CSCI-B 565 DATA MINING
Homework 2
Computer Science Core
Spring
Indiana University,
Bloomington, IN

Yuanzhi Bao
baoyu@indiana.edu

Sept.23 2016

All the work herein is solely mine.

Questions

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

   (a) k-means always converge but not guarantee global optimum. k-means need exponentially iterations if we dont give it a threshold to tell it when to stop.

   (b) The bound of iterations is determined by a finit number. For the reason that k-means will difinetly give us a converger result. We can set up a large number based on the dataset we are facing so when the iteration number hits the number, k-means can stop.

2. Lines 12-16 of the k-means algorithm describe initialization of the centroids. Why is this code problematic? What are some implications of using k-means?
   Answer:

   (a) We don't know k-value at first, all we can do is guessing, which makes this code problmatic.

   (b) i. Advantages
      A. If we have a large set of base number, then k-means would easily give us a faster computational result then hierarchical clustering.
      B. k-means could give us a tighter clusters then hierarchical clustering, especially when the clusters are globular.

   ii. Disadventages
      A. It's very difficult to predict k value
      B. It doesn't work well with global cluster.
      C. It could end up with different results when we gave different initial partitions.
      D. It's a NP hard cluster, which means even if we get a good result, we still dont know if there is better result out there.

3. What is the run-time of this algorithm (include your new parameter from Question 1).
   Answer:
   $O(n * K * I)$
   n: number of points
   k: number of clusters
   I: number of iterations

4. We describe two problems that arise when using k-means in practice. Assume the datum is $\delta \in \Delta$, the centroids are c_i, c_j for $i \neq j$ and distance d.

   - *Ties* occur when $d(c_i, \delta) = d(c_j, \delta)$. Of course, there can be threeway, fourway, ..., k-way ties. One solution is to randomly assign the datum to one of the two centroids. What are two other solutions to this problem?
     Answer:
     (a) We can simply assign the datum to the "first best", which means when $Ties$ happens, just assin the datum to the first one.
     (b) We can also simply assign the datum to the "last best", which means when $Ties$ happens, just assin the datum to the last one.
   - *Centroid collapse* occurs when $d(c_i, c_j) \sim 0$. Like ties, this can include more than two. One is to find the median m of the union of the two centroids and then assign values less than the median to one and values greater than the median to the other, taking into account an odd number will be the problem above. What are two other solutions? Observe that an additional threshold on centroids, $\tau_c > 0$, is needed, to determine whether $d(c_i, c_j) \leq \tau_c$ is true. First, how would τ_c be determined? Second, where in the algorithm should this be checked?
     Answer:
     (a) Two other solutions:
       i. We can simply just delete all the other same centroids and assign the all data to just one.
       ii. We can also just ramdonly pick $1/n$ (n is the number of the same centroids) to every centroids which have tha same value
     (b) To detemrmine τ_c is as same as to determine how seperate you want to centtroids to be.
     If you want them to be more seperate than τ_c should be reasonly big, otherwise τ_c should be reasonly small.
     (c) It should be before we update centriods with average. So we can implement this before line 25 and after line 24
   - Modify the k-means algorithm to address ties and collapsing centroids. Explicitly add pseudo-code to the algorithm and call this k-meansr.

```
1: ALGORITHM k-meansr
2: INPUT (data Δ, distance d : Δ^2 → R_≥0, centoid number k, threshold τ)
3: OUTPUT (Set of centoids {c_1, c_2, ..., c_k})
4:
5: *** Dom(Δ) denotes domain of data.
6:
7: *** Assume centroid is structure c = (v ∈ DOM(Δ), B ⊆ Δ)
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:     c_j^i.v ← random(Dom(Δ))
```

