

Homework 2

Computer Science

Fall 2016

B565

Anusha Ramamurthy
anuramam@iu.edu

September 23, 2016

All the work herein is solely mine.

Directions

Please follow the syllabus guidelines in turning in your homework. I will provide the L^AT_EX of this document too. You may use it or create one of your own. This homework should be started quickly. Sometimes there are natural questions arising from code. 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 of the problem asks you to reflect on k -means and work through its theoretical elements. I have written algorithm below. Answer the subsequent questions.

```
1: ALGORITHM  $k$ -means
2: INPUT (data  $\Delta$ , distance  $d : \Delta^2 \rightarrow \mathbb{R}_{\geq 0}$ , centroid number  $k$ , threshold  $\tau$ )
3: OUTPUT (Set of centroids  $\{c_1, c_2, \dots, c_k\}$ )
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:    $c_j^i.v \leftarrow random(Dom(\Delta))$ 
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:      $c_j^i.B \leftarrow c.B \cup \{\delta\}$ , where  $\min_{c_j^i} \{d(\delta, c_j^i.v)\}$ 
```

```

23:   end for
24:   for  $j = 1, k$  do
25:       *** 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  $(\{c_1^i, c_2^i, \dots, c_k^i\})$ 

```

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

```

1:  $i \leftarrow 0$ 
2: *** Randomly assign value to first centroid.
3:  $c_1^0.v \leftarrow \text{random}(\text{Dom}(\Delta)) = (16, 19)$ 
4: *** Randomly assign value to second centroid.
5:  $c_2^0.v \leftarrow \text{random}(\text{Dom}(\Delta)) = (2, 5)$ 
6:  $i \leftarrow i + 1$ 
7: *** Associate each datum with nearest centroid
8:  $c_1^1.B = \{(22, 55), (42, 12), (15, 16)\}$ 
9:  $c_2^1.B = \{(2, 5), (1, 5)\}$ 
10: *** Update centroids
11:  $c_1^1.v \leftarrow (26.3, 27.7) = (1/3)((22, 55) + (42, 12) + (15, 16))$ 
12:  $c_2^1.v \leftarrow (1.5, 5) = (1/2)((2, 5) + (1, 5))$ 
13: *** The convergence condition is split over the next few lines to explicitly show the calculations
14:  $(1/k) \sum_{j=1}^k \|c_j^{i-1} - c_j^i\| = (1/2)(\|c_1^0 - c_1^1\| + \|c_2^0 - c_2^1\|) = (1/2)(\| \binom{2}{5} - \binom{1.5}{5} \| + \| \binom{16}{19} - \binom{26.3}{27.7} \|)$ 
15:  $= (1/2)[(\binom{5}{0}^\top \binom{5}{0})^{(1/2)} + ((\binom{-9.7}{-8.7})^\top \binom{-9.7}{-8.7})^{(1/2)}] = (1/2)(\sqrt{.5} + \sqrt{169.7}) \sim (1/2)(13.7) = 6.9$ 
16: Since the threshold is met ( $6.9 < 10$ ),  $k$ -means stops, returning  $\{(26.3, 27.7), (1.5, 5)\}$ 

```

Questions

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

Answer: k -means does not always converge. For some combinations of distance metrics and the centroids chosen initially, k -means does converge. However, since different methods of choosing the initial centroids exist, we can not always say it will converge. Especially if the initial centroids are chosen at random. Sometimes the condition $((1/k) \sum_{j=1}^k \|c_j^{i-1} - c_j^i\|) < \tau$ is never satisfied and k -means does not converge. We can create a bound on the iterate to repeat until just 1 percent of the points change clusters or to run it about a 100 to 1000 times depending on the dataset and the computational capabilities available

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: Random selection of centroids may lead to k -means not converging. It also leads to different outcomes at each run. It may so happen that we have outliers in our datasets which may be selected as our initial centroids, which increases the time required for k -means to converge.

