Shriram Rajasekar
Alex Miller
Kenan Stredic

Exploring Semi-Supervised Learning Techniques for Digit Recognition using MNIST Dataset

Project Overview:

The project aims to explore the effectiveness of semi-supervised learning techniques, specifically the Expectation-Maximization (EM) algorithm, in digit recognition using the MNIST dataset. The MNIST dataset is a collection of handwritten digits widely used for training various image processing systems. We propose to preprocess the dataset, apply Principal Component Analysis (PCA) for feature extraction, and then implement the EM algorithm for semi-supervised learning to classify the digits. By leveraging both labeled and unlabeled data, we anticipate achieving improved accuracy in digit classification compared to purely supervised approaches. Additionally, the project will provide insights into the performance of semi-supervised learning techniques in image classification tasks.

Dataset Description:

The MNIST dataset consists of grayscale images of handwritten digits (0-9), each image being 28x28 pixels. The dataset includes a total of 70,000 images, divided into 60,000 training samples and 10,000 test samples. Each pixel intensity ranges from 0 to 255, representing the darkness of the pixel.

Dataset Variables:

• Image Pixel Values: Numeric (integers representing pixel intensities ranging from 0 to 255)
• Label: Categorical (integers from 0 to 9, representing the digit each image represents)

Project Workflow:

1. Data Preprocessing:
    - Convert images into numerical data, representing pixel intensities.

2. Feature Extraction:
    - Apply Principal Component Analysis (PCA) to identify the most important features of the image data.

3. Labeling:
    - Manually label a subset of the data with their corresponding digit to serve as the initial training set.

4. EM Algorithm for Semi-Supervised Learning:

- E-step: Use a Gaussian Mixture Model to estimate the posterior probability of each unlabeled image for every digit label.
- M-step: Update the parameters of the Gaussian Mixture Model.

5. Thresholding:

- Assign a label to an unlabeled image only if the probability of the image belonging to that label is above a certain threshold.

6. Model Training:

- Combine the manually labeled dataset with the images labeled by the EM algorithm.
- Train a classification model using this combined dataset.

Dataset Link:
https://www.kaggle.com/datasets/oddrationale/mnist-in-csv