Homework 3
Introduction to Data Analysis and Mining
Spring 2018
CSCI-B 365

Instructor: Hasan Kurban

March 14, 2018

Directions

Please follow the syllabus guidelines in turning in your homework. I am providing the LaTeX of this document too. This homework is due Monday, Feb 26, 2018 10:00p.m. **OBSERVE THE TIME**. Absolutely no homework will be accepted after that time. All the work should be your own. Within a week, AIs can contact students to examine code; students must meet within three days. The session will last no longer than 5 minutes. If the code does not work, the grade for the program may be reduced. Lastly, source code cannot be modified post due date.

k-means Algorithm in Theory

This part is provided to help you implement k-means clustering algorithm.

1: **ALGORITHM** `k-means`
2: **INPUT** (data Δ, distance $d : \Delta^2 \rightarrow \mathbb{R}_{\geq 0}$, centoid number k, threshold τ)
3: **OUTPUT** (Set of centoids $\{c_1, c_2, \ldots, c_k\}$)
4:
5: *** $Dom(\Delta)$ denotes domain of data.
6:
7: *** Assume centroid is structure $c = (v \in DOM(\Delta), B \subseteq \Delta)$
8: *** $c.v$ is the centroid value and $c.B$ is the set of nearest points.
9: *** c^i means centroid at i^{th} iteration.
10:
11: $i = 0$
12: *** Initialize Centroids
13: **for** $j = 1, k$ **do**
14: $\quad c_j^i.v \leftarrow random(Dom(\Delta))$
15: $\quad c_j^i.B \leftarrow \emptyset$
16: **end for**
17:
18: **repeat**
19: $\quad i \leftarrow i + 1$
20: $\quad$ *** Assign data point to *nearest* centroid
21: $\quad$ **for** $\delta \in \Delta$ **do**
22: $\quad\quad c_j^i.B \leftarrow c.B \cup \{\delta\}$, where $\min_{c_j^i}\{d(\delta, c_j^i.v)\}$
23: $\quad$ **end for**
24: $\quad$ **for** $j = 1, k$ **do**
25: $\quad\quad$ *** Get size of centroid

26: $n \leftarrow |c_j^i.B|$
27: *** Update centroid with average
28: $c_j^i.v \leftarrow (1/n) \sum_{\delta \in c_j^i.B} \delta$
29: *** Remove data from centroid
30: $c_j^i.B \leftarrow \emptyset$
31: **end for**
32: *** Calculate scalar product (abuse notation and structure slightly)
33: *** See notes
34: **until** $((1/k) \sum_{j=1}^{k} \ ||c_j^{i-1} - c_j^i||) < \tau$
35: **return** $(\{c_1^i, c_2^i, \ldots, c_k^i\})$

k-means on a tiny data set.

Here are the inputs:

$$\Delta = \{(2,5),(1,5),(22,55),(42,12),(15,16)\} \quad (1)$$
$$d((x_1,y_1),(x_2,y_2)) = [(x_1-x_2)^2 + (y_1-y_2)^2)]^{1/2} \quad (2)$$
$$k = 2 \quad (3)$$
$$\tau = 10 \quad (4)$$

Observe that $Dom(\Delta) = \mathbb{R}^2$. We now work through k-means. We ignore the uninformative assignments. We remind the reader that T means transpose.

$i \leftarrow 0$
*** Randomly assign value to first centroid.
$c_1^0.v \leftarrow random(Dom(\Delta)) = (16, 19)$
*** Randomly assign value to second centroid.
$c_2^0.v \leftarrow random(Dom(\Delta)) = (2, 5)$
$i \leftarrow i + 1$
*** Associate each datum with nearest centroid
$c_1^1.B = \{(22,55),(42,12),(15,16)\}$
$c_2^1.B = \{(2,5),(1,5)\}$
*** Update centroids
$c_1^1.v \leftarrow (26.3, 27.7) = (1/3)((22,55) + (42,12) + (15,16))$
$c_2^1.v \leftarrow (1.5, 5) = (1/2)((2,5) + (1,5))$
*** The convergence condition is split over the next few lines to explicitly show the calculations
$(1/k) \sum_{j=1}^{k} \ ||c_j^{i-1} - c_j^i|| = (1/2)(||c_1^0 - c_1^1|| + ||c_2^0 - c_2^1||) = (1/2)(||\binom{2}{5} - \binom{1.5}{5}|| + ||\binom{16}{19} - \binom{26.3}{27.7}||)$
$= (1/2)[(\binom{.5}{0}^{\mathsf{T}} \binom{.5}{0})^{(1/2)} + (\binom{-9.7}{-8.7}^{\mathsf{T}} \binom{-9.7}{-8.7})^{(1/2)})] = (1/2)(\sqrt{.5} + \sqrt{169.7}) \sim (1/2)(13.7) = 6.9$
Since the threshold is met $(6.9 < 10)$, k-means stops, returning $\{(26.3, 27.7), (1.5, 5)\}$

