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

Student Name: Digambar Singh
Roll Number: 200337
Date: November 17, 2023

QUESTION 1

Step 1: Assigning the Data Point x_n Greedily to the Best Cluster

Randomly select an example x_n where $n \in \{1, 2, \ldots, N\}$, assuming N is the total number of data examples. Assign x_n to the closest cluster n_k'' by finding the smallest of all distances between x_n and all cluster centers μ_k where $k \in \{1, 2, \ldots, K\}$, assuming K is the total number of clusters using the following:

$$\text{Predicted cluster for point } x_n \text{ is } C_k : C_k = \arg\min_k \|x_n - \mu_k\|^2 \tag{1}$$

Step 2: The SGD-based Cluster Mean Update Equation

We increment n_k by 1 every time a point x_n is assigned to cluster k, i.e.,

$n_k = n_k + 1$,

where n_k is the size of cluster k Ṡince we are using a single data point x_n and assigning it to a cluster k, we only need to update the k-th cluster mean as all other cluster means will remain the same. So our Loss function is as follows:

$$L = \sum_{n_k} z_{nk} \|x_n - \mu_k\|^2 \quad \text{where } z_{nk} = 1 \text{ as } x_n \text{ belongs to cluster } k \tag{2}$$

Therefore, the loss function can be rewritten as:

$$C = \sum_{n_k} \|x_n - \mu_k\|^2 \tag{3}$$

We need to minimize the loss function with respect to μ_k. Therefore, the gradient with respect to μ_k will be:

$$g_{\mu_k} = \frac{\partial L}{\partial \mu_k} = -2 \sum_{n_k} (x_n - \mu_k) = -2 \left(\sum_{n_k} (x_n - \mu_k) \right) \tag{4}$$

Therefore,

$$g_{\mu_k} = -2(\mu_k(n_k - 1) + x_n - n_k \mu_k) = -2(x_n - \mu_k) \tag{5}$$

The updated equation for SGD will be:

$$\mu_k^{t+1} = \mu_k^t - \eta_t g_{\mu_k}^t \tag{6}$$

where η_t is the learning rate at time t.

Therefore, the updated equation on substituting is:

$$\mu_k^{t+1} = \mu_k^t + \eta_t (x_n - \mu_k^t) \tag{7}$$

This update equation makes sense intuitively because the updated mean of the cluster to which x_n was assigned must consider the mean of all the n_{nk} examples present in that cluster (including point x_n). Also, the update on the mean (i.e., the cluster center) is controlled by choosing an appropriate learning rate η_t.

A good choice of learning rate will be as follows:

We should set the learning rate $\eta_t = \frac{1}{2n_k}$,

because when we set the above learning rate, the updated equation in SGD becomes:

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

The above update equation is exactly the same as obtaining the new average of all data points in the k-th cluster to which our example x_n was assigned. The learning rate is good because, as more and more data points get assigned to a cluster (i.e., the cluster size increases, n_k increases), the learning rate $\frac{1}{2n_k}$ decreases as we move closer to the exact solution of μ_k.

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

Student Name: Digambar Singh
Roll Number: 200337
Date: November 17, 2023

QUESTION 2

we need to find an objective function which reduces intra-class distance or spread and increase inter-class distance. we can do so by defining a loss function as follows:

$$\mathcal{L}(w) = \frac{w^T(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w}{w^T(\sum_{i \in C_+}(x_i - \mu_1)(x_i - \mu_1)^T + \sum_{i \in C_-}(x_i - \mu_2)(x_i - \mu_2)^T)w}$$

where:

- w is the vector along the projection direction (a unit vector).
- μ_1 and μ_2 are the means of the data points in class +1 and -1, respectively.

The goal is to optimize the ratio of the scatter between classes to the scatter within classes. This aims to find a projection direction that maximizes the difference between the means of the two classes while minimizing the spread within each class. Also, this positions the class means far apart, making the separation effective in one direction.

