SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

01.112 Machine Learning, Fall 2019
Homework 2

Due 19 Oct 2019, 11:59 pm

This homework will be graded by Chen Zihan

1. K-Means [30 points]

Consider the data in the file "hw2-image.txt". This file contains a large number (210,012) of length 3 vectors, each on one line. Each vector represents the red, green, and blue intensity values of one of the pixels in the image shown (Fig 1). The image has 516 rows and 407 columns. The pixels in the file are listed row by row from top to bottom, and within each row from left to right. For example, the first pixel in the file is the uppermost left pixel in the image. The second line of the file contains the pixel to the right of that one, and so on. In this assignment, we will explore clustering methods, applying them in particular to the problem of dividing the pixels of the image into a small number of similar clusters. Consider the K-means clustering algorithm, as described in class. In particular, consider a version in which the inputs to the algorithm are:

- The set of data to be clustered. (i.e., the vectors $x^{(1)}, x^{(2)}, x^{(3)}, ...$)
- The desired number of clusters, K.
- Initial centroids for the K clusters.

Then the algorithm proceeds by alternating: (1) assigning each instance to the class with the nearest centroid, and (2) recomputing the centroids of each class—until the assignments and centroids stop changing. Please use squared Euclidean distance (Lecture 5, Eq. 2) as the metric for clustering.

There are many implementations of K-means publicly available. However, please implement K-Means on your own. Then, use your implementation to cluster the data in the file mentioned above ("hw2-image.txt"), using K = 8, and the initial centroids as given below in the table:

R	G	B
255	255	255
255	0	0
128	0	0
0	255	0
0	128	0
0	0	255
0	0	128
0	0	0

Turn in your code, as well as a report on all of the following:

(a) How many clusters there are in the end. (A cluster can "disappear" in one iteration of the algorithm if no vectors are closest to its centroid.)

(b) The final centroids of each cluster.

(c) The number of pixels associated to each cluster.

(d) Plot the sum of squared Euclidean distance of each pixel to the nearest centroid (Lecture 5, Eq. 8) against the iteration number of the algorithm.

Visualize your result by replacing each pixel with the centroid to which it is closest, and displaying the resulting image.

See implementation attached.

2. K-Mediods [10 points]

In clustering, Euclidean distance is not the only way to measure the distance between two points/vectors. l_p norms is a family of distance measures that are parameterized by $p \geq 1$. The l_p norm of a vector is:

$$\|x\|_p = \left(\sum_j |x_j|^p \right)^{\frac{1}{p}}.$$

Euclidean distance is the l_2 norm of the vector difference between two points, i.e.,

$$\|x - y\|_2 = \left(\sum_j |x_j - y_j|^2 \right)^{\frac{1}{2}}.$$

The Manhattan distance is the l_1 norm of the vector difference between two points, i.e.,

$$\|x - y\|_1 = \sum_j |x_j - y_j|.$$

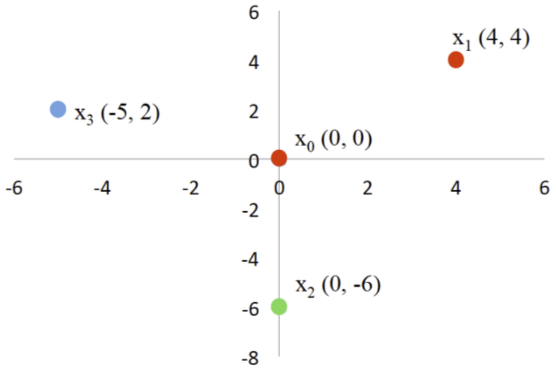
The l_∞ distance is the maximum absolute element in the vector difference between two points, i.e.,

$$\|x - y\|_\infty = \max_j |x_j - y_j|.$$

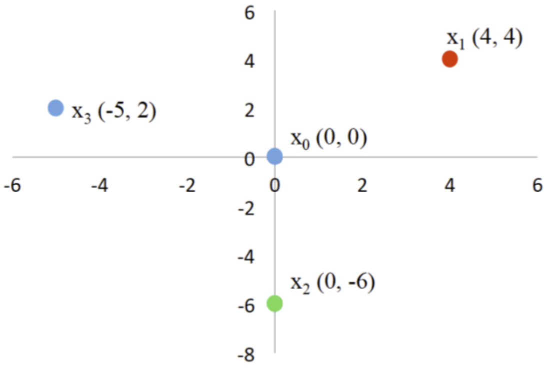
The following figures (points in the same cluster have the same color) are produced by the k-medoids algorithm for $k = 3$ clusters using l_1 , l_2 , and l_∞ distance measures. Indicate which distance measure is used for each figure.

$A : l_\infty, B : l_2, C : l_1$

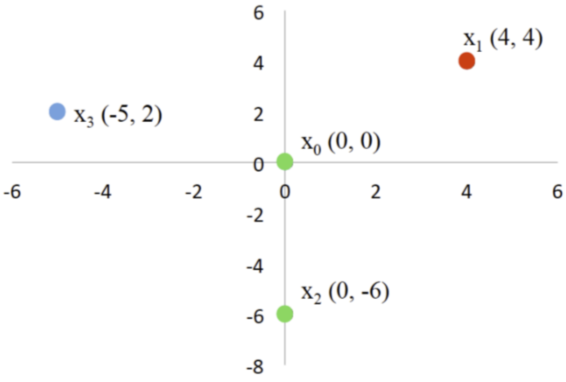
A.



B.



C.



3. K-Means vs K-Mediods [10 points]

K-means clustering creates cluster centroids that do not correspond to any real data points whereas, K-Mediods selects real data points as cluster centers. What are the advantages and disadvantages of K-medoids, compared to K-means?

Advantages:

(a) K-mediods is applicable to arbitrary objects and distance functions (e.g. categorical data) which is not possible for k-means.

(b) It is less sensitive to noisy data when compared to k-means since a mediod is less influenced by outlier than a mean.

(c) It can be applied even when true data points are not available and only their pair-wise distances are provided, unlike k-means.

(d) It can be applied to both continuous and discrete domains as against k-means which can be applied only to the continuous domain since the mean is not necessarily a data point.

Disadvantages:

(a) K-mediods algorithm has longer run time than k-means, especially for large and random datasets.

(b) The mean value has a true geometrical and statistical meaning unlike the mediod which are neither medians nor geometric median.