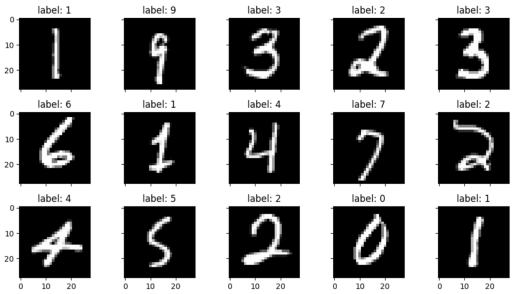
Home Assignment 1 – Unsupervised Learning

April 19, 2023

1 Exploring the Data

The MNIST database comprises of images of handwritten digits. There are total of 70,000 images in the database labeled by digits between 0 and 10. Each label has over 6000 samples. The database was splitted to 60.000 training samples and 10,000 test samples. digits are shown in figure 1

Figure 1: MNIST Digits



2 Clustering

In this part we choose Kmeans, Agglomerative Clustering and Gaussian Mixture as our clustring models. As the computational complexity for Agglomerative Clustering is $O(n^3)$, we used only a fraction of the database - drawing 400 samples for each label. Furthermore, Agglomerative Clustering does not partition the input space, thus assigning new test point to this model is not possible. Thus, in order to deal with the matter, we fitted KNN model on the training set using the clusters, found by Agglomerative Clustering over the training set, as labels. We set the number of neighbors to be 1, which is similar as using linkage = single, allocating the test samples to the nearest cluster.

In all three models, the number of clusters was set be equal to the number of unique labels, i.e. 10 clusters. First we clustered the training set and used the true labels to map the clusters to labels. We used two methods to associate labels to clusters:

- **Correlation:** In this method we first created a correlation matrix between the true labels and the clusters. Then we reduced the "noise" of self correlation by a threshold, between a cluster's samples to themselves and finally assigned a label with highest correlation to the cluster.
- **Majority:** In this method we mapped a cluster to label which was the most common in the cluster. Finally, we measured the classification predictions using accuracy.

Summary table of accuracy results using both methods is shown in figure 2. The results demonstrate that using clustering for classification produces inaccurate results, far less than designated classification methods like KNN or CNN.

Figure 2: Classification Accuracy Using Clustering

Classification Accuracy Using Clustering		
	Correlation Method	Majority Method
Kmeans	45.01%	59.46%
Agglomerative Clustering	47.55%	61.72%
Gaussian Mixture	40.33%	58.97%
KNN	97%	
CNN	96%	

2.1 Experiments

Kmeans: We experimented with several penalty thresholds for reducing self correlation, the results indicate that the higher the penalty threshold the higher the accuracy (47.55%).
Agglomerative Clustering: We experimented with several linkage methods used when determining the distance between clusters. the linkage methods used were: ['ward', 'complete', 'average', 'single']. It can be concluded from the results that 'ward' method produced the highest accuracy, in both mapping methods (47% and 62% respectively).
Gaussian Mixture: We conducted experiment with different types of covariance matrices. Using covariance type = 'spherical' means that all the Gaussians have their own single variance, resulting in a shape of a sphere, produces the best accuracy (45% 58%).

3 Dimensionality Reduction

In this section the following methods were used for Dimensionality Reduction: PCA, T-SNE and Gaussian Random Projection. Using these methods, we mapped both the training and test set to a lower dimension space and used KNN for classification. T-SNE does not learn a mapping function on a lower dimensional space and thus doesn't have a transform method that can be used to transform the test set. In order to overcome this issue, we first found, for each test sample, 3 nearest Neighbors from the training set in the higher dimension. Then, we averaged the corresponding training samples in the lower dimension, to be the estimated test sample in the lower dimension. Summary table of accuracy results using both methods is shown in figure 3. The results demonstrate that preforming classification after Dimensionality Reduction can produce excellent results, depending on the lower space Dimensionality and can even surpass results of classification methods like KNN or CNN.

Figure 3: Classification Accuracy after Dimensionality Reduction

Classification Accuracy after Dimensionality Reduction		
	2 Dimensions	50 Dimensions
PCA	40.43%	97.52%
T-SNE	94.51%	N/A
Gaussian Random Projection	33.98%	93.86%
KNN	97%	
CNN	96%	

3.1 Experiments

PCA: We experimented with different number of dimensions in the target space and whether to use whitening. Whitening will transform the data so that it has zero mean and identity covariance matrix which might improve the predictive accuracy when utilize classification estimators. The results indicates that choosing 10+ dimensions over 2 dimensions in the lower space, immensely improve accuracy, while whitening did not have significant impact on the results.
T-SNE: For this algorithm we conducted experiments using different perplexity values: [2, 5, 30, 50, 100]. Higher accuracy, of over 94%, was demonstrated when perplexity $\in [5, 30, 50]$
Gaussian Random Projection: We experimented with different number of dimensions in the target space. When choosing 2 dimension as the lower space dimension the model produce low accuracy of 27.18%. As the number of dimensions increases, the accuracy approves dramatically, equal to 93.86% and 95.81% when choosing 50 and 100 dimension for the lower space respectively.

4 summery

In this assignment we demonstrate that clustering should not be used as a method for classification and should only be utilize to find structures in a given data, which should subsequently be validated according to business purposes. In addition, Dimensionality Reduction could be use to reduce data redundancies and to reduce computational complexity without undermining, and even assist, downstream classifiers performance.