15: $c_j^i.B \leftarrow \emptyset$
16: **end for**
17:
18: **repeat**
19: $i \leftarrow i + 1$
20: *** Assign data point to *nearest* centroid
21: **for** $\delta \in \Delta$ **do**
22: **if** t **then**here are more than one min centroids which $d(\delta, c_j^i.v) = d(\delta, c_k^i.v)$
23: $r \leftarrow$ random chose one from all the centroids which have the same d
24: $c_r^i.B \leftarrow c.B \cup \{\delta\}$
25: **else**
26: $c_j^i.B \leftarrow c.B \cup \{\delta\}$, where $\min_{c_j^i}\{d(\delta, c_j^i.v)\}$
27: **end if**
28: **end for**
29: **for** $j = 1, k$ **do**
30: *** Check for *Centroid collaspe*
31: **if** T **then**here are more than one centroids $d(c_i, c_j) \sim 0$
32: Randomly choose $1/n$(n is the number of the centriods) to every centriods.
33: **end if**
34: *** Get size of centroid
35: $n \leftarrow |c_j^i.B|$
36: *** Update centroid with average
37: $c_j^i.v \leftarrow (1/n)\sum_{\delta \in c_j^i.B} \delta$
38: *** Remove data from centroid
39: $c_j^i.B \leftarrow \emptyset$
40: **end for**
41: *** Calculate scalar product (abuse notation and structure slightly)
42: *** See notes
43: **until** $((1/k)\sum_{j=1}^{k} ||c_j^{i-1} - c_j^i||) < \tau$
44: **return** $(\{c_1^i, c_2^i, \ldots, c_k^i\})$

Integration

We will look at the problem of integrating two pieces of data through a metric. The data are described by $([X : t], d_x), ([Y : u], d_u)$ where $X : t$ means it is type t, $Y : u$ is type u, and d_x, d_y distance metrics. We integrate the data and now need a metric $([X : t] \times [Y : u], d)$. Is this possible? We need to prove that d is a metric. To make notation easier, assume $Z = [X : t] \times [Y : u]$. For $(a, b) \in Z^2$, we write a_0 to mean the t type leftside of the product and b_0 for the t type rightside. For example, $Z = [N : \mathsf{int}] \times [S : \mathsf{string}]$. $(a, b) = ((34, \texttt{two}), (100, \texttt{three}))$, then $a_0 = 34, b_0 = 100$ and $a_1 = \texttt{two}, b_1 = \texttt{three}$.

Let's define one of the simplest metrics. $d : Z^2 \to \mathbb{R}_{\geq 0}$ where:

$$d(a, b) \quad = \quad d_x(a_0, b_0) + d_y(a_1, b_1)$$

Now we show reflexivity, symmetry, and transitivity.

- $(\forall, a \in Z)\ d(a, a) = 0$. Then $d(a, a) = d_x(a_0, a_0) + d_y(a_1, a_1) = 0$
- $(\forall\, a, b)\ d(a, b) \to d(b, a)$.

$$d(a, b) = d_x(a_0, b_0) + d_y(a_1, b_1) = d_x(b_0, a_0) + d_x(b_1, a_1) = d(b, a)$$

- $(\forall\, a, b, c)\ d(a, b) + d(b, c) \geq d(a, c)$

$$\begin{aligned} d(a, b) + d(b, c) &= d_x(a_0, b_0) + d_x(b_0, c_0) + d_y(a_1, b_1) + d_y(b_1, c_1) \\ &\geq d_x(a_0, c_0) + d_y(a_1, c_1) = d(a, c) \end{aligned}$$

Suppose we have $[X : \mathsf{int}]$ are the number of cable subscription cancelations (say, *per* hour). We find data $[Y : \mathsf{char}]$ that indicates whether there was "good" programming at that time (we're purposely being vague). The ordering is $\mathtt{n} < \mathtt{o} < \mathtt{g} < \mathtt{e}$, $\mathtt{e}$ being the best. We integrate this and get:

X	Y
14	g
45	o
54	g
21	n
60	o

Although we didn't need to use the type information explicitly, its presence shows that we can build metrics over disparate kinds of integrated data. Design a simple metric, different from the one above, for this integrated data. Prove it is a metric.
Answer:

1. Design a simple metric:

   (a) We can assume that

$$d_x(i,j) = \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases} \quad \text{For objects } i, j. \tag{1}$$

(2)

$$d_y(i,j) = \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases} \quad \text{For objects } i, j. \tag{3}$$

(4)

   (b) The integrate matric

$$d(a,b) = d_x(a_0,b_0) + 3 * d_y(a_1,b_1)$$

2. Prove:

   (a) refexivity: This is obvious
   $(\forall, a \in Z)\ d(a,a) = 0$. Then $d(a,a) = d_x(a_0,a_0) + 3 * d_y(a_1,a_1) = 0$

