Machine learning first programming exercise: Kmeans clustering

Atit Bashyal
Rana Hamza Intisar

March 2019

1 Task One

We picked digit number seven, which was indexed by row number 1400 to 1599 in the digits data set. As outlined in the task sheet we carried out a K-means clustering on the chosen sample of digit seven in four runs of the algorithm. For these four runs, we set K=1,2,3 and 200. The following report discusses the visualizations of images that were coded in the respective code book vectors, that we obtained by the four runs of our algorithm.

1.1 K=1 Clustering

For the case where k=1, the algorithm generates a single code book vector C_1 which represents the mean of all the vector points in our sample set. In mathematical formalization this can be represented in the following way, let C_1 be the code-book vector for k=1 then mathematically,

$$C_1 =\mid S_j \mid^{-1} \sum_{x_i \in s_j} x \tag{1}$$

where, S_j is the set of data points from the digits data set which contains the image vector points for the digit seven. The code-book vector output as an image for the case K=1 that we obtained is shown below in figure 1.

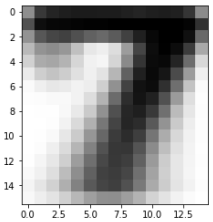

Figure 1: Code-book vector for K=1

Each pixel of the code-book vector image contains information on the average pixel value of all the 200 images of digit seven. Therefore, the code-book image (for k=1) we plot, does not show a definite edge boundary of digit seven.

1.2 K=2 Clustering

For the case where k=2, the algorithm generates two code book vectors C_1 and C_2 which represent the mean of all the image vector points falling in the respective cluster sets S_1 and S_2. In mathematical formalization this can be represented in the following way, let C_i $i \in 1, 2$ be the code-book vectors for k=2 then mathematically,

$$C_i = \mid S_j \mid^{-1} \sum_{x_i \in s_j} x \mid j \in 1, 2 \tag{2}$$

The code-book vector image outputs for the case K=2 that we obtained are shown below in figure 2.Similar to the code-book image (for k=1), the code-book images of (k=2) also do not show a definite edge boundaries.

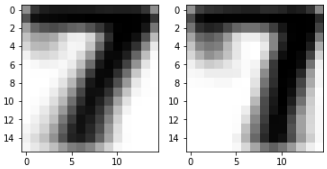

Figure 2: Code-book vectors for K=2, the left image is C_1 and right image is C_2

1.3 K=3 Clustering

For the case where k=3, the algorithm generates three code book vectors C_1 , C_2 and C_3 which represent the mean of all the image vector points falling in the respective cluster sets S_1 , S_2 and S_3. In mathematical formalization this can be represented in the following way, let C_i $i \in 1, 2, 3$ be the code-book vectors for k=3 then mathematically,

$$C_i = \mid S_j \mid^{-1} \sum_{x_i \in s_j} x \mid j \in 1, 2, 3 \tag{3}$$

The code-book vector image outputs for the case K=3 that we obtained are shown below in figure 3 .Similar to the code-book images for k=1 and k=2, the code-book images of k=3 also do not show definite edge boundaries. An interesting feature can be seen in the code-book vector C_1 where, the code-book vector contains pixel information for an particular image of digit seven, where the digit has been written/rotated diagonally. this feature, is not prominent in the code-book vector image, i.e is seen with a low pixel intensity, because the instances in which the digit is written diagonally is less compared to the instances when the digit is written vertically.

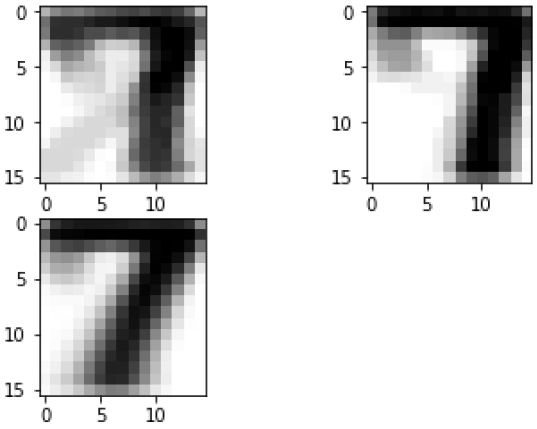

Figure 3: Code-book vectors for K=2, the upper left image is C_1 and upper right image is C_2 and the lower left image is C_3

