Homework 2
Computer Science
Fall 2016
B565

Jivitesh Poojary

September 23, 2016

Declaration

All the work done in this assignment is my and my alone. The *R* Programs have been coded and complied in *RStudio*. The k-means implementation is done in *Python*. I have made use of *knitr* package of *RStudio* to embed the programs in the LaTeX file. The database used for the programming questions is *Postgresql*.

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**
        *** Get size of centroid
        $n \leftarrow |c_j^i.B|$
        *** 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 - For some combinations of proximity functions and types of centroids, K-means always converges to a solution; i.e. K-means reaches a state in which no points are shifting from one cluster to another, and hence, the centroids dont change. This is the case because we will be having only finite number of clusters and datapoints. However, the process may take long time to arrive at the convergence. Another point that has to be noted is that k-means will converge to a local minima, this may not be true globally.

   However, this is not the case always. It may be the case that the algorithm may cycle in loop without a substantial change in the distance. To resolve this issue, we can keep a condition wherein the algorithm breaks when most of the points have been correctly clustered and only a tiny proportion of the points change clusters.

   Yes, a bound on the iterate must be included. The bound on the iterate limits the runtime of the

clustering algorithm. However, care must be taken while putting the limit as the quality of the clustering may be poor as a result of fewer iterations.

- Depending on the size of our data we can keep a count on the number of points that are changing clusters. If the number of points changing clusters is less than a predefined percentage of total data points we can stop the iteration.
- We can calculate the change in the centroid values. If the centroids are not changing significantly there is a high possibility that we have almost reached the solution.

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 -
Problem in code:
The code in lines 12-16 tells us to select random points from the domain of the dataset and initializes the block as empty. There are few issues with the code:

- The random value generator function may produce the same values, there has to be a check somewhere in future code that stores the centroid that have already been assigned and each time a new centroid value is generated that is unique.
- The value generated may be an outlier which will degrade the clustering process by making empty or singleton clusters. In this case the centroid selected is far away from other data points and very few or none points will be part of the block partition of the centroid.

Solutions to the problem:

- We can resolve this issue by making multiple runs, each time we should be choosing a different set of centroids and calculating the overall inter-block distance of all the data points in the cluster. And finally choosing a solution with the minimum distance. This may not work always work depending on the final objective, the available data and the required number of clusters. Besides this process adds an element of complexity to the process.
- Another way of resolving the issue is by making sure that the value generated is not an outlier to the range of values in the feature variable.
- - Hierarchical clustering can be used in this case by taking a small subset of data from the source. This solution works well only if the sample we are taking is relatively small and the number of clusters we are looking for are significantly lesser than the subset of datapoints selected.

Implications of using k-means:

- K-means algorithm is simple in nature and the algorithm has a lesser complexity.
- It is also an efficient algorithm, however we can improve the efficiency even further by making use of some techniques and variants in implementation.
- K-means algorithm heavily relies on the initialization of centroids in space. If the initial centroids are not selected correctly, then the whole clustering process may go hay wire. As a result, we may get sub-optimal or undesired clusters.
- We may also get empty clusters if no points are allocated to a cluster during the assignment step.
- - There is a possibility of clusters being formed for outlier values when not desired. Outliers can drastically influence the clustering process. The centroid obtained via this process may not be a true representatives of the dataset as they are supposed to be. Another problem of having outliers it that we may have a singleton cluster where the cluster has only one data point. This is again a sub-optimal solution with a high chance of a better solution with lower intrablock distance.
- If the number of clusters sought are large and the data available is not comparable in nature, then we may get clusters that are too close. There may not be a clear distinction in the clusters.

- It depends on the choice of distance metric that is been used in the algorithm. The Euclidean distance metric is usually preferred in the implementation of k-means. We might get different solutions if applied using Manhattan, Jaccard, Cosine or Bregman divergence distance metrics.
- The implementation also depends on our requirement, if we are interested in squared error distance in hyperspace or only the angle.
- The performance might vary depending on the type of data available for instance if we have data like address or DNA sequence, it is not possible to find a centroid for such a data.
- Another thing that had to be kept in mind while using k-mean is that there is a problem of distance between points in concave and convex distribution not been calculated correctly. This may a problem when we have multidimensional data but because

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

- The space complexity of the algorithm is *O((p + K) * n)*
- The time complexity of the algorithm is *O(I * p * K * n)*,
Where, I is the number of iterations
p is the number of data points that have to be clustered
K is the required number of clusters
n is the number of data attributes that have been considered for clustering

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

Answer -
a) Solution to the tied problem:

- Randomly assign the datum to one of the two centroids
- We can find the cluster with most number of points and assign the point to that cluster. This solution however has a bias over the order of data points that arrive for clustering and may not be the best possible solution.
- The first iteration will have a random assignment and for iterations after the initialization of the algorithm we can store the history of cluster assignment and whenever there is a tie, assign the datum to the cluster it was assigned previously. This solution however has some bias over the previous assignment and again may not be the best possible solution.
- Another possible solution could be that if there is a tie we calculate the mean distance from other points in the cluster, if the tie still persists then do a random assignment

- There is a concept of fuzzy k-means where we have clusters which overlap each other. This a perfect example for that kind of implementation of k-means.

b) Solution to the centroid collapse problem:

- Find the median m of the two centroids and then assign values less than the median to one and greater than the median to another, taking into account an odd number will be a problem here.
- Centroid collapse is not always a bad thing from the global perspective. If we getting a suboptimal solution by keeping centroids which are very close it is better to merge the centroids in this case. In this case we can continuously check the distance between two centroids and when they are close to merging we can drop one of the centroids and create a new centroid randomly again. The process repeats again as per the above algorithm.
- ——We can make use of hierarchical clustering algorithm here

Additional Threshold parameter:

- The k-means clusters are spherical in shape this is because we recursively calculate the mean of centroids and find an average distance between the centroids. This gives us some idea about the τ_c. The only way we can calculate the value the value of this parameter is through rigourous testing on the dataset and trying different values which give us a better solution.
- τ_c can be calculated using multiple techniques, some of them reduce the efficiency at the cost of convergence reduction. While some of them increase the complexity of the program.
- We can find the range of each attribute and divide the range by (2*number of attributes). This should divide the range into equal regions such that each centroid has equal space.
- We should also consider the size of the data that is available, if the size is greater huge we can relax the condition so that smaller blocks are generated an convergence is reached. however for a small size of data