   (b) symmetry: This is obvious too
   $(\forall\, a, b)\ d(a,b) \to d(b,a)$.

$$d(a,b) = d_x(a_0,b_0) + 3 * d_y(a_1,b_1) = d_x(b_0,a_0) + 3 * d_y(b_1,a_1) = d(b,a)$$

   (c) transitivity:
   $(\forall\, a, b, c)\ d(a,b) + d(b,c) \geq d(a,c)$
   $d(a,b) = d_x(a_i,b_i) + 3 * d_y(a_j,b_j)$
   $d(b,c) = d_x(b_i,c_i) + 3 * d_y(b_j,c_j)$
   $d(a,c) = d_x(a_i,c_i) + 3 * d_y(a_j,c_j)$
   Based on the matric above we can simplify the provef to prove
   $d(a,b) + d(b,c) \geq d(a,c)$
   if $a = b = c$
   $d(a,b) + d(b,c) = d(a,c) = 0$
   if $a \neq b \neq c$
   $d(a,b) + d(b,c) = 2 > d(a,c)$
   if $a \neq b, b \neq c, a \neq c$
   $d(a,b) + d(b,c) = 2 > d(a,c) = 0$
   There is impossbile that $a = b, b = c, a \neq c$ So it has transitivity.

1. We can combine multiple metrics to built more sophisticated measures of dissimilarity. This problem has to do with different metrics over the same data. Let $x = \{a, b, c, d\}, y = \{a, b, e\}, z = \{b, f\}, w = \{a, d, f, e\}$. Here are several metrics:

$$
\begin{aligned}
d_1(x, y) &= \begin{cases} 0, & x = y \\ 1, & x \neq y \end{cases} \quad \text{For objects } x, y. \\
J(x, y) &= |x \cap y| / |x \cup y| \quad \text{For sets } x, y. \\
d_2(x, y) &= 1 - J(x, y) \quad \text{For sets } x, y. \\
c(x, y) &= \begin{cases} 0, & x = y \\ 1, & otherwise \end{cases} \quad \text{for individual characters, } e.g., \texttt{a} = \texttt{b} \\
d_3(\mathbf{x}, \mathbf{y}) &= \Sigma_{i=0}^{n-1} c(\mathbf{x}[i], \mathbf{y}[i]) \quad n = ||\mathbf{x}||, \text{the length of the string.} \\
d_4(\mathbf{x}, \mathbf{y}) &= \left| \frac{\mathbf{x}^\mathsf{T} \mathbf{y}}{||\mathbf{x}|| \, ||\mathbf{y}||} \right| \quad \text{for vectors } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n
\end{aligned}
$$

Calculate the following:

(a) For every i, find $d_i(x, w)$
$d_1(x, w) = 0$ Since the x, w don't contain the same elements
$d_2(x, w) = 1 - 1/3 = 2/3$
$d_3(x, w) = 0 + 1 + 1 + 1 = 3$
$d_4(x, w) = (a^2 + bd + cf + ed)/(\sqrt{a^2 + b^2 + c^2 + d^2} * \sqrt{a^2 + d^2 + f^2 + e^2})$

(b) Find the d_i that has the minimum value for x, z.

i. $d_1(x, z) = 1$
ii. $d_2(x, z) = 1 - 1/5 = 4/5$
iii. $d_3(x, z) > 1$
iv. $d_4(x, z)$ is a range of 0 to 1
v. Conclusion:
Since d_2 is for sure number d_4 is not. The minimum value shoule be d_2

(c) Which distance gives the the maximum value for any pairs?
Answer:
d_3 gives the maximum value for any pairs for the reason that others all have the limitation of 1, but d_3 gives value bigger than 1

(d) True or False. For any set v, $d_1(v, v) = d_2(v, v) = d_3(v, v) = d_4(v, v)$.
Answer:
False
look at the a problem we can say it's false.

2. We have shown that metrics can be combined. Why is the important to integration? Prove or disprove the following are metrics (using d_i from above):
Answer:
For the reason that when we intergration different sets, we can have the distance between different sets. For this reason we can combine different features that represent one data. So we can make more accurate answer or sophisticated answer.
To prove or disprove the following are matrics we have to prove d_1 *to* d_4 is matric or not. The prove statement are show following.

