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

Student Name: Hemang M Khatri
Roll Number: 231110016
Date: November 17, 2023

QUESTION 1

Step 1: Assigning x_n to the "best" cluster

The assignment variable z_{nk} is a binary indicator of whether data point x_n belongs to cluster k. It is defined as:

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

Here, $\|x_n - \mu_j\|^2$ represents the squared Euclidean distance between x_n and the cluster mean μ_j. The data point x_n is assigned to the cluster with the smallest squared Euclidean distance.

Step 2: Updating the cluster means using SGD

The update equation for the cluster means using stochastic gradient descent is as follows:

$$\mu_k = \mu_k + \alpha \cdot z_{nk} \cdot (x_n - \mu_k)$$

In this equation:

μ_k : current mean of cluster k
α : learning rate
z_{nk} : assignment variable (1 if x_n belongs to cluster k, 0 otherwise)
x_n : current data point

This equation can be broken down into two parts:

1. **Assignment indicator (z_{nk}):** This term determines whether the data point x_n belongs to cluster k. If x_n is assigned to cluster k, $z_{nk} = 1$; otherwise, $z_{nk} = 0$.
2. **Update term ($(x_n - \mu_k)$):** This term represents the difference between the current data point x_n and the current cluster mean μ_k. The update is performed in the direction of this difference.

Multiplying these two terms by the learning rate (α) controls the step size of the update.

Intuition behind the update equation

The update equation is intuitive:

- If x_n belongs to cluster k ($z_{nk} = 1$), the mean μ_k is updated towards x_n.
- If x_n does not belong to cluster k ($z_{nk} = 0$), the mean μ_k remains unchanged.

This process ensures that the cluster means are adjusted to better represent the characteristics of the data, with each update reflecting the influence of the current data point.

Choice of step size (α)

The learning rate (α) determines the size of the step taken in the direction of the gradient. A common heuristic is to use a decreasing schedule for the learning rate:

$$\alpha = \frac{c}{t}$$

where c is a constant and t is the iteration number. This schedule starts with a larger learning rate and gradually decreases it over time. The rationale is to take larger steps initially for faster convergence and smaller steps later to prevent overshooting and oscillations. The specific choice of c depends on the problem and often requires experimentation.

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

Student Name: Hemang M Khatri
Roll Number: 231110016
Date: November 17, 2023

QUESTION 2

The goal is to maximize Fisher's Linear Discriminant Analysis (LDA) objective, given by the ratio of the between-class scatter to the within-class scatter:

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

Where:

- w is the projection vector,
- S_B is the between-class scatter matrix,
- S_W is the within-class scatter matrix.

The between-class scatter matrix S_B is defined as:

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$

Where:

- μ_1 and μ_2 are the means of class +1 and -1, respectively.

The within-class scatter matrix S_W is given by:

$$S_W = \sum_{i=1}^{C} \sum_{n=1}^{N_i} (x_n^i - \mu_i)(x_n^i - \mu_i)^T$$

Where:

- C is the number of classes,
- N_i is the number of samples in class i,
- x_n^i is the n-th sample in class i,
- μ_i is the mean of class i.

The total scatter matrix S_T is defined as the sum of the within-class scatter matrix and the between-class scatter matrix:

$$S_T = S_W + S_B$$

To find the optimal projection vector w, you need to solve the generalized eigenvalue problem:

$$S_B w = \lambda S_W w$$

The optimal w is the eigenvector corresponding to the largest eigenvalue λ. This w maximizes the ratio $J(w)$, leading to the desired projection that maximizes the distance between class means and minimizes the scatter within each class.

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

Student Name: Hemang M Khatri
Roll Number: 231110016
Date: November 17, 2023

QUESTION 3

The relationship between u and v is given by $X^T u = \lambda v$, where λ is the corresponding eigenvalue.

Now, let's express u in terms of X and v. Pre-multiply both sides by X:

$$XX^T u = \lambda X v$$

Now, notice that $XX^T u$ is a multiplication involving the centered data matrix X^T. To obtain u in terms of the original data matrix X, we can use the relationship $X^T = \frac{1}{N} X^T X$:

$$N \cdot \frac{1}{N} X^T X u = \lambda X v$$

Simplifying, we get:

$$X^T X u = \lambda X v$$

Now, compare this with the eigenvalue equation for S:

$$Su = \lambda u$$

It's clear that $u = Xv$ is an eigenvector of S with the same eigenvalue λ. Therefore, if someone gives you an eigenvector v of $\frac{1}{N}(XX^T)$, you can use it to obtain an eigenvector $u = Xv$ of the covariance matrix S.

The advantage of this approach lies in the fact that it allows you to work directly with the original data matrix X without explicitly forming the large covariance matrix S, making it computationally more efficient for high-dimensional datasets.

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

Student Name: Hemang M Khatri
Roll Number: 231110016
Date: November 17, 2023

QUESTION

4

(1)

The introduced model is a probabilistic mixture of linear regressions. Let $z_n \in \{1, 2, ..., K\}$ be a latent variable indicating the cluster to which x_n belongs. The model assumes a multinoulli prior on z_n: $p(z_n) = \text{multinoulli}(z_n|\boldsymbol{\pi})$, where $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)$ represents the probabilities of each cluster.

The global parameters of the model are $\Theta = \{(\mathbf{w}_1, ..., \mathbf{w}_K), (\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_K)\}$, where $\mathbf{w}_k$ is the weight vector associated with cluster k, and $\boldsymbol{\pi}_k$ is the prior probability of cluster k.

The likelihood of the response variable y_n given the latent variable and parameters is defined as:

$$p(y_n|z_n, \mathbf{x}_n, \mathbf{W}) = \mathcal{N}(y_n|\mathbf{w}_{z_n}^T \mathbf{x}_n, \beta^{-1})$$

Here, $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_K]$ is the matrix of weight vectors. The model is more expressive than the standard linear regression model, as it can capture different regression patterns ($\mathbf{w}_k$) for each cluster (z_n). This is particularly useful when the data can be naturally divided into distinct clusters with different underlying relationships between inputs and outputs.

(2)

Step 1: Update Latent Variables Z

For each data point $n = 1, ..., N$, update the latent variable z_n using Bayes' theorem:

$$p(z_n = k|x_n, y_n, \Theta) \propto \pi_k \mathcal{N}(y_n|w_k^T x_n, \beta^{-1})$$

Normalize the probabilities for each data point:

$$p(z_n|x_n, y_n, \Theta) = \frac{\pi_k \mathcal{N}(y_n|w_k^T x_n, \beta^{-1})}{\sum_{j=1}^{K} \pi_j \mathcal{N}(y_n|w_j^T x_n, \beta^{-1})}$$

Step 2: Update Weight Vectors w_k

For each cluster $k = 1, ..., K$, update the weight vector w_k using the maximum likelihood estimate:

$$w_k = \left(\sum_{n=1}^{N} \mathbb{1}[z_n = k] x_n x_n^T + \lambda I \right)^{-1} \sum_{n=1}^{N} \mathbb{1}[z_n = k] y_n x_n$$

Step 3: Update Cluster Priors π_k

Update the cluster priors using the maximum likelihood estimate:

$$\pi_k = \frac{\sum_{n=1}^{N} \mathbb{1}[z_n = k]}{N}$$

Special Case: $\pi_k = \frac{1}{K}, \forall k$

If $\pi_k = \frac{1}{K}$ for all clusters, the update equation for z_n simplifies to:

$$p(z_n = k|x_n, y_n, \Theta) \propto \mathcal{N}(y_n|w_k^T x_n, \beta^{-1})$$

Intuitive Explanation

- **Updating Latent Variables z_n:** This step computes the probability of each data point belonging to each cluster based on the prior belief and the likelihood of the observed data given the current model. It's a soft assignment of data points to clusters.
- **Updating Weight Vectors w_k:** This step refines the parameters of each cluster's regression model based on the data points assigned to that cluster. The regularization term helps prevent overfitting.
- **Updating Cluster Priors π_k:** This step adjusts the prior probabilities based on the proportion of data points currently assigned to each cluster, reflecting the responsibility of each cluster in explaining the overall dataset.

