Submitted by:

Ketaki Sahasrabudhe MDS202422
Sampriti Mahapatra MDS202433

Report on DMML Assignment 3: Semi-Supervised Learning with Clustering

Objective

This assignment explores how to enhance classification performance in scenarios with limited labeled data using **semi-supervised learning** techniques. Specifically, it leverages **K-Means clustering** to propagate labels to unlabeled samples and assesses different propagation strategies.

Dataset and Setup

- **Fashion MNIST** and **Overhead MNIST (OMNIST)** datasets were used.
- All images were normalized to a [0, 1] pixel range.
- A small set of labeled examples was used to bootstrap learning.
- **K-Means** clustering with varying numbers of clusters (`k = 50, 100, 200`) was applied.
- Three label propagation strategies were evaluated:
    - **Representatives Only**: Only cluster centers are labeled.
    - **Full Propagation**: All samples in a cluster inherit the label of the closest representative.
    - **Partial Propagation**: Labels are propagated to a subset (likely based on confidence or proximity).

Fashion MNIST Results

Accuracy vs. Number of Clusters

Number of Clusters (k)	Representatives Only	Full Propagation	Partial Propagation
50	~0.675	~0.672	~0.671
100	~0.700	~0.698	~0.695
200	~0.760	~0.761	~0.748

Observations

- All three strategies showed **consistent improvements** in accuracy as `k` increased.
- **Full propagation slightly outperformed** other methods at `k=200`.
- **Partial propagation lagged behind** both other methods slightly at all cluster sizes.

Overhead MNIST (OMNIST) Results

Accuracy vs. Number of Clusters

Number of Clusters (k)	Representatives Only	Full Propagation	Partial Propagation
50	~0.245	~0.270	~0.450
100	~0.325	~0.320	~0.440
200	~0.375	~0.335	~0.447

Observations

- **Partial propagation significantly outperformed** the other two methods across all cluster sizes.
- **Representatives only** showed improvement with increasing k, but accuracy remained modest.
- **Full propagation** was consistently lower than partial propagation and similar to or slightly below the representatives-only approach.

Conclusions

- **Label propagation using clustering** is effective in low-label settings.
- For **Fashion MNIST**, **full propagation** at high cluster granularity (`k=200`) provides the best results.
- For **OMNIST**, **partial propagation** clearly leads in performance, suggesting the nature of the data benefits from more conservative or selective labeling.
- The optimal label propagation strategy may **depend on dataset complexity and intra-class variance**.