(a) d_1

i. refexivity:
$d_1(a, a) = 0$

ii. symmetry: :
if $a = b$
$d_1(a, b) = d_1(b, a) = 0$
if $a \neq b$
$d_1(a, b) = d_1(b, a) = 1$
proved

iii. transitivity:
if $a = b = c$
$d_1(a, c) = d_1(a, b) + d_1(b, c) = 0$
if $a \neq b \neq c$
$d_1(a, c) = 1 < d_1(a, b) + d_1(b, c) = 2$
if $a \neq b = c$ or $a = b \neq c$
$d_1(a, c) = 1 = d_1(a, b) + d_1(b, c) = 1$
proved

(b) d_2

i. refexivity:
$d_2(a, a) = 0$

ii. symmetry: :
$d_2(a, b) = d_2(b, a)$ for the reason the same elements in two sets won't change when they change the order. So the answer remind the same.

iii. transitivity:
As we know its Jaccard distance, we can say that
$d_2(a, b) + d_2(b, c) \geq d_2(a, c)$

(c) d_3

i. For the reason that if x and y are not the same length, d_3 became a not well defined "metric". Which made it not a metric.

(d) d_4

i. refexivity:
$d(a, a) = 1$
So it's not a metric.

Now we can prove or disprove follows are metric or not

(a) $d_{i'}(x, y) = \frac{d_i(x,y)}{1+d_i(x,y)}$ for every i.

i. d_1

A. refexivity:
$d(a, a) = \frac{d_(a,a)}{1+d_(a,a)} = \frac{0}{1+0} = 0$
proved

B. symmetry:
if $a = b$
$d(a, b) = d(b, a) = 0$ if $a \neq b$
$d(a, b) = d(b, a) = 1/2$
proved

C. transitivity:
if $a = b = c$
$d(a, b) = d(b, a) = d(a, c) = 0$
if $a \neq b = c$
$d(a, b) + d(b, c) = 1/2 + 0 = 1/2 > d(a, c) = 0$
if $a \neq b \neq c$
$d(a, b) + d(b, c) = 1/2 + 1/2 = 1 > d(a, c) = 1/2$
if $a = b \neq c$
$d(a, b) + d(b, c) = 0 + 1/2 = 1/2 = d(a, c) = 1/2$
proved So all cases a, b, c
We have d_1 is a metric

ii. d_2

A. refexivity:
$d(a, a) = \frac{d_(a,a)}{1+d_(a,a)} = \frac{0}{1+0} = 0$
proved

B. symmetry:
if $a = b$
$d(a, b) = d(b, a) = 0$ if $a \neq b$
$d(a, b) = d(b, a) \neq 0$
proved

C. transitivity:
The only way I can demenstrate is based on it is metric as d_2, $d_{i'}(a, b)$ keeps the basic as d_2. So we can say it is refexivity proved.

iii. d_3
It's not a metric because it is not well defined, as x, y have different length we don't know what to do.

iv. d_4
It's not a metric because $d(a, a) \neq 0$

(b) $d_{i'}(x, y) = \alpha d_i(x, y)$ for $\alpha \in \mathbb{R}_{>0}$

i. d_1

A. refexivity:
$\alpha d(a, a) = \alpha * 0$
proved

B. symmetry:
if $a = b$
$\alpha * d(a, b) = \alpha * d(b, a) = 0$
if $a \neq b$
$\alpha * d(a, b) = \alpha * d(b, a) \neq 0$, we know they are equal but value is variety based on different data sets.
proved

