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

Student Name: Sanjeev Kumar
Roll Number: 231110044
Date: November 16, 2023

QUESTION 1

1. Solving Step 1 (Assigning $\mathbf{x}_n$ to the Nearest Cluster): Calculate the distance between the new data point $\mathbf{x}_n$ and the cluster means $\{\mu_k\}_{k=1}^{K}$. Assign $\mathbf{x}_n$ to the cluster k with the closest mean using the Euclidean distance.

Mathematically, z_{nk} can be found as below:
$z_{nk} = \arg\, min_k\ ||\mathbf{x}_n - \mu_k||^2$

2. SGD-based Cluster Mean Update Equations for Step 2: The update equation for the cluster mean μ_k using SGD can be derived from the K-means objective function. Taking the gradient of $\mathcal{L}$ with respect to μ_k, we get:

$$\frac{\partial \mathcal{L}}{\partial \mu_k} = -2\sum_{n=1}^{N} z_{nk}(x_n - \mu_k)$$

The update equation for μ_k using SGD with a step size η becomes:

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

where $\mu_k^{(t)}$ represents the value of μ_k at iteration t.

Intuition behind the update equation: The update equation makes sense because it moves the cluster mean μ_k towards the data point x_n proportional to the responsibility z_{nk} (which measures the influence of x_n on cluster k) and the difference between the current mean and the data point.

3. Choice of Step Size:
Since loss function here is non-convex and We are using SGD approach, for converging to good optima, we can use adaptive step size methods such as learning rate schedule which dynamically adjust the step size based on the gradient update history.

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

Student Name: Sanjeev Kumar
Roll Number: 231110044
Date: November 16, 2023

QUESTION 2

Inputs $\mathbf{x}_n \in \mathbb{R}^D$ and labels $y_n \in \{+1, -1\}$.

We want to project the inputs into one dimension using a projection direction given by $\mathbf{w} \in \mathbb{R}^D$ such that, after the projection, the distance between the means of the inputs from the two classes becomes as large possible, and the inputs within each class become as close to each other as possible.

Suppose μ_{+1} and μ_{-1} denote the means of positive and negative class samples after projection.

Loss function for above problem can be written as the combination of two terms: one that maximizes the distance between means of the classes after projection, second that minimizes the spread of within class input samples after projection. Let $\mathcal{L}1$ and $\mathcal{L}2$ denote the first and second objective functions respectively.

For maximizing the distance between inter-class means of two classes after projection, we can maximize the distance between the means μ_{+1} and μ_{-1}. Thus,
maximize $\mathcal{L}1(\mathbf{w}) = (\mu_{+1} - \mu_{-1})^2$

For minimizing the within class spread of inputs after projection, we can minimize the distance between the samples of a class and it's mean. Thus,
minimize $\mathcal{L}2(\mathbf{w}) = \sum_{n, y_n = +1} ||\mathbf{w}^T\mathbf{x}_n - \mu_{+1}||^2 + \sum_{n, y_n = -1} ||\mathbf{w}^T\mathbf{x}_n - \mu_{-1}||^2$

Therefore, the overall objective/loss function for the problem will be (say $\mathcal{L}(\mathbf{w})$):
maximize $\mathcal{L}(\mathbf{w}) = \mathcal{L}1(\mathbf{w}) - \lambda\mathcal{L}2(\mathbf{w})$

Here, λ is a trade-off parameter that controls the importance of maximizing the inter-class distance relative to minimizing the within class distances. This parameter is crucial as it balances the contribution of the two terms in the objective function.

By optimizing this objective function with respect to the projection vector $\mathbf{w}$, we can find the optimal projection direction that maximizes the distance between class means and minimizes the spread within each class after projection.

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

Student Name: Sanjeev Kumar
Roll Number: 231110044
Date: November 16, 2023

QUESTION 3

For doing PCA on a NxD matrix $\mathbf{X}$, we need to find the eigenvectors of the covariance matrix $\mathbf{S} = \frac{1}{N}\mathbf{X}^T\mathbf{X}$ (assuming centered data).

