ISYE 6740 Spring 2021
Homework 1

Student: Homayoun Gerami

January 2021

1 K-means clustering [60 points]

Given m data points x^i, $i = 1, \ldots, m$, K-means clustering algorithm groups them into k clusters by minimizing the distortion function over $\{r^{ij}, \mu^j\}$

$$J = \sum_{i=1}^{m} \sum_{j=1}^{k} r^{ij} \|\mathrm{x}^i - \mu^j\|^2, \tag{1}$$

where $r^{ij} = 1$ if x^i belongs to the j-th cluster and $r^{ij} = 0$ otherwise.

1. (10 points) Derive mathematically that using the squared Euclidean distance $\|\mathrm{x}^i - \mu^j\|^2$ as the dissimilarity function, the centeroid that minimizes the distortion function J for given assignments r^{ij} are given by

   $$\mu^j = \frac{\sum_i r^{ij} \mathrm{x}^i}{\sum_i r^{ij}}.$$

   That is, μ^j is the center of j-th cluster. Hint: You may start by taking the partial derivative of J with respect to μ^j, with r^{ij} fixed.

   student answer:

   we perform $\frac{\partial J}{\partial \mu}$ while r^{ij} is constant, and set the derivative to zero, since we are looking for the μ^js that minimize the J, and then formulate the μ:

   $$\frac{\partial J}{\partial \mu} = 2 \sum_{i=1}^{m} r^{ij} (x^i - \mu^j) \ = 0 \quad \Rightarrow \quad \mu^j = \frac{\sum_i r^{ij} \mathrm{x}^i}{\sum_i r^{ij}}.$$

2. (10 points) Derive mathematically what should be the assignment variables r^{ij} be to minimize the distortion function J, when the centroids μ^j are fixed.

   student answer:

   Because J is a linear function of r^{ij} this optimization can be performed easily to give a closed form solution. The terms involving different i are independent and so we can optimize for each i separately by choosing r^{ij} to be 1 for whichever value of j gives the minimum value of $\| \ x^i - \mu^j\|^2$. hence we can write r^{ij} as :

   $$r^{ij} = \begin{cases} 1, & \text{if } j = argmin^j \|\mathrm{x}^i - \mu^j\|^2 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

3. (10 points) Write down a pseudocode for K-means algorithm here, based on your derived results.

student answer:

(a) start

(b) Read the input data X^i, m is the number of samples, $i \in (1, ..., m)$

(c) Specify number of clusters that we would like to have K

(d) Random selection of K number of clusters' centers(centeroids) μ^j, $j \in (1, ..., K)$, for simplification $\mu^j \in (X^1, ..., X^m)$ but μ^j can belong to any desired valid range

(e) **Repeat:**

- Assign each data point, X^i to its closest centeroid; data sample X^i belongs to cluster j if the $\|\mathrm{x}^i - \mu^j\|^2$ norm distance is minimized
- compute the new centeroid (mean) of each cluster as per the following derived formula: $\mu^j = \frac{\sum_i r^{ij} \mathrm{x}^i}{\sum_i r^{ij}}$

(f) **until:** the centeroid positions do not change

(g) end

4. (10 points) Explain why K-means algorithm converges to a local optimum in finite steps.

student answer:

In the K-means algorithm, the following two steps are repeated until convergence:
1. Minimize J with respect to r^{ij}
2. Minimize J with respect to μ^j

The cost function, J, decreases with every iteration and that guarantees the convergence.
Here is the snapshot from Prof. Yui lecture, demonstrating this phenomena:

- The minimum value of the objective is finite
- Each iteration of kmeans algorithm decrease the objective
  - Cluster assignment step decreases objective
    - $\pi(i) = argmin_{j=1,\ldots,k} \left\|x^i - c^j\right\|^2$ for each data point i
  - Center adjustment step decreases objective
    - $c^j = \frac{1}{|\{i:\pi(i)=j\}|} \sum_{i:\pi(i)=j} x^i = argmin_c \sum_{i:\pi(i)=j} \left\|x^i - c\right\|^2$

Figure 1: Convergence of kmeans algorithm

5. (20 points) Calculate k-means by hands using Euclidean distance, i.e., the set up in Equation (**??**). Given 5 data points configuration in Figure 1. Assume $k = 2$. Assuming the initialization of centroid as shown.