ALGORITHM k-`meansr`
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
*****To Deal with Centroid Collapse i.e. $d(c_1, c_2) \sim 0$
for $j = 1, k$ **do**
    $nCen \leftarrow random(Dom(\Delta))$
    if $nCen \in c\{1, \ldots k\} = TRUE$ **then** Check to see if we have a new list of centroids in our region
        $j \leftarrow j - 1$
    else
        $c^i_j.v \leftarrow nCen$
        $c^i_j.B \leftarrow \emptyset$
    end if
end for

repeat
  $i \leftarrow i + 1$
  *** Assign data point to *nearest* centroid
  for $\delta \in \Delta$ **do**
    ***** To Deal with the tie problem, where the point can go in either direction.
    if $c_x \in \min_{c_j^i}\{d(\delta, c_j^i.v)\} > 1$ **then**
      ▷ Multiple centroids are near the point at an equidistant range ▷ $c_x \subset c_i$
      $c_{i'}.B \leftarrow c.B \cup \{\delta\}$, $c_{i'} \leftarrow random(c_x)$,
    ▷ This is the new centoid that we have obtained using the random funtion form the domain.
    else
      $c_j^i.B \leftarrow c.B \cup \{\delta\}$, where $\min_{c_j^i}\{d(\delta, c_j^i.v)\}$
    end if
  end for
  for $j = 1, k$ **do**
    *** Get size of centroid
    $n \leftarrow |c_j^i.B|$
    *** Update centroid with average
    $c_j^i.v \leftarrow (1/n)\sum_{\delta \in c_j^i.B} \delta$
    *** Remove data from centroid
    $c_j^i.B \leftarrow \emptyset$
  end for
  *** Calculate scalar product (abuse notation and structure slightly)
  *** See notes
until $((1/k)\sum_{j=1}^{k} ||c_j^{i-1} - c_j^i||) < \tau$
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.

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:

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

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

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

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

Let us create a frequency table:

Element	x	y	z	w
a	1	1	0	1
b	1	1	1	0
c	1	0	0	0
d	1	0	0	1
e	0	1	0	1
f	0	0	1	1

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

Answer -

Given that $x = \{a, b, c, d\}$ and $w = \{a, d, f, e\}$

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

$$\begin{aligned} x &\neq w \\ d_1(x,w) &= 1 \end{aligned}$$

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

$$\begin{aligned} J(x,w) &= |x \cap w|/|x \cup w| \quad \text{For sets } x, w. \\ |x \cap w| &= |\{a,d\}| = 2 \\ |x \cup w| &= |\{a,b,c,d,e,f\}| = 6 \\ d_2(x,w) &= 1 - J(x,w) = 1 - 2/6 \\ d_2(x,w) &= 2/3 \end{aligned}$$

$$\begin{aligned} c(x,y) &= \begin{cases} 0, & x = y \\ 1, & otherwise \end{cases} \quad \text{for individual characters, } e.g., \texttt{a} = \texttt{b} \\ d_3(\mathbf{x},\mathbf{y}) &= \Sigma_{i=0}^{n-1} c(\mathbf{x}[i], \mathbf{y}[i]) \quad n = ||\mathbf{x}||, \text{the length of the string.} \\ d_3(\mathbf{x},\mathbf{w}) &= \Sigma_{i=0}^{n-1} c(\mathbf{x}[i], \mathbf{w}[i]) \\ &= c(a,a) + c(b,d) + c(c,f) + c(d,e) \\ &= 0 + 1 + 1 + 1 \\ d_3(\mathbf{x},\mathbf{w}) &= 3 \end{aligned}$$

$$\begin{aligned} d_4(\mathbf{x},\mathbf{y}) &= \left| \frac{\mathbf{x}^\mathsf{T}\mathbf{y}}{||\mathbf{x}||\,||\mathbf{y}||} \right| \quad \text{for vectors } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \\ d_4(\mathbf{x},\mathbf{w}) &= \left| \frac{\mathbf{x}^\mathsf{T}\mathbf{w}}{||\mathbf{x}||\,||\mathbf{w}||} \right| \\ \mathbf{x}^\mathsf{T}\mathbf{w} &= a.a + b.d + c.f + d.e \\ ||\mathbf{x}|| &= \sqrt{a^2 + b^2 + c^2 + d^2} \\ ||\mathbf{w}|| &= \sqrt{a^2 + d^2 + f^2 + e^2} \\ d_4(\mathbf{x},\mathbf{y}) &= \left| \frac{a.a + b.d + c.f + d.e}{\sqrt{a^2 + b^2 + c^2 + d^2}\sqrt{a^2 + d^2 + f^2 + e^2}} \right| = \left| \frac{1 + 0 + 0 + 0}{\sqrt{1+1+1+1}\sqrt{1+1+1+1}} \right| \\ d_4(\mathbf{x},\mathbf{y}) &= 1/4 \end{aligned}$$

(b) Find the d_i that has the minimum value for x, z.
Answer -
Given that $x = \{a, b, c, d\}$ and $z = \{b, f\}$

$$
\begin{aligned}
d_1(x,y) &= \begin{cases} 0, & x = y \\ 1, & x \neq y \end{cases} \quad \text{For objects } x, y. \\
x &\neq z \\
d_1(x,z) &= 0 \\
\\
J(x,y) &= |x \cap y|/|x \cup y| \quad \text{For sets } x, y. \\
d_2(x,y) &= 1 - J(x,y) \quad \text{For sets } x, y. \\
\\
J(x,w) &= |x \cap w|/|x \cup w| \quad \text{For sets } x, w. \\
|x \cap z| &= |\{\}| = 0 \\
|x \cup z| &= |\{b\}| = 1 \\
d_2(x,z) &= 1 - J(x,z) = 1 - 0/1 \\
d_2(x,z) &= 1 \\
\\
c(x,y) &= \begin{cases} 0, & x = y \\ 1, & otherwise \end{cases} \quad \text{for individual characters, } e.g., \texttt{a} = \texttt{b} \\
d_3(\mathbf{x},\mathbf{y}) &= \Sigma_{i=0}^{n-1} c(\mathbf{x}[i], \mathbf{y}[i]) \quad n = ||\mathbf{x}||, \text{the length of the string.} \\
d_3(\mathbf{x},\mathbf{z}) &= \Sigma_{i=0}^{n-1} c(\mathbf{x}[i], \mathbf{z}[i]) \\
&= c(a,b) + c(b,f) + c(c,) + c(d,) \\
&= 1 + 1 + 0 + 0 \\
d_3(\mathbf{x},\mathbf{z}) &= 2 \\
\\
d_4(\mathbf{x},\mathbf{y}) &= \left| \frac{\mathbf{x}^\mathsf{T}\mathbf{y}}{||\mathbf{x}||\,||\mathbf{y}||} \right| \quad \text{for vectors } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \\
d_4(\mathbf{x},\mathbf{z}) &= \left| \frac{\mathbf{x}^\mathsf{T}\mathbf{z}}{||\mathbf{x}||\,||\mathbf{z}||} \right| \\
\mathbf{x}^\mathsf{T}\mathbf{z} &= a.b + b.f + c. + d. \\
||\mathbf{x}|| &= \sqrt{a^2 + b^2 + c^2 + d^2} \\
||\mathbf{z}|| &= \sqrt{b^2 + f^2} \\
d_4(\mathbf{x},\mathbf{z}) &= \left| \frac{a.b + b.f + c. + d.}{\sqrt{a^2 + b^2 + c^2 + d^2}\sqrt{b^2 + f^2}} \right| \\
d_4(\mathbf{x},\mathbf{z}) &=
\end{aligned}
$$