For finding eigenvector $\mathbf{u} \in \mathbb{R}^D$ of S we can use an eigenvector v of matrix C as below:
where, $\mathbf{C} = \frac{1}{N}\mathbf{X}\mathbf{X}^T$;

If $\mathbf{v}$ is the eigenvector of $\mathbf{C}$ then, $\mathbf{Cv} = \lambda\mathbf{v}$, where λ is the eigenvalue corresponding to the eigenvector $\mathbf{v}$.

Multiplying both the sides of above equation by $\mathbf{X}^T$.

$$\mathbf{X}^T\mathbf{Cv} = \lambda\mathbf{X}^T\mathbf{v}$$

substituting $\mathbf{C}$ in above equation, we get:

$$\mathbf{X}^T(\tfrac{1}{N}\mathbf{X}\mathbf{X}^T)\mathbf{v} = \lambda\mathbf{X}^T\mathbf{v}$$

$$(\tfrac{1}{N}\mathbf{X}^T\mathbf{X})\mathbf{X}^T\mathbf{v} = \lambda\mathbf{X}^T\mathbf{v}$$

$$\mathbf{S}\mathbf{X}^T\mathbf{v} = \lambda\mathbf{X}^T\mathbf{v}$$

let $\mathbf{X}^T\mathbf{v} = \mathbf{u}$

$$\text{then, } \mathbf{Su} = \lambda\mathbf{u}$$

Therefore, $\mathbf{u}$ is an eigenvector of covariance matrix $\mathbf{S}$.

Advantage:
The advantage of obtaining eigenvectors of $\mathbf{S}$ in this way is computationally efficient, when $D > N$.

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

Student Name: Sanjeev Kumar
Roll Number: 231110044
Date: November 16, 2023

QUESTION 4

1. What this model is doing and why it is better than standard probabilistic model which uses a single weight vector to model each response:

This model by introducing a latent variable z_n (which captures the cluster of the sample) for each training example $(\mathbf{x}_n, y_n)$ and K weight vectors $\mathbf{w}_k$ for each of the clusters, is trying to capture different relationships between inputs and outputs for different clusters, enabling a more flexible representation compared to the standard probabilistic linear model. We are doing this because as per problem statement using a single weight vector is insufficient to capture the different relationships between input and output.

2. ALT-OPT algorithm to estimate Z and (MLE of) Θ:

For Gaussian Model: $\beta^{-1} = \sigma^2$

So we have to derive the ALT-OPT algorithm for estimating latent variables $\mathbf{Z}$ and parameters $\{\mathbf{w}_k\}$ for the given latent variable model.Since β is same for each of the cluster as per question, we don't need to update it.

Initialization:

- Initialize latent variables $\mathbf{Z}$ randomly as per given prior distribution.
- Initialize parameters $\{\mathbf{w}_k\}$ randomly.

E-step (Update Z): Compute the posterior probabilities $p(z_n = k|\mathbf{x}_n, y_n, \{\mathbf{w}_k\})$ for each data point $\mathbf{x}_n$ and each cluster k. The posterior probability is proportional to the product of prior of z_n and likelihood of data point $\mathbf{x}_n$ belonging to cluster k:

$$p(z_n = k|\mathbf{x}_n, y_n, \{\mathbf{w}_k\}, \propto \pi_k \exp(-||y_n - \mathbf{w}_k^T\mathbf{x}_n||^2)$$

After computing these probabilities for each cluster, the data point $\mathbf{x}_n$ is assigned to the cluster k with the highest probability.

M-Step (Update $\{\mathbf{w}_k\}$):

- **Update cluster specific $\mathbf{w}_k$:** For each cluster k, the weight vector $\mathbf{w}_k$ is updated by taking the weighted average of the data points assigned to cluster k:

  $$\mathbf{w}_k = \frac{\sum_{n=1}^{N} \mathbb{I}(z_n=k) y_n \mathbf{x}_n}{\sum_{n=1}^{N} \mathbb{I}(z_n=k) \mathbf{x}_n^T \mathbf{x}_n}$$

  After updating $\{\mathbf{w}_k\}$, we reassign each data point $\mathbf{x}_n$ to the cluster k where the likelihood $\exp(-||y_n - \mathbf{w}_k^T\mathbf{x}_n||^2)$ is highest. This step ensures that the latent variables $\mathbf{Z}$ are updated to reflect the most likely cluster assignments given the updated parameters.

