Data Mining and Machine Learning
Assignment 3

Paarth Iyer (MCS202218), Varun Agrawal (MDS202251)

The Dataset:

The datasets were downloaded in the ubyte format and then loaded locally using idx2numpy package.

1. The Fashion MNIST dataset was used in this assignment. In this dataset, each image is a grey-scale image, 28 pixels in height and 28 pixels in width. 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. The training dataset has 60,000 samples while the designated test dataset has 10,000 samples. Each sample has a label from 0-9 (10 classes), viz., 'T-shirt/top', 'Trouser', 'Pullover', 'Dress', 'Coat', 'Sandal', 'Shirt', 'Sneaker', 'Bag', 'Ankle Boot'.
2. The Overhead-MNIST dataset is composed of around 10000 grayscale satellite images of shape (28, 28, 1. There are 784 pixels per picture. The classes it has are car, harbour, oil and gas field, parking lot, plane, ship, and storage tank, runway mark, and stadium, with around equal distribution.

The Task:

The task was to use clustering for semi-supervised learning of the Fashion MNIST dataset and the Overhead-MNIST dataset. K-Means clustering was used to cluster subset of labelled images to seed the classification process.

Function Descriptions:

1. print_imgs(imgs , n) : Take set of images as input and then displays those images in n columns.
2. flatten_imgs(imgs) : Take set of images as input and then flattens it and normalize it.
3. clusters_and_labels (k, train_imgs, train_labels) : takes value for k and fits a kmeans cluster model for given k on train images and returns the model and closest cluster labels to the train labels and their indices.
4. predict_labels(imgs , n) : Predicts in which clusters each test images lie

propogate(kmeans , cluster_labels, x_train, x_cluster_dists, prop_frac) : propogates the cluster labels to closest prop_frac*100% of points to the cluster.

Procedure (For both datasets we used the same procedure):

1. 10% of the samples from the training dataset were set aside as the validation samples, on which we tuned the parameters of models to maximize the accuracy in both cases.
2. Images were flattened and normalized.
3. The image vectors were reduced in dimensions using PCA. Values for the reduced dimensions were found based on the accuracy of the validation set.
4. Kmeans clustering was performed on the train set. For each cluster the closest points were found and the cluster was given the label using clusters_and_labels().
5. Each of the representative images were printed so that they could be labelled. In this scenario we used the provided labels instead of manual labelling.
6. We also make new datasets using the propagation of labels to given percent of the closest points to the centre of the cluster.
7. We tried placing the labels on the points according to clusters , logistic regression on the propagated dataset and svm on propogate dataset and we tried different % of propagation to find highest accuracy on the validation set.
8. We find the model with highest accuracy and then fit that on the test set to find our final performance.

Results:

Fashion MNIST :

1. The best value for the n_components in PCA was found to be around 50. Going above or below that value generally resulted in slightly lower performance for this dataset.
2. Using **k = 300,** which is 0.5% of training dataset, we achieved an accuracy of 80.46% on the validation set using 5% propagation and using SVM to make the model. This means manually labelling 300 images and propagating to approximately 2700 images. We tried to get as close to 80% percent as possible without having too many manually labelled images.
3. This results in an accuracy of 80.12% on the test dataset using the above model.
4. Just clustering and Logistic regression both performed worse than the SVM, but both being above 74% on the validation set. They were not tested on the test dataset as they did not give the best performance.
5. Looking at the classification report of the model, we see that all the categories except for pullover and shirt had good overall performance. The coat and t-shirt also have slightly worse recall and accuracy compared to others. This could be because of their similar 'look' in the images.
6. For the propagation percentages, we found that going lower or higher than the optimal percentages resulted in lower performance.

Over MNIST :

1. The best value for the n_components in PCA was found to be around 25. Going above or below that value generally resulted in slightly lower performance for this dataset.

2. In this one, our goal was to hit close to 50% without too many clusters. This resulted in picking **k=150,**
   which is around 1.7% of the training set. Using Svm with a 30% propagation gave us 49.29% accuracy in the validation set.
3. This results in an accuracy of 48.35% on the test dataset using the above model.
4. The logistic regression model performed worse than just clustering for all propagation values on the validation set. Both performed way worse than SVM model
5. According to the classification report, parking lot has the worst performance, followed by runway mark, stadium and oil gas field. All of these are not that descriptive when looking at their images.
6. For the propagation percentages, we found that going lower or higher than the optimal percentages resulted in lower performance.