Assignment 3: Semi-Supervised Learning for Fashion MNIST Dataset

Group Members: Aniket Santra - MDS202106 ; Soham Biswas - MDS202147

Introduction:

In this assignment we're provided with the Fashion-MNIST dataset. Fashion-MNIST is a dataset of Zalando's article images — consisting of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28x28 grayscale image, associated with a label from 10 classes. Each image is 28 pixels in height and 28 pixels in width, for a total of 784 pixels in total. Each pixel has a single pixel-value associated with it, indicating the lightness or darkness of that pixel, with higher numbers meaning darker. This pixel-value is an integer between 0 and 255. Zalando intends Fashion-MNIST to serve as a direct drop-in replacement for the original MNIST dataset for benchmarking machine learning algorithms. It shares the same image size and structure of training and testing splits.

Clustering algorithms are unsupervised machine learning approaches for grouping data based on similarity. We'll use the clustering approach to locate the most relevant samples in our data collection. We can then label them and use them to train our categorization supervised machine learning model.

Procedure:

- Fashion-MNIST dataset contains the 28x28 grayscale images and the total 784 pixel-values are stored in a 28x28 matrix. So, in the pre-processing step we've reshaped the image matrices into vectors to perform the logistic regression.

- First we've used a logistic regression model to classify the Fashion-MNIST dataset. We've trained the model with the whole training dataset consisting of 60,000 rows (or instances) and evaluated the model to the test dataset consisting of 10,000 rows or instances. We've got 84.1% accuracy. That's our baseline. We've tried to do better by using K-Means as a preprocessing step.

- In the next step we've created a pipeline that first clustered the training set into certain no. of clusters and replaced the images with their distances to the clusters, then applied a logistic regression model. We did this for randomly chosen 100, 200, 300 no. of clusters successively and checked the accuracies. We've observed that the accuracy of the model increases as the no. of clusters increases.

No. of Clusters	Accuracy
100	82.62%
200	83.89%
300	84.56%

- Then we've moved into the semi-supervised learning part. We considered that we've very small no. of labelled instances. We've taken random no. of rows or instances like 1000, 2000, 3000 from the training dataset and checked the accuracy of our model.

No. of Labeled Instances	Accuracy
1000	78.86%
2000	80.52%
3000	80.87%

We've seen that our logistic regression model performed better for large no. of instances, so for all the next experimentations we've worked with large cluster sizes. Also we've observed earlier that the model performed better for large no. of clusters too.

- We've clustered our whole training dataset into 3000 clusters and for each cluster we've found the image closest to the centroid. We called these images the representative images. So now instead of completely random instances there were 3000 representative images of each cluster. We've applied a logistic regression model using the representative images. We got 80.84% accuracy.

- Next we've propagated the labels to all the other instances in the same cluster and trained our model again. After testing we got 81% accuracy.

- We should probably have propagated the labels only to the instances closest to the centroid, because by propagating to the full cluster, we have certainly included some outliers. So, next we've only propagated the labels to the 25th, 50th and 75th percentile closest to the centroid successively and got the accuracies in each case.

Percentile Closest	Accuracy
25	80.58%
50	81.11%
75	80.77%

We've got highest accuracy (81.11%) for the propagated labelled instances which are 50th percentile closest to the centroid.

Thus we could get closer to the performance of logistic regression on the fully labelled Fashion-MNIST dataset (which was 84.1%).

Conclusion:

Here, we couldn't increase the accuracy too much by using semi-supervised learning techniques. Still in some situations when we have completely unlabelled data, using semi-supervised learning techniques we can boost the accuracy of our models by labeling a small part of our data manually, in prior.