# **HW Assignment 7: Unsupervised Learning (Clustering)** CS6140: Machine Learning Spring 2023

Due Date: Sunday, April 2, 2023  
**(10 points)**

# MNIST database - Wikipedia

# <https://en.wikipedia.org/wiki/MNIST_database#:~:text=The%20MNIST%20database%20(Modified%20National,the%20field%20of%20machine%20learning>.

# Scenario

The second module of this course relates to Unsupervised Learning. The central idea is to gain insights from the data when the response variable is absent, i.e., only from the feature space. One can use Unsupervised Learning for grouping rows to find homogeneity or columns to tackle non-orthogonality. The techniques for the former are called clustering, and for the latter, called dimension reduction.

This assignment is about using various Clustering algorithms on a popular data called MNIST. Clustering is a group of machine learning techniques that involve grouping similar data observations (rows) together based on their characteristics. This allows for easier analysis and can help identify patterns (and outliers) in the data. Clustering algorithms use different techniques to identify clusters in the data. This may involve identifying points that are close together in space, or looking for similarities in the attributes of the data points. Clustering algorithms often use distance measures to determine how similar or dissimilar data points are from each other. The choice of features used to cluster the data can have a big impact on the results. Careful feature selection can help ensure that the clusters are meaningful and useful for analysis. As discussed in the class, clustering techniques can produce biased results due to the selection of the algorithm and the features.

# Tasks

You are given a sample program to import, visualize and manipulate (add noise) to the MNIST data. Use that program as the starting point to import and explore the data. You are asked to cluster the digits data in two different ways:

* On the original data
* On the noisy data. You have a choice of adding noise either using normal distribution or uniform distribution.

You are asked to use k-means, mini-batch k-means, DBSCAN and HDBSCAN techniques for this assignment. You can ignore the Hierarchical clustering for the assignment.

**Here is a list of minimum essential steps for the assignment:**

1. Import and explore MNIST data
2. Standardize the data. Since all columns are on a scale of 0 to 255, you may simply divide the columns by 255 to bring them to scale.
   1. *There is no need to split into test and train, only for this assignment.*
3. Identify outliers and create their own cluster.
4. Filter out outliers and perform clustering using the above techniques.
5. Choose 10 cluster solutions for all the above techniques, because of the obvious nature of the data.
6. Save the clusters and compare them with the actual response.
7. Tune the algorithms by repeating steps 5 and 6.
8. Comment on the best outcomes and the associated algorithm(s).

# **Expected Output**

**Please submit a fully executed jupyter notebook identifying question number and steps. Make sure to add comments to your solution.**

**Accompanying programs:**

* P9\_Clustering\_Algorithms.ipynb
* P9\_MNIST\_data.ipynb