  The iterative nature of these E-step and M-step updates allows the algorithm to refine the cluster assignments $\mathbf{Z}$ and the cluster-specific parameters $\{\mathbf{w}_k\}$ until convergence, leading to a model that captures the underlying patterns and relationships within the data.

Update of z_n if prior distribution is uniform:

If the prior distribution of z_n is uniform i.e., $\pi_k = \frac{1}{K}$, where K is the no of clusters, instead of multinaulli, the assignment of z_n to a specific cluster is solely based on the likelihood term (the Gaussian distribution term in the given model). The uniform prior does not bias the assignment towards any particular cluster, ensuring that the assignment of z_n is solely driven by the data likelihood given the cluster parameters.

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

Student Name: Sanjeev Kumar
Roll Number: 231110044
Date: November 16, 2023

QUESTION

5

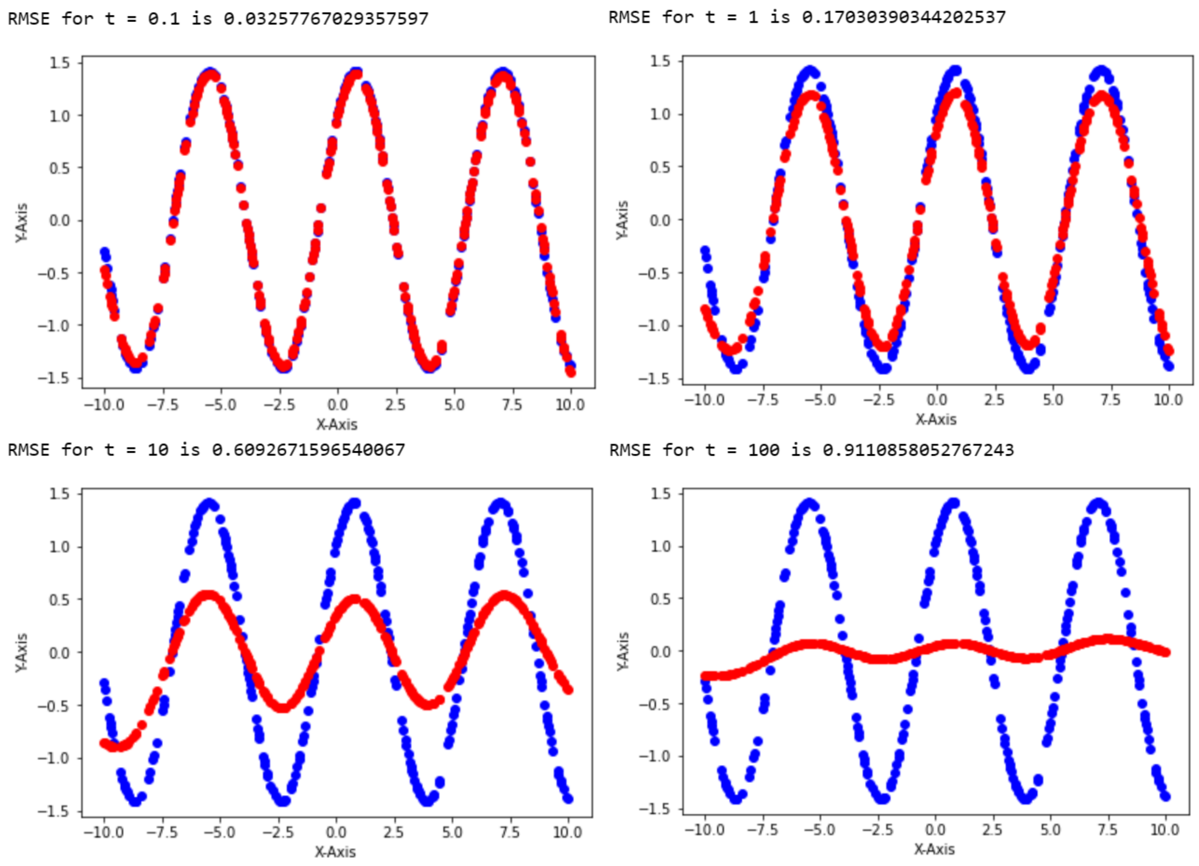
1 Part1:

1.1 Kernel Ridge Regression:

As we increase the hyper-parameter λ underfitting of data increases due to higher weightage to regularizer.

