Homework 1
Computer Science
Fall 2016
B565

Jivitesh Poojary

September 3, 2016

Declaration

All the work done in this assignment is my and my alone

Notation and Definitions

- Mathematical structures are italicized. For example, we write a set as X, not X.
- Use standard notation. For example, $\cup, \cap$ for union and intersection.
- Let Π be a partition of X. We define a binary relation $\sim_\Pi$ on X as follows:

$$x \sim_\Pi y \quad \leftrightarrow \quad x, y \in B, B \in \Pi$$

  We write $\sim_{\Pi[x]}$ to mean $B \in \Pi$ such that $x \in B$

Definition Equivalence Relation.
Let Π be a partition of X and $\sim_\Pi$ be a on relation on X. $\sim_\Pi$ is an equivalence relation if:

- Reflexivity $(\forall x \in X)\ x \sim_\Pi x$
- Symmetry $(\forall x, y \in X)\ x \sim_\Pi y \rightarrow y \sim_\Pi x$
- Transitivity $(\forall x, y, z \in X)\ x \sim_\Pi y \wedge y \sim_\Pi z \rightarrow x \sim_\Pi z$

Problems and Answers

1. Define the following terms

   (a) Partition of a non-empty (finite) set X.
   Answer - Partiiton of a non-empty (finite) set X is defined as splitting of the set X into non-empty blocks such that each block has a a unique characteristic that binds all the elements of that block in such a way that each element is unique to that block.

   (b) Distance metric d over X.
   Answer - The distance metric d over X is defined as a function for computing the distance between two elements of a set. It is used to find elements which share similar characteristics. Some of properties that are satified by the distance metric are -

Positivity - $(\forall\, x)\ d(x,x) = 0$.
Symmetry - $(\forall\, x, y)\ d(x,y) = d(y,x)$
Transitivity - $(\forall\, x, y, z)\ d(x,y) + d(y,z) \geq d(x,z)$

(c) Show that given an equivalence relation $\sim$ over a non-empty (finite) set X, there is an associated *unique* partition Π.

(d) Given a set X, a partition Π, and the equivalence relation $\sim_\Pi$, and distance metric d over X, choose two distinct points $x, y \in X$ such that

i. $x, y \in B \in \Pi$ where $d(x,y) = k$. Then there must be two distinct points $a, b \in X$ where $a \sim_{\Pi[x]} b$ such that $d(a,b) = k$. TRUE OR FALSE
Answer - TRUE x and y are two elements in the same block B which is present in partition . Elements a and b have an equivalence relationship and are present in the same block. Since the equation $d(x,y) = k$ is true for x and y it must also be true for $d(x,y) = k$.

ii. $x \in B, y \in B'$ where $B, B' \in \Pi \wedge B \neq B'$ and $d(x,y) = k$. Then there may not be two distinct points $a, b \in X$ such that $d(a,b) = k$. TRUE OR FALSE
Answer - FALSE x is an element of block B and y is an element of block B' , blocks B and B' are distinct and the equation $d(x,y) = k$ holds true.

iii. If $|X| = n$, then there must be n distinct, non-empty blocks $B_1, B_2, \ldots, B_n \in \Pi$ TRUE OR FALSE
Answer - FALSE The cardinality of set X is n, however this does not necessary mean that there has to be n non-empty blocks part of the partition. The key word here is must which makes the statement false.

iv. If $|\Pi| = n$, then there are n distinct $x_1, x_2, \ldots, x_n \in X$. TRUE OR FALSE
Answer - TRUE If the cardinality of the partition is n then all these elements have to be part of the set X. The total number of elements in the set may be more and each block in the partition may have more than one element. But as the blocks of the partiton cannot be empty we are guaranteed to have atleast n distinct elements.

v. $x \in B, y \in B'$ where $B, B' \in \Pi \wedge B \neq B'$ and $d(x,y) = k$. Then $(\forall x' \in X)\ x' \in B \wedge x \neq x' \rightarrow d(x,x') < k$. TRUE OR FALSE
Answer - FALSE x is an element of block B and y is an element of block B. Both the blocks are of the same partition and they are not equal to each other. The distance between the two points x and y is k. Then for element x belonging to block B the statement d(x,x) ¡k does not necessarily hold true. This statement gives an example of one of the possibilities and cannot be said to be true always.