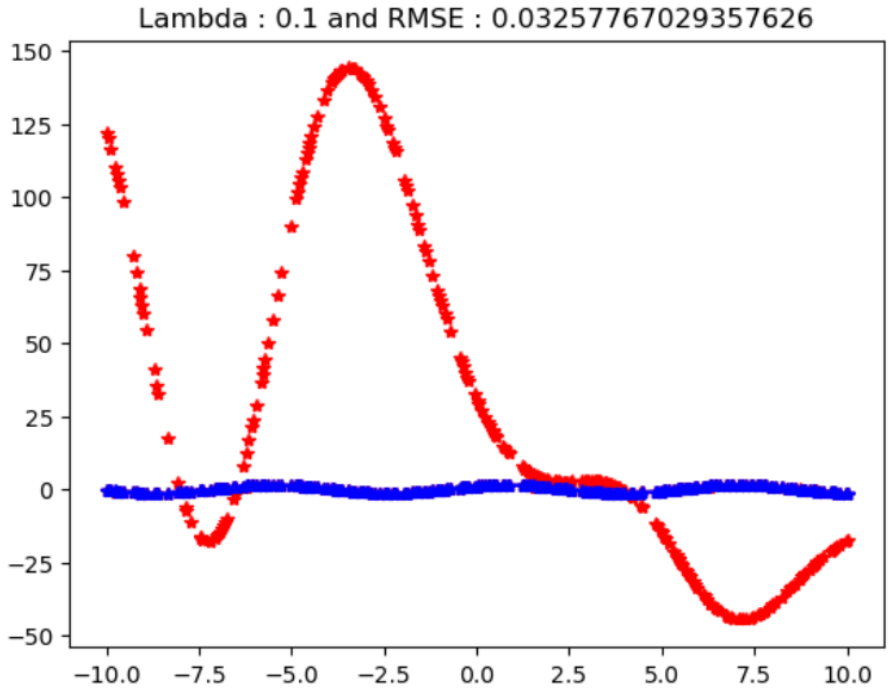
Introduction to ML (CS771), Autumn 2023
Indian Institute of Technology Kanpur
Homework Assignment Number 2

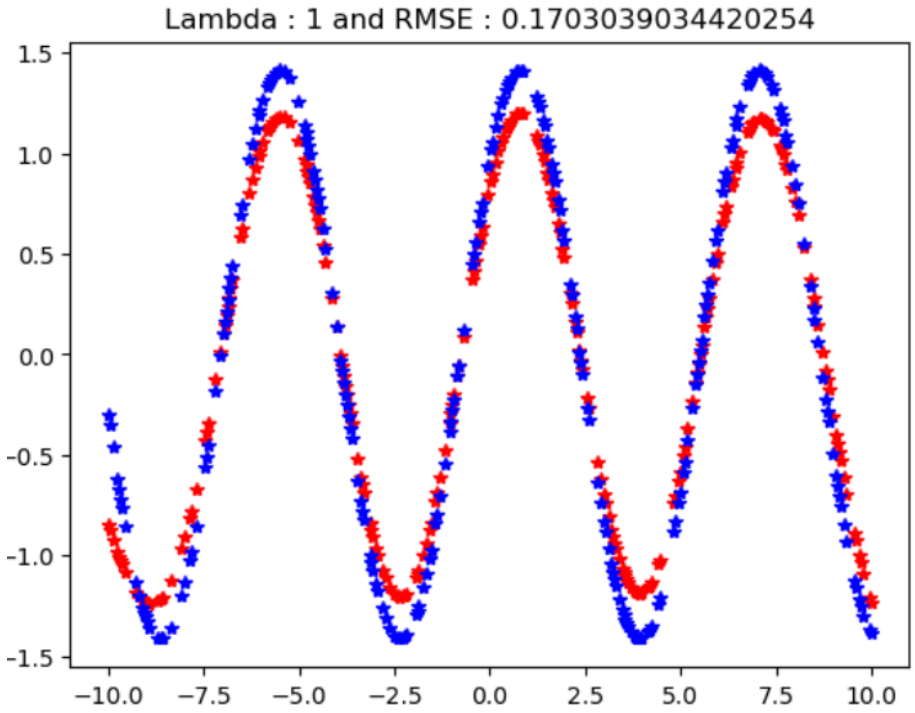
Student Name: Hemang M Khatri
Roll Number: 231110016
Date: November 17, 2023

