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

Student Name: Shrilakshmi S K
Roll Number: 211012
Date: November 17, 2023

QUESTION 1

(1) **Assigning input points greedily to best cluster:**
Let input points be $\mathbf{x}_1, \mathbf{x}_2 ... \mathbf{x}_n$.

Assign the input to closest cluster mean. For this, we can just find the Euclidean distance of x_n from the means of all clusters and assign x_n to the cluster whose mean is the closest to x_n. Mathematically, it can be written as

$$z_n = argmin_k ||x_n - \mu_k||^2$$

Note that $z_n = k$ means $z_{nk} = 1$ in one-hot representation of z_n

(2) **SGD based cluster update:**
The iteration t goes as follows.
From the dataset, pick a random example, say $\mathbf{x}_k$. Suppose, it's closest cluster mean in previous iteration $(t-1)th$ iteration was μ_k^{t-1}

If $i = k, \mu_k^t = \mu_k^{t-1} + \eta_t(\mathbf{x}_k - \mu_k^{t-1})$ Else $\mu_k^t = \mu_k^{t-1}$

Repeat iterations until convergence. In every iteration, the cluster mean is shifting near randomly picked point $\mathbf{x}_k$.

(3) Step size should be decreasing, so η should be inversely proportional to t.

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

where k is proportionality constant.

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

Student Name: Shrilakshmi S K
Roll Number: 211012
Date: November 17, 2023

QUESTION 2

Let N_- be number of inputs with labels -1. And let N_+ be number of inputs with labels +1. Define

$$\mu_- = \frac{1}{N_-} \sum_{\mathbf{x}_i | y_i = -1} \mathbf{x}_i$$

$$\mu_+ = \frac{1}{N_+} \sum_{\mathbf{x}_i | y_i = +1} \mathbf{x}_i$$

Scalar representation of vector $\mathbf{x}$ projected along $\mathbf{w}$ where $\mathbf{w}^\top \mathbf{w} = 1$ is $\mathbf{w}^\top \mathbf{x}$

So the class mean of projected inputs are $\mathbf{w}^\top \mu_-$ and $\mathbf{w}^\top \mu_+$ for classes with $y = -1$ and $y = 1$ respectively. Distance between projected means is $d = ||\mathbf{w}^\top (\mu_- - \mu_+)||$

Let us now calculate class variances σ_- and σ_+

$$\sigma_- = \frac{1}{N_-} \sum_{\mathbf{x}_i | y_i = -1} (\mathbf{w}^\top (\mathbf{x}_i - \mu_-))^2$$

$$\sigma_+ = \frac{1}{N_+} \sum_{\mathbf{x}_i | y_i = +1} (\mathbf{w}^\top (\mathbf{x}_i - \mu_+))^2$$

Loss function to minimize the variance and maximize the distance can be as following:

$$\mathbf{L(w)} = \frac{\sigma_+ + \sigma_-}{||\mathbf{w}^\top (\mu_- - \mu_+)||}$$

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

Student Name: Shrilakshmi S K
Roll Number: 211012
Date: November 17, 2023

QUESTION 3

Dimensions of matrix $\mathbf{X}$ is $N \times D$
Dimensions of matrix $\mathbf{X}^\top\mathbf{X}$ is $D \times D$
Dimensions of matrix $\mathbf{X}\mathbf{X}^\top$ is $N \times N$

We have assumed $D > N$. So calculating eigenvectors (eigendecomposition) of $\mathbf{X}\mathbf{X}^\top$ is more efficient that calculating eigenvectors of $\mathbf{X}^\top\mathbf{X}$.

Now, let us see how can we derive eigenvectors of $\mathbf{X}^\top\mathbf{X}$ from eigenvectors of $\mathbf{X}\mathbf{X}^\top$.
Let

$$\mathbf{T} = \frac{1}{N}\mathbf{X}\mathbf{X}^\top$$

Let $\boldsymbol{v} \in \mathbf{R}^N$ be an eigenvector of $\mathbf{T}$

$$\mathbf{T}\boldsymbol{v} = \lambda\boldsymbol{v}$$

$$\frac{1}{N}\mathbf{X}\mathbf{X}^\top\boldsymbol{v} = \lambda\boldsymbol{v}$$

Multiplying by $\mathbf{X}^\top$ on the left of both LHS and RHS,

$$\frac{1}{N}\mathbf{X}^\top\mathbf{X}\mathbf{X}^\top\boldsymbol{v} = \lambda\mathbf{X}^\top\boldsymbol{v}$$

$$\frac{1}{N}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\boldsymbol{v}) = \lambda(\mathbf{X}^\top\boldsymbol{v})$$