1.4 K=200 Clustering

For the case where k=200, the algorithm generates two hundred code-book vectors C_i where $i \in \{1, 2, 3, 4, ..., 200\}$ which represent each of the image vector points in the set S_j where, S_j is the set of data points from the digits data set which contains the image vector points for the digit seven. In mathematical formalization this can be represented in the following way, let C_i $i \in \{1, 2, 3, 4, ..., 200\}$ be the code-book vectors for k=200 then mathematically,

$$C_i = x_i \in S_j \mid i \in \{1, 2, 3, 4, ..., 200\} \tag{4}$$

The code-book vector image outputs for the case K=200 that we obtained are shown below in figure 4 .

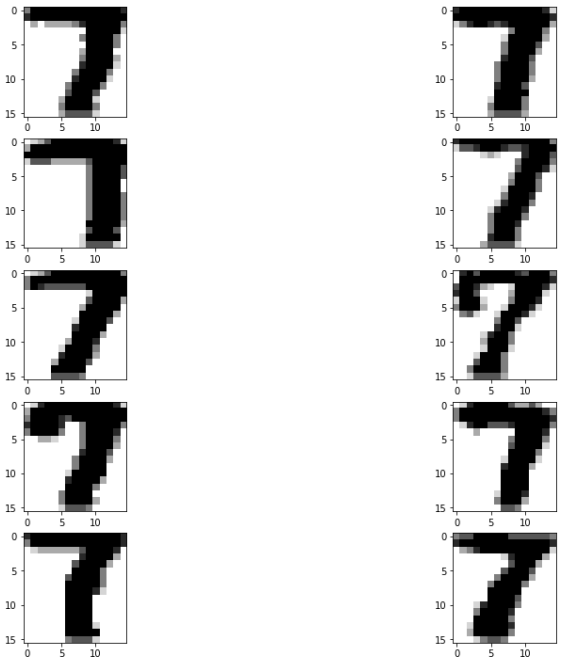

Figure 4: Code-book vectors for K=200, starting from the upper left image and moving to the right images are for code-book vectors 1,20,40,60,80,100,120,140,160 and 180

2 Bonus Task: clustering with the complete data set

In this task we used the complete the digits data set. We carried out a K-means clustering on the data set in three runs of the algorithm. For these four runs, we set K=10,20,30. The following visualizations of images are images that were coded in the respective code book vectors, that we obtained by the four runs of our algorithm.

2.1 K=10 Clustering

We choose to run the algorithm with 10 clusters since we have 10 different digits in the data set. The images of the code-book vectors generated show, that the algorithm does not do a good job in identifying the digits. The clusters that are formed have code-book vectors containing pixel information of different digits.

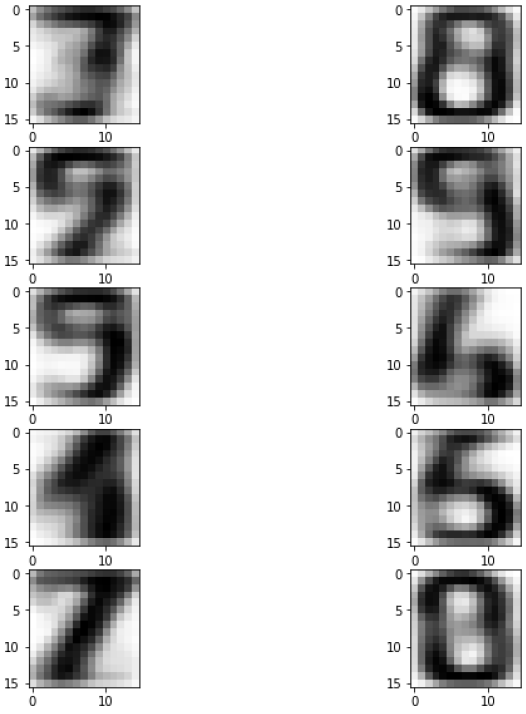

Figure 5: Code-book vectors for K=10, on the digits data set

2.2 K=20 Clustering

We choose to run the algorithm next with 20 clusters . The images of the code-book vectors generated show, that the algorithm does a better job in identifying the digits compared to the k=10 case. Some of the clusters that are formed still have code-book vectors containing pixel information of different digits.

