5️⃣

# Assignment - 5

Pradyoth Singenahalli Prabhu – 02071847

Woodi Raghavendra Varun – 02070206

Bharath Anand – 02044023

## Question - 1

a.

```
Kernel: linear, Training size: 5, Accuracy: 0.7252631578947368
Kernel: poly, Training size: 5, Accuracy: 0.3989473684210526
Kernel: rbf, Training size: 5, Accuracy: 0.19052631578947368
Kernel: linear, Training size: 10, Accuracy: 0.7666666666666667
Kernel: poly, Training size: 10, Accuracy: 0.6
Kernel: rbf, Training size: 10, Accuracy: 0.21555555555555556
Kernel: linear, Training size: 15, Accuracy: 0.8047058823529412
Kernel: poly, Training size: 15, Accuracy: 0.7470588235294118
Kernel: rbf, Training size: 15, Accuracy: 0.25176470588235295
```

b.

```
k: 1, Training size: 5, Accuracy: 0.6168421052631579
k: 3, Training size: 5, Accuracy: 0.5631578947368421
k: 5, Training size: 5, Accuracy: 0.49473684210526314
k: 1, Training size: 10, Accuracy: 0.7155555555555555
k: 3, Training size: 10, Accuracy: 0.6866666666666666
k: 5, Training size: 10, Accuracy: 0.6422222222222222
k: 1, Training size: 15, Accuracy: 0.7223529411764706
k: 3, Training size: 15, Accuracy: 0.7176470588235294
k: 5, Training size: 15, Accuracy: 0.7011764705882353
```

## Question - 2

Assigns $x1$ to cluster 1 and all other points to cluster 2.

Apply K-means algorithm:

## 1. Calculating the centroid:

**Cluster 1:**

$$\mu_1 = x_1 = (0, 16)$$

**Cluster 2:**

$$\mu_2 = \left(\frac{(0-4+0)}{3}, \frac{(9+0+4)}{3}\right)$$

$$\mu_2 = \left(\frac{-4}{3}, \frac{13}{3}\right)$$

## 2. Using Euclidean distance to calculate the closest point

- Calculate the Euclidean distance between each point and the two centroids.
- Assign each point to the cluster with the centroid that is closer.

For example,

- x1 is closer to $\mu$1, so it is assigned to cluster 1.
- x2 is closer to $\mu$2, so it is assigned to cluster 2.
- x3 is closer to $\mu$2, so it is assigned to cluster 2.
- x4 is closer to $\mu$1, so it is assigned to cluster 1.

## 3. Calculate the centroids again, until convergence:

**Cluster 1:**

$$\mu_1 = \left(\frac{(0-4)}{2}, \frac{(16+0)}{2}\right)$$
$$\mu_1 = (-2, -8)$$

**Cluster 2:**

$$\mu_1 = \left(\frac{(0+0}{1}, \frac{9+4}{2}\right)$$
$$\mu_1 = (0, 6.5)$$

## 4. Finding the Closest Points using Euclidean Distance After calculating the centroids in the previous step, we use the

**Euclidean distance metric to determine the closest points to each centroid. Following are the assignments of points to clusters based on their proximity to the centroids:**

- x1 is assigned to cluster 1 as it belongs to $\mu$1
- x2 is assigned to cluster 2 as it is closer to $\mu$2
- x3 is assigned to cluster 2 as it is closer to $\mu$2
- x4 is assigned to cluster 1 as it is closer to $\mu$1

Upon observing the above assignments, we can see that there is no change in the assignments from the previous iteration. Hence, we have achieved convergence.

## Q. What will be the final assignment of points to clusters computed by Kmeans?

**Cluster 1:**

$$x_1 = (0, 16)$$
$$x_4 = (4, 0)$$

**Cluster 2:**

$$x_2 = (0, 9)$$
$$x_3 = (-4, 0)$$

## Q. In this case, does Kmeans find the assignment of points to clusters that minimizes the following Equation?

Indeed, the k-means algorithm aims to minimize the following equation by finding the optimal assignment of points to clusters.