3. What is the run-time of this algorithm (include your new parameter from Question 1).

Answer: If we use the parameter I as the iterator, the runtime is $O(I * K * M * N)$ where I is the number of Iterations, K is the number of centroids, M is the size of the dataset and N are the number of attributes of the dataset.

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: Another way is to assign it to the cluster that has lesser number of points. This can overcome empty clusters. If a significant numbers of points appear to be equidistant from a set of clusters, we can also consider the method of merging clusters.

- *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: Another method is to merge these clusters to be one large cluster. Alternatively we can update the centroid values to be the average value of the blocks in the clusters and recalculate the distance from the new centroids. τ_c can be determined as the Euclidean distance between the centroids. This should be checked at each iteration just before our stopping condition

- Modify the k -means algorithm to address ties and collapsing centroids. Explicitly add pseudo-code to the algorithm and call this k -meansr. **Answer:**

```

1: ALGORITHM  $k$ -meansr
2: INPUT (data  $\Delta$ , distance  $d : \Delta^2 \rightarrow \mathbb{R}_{\geq 0}$ , centroid number  $k$ , threshold  $\tau$ )
3: OUTPUT (Set of centroids  $\{c_1, c_2, \dots, c_k\}$ )
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:    $c_j^i.v \leftarrow random(Dom(\Delta))$ 
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:    $c_j^i.B \leftarrow c.B \cup \{\delta\}$ , where  $\min_{c_j^i} \{d(\delta, c_j^i.v)\}$ 
23:   If(Ties)
24:   *** Assign values to the centroid with lesser number of points
25: end for
26: for  $j = 1, k$  do
27:   *** Get size of centroid
28:    $n \leftarrow |c_j^i.B|$ 
29:   *** Update centroid with average
30:    $c_j^i.v \leftarrow (1/n) \sum_{\delta \in c_j^i.B} \delta$ 
31:   *** Remove data from centroid
32:    $c_j^i.B \leftarrow \emptyset$ 
33: end for
34: *** Calculate scalar product (abuse notation and structure slightly)
35: *** See notes
36: If  $\text{True}(d(c_i, c_j) \leq \tau_c)$ 
37:   *** Reassign values between clusters
38:   Merge Clusters  $c_i, c_j$ 
39: until  $((1/k) \sum_{j=1}^k \|c_j^{i-1} - c_j^i\|) < \tau$ 
40: return  $\{c_1^i, c_2^i, \dots, c_k^i\}$ 

```

***k*-means on a tiny data set.**

Here are the inputs:

$$\Delta = \{(2, 5), (1, 5), (22, 55), (42, 12), (15, 16)\} \quad (5)$$

$$d((x_1, y_1), (x_2, y_2)) = [(x_1 - x_2)^2 + (y_1 - y_2)^2]^{1/2} \quad (6)$$

$$k = 2 \quad (7)$$

$$\tau = 10 \quad (8)$$

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, $Z = [N : \text{int}] \times [S : \text{string}]$. $(a, b) = ((34, \text{two}), (100, \text{three}))$, then $a_0 = 34, b_0 = 100$ and $a_1 = \text{two}, b_1 = \text{three}$.

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

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

Now we show reflexivity, symmetry, and transitivity.

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

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

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

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

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

X	Y
14	\mathbf{g}
45	\mathbf{o}
54	\mathbf{g}
21	\mathbf{n}
60	\mathbf{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:

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

$$d(a, b) = \left| \frac{d_x(a_0, b_0)}{1 + |d_x(a_0, b_0)|} - \frac{d_y(a_1, b_1)}{1 + |d_y(a_1, b_1)|} \right|$$

Now we show reflexivity, symmetry, and transitivity.

