DMML ASSIGNMENT(3) REPORT

ESHA BHATTACHARYA(MDS202324)
MALDE DHARMI(MDS202333)

CLUSTERING & CLASSIFICATION ON THE FASHION MNIST DATASET:

INTRODUCTION:

Here we are using clustering techniques to group similar images together based on their pixel intensities. By clustering, we aim to assign labels to entire clusters rather than individual images. This approach enables us to streamline the labeling process, as labeling all 60,000 images individually would be impractical.

Following the clustering phase, the labeled clusters serve as training data for a multi-layered perceptron (MLP) classification model. The MLP is designed to learn patterns and relationships within the clustered data and classify images into their respective fashion categories. Through this combined clustering and classification approach, we endeavor to develop an efficient and accurate system for labeling and classifying fashion items within the dataset.

It comprises 60,000 grayscale images, each with dimensions of 28x28 pixels. These images are categorized into 10 classes, each representing a distinct type of clothing item. The classes are as follows: "T-shirt/top", "Trouser", "Pullover", "Dress", "Coat", "Sandal", "Shirt", "Sneaker", "Bag", and "Ankle boot". Each image is labeled with an integer from 0 to 9, corresponding to these classes..

CLUSTERING METHODOLOGY:

.
In our analysis, we employed the K-means clustering algorithm to group the images within the Fashion MNIST dataset into distinct clusters.

After applying K-means clustering, each image was assigned to the cluster represented by its nearest centroid. To label the clusters, we adopted a strategy that leveraged the labels of the nearest neighbors to the centroid. Specifically, for each cluster, we identified the 10 nearest neighbors to its centroid and computed the mode of their labels. The mode represents the most frequent label among the nearest neighbors.

By selecting the mode label, we aimed to capture the predominant fashion category within each cluster, thus providing a meaningful and representative label for the entire cluster. This approach ensures that the labeling process is robust and reflective of the dominant characteristics of the images within each cluster.

CLASSIFICATION MODEL:

The multi-layer perceptron (MLP) utilized in our analysis comprises several layers designed to effectively learn and classify the labeled data obtained from clustering. The architecture of the MLP is as follows:

Input Layer (Flatten): The input layer flattens the 28x28 pixel images into a one-dimensional array, allowing the neural network to process them as input.

Hidden Layers:
Dense Layer 1: Consisting of 100 neurons, this layer applies the rectified linear unit (ReLU) activation function to introduce non-linearity into the model and extract relevant features from the input data.
Dense Layer 2: Comprising 50 neurons, this layer further refines the features learned by the preceding layer using the ReLU activation function.

Output Layer:
Dense Layer 3: The output layer consists of 10 neurons, each corresponding to one of the fashion categories in the dataset. The softmax activation function is applied to compute the probability distribution over the classes, enabling the model to output the likelihood of each image belonging to a particular category.

TRAINING:

Once the model architecture is defined and compiled, the training process begins to optimize the model's parameters based on the defined loss function and chosen optimizer. Here's how the training is performed:

Loss Function: The model is optimized using the sparse categorical cross-entropy loss function, suitable for multi-class classification tasks.

Optimizer: Stochastic Gradient Descent (SGD) optimizer is employed to update the model's parameters based on the computed gradients, aiming to minimize the loss.

Metrics: The model's performance is evaluated based on accuracy, measuring the proportion of correctly classified images.

RESULTS:

The highest accuracy obtained is for 300 clusters, and the accuracy is 78%(approximately).

CLUSTERING & CLASSIFICATION ON THE OVERHEAD MNIST DATASET:

The dataset consists of images represented as arrays. Each image is 28 x 28 pixels, resulting in an array of size 784 for each image. The labels are associated with each data point and represent the class or category of each input.
Initially, the data was scaled down from values ranging from 0 to 255 to values ranging from zero to one. This normalization is important for ensuring consistent behavior during the clustering and classification processes.

APPROACH:

The project employs the following approach:

1. Clustering with K-Means: The training data is clustered into k groups using the k-means algorithm. The process identifies representative labels for each cluster based on the labels of the data points closest to the cluster centroids. The labels are then propagated to the rest of the data points in the same cluster.
2. Training MLP Classifier: An MLP classifier is trained on the partially labeled data obtained from the clustering stage. This approach uses the partially labeled data to improve the model's accuracy.

The process is repeated for different values of k (number of clusters) to observe the effect on final accuracy.

RESULTS:

A plot of the number of clusters (k) against final accuracy was created to visualize the relationship. If we choose a small number of clusters, the amount of labeled data available will be minimal, leaving a significant portion of the dataset unlabeled. Training the model using only these clusters may result in lower accuracy compared to using the full training dataset. Conversely, selecting a large number of clusters relative to the size of the training data can improve accuracy. However, as the number of clusters increases, the amount of labeled data needed also rises, leading to higher costs due to the necessity of obtaining a large volume of manually labeled data. So, the highest accuracy achieved is 36%(approximately) for values of k=50,100,50,200,250 and 300 is for k=300