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

*Student Name:* Vikram Kumar
*Roll Number:* 231280003
*Date:* November 16, 2023

# QUESTION 1

The standard k-means loss function is given as:

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

For the SGD K-means:

**Step 1:**To bring it online, randomly select an example $x_n$ at each step and then assign $x_n$ to the "best" cluster in a "greedy" manner.

Using **ALT-OPT** technique

put $\mu = \hat{\mu}$ and solve for $\hat{z}_n$:

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

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

To perform Step 1, we need to assign a cluster to $x_n$ using the above equation for each of the examples $\{x_n\}$ where $n = 1, \ldots, N$.

**Step 2:** To update the cluster means.

Solving for $\mu$ using SGD by puting $z = \hat{z}$:

$$\hat{\mu} = \arg\min_{\mu} \mathcal{L}(X, \hat{Z}, \mu) = \arg\min_{\mu} \sum_{n=1}^{N} \sum_{n:\hat{z}_n = k} ||x_n - \mu_k||^2$$

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

At any iteration $t$, randomly select an example $x_n$ uniformly, and approximate $g$ as follows:

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

Now the mean can be updated as:

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

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

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, so that the updated mean would also be in ratio of the sum of features of every data point to the total number of data points in that cluster.

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

*Student Name:* Vikram Kumar
*Roll Number:* 231280003
*Date:* November 16, 2023

# QUESTION 2

The objective/loss function to achieve the described goal is typically formulated using Fisher's Linear Discriminant Analysis (LDA). LDA aims to maximize the ratio of the between-class variance to the within-class variance. The objective function can be defined as follows:

$$J(w) = \frac{w^T S_B w}{w^T S_W w} \tag{1}$$

where:

- $w$ is the projection direction vector (a column vector in $\mathbb{R}^D$),
- $S_B$ is the between-class scatter matrix,
- $S_W$ is the within-class scatter matrix.

The between-class scatter matrix $S_B$ is defined as:

$$S_B = (\mu_+ - \mu_-)(\mu_+ - \mu_-)^T \tag{2}$$

The within-class scatter matrix $S_W$ is defined as:

$$S_W = \sum_{i \in \text{class } +} (x_i - \mu_+)(x_i - \mu_+)^T + \sum_{i \in \text{class } -} (x_i - \mu_-)(x_i - \mu_-)^T \tag{3}$$

where:

- $\mu_+$ and $\mu_-$ are the means of the positive and negative class samples, respectively,
- $x_i$ represents individual samples in the respective classes.

The objective function $J(w)$is designed to maximize the separation between the means of different classes in the projected space while minimizing the scatter within each class.

Maximizing this ratio results in a projection where the distance between the means of the classes is maximized, and the scatter within each class is minimized, achieving the desired goal of separating the classes in the one-dimensional projection.

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

*Student Name:* Vikram Kumar
*Roll Number:* 231280003
*Date:* November 16, 2023

# QUESTION 3

Let's assume we have a eigen vector v of matrix

$$S = \frac{1}{N} X X^T$$

It will satisfy $Sv = \lambda v$

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

where $\lambda$ is eigen value
multiply both side by $X^T$

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

put $X^T v = u$

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

$$Su = \lambda u$$

$u = X^T v$ is nothing else but an eigenvector for $\frac{1}{N} XX^T$

for the normal case complexity to compute k eigen vectors for matrix $S = \frac{1}{N} XX^T$ is $O(KD^2)$, and for the given case (D >N) it will be $O(KD^2) + O(KND) = O(KND)$ which is better than the normal case $O(KD^2)$

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

*Student Name:* Vikram Kumar
*Roll Number:* 231280003
*Date:* November 16, 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 it first clusters the data into K different linear curves, and then makes predictions for y. This will also help in the reduction of outliers in a linear curve, as the outliers may get separated due to clustering.

Part 2 and 3: Here, our latent variable model becomes:

$$p(z_n = k|y_n, \theta) = \frac{p(z_n = k)p(y_n|z_n = k, \theta)}{\sum_{l=1}^{K} p(z_n = l)p(y_n|z_n = l, \theta)} \tag{25}$$

$$p(y_n, z_n|\theta) = p(y_n|z_n, \theta)p(z_n|\theta) \tag{26}$$

where:

$$p(z_n = k) = \pi_k \tag{27}$$

$$p(y_n|z_n, \theta) = \mathcal{N}(w_{z_n}^T x_n, \beta^{-1}) \tag{28}$$

**ALT-OPT Algorithm:**

Step 1: Find the best $z_n$:

$$z_n = \arg\max_{z_n} \frac{\pi_k \mathcal{N}(w_{z_n}^T x_n, \beta^{-1})}{\sum_{l=1}^{K} \pi_l \mathcal{N}(w_l^T x_n, \beta^{-1})} \tag{29}$$

$$\Rightarrow z_n = \arg\max_{z_n} \frac{\pi_k \exp\left(-\frac{\beta}{2}(y_n - w_{z_n}^T x_n)^2\right)}{\sum_{l=1}^{K} \pi_l \exp\left(-\frac{\beta}{2}(y_n - w_l^T x_n)^2\right)} \tag{30}$$

Step 2: Re-estimate the parameters:

$$N_k = \sum_{n=1}^{N} z_{nk} \tag{31}$$

$$w_k = (X_k^T X_k)^{-1} X_k^T y_k \tag{32}$$

$$\pi_k = \frac{N_k}{N} \tag{33}$$

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

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

$$z_n = \arg\max_{z_n} \frac{\exp\left(-\frac{\beta}{2}(y_n - w_{z_n}^T x_n)^2\right)}{\sum_{l=1}^{K} \exp\left(-\frac{\beta}{2}(y_n - w_l^T x_n)^2\right)} \tag{34}$$

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:* Vikram Kumar
*Roll Number:* 231280003
*Date:* November 16, 2023

**QUESTION**

# 5

**Problem 1:**

**Part (1)** Increasing the regularization hyperparameter leads to a rise in error. This could be attributed to both the training and test sets being drawn from the same sine curve with minimal outliers. Consequently, lower regularization results in a more accurate fit to the training data, and by extension, a more favorable fit to the test data.

**Part (2)** A lower value of $L$ corresponds to increased prediction error, given the fewer feature points considered. An $L$ value of 50 proves sufficient, as raising it to 100 only marginally alters the root mean square error (RMSE) by 0.0121.

**Problem 2:**

**Part (1)** Upon visualizing the data, it becomes evident that the clusters are arranged radially around the origin. To address this, a handcrafted feature transformation was employed, representing the distance from the origin.

**Part (2)** Utilizing a feature transformation based on the point's distance from the origin facilitates easy clustering of the provided data. The effectiveness of a chosen landmark in clustering depends on its proximity to the origin; a closer landmark tends to cluster well, whereas a distant one may not perform as effectively, given the radial distribution of the data around the origin. Note that landmark points are highlighted in red with a star marker, while blue points may not be visible when the landmark belongs to the blue clusters.

**Problem 3:** In the context of the provided dataset, t-SNE provide better performance compared to PCA in clustering. It offers a more efficient representation of local structures and achieves a distinct separation between clusters.