Let $\boldsymbol{u} = \mathbf{X}^\top\boldsymbol{v}$. We can see from above equation that $\boldsymbol{u}$ is an eigenvector of $\mathbf{S} = \frac{1}{N}\mathbf{X}^\top\mathbf{X}$

Time complexity of calculating eigenvectors of $\mathbf{S}$ using this method $= O(N^2) + O(ND) = O(ND)$

Time complexity of direct eigendecomposition of $\mathbf{S} = O(D^2)$

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

Student Name: Shrilakshmi S K
Roll Number: 211012
Date: November 17, 2023

QUESTION 4

(1) Standard linear model uses a single weight vector to model the data. Above model instead uses a combination of K different weight vectors to model the data. It first divides the data into clusters and depending upon the cluster it uses one of the K weight vectors to model the data of that cluster. It is hard for single weight vector to fit all the training data into single curve effectively. Above model reduces the number of outliers because it divides the data into clusters and applies regression on different clusters with different weight vectors.

(2) ALT-OPT:

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

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

Initialize $\theta = \hat{\theta}$
Finding the optimal z_n

$$\hat{z_n} = argmax_k \frac{\pi_k N(\mathbf{w}_k^\top x_n, \beta^{-1})}{\sum_{l=1}^{K} \pi_l N(\mathbf{w}_l^\top x_n, \beta^{-1})}$$

where $N(\mathbf{w}_j^\top x_n, \beta^{-1}) = exp(-\frac{\beta}{2}(y_n - \mathbf{w}_j^\top x_n)^2)$

$$\hat{z_n} = argmax_k \frac{\pi_k exp(-\frac{\beta}{2}(y_n - \mathbf{w}_k^\top x_n)^2)}{\sum_{l=1}^{K} \pi_l exp(-\frac{\beta}{2}(y_n - \mathbf{w}_l^\top x_n)^2)}$$

Update θ

$$\hat{\theta} = argmax_\theta \log p(\mathbf{Y}, \hat{\mathbf{Z}}|\theta)$$

$$\hat{\theta} = argmax_\theta \sum_{n=1}^{N} \sum_{k=1}^{K} \hat{z_{nk}} (\log(\pi_k) + \log N(\mathbf{w}_k^\top x_n, \beta^{-1})$$

Update $\mathbf{w}_k$

$$\hat{\mathbf{w}}_k = argmax - \mathbf{w}_k \sum_{n:\hat{z}_{nk}=1} logN(\mathbf{w}_k^\top x_n, \beta^{-1})$$

$$\hat{\mathbf{w}}_k = (\mathbf{X}_k^\top \mathbf{X}_k)^{-1} \mathbf{X}_k^\top \mathbf{y}_k$$

Update π_k

$$\hat{\pi}_k = \frac{\sum_{n=1}^{N} z_{nk}}{N}$$

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

Repeat the above process until it converges.

If $\pi_k = \frac{1}{K}$, then update z_n

$$\hat{z_n} = argmax_k \frac{exp(-\frac{\beta}{2}(y_n - \mathbf{w}_k^\top x_n)^2)}{\sum_{l=1}^{K} exp(-\frac{\beta}{2}(y_n - \mathbf{w}_l^\top x_n)^2)}$$

This ensures that new $\mathbf{w}_k$ results in least square error for $(\mathbf{x}_n, y_n)$ example, which is similar behaviour to logistic regression.

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

Student Name: Shrilakshmi S K
Roll Number: 211012
Date: November 17, 2023

QUESTION 5

All plots are included in code zip file (as per Piazza post)

Part 1:
The predicted output is in red, true output is in green. The RMSEs are included in the plots.

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

From the plots, we can observe that as λ increases from 0.1 to 100, the RMSE increases. Hence we can conclude that $\lambda = 0.1$ is the best choice.

$L = 2, RMSE = 0.9725$
$L = 5, RMSE = 0.9308$
$L = 20, RMSE = 0.1790$
$L = 50, RMSE = 0.0779$
$L = 100, RMSE = 0.0794$

We can see that as the number of landmark points increase, the RMSE decreases until L=50. The decrease is also very steep. It increases slightly for L=100 again. L=50 is good enough because it has the least RMSE value.

Part 2:
The given kernel function has distance term in the exponent. This means that if the landmark is chosen near inner cluster, the clusters are separated out perfectly as inner cluster and outer cluster have significant distance between them. But if the landmark is chosen near outer cluster, inner cluster is still close to outer cluster near the chosen landmark, whereas the diametrically opposite outer cluster is far away from landmark point, Hence, the assigned clusters are not accurate.

Part 3:
The difference between PCA and tSNE plots is that PCA plots overlap, whereas tSNE has better boundaries and clusters are more accurate. In this case, tSNA is a better projection algorithm.

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

Student Name: Shrilakshmi S K
Roll Number: 211012
Date: November 17, 2023

QUESTION 6