- $(\forall a \in Z) d(a, a) = 0$. Then

$$d(a, a) = \left| \frac{d_x(a_0, a_0)}{1 + |d_x(a_0, a_0)|} - \frac{d_y(a_1, a_1)}{1 + |d_y(a_1, a_1)|} \right| = 0$$

- $(\forall a, b) d(a, b) \rightarrow d(b, a)$.

$$d(a, b) = \left| \frac{d_x(a_0, b_0)}{1 + |d_x(a_0, b_0)|} - \frac{d_y(a_1, b_1)}{1 + |d_y(a_1, b_1)|} \right| = \left| \frac{d_x(b_0, a_0)}{1 + |d_x(b_0, a_0)|} - \frac{d_y(b_1, a_1)}{1 + |d_y(b_1, a_1)|} \right| = d(b, a)$$

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

$$\begin{aligned} d(a, b) + d(b, c) &= \left| \frac{d_x(a_0, c_0)}{1 + |d_x(a_0, c_0)|} - \frac{d_y(a_1, c_1)}{1 + |d_y(a_1, c_1)|} \right| + \left| \frac{d_x(b_0, c_0)}{1 + |d_x(b_0, c_0)|} - \frac{d_y(b_1, c_1)}{1 + |d_y(b_1, c_1)|} \right| \\ &\geq \left| \frac{d_x(a_0, c_0)}{1 + |d_x(a_0, c_0)|} - \frac{d_y(a_1, c_1)}{1 + |d_y(a_1, c_1)|} + \frac{d_x(b_0, c_0)}{1 + |d_x(b_0, c_0)|} - \frac{d_y(b_1, c_1)}{1 + |d_y(b_1, c_1)|} \right| \\ &\geq \left| \frac{d_x(a_0, c_0)}{1 + |d_x(a_0, c_0)|} - \frac{d_y(a_1, c_1)}{1 + |d_y(a_1, c_1)|} \right| \\ &\geq d(a, c) \end{aligned}$$

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:

$$d_1(x, y) = \begin{cases} 0, & x = y \\ 1, & x \neq y \end{cases} \quad \text{For objects } x, y.$$

$$J(x, y) = |x \cap y| / |x \cup y| \quad \text{For sets } x, y.$$

$$d_2(x, y) = 1 - J(x, y) \quad \text{For sets } x, y.$$

$$c(x, y) = \begin{cases} 0, & x = y \\ 1, & \text{otherwise} \end{cases} \quad \text{for individual characters, e.g., } \mathbf{a} = \mathbf{b}$$

$$d_3(\mathbf{x}, \mathbf{y}) = \sum_{i=0}^{n-1} c(\mathbf{x}[i], \mathbf{y}[i]) \quad n = \|\mathbf{x}\|, \text{ the length of the string.}$$

$$d_4(\mathbf{x}, \mathbf{y}) = \left| \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} \right| \quad \text{for vectors } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$$

Calculate the following:

- (a) For every i , find $d_i(x, w)$

Answer :

Assuming x and w are two strings of letters, where $x = \{a, b, c, d\}$ and $w = \{a, d, f, e\}$

From the definition,

$$d_1(x, w) = d_1(\{a, b, c, d\}, \{a, d, f, e\}) = 1$$

For d_2 we assume that x and w are sets, hence

$$d_2(x, w) = 1 - \left(\frac{2}{6}\right) = 1 - \left(\frac{1}{3}\right) = \frac{2}{3}$$

For d_3 we assume that x and w are strings, hence

$$d_3(x, w) = ([a, a]^{i=0} + [ab, ad]^{i=1} + [abc, adf]^{i=2} + [abcd, adfe]^{i=3})$$

$$d_3(x, w) = 0 + 1 + 1 + 1 = 3$$

For d_4 we assume that x and w are vectors and assign values

$$a = 1, b = 2, c = 3, d = 4, e = 5, f = 6$$

$$d_4(x, w) = \frac{\left(\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} * \begin{bmatrix} 1 & 4 & 6 & 5 \end{bmatrix} \right)}{\sqrt{(1^2+2^2+3^3+4^4)*(1^2+4^2+6^2+5^2)}} = 0.97$$

