# DMML - Assignment 3

Alena Maria Thomas - MDS202303
Ananya Kaushal - MDS202306

April 21, 2024

## Overview

In class, we saw an example of using clustering for semi-supervised learning of the MNIST dataset, where we used K-Means clustering to identify a small subset of labelled images to seed the classification process.

The task is to conduct a similar experiment with the following two datasets.

1. The Fashion MNIST dataset for which you must build a neural network (multi-layer perceptron, MLP) model.
2. The Overhead MNIST dataset for which you can find a standard neural network (multi-layer perceptron, MLP) model here.

The MNIST example started with 50 clusters. Experiment with different (relatively small) values of K for these two datasets.

## Task 1 - Fashion MNIST dataset

### Implementation

1. Loaded the dataset using the `keras` package.
2. Preprocessed the dataset for training.
   - Normalized the inputs so that the pixel values are in the $[0-1]$ range, rather than $[0-255]$
3. Defined the following functions:
   (a) `KMeans_seed_set(X, y, k)`: Performs K-Means clustering for the specified dataset (X, y) and number of clusters (k) and returns a smaller representative labeled subset.
      i. Each element of the dataset is a 28x28 grayscale image. Since the input of clustering must be 1D vectors, we reshaped the input so that each 28x28 image becomes a single 784 dimensional vector.
      ii. Fit the K-Means on the scaled dataset and obtained the centroids and cluster labels.
      iii. For each cluster, identified the image closest to the centroid of that cluster, and collected all such images to obtain a small representative labeled subset of the dataset. This smaller dataset will be used to seed the classification process using MLP model and is returned.
   (b) `create_mlp_model()`: Returns a simple MLP model
   (c) `train_mlp_model(model, X, y`:
      i. One-hot encoded the target variable since it is categorical (takes values 0-9 for the 10 possible classes in the dataset).
      ii. Fits the provided MLP model on the training dataset (X, y) and returns the history object.
4. Performed K-Means clustering to obtain a small representative labeled subset of the training dataset.
5. Built an MLP model and fit it on the above obtained small representative dataset.
6. Evaluated the performance of the MLP model on the original test dataset.
7. Repeated steps 4-6 for different values of $K$ and compared the performance.

### Results

| Number of clusters (K) | Training accuracy | Testing accuracy |
|---|---|---|
| 10 | 0.9234 | 0.8962 |
| 20 | 0.9231 | 0.8949 |
| 30 | 0.9228 | 0.8954 |
| 40 | 0.9240 | 0.8941 |
| 50 | 0.9237 | 0.8939 |

## Task 2 - Overhead MNIST dataset

### Implementation

We repeated the same implementation as performed for Task 1.

### Results

| Number of clusters (K) | Training accuracy | Testing accuracy |
|---|---|---|
| 10 | 0.7264 | 0.4329 |
| 20 | 0.7264 | 0.4329 |
| 30 | 0.7264 | 0.4329 |
| 40 | 0.7264 | 0.4329 |
| 50 | 0.7264 | 0.4329 |