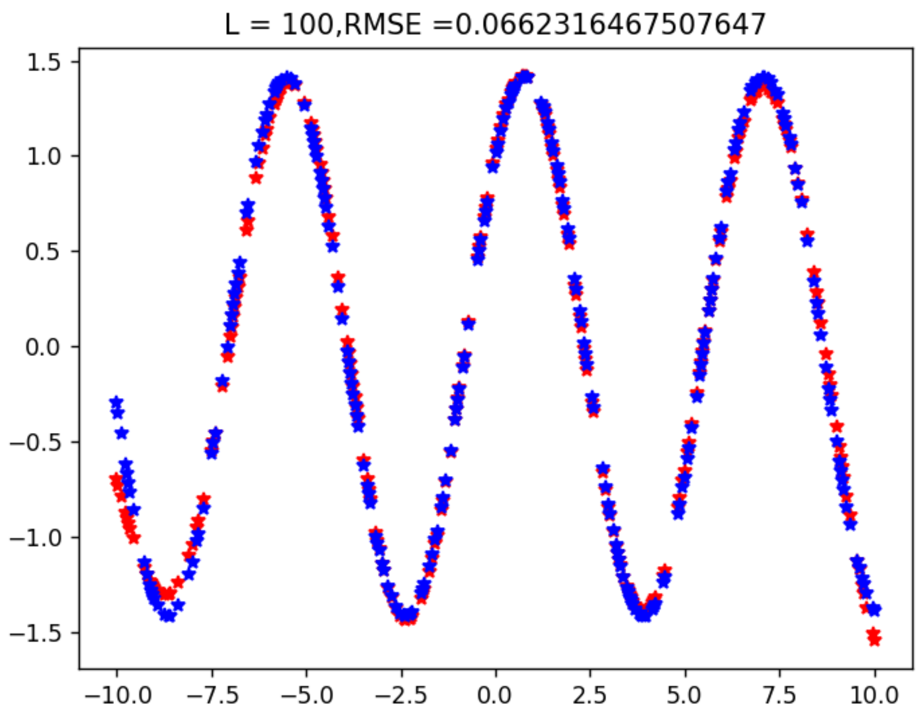

Figure 9: for L=100

Programming Problem: Part 2

1. Using Hand-crafted Features::
Here is the original data plot before transformation:

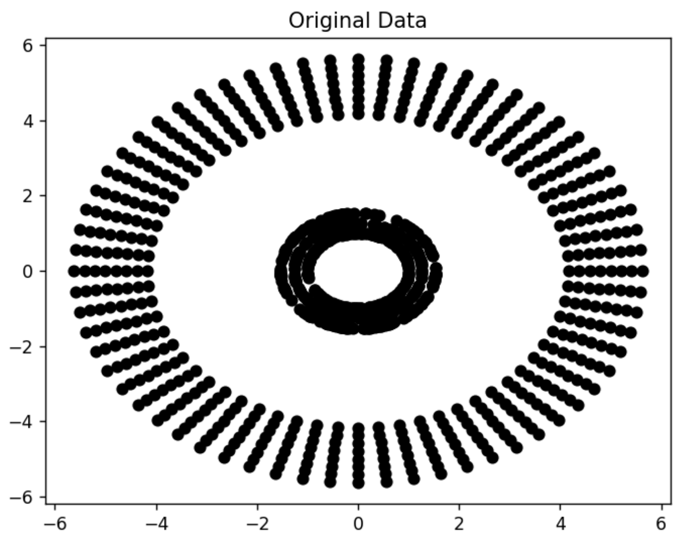

Figure 10: Original data before applying transformation

Originally we have 2D data and Now we are transforming it to 3D like following::

$$[x, y] \Rightarrow [x^2, \sqrt{2 \cdot |xy|}, y^2]$$

After applying transformation we apply K-means on that and plot clustering result in original 2D space which is following:

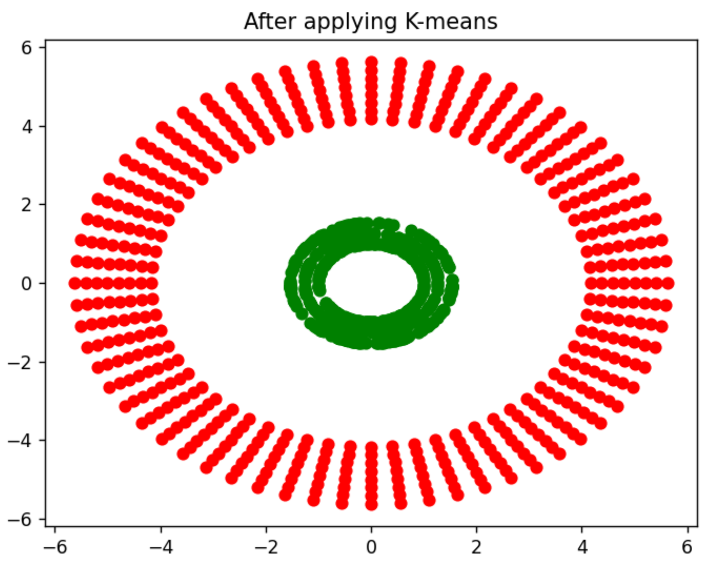

Figure 11: K-means result on original 2D data

2. Using Kernels::

We are choosing landmark uniformly randomly from the data set so every run of the program may give different answer.