2. Let $X = \{1,2,3,4,5,6,7\}$. Find the *smallest* equivalence relation $\sim$ such that:

$$\begin{aligned} (1,2) &\in \sim \\ (2,3) &\in \sim \\ (5,2) &\in \sim \\ (4,6) &\in \sim \\ (7,7) &\in \sim \end{aligned}$$

Why would a data scientist be interested in the smallest?
Answer - The set X is defined as X=1,2,3,4,5,6,7. Some of the elements showing the equivalence relations are (1,2), (2,3), (5,2), (4,6), (7,7).
Using the symmetry rule of equivalence relations
We get the following elements (2,1), (3,2), (2,5), (6,4)

Further, we apply the transitivity rule to the equivalence relationships
We get the following elements (1,5), (5,3)
From the above two examples we can conclude that elements 1,2,3 and 5 belong to the same block, similarly elements 4 and 6 belong to the same block and finally the element 7 belongs to different block.
$\Pi = \{\{1, 2, 3, 5\}, \{4, 6\}, \{7\}\}$

A data scientist would be interested in the smallest number of equivalence set is to avoid the problem of saturation or overfitting. If we have all the outcomes or possible elements of the superset we would be memorizing the outcome instead of actually the training the system. With the smallest number of equivalence relations in place we can use our algorithms and techniques in a more efficient manner and make an actual prediction instead of mapping a question to an answer.

3. Let $X = \{0, 3, 7, 8, 9\}$. Form a partition that has three blocks such that

$$d(x, y) = [(x - y)^2]^{1/2}$$

has a minimum intrablock distance.

INPUT `TOTIntraBlockDis`(Set $X = \{B_1, B_2, \ldots, B_n\}$, distance d over X)
▷ X is a partition and B_i are the blocks.
OUTPUT $R_{\geq 0}$ v
$v \leftarrow 0$
for $i = 1, n$ **do**
  $v \leftarrow v +$ `IntraBlockDis`(B_i, d)
end for
return v

INPUT `IntraBlockDis`(Set $X = \{x_1, x_2, \ldots, x_n\}$, distance d)
OUTPUT $R_{\geq 0}$ v
$v \leftarrow 0$
for $i = 1, n - 1$ **do**
  for $j = i + 1, n$ **do**
    $v \leftarrow v + d(i, j)$
  end for
end for
return v

Example. Assume $X = \{1, 2, 3, 4, 5\}$ and $d(x, y) = |x - y|$. Then `IntraBlockDis`$(X, d) = 20$. The calculation is shown below:

i	j	$d(i,j)$	v
1	2	1	1
	3	2	3
	4	3	6
	5	4	10
2	3	1	11
	4	2	13
	5	3	16
3	4	1	17
	5	2	19
4	5	1	20

Answer - For the above question since we need to form a partition that has three blocks -there are two possible solutions, the solutions depend on the grouping method we use for creating blocks.

SOLUTION 1 $\Pi = \{\{0,3\},\{7,8\},\{9\}\}$
With the above partition we can calculate the `IntraBlockDis` as follows

i	j	$d(i,j)$	v
0	3	3	3
7	8	1	4
9	9	0	4

SOLUTION 2 $\Pi = \{\{0,3\},\{7\},\{8,9\}\}$
With the above partition we can calculate the `IntraBlockDis` as follows

i	j	$d(i,j)$	v
0	3	3	3
7	7	0	3
8	9	1	4

In both the above solutions the `IntraBlockDis`(X,d) comes out to be 4. As a result both of them are valid solutions

4. Show the results of `TOTIntraBlockDis`$(\{\{1,2\},\{3\},\{4,10\}\}, d(x,y) = |x-y|)$.
Answer - We are required to compute the total intrablock distance of partition containing three blocks given the elements in the individual blocks and the distance metric
Block 1 1,2
Block 2 3
Block 3 4,10

Block	i	j	$d(i,j)$	v
Block 1	1	2	1	1
Block 2	3	3	0	1
Block 3	4	10	6	7

5. Write the `InterBlockDis` algorithm that takes a partition and distance function and returns the distance between blocks. Use this function to calculation the interblock distance on the partition in Problem 3.
Answer - Please find below the algorithm for computing the inter block distance.