   (a) (15 points) Complete the following table for all iterations until the algorithm converges.

   student answer:

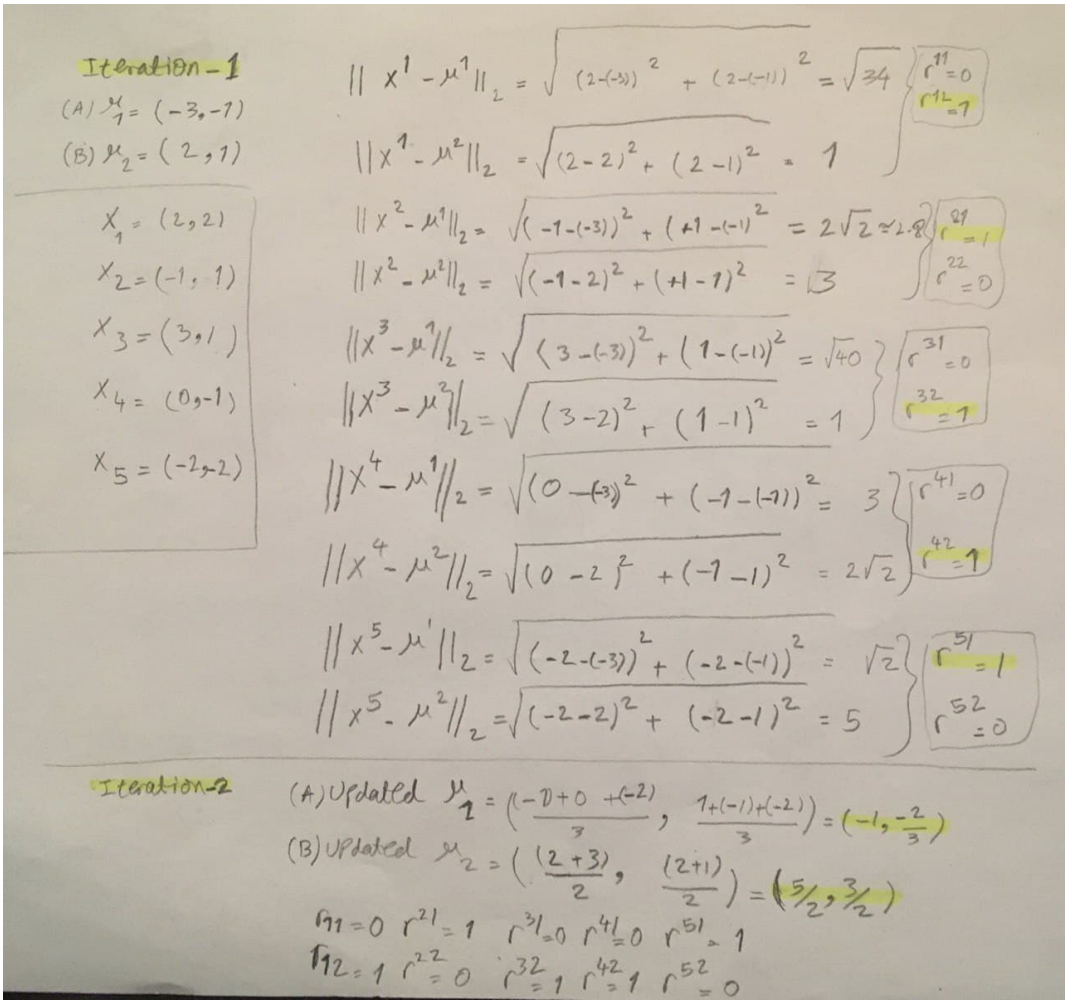

Figure 2: k-means by hands using Euclidean distance

Iteration No.	μ^1	μ^2	r^{11}	r^{21}	r^{31}	r^{41}	r^{51}
1	(-3,-1)	(2,1)	0	1	0	0	1
2	$(-1, \frac{-2}{3})$	$(\frac{5}{2}, \frac{3}{2})$	0	1	0	0	1

   (b) (5) points How many iterations it takes for k-means to converge?

student answer:

If we count the assignment of the centroids A,B and association of the points to the two clusters as the Iteration number-1, then the k-means simply converges with the next iteration, number-2, so my answer is two iterations.