Problem 1 [10 points]

Answer the following questions for Ionosphere Data Set:

1.1 Briefly describe this data set–what is its purpose? How should it be used? What are the kinds of data it's using?

The data is collecting the information of radar signal, and the system that recieved the data consist of a phased array of 16 high frequency antennas.

First 34 attributes are all numeric values, and the last attribute is a character, which is either "g" or "b".

1.2 Using R, show code that answers the following questions:

1.2.1 How many entries are in the data set?

```
entries <- nrow(mydata)
> entries
[1] 351
```

1.2.2 How many unknown or missing data are in the data set?

```
> sum(is.na(mydata))
[1] 0
```

1.2.3 Create a bar plot of 1st, 2nd, and 35th variables. Label the plots properly. Discuss the distribution of values *e.g.*, are uniform, skewed, normal. Place images of these bar plots into the document. Show the R code that you used below and discussion below that.

```
barplot(table(mydata$V1), xlab = "True or False", ylab = "Count")
barplot(table(mydata$V2), xlab = "Value", ylab = "Count")
barplot(table(mydata$V35), xlab = "Good or Bad", ylab = "Count")
```

1.2.4 Make a scatter plots of $[V22, V20]$ and $[V1, V2]$ variables and color the data points with the class variable $[V35]$. Discuss the plots, i.e., do you observe any relationships between variables?

```
plot(mydata$V22,mydata$V20,col= ifelse(mydata$V35 == 'g', 'red', ifelse(mydata$V35
plot(mydata$V1,mydata$V2,col= ifelse(mydata$V35 == 'g', 'red', ifelse(mydata$V35 =
```

Problem 2 [10 points]

The pseudo-code for k-means and a running example of k-means on a small data set are provided above. Answer the following questions:

2.1 Does k-means always converge? Given your answer, a bound on the iterate must be included. How is its value determined?

By definition, $k - means$ always converge. Given a data set, we first initialize n^{th} random centroids and set a termination condition, then assign data point to its closet centroid. Check whether the data meets the termination condition (after i^{th} iteration or already optimitized). If the condition is not met, find the mid point of each centroid data set, and repeat steps above.