The assignment of x1 is to cluster 1, with the centroid $\mu$1.

$$distance = 0$$

x2 is assigned to cluster 2 and centroid $\mu$2,

$$\begin{aligned}\text{distance} &= \sqrt{(0,0)^2 + (9-6.5)^2} \\ &= 2.5 \\ &= (2.5)^2 \\ &= 6.25\end{aligned}$$

x3 is assigned to cluster 2 and centroid $\mu$2,

$$\begin{aligned}\text{distance} &= \sqrt{(0+4)^2 + (16-0)^2} \\ &= 16.97 \\ &= (16.97)^2 \\ &= 287.96\end{aligned}$$

x4 is assigned to cluster 1 and centroid $\mu$1,

$$\begin{aligned}\text{distance} &= \sqrt{(0-0)^2 + (4-6.5)^2} \\ &= 2.5 \\ &= (2.5)^2 \\ &= 6.25\end{aligned}$$

Therefore, sum of squared distance is:

$$\begin{aligned}&= 0 + 6.25 + 287.96 + 6.25 \\ &= 300.46 \\ &= \frac{300.46}{4} \\ &= 75.115\end{aligned}$$

# Question - 3

a.

```
K-means Clustering Results:
Average Acc: 0.31728887181175935, Std Acc: 0.02397677401492886
Average NMI: 0.4785060090771195, Std NMI: 0.02636470160554388
Average Running Time: 0.14132413864135743, Std Running Time: 0.02686845716306815
```

b.

```
Hierarchical Clustering (Single):
Acc: 1.5502612709029693e-05
NMI: 0.01752436468371601
Running Time: 0.07271289825439453

Hierarchical Clustering (Complete):
Acc: 0.21759559613437726
NMI: 0.42646243629848035
Running Time: 0.049822092056274414

Hierarchical Clustering (Weighted):
Acc: 0.19564417810688814
NMI: 0.40202703094583164
Running Time: 0.04734921455383301

Hierarchical Clustering (Ward):
Acc: 0.3780245398844174
NMI: 0.5531500593557892
Running Time: 0.04751276969909668
```

# Conclusion:

1. The task is to classify hand-written digit images in the USPS dataset using SVM and kNN classification methods. For SVM, the experiment uses 1K images of 10 digits, with 3 different kernels (linear, polynomial, and RBF) and 3 different training settings, resulting in 9 different results. For kNN, the experiment uses the same dataset with 3 different training settings and k values of 1, 3, and 5, also resulting in 9 experimental results. The goal is to compare the performance of SVM and kNN for hand-written digit classification in the USPS dataset, using different training settings and parameter configurations.

2. The task is to apply K-Means algorithm to cluster a set of points into two clusters. The initial assignment of points to clusters is random, and the algorithm takes several steps to find the final cluster assignments based on the centroids of the clusters and Euclidean distance as the distance metric. In this case, the initial assignment of $x$1 to cluster 1 and all other points to cluster 2 results in the final cluster assignment with $x$1 in cluster 1 and all other points in cluster 2. However, this assignment may not necessarily minimize the given equation since the initialization of K-Means is random and can lead to different local optima.

3. The task involves evaluating clustering accuracy, NMI, and running time for Kmeans and Hierarchical Clustering on the USPS dataset. Kmeans will be

evaluated for clustering accuracy, NMI, and end-to-end running time for 10 clusters, with five trials and reporting of average and standard deviation for each. Hierarchical Clustering will be evaluated for clustering accuracy, NMI, and end-to-end running time for 10 clusters using four different criteria: single (Min), complete (Max), weighted, and ward. Therefore, a total of 12 different results will be reported to compare the performance of Kmeans and Hierarchical Clustering on the USPS dataset.