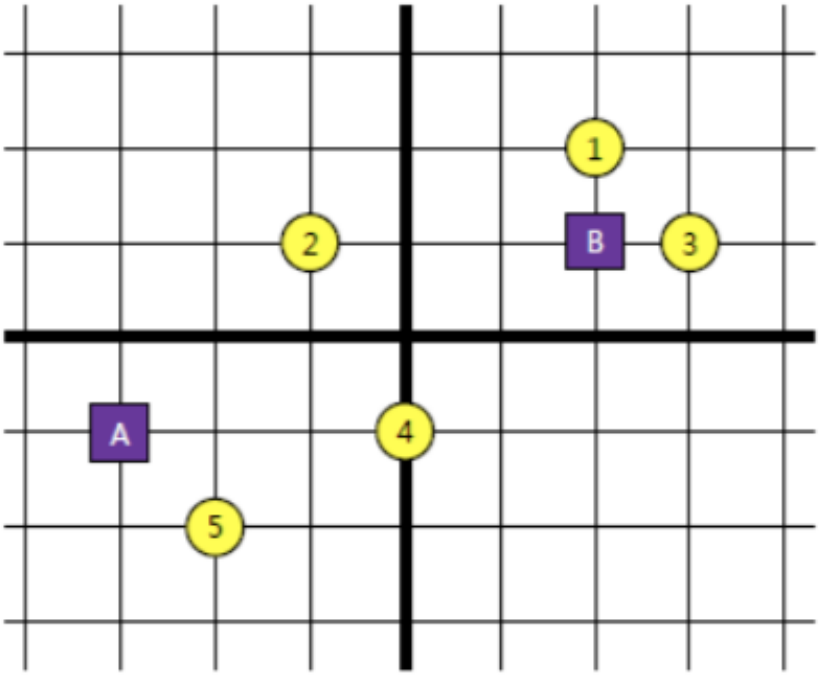

Figure 3: K-means.

2 Image compression using clustering [40 points]

In this programming assignment, you are going to apply clustering algorithms for image compression. Your task is implementing *K-means* for this purpose. **It is required you implementing the algorithms yourself rather than calling from a package.**

Formatting instruction

Input

- `pixels`: the input image representation. Each row contains one data point (pixel). For image dataset, it contains 3 columns, each column corresponding to Red, Green, and Blue component. Each component has an integer value between 0 and 255.
- `k`: the number of desired clusters. Too high value of K may result in empty cluster error. Then, you need to reduce it.

Output

- `class`: cluster assignment of each data point in pixels. The assignment should be 1, 2, 3, etc. For $k = 5$, for example, each cell of class should be either 1, 2, 3, 4, or 5. The output should be a column vector with `size(pixels, 1)` elements.
- `centroid`: location of k centroids (or representatives) in your result. With images, each centroid corresponds to the representative color of each cluster. The output should be a matrix with K rows and 3 columns. The range of values should be [0, 255], possibly floating point numbers.

Hand-in

Both of your code and report will be evaluated. Upload them together as a zip file. In your report, answer to the following questions:

1. (30 points) Compress pictures using k-means, for `beach.bmp` and `football.bmp` and also choose a third picture of your own to work on. We recommend size of 320×240 or smaller. Run your k-means implementation with these pictures, with several different $k = 2, 4, 8, 16$. How long does it take to converge for each k (report the number of iterations, as well as actual running time)? Please write in your report, and also include the resulted compressed pictures for each k.
2. (10 points) Run your k-means implementation with different initialization centroids. How does this it affect your final result? (We usually randomize initial location of centroids in general. To answer this question, an intentional poor assignment may be useful.) Please write in your report.

Note

- You may see some error message about empty clusters when you use too large k. Your implementation should treat this exception as well. That is, do not terminate even if you have an empty cluster, but use smaller number of clusters in that case.
- We recommend you to test your code with several different pictures so that you can detect some problems that might happen occasionally.
- If we detect copy from any other student's code or from the web, you will not be eligible for any credit for the entire homework, not just for the programming part. Also, directly calling built-in functions or from other package functions is not allowed.

Student Answer:

Please refer to the next pages for the code, and my associated analysis.