2.2 What is the run-time of this algorithm?

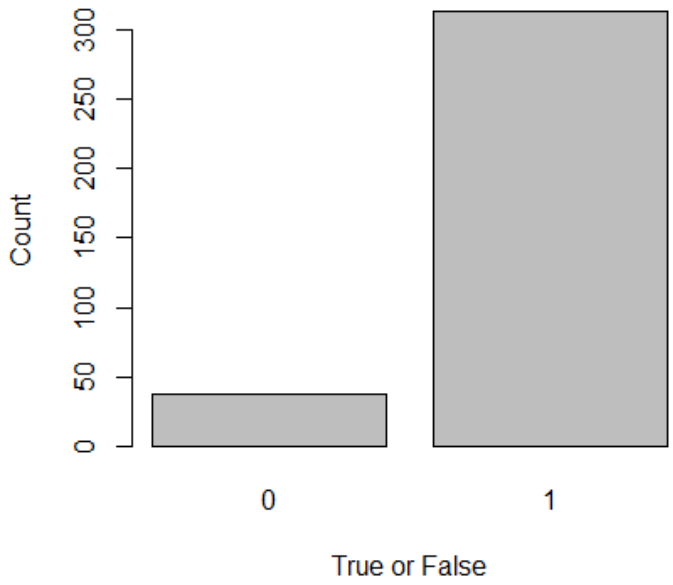

Figure 1: V1 barplot

Figure 2: *

right skewed distribution

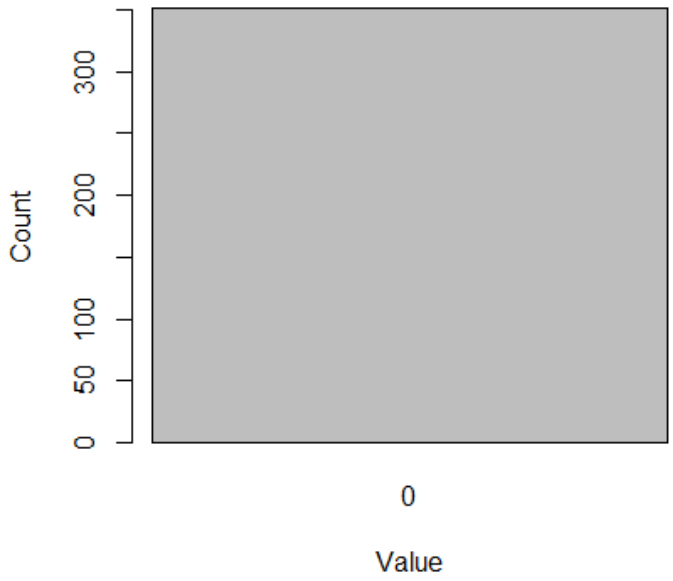

Figure 3: V2 barplot

Figure 4: *

uniform distribution

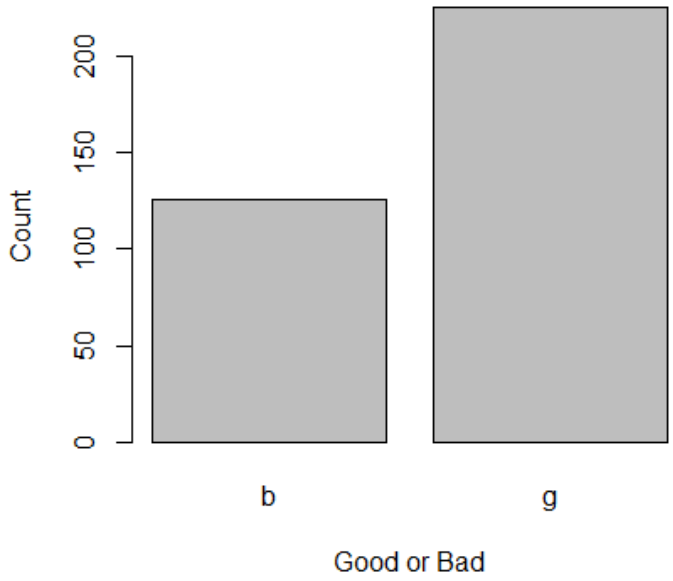

Figure 5: V35 barplot

Figure 6: *

right skewed distribution

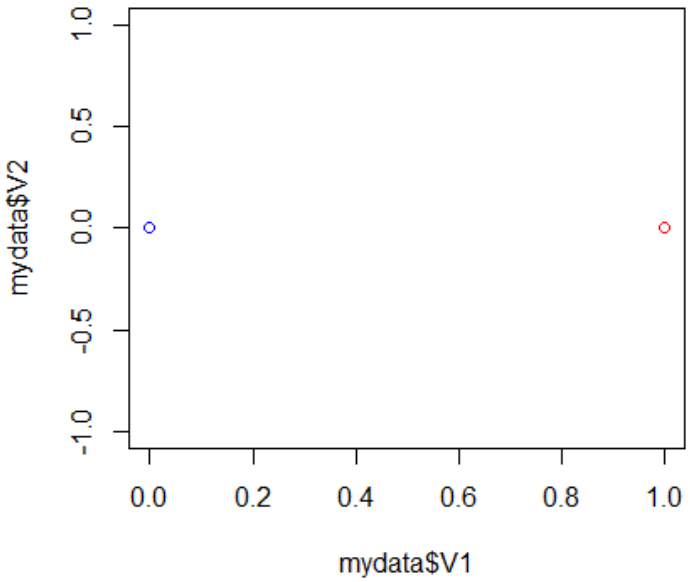

Figure 7: V1 V2 scatterplot

Figure 8: *

There are only two points on the graph, "good" or "bad" value depends on variable 1, if V1 equals 1, then it is good value. If V1 equals 0, then it is bad value.

