Assignment 3

Project Report

Submitted By - Arnab Sen (MDS202009)

Sourendu Bagchi (MDS202038)

We used clustering as a step of semi-supervised learning on the Fashion\_MNIST dataset to classify each data to some particular class ( 10 total classes available e.g. “Coat”, “Sneaker”, “Trouser”, “Bag” ).

First we apply usual supervised learning with Logistic Regression on the entire training dataset. The accuracy of the model comes out to be 83.21%.

But to achieve this one needed to lable all 60000 data present in the training dataset, which is a tedious and erroneous task. Hence, to ease this labling time and effort, we try to see whether a similar level of accuracy could be achieved by only labling a portion of the dataset.

If we randomly lable 50 datapoints and see that the accuracy reduces drastically. To better our accuracy, instead of labling random datapoints from the dataset we cluster the dataset into 50 clusters, and use the datapoints closest to the 50 centroids as representatives of those clusters and give all the datapoints in a single cluster the lable of its representative.

Thus we observe that our accuracy has improved slightly to 63%.

We then try the same for differenent no. of clusters such as 100, 150, 200, 250, 300. We obsereve that the accuracy comes out to be the hihgest near 200.

So we check on shorter intervals near 200 and observe that the accuracy is highest at 220.

Hence we choose 220 as our no. of clusters and propegate its lables to the entire cluster. Accuracy is 77.09%.

Taking 20 percentile nearest to centroid datapoints in each cluster does not improve the accuracy significantly. Hence we rather choose to propegate to the entire cluster.