CMI/2022/Assignment2 Deadline: 23May, 2022

Data Mining & Machine Learning

***Group Members*:**

1. Avik Das - MDS202112
2. Meghna Mondal – MDS202123

Assignment 3: Semi-Supervised Learning

(On Fashion MNIST Dataset)

We are given the Fashion-MNIST is a dataset, 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. Now, we will use K-Means clustering for semi-supervised learning of the MNIST dataset, to identify a small subset of labelled images to see the classification process.

Procedure of using K-Means Clustering for Semi-Supervised Learning:

- **Pre-processing of the Dataset**:

Each image in the dataset 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.So, we first normalise the dataset so that it lies between [0, 1] and then flatten each image i.e., converting each image to a column vector.

- **Fitting a Logistic Regression:**

First, we have fit a Logistic Regression Model on the train data and evaluate it on the test set. We got an accuracy of 0.8416.

- **Using Clustering for Pre-processing:**

Now, we can use K-Means as a pre-processing step. We will create a pipeline that will first cluster the training set into 200, 300, 500 clusters and replace the images with their distances to the clusters, then apply a logistic regression model. Here, we get the accuracy as –

NO. OF CLUSTERS	ACCURACY
200	0.8366
300	0.8425
500	0.8442

- **Choosing optimum no. of Clusters:**

Increasing the cluster size, we are getting better accuracy, so we take the cluster size to be 1000 and 2000 and for cluster size 2000, we get the result with the best classification performance. So, we take the cluster size to be 2000 and create the pipeline with that. Now, we get an accuracy of 0.8508 i.e., we reduce the error rate by 5.8% by using Clustering.

- **Random Labelled Instances:**

For semi supervised learning, we have plenty of unlabelled instances and very few labelled instances. We have taken 1000, 1500, 2000 labelled instances and we fit a logistic regression model and get the accuracies as –

LABELLED INSTANCES	ACCURACY
1000	0.7877
1500	0.8009
2000	0.8041

- **Using Centroids to fit Logistic Regression:**

It's much less than earlier that's why we will now cluster the training set into 200 clusters, then for each cluster we will find the image closest to the centroid. We will call these images the representative images. We have plotted the representative images also.

For 200 clusters we get an accuracy of 0.7673. Now, taking no. of clusters as 2000, we get 0.814 following the same process of fitting Logistic Regression model to the representative sets.

- **Propagating Same Labels to each Datapoint in a cluster:**

Now to improve the accuracy even more, we can propagate the labels to all the other instances in the same cluster as of the cluster centroids. We got a tiny little accuracy boost to 0.8121.

- **Partially Propagating Same Labels to each Datapoint in a cluster:**

From this observation, we suspect that 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, now we check the accuracies by propagating only the labels to the 25, 50, 75th percentile closest to the centroid and we get the accuracy as –

PERCENTILE CLOSEST	ACCURACY
25	0.8042
50	0.809
75	0.8124
80	0.8136
85	0.8129

So, finally we propagate the labels only to 80th percentile closest to the instances closest to the centroid. And here we get an accuracy of 0.8136.

Now, we also compute that only 10% of the propagated labels are actually correct.

So, here with just 2000 labelled instances, we got 81.36% performance, which is getting closer to the performance of logistic regression on the fully labelled *Fashion MNIST* dataset which was 84.2%.