MACHINE LEARNING ASSIGNMENT 3

Ankush Dey (MDS202108)
Anjali Pugalia (MDS202107)

Introduction:

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. 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. We will treat this as a semi supervised data, and perform clustering on it.

Steps used to Classify the Data:

- We first import all the necessary libraries.
- Next we load the dataset which is divided into test and train parts.
- The data points are divided by 255 to normalise.
- Next we flatten the data points using the reshape function to convert it to two dimensions.
- We then run a Logistic Regression on the data with 5000 iterations. We fit the test data and get an accuracy of 84%.
- A kmeans algorithm is run on the data to create 100,200 and 300 clusters then a logistic regression model is used on each of them. We get accuracy levels 81.81%, 82.51% and 83.06% respectively.
- Then we checked the accuracy of the logistic regression for the first 500, 1000, and 2000 labelled data. We get accuracy levels of 78.52%, 79.24% and 80.89% respectively.
- Since the accuracy was maximum for the 2000 labelled instance. So we did the kmeans for 2000 clusters.

- So when we did the logistic regression using these clusters, we got an accuracy of 81.75%.
- Since it is often costly and painful to label instances, especially when it has to be done manually by experts, it is a good idea to label representative instances rather than just random instances.
- We can go one step further if we propagate the labels to all the other instances in the same cluster, this is called label propagation.
- Here we propagated each representative instance's label to all the instances in the same cluster, including the instances located close to the cluster boundaries, which are more likely to be mislabeled.
- Now we only propagate the labels to the 20% , 50% and 75% and of the instances that are closest to the centroids.
- We get an accuracy of 80.46% ,80.90% and 81.26% respectively . We expected a higher accuracy for this result. One of the possible reasons for getting a poor accuracy percentage could be due to overfitting of the data.
- We can improve the results by reducing the number of clusters. However due to time constraints we were unable to try this. The model ran for several hours before giving a result.
- We could also look at different methods such as using neural networks to improve accuracy levels.