Our observation is :Because the provided data is easily clusterable, we use a feature transform that is based on the point's distance from the origin. When a landmark is closer to the origin, the data sometimes clusters around it more effectively than when it is farther away since the data are dispersed radially around the origin!

Following are the plots:

blue square is landmark point.

Note:-L=trainingData[53,:] means landmark is 53rd index row is selected as landmark.

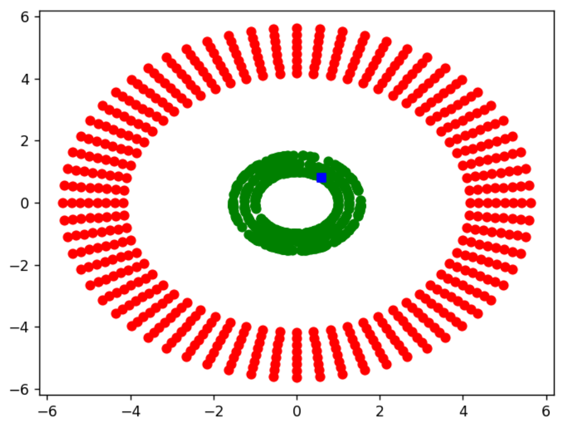

Figure 12: L=trainingData[53,:]

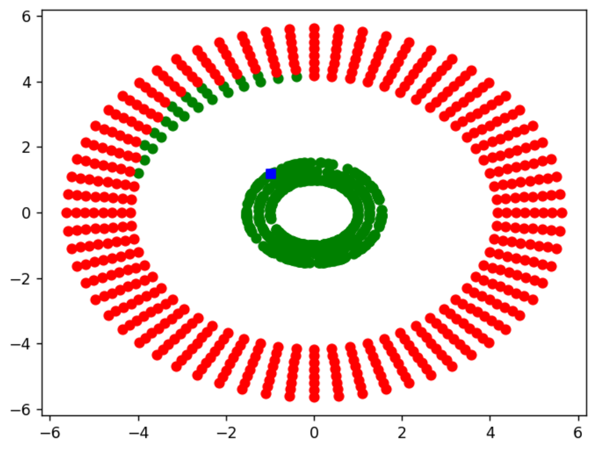

Figure 13: L=trainingData[61,:]

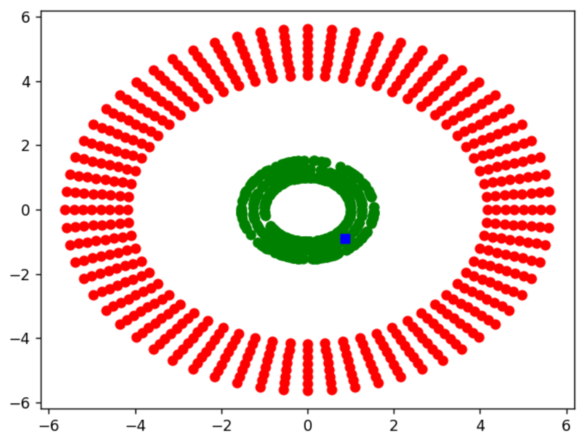

Figure 14: L=trainingData[749,:]

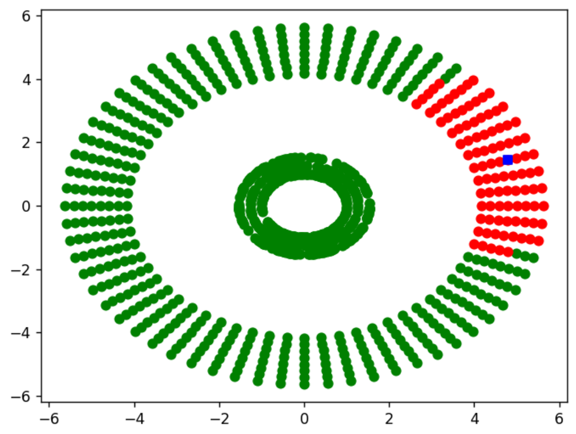

Figure 15: L=trainingData[250,:]

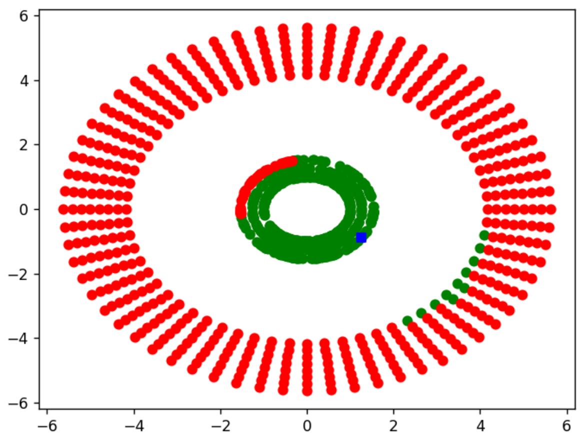

Figure 16: L=trainingData[255,:]

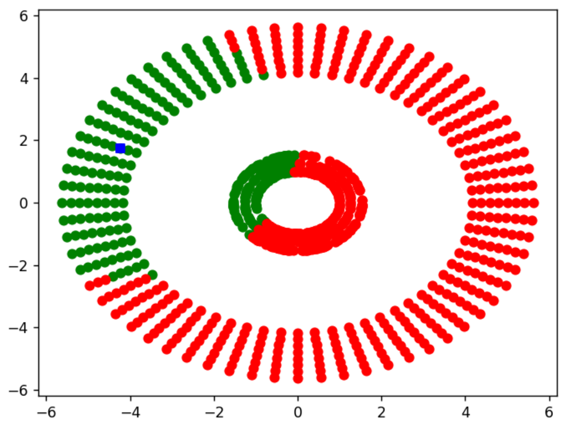

Figure 17: L=trainingData[369,:]

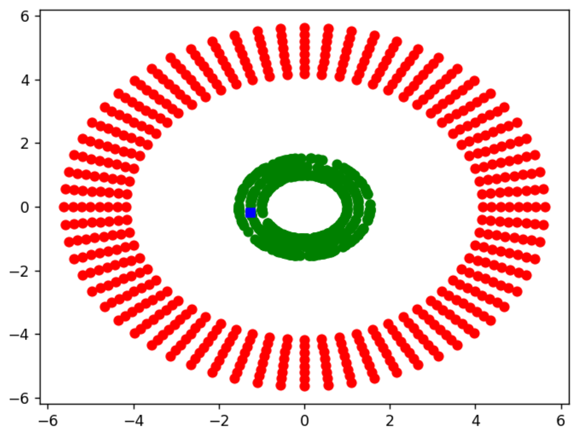

Figure 18: L=trainingData[300,:]

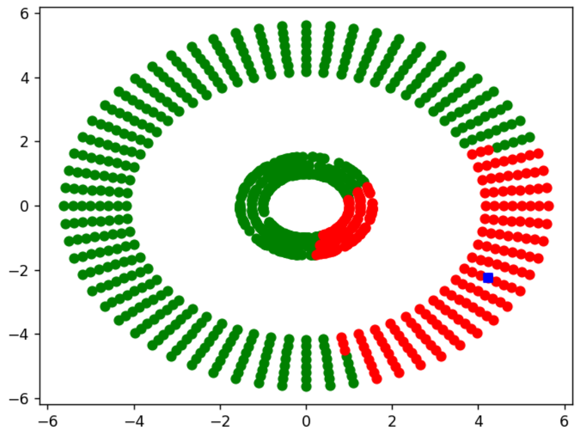
Figure 19: L=trainingData[716,:]

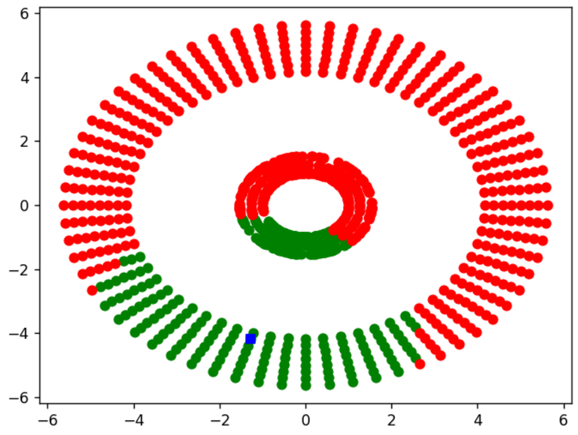
Figure 20: L=trainingData[838,:]

