Machine Learning
COMP 5630/ COMP 6630/ COMP 6630 - D01

Instructor: Dr. Shubhra ("Santu") Karmaker
TA 1: Dongji Feng
TA 2: Souvika Sarkar
Department of Computer Science and Software Engineering
Auburn University
Fall, 2022

October 26, 2022

Assignment #7
K-Means Clustering

Submission Instructions

This assignment is due Tuesday, November 8, 2022, at 11:59 pm. Please submit your solutions via Canvas (https://auburn.instructure.com/). You should submit your assignment as a PDF file. Please do not include blurry scanned/photographed equations as they are difficult for us to grade.

Late Submission Policy

The late submission policy for assignments will be as follows unless otherwise specified:

1. 75% credit within 0-48 hours after the submission deadline.
2. 50% credit within 48-96 hours after the submission deadline.
3. 0% credit beyond 96 hours after the submission deadline.

Tasks

1 K-Means Implementation [50 pts]

In this problem, you will implement Lloyd's method for the k-means clustering problem and answer several questions about the k-means objective, Lloyd's method, and k-means++. Recall that given a set $S = x_1, ..., x_n \rightarrow Rd$ of n points in d-dimensional space, the goal of the k-means clustering is to find a set of centers $c_1, ..., c_k \in \mathbb{R}^d$ that minimize the k-means objective:

$$\sum_{j=1}^{n} \min_{i \in \{1,...,k\}} ||x_j - c_i||^2$$

which measures the sum of squared distances from each point x_j to its nearest center.

Consider the following simple brute-force algorithm for minimizing the k-means objective: enumerate all the possible partitionings of the n points into k clusters. For each possible partitioning, compute the optimal centers $c_1, ..., c_k$ by taking the mean of the points in each cluster and computing the corresponding k-means objective value. Output the best clustering found. This algorithm is guaranteed to output the optimal set of centers, but unfortunately, its running time is exponential in the number of data points.

a) **[10 pts]:** For the case $k = 2$, argue that the running time of the brute-force algorithm above is exponential in the number of data points n.

In class, we discussed that finding the optimal centers for the k-means objective is NP-hard, which means that there is likely no algorithm that can efficiently compute the optimal centers. Instead, we often use Lloyd's method, which is a heuristic algorithm for minimizing the k-means objective that is efficient in practice and often outputs reasonably good clusterings. Lloyd's method maintains a set of centers $c_1, ..., c_k$ and a partitioning of the data S into k clusters, $C_1, ..., C_k$. The algorithm alternates between two steps: (i) improving the partitioning $C_1, ..., C_k$ by reassigning each point to the cluster with the nearest center, and (ii) improving the centers $c_1, ..., c_k$ by setting c_i to be the mean of those points in the set C_i for $i = 1, ..., k$. Typically, these two steps are repeated until the clustering converges (i.e, the partitioning $C_1, ..., C_k$ remains unchanged after an update). Pseudocode is given below:

(a) Initialize the centers $c_1, ..., c_k$ and the partition $C_1, ..., C_k$ arbitrarily.

(b) Do the following until the partitioning C1,..., Ck does not change:

i. For each cluster-index i, let $C_i = \{x \in S : x$ is closer to c_i than any other center$\}$, breaking ties arbitrarily but consistently.

ii. For each cluster index i, let $c_i = \frac{1}{|Ci|} \sum_{x \in C_i} x$.

In the remainder of this problem, you will implement and experiment with of Lloyd's k-means clustering algorithm for image segmentation. Specifically, You will work on the

"k-means.py" python file along with the image dataset and template package provided to you. Using PIL, this program will load a selected image, represent each pixel using its RGB values and analyze pixel-by-pixel RGB values to find the centroid values of the image. K represents the number of centroids that are initialized randomly within the min and max RGB values of the image. Once the centroid values have been optimized using k-means, the program will produce and display the segmented image with the found RGB centroid values.

(a) **[10 pts]:** Complete the function $assignPixels(centroids)$. The input $centroids$ is a list of current centroids. The function assigns each pixel to the current centroids for the algorithm. Specifically, this method finds the closest centroid to the given pixel, then assigns that centroid to the pixel. The function should return a dictionary $clusters$ where the keys are the unique centroids, and values are the pixels assigned to each centroid. Note that, this process can reduce the number of clusters if there are duplicate centroids.

(b) **[10 pts]:** Complete the function $adjustCentroids(clusters)$ that is used to re-center the centroids according to the pixels assigned to each. A mean average is applied to each cluster's RGB values, which are then set as the new centroids. Output is the list of new centroids.

(c) **[10 pts]:** Complete the function $initializeKmeans(someK)$ which creates a list of K centroids and initializes them with the RGB values of randomly selected K different pixels. That means, first sample K sample pixels, i.e. (x,y) locations, read their RGB values and used those RGB values for initializing the K centroids. Finally, return the list of K centroids.

(d) **[10 pts]:** Complete the function $iterateKmeans(centroids)$ that iterates the k-means clustering steps for maximum 20 iterations. However, you can stop early if $converged(centroids, old\_centroids)$ returns $True$. Converged is a function provided to you that will help determine if the centroids have converged or not.

2 Analyze the Effect of k on Lloyd's method [20 pts]

In this part, you will investigate the effect of the number of clusters k on Lloyd's method and a simple heuristic for choosing a good value for k. Your dataset will still be the 12 different 2D images provided in the "img" folder.

(a) **[10 pts]:** Write a short script to plot the k-means objective value obtained for each value of k in $1, ..., 20$. The x-axis of your plot should be the value of k used, and the y-axis should be the k-means objective. Include both the plot and script in your written report.

(b) **[10 pts]:** One heuristic for choosing the value of k is to pick the value at the "elbow" or bend of the plot produced in part (a), since increasing k beyond this point results in a relatively little gain. Based on your plot, what value of k would you choose? Does this agree with your intuition when looking at the data? Why or why not?

3 Analyze the Effect of Initialization on Lloyd's method [30 pts]

Now, you will investigate the effect of centroid initialization on Lloyd's method. In particular, you will compare random initialization with Lloyd's method. Note that the provided k-means script has the randomized initialization already implemented. In this problem, you just need to run the code and answer questions.

a) **[5 pts]:** For test image 5, set $k = 2$ and run Lloyd's method 10 times (each time with random initialization) and save the 10 output plots. What do you see? Can you explain the output?

b) **[5 pts]:** For test image 5, set $k = 10$ and run Lloyd's method 10 times (each time with random initialization) and save the 10 output plots. What do you see? Can you explain the output?

c) **[10 pts]:** For test image 8, set $k = 2$ and run Lloyd's method 10 times (each time with random initialization) and save the 10 output plots. Pick an image that shows the highest contrast. Repeat this process for $k = \{3, 4, 5\}$. Again, pick the highest contrast image for each k. Add the highest contrast images to your report. You should have 4 images.

d) **[10 pts]:** Based on all these experiments, propose some heuristics for initialization of centroids (apart from random initialization) which can help achieve meaningful segmentation results.

Disclaimers: This assignment re-uses some materials from the publicly available website: CMU Introduction to Machine Learning Course, 10-315, Spring 2019. I personally thank Prof. Maria-Florina Balcan for sharing her teaching materials publicly. This assignment is exclusively used for instructional purposes.