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

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

i. $d_1(x,y)$ will have a maximum value of 1 and minimum value of 0 when $x \neq y$ and $x = y$ respectively.

ii. $d_2(x,y)$ will have a maximum value of 1 and minimum value of 0 when $x \neq y$ (when a and y are disjoint sets) and $x = y$ (when both the sets are equal) respectively.

iii. $d_3(x,y)$ will have a maximum value ∞ and minimum value of 0, when $x \neq y$ c(x,y) gives 1 and when $x = y$, c(x,y) gives 0 for each element.The range of values for $d_3(x,y)$ is $[0, \infty)$.

iv. $d_4(x,y)$ will have a maximum value of 1 and minimum value of 0 when $x \neq y$ and $x = y$ respectively.When either of the vectors is null then we get $0/0$ of an indeterminate answer.

Hence the distance metric d_3 gives the maximum value for any pairs for most of the cases. When both the vectors are same the value of $d_4(x,y)$ has the maximum value of 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

Let be any set such that $v = \{a_1, a_2, ..., a_n\}$

$$
\begin{aligned}
d_1(x,y) &= \begin{cases} 0, & x = y \\ 1, & x \neq y \end{cases} \quad \text{For objects } x, y. \\
\\
v &= v \\
d_1(v,v) &= 0 \\
\\
J(x,y) &= |x \cap y|/|x \cup y| \quad \text{For sets } x, y. \\
d_2(x,y) &= 1 - J(x,y) \quad \text{For sets } x, y. \\
\\
J(v,v) &= |v \cap v|/|v \cup v| \quad \text{For sets } v, v. \\
|v \cap v| &= |\{a_1, a_2, ..., a_n\}| = n \\
|v \cup v| &= |\{a_1, a_2, ..., a_n\}| = n \\
d_2(v,v) &= 1 - J(v,v) = 1 - n/n \\
d_2(v,v) &= 0 \\
\\
c(x,y) &= \begin{cases} 0, & x = y \\ 1, & otherwise \end{cases} \quad \text{for individual characters, } e.g., \texttt{a} = \texttt{b} \\
d_3(\mathbf{x},\mathbf{y}) &= \Sigma_{i=0}^{n-1} c(\mathbf{x}[i], \mathbf{y}[i]) \quad n = ||\mathbf{x}||, \text{the length of the string.} \\
d_3(\mathbf{v},\mathbf{v}) &= \Sigma_{i=0}^{n-1} c(\mathbf{v}[i], \mathbf{v}[i]) \\
&= c(a_1, a_1) + c(a_2, a_2) + c(a_3, a_3) + ... + c(a_n, a_n) \\
&= 0 + 0 + .. + 0 + 0 \\
d_3(\mathbf{v},\mathbf{v}) &= 0 \\
\\
d_4(\mathbf{x},\mathbf{y}) &= \left| \frac{\mathbf{x}^\mathsf{T}\mathbf{y}}{||\mathbf{x}||\,||\mathbf{y}||} \right| \quad \text{for vectors } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \\
d_4(\mathbf{v},\mathbf{v}) &= \left| \frac{\mathbf{v}^\mathsf{T}\mathbf{v}}{||\mathbf{v}||\,||\mathbf{v}||} \right| \\
\mathbf{v}^\mathsf{T}\mathbf{v} &= a_1^2 + a_2^2 + a_3^2 + ... + a_n^2 \\
||\mathbf{v}|| &= \sqrt{a_1^2 + a_2^2 + a_3^2 + ... + a_n^2} \\
||\mathbf{v}|| &= \sqrt{a_1^2 + a_2^2 + a_3^2 + ... + a_n^2} \\
d_4(\mathbf{v},\mathbf{v}) &= \left| \frac{a_1^2 + a_2^2 + a_3^2 + ... + a_n^2}{\sqrt{a_1^2 + a_2^2 + a_3^2 + ... + a_n^2}\sqrt{a_1^2 + a_2^2 + a_3^2 + ... + a_n^2}} \right| \\
d_4(\mathbf{v},\mathbf{v}) &= \left| \frac{a_1^2 + a_2^2 + a_3^2 + ... + a_n^2}{a_1^2 + a_2^2 + a_3^2 + ... + a_n^2} \right| \\
d_4(\mathbf{v},\mathbf{v}) &= 1
\end{aligned}
$$

$d_1(v,v) = d_2(v,v) = d_3(v,v) = 0$ and $d_4(v,v) = 1$, this the above statement is FALSE. We can also see that the test considered is the test of reflexivity, our metric d_4 fails in this test and is not a valid distance metric.

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.

Considering $d_1(x, y)$

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

- $(\forall, a \in Z)\ d(a,a) = 0.$
  Then $d_{i'}(a,a) = \frac{d_1(a,a)}{1+d_1(a,a)} = \frac{0}{1+0} = 0$

Thus, $d_{i'}(x, y)$ satisfies the property of reflexivity

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

$$d_{i'}(a,b) = \frac{d_1(a,b)}{1+d_1(a,b)} = \frac{d_1(b,a)}{1+d_1(b,a)} = d_1(b,a) = d_{i'}(b,a)$$

- Since, $d_1(a,b) = d_1(b,a)$

Thus, $d_{i'}(x, y)$ satisfies the property of symmetry

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

$$\begin{aligned} d_{i'}(a,b) + d_{i'}(b,c) &= \frac{d_1(a,b)}{1+d_1(a,b)} + \frac{d_1(b,c)}{1+d_1(b,c)} \\ &\geq \frac{d_1(a,c)}{1+d_1(a,c)} = d_{i'}(a,c) \\ LHS &= \frac{d_1(a,b)}{1+d_1(a,b)} + \frac{d_1(b,c)}{1+d_1(b,c)} \\ RHS &= \frac{d_1(a,c)}{1+d_1(a,c)} \end{aligned}$$

For $a \neq b \neq c$

$$\begin{aligned} LHS &= \frac{1}{1+1} + \frac{1}{1+1} = 1 \\ RHS &= \frac{1}{1+1} = \frac{1}{2} \\ LHS &> RHS \end{aligned}$$

For $a = b \neq c$

$$\begin{aligned} LHS &= \frac{0}{1+0} + \frac{1}{1+1} = \frac{1}{2} \\ RHS &= \frac{1}{1+1} = \frac{1}{2} \\ LHS &= RHS \end{aligned}$$

For $a = b = c$

$$\begin{aligned} LHS &= \frac{0}{1+0} + \frac{0}{1+0} = 0 \\ RHS &= \frac{0}{1+0} = 0 \\ LHS &= RHS \end{aligned}$$

From the above three conditions we can see that

$$\begin{aligned} LHS &\geq RHS \\ \frac{d_1(a,b)}{1+d_1(a,b)} + \frac{d_1(b,c)}{1+d_1(b,c)} &\geq \frac{d_1(a,c)}{1+d_1(a,c)} \\ d_{i'}(a,b) + d_{i'}(b,c) &\geq d_{i'}(a,c) \end{aligned}$$

Thus, $d_{i'}(x,y)$ satisfies the property of transitivity

Considering $d_2(x,y)$

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

Considering $d_3(x,y)$

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

Considering $d_4(x,y)$

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

All the three properties of the distance metric have been proved for d_1 distance metric. As we have proved the required for one distance metric, it will be true for all the other distance metrics also.

(b) $d_{i'}(x,y) = \alpha d_i(x,y)$ for $\alpha \in \mathbb{R}_{>0}$
Considering i as 1 hence our distance metric is $d_1(x,y)$

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

- $(\forall, a \in Z)\ d(a,a) = 0$. Then $d_{i'}(a,a) = \alpha d_1(a_0,a_0) = \alpha.0 + \alpha.0 = 0$

Thus, $d_{i'}(x, y)$ satisfies the property of reflexivity

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

$$d_{i'}(a, b) = \alpha d_1(a, b) = \alpha d_1(b, a) = d_{i'}(b, a)$$

- Since, $d_1(a, b) = d_1(b, a)$

Thus, $d_{i'}(x, y)$ satisfies the property of symmetry

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

$$\begin{aligned} d_{i'}(a, b) + d_{i'}(b, c) &= \alpha d_1(a, b) + \alpha d_1(b, c) \\ &= \alpha[d_1(a, b) + d_1(b, c)] \\ &\geq \alpha d_1(a, c) = d_{i'}(a, c) \end{aligned}$$

Thus, $d_{i'}(x, y)$ satisfies the property of transitivity

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

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

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

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

$$\begin{aligned} d_5(a, a) &= d_1(a_0, a_0) + 3d_2(a_1, a_1) \\ &= 0 + 3[1 - J(a_1, a_1)] \\ J(a_1, a_1) &= |a_1 \cap a_1|/|a_1 \cup a_1| = 1 \end{aligned}$$

$$d_5(a, a) = 0 + 3[1 - 1] = 0$$

Thus, $d_5(x, y)$ satisfies the property of reflexivity

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

$$\begin{aligned} Since, |a \cap b| &= |b \cap a| \\ |a \cup b| &= |b \cup a| \\ |b \cap a|/|b \cup a| &= |a \cap b|/|a \cup b| \\ J(b, a) &= J(a, b) \\ d_5(a, b) &= d_1(a_0, b_0) + 3d_2(a_1, b_1) \\ &= d_1(a_0, b_0) + 3[1 - J(a_1, b_1)] \\ &= d_1(b_0, a_0) + 3[1 - J(b_1, a_1) \\ &= d_1(b_0, a_0) + 3d_2(b_1, a_1) \\ d_5(a, b) &= d_5(b, a) \end{aligned}$$

Thus, $d_5(x, y)$ satisfies the property of symmetry

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

$$\begin{aligned}
Since, |a \cap b| &= |b \cap a| \\
|a \cup b| &= |b \cup a| \\
|b \cap a|/|b \cup a| &= |a \cap b|/|a \cup b| \\
J(b,a) &= J(a,b) \\
d_5(a,b) + d_5(b,c) &= d_1(a_0,b_0) + d_1(b_0,c_0) + 3d_2(a_1,b_1) + 3d_2(b_1,c_1) \\
&= d_1(a_0,b_0) + d_1(b_0,c_0) + 3[1 - J(b,a)] + 3[1 - J(c,b)] \\
&= d_1(a_0,b_0) + d_1(b_0,c_0) + 3[1 - |b \cap a|/|b \cup a|] + 3[1 - |c \cap b|/|c \cup b|] \\
&= d_1(a_0,b_0) + d_1(b_0,c_0) + 3[1 - |a \cap b|/|a \cup b| + 1 - |b \cap c|/|b \cup c|] \\
&\geq d_1(a_0,c_0) + 3d_2(a_1,c_1) = d_5(a,c)
\end{aligned}$$

Thus, $d_5(x,y)$ satisfies the property of transitivity, and hence it is a metric

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

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

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

$$\begin{aligned}
d_6(a,a) &= d_2(a,a) \\
&= 1 - J(a,a) \\
J(a,a) &= |a \cap a|/|a \cup a| = 1 \\
d_6(a,a) &= 1 - 1 = 0
\end{aligned}$$

Thus, $d_6(x,y)$ satisfies the property of reflexivity

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

$$\begin{aligned}
Since, |a \cap b| &= |b \cap a| \\
|a \cup b| &= |b \cup a| \\
|b \cap a|/|b \cup a| &= |a \cap b|/|a \cup b| \\
J(b,a) &= J(a,b) \\
d_6(a,b) &= d_2(b,a) \\
&= 1 - J(b,a) \\
&= 1 - J(a,b) \\
&= d_2(a,b) \\
d_6(a,b) &= d_6(b,a)
\end{aligned}$$

Thus, $d_6(x,y)$ satisfies the property of symmetry

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

$$\begin{aligned}
Since, |a \cap b| &= |b \cap a| \\
|a \cup b| &= |b \cup a| \\
|b \cap a|/|b \cup a| &= |a \cap b|/|a \cup b| \\
J(b,a) &= J(a,b) \\
d_6(a,b) + d_6(b,c) &= d_2(b,a) + d_2(c,b) \\
&= 1 - J(b,a) + 1 - J(c,b) \\
&= 1 - |b \cap a|/|b \cup a| + 1 - |c \cap b|/|c \cup b| \\
&= 1 - |a \cap b|/|a \cup b| + 1 - |b \cap c|/|b \cup c| \\
&= d_2(a,b) + d_2(b,c) \\
&\geq d_2(c,a) = d_6(a,c)
\end{aligned}$$

Thus, $d_6(x,y)$ satisfies the property of transitivity, and hence it is a metric

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

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

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

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

$$\begin{aligned}
d_7(a,a) &= d_3(a_0,a_0)d_2(a_1,a_1) \\
d_3(a_0,a_0) &= 0 \\
d_2(a_1,a_1) &= 0 \\
d_7(a,a) &= 0.0 = 0
\end{aligned}$$

Thus, $d_7(x,y)$ satisfies the property of reflexivity

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

$$\begin{aligned}
d_7(a,b) &= d_3(a_0,b_0)d_2(a_1,b_1) \\
&= [\Sigma_{i=0}^{n-1} c(\mathbf{a_0}[i], \mathbf{b_0}[i])].[1 - J(a_1,b_1)] \\
&= [\Sigma_{i=0}^{n-1} c(\mathbf{b_0}[i], \mathbf{a_0}[i])].[1 - J(b_1,a_1)] \\
&= d_3(b_0,a_0)d_2(b_1,b_1) \\
&= d_7(b,a) \\
d_7(a,b) &= d_7(b,a)
\end{aligned}$$

Thus, $d_7(x,y)$ satisfies the property of symmetry

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

$$\begin{aligned} d_7(a,b) + d_7(b,c) &= d_3(a_0,b_0)d_2(a_1,b_1) + d_3(b_0,c_0)d_2(b_1,c_1) \\ &= [\Sigma_{i=0}^{n-1} c(\mathbf{a_0}[i], \mathbf{b_0}[i])].[1 - J(a_1,b_1)] + [\Sigma_{i=0}^{n-1} c(\mathbf{b_0}[i], \mathbf{c_0}[i])].[1 - J(b_1,c_1)] \\ &= - - - - - - - - - - - - - \\ &\geq d_x(a_0,c_0) + d_y(a_1,c_1) = d(a,c) \end{aligned}$$

Thus, $d_7(x,y)$ does not satisfies the property of transitivity, and hence it is not a metric

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

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

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

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

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

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

$$\begin{aligned} d_8(a,a) &= d_1(a_0,a_0) + d_2(a_1,a_1) + d_3(a_2,a_2) + d_4(a_3,a_3) \\ d_8(a,a) &= 0 + 0 + 0 + 1 \\ d_8(a,a) &= 1 \neq 0 \end{aligned}$$

This fails the reflexivity test, hence $d_8(x,y)$ is not a metric.

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 -
The authors have surveyed the problem of comparing labeled trees based on simple local operations of deleting, inserting, and relabeling nodes. These operations lead to the tree edit distance, alignment distance, and inclusion problem. For each problem they have reviewed the results available and present, in detail, one or more of the central algorithms for solving the problem. The authors have discussed concepts liek the tree edit distance and tree inclusion criterias. Recently however, more advanced techniques such as fast matrix multiplication have been applied to the tree edit distance problem, the paper helps us to understand it in more detail.

The authors have also discussed the various other algorithms like the Zhang and Shashas algorithm, Kleins algorithm and Jiang, Wang, and Zhangs algorithm for our mathematical understanding. They have also covered imprtant concept of Tree Alignment Distance, Tree inclusion. They have surveyed the tree edit distance, alignment distance, and inclusion problems. Furthermore, we have presented, in our opinion, the central algorithms for each of the problems. There are several open problems, which may be the topic of further research.

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 -

   i. The cost of a single biopsy is from \$1000 to \$5000 and there are in all 699 records in our dataset. As the cost is a range, we can say that the cost of bipsies in when done by a pathalogist will be (\$1000 to \$5000) x 699, i.e. \$699,000 to \$3,495,000.

   ii. If the computer can identify 90% of the cases to nearly 100% accuracy. We can conclude that 90% of the 699 records have been correctly classified. Hence the cost of the computer program will be \$10 x 699, i.e. \$6,990. The cost of the computer program does not depend on the accuracy, irrespective of the accuracy of the program it is going to be executed for that tuple.

   (b) What would have been the likely total cost of masectomies?
   Answer -
   Out of the 699 records 241 readings show malignant classification (this information is given as part of the dataset description) and the cost of a masectomy is \$15,000 to \$55,000. Thus we can say that the likely cost of masectomy is (\$15,000 to \$55,000) x 241, i.e. \$ 3,615,000 to \$13,255,000

   (c) Assuming a 70% mortality rate for untreated in year five, how many deaths does the data suggest in five years?
   Answer -
   241 patient records showed malignant features for them. If the assumption that "a 70% mortality rate for untreated in year five" hold true we can conclude that (241 x 70%), i.e. 169 deaths will happen in five years.

(d) Compose a succint problem statement that you imagine is pertinent to this scenario.
Answer -
—— Understand the features of benign and malignant cancers and seperate them out by executing a computer program using sample data of the patients breasts.

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 - Δ has 9 attributes (or 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 - The total number of missing values is 16. The size of Δ^{miss} is also 16.

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

(d) Give the SCNs for that have missing values.
Answer - The following are the SCNs for missing value: "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 - Yes, I would recommend the women to undergo re-examination. The women are at a risk of disease progression and may miss an opportunity for early treatment which will have grave consequences.

i. *Cost for the pathologist* - As there are 16 records which have missing values in them the cost of the computer program will be in a range ($1000 to $5000) x 16 = $16000 to $80000.
ii. *Cost for the computer program* - As there are 16 records which have missing values in them the cost of the computer program will be $10 x 16 = $160

(f) Is the amount of missing data significant from an algorithmic perspective?
Answer - —-No, the amount of missing data is insignificant from an algorithmic perspective. We have an algorithm which itself has some complexity as it is very iterative in nature. So the small amount of missing data wont have a drastic effect from an algorithmic point of view.

(g) Assess the significance of either keeping or removing the tuples with unknown data. You should consider the human element too.
Answer - The tuples with unknown data has to be kept for our analysis.

i. *Significance of removing the unknown data* - If the unknown data was very small in comparison with the total dataset size we would be looking at removing the tuples. Similarly if the data we were having was abundant in nature and freely available externally we would have considered removing it.
ii. *Significance of keeping the unknown data* - The proportion of missing data in this case is (16/699) = 2.289%. This may be relatively small proportion but each tuple in the dataset represents a person. This person can be any women, who may be suffering from breast cancer or may have a risk of developing one in the future. If we were to leave out these tuples we would be letting down the indivisuals who have shared this personal data. Besides, we would also be missing out on an opportunity to treat other women who may have the same condition but have not shared thier data.

iii. I'm keeping the data in my dataset for the execution ofmy computer program. Breast cancer is one of the leading cancers in women around the world. Early detection and improved therapy planning are crucial for increasing the survival rates of cancer patients. So I would say that we should be keeping the tuples of unknown data in our dataset.

(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}.

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.

(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 - The relation between $cos\theta$ and the Pearson correlation coefficent is that, both of them give us some idea about the similarity of the attributes in the data set. This process of data cleaning is important as we will be predicting with a bias the same thing if we were to use both the attributes. They do not provide any extra information to understand our classification process. Once we remove the highly correlational attributes we get a model that a more efficient at predicting our end classification.

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:

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

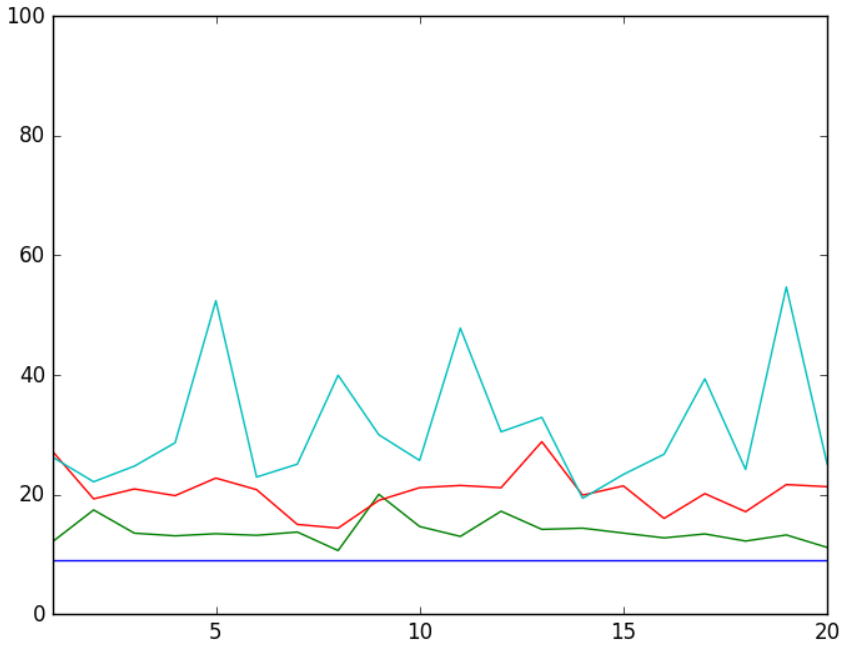

Figure 1: KMeans Error Graph for different values of K

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 - The problem statement of the excercise was Understand the features of benign and malignant cancers and seperate them out by executing a computer program using sample data of the patients breasts. There are two figures below, one is with a regulated threshold for inter centroid distance the other is for a hard break with a fixed limit on iterations. The results obtained from the program show the following:

1. From the graphs we can see that as there curves progress they are not linear most of the times. There are peaks and valleys in our graph. This is because our program on kmeans heavily depends on the initial points. More farther the points are from and optimum distance greater is the chance of miss-classification and thus resulting in a higher error rate.
2. The error rate increases with larger number of centroids i.e. larger values of k.

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.

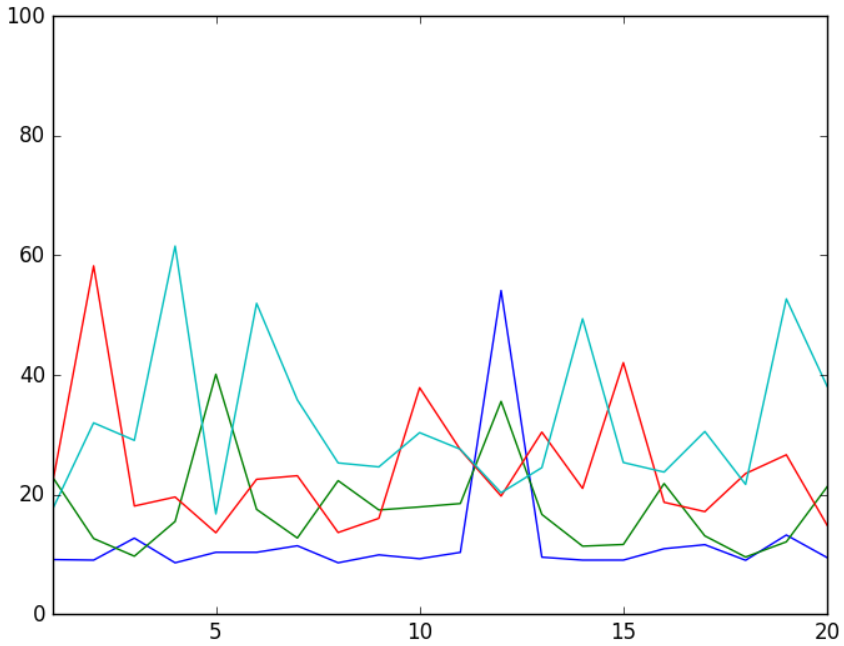

Figure 2: KMeans Error Graph for different values of K with Threshold

Extra analysis on Medical Data

Information on the data set:

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

1. **Clump thickness:** Benign cells tend to be grouped in monolayers, while cancerous cells are often grouped in multilayers.
2. **Uniformity of cell size/shape:** Cancer cells tend to vary in size and shape. That is why these parameters are valuable in determining whether the cells are cancerous or not.
3. **Marginal adhesion:** Normal cells tend to stick together. Cancer cells tend to lose this ability. So loss of adhesion is a sign of malignancy.
4. **Single epithelial cell size:** Is related to the uniformity mentioned above. Epithelial cells that are significantly enlarged may be a malignant cell.

5. **Bare nuclei:** This is a term used for nuclei that is not surrounded by cytoplasm (the rest of the cell). Those are typically seen in benign tumours.

6. **Bland Chromatin:** Describes a uniform "texture" of the nucleus seen in benign cells. In cancer cells the chromatin tends to be coarser.

7. **Normal nucleoli:** Nucleoli are small structures seen in the nucleus. In normal cells the nucleolus is usually very small if visible at all. In cancer cells the nucleoli become more prominent, and sometimes there are more of them.

8. **Mitoses:** Mitotic rate is acknowledged to be an important independent adverse predictor of survival; as the number of mitoses/mm2 increases, cancer survival decreases.

These are terms used in a pathology report on fine needle aspirations to assess whether a lump in a breast could be malignant (cancerous) or benign (non-cancerous). For instance cancer cells tend to vary in size and shape. So uniformity of cell size/shape points in a benign direction. Also bare nuclei, bland chromatin and normal nucleoli are signs of benignity.

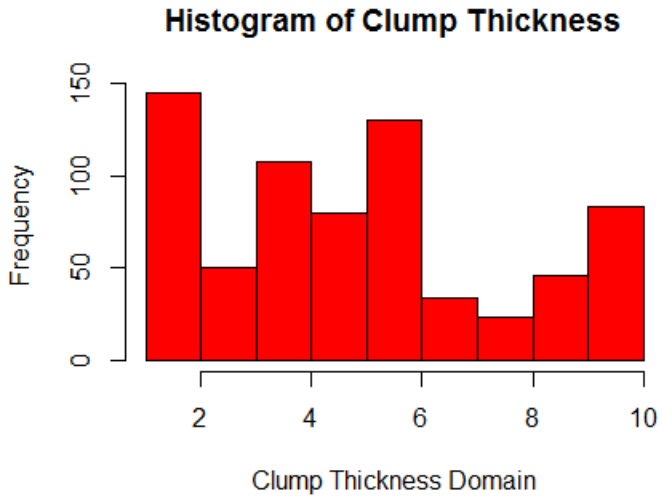

Figure 3: Histogram of Clump Thickness

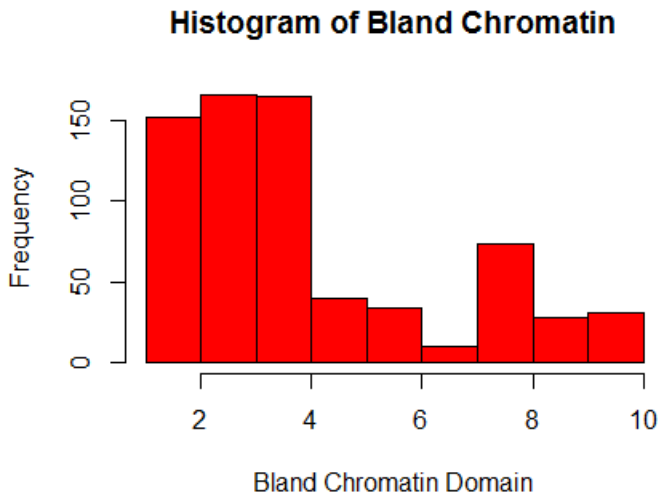

Figure 4: Histogram of Chromatin

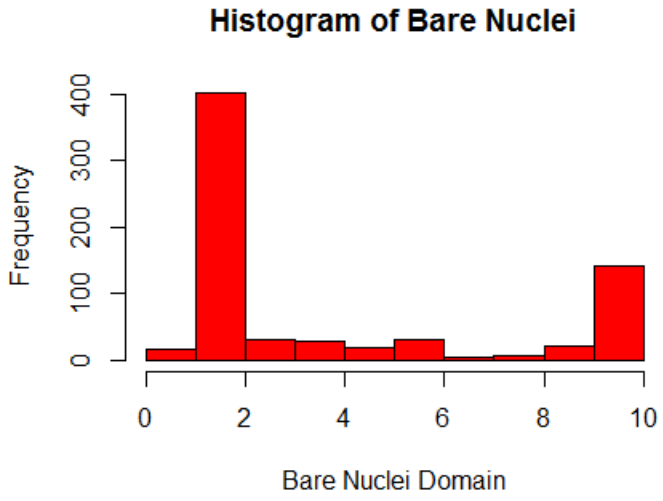

Figure 5: Histogram of Bare Nuclei

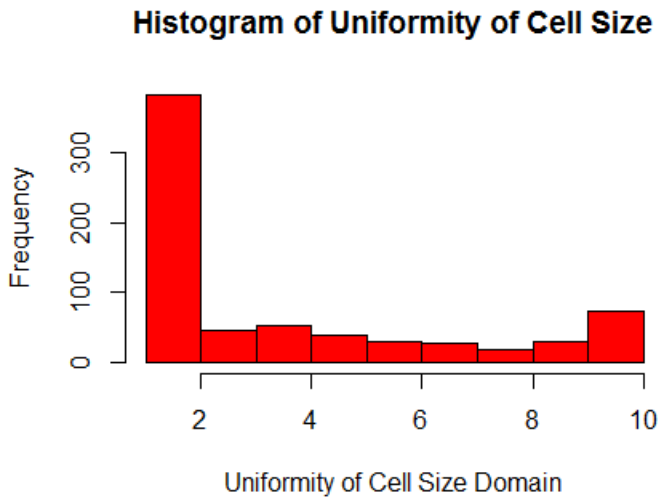

Figure 6: Histogram of Cell Size

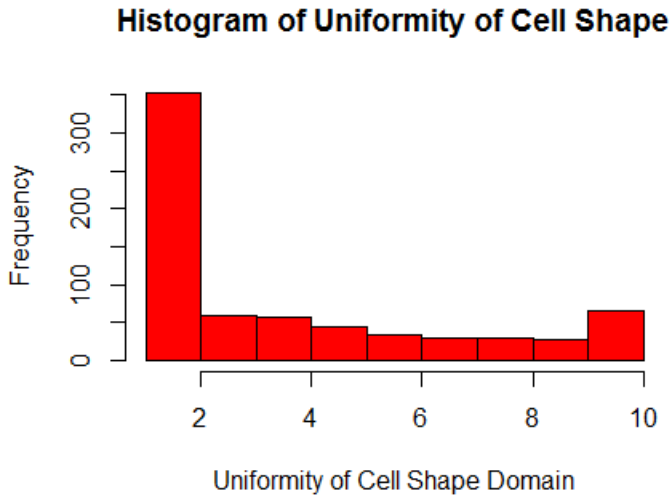

Figure 7: Histogram of Cell Shape

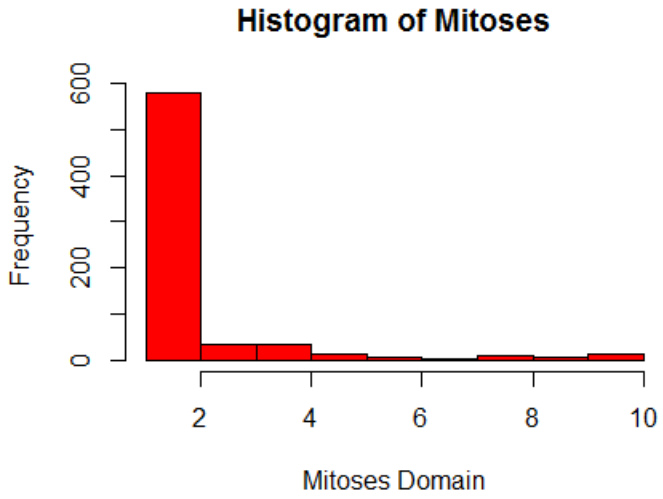

Figure 8: Histogram of Mitoses

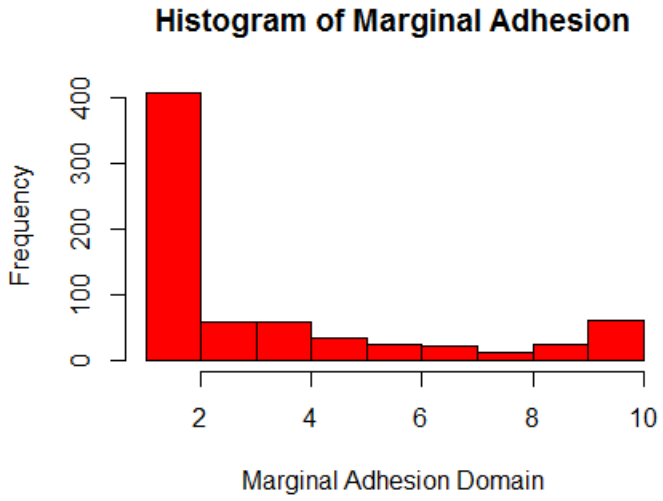

Figure 9: Histogram of Marginal Adhesion

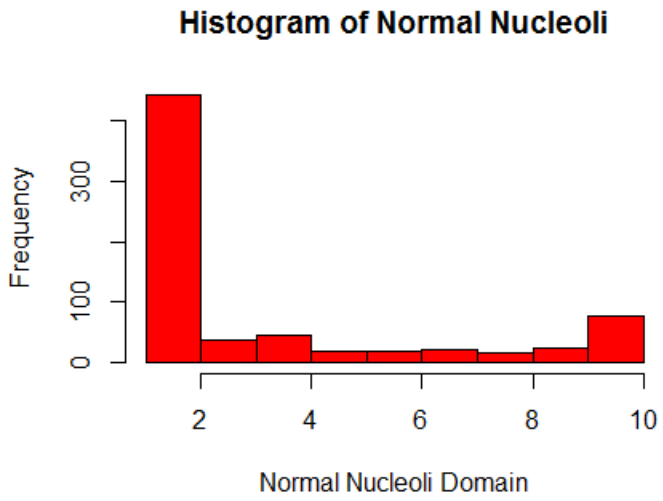

Figure 10: Histogram of Normal Nucleoli

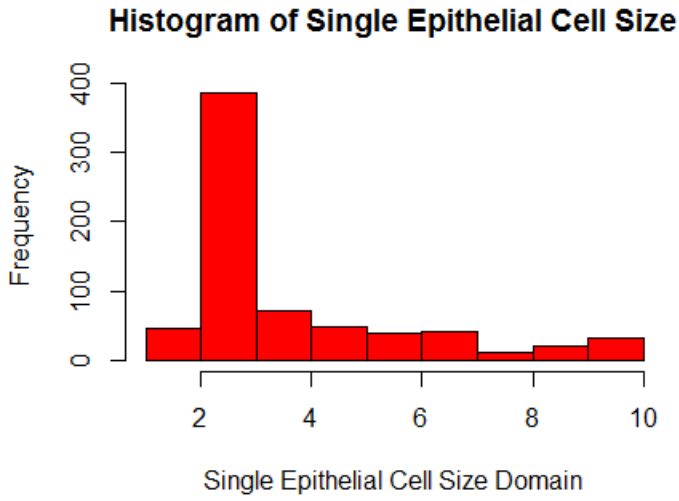

Figure 11: Histogram of Single Epithelial Cell Size

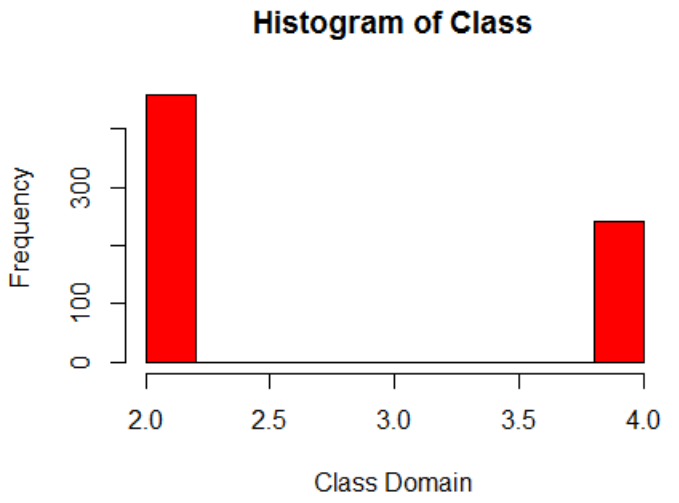

Figure 12: Histogram of Clump Thickness