# CSE552 – Machine Learning (Spring 2020)

# Homework #4

**Due**: May 17, 2022.

**Hand-in Policy**: Via Teams. No late submissions will be accepted.
**Collaboration Policy**: No collaboration is permitted.
**Grading**: This homework will be graded on the scale 100.

**Description**: The aim of this homework is to explore clustering techniques. Use the following data for testing your implementation: (MNIST Digit Recognitions Data – available through `mnist.load_data()` in Keras).

### Part I: Apply PCA to MNIST Data

Use the function **`pca(X)`** from Homework 3 or from an existing library. Recall that this function takes an $n \times d$ matrix $\boldsymbol{X}$ and returns **`mean`**, **`weights`** and **`vectors`**. $\boldsymbol{X}$ has in each row the pixel of an input image. The mean is the mean of the columns of X. The principal components of X are in **`vectors`**. The corresponding eigenvalues are in **`weights`**. Using only several components, obtain a new data matrix $\boldsymbol{X}'$. Use this new matrix $\boldsymbol{X}'$ in the next part.

### Part II: Clustering

Use an existing k-means algorithm with three different distance metric: 1) L2 norm (Euclidean distance), 2) L1 norm (Manhattan distance), and 3) Cosine distance.

Using the transformed data apply k-means algorithm (use k=10 for ten digits) to cluster 80% of the data and test the result on the remaining 20% of the data (repeat this 5 times for cross validation). Report the performance of the clustering using the following measurement.

- Labeling of clusters:
  - Using the given labels for the training data form the following table:

| | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ | $C_7$ | $C_8$ | $C_9$ | $C_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Label 0 | $n_{1,0}$ | $n_{2,0}$ | $n_{3,0}$ | | | | | | | |
| Label 1 | | | | | | | | | | |
| Label 2 | | | | | | | | | | |
| Label 3 | | | | | | | | | | |
| Label 4 | | | | | $n_{5,4}$ | | | | | |
| Label 5 | | | | | | | | | | |
| Label 6 | | | | | | | | | | |
| Label 7 | | | | | | | | | | |
| Label 8 | | | | | | | | | | |
| Label 9 | | | | | | | | | | $n_{10,9}$ |

    Where $n_{i,j}$ indicates how many of the training data with label j falls into the cluster i.

  - Find the maximum $n_{i,j}$ in the table and label cluster $i$ with label $j$. Find the next maximum $n_{i,j}$ and if cluster $i$ is not already labeled or label j is not yet assigned, label it with $j$. Otherwise move to the next maximum $n_{i,j}$ and label if not already labeled or the label is not yet assigned. Repeat this until all the clusters are labeled.

For example, the following incomplete table of clustering result will have the given labels.

| | $C_1$ | $C_2$ | $C_3$ | $C_4$<br>L0 | $C_5$ | $C_6$ | $C_7$<br>L2 | $C_8$<br>L1 | $C_9$ | $C_{10}$<br>L3 |
|---|---|---|---|---|---|---|---|---|---|---|
| Label 0 | 0 | 0 | 100 | 300 | 100 | 100 | 0 | 0 | 0 | 0 |
| Label 1 | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 400 | 100 | 0 |
| Label 2 | 190 | 0 | 0 | 100 | 0 | 0 | 310 | 0 | 0 | 0 |
| Label 3 | 100 | 100 | 100 | 0 | 140 | 0 | 0 | 0 | 0 | 160 |

The maximum 400 will label cluster 8 as label 1. The next maximum 310 will label cluster 7 as label 2. The next maximum 300 will label cluster 4 as label 0. The next maximum 190 will not label cluster 1 as label 2 since label 2 is already assigned. The next maximum 160 will label cluster 10 as label 3.

- Training error:
    - ○ Once the clusters are labeled, for each training data, construct the confusion matrix and calculate the accuracy.
- Test error:
    - ○ For the test data, use 1-nn to decide which cluster the data will fall into. And construct the confusion matrix and calculate the accuracy.

**What to hand in:** You are expected to hand in one of the following

**HW5_lastname_firstname_studentnumber_code.ipynb**. Your notebook should have:

**Part I: Code**

Results:

Conclusions:

**Part II: Code**

Results:

Conclusions: