Homework 2
Computer Science
Fall 2016
B565

Saheli Saha
sasaha@iu.edu

Spetember 23rd 2016

Directions

Please follow the syllabus guidelines in turning in your homework. I will provide the LaTeX 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.

ALGORITHM `k-means`
INPUT (data Δ, distance $d : \Delta^2 \to \mathbb{R}_{\geq 0}$, centoid number k, threshold τ)
OUTPUT (Set of centoids $\{c_1, c_2, \ldots, c_k\}$)

*** $Dom(\Delta)$ denotes domain of data.

*** Assume centroid is structure $c = (v \in DOM(\Delta), B \subseteq \Delta)$
*** $c.v$ is the centroid value and $c.B$ is the set of nearest points.
*** c^i means centroid at i^{th} iteration.

$i = 0$
*** Initialize Centroids
for $j = 1, k$ **do**
    $c_j^i.v \leftarrow random(Dom(\Delta))$
    $c_j^i.B \leftarrow \emptyset$
end for

repeat
    $i \leftarrow i + 1$
    *** Assign data point to *nearest* centroid
    for $\delta \in \Delta$ **do**
        $c_j^i.B \leftarrow c.B \cup \{\delta\}$, where $\min_{c_j^i}\{d(\delta, c_j^i.v)\}$
    end for
    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** $(\{c_1^i, c_2^i, \ldots, c_k^i\})$

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.

1: $i \leftarrow 0$
2: *** Randomly assign value to first centroid.
3: $c_1^0.v \leftarrow random(Dom(\Delta)) = (16,19)$
4: *** Randomly assign value to second centroid.
5: $c_2^0.v \leftarrow random(Dom(\Delta)) = (2,5)$
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: $= (1/2)[(\binom{.5}{0}^{\mathsf{T}}\binom{.5}{0})^{(1/2)} + (\binom{-9.7}{-8.7}^{\mathsf{T}}\binom{-9.7}{-8.7})^{(1/2)})] = (1/2)(\sqrt{.5}+\sqrt{169.7}) \sim (1/2)(13.7) = 6.9$
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: Convergence of K means depends on the data set. In short K means algorithm goes as follows:
   Step 1. For any given data set choose K i.e number of clusters.
   Step 2. Choose K number of centroids randomly depending on the data-set.
   Step 3. Assign each of the remaining elements to the cluster to it's nearest centroid.
   Step 4. After each assignment calculate cluster centers by finding mean of the data points belonging to the same cluster.
   If the difference between consecutive centroids are in descending mode then there is a high chance of convergence. For any local data which can be manipulated easily the data will converge.

While considering real life data which is often hi-dimensional, we cannot conclude that it will defiantly converge, in fact as we are choosing centroids randomly there is a high chance that K means will not converge.

Iteration value is solely depends on the nature of the data set. Best solution will be to run the algorithm for few times to understand the rate of convergence. If the data converges easily then we can decide the iteration value.

So for the above mentioned K-Means algorithm can be modified with bound in the following way:-

K means Algorithm with bound

1: **ALGORITHM** `k-means`
2: **INPUT** (data Δ, distance $d : \Delta^2 \to \mathbb{R}_{\geq 0}$, centoid number k, threshold τ, bound for iteration ω)
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: *** I_{count} means iteration counter which will be updated after each clustering while again updating the centroid.
11:
12: $i = 0$
13: *** Initialize iteration Count with 0
14: $I_{count} = 0$
15: *** Initialize Centroids
16: **for** $j = 1, k$ **do**
17: $c_j^i.v \leftarrow random(Dom(\Delta))$
18: $c_j^i.B \leftarrow \emptyset$
19: **end for**
20:
21: **repeat**
22: $i \leftarrow i + 1$
23: *** Assign data point to *nearest* centroid
24: **for** $\delta \in \Delta$ **do**
25: $c_j^i.B \leftarrow c.B \cup \{\delta\}$, where $\min_{c_j^i}\{d(\delta, c_j^i.v)\}$
26: **end for**
27: **for** $j = 1, k$ **do**
28: *** Get size of centroid
29: $n \leftarrow |c_j^i.B|$
30: *** Update centroid with average
31: $c_j^i.v \leftarrow (1/n) \sum_{\delta \in c_j^i.B} \delta$
32: $c_{count} \leftarrow c_{count} + 1$
33: *** Remove data from centroid
34: $c_j^i.B \leftarrow \emptyset$
35: **end for**
36: *** Calculate scalar product (abuse notation and structure slightly)
37:
38: **until** $((1/k) \sum_{j=1}^{k} \ ||c_j^{i-1} - c_j^i||) < \tau \ OR \ I_{count} < \omega$
39: **return** $(\{c_1^i, c_2^i, \ldots, c_k^i\})$

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: There are various reason for the code being problematic:-

- In K Means we are choosing the centroid randomly. So for different person the K-Means algorithm will work differently, as a result the clustering will also vary.
- Randomly selected these centroid may be poor selection. For huge amount of data when K value is fixed choosing centroid is a hectic job and we may choose a poor centroid.One of the common way of choosing initial centroid is to choose centroids first then run the algorithm. Whichever gives minimum SSE(Standard Squared Error) those are considered as initial centroids.But this may not work very well for every problem.
- Randomly initialized centroids can be near to each other.
- Randomly initialized centroids may collapse at some distance.
- K Means has problem when clusters are of different sizes, densities. Specially while dealing with unstructured data.

Implications of k-means algorithm are:

(a) The prior probability for all K cluster is the same.

(b) Depending on the random cluster you select you end up determining the final points of the centroid.

(c) K means assumes the variance of the distribution of each attribute is spherical.

(d) If the cluster points are symmetric in Euclidian distance there is a high chance the centroid values will get stagnant.

(e) K means has problem when the densities are of different sizes,densities and non granular shape.

(f) K means clusters tend to produce clusters in uniform sizes.

3. What is the run-time of this algorithm (include your new parameter from Question 1).
Answer: Run-time of this algorithm depends on the 4 factor:-

- n : number of points
- K : number of clusters
- I_{count} : number of iterations
- d : number of attributes

In Question 1. I have added the iteration count I_{count}. So adding this run-time of this algorithm can be calculated with O(n * K * I_{count} * d).

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: For ties problem there can be two other solution:-
  - ⋆ If the distance of two centroids are same from a single data then I will see the cluster that which has less element I will assign this data to that cluster.
  - ⋆ We can choose the one cluster at a time and can calculate the mean and again we can choose another cluster and add that item to that cluster. After the calculation we can check for which data it is converging more. Then we can add that data to that cluster. In the worst case we do the randomization again to assign new cluster where the any data would not be in the same distance from the centroid point.

- *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: For centroid collapse two other solutions:-
  - ⋆ For the centroid collapse one solution can be the two centroids which are very close to each other, discard those centroids and take the next immediate points for centroids. For this the cluster may vary slightly which should not effect much as K-Means is all about randomization.
  - ⋆ Instead of finding the median we can also calculate the mean then assign the points to the centroids.
  - ⋆ Another solution can be, we can assign a threshold for the distance between the centroids if the distance between the centroids is less than the threshold value then we can perform the randomization again.

  First: The threshold value (τ_c) can be determined by observing the data nature. The average of differences of data points should be calculated first. τ_c value should be determined close to that value. **S**econd: After updating the centroid we can check the centroid difference. This can be done after line 28 in the above mentioned algorithm.
- Modify the k-means algorithm to address ties and collapsing centroids. Explicitly add pseudo-code to the algorithm and call this k-meansr.

Answer: For addressing ties and centroid collapsing I am using one approach for each.

- ⋆ For ties we can see the clusters which ever has less data we can assign that point to that smaller cluster.
- ⋆ For collapsing I am adding one threshold (τ_c) for centroid. Which will check the distance between the centroid if it is less than the centroid then will perform the randomization again to assign the centroid.

k-means Algorithm after adding ties and collapsing solution

1: **ALGORITHM** `k-means`
2: **INPUT** (data Δ, distance $d : \Delta^2 \to \mathbb{R}_{\geq 0}$, 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: **end for**
24: **for** $j = 1, k$ **do**
25: *** Get size of centroid and compare the size then assign the common data to the cluster who has less data points.
26: $n_1 \leftarrow |c_j^{i_1}.B|$
27: $n_2 \leftarrow |c_j^{i_2}.B|$
28: ** Common data δ_c
29: if $n_1 > n_2$ then $|c_j^{i_2}.B| \leftarrow \delta_c$
30: else $|c_j^{i_1}.B| \leftarrow \delta_c$
31: *** Update centroid with average and check with the threshold value τ_c
32: $c_j^{i_1}.v \leftarrow (1/n) \sum_{\delta \in c_j^{i_1}.B} \delta$
33: $c_j^{i_2}.v \leftarrow (1/n) \sum_{\delta \in c_j^{i_2}.B} \delta$
34: *** Multiple centroid updated now checking the difference with the threshold value τ_c
35: If $d(c_j^{i_1}, c_j^{i_2}) < \tau_c$
36: Repeat steps 27, 28 again to update another set of centroids.
37: *** Remove data from centroid
38: $c_j^i.B \leftarrow \emptyset$
39: **end for**
40: *** Calculate scalar product (abuse notation and structure slightly)
41: *** See notes
42: **until** $((1/k) \sum_{j=1}^{k} \ ||c_j^{i-1} - c_j^i||) < \tau$
43: **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 $\texttt{n} < \texttt{o} < \texttt{g} < \texttt{e}$, $\texttt{e}$ 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: 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 n < o < g < e, 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: Based on the question given I am forming a similar metric to prove all the conditions.

X	Y	R
20	g	3
35	n	1
60	e	4
15	g	3
42	o	2

As nothing mentioned about the allocation of Y column to X column so I am assuming it has been randomly assigned.As per the grading mentioned in the question value of R is assigned in that order. i.e,.The ordering is n < o < g < e, e being the best, thus the elements in column R is reflecting that n = 1 o = 2 g = 3 e = 4.The character value for column Y is n = 14 o = 15 g = 7 e = 5 based the Hamming distance, taken the positional value of the data in the alphabet series.
So Z = (a,b) = (20,7),(35,14),(60,5),(15,7),(42,2).
To prove the metric the 3 conditions of distance metric needs to be satisfied.

- **Reflexivity** = $(\forall, a \in Z)\ d(a, a) = 0$.
  Then $d(a, a) = d_x(a_0, a_0) + d_y(a_1, a_1) = d_x(20,20) + d_x(35,35) = 0$. Hence, proving reflexivity.

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

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

$$d(a, b) = d_x(20, 7) + d_y(35, 14) = 34 = d_x(7, 20) + d_y(14, 35)$$

  Calculating using the one Dimensional Euclidean distance.
  Reference: *https* : *//en.wikipedia.org/wiki/Euclidean_distance*
  Hence, symmetry proved.

- **Transitivity** = $(\forall\, a, b, c)\ d(a, b) + d(b, c) \geq d(a, c)$
  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)
  d(a,b) + d(b,c) = $d_x(20,7) + d_x(7,3) + d_y(35,14) + d_y(14,1) = 51 = d_x(20,3) + d_y(35,1)$
  Hence, Transitivity proved.

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:

   The distances calculated above prove that all the four distance measurements are not equal.

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

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

$$c(x, y) = \begin{cases} 0, & x = y \\ 1, & otherwise \end{cases} \quad \text{for individual characters, } e.g., \texttt{a = b}$$
$$d_3(\mathbf{x}, \mathbf{y}) = \Sigma_{i=0}^{n-1} c(\mathbf{x}[i], \mathbf{y}[i]) \quad n = ||\mathbf{x}||, \text{the length of the string.}$$

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

`Calculate the following:`

Answer: Before calculating the value for d_1, d_2, d_3 I am assigning numeric value for d_4. I have considered hamming distance for assigning the numeric values. It will take the positional value of the alphabet from the english alphabet series.
eg: b=2,g=7,p=16. Calculate the following:

1. For every i, find $d_i(x, w)$

   (a) For i = 1 . $d_1 = 1$ .Since $x \neq w$.

   (b) For i = 2. d_2 = 1 - 1-J(x,w).
   J(x,w) = $|x \cap w|/|x \cup w|$ = [a,d]/[a,b,c,d,e,f] = [2]/[6]
   $d_2 = 2/3 = 0.66$

   (c) For i= 3. d_3 = c(x,w)

   - c_1(a,a) = o . as it equal for both x and w.
   - c_2(b,d) = 1. as $x \neq w$.
   - c_3(c,f) = 1.
   - c_4(d,e) = 1.

   c(x,w) = 0+1+1+1 = 3

   (d) For i = 4.
   $d_4(x, w) = \left| \frac{\mathbf{x}^\mathsf{T}\mathbf{w}}{||\mathbf{x}||\,||\mathbf{w}||} \right|$ for vectors $\mathbf{x}, \mathbf{w} \in \mathbb{R}^n$
   $\mathbf{x}^\mathsf{T}$w = (1 2 3 4)^T (1 4 6 5)
   =1/16 *[1(1)+2(4)+3(6)+4(5)] = 44/16 = 2.98

2. Find the d_i that has the minimum value for x, z.
   Minimum value for x, z.

   - $d_1 = 0$. As $x \neq z$
   - d_2 = 1 - J(x,z)
     J(x,z) = $|x \cap z|/|x \cup z|$ = [b]/[a,b,c,d,f] = [1]/[5]
     $d_3 = 4/5 = 0.8$

- d_3 = c(x,z)
  c_1(a,b) = 1 + c_2(b,f) = 1
  c(x,z) = 2.
- $d_4(x,z) = \left|\frac{\mathbf{x}^\mathsf{T}\mathbf{z}}{||\mathbf{x}||\,||\mathbf{z}||}\right|$ for vectors $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$
  $\mathbf{x}^\mathsf{T}$z = (1 2 3 4)^T (2 6)
  =1/8 *[1(2)+2(6)] = 14/8 = 1.75

Hence, minimum value of x, z is for d_1

3. Which distance gives the the maximum value for any pairs?
   For checking this a pair of dissimilar length of elements such as y, z has been taken.

   - $d_1 = 0$. As $y \neq z$
   - d_2 = 1 - J(y,z)
     J(y,z) = $|y \cap z|/|y \cup z|$ = [b]/[a,b,c,f] = [1]/[4]
     $d_2 = 3/4 = 0.75$
   - d_3 = c(y,z)
     c_1(a,b) = 1 + c_2(b,f) = 1
     d_3 = 1+1 = 2
   - $d_4(y,z) = \left|\frac{\mathbf{y}^\mathsf{T}\mathbf{z}}{||\mathbf{y}||\,||\mathbf{z}||}\right|$ for vectors $\mathbf{y}, \mathbf{z} \in \mathbb{R}^n$
     $\mathbf{y}^\mathsf{T}$z = (1 2 3)^T (2 6)
     =1/6 *[1(2)+2(6)] = 14/6 = 2.33

   Hence, after calculating the equations it can be concluded that d_3 has the maximum value.

4. True or False. For any set v, $d_1(v,v) = d_2(v,v) = d_3(v,v) = d_4(v,v)$.
   False. Let, v= z

   - For d_1(z,z) =1 as z = z
   - For d_2(z,z) = 1 - J(z,z)
     J(z,z) = $|z \cap z|/|z \cup z|$ = [b,f]/[b,f] = [2]/[2]
     $d_2 = 0$
   - d_3 = c(z,z) = 0
   - $d_4(y,z) = \left|\frac{\mathbf{z}^\mathsf{T}\mathbf{z}}{||\mathbf{z}||\,||\mathbf{z}||}\right|$ for vectors $\mathbf{z}, \mathbf{z} \in \mathbb{R}^n$
     $\mathbf{z}^\mathsf{T}$z = (2 6)^T (2 6)
     =1/4 *[2(2)+6(6)] = 40/4 = 10

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

   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:

   Let $x = \{1,2,3\}, y = \{2,4,6\}$ and $z = \{2,5,7\}$

(a) $d_{i'}(x,y) = \frac{d_i(x,y)}{1+d_i(x,y)}$ for every i.
Proving reflexivity for i = 1,2,3,4.
For i = 1 . $d_1(x,y) = \frac{d_1(x,x)}{1+d_1(x,x)} = 0$.
For i = 2 . $d_2(x,x) = \frac{d_2(x,x)}{1+d_2(x,x)} = 0$.
For i = 3 . $d_3(x,x) = \frac{d_3(x,x)}{1+d_3(x,x)} = 0$.
For i = 4 . $d_4(x,x) = \frac{d_4(x,x)}{1+d_4(x,x)} = 14/9$.
To be in reflexive relation d(x,x) should be equal to 0 as d_4 does not meet the condition $d_{i'}(x,y) = \frac{d_i(x,y)}{1+d_i(x,y)}$ is not proving for the distance metric.

(b) $d_{i'}(x,y) = \alpha d_i(x,y)$ for $\alpha \in \mathbb{R}_{>0}$
If α is a real number as per the relation defined d_1, d_2 and d_3 will be 0 hence satisfying the reflexive property. Although, d_4 = 14/9 * α.
Hence disproved.

(c) $d_5(x,y) = d_1(x,y) + 3d_2(x,y)$
d_5(x,y) = 1 + 3 * 0.8 = 3.4.
Reflexivity : $d_5(x,y) = 0+0$.
Symmetric :$d_5(x,y) = 3.4 = d_5(y,x)$
Transitive : $d_5(x,y) + d_5(y,z) \geq d_5(x,z) \rightarrow 3.4 + 3.4 \geq 3.4$.
This function satisfying all the three conditions so this is an distance metric.

(d) $d_6(x,y) = d_2(y,x)$
$d_6(x,y) = 0.8$
Reflexivity : $d_6(x,x) = 0$.
Symmetric : $d_6(x,y) = d_6(y,x) \rightarrow d_2(y,x) = d_2(x,y) = 0.8$
Transitive : $d_6(x,y) + d_6(y,z) \geq d_6(x,z) \rightarrow 0.8 + 0.8 \geq 0.8$

(e) $d_7(x,y) = d_3(x,y)d_2(x,y)$
$d_7(x,y) = 3 * 0.8 = 2.4$
Reflexivity : $d_7(x,x) = d_3(x,x)d_2(x,x) = 0$
Symmetry : $d_7(x,y) = d_7(y,x) = 2.4$
Transitivity : $d_7(x,y) + d_7(y,z) \geq d_7(x,z) \rightarrow 2.4 + 0.8*2 \geq 0.8 * 3$
Hence, d_7 is a distance metric.

(f) $d_8(x,y) = \sum_{i=1}^{4} d(x,y)$
$\sum_{i=1}^{4} d(x,y) \rightarrow d_1(x,y) + d_2(x,y) + d_3(x,y) + d_4(x,y)$
= 1+0.8+3+3.11 = 7.91
Reflexivity : $d_8(x,x) = 0 + 0 + 0 + 14/9$.
As this function does not satisfy the reflexivity condition. So, it cannot be a distance metric.

6. 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: The article mainly focuses on the trees. Tree faces problem of comparing labeled trees based on simple local operations of deleting, inserting, and relabeling nodes.The paper focuses on using tree function to solve computational problem. The paper mainly explained the issues with labeled trees. Labeled tree with siblings from left to right order is called an ordered tree. The paper first focuses on the tree edit distance operation: The tree edit distance problem is to compute the edit distance and a corresponding edit script. The edit script between two trees is the cost of transforming on tree into another tree after all the changes. Many algorithms are discussed for the operation, particular one algorithm is choosen as the best suited to create dynamic programming. The algorithm presented by Zhang and Shasha which takes a different approach to resolve the algorithm by taking key roots of the present roots of the tree. This key values mainly helpful for reducing space and time complexity.But for worst case Zhang and Shasha's algorithm will be time consuming just like the simple algorithm. So to address the problem Kleins algorithm has been discussed. It obtained a better time bound for worst case.

Finally, for trees with linear depth and linear number of leaves we use Kleins algorithm. The main idea of Kleins algorithm is to decompose the tree into disjoint paths which are known as heavy path. Some restriction sets must be included to make the operation solvabe. The disjoint sub trees need to be mapped. Next discussed Tree Alignment Distance: Tree edit distance is the main part of this algorithm. Basically it is an extension of the same. It mainly focuses on creating an optimal alignment between two trees. To deduce the result it uses time and space metric and it is efficient mainly for small degree trees. Lastly, there is a study on tree inclusion operation.Which is also a special case of tree edit distance operation. It mainly focus on the inclusion of one tree into another. Although the paper addresses different kind of tree related problem it requires further research. It also proposes some approaches to work on those problems.
As a data mining student this study is really helpful and it shows one of the practical area to focus on for data mining. Creating nodes and branches are somehow relevant to create partition and centroid.
The study of the tree and different methodologies under it looks highly relevant to data mining. The approach of creating nodes and branches is somewhere similar to creating partition and centroids for them. Additionally tree may be an useful measurement to calculate high dimensional data in a feasible time. As we all know data mining mainly focuses on high dimensional data so this topic is highly relevant to data mining and important for data scientist.

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: number of instances is 699 while cost of biopsies range from $\$1000 to \$5000$ so, The total number of biopsies will range from $\$699,000 to \$3,495,000$. 90%of the cases of 699 instances is 69 cases . Hence, cost of the computer programing will be $629 * 10 = \$6290$.

   (b) What would have been the likely total cost of masectomies?
   Answer: The cost of Masectomy ranges between $\$15,000$ to $\$55,000$and there are 699 instances . Hence, total cost of Masectomy ranges from $\$10,485,000$ to $\$38,445,000$.

(c) Assuming a 70% mortality rate for untreated in year five, how many deaths does the data suggest in five years?
Answer: As per data analysis part, the total number of patients with malignant disease is 241 and mortality rate for untreated in 5 years is 70% . Hence, the number of deaths suggested in five years is approximately 169.

(d) Compose a succint problem statement that you imagine is pertinent to this scenario.
Answer: Based on the data provided about the it seems that the computer programming is 90% accurate. Current year the number of patients is 699. As mentioned the computer's accuracy is 90% so the clinic makes a loss of approximately 69 patients. Although, if the number of patient increases to 2000 then the loss would have been as large as 200. The accuracy should be above 95% to make more profit. There are additional cost of external pathology is also quite high if the hospital can avail a an internal one then it will save the cost. The cost of a masectomy is also high so if they can reduce the cost of these and can bring more profit to the clinic.

2. **Data Preparation**

I have taken all the data in cancer_data structure and then after cleaning the data I have stored them in cleaned_cancer_data using R. Rest operation are described below as the answers of the questions.

Ignoring the `Sample code number` (SCN),

(a) Ignoring the SCN and C columns, how many attributes (or features) does Δ have?
Answer: Except SCN and C there are nine attributes is in Δ.
$ncol(cleaned\_cancer\_data) = 9$

(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 value exist. I have assigned the Δ^{miss} to $missing\_data$ as follows.
$missing\_data \leftarrow cancer\_data[cancer\_data\$nuclei\_chromatin == 0,]$
$dim(missing\_data) = (16)(11)$

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

(d) Give the SCNs for that have missing values.
Answer: SCNs for those missing values are as follows.
$missing\_data\$SCN$
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?
Answer: As the data is about breast cancer I am considering most of the patients are woman. If we review gender based breast cancer statistics less than 5% patients are men. Then the cost pathologist and computer program as mentioned is quite high as calculated above. As per the given data for 699 instances the cost of biopsies are more that $300,000. Plus for 90% cases of 699, that is 629.Hence, the cost of computer programming $6290. So the cost is quite high plus these kind of treatment also effects woman health badly. So I would not recommend re-examination.

(f) Is the amount of missing data significant from an algorithmic perspective?
Answer: There are total 699 samples and the size of missing data is 16, which is approximately $2 - 3\%$ of the entire data. From an algorithmic perspective, the amount of missing data is insignificant as the percentage of data missing is very less. Thus Δ^{miss} would not be affecting the outcome or run time of the algorithm.

(g) Assess the significance of either keeping or removing the tuples with unknown data. You should consider the human element too.
Answer: This dataset has information about other attributes, the missing values will effect

directly the categories under the class attribute and it has an insignificant amount of missing values. So removing the missing value tuples will may effect the the attribute's value significantly accuracy of the output may hamper.
If we keep the data then we must replace these data with some relevant values like mean of the column. After that we must evaluate the effect of the replacement on the output. Based on the output we can decide the added value improved the accuracy or not.

(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: As discussed in the class I am replacing the missing value with the mean of the corresponding column ($nuclei\_chromatin$).
$cleaned\_cancer\_data \leftarrow cancer\_data[, 2:10]$
Replacing the missing value with the mean the column i.e; nuclei_chromatin column of the cleaned_cancer_data.
$cleaned\_cancer\_data\$nuclei\_chromatin[cleaned\_cancer\_data\$nuclei\_chromatin == 0]$
$\leftarrow round(mean(cancer\_data\$nuclei\_chromatin))$
After replacing with the mean value we got the Δ_1^{clean} data in the csv format as DeltaFix.csv file.
$write.csv(cleaned\_cancer\_data, file = "DeltaFix.csv")$

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`,
Answer: I have used PostgreSQL to load data. I have created table first in PostgreSQL then imported the data in the table. After that I have made the connection with R using the following command.
$drv \leftarrow dbDriver("PostgreSQL")$
con = dbConnect(drv, user="postgres", password=" ",
host="localhost", port=5433, dbname="postgres")
`Reading data from Database:`
`cancer_data` = dbReadTable(con, "RestData")
$cleaned\_cancer\_data \leftarrow cancer\_data[, 2:10]$
Replacing the missing value with the mean the column i.e; nuclei_chromatin column of the cleaned_cancer_data.
$cleaned\_cancer\_data\$nuclei\_chromatin[cleaned\_cancer\_data\$nuclei\_chromatin == 0]$
$\leftarrow round(mean(cancer\_data\$nuclei\_chromatin))$

(b) Plot histograms for each attribute and C.
Answer All the plots are kept in the word file titled as `Histogram Plotting for Delta.docx`.

(c) Find the mean, median, mode, and variance of each attribute.
Answer: To calculate the mode value defined one mode function:

- $getmode \leftarrow function(v)\{$
- $uniqv \leftarrow unique(v)$
- $uniqv[which.max(tabulate(match(v, uniqv)))]$ }
  Reference: $https://www.tutorialspoint.com/r/r\_mean\_median\_mode.htm$

Mean, median, mode and variance for each attribute as follows:

- ⋆ mean $\big($cancer_data\$clump_thickness$\big) = 4.41774$
- ⋆ median $\big($cancer_data\$clump_thickness$\big) = 4$
- ⋆ getmode $\big($cancer_data\$clump_thickness$\big) = 1$