INPUT `TOTInterBlockDis`(Set $X = \{B_1, B_2, \ldots, B_n\}$, distance d over X)
▷ X is a partition and B_i are the blocks.
OUTPUT $R_{\geq 0}$ v
$v \leftarrow 0$
for $i = 1, n-1$ **do**
  for $j = i+1, n$ **do**
    $v \leftarrow v +$ `InterBlockDis`(B_i, B_j, d)
  end for
end for
return v

INPUT `InterBlockDis`(Set $X = \{x_1, x_2, \ldots, x_n\}$, Set $Y = \{y_1, y_2, \ldots, y_n\}$, distance d)
OUTPUT $R_{\geq 0}$ v
$v \leftarrow 0$

for $i = 1, n$ **do**
  for $j = 1, n$ **do**
    $v \leftarrow v + d(x_i, y_j)$
  end for
end for
return v

As per Problem 3 the three blocks that were obtained by us were $\{0, 3\}, \{7, 8\}$ and $\{9\}$. If were to apply the `InterBlockDis` algorithm in this case we would get the following calculations.

i	j	$d(i,j)$	v
0	7	7	7
	8	8	15
	9	9	24
3	7	4	28
	8	5	33
	9	6	39
7	9	2	41
8	9	1	42

6. This problem asks you to prove (or disprove) that a function d is a metric. We give an example of a proof first.

   Prove (or disprove with a counter example) that d defined below is a metric.

   Proof

   Let $d : R^2_{\geq 0} \to R_{\geq 0}$ such that

$$d(x, y) = \begin{cases} |x - y| / \max\{x, y\}, & x + y > 0 \\ 0 \quad o.w. \end{cases}$$

   $(\forall\, x)\ d(x, x) = 0$.
   Assume $a = 0$
   $d(a, a) = 0$ by definition.
   Assume $a > 0$ w.l.o.g.
   $d(a, a) = |a - a|/a = 0$
   $(\forall\, x, y)\ d(x, y) = d(y, x)$
   Assume $a \leq b$. Then $\max\{a, b\} = b$.
   Since $d(a, b) = (b - a)/b$ and $d(b, a) = (b - a)/b$, then $d(a, b) = d(b, a)$
   $(\forall\, x, y, z)\ d(x, y) + d(y, z) \geq d(x, z)$
   Assume $a \leq b \leq c$ w.l.o.g.
   $(b - a)/b + (c - b)/c \geq (c - a)/c$
   $(b - a)/b \geq (b - a)/c$
   $c \geq b$

   Prove (or disprove with a counter example) that d is a metric.

   Let $d : R^2 \to R_{\geq 0}$ such that

$$d(x, y) = \left| \frac{x}{1 + |x|} - \frac{y}{1 + |y|} \right|$$

Answer - We have to prove d is also a distance metric given that Since we are using a L2 Norm with R ≥ 0 both the elements x ≥ 0 and y ≥ 0

Proof

Let $d : R^2_{\geq 0} \to R_{\geq 0}$ such that

$$d(x, y) = \left| \frac{x}{1 + |x|} - \frac{y}{1 + |y|} \right|$$

$(\forall\, x)\ d(x, x) = 0.$
Assume $a = 0$
$d(a, a) = \left| \frac{0}{1+|0|} - \frac{0}{1+|0|} \right| = 0$
Assume $a > 0$ w.l.o.g.
$d(a, a) = \left| \frac{a}{1+a} - \frac{a}{1+a} \right| = 0$
$(\forall\, x, y)\ d(x, y) = d(y, x)$
Assume $a \geq b$.
Since $d(a, b) = \left| \frac{a}{1+a} - \frac{b}{1+b} \right|$ and $d(b, a) = \left| \frac{b}{1+b} - \frac{a}{1+a} \right| = \left| \frac{a}{1+a} - \frac{b}{1+b} \right|$, then $d(a, b) = d(b, a)$
$(\forall\, x, y, z)\ d(x, y) + d(y, z) \geq d(x, z)$
Assume $a \geq b \geq c$ w.l.o.g.
$\left| \frac{a}{1+a} - \frac{b}{1+b} \right| + \left| \frac{b}{1+b} - \frac{c}{1+c} \right| \geq \left| \frac{c}{1+c} - \frac{a}{1+a} \right|$
$c \geq b$

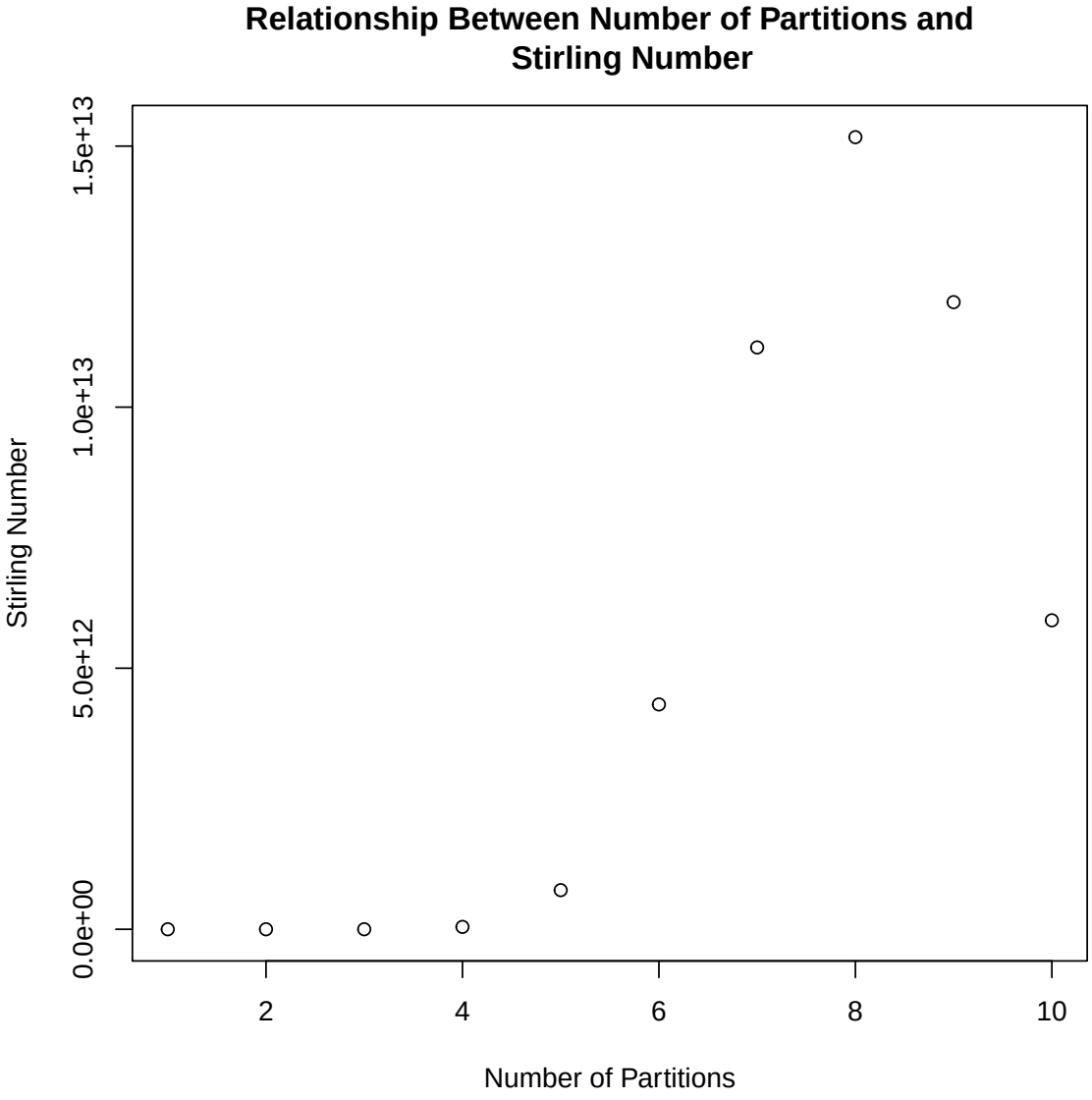
7. Stirling numbers of the second kind gives the number of ways to partition a set of n elements into k blocks, written $S(n, k)$ and is the sum