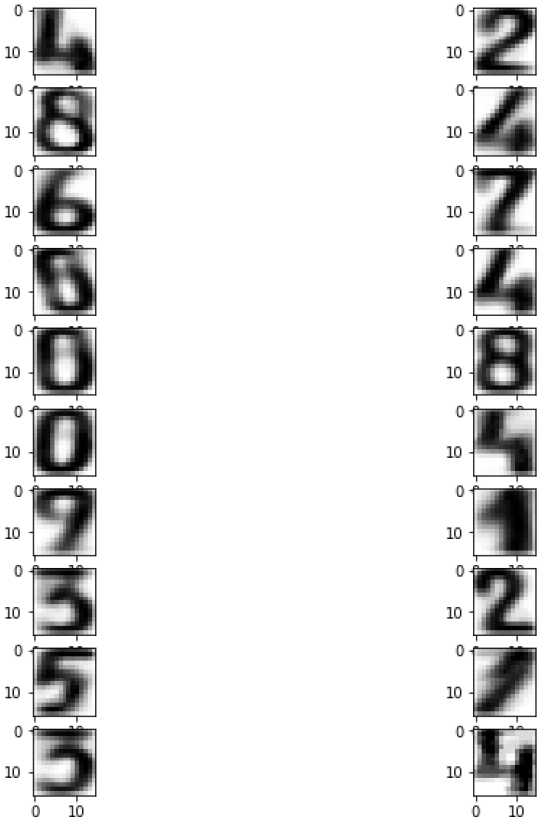

Figure 6: Code-book vectors for K=20, on the digits data set

2.3 K=30 Clustering

We choose to run the algorithm next with 30 clusters . The images of the code-book vectors generated show, that the algorithm does a good job in identifying the digits compared to the k=10 and k=20 cases. Some of the clusters that are formed still have code-book vectors containing pixel information of different digits. At this point it is easier to notice that the algorithm has been forming some clusters with information on different digits which have some similarities in the way they are written for example the algorithm clusters digit 1 with digit 7, digit 0 with digit 8, digit 5 with digit 6 and digit 9 with digit 2.

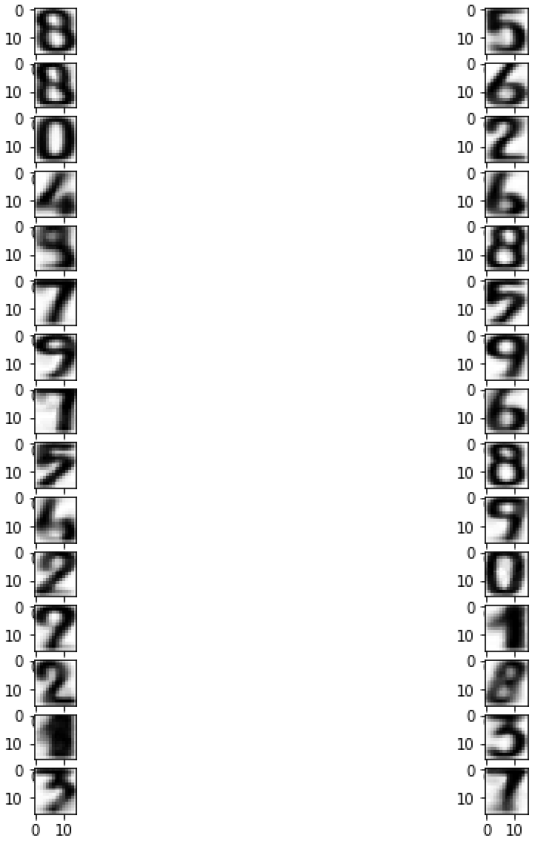

Figure 7: Code-book vectors for K=30, on the digits data set