C. transitivity:
if $a = b = c$
$\alpha * d(a, b) = \alpha d(b, a) = \alpha d(a, c) = 0$
if $a \neq b = c$
$\alpha * d(a, b) + \alpha * d(b, c) = \alpha * 1 + 0 = \alpha > d(a, c) = 0$
if $a \neq b \neq c$
$\alpha * d(a, b) + \alpha * d(b, c) = \alpha * 1 + \alpha * 1 = \alpha * 2 > d(a, c) = \alpha$ Since $\alpha > 0$

if $a = b \neq c$
$\alpha * d(a,b) + \alpha * d(b,c) = 0 + \alpha * 1 = \alpha = d(a,c) = \alpha$
proved So all cases a, b, c
We have d_1 is a metric

ii. d_2

A. refexivity:
Since d_2 satisfies this qulity, multiple with$\alpha > 0$ will keep this qulity. proved

B. symmetry:
if $a = b$
$\alpha * d(a,b) = \alpha * d(b,a) = 0$ if $a \neq b$
$d(a,b) = d(b,a) \neq 0$
proved

C. transivity:
The only way I can demenstrate is based on it is metric as d_2, $d_{i'}(a,b)$ keeps the basic as d_2. So we can say it is refexivity proved.

iii. d_3
It's not a metric because it is not well defined, as x, y have different length we don't know what to do.

iv. d_4
It's not a metric because $\alpha * d(a,a) \neq 0$

(c) $d_5(x,y) = d_1(x,y) + 3d_2(x,y)$
It is metric

i. refexivity:
$d_1(a,a) + 3d_2(a,a) = 0 + 0 = 0$
proved

ii. symmetry:
if $a = b$
$d_1(a,b) + 3 * d_2(a,b) = 0 + 0 = 0$
if $a \neq b$
$d_1(a,b) + 3 * d_2(a,b) = 1 + 3 * m$ we dont know this m, but $d_2(a,b) = d_2(b,c)$ based on aboved
$d_1(a,b) + 3 * d_2(a,b) = 1 + 3 * m$ proved

iii. transitivity:
Based on we already know that $\alpha * d_2(a,b)$ is valide, $d_1(a,b) + 3 * d_3(a,b)$ doesn't change the quality of this two metrics, we can say it still transitivity.

(d) $d_6(x,y) = d_2(y,x)$
It is metric since change the positions of x, y wont change the property of d_2

(e) $d_7(x,y) = d_3(x,y)d_2(x,y)$
It is not metric since d_3 is not well difined as we mentioned above

(f) $d_8(x,y) = \sum_{i=1}^{4} d(x,y)$
it is Not metric since d_3 and d_4 are not metric, sum up them wont change their property. Specially d_3 is not well defined.

3. Read the paper, "A Survey on Tree Edit Distance and Related Problems," by Bille[**?**]. In no more than two paragraphs, discuss what is *most* relevant to either datamining or data science.
Answer:
In either datamining or data science we have the following relevant important directions based on the paper:

The problems we are talking about here are tree edit distance, alignment distance, and inclusion problems. So when we talk about problems we are talking about these.

(a) There are some orderred versinos of the problems above are NP-complete. Using defferent types of mappings would give these NP-complete problem a lot improvement of understandng.

(b) We now have the lower bound and woesr upper bound for the ordered tree edit distance, But there is a big gap between these two which means we have a lot work to do to improve this.

(c) We should consider more than edit operations other than the operations mentioned above.

Application of k-means and Data Prepartion to Medical Data

This problem examines Wolberg's breast cancer data[**?**] that we will denote by Δ. This set, though tiny, provides a good start for k-means and preprocessing. Δ is found at
`http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/`

data	`breast-cancer-wisconsin.data`
description	`breast-cancer-wisconsin.names`

While you will read the data description to more fully understand the format, we create some attribute names to make discussion easier.

ID	Description	Domain	Attribute Name
1.	Sample code number	string	SCN
2.	Clump Thickness	$\mathbb{N}$	A_2
3.	Uniformity of Cell Size	$\mathbb{N}$	A_3
4.	Uniformity of Cell Shape	$\mathbb{N}$	A_4
5.	Marginal Adhesion	$\mathbb{N}$	A_5
6.	Single Epithelial Cell Size	$\mathbb{N}$	A_6
7.	Bare Nuclei	$\mathbb{N}$	A_7
8.	Bland Chromatin	$\mathbb{N}$	A_8
9.	Normal Nucleoli	$\mathbb{N}$	A_9
10.	Mitoses	$\mathbb{N}$	A_{10}
11.	Class:	char	C

1. **Datamining Problem** Suppose you're working to help a clinic serve a community that has limited resources to identify and treat breast cancer. The cost of a biopsy is from \$1000 to \$5000, since it requires a pathologist. The cost of a masectomy is \$15,000 to \$55,000 (these are representative costs in 2016). The cost of a computer program, ignoring the modest fixed cost of machine *etc.*, is \$10.