$$S(n, k) = \frac{1}{k!} \sum_{i=0}^{k} (-1)^{k-i} \binom{k}{i} i^n$$

Implement this function in `R` and create a plot for $n = 20$, $k = 1, 2, \ldots, 10$. You will turn in your `R` source code called `Stirling` and attach the visualization to your homework.
Answer -

```
Stirling <- function(n,k)
{
        x<-c(0:k)
        y<-array(0:0,dim=c(k))
        count<-1
        while(count<=k){
                x<-c(0:count)
                y[count]<-(1/factorial(count))*sum(((-1)^(count-x))*
                          ((factorial(count)/(factorial(x)*factorial(count-x))))*(x^n))
                count<-count+1
        }
        return(y)
}

plot(1:10,Stirling(20,10),main="Relationship Between Number of Partitions and
     Stirling Number",xlab="Number of Partitions",ylab="Stirling Number")
```



8. In no more than a paragraph, summarize the paper, "On the Surprising Behavior of Distance Metrics in High Dimensional Space."
   Answer - The paper talks about the general behaviour of the of the commonly used L_k norm $(x, y \in R^d, k \in Z, L_k(x, y) =)$ in high dimensional space. Recent results suggest that the L_k-norm may be relavant for $k = 1$ or 2 than values of $k \geq 3$. The authors have examined the behaviour of fractional distance metric for cases where k is allowed to be a fraction smaller tham 1. Higher norm parameters provide poorer contrast between the furthest and nearest neighbor. Among the distance metrics with integral norms, the manhattan distance metric is the method of choise for providing the bets contrast between the different points. Smaller the fraction, the greater the rat eof absolute divergence between the maximum and minimum value. The results of the paper are likely to have a powerful impact on the particular choise of distance metric which is used from problems such as clustering, categorization, and similarity search; all of which depend upon some notion of proximity.

