ROB311 - TD6 - Apprentissage pour la Robotique

Araujo Belén, Victor
De Carvalho Ferreira Sula, Julia

October 2019

1. K-Means Clustering

K-Means is a unsupervised algorithm which the basic objective is to group similar points together, known as a cluster.

The algorithm does so by identifying the cluster centers, the centroid that is a representative of the whole cluster, and assigning every data point to the closer cluster. As this iterative process happens, new centroids are calculated by averaging the cluster data.

As many unsupervised algorithms K-Means is excellent for cases where the classes are unknown.K-Means has also two phases: training and testing. The first phase aims to finding the best centroids to represent the data, the second phase uses the centroids to classify the test data.

2. The Problem: Optical Recognition of Handwritten Digits

In this TD, the goal is to use K-means to identify a digit from a Handwritten photo of a Number.

Firstly, in order to facilitate the centroid choosing process, the initial cluster center is chosen as a member of each of the classes, that is to say, a representative of one of the digits(0 to 9).

Sequentially, the fit is done, resulting in a cluster center that better represents the data. This cluster center is then applied to the testing data, classifying it.

3. Results

Applying the K-means and the scklearn metrics, the prediction and classifying process can be evaluated.

Firstly the confusion matrix of both the training and the testing process is obtained as shown bellow. It is easily noticed that for the digit 4 and 9 the number of correct answers is not as large as expected.

$$C_{training} = \begin{bmatrix} 0 & 156 & 17 & 9 & 0 & 1 & 2 & 0 & 201 & 3 \\ 0 & 0 & 343 & 12 & 0 & 0 & 1 & 11 & 12 & 1 \\ 0 & 0 & 5 & 351 & 0 & 5 & 0 & 10 & 11 & 7 \\ 0 & 14 & 0 & 0 & 272 & 1 & 4 & 13 & 2 & 81 \\ 0 & 0 & 0 & 72 & 0 & 289 & 1 & 0 & 6 & 8 \\ 0 & 1 & 0 & 0 & 1 & 0 & 374 & 0 & 1 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 & 340 & 4 & 40 \\ 1 & 2 & 1 & 88 & 0 & 2 & 5 & 1 & 274 & 6 \\ 0 & 5 & 0 & 246 & 0 & 0 & 0 & 8 & 2 & 121 \end{bmatrix} \quad (1)$$

$$C_{testing} = \begin{bmatrix} 176 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 62 & 21 & 1 & 0 & 0 & 4 & 0 & 94 & 0 \\ 1 & 2 & 150 & 8 & 0 & 0 & 0 & 3 & 13 & 0 \\ 0 & 0 & 0 & 165 & 0 & 1 & 0 & 7 & 8 & 2 \\ 0 & 4 & 0 & 0 & 144 & 0 & 0 & 1 & 7 & 25 \\ 0 & 0 & 0 & 26 & 1 & 152 & 1 & 0 & 0 & 2 \\ 1 & 1 & 0 & 0 & 1 & 0 & 176 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 140 & 3 & 36 \\ 0 & 6 & 1 & 31 & 0 & 2 & 2 & 1 & 121 & 10 \\ 0 & 3 & 0 & 145 & 0 & 3 & 0 & 1 & 2 & 26 \end{bmatrix} \quad (2)$$

Calculating the score of the classification, this becomes even more clearly , as the percentage for digits 4 and 9 are incredibly low compared to the other digits as shown by the result bellow.

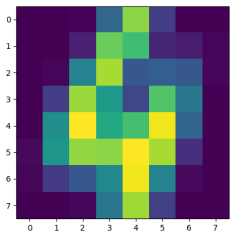
```
Score for training data [0.99332443 0.54736842 0.91957105
0.60154242 0.82299546 0.85756677 0.97777778 0.88311688
0.61297539 0.37230769]
```

```
Score for testing data [0.98876404 0.47692308 0.85959885
0.59033989 0.87537994 0.89411765 0.96703297 0.84337349
0.57075472 0.18505338]
```

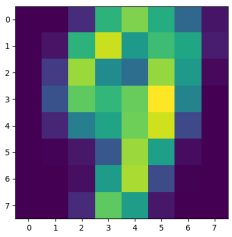
This behavior is due the fact that there any many styles of writing this both numbers, which can be fairly similar, as shown by the figure 1 that represents the centroid of each of this digits.

The cluster center of every digit is shown in figure 2.

In order to increase efficacy one can create more classes in which the centroids will be defined by digits 4 and 9 in different styles. This allows to create a greater range of cluster centers and, therefore, increase the separability of the digits. Doing so, the score of the training and testing process were higher,the average was 0.8278151201767853 and 0.801787313708403 , respectively, in comparison to the first scores that were 0.7576087602419194 and 0.724970578744805.

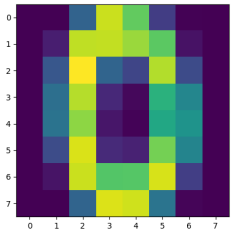


(a) Cluster center digit 4

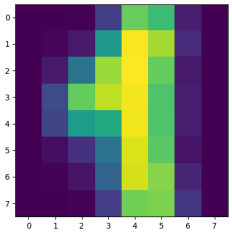


(b) Cluster center digit 9

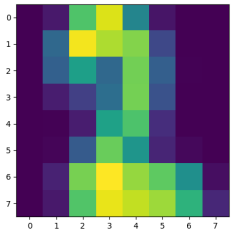
Figura 1



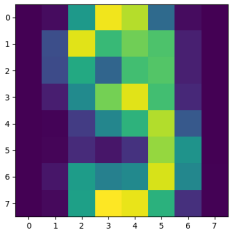
(a) Cluster center digit 0



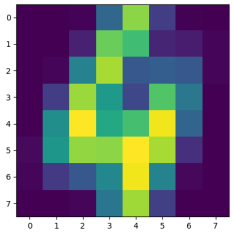
(b) Cluster center digit 1



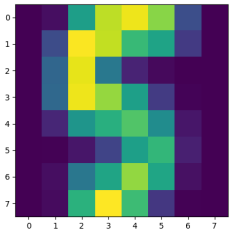
(c) Cluster center digit 2



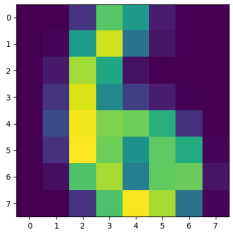
(d) Cluster center digit 3



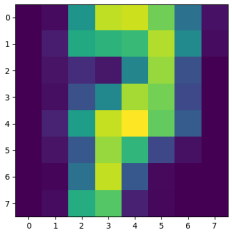
(e) Cluster center digit 4



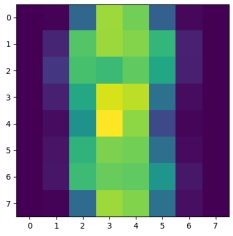
(f) Cluster center digit 5



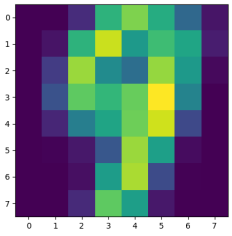
(g) Cluster center digit 6



(h) Cluster center digit 7



(i) Cluster center digit 8



(j) Cluster center digit 9

Figura 2: Centroids of each digits