   (a) What is the total cost of the biopsies in Δ when done by a pathologist? Assume the computer can identify 90% of the cases to nearly 100% accuracy. What is the cost of the computer program?
   Answer:
   The reason I can only give a range for this question is because we dont know what excatly cost for every one. So a range is a reasonable answer. I know someone may use mean as the number. But for human elements we should be careful about this.

      i. We have 669 instances in this data set so the cost is between $1000 * 669$ to $5000 * 669$ is 669000\$ to 3345000\$

      ii. As we know now cumpyter give 90% right answer. Which means everytime computer runs the program, only 10% is wrong. For a single point it has 10% get the wrong answer. So if we run the program three times. For each single point which doesn't change the reasult for this three times. It would be only $10\% * 10\% * 10\%$ is 0.1% to get the wrong answer. Which give us the confidence to say it's right.
      So the cost for this strategy is $10\$ * 669 * 3 = 20070\$$

(b) What would have been the likely total cost of masectomies?
Answer:
Cost $= 15000 * number\ of\ Malignant$ to $55000 * number\ of\ Malignant = 3615000$ to 13255000

(c) Assuming a 70% mortality rate for untreated in year five, how many deaths does the data suggest in five years?
Answer:
We have the 10% wrong number is we dont exam the data serveral times.
result $= 10\% * number\ of\ Benign * 70\% = 0.1 * 458 * 0.7 = 32$

(d) Compose a succint problem statement that you imagine is pertinent to this scenario.
Answer:
From my perspect, the result of computer can not be trusted 100%. So the best way to solve this problem is to combine computer and human pathologist. Let the computer learn to decide the result it can be positive sure but pass the unsure data to the human pathologist.
Also It is not just bi or ma. Some people condition is between these two.

2. **Data Preparation** Ignoring the `Sample code number` (SCN),

(a) Ignoring the SCN and C columns, how many attributes (or features) does Δ have?
Answer:
9 features

(b) Let $\Delta^{miss} \subset \Delta$ be the data that has missing values. How many missing values exist (total)? What is the size of Δ^{miss}?
Answer:

i. only one missing value exist. it's the valule of Bare Nuclei

ii. 16 points have the missing value.
$\Delta^{miss} = 16$

(c) How many patients have missing values?
Answer:
16

(d) Give the SCNs for that have missing values.
1057013 1096800 1183246 1184840 1193683 1197510 1241232 169356 432809 563649 606140 61634 704168 733639 1238464 1057067

(e) Of these data, would you have recommended re-examination for the women? What would be the costs both for the pathologist and computer program?

i. I would like to recommended re-exnimation these data. For the reason that based on later data Correlation Coefficient Uniformity of Cell Shape is the one shoule be deleted, which means this data is relavantly important.

ii. The cost for pathologist:
$1000 * 16$ to $5000 * 16$ is 16000$ to 80000$
The cost for computer:
$10 * 16 = 160$

(f) Is the amount of missing data significant from an algorithmic perspective?
Answer:
Based on the statement above I would say it's not significant from an algorithmic perspective. For the reason that the result not defined just by one value, but by all the 9 values. Missing one of them won't change the result significanly based on what I tried later in the K-means algorithm.

(g) Assess the significance of either keeping or removing the tuples with unknown data. You should consider the human element too.
Answer:

i. At first it seems like it's not siginificance if we just remove the tuples with unknow data. For the reason that in almost 700 data lose 16 is not matter.

ii. In this assignment, I did remove the tuples with unknow data in the K-means implement at first and got a very good result. 95% correct. The reason I did this is just because it would gave us a better and more clear answer if we just removed the "imperfect" data.

iii. But we should consider this in different situations. For example 16 may seem like a small number when it compare to 700 but if these 16 people have different situation than others, Then they become significant for the whole data.

iv. In the mean while, They are real people. We should not just removing they like that. At least we could fix the data or found out the reason they have unknow data.