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

Student Name: Digambar Singh
Roll Number: 200337
Date: November 17, 2023

QUESTION 3

Given an eigenvector v of a matrix $S = \frac{1}{N}XX^T$, we know that $Sv = \lambda v$, where λ is the eigenvalue of matrix S.

$$\Rightarrow \frac{1}{N}(XX^T)v = \lambda v$$

Multiplying both sides by X^T:

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

Let us substitute X^Tv as u:

$$\Rightarrow \frac{1}{N}(XX^T)u = \lambda u$$

But $u = X^Tv$ is an eigenvector of S.

We know that to compute k eigenvalues of a matrix in the form of S, the time complexity is $O(KD^2)$. However, in this case, the time complexity will be $O(KND)$ for matrix multiplication $O(KN^2)$ for eigen decomposition of the matrix. Since $N < D$, the time complexity in this case is less than the traditional case, i.e., $O(KND) < O(KD^2)$.

The advantage is that the computation is faster with an overall time complexity of $O(KND)$.

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

Student Name: Digambar Singh
Roll Number: 200337
Date: November 17, 2023

QUESTION

4

Part 1

A standard linear model will only work for those solutions where we have to regress a linear curve, whereas this model can be a combination of K different linear curves. Basically, what the model is doing is that, at first, it is clustering the data on K different linear curves, and then the predictions are made for y. This will also help in the reduction of outliers in a linear curve, as the outliers may get separated out due to clustering.

Part 2

Here, our latent variable model becomes:

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

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

where:

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

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

ALT-OPT Algorithm

Step 1: Find the best z_n:

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

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

Step 2: Re-estimate the parameters:

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

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

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

Here, X_k is an $N_k \times D$ matrix containing training sets clustered in class k, and y_n are $N_k \times 1$ vectors containing training set labels clustered in class k.

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

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

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: Digambar Singh
Roll Number: 200337
Date: November 17, 2023

QUESTION

5

Part 1 a:

In this part, kernel ridge regression was employed to compute various RMSE values for different λ values:

RMSE score for $\lambda = 0$	: 73.96865143236798
RMSE score for $\lambda = 0.1$	: 0.03257767029357438
RMSE score for $\lambda = 1$	: 0.1703039034420251
RMSE score for $\lambda = 10$	: 0.6092671596540066
RMSE score for $\lambda = 100$	: 0.9110858052767243

Around $\lambda = 1$, it indicates a good result. To illustrate further, consider fitting all the errors.

Part 1 b:

In this part, landmark ridge regression was utilized for $\lambda = 0.1$. The obtained RMSE values for various λ values are as follows:

RMSE score for $\lambda = 2$	: 0.964892852994235
RMSE score for $\lambda = 5$	: 0.848498543561041
RMSE score for $\lambda = 20$	: 0.11661472240805554
RMSE score for $\lambda = 50$	: 0.07405337773918906
RMSE score for $\lambda = 100$	: 0.06992706556431774

Around $\lambda = 100$ and $\lambda = 50$, it indicates good results.

Part 2 a:

In this part, clustering was performed on two classes Using Hand-crafted Features, yielding favorable results:

The clustering process successfully separated the data into distinct groups, effectively distinguishing between the two classes.

Part 2 b:

In this segment, clustering was conducted within two classes using kernel methods. The process was iterated ten times with random initialization points to evaluate the impact of initialization on the results. It was observed that initializing the points effectively significantly influenced the outcomes. Notably, better initialization led to better results, underscoring the importance of proper initialization in achieving favorable clustering outcomes.

Part 3:

Principal Component Analysis (PCA) primarily concentrates on capturing the maximum variance in the dataset along principal axes, resulting in a spread-out representation. However, its effectiveness in separating different classes might be limited.

Contrarily, t-distributed Stochastic Neighbor Embedding (t-SNE) prioritizes the preservation of local structures within the data. It often yields superior separation of different classes in lower-dimensional space when compared to PCA. Particularly for intricate and non-linear relationships in the data, t-SNE tends to cluster points of the same class more distinctly than PCA. This distinction makes t-SNE plots more effective in showcasing class separations.