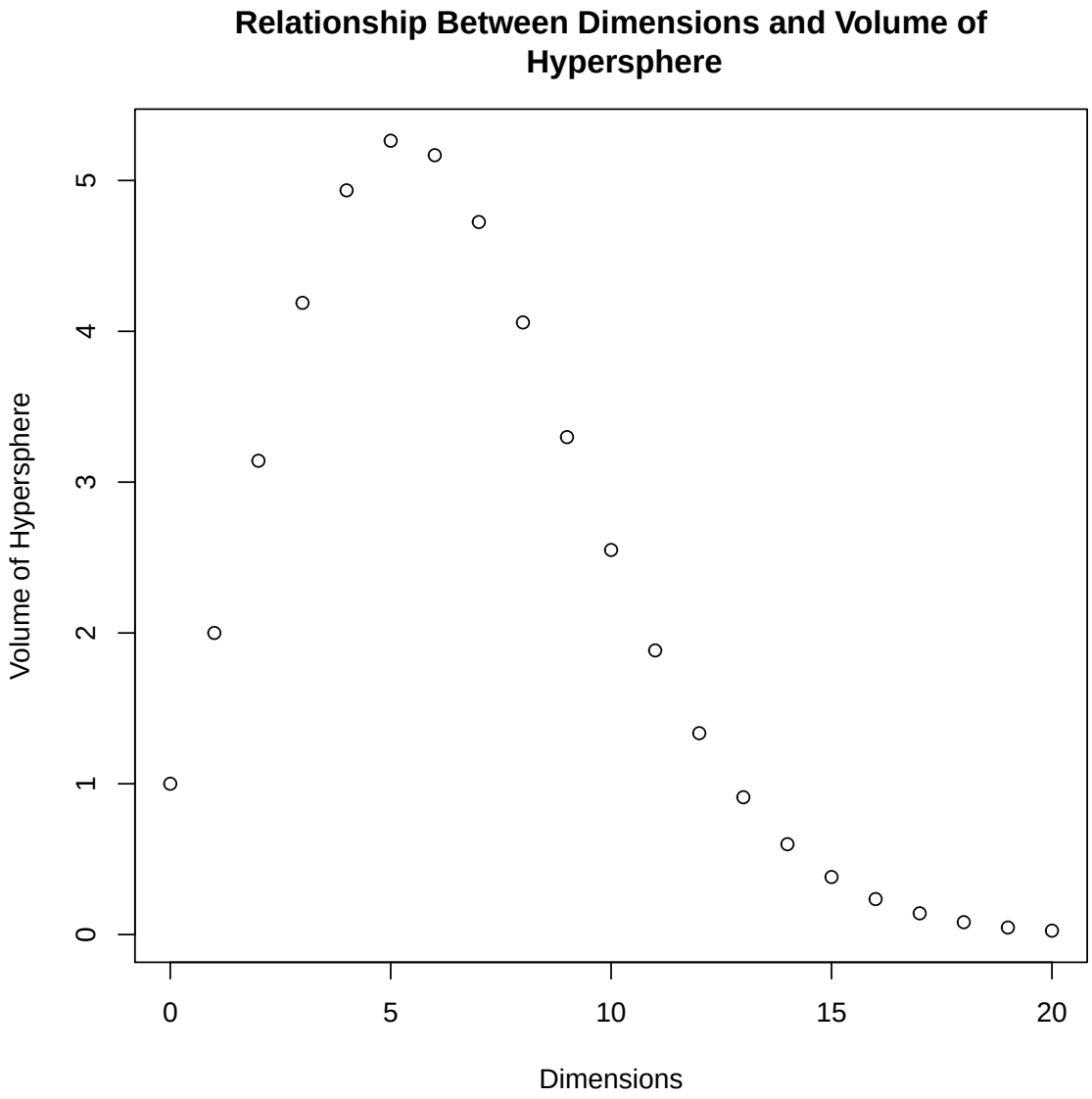
9. Curse of Dimensionality. A hypersphere describes the set of points within a fixed distance from a given point. We can write the volume of a hypersphere in n dimensions of unit radii as the recursion:

$$V_0 = 1 \quad (1)$$
$$V_1 = 2 \quad (2)$$
$$V_n = \frac{2\pi}{n} V_{n-2} \quad (3)$$

Using `R`, plot the volume of the hypersphere in $n = 0, 1, \ldots, 20$ dimensions of unit radii. Discuss the plot and how it relates to the paper in the previous question. The `R` code is called `CoD`.
Answer -

```
CoD <- function(x)
{
        i<-3
        y<-array(0:0,dim=c(x))
        y[1]<-1
        y[2]<-2
        while(i<=(x+1)){
                y[i]<-(2*pi*y[i-2])/(i-1)
                i<-i+1
        }
        return(y)
}

plot(0:20,CoD(20),main="Relationship Between Dimensions and Volume of
   Hypersphere",xlab="Dimensions",ylab="Volume of Hypersphere")
```



As we can see from the graph the volume of the hypersphere increases upto 5 dimensions and starts decresing after that. We can also observe that for dimensions higher than 13 the volume falls below 1.And for all higher dimenions the value of the volume converges to 0 where we cannot actually distinguish between the volume change with increase in dimension.

As discussed in the previous paper for smaller vales of dimensions we can observe contrast between values and can more accurately identify the nearest neighbours. Maximum contrast is observed between the range of 0 to 2. With contrast gently increasing again between 5 to 10. However for for larger values of dimensions the contrast being extremely poor in identifying the nearest neighbours. This hinders the process of creation of a model that reliably classifies the data objects into models.

Thus we can conclude that as the dimentionality increases we will be having the trouble of reduced classification accuracy and poorly defined clusters.