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

Answer :

Assuming x and z are two strings of letters, where $x = \{a, b, c, d\}$ and $z = \{b, f\}$

From the definition,

$$d_1(x, z) = d_1(\{a, b, c, d\}, \{b, f\}) = 1$$

For d_2 we assume that x and w are sets, hence

$$d_2(x, z) = 1 - \left(\frac{1}{6}\right) = 0$$

For d_3 we assume that x and w are strings, hence

$$d_3(x, z) = ([a, b]^{i=0} + [ab, bf]^{i=1} + [abc, bff]^{i=2} + [abcd, bff]^{i=3})$$

$$d_3(x, z) = 1 + 1 + 1 + 1 = 4$$

For d_4 we assume that x and w are vectors and assign values

$$a = 1, b = 2, c = 3, d = 4, e = 5, f = 6$$

$$d_4(x, z) = \frac{\left(\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} * \begin{bmatrix} 2 & 6 & 0 & 0 \end{bmatrix} \right)}{\sqrt{(1^2+2^2+3^3+4^4)*(2^2+6^2+0^2+0^2)}} = 1.212$$

- (c) Which distance gives the the maximum value for any pairs?

Answer: d_3 gives the maximum distance for any pairs as per the assumptions used by me.

- (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. If we put in values for $d_4(v, v)$ we get values other than rest of the metrics.

Assuming $v = \{a, b, c, d\}$

From the definition,

$$d_1(v, v) = d_1(\{a, b, c, d\}, \{a, b, c, d\}) = 0$$

For d_2 v is a set, hence

$$d_2(v, v) = 1 - \left(\frac{4}{4}\right) = 0$$

For d_3 we assume that v is a string, hence

$$d_3(v, v) = ([a, a]^{i=0} + [ab, ab]^{i=1} + [abc, abc]^{i=2} + [abcd, abcd]^{i=3})$$

$$d_3(v, v) = 0 + 0 + 0 + 0 = 0$$

For d_4 we assume v is a vector and assign values

$$a = 1, b = 2, c = 3, d = 4, e = 5, f = 6$$

$$d_4(v, v) = \frac{\left(\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} * [1 \quad 2 \quad 3 \quad 4] \right)}{\sqrt{(1^2+2^2+3^3+4^4)*(1^2+2^2+3^2+4^2)}} = 1$$

However, if we choose to ignore $d_4(v, v)$, the relation $d_1(v, v) = d_2(v, v) = d_3(v, v)$ holds, and is True.

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):

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

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

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

(d) $d_6(x, y) = d_2(y, x)$

(e) $d_7(x, y) = d_3(x, y)d_2(x, y)$

(f) $d_8(x, y) = \sum_{i=1}^4 d(x, y)$

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 data mining or data science.

Answer: The paper discusses the different problems with regard to Trees, structures most commonly used in the field of analysis. They are most widely used to represent information. They are in biology, topology, linguistics and other fields, as they model real world relationships. The paper considers the various operations that are performed on Trees and provides different approaches to perform these operations in an efficient manner. These problems are computationally hard and some fall into the category of NP-hard and SNP-hard problems. However the author prescribes a methods where, by placing certain restrictions on preserving the structure, these problems can be solved in polynomial time.

They are relevant to our field of study, Data Mining and Analysis because we study real world behaviors and relationships. Most relationships can be modeled by trees. It is inevitable that we will have to compare the similarity or dissimilarities of different Trees. How do we go about doing so if they are NP-Hard problems? We do have much faster machines and the cost of storage has greatly reduced, but we still have find methods to perform analysis at a much faster pace. It is almost natural to construct a tree or forest structure to our data for easier analysis, but as surveyed by the author and a lot of other computer engineers, tree operations are CPU intensive. They require restrictions to be applied in order to make them polynomial time problems.

