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

Student Name: SHIVAM MISHRA
Roll Number: 231110047
Date: November 16, 2023

QUESTION 1

The standard k-means loss function is given as:

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

Now Applying SGD K means

STEP 1: We would be taking a random example x_n at a time, and then assigning x_n greedily to the best cluster based on Euclidean distance, then will apply ALTERNATING OPTIMIZATION(ALT-OPT) technique(because of 2 variables).
Fix $\mu = \hat{\mu}$ and solve for z_n.

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

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

To perform Step 1, we need to assign a cluster to x_n using the above equation for each example x_n in the set $\{x_n\}_{n=1}^{N}$.

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

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

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

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

we will uniformly randomly choose x_n and approximate gradient g as we do in SGD at any iteration t and let $g \approx g_n$

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

Now the updated equation of mean in SGD is as follows:

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

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

η is the learning rate.

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

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, is determined based on the ratio of the sum of features of every data point within that cluster to the total number of data points in that cluster.

This step size is a good choice because the learning rate η 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.

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

Student Name: SHIVAM MISHRA
Roll Number: 231110047
Date: November 16, 2023

QUESTION 2

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

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

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

S_w is within the class covariance matrix. The numerator $\|\mathbf{w}^T \mu_+ - \mathbf{w}^T \mu_-\|^2$ 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 enhances the separation between the means of the two classes, leading to improved distinction.
2. Reducing the denominator ensures that the data points within each class are more tightly packed, diminishing overlap and enhancing the discrimination between classes.

This formulation reflects the concept of increasing the distance between the means of different classes while simultaneously bringing the data points within each class closer together.

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

Student Name: SHIVAM MISHRA
Roll Number: 231110047
Date: November 16, 2023

QUESTION 3

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

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

$$Sv = \lambda v$$

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

Now we multiply X^T on both side of the equation to get :

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

Now substitute $u = X^T v$

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

u is eigenvector of matrix S.
In a Normal scenario, computing K eigenvectors with maximum variance of matrix S takes $O(KD^2)$ time.
But in this approach time complexity is $O(KN^2) + O(KND)$

$$D > N$$

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

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

Student Name: SHIVAM MISHRA
Roll Number: 231110047
Date: November 16, 2023

QUESTION 4

(1)

This technique integrates multiple linear curves to construct a model featuring K distinct linear components, providing flexibility beyond traditional linear models that are limited to regression on a single linear curve.
Essentially, the model initiates by categorizing the data along K separate linear curves before making predictions for the output variable (y). Utilizing latent variables to assign each point to its cluster makes the method robust against outliers. This strategy proves advantageous in mitigating the impact of outliers within a linear curve, as it introduces the potential for separating outliers during the clustering process.

(2)

given ::

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

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

Latent Variable Model :

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

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

Applying the ALT-OPT algorithm

Step 1:
Find optimum z$_n$:

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

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

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

Step 2:
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 k classes and are represented by X_k which has $N_k \times D$ dimension, whereas training set labels are also clustered in k classes 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 expression is similar to softmax classification and 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: SHIVAM MISHRA
Roll Number: 231110047
Date: November 16, 2023

QUESTION

5

Programming Problem: Part 1

1. Kernel Ridge

λ	RMSE
0.1	0.032577670293574425
1	0.1703039034420253
10	0.6092671596540067
100	0.9110858052767243

Elevating the regularization hyperparameter is associated with a rise in errors. This phenomenon can be explained by the fact that both the training and test sets are derived from a shared sine curve, exhibiting minimal outliers. As regularization strength increases, the model becomes more focused on precisely fitting the training data. Consequently, this heightened emphasis on training data details may lead to overfitting, causing the model to perform poorly on the generalization task of the test set. In essence, a lower regularization value facilitates a more flexible model that captures the inherent pattern of the sine curve, resulting in a superior fit on both the training and test data.

2. Landmark Ridge

λ	RMSE
2	0.939963050174282
5	0.8869212316019458
20	0.18660312692554937
50	0.06271701217985742
100	0.06291427915461083

A smaller value of L corresponds to a higher prediction error due to the limited number of feature points considered. The choice of L=100 proves to be advantageous as observed from the plots and data.

Programming Problem: Part 2

1. Using Hand-crafted Features

The data given in the file is not linearly separable the data needs to be elevated to some higher dimension and then back to two dimensions after separating the clusters.

2. Using Kernels

We randomly select landmarks from the dataset, and due to this random process, each program run may yield different results. Our observation is that the provided data is easily separable into clusters. To leverage this property, we employ a feature transformation based on the distance of each point from the origin. Specifically, when a landmark is in close proximity to the origin, the data tends to cluster around it more effectively compared to cases where the landmark is farther away. This phenomenon is attributed to the radial dispersion of data around the origin.

Programming Problem: Part 3

Clusters in t-SNE exhibit a visually more pronounced separation compared to PCA. Furthermore, when ten different initializations are performed, t-SNE consistently demonstrates lower errors than those observed in PCA.