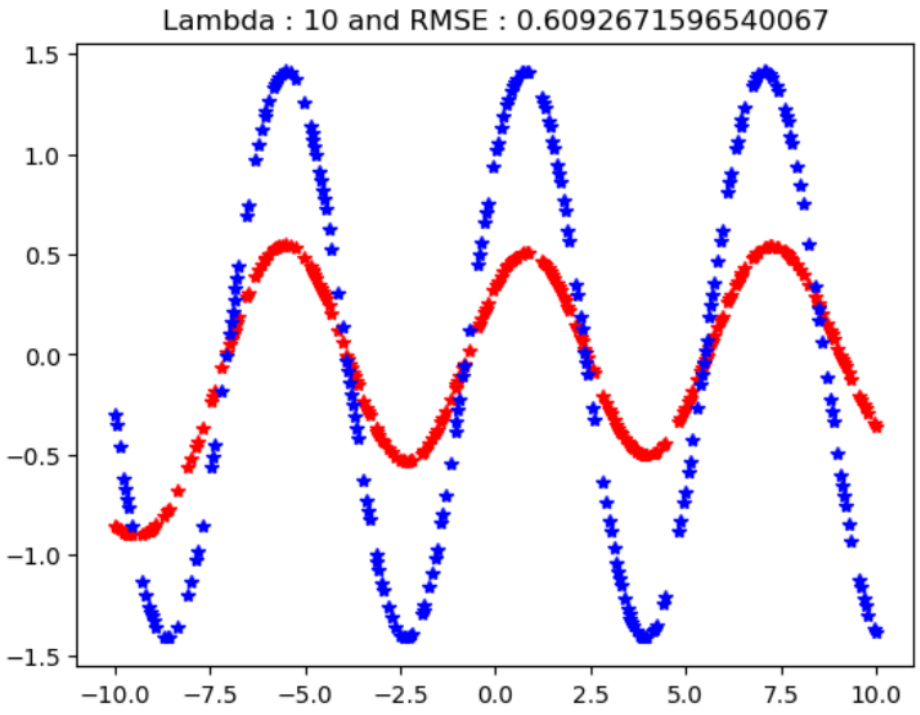
QUESTION
5

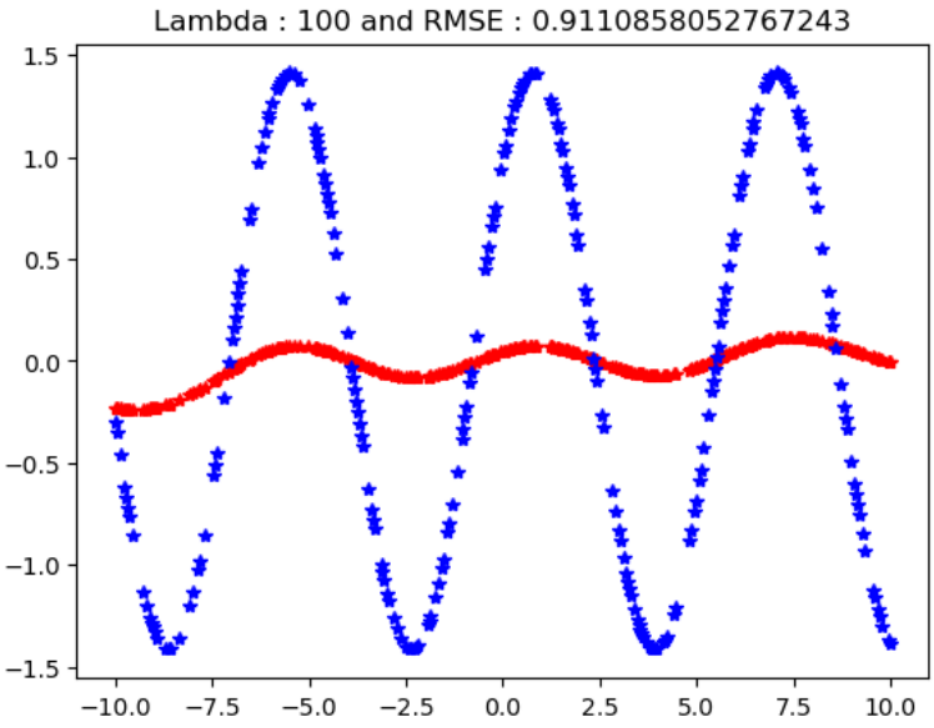
1 Part 1.1

```
RMSE for the given lambda value 0 : 65.26243117413642
RMSE for the given lambda value 0.1 : 0.03257767029357626
RMSE for the given lambda value 1 : 0.1703039034420254
RMSE for the given lambda value 10 : 0.6092671596540067
RMSE for the given lambda value 100 : 0.9110858052767243
```