RMSE for different λ is as below:

- $\lambda = 0.1 \implies RMSE = 0.0326$
- $\lambda = 1 \implies RMSE = 0.1703$
- $\lambda = 10 \implies RMSE = 0.6093$
- $\lambda = 100 \implies RMSE = 0.9111$



```
RMSE for t = 0.1 is 0.032577670293575977
RMSE for t = 1 is 0.17030390344202537
RMSE for t = 10 is 0.6092671596540067
RMSE for t = 100 is 0.9110858052767243
```

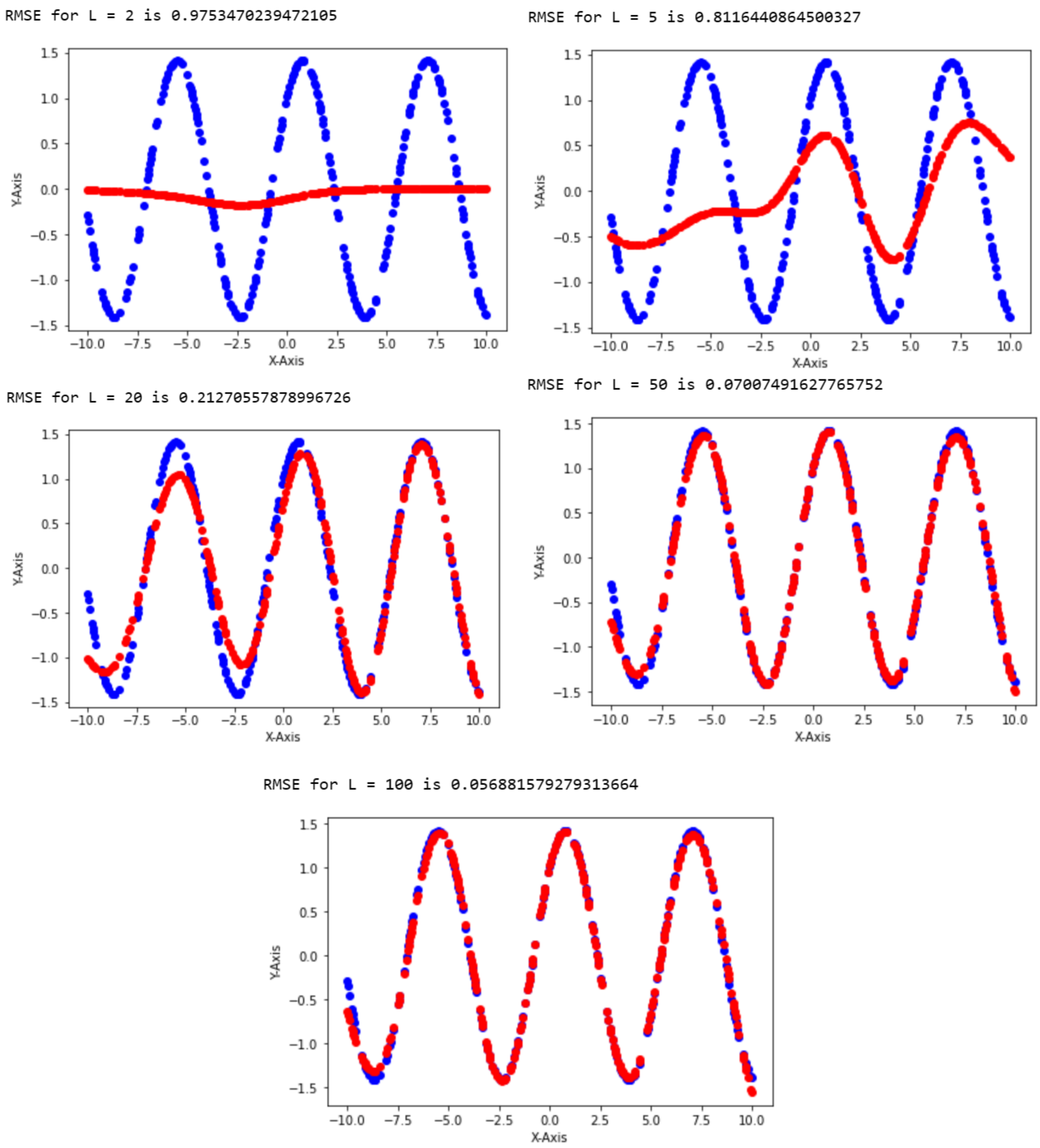
Figure 1: Kernel Ridge Regression

1.2 Landmark Ridge Regression:

As we increase the no of landmarks prediction fits to true values.

- $L = 2 \implies RMSE = 0.9753$
- $L = 5 \implies RMSE = 0.8116$
- $L = 20 \implies RMSE = 0.2127$
- $L = 50 \implies RMSE = 0.0701$
- $L = 100 \implies RMSE = 0.0569$

L = 100 seems the best choice out of all the given choices.



```
RMSE for L = 2 is 0.9753470239472105
RMSE for L = 5 is 0.8116440864500327
RMSE for L = 20 is 0.21270557878996726
RMSE for L = 50 is 0.07007491627765752
RMSE for L = 100 is 0.056881579279313664
```

Figure 2: Landmark Ridge Regression