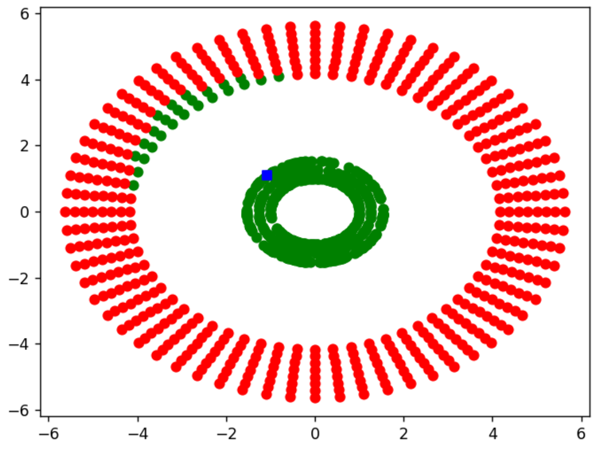
Figure 21: L=trainingData[940,:]

Programming Problem: Part 3

Clusters in tSNE appear to be more separated visually then PCA, and errors in tSNE seem to be lower than those in PCA when ten different initializations are done!

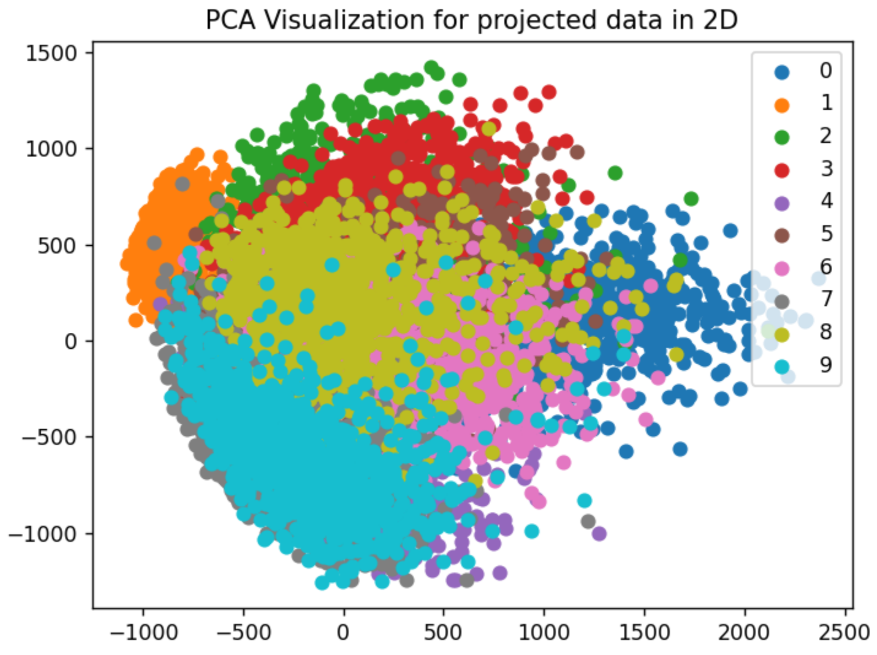

Figure 22: Visulaization the projected data in two dimensions using PCA method

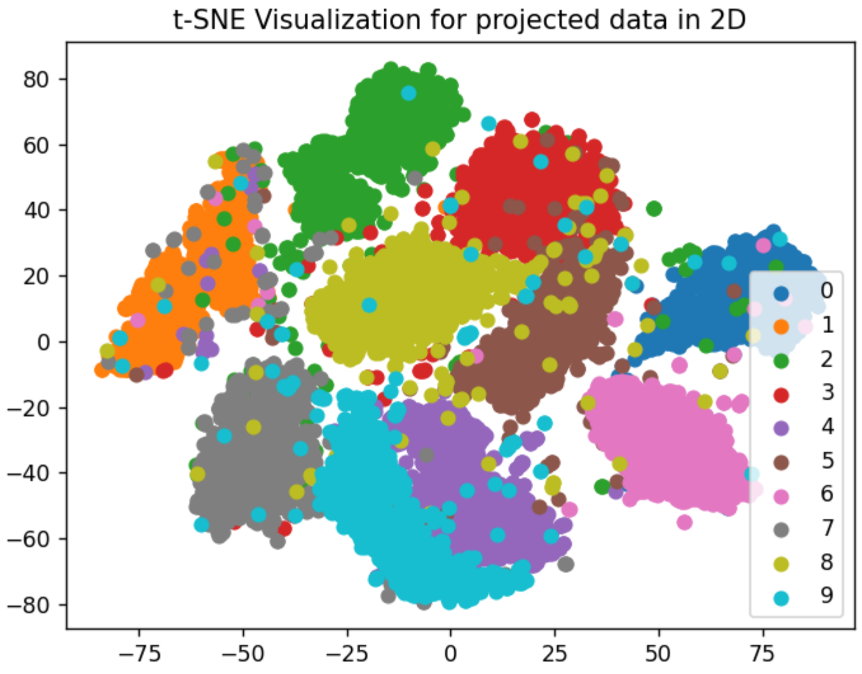

Figure 23: Visulaization the projected data in two dimensions using tSNE method