Homework 2

Daniele Napolitano

November 25, 2022

---

1 PCA and LDA comparison

The task for this exercise is to compare PCA and LDA in their ability to cluster when projecting very high-dimensional datapoints to 2 or 3 dimensions. In particular, consider the dataset MNIST.
It contains 42000 images, each of resolution 28*28 pixels (784 flattened pixels per image), and each image is labeled, giving a total of 42000 rows and 785 columns.

Figure 1: Some of the digit images in the MNIST dataset

1.1 PCA

Principal Component Analysis is a dimentionality reductyion technique that aims to maintain the highest amount of information, by changing basis and applying the SVD (Singular Value Decomposition) method. In particular, given $X \in \Re^{dxN}, k < d$ And applying the truncated SVD:

$$X_k = U_k \Sigma_k V_k^T \tag{1}$$

$U_k \in \Re^{dxk}$ will represent the projection matrix, that will project the original dataset of d dimensions in k dimensions:

$$Z = U_k^T X \tag{2}$$

If k=2, the projection can be plotted on a 2D graph.
As shown in Fig.2, clusters overlap on each other, for that reason it's visually difficult to distinguish them. Calculating the truncated SVD and the projection Z took only 0.1s, making PCA a fast method for dimentionality reduction.
To test the efficiency of PCA as a clustering method, the test dataset is used to try to predict the label of each data point, and then comparing the prediction with the actual label (accuracy estimation).
The label is guessed by looking at the distances of the data point with respect to all the cluster centroids, and assigning it to the closest one.
Out of 4983 data points, only 2712 were correctly labelled, giving an accuracy of 0.5406381697772427.
As shown in Fig. 2, since clusters are overlapped, also centroids are close to each others, making the prediction more difficult and prone to error.

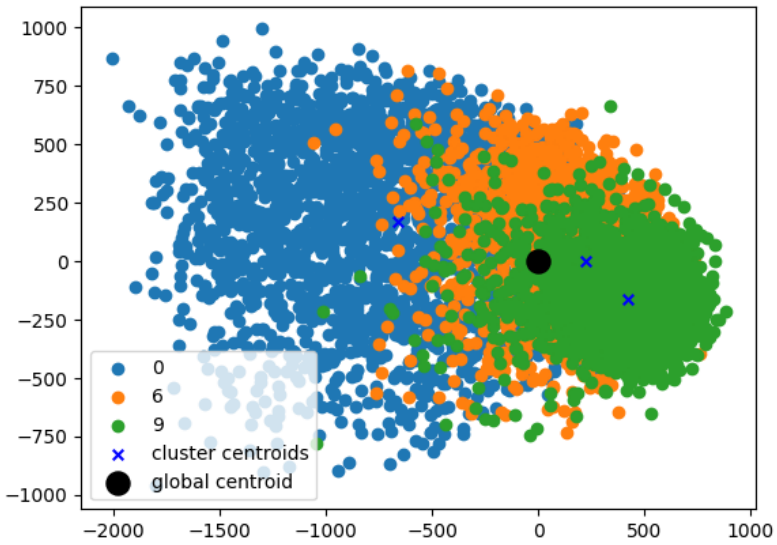

Figure 2: MNIST training dataset plotted in 2 dimensions using PCA

1.1.1 Other experiments with PCA

For k=3, the situation does not change, in fact as shown in Fig.3, clusters are still overlapped (being able to distingush them is even more difficult), and prediction accuracy goes down to 0.32329921733895245, while the time to compute the truncated SVD and projection got up to 0.6s. A lower k value seems to be better for both accuracy and time.

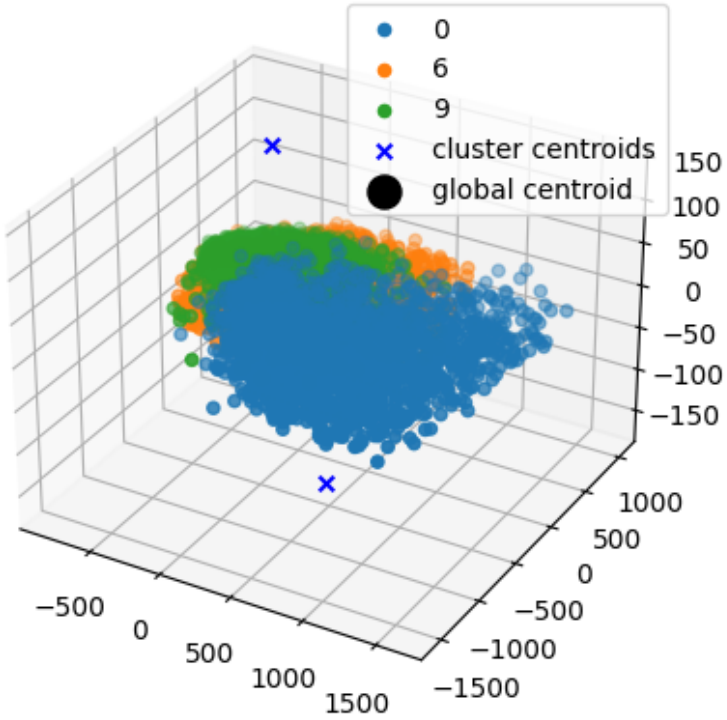

Figure 3: MNIST training dataset plotted in 3 dimensions using PCA

1.2 LDA

Linear Discriminant Analysis is a clustering technique, which attempts to better distinguish classes (when compared to PCA).