2 Part2:

2.1 Using Hand Crafted Features:

On plotting the sample data,we can see two clusters of circular form with different radius are present. For applying standard K-means we can convert the two dimensional input into one dimensional input by taking the Euclidean norm thus samples from one cluster 1 having small radius will be one side and samples from Cluster 2 having large radius will be the other side.

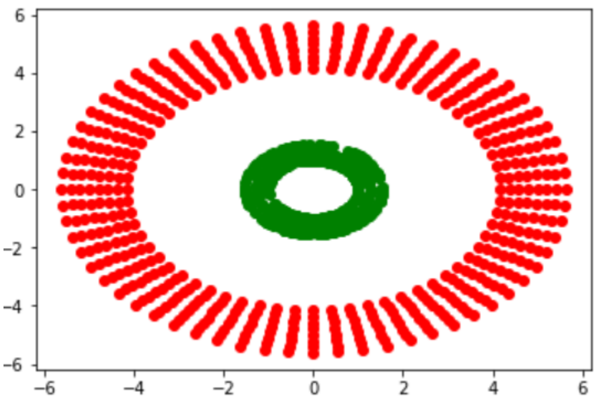


Figure 3: Using Hand Crafted Features

2.2 Using Kernels:

Correct clustering in some cases and not so correct in other cases can be attributed to the characteristics of RBF kernel and placement of landmark. It assigns higher weights to the points that are close to the landmark and decreases the weight exponentially as the distance increases. If the landmark is placed within or near one cluster, the RBF kernel features will be more responsive to the points within that cluster, effectively emphasizing the local structure.

