HW1 - Advanced Machine Learning Report

In this assignment, we explored various clustering, dimensionality reduction & classification techniques on the MNIST dataset. We started the assignment with a short EDA to assure the dataset is balanced in terms of the different classes.

1. **Dimensionality Reduction**: We interchanged Q1 and Q2 so we can first use dimensionality reduction to visualize the clusters. Initially, we transformed the data's 784 features into the first 2 components of dimensional reduction to produce a visual representation. To enable unlabelled image visualization, we replaced the figure markers with the original images. Each dimensional reduction technique listed below was succeeded by a random forest classifier and was evaluated using the test data. Prior to implementing the dimensional reduction, we evaluated several benchmarks, including the Dummy Classifier and Random Forest, which achieved 10% and 97% accuracy (supervised, with the entire feature space), respectively.
    a. T-SNE - the visual representation was best, but in terms of the classifier performance the accuracy was 89%, due to that we had to imitate the transformation of the test set with Nearest Neighbors Algorithm (which is the unsupervised learner for neighbor search) to approximate the corresponding T-SNE values of a given observation.
    b. PCA - The visual representation was unclear, but the Random Forest classifier achieved a 91% accuracy.
    c. Isomap - visual representation was slightly better than PCA but accuracy was only 49%.
    d. LDA - visual representation was good and accuracy was 91% (similar to PCA), little bit of surprising because LDA is a supervised approach.

    Since T-SNE provided the most visually distinct separation of the true clusters (despite the distorted distances), we chose it to evaluate the clustering algorithms.

2. **Clustering**: In this section, we decided to use an unsupervised approach and create the labels for the classifier using the clustering results (as in unsupervised tasks). To accomplish this, we displayed the images from each cluster on the T-SNE embedded space.
    a. K-means - the algorithm didn't converge with a clear result in terms of clusters and the performance of the classifier was aligned with the visual representation (54% accuracy).
    b. DBSCAN - the algorithm failed to produce a result with more than 2 clusters (even with different hyperparameters), hence there was no need to train a classifier as the results were bound to be poor.
    c. Agglomerative clustering - formed the most distinguished clusters and after combining small clusters into a larger cluster (e.g. for the digit 1 two clusters were formed, a regular one and a tilted one), the accuracy reached 75%.
    d. Gaussian Mixture Model - had a mixed result in terms of visual representation of the cluster. The performance of the classifier was 58% (slightly better than K-means.

    The clustering techniques struggled to accurately identify the digits 4, 5, and 8, which were the most challenging to classify due to their similarity (in pixel representation).

3. **Classification**:
    We experimented with the SVM classifier to classify the images to digits, the classifier performance was impressive, even without tuning we reached accuracy of 98%. Exploring different parameters with grid search didn't improve the results.