- ⋆ var (cancer_data$clump_thickness) = 7.928395
- ⋆ mean (cancer_data$cell_size) = 3.134478
- ⋆ median (cancer_data$cell_size) = 1
- ⋆ getmode (cancer_data$cell_size) = 1
- ⋆ var (cancer_data$cell_size) = 9.311403
- ⋆ mean (cancer_data$cell_shape) = 3.207439
- ⋆ median (cancer_data$cell_shape) = 1
- ⋆ getmode (cancer_data$cell_shape) = 1
- ⋆ var (cancer_data$cell_shape) = 8.832265
- ⋆ mean (cancer_data$adhesion) = 2.806867
- ⋆ median (cancer_data$adhesion) = 1
- ⋆ getmode (cancer_data$adhesion) = 1
- ⋆ var (cancer_data$adhesion) = 8.153191
- ⋆ mean (cancer_data$epithelial_cell_size) = 3.216023
- ⋆ median (cancer_data$epithelial_cell_size) = 2
- ⋆ getmode (cancer_data$epithelial_cell_size) = 2
- ⋆ var (cancer_data$epithelial_cell_size) = 4.903124
- ⋆ mean (cancer_data$nuclei_chromatin) = 3.463519
- ⋆ median (cancer_data$nuclei_chromatin) = 1
- ⋆ getmode (cancer_data$nuclei_chromatin) = 1
- ⋆ var (cancer_data$nuclei_chromatin) = 13.25476
- ⋆ mean (cancer_data$chromatin) = 3.437768
- ⋆ median (cancer_data$chromatin) = 3
- ⋆ getmode (cancer_data$chromatin) = 2
- ⋆ var (cancer_data$chromatin) = 5.94562
- ⋆ mean (cancer_data$normal_nucleoli) = 2.866953
- ⋆ median (cancer_data$normal_nucleoli) = 1
- ⋆ getmode (cancer_data$normal_nucleoli) = 1
- ⋆ var (cancer_data$normal_nucleoli) = 9.32468
- ⋆ mean (cancer_data$mitoses_class) = 1.589413
- ⋆ median (cancer_data$mitoses_class) = 1
- ⋆ getmode (cancer_data$mitoses_class) = 1
- ⋆ var (cancer_data$mitoses_class) = 2.941492
- ⋆ mean (cancer_data$class) = 2.689557
- ⋆ median (cancer_data$class) = 2
- ⋆ getmode (cancer_data$class) = 2
- ⋆ var (cancer_data$class) = 0.9049194

(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: Pearson correlation defined as the measure of the linear dependence between two variables X and Y for values ranging between -1 to +1. Where as cosine is the angle difference between two vectors.Both of these can be viewed as variance of the inner product and both of them are used for centering different magnitudes. The two pairing can get a thicker range of output data in the positive end.

Person's co-efficient for all the variables is attached with the assignment named as "$Pearson\_Corr.csv$".

$correlated\_values \leftarrow cor(cleaned\_cancer\_data, method = "pearson")$

$write.csv(corr\_values, file = "Pearson\_Corr.csv")$

After cleaning the data I got the Δ_2^{clean} $Delta\_2\_clean\_Pearson.csv$ The following R code I have used to get the Δ_2^{clean}. The pair of attributes that are strongly related is replaced by 0.

$delta\_2\_clean \leftarrow replace(x = correlated\_values, correlated\_values == 1, 0)$

After replacing I got the Δ_2^{clean} and have wriiten it in a csv file with the following command.

$write.csv(delta\_2\_clean, file \leftarrow "Delta\_2\_clean\_Pearson.csv")$ The purpose of this step is to remove the attributes which are majorly related to the rest of the value, thereby, so that we can do out calculations as quick as possible.

Reference: $https://brenocon.com/blog/2012/03/cosine-similarity-pearson-correlation-and-ols-coefficients/$

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: I have written the code named as `K_Means Code.py` and is attached.

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.