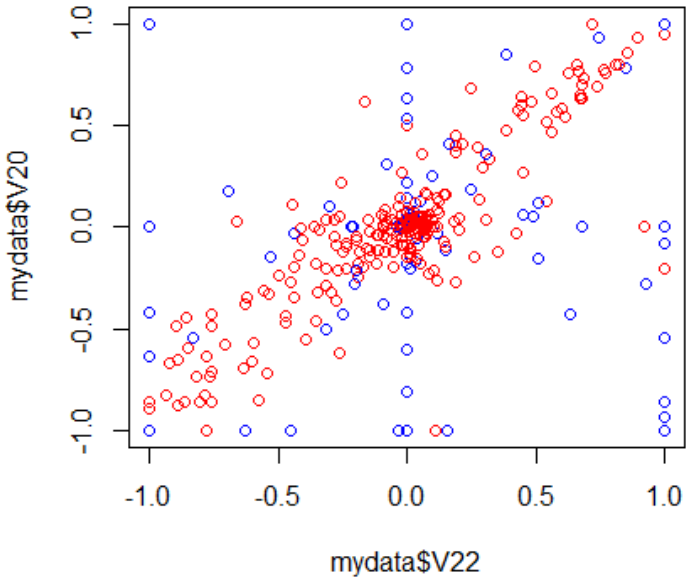

Figure 9: V20 V22 scatterplot

Figure 10: *
All the good value are linearly correlated, and the bad values are not correlated

Let n be the number of data point, i be the number of iteration. Complexity is O(n * K * I * d), n = number of points, K = number of clusters, I = number of iterations, d = number of attributes

Problem 3 [20 points]

Implement Lloyd's algorithm for k-means (see algorithm k-means above) in R and call this program C_k. As you present your code explain your protocol for

3.1 initializing centroids

3.2 maintaining k centroids

3.3 deciding ties

3.4 stopping criteria

Problem 4 [40 points]

In this question, you are asked to run your program, C_k, against Ionosphere Data Set. Upon stopping, you will calculate the quality of the centroids and of the partition. For each centroid c_i, form two counts:

$$g_i \leftarrow \sum_{\delta \in c_i.B} [\delta.C = \text{“g”}], \quad \text{good}$$

$$b_i \leftarrow \sum_{\delta \in c_i.B} [\delta.C == \text{“b”}], \quad \text{bad}$$

where $[x = y]$ returns 1 if True, 0 otherwise. For example, $[2 = 3] + [0 = 0] + [34 = 34] = 2$

The centroid c_i is classified as good if $g_i > b_i$ and bad otherwise. We can now calculate a simple error rate. Assume c_i is good. Then the error is:

$$error(c_i) = \frac{b_i}{b_i + g_i}$$

We can find the total error rate easily:

$$Error(\{c_1, c_2, \ldots, c_k\}) = \sum_{i=1}^{k} error(c_i)$$

Report the total error rates for $k = 2, \ldots 5$ for 20 runs each, presenting the results that are easily understandable. Plots are generally a good way to convey complex ideas quickly, i.e., box plot. Discuss your results.
As we can see from the graph, No matter how many clusters there are, the error rate will always be the same.

Problem 5 [10 points]

In this question, you are asked to make use of the R package for k-means. Elbow method is one of the techniques to decide the optimal cluster number. Find the optimal cluster number for Ionosphere data set using elbow method (for $2 \leq k \leq 15$). Provide a plot that shows the total SSE for each k. Discuss your results.

