Stat 432 Homework 2

Assigned: Feb 3, 2019; Due: 11:59pm Feb 8, 2019

Question 1 (understand k-means)

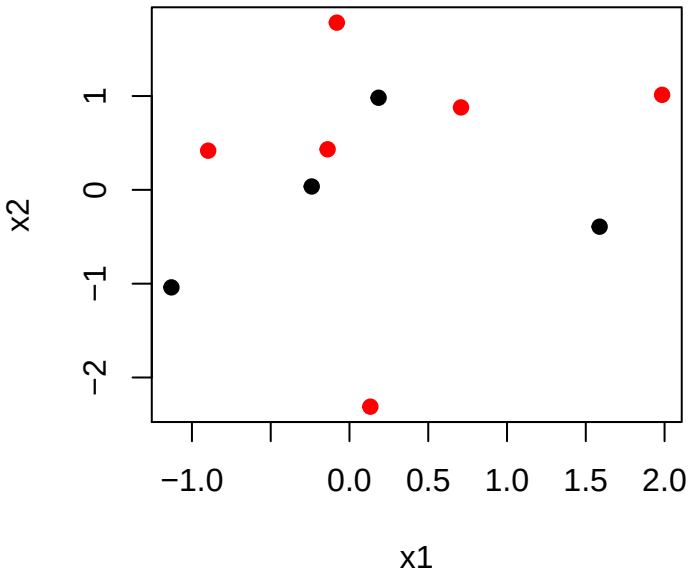
k-means is a relatively simple algorithm that can write by ourselves. For this question, you are not allowed to use any existing functions that perform k-means directly. Let's first generate some data. You should copy this exact code to generate the same dataset and the initial cluster assignment labels.

```
set.seed(2)
n = 10

# first coordinate (variable) of each observation
x1 = rnorm(n)

# second coordinate (variable) of each observation
x2 = rnorm(n)

# we also generate an initial value of the cluster assignments
C = sample(1:2, n, replace = TRUE)
C
```

```
##  [1] 2 1 1 1 2 2 2 1 2 2
```

The above code means that we consider just two clusters, and if `C[i] = 1`, we are currently assigning observation `i` to cluster 1, otherwise, its assigned to cluster 2. Hence, we can view this vector C as a cluster assignment function. To visualize the current cluster assignment, you can do the following:

```
plot(x1, x2, col = C, pch = 19)
```



We know that in each iteration of the k-means algorithm, we first fix the cluster assignment function and update the cluster means m_k, for $k = 1, \ldots, K$; then, fix the cluster means and update the cluster assignment function.

a. [2 points] Do this iteration once, and output the new cluster assignment function (both the value of the vector `C` and plot it) and the cluster means for both clusters.

b. [2 points] Write the above two steps into a single function. Repeatedly call this function to update `C` and the cluster means. When they do not change anymore, stop the algorithm. You should not have an excessively long output for this part. Only output the final result.

c. [2 points] Based on your final result, calculate and report the within-cluster distance of the k-mean algorithm, which is also the objective function used for k-means.

d. [2 points] Randomly generate another set of initial values for `C` and repeat the above steps. Observe if the two runs lead to the same clustering result. Comment on your findings.

e. [2 points] Apply any clustering algorithm discussed in the lecture other than k-means on the same data set. Compare the result by using this algorithm with what you got by using k-means.

Question 2 (bonus: k-means of a picture)

Pick your favorite picture and perform a k-means clustering of the pixels (cluster the 3d points of RGB colors). You can use any exsiting functions for this question. Your picture should not be too large, ideally less than 500×500 pixels. You can shrink your original picture to fit this size. You should perform the following steps. If you need an example of this, read the `R` supplementary file of k-means. You are free to google the topic `clustering images` and use existing codes.

a. [1 points] Plot the original picture.

b. [2 points] For $k = 2, 3, 5, 10$, perform k-means on the pixels and plot the resulting picture by replacing each pixel with their corresponding cluster mean.

c. [1 points] If you have to choose a best value for k (not limited to the four settings in the previous question), what would you do? Explain the basis of your decision.