(h) Repair Δ^{miss} by replacing unknown data using one of the techniques we discussed in class. This will be presented as $(\texttt{SCN}, A_i, v)$ where `SCN` is the tuple key, A_i is the attribute, and v is the new value. Create a `CSV` file `DeltaFix.csv` for this data. Call the entire data set, including the values that have been replaced, as Δ_1^{clean}.
Answer:
Using the mean of the feater in the all data that with no-missing value to implement the missing value. The fixed data is as below:

SCN	A_i	v
1057013	A_7	4
1096800	A_7	4
1183246	A_7	4
1184840	A_7	4
1193683	A_7	4
1197510	A_7	4
1241232	A_7	4
169356	A_7	4
432809	A_7	4
563649	A_7	4
606140	A_7	4
61634	A_7	4
704168	A_7	4
733639	A_7	4
1238464	A_7	4
1057067	A_7	4

3. **Data Analysis**

(a) Using either `MySQL`, `SQL Server` or `PostgreSQL`, built a table and load the fixed data set. Connect to `R` so that you can quickly and easily perform analysis. Using `R`,

(b) Plot histograms for each attribute and C.

(c) Find the mean, median, mode, and variance of each attribute.
Answer:
The code are writen in R file saved in matrix C

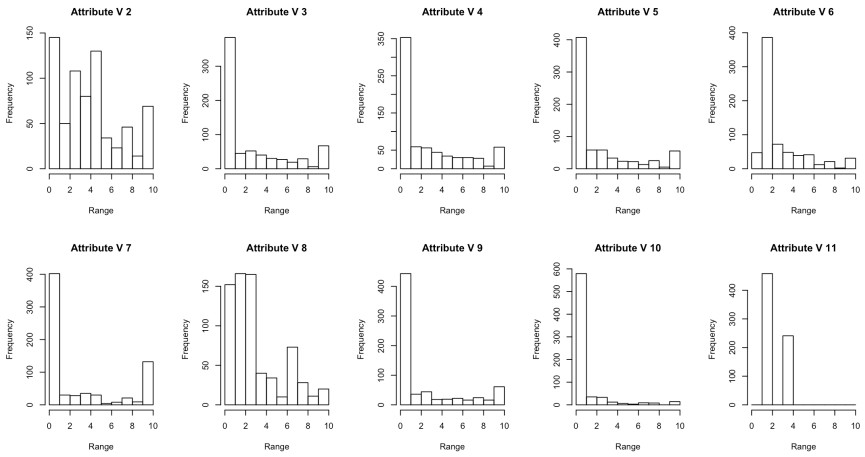

Figure 1:

A_i	mean	median	mode	variance
A_i	4.417	4	1	7.92
A_i	3.13	1	1	9.31
A_i	3.20	1	1	8.83
A_i	2.80	1	1	8.15
A_i	3.21	2	2	4.90
A_i	3.55	1	1	12.97
A_i	3.43	3	2	5.94
A_i	2.86	1	1	9.32
A_i	1.58	1	1	2.94

(d) For each pair A_i, A_j, $i \neq j$, find the Pearson's correlation coefficient. This provides an insight to the linearity of the attributes. To remind you,

$$\rho_{X,Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

σ is the standard deviation

μ is the mean

Eis the expectation

How is ρ related to $cos\theta = \frac{\mathbf{xy}}{||\mathbf{x}||||\mathbf{y}||}$? Remove one of the pairs of attributes that are strongly linearly related for every pair of attributes. Call this Δ_2^{clean}. What is the purpose of this step?
Answer:

i. Correlation is the cosine similarity between For each pair A_i, A_j, $i \neq j$. Which means they are basicly represent the same things. Although they may have the different value but the core reason inside these two are pretty much the same.

ii. The picture generated for R file to see the Pearson's Correlation Coefficient for each pair of Value. As we can see A_3 and A_4 have the biggest value. So we can remove either one of them.
Also about this one problem I have a question is whether we should remove the pair of attributes or just one value of this pair. Based on the class AI gave to us keep one should remind the most of the information but it says remove one of the pairs from the attributes. Which makes not that sense to me.