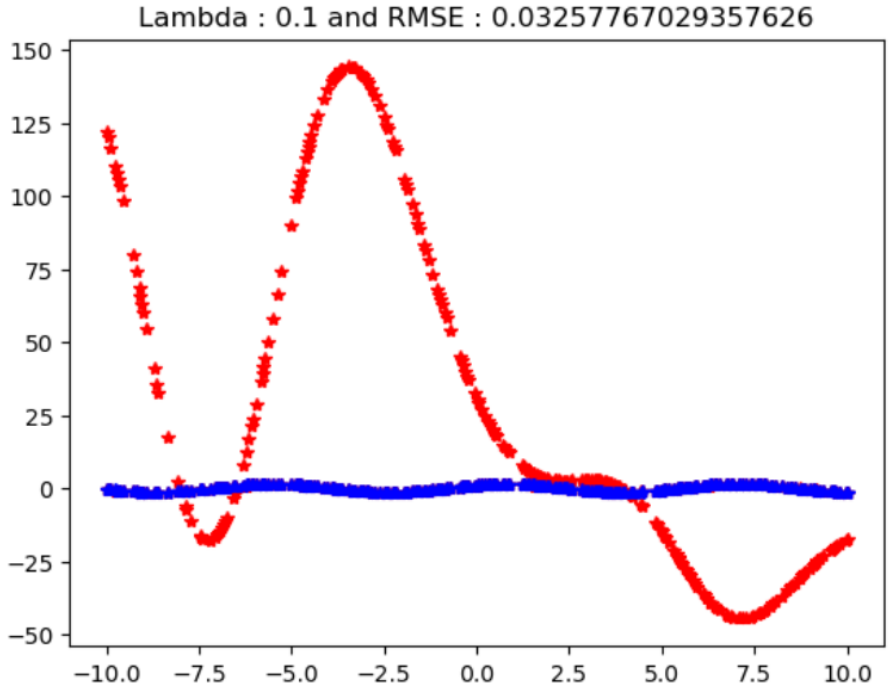

Lambda : 1 and RMSE : 0.1703039034420254
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0


Lambda : 10 and RMSE : 0.6092671596540067
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

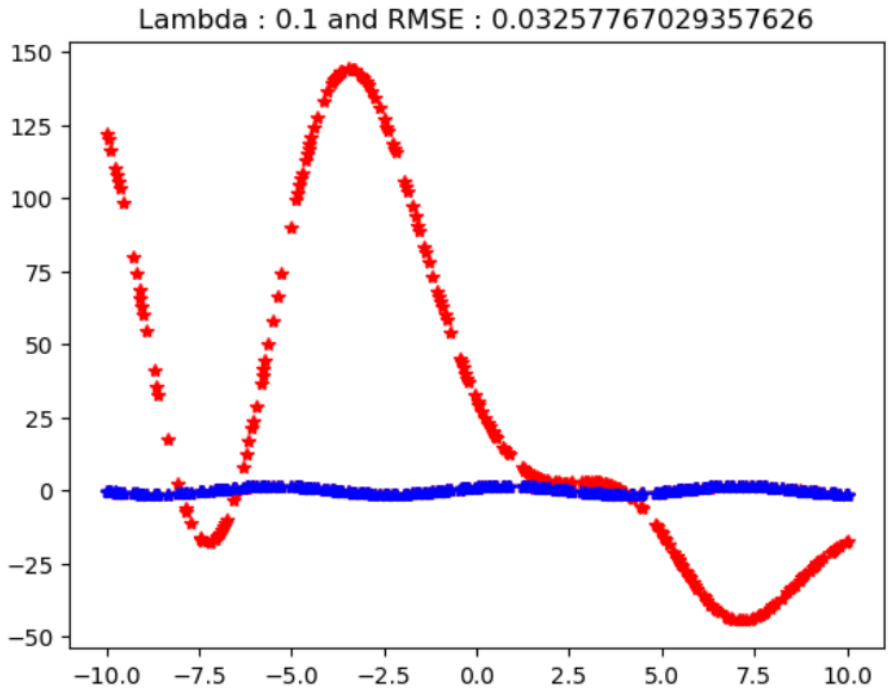

Lambda : 100 and RMSE : 0.9110858052767243
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

2 Part 1.2

```
RMSE for the given lambda value 0 : 65.26243117413642
RMSE for the given lambda value 0.1 : 0.03257767029357626
RMSE for the given lambda value 1 : 0.1703039034420254
RMSE for the given lambda value 10 : 0.6092671596540067
RMSE for the given lambda value 100 : 0.9110858052767243
```



```
RMSE for the given lambda value 0 : 65.26243117413642
RMSE for the given lambda value 0.1 : 0.03257767029357626
RMSE for the given lambda value 1 : 0.1703039034420254
RMSE for the given lambda value 10 : 0.6092671596540067
RMSE for the given lambda value 100 : 0.9110858052767243
```



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

Student Name: Hemang M Khatri
Roll Number: 231110016
Date: November 17, 2023

QUESTION 6

My solution to problem 6