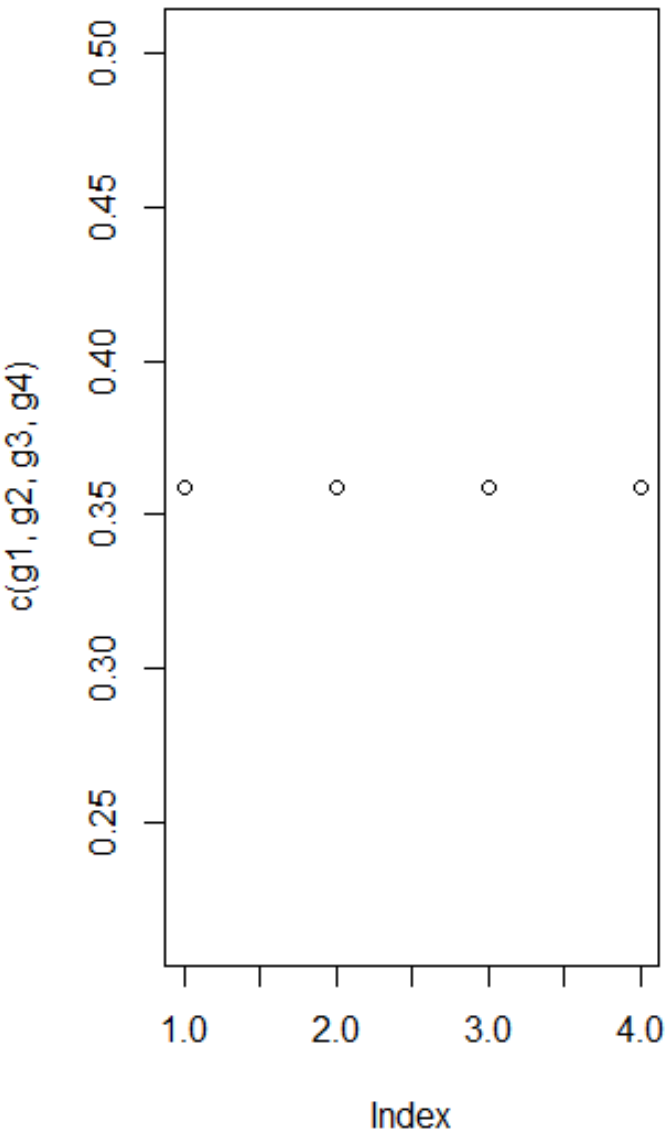

Figure 11: SSE

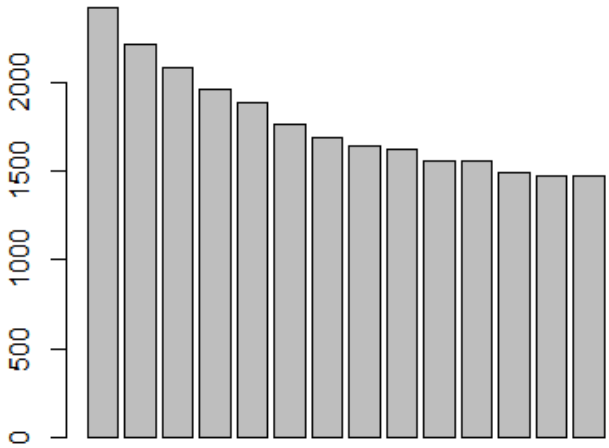

Figure 12: SSE

Problem 6 [10 points]

Let $X \subset \mathbb{R}^n$ ($\mathbb{R}$ is the set of reals) for positive integer $n > 0$. Define a distance $d : X \times X \to \mathbb{R}_{\geq 0}$ as

$$d(x, y) = max\{|x_i - y_i|\}, \forall i \ 1 \leq i \leq n$$

Prove/disprove d is a metric?

1.$(Nonnegativity) : as|a - b|isalways >= 0, it'sTrue$
2.$(Definiteness) : If|a - b| = 0, amustequalsb.SothisisTrue$
3.$(Symmetry) : Since|a - b| = a - b = b - a.Therefore, d(x, y) = d(y, x)$
4.$d(x, y) \leq d(y, z) + d(x, z)$
$max|x - z| \leq max|x - y| + |y - z|$
$max|x - y| + max|y - z| = d(x, y) + d(y, z)$

Extra credit [30 points]

This part is optional.

1 The k-means algorithm provided above stops when centroids become stable (Line 34). In theory, k-means converges once SSE is minimized

$$SSE = \sum_{j}^{k} \sum_{x \in c_j.B} ||\mathbf{x} - c_j.v||_2^2$$

In this question, you are asked to use SSE as stopping criterion. Run k-means over Breast Cancer Wisconsin Data Set and report the total SSE in a plot for $k = 2, \ldots, 5$ for 20 runs each. Discuss your results. [15 points].

2 Traditional k-means initialization is based on choosing values from a uniform distribution. In this question, you are asked to improve k-means through initialization. k-means ++ is an extended k-means clustering algorithm and induces non-uniform distributions over the data that serve as the initial centroids. Read the paper and discuss the idea in a paragraph. Implement this idea to improve your k-means program. [15 points]

What to Turn-in

Submit a .zip file that includes the files below. Name the .zip file as "usename-section number", i.e., hakurban-B365.

- The *tex and *pdf of the written answers to this document.
- *`R`files for:
  - implementation of k-means [Problem 4]
  - `R` package usage [Problem 5]
  - extra credit questions–optional
- A README file that explains how to run your code and other files in the folder