	A2	A3	A4	A5	A6	A7	A8	A9	A10
A2	1.0000000	0.5365269	0.5341678	0.4334475	0.4700347	0.4629008	0.4178644	0.4623299	0.3531449
A3	0.5365269	1.0000000	0.8148055	0.6396170	0.6819673	0.6415777	0.5994902	0.6634772	0.4553330
A4	0.5341678	0.8148055	1.0000000	0.6051535	0.6501500	0.6323085	0.5715317	0.6281452	0.4186323
A5	0.4334475	0.6396170	0.6051535	1.0000000	0.5708719	0.5887504	0.5114875	0.5470115	0.4002850
A6	0.4700347	0.6819673	0.6501500	0.5708719	1.0000000	0.5748881	0.5220132	0.6086361	0.4312051
A7	0.4629008	0.6415777	0.6323085	0.5887504	0.5748881	1.0000000	0.5557728	0.5566875	0.4134997
A8	0.4178644	0.5994902	0.5715317	0.5114875	0.5220132	0.5557728	1.0000000	0.5614033	0.3332470
A9	0.4623299	0.6634772	0.6281452	0.5470115	0.6086361	0.5566875	0.5614033	1.0000000	0.4536334
A10	0.3531449	0.4553330	0.4186323	0.4002850	0.4312051	0.4134997	0.3332470	0.4536334	1.0000000

Figure 2:

4. Implement k-means so that you can cluster Δ_2^{clean} without using C. Upon stopping, you will calculate the quality of the centroids and of the partition. For each centroid c_i, form two counts:

$$b_i \leftarrow \sum_{\delta \in c_i.B} [\delta.C = 2], \quad \text{benign}$$

$$m_i \leftarrow \sum_{\delta \in c_i.B} [\delta.C = 4], \quad \text{malignant}$$

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 benign if $b_i > m_i$ and malignant otherwise. We can now calculate a simple error rate. Assume c_i is benign. Then the error is:

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

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. Discuss your results and include your initial problem statement.
Answer:

(a) When k = 2, I initialzed the centroid by choosen the two datapoint in the dataset which has the biggest distance in the all set. Then do the k-means. This part of code put in the txt file whin the java.zip file. For the reason that late k = 3, 4, 5 I implemented a method to initialzed the controids randomly from the data set.
Use this method I got a very good result as $error = 0.032$
However when I initialzed the centroids by randomly choose from the dataset, the *error* is always around 0.1.
Which combined with the statement before. Gave us a conclusion. When we knew the cluster number before implemented the data. And we can make more customerlized centroid which might gave us a better result then we just randomly choose the centroid.

(b) when $k = 3$
The answer for 20 runs is:
0.60 0.61 0.57 0.37 0.23 0.17 0.17 0.17 0.16 0.17 0.17 0.17 0.17 0.17 0.17 0.17 0.17 0.17 0.17 0.17

(c) when $k = 4$
The answer for 20 runs is:
0.49 0.27 0.19 0.15 0.16 0.15 0.15 0.15 0.15 0.15 0.15 0.15 0.15 0.15 0.15 0.15 0.15 0.15 0.15 0.15

(d) when $k = 5$
The answer for 20 runs is:
0.65 0.33 0.22 0.19 0.19 0.19 0.19 0.19 0.19 0.19 0.19 0.19 0.19 0.19 0.19 0.19 0.19 0.19 0.19 0.19

(e) Discussion:

i. When $k > 2$, the centroids are initialzed randomly from the dataset. And as I run more times the program. It seems like as then k getting bigger when result gets bigger too. which means the error gets bigger. But when we run many times. The result get down as the program get going.

ii. As I implement many times for different k, I found out that when k is big as 4,5. It more likely gave us a big error number for example 0.86 as so on. Which is unreasonable. So it probably would make the result unstable when k is bigger than 3.

iii. As the k gets bigger, as 6, 7, 8. The error getting bigger. And there is a funny thing comes out when k keeping getting bigger as more than 10. Is would make some centroid assign to no datapoints. which means we may just pick useless datapoint for cluster.

iv. As the statement mentioned above. When k getting bigger. It would be more likely that Centroid Collapse happes.

What to Turn-in

- The *pdf of the written answers to this document.
- The code for k-means, `R`.
- The AIs can schedule a time to verify your codes works. If there is a subsequent time-stamp to the due date of the source code, the grade may be reduced.