> *"The idea of LDA is to build a projector $Q \in \Re^{dxk}$ such that the within-clusters distance of the projected data is as small as possible, while the between-clusters distance of the projected data is as big as possible."*

The projection matrix is computed using the Cholesky decomposition method, which turns a positive-definite symmetrical matrix into the product of a lower triangular matrix (L) and its transpose (L^T): $S_w = LL^T$. The projection matrix $Q^T \in \Re^{kxd}$ is obtained as:

$$Q = L^{-T}W \tag{3}$$

Where W is the matrix which contains the first k eigenvalues of $L^{-1}S_bL$, and $S_b = X_cX_c^T$.
The other big difference with PCA is in fact that it's required to compute the class centroid for each cluster, and center each cluster accordingly ($X_c = X - c(X)$), while PCA only needed the global centroid.
Finally, the projected matrix is computed as:

$$Z = Q^T X \tag{4}$$

LDA requires more operations than PCA, so it's slower (3.8s to perform all the operations described above in order to compute Q and the projection Z), but the clusters are much easier to distinguish, also class centroids are further away from each other (Fig.4)

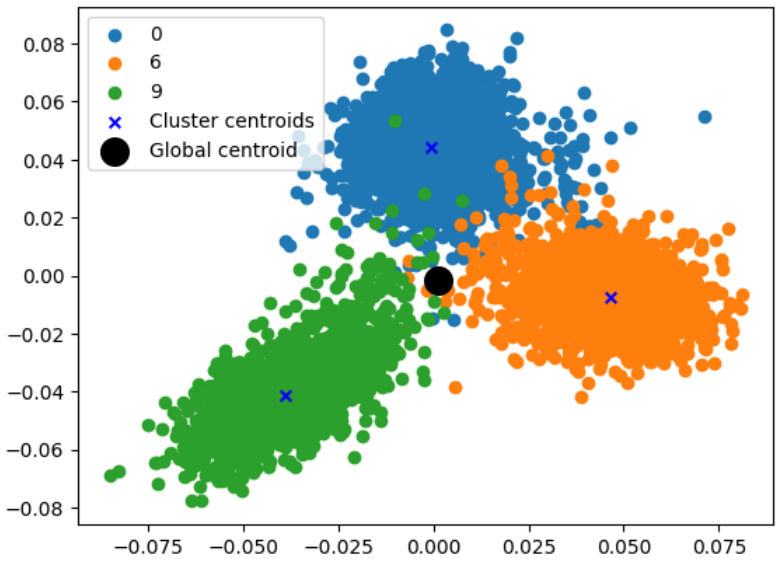

Figure 4: MNIST training dataset plotted in 2 dimensions using LDA

Out of 4983 data points in the test set, 3312 of them were correctly labelled, giving an accuracy of 0.6646598434677905.

1.2.1 Other experiments with LDA

For k=3, it took 2.2s to compute Q and Z, and the accuracy remained stable (0.6646598434677905). Clusters are still visually easy to distinguish (Fig.5).

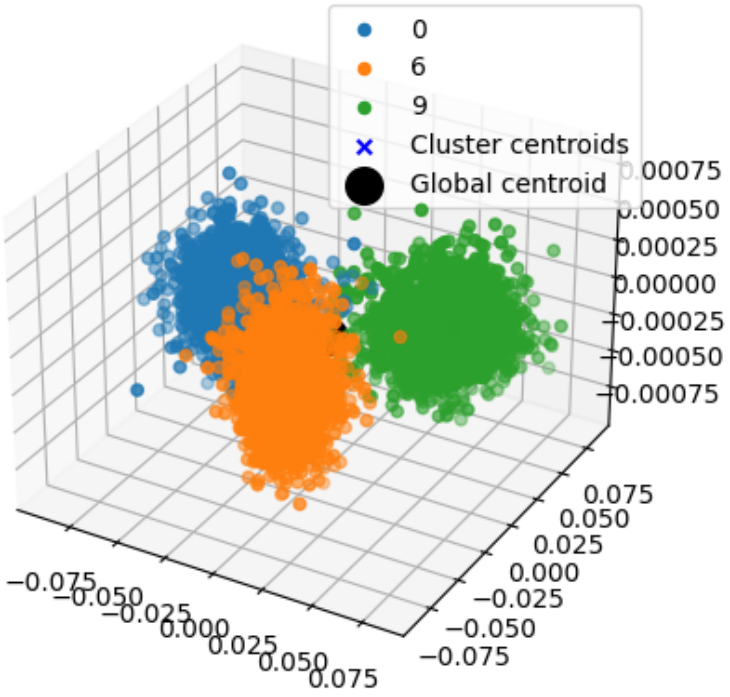

Figure 5: MNIST training dataset plotted in 3 dimensions using LDA