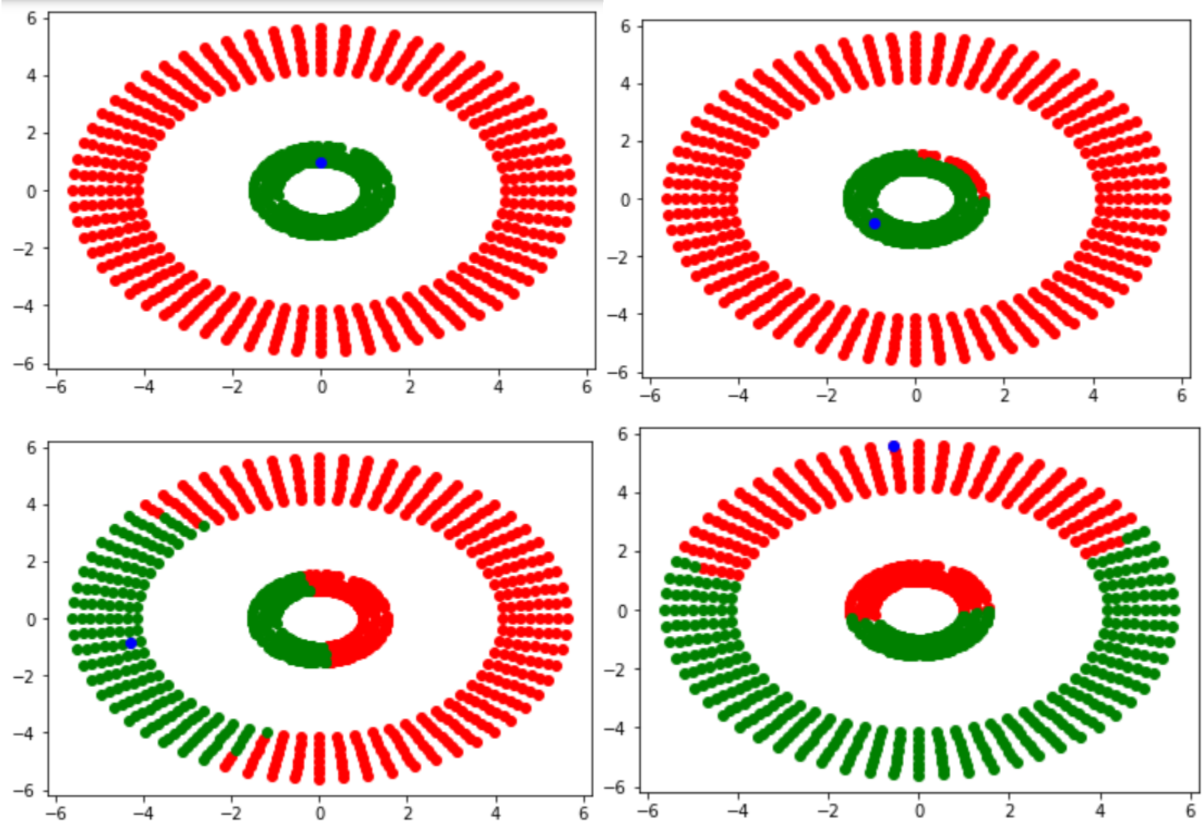


Figure 4: Using Kernels

Some sample plots attached. All 10 plots can be generated using code attached.

3 Part3:

PCA being the linear method of dimensionality reduction overlaps the samples from different clusters while t-SNE being non-linear method has almost segregated all the clusters perfectly.

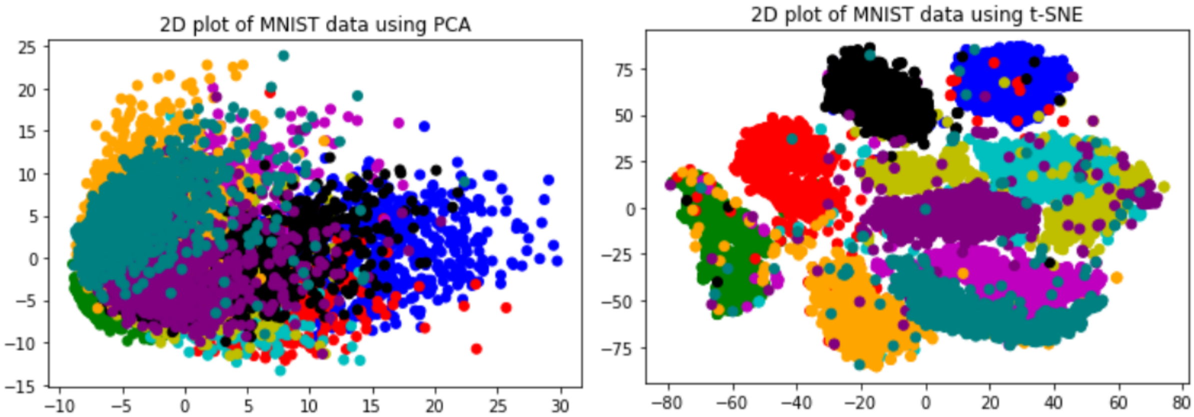


Figure 5: PCA vs t-SNE on MNIST