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	N	A_2
3.	Uniformity of Cell Size	N	A_3
4.	Uniformity of Cell Shape	N	A_4
5.	Marginal Adhesion	N	A_5
6.	Single Epithelial Cell Size	N	A_6
7.	Bare Nuclei	N	A_7
8.	Bland Chromatin	N	A_8
9.	Normal Nucleoli	N	A_9
10.	Mitoses	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: There are 8 duplicate rows in the dataset given to us. If we run our algorithm only for the 691 rows of data. The cost of Biopsies is from \$6,91,000 to \$34,55,000 . While the cost of the computer program would only be around \$6,910

- (b) What would have been the likely total cost of mastectomies?

Answer: We have 238 rows which are labeled as 4. These indicate that mastectomies have been performed. The total cost of mastectomies are around from \$35,70,000 to \$130,90,000.

- (c) Assuming a 70% mortality rate for untreated in year five, how many deaths does the data suggest in five years?

Answer: There were 238 mastectomies performed. Supposing these were undetected, we would have about 72 deaths in the next five years if left untreated.

- (d) Compose a succinct problem statement that you imagine is pertinent to this scenario.

Answer: Healthcare is expensive and not everyone can afford it. Our problem statement can be defined as from the historical data, identifying a true estimator of the patient needing to undergo a biopsy.

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: There are 9 attributes after ignoring the SCN and C columns

- (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: There are 16 missing values. $\Delta^{miss} = 16$
- (c) How many patients have missing values?
Answer: There is one missing value in each of the 16 rows. Hence 16 patients have missing values
- (d) Give the SCNs for that have missing values.
Answer: Give the SCNs for that have missing data. SCNs for missing data are 1057013, 1096800, 1183246, 1184840, 1193683, 1197510, 1241232, 169356, 432809, 563649, 606140, 61634, 704168, 733639, 1238464 and 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?
Answer: Since its the question of an undetected tumor and the life of the patient is at stake, I would definitely say a re-examination is required. The cost for the pathologist would be around \$16,000 to \$80,000. It would be around \$160 for the programmer.
- (f) Is the amount of missing data significant from an algorithmic perspective?
Answer: The missing values are significant from an algorithmic perspective as we have insufficient data. We have no way of being very sure that our results are close to any form of accuracy.
- (g) Assess the significance of either keeping or removing the tuples with unknown data. You should consider the human element too.
Answer: We should definitely keep the missing values as removing the tuples will affect our estimates of other attributes as well. From the human element, we should definitely ask for the patients with missing values to retake the test, as there is no way to confirm where the error occurred and what caused the missing values to be a part of the data.
- (h) Repair Δ^{miss} by replacing unknown data using one of the techniques we discussed in class. This will be presented as (SCN, A_i, v) where SCN is the tuple key, A_i is the attribute, and v is the new value.
Answer: I created a table caled BCancerDB on MySQL and inserted the dataset using the Import function available. Missing values were automatically replaced by 0 in MySQL as I specified the attribute values to be of Integer Datatypes
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: There are multiple ways to replace missing values. I have included them in the code file. We can either replace it by the mean value of the Attribute, ie Bare Nuclei. Another method is to replace it with the Mode, ie. is the most frequent value in the column. However both these methods change the distribution and I dont agree with them. I found the correlation of the attribute Bare Nuclei and Uniformity of Cell Shape to be the highest. I hence replaced the missing values of the Bare Nuclei column with the corresponding values od Uniformity of Cell Shape attribute. You can find this in `DeltaFix.csv`

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: All code and results are in the `DataAnalysis.R` file.

- (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,

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

σ is the standard deviation

μ is the mean

E is the expectation

How is ρ related to $\cos\theta = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$? 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: The Pearson correlation is related to cosine similarity. The cosine similarity is the inner product of vectors bounded between 0 to 1. The Pearson correlation is the cosine similarity of centered vectors bounded between 0 and 1.

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:

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

We can find the total error rate easily:

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

Report the total error rates for $k = 2, \dots, 5$ 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.

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.