Homework 1
Computer Science
Fall 2016
B565

Anusha Ramamurthy
anuramam@iu.edu

October 16, 2016

All the work herein is solely mine.

Directions

Please follow the syllabus guidelines for your homework. I will provide the LaTeX of this document too. **The homework is due Sunday, September 4 by 5:00 p.m.**

Notation an 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

1. Define the following terms

   (a) Partition of a non-empty (finite) set X.
   A partition can be defined as a function that divides a non-empty set X into finite or smaller subsets such that if Π be the partition of X into blocks A_1,A_2,A_3,....,A_n

      i. $A_1 \cup A_2 \cup A_3 \cup .... \cup A_n = X$
      ii. $A_1 \cap A_2 \cap A_3 \cap .... \cap A_n = \emptyset$

(b) Distance metric d over X.

A distance metric is a function that gives the measure of similarity or dissimilarity between any given pair of elements in a set. A distance metric $d{:}X{\times}X = [\,0,\infty)$ and $\forall\ x{,}y \in X$ has to satisfy four conditions.

i. $d(\,x,y)\ \geq 0$

ii. $d(\,x,x)\ = 0$

iii. $d(\,x,y)\ = d(\,y,x)$

iv. $d(\,x,z)\ \leq d(\,x,y)\ + d(\,y,z)$

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

We have $\sim$, an equivalence relation over a non empty set X and let a,b $\in X$. An equivalence class of a is defined as

$[\,a]\ = \{x \in \mathrm{X}|xRa\}$

$\forall b \in [\,a]\ ,\ b = a$

$\forall[\,a] \cup [\,b] \cup ... \cup [\,z]\ = [\,X]$

$\forall[\,a] \cap [\,b] \cap ... \cap [\,z]\ = \emptyset$

Each equivalence class has a finite set of elements and is none empty. The union of equivalence classes gives the set of all elements in the equivalence space. Since each equivalence relation is unique, the intersection of equivalence classes is a null set. This is equivalent to the definition of partition. Hence for an a given equivalence relation over a non-empty (finite) set X, creates an*unique* partition Π of the set $X//$

(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

True. By definition of a partition

A. $A_1 \cup A_2 \cup A_3 \cup \ ....\ \cup A_n = X$, if A_1,A_2,...,A_n so on are the Π of X

Hence if $B \in \Pi$, all points in B must exist in X

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.

Lets assume the set $X = \{1, 2, 4, 5\}$. We define an equivalence relation $\sim_\Pi$ on X and a distance metric d, such that $\forall x, y \in X$, $d\ (\,x, y)\ = |x - y| = k$ where k is divisible by 3. Thus we have two distinct partitions B and B' where $B = \{1, 5\}$ and $B' = \{2, 4\}$.

$d(\,1, 2)\ = |1 - 2| = 1$

$\{4, 5\} \in X$, $d(\,5, 4)\ = |5 - 4| = 1$ which is divisible by 3

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.

As the above example X has 4 elements, but only 2 distinct partitions. Hence there can only be n or less than n distinct partitions.

iv. If $|\Pi| = n$, then there are n distinct $x_1, x_2, \ldots, x_n \in X$. TRUE OR FALSE

Answer: True.

Since a partition is defined on finite sets, If there are n partitions there are at least n distinct elements in X. If a partition over X, Π has only 1 element , then by the definition of partitions X has at least 1 distinct element.

v. $x \in B, y \in B'$ where $B, B' \in \Pi \wedge B \neq B'$ and $d(x, y) = k$. Then $(\forall x' \in X)\ x' \in B \wedge x \neq x' \to d(x, x') < k$. TRUE OR FALSE

Answer: False.

Lets assume the set $X = \{1, 2, 4, 5\}$, We define an equivalence relation $\sim_\Pi$ on X and a distance metric d, such that $\forall x, y \in X$ and $d\ (\,x, y)\ = |x - y| = k$ where k is divisible by 3. Thus we have two distinct partitions B and B' where $B = \{1, 5\}$ and $B' = \{2, 4\}$. Thus though elements 1 and 5 belong to the same partition,

$$\begin{aligned} d(1, 5) &= |1 - 5| \\ &= | - 4| \\ &= 4 \end{aligned}$$

$$\text{and } 4 \nleq k \text{ , where } k = 1$$

2. Let $X = \{1, 2, 3, 4, 5, 6\}$. 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?
Solution: $(1, 2) ,(1, 3) ,(1, 5) ,(3, 5) ,(4, 6) ,(7, 7)$ is the smallest equivalence relation. A data scientist is interested in the smallest because if we have the smallest set of equivalence relations we can identify the minimum number of features or points required for us to find similarities or dissimilarities in the data

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

Solution:

i	j	$d(i,j)$	v
0	3	3	3
	7	7	10
	8	8	18
	9	9	27
3	7	4	31
	8	5	36
	9	6	42
7	8	1	43
	9	2	45
8	9	1	46

From the table we can see that blocks of $\{0\}, \{3\}$ and $\{7, 8, 9\}$ have minimum Intrablock distance.

4. Show the results of `TOTIntraBlockDis`($\{\{1, 2\}, \{3\}, \{4, 10\}\}, d(x, y) = |x - y|$).
Solution:

i	j	$d(i,j)$	v
1	2	1	1

i	j	$d(i,j)$	v
3	3	0	0

i	j	$d(i,j)$	v
4	10	6	6

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.

```
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 $i = 1, length(i)$ do
    for $j = 1, length(j)$ do
        $v \leftarrow v + \mathtt{d}(i, j)$
    end for
```

```
  **end for**
  **return** $v$
```

`InterBlockDis` algorithm that takes a partition and distance function and returns the distance between blocks each pair of blocks.

i	j	$d(i,j)$	v
0	3	3	3
	7	7	7
	8	8	15
	9	9	24
3	7	4	4
	8	5	9
	9	6	15

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

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

   $(\forall\, x)\ d(x,x) = 0.$
   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 $d : R^2 \rightarrow R_{\geq 0}$ such that

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

   Solution: **Proof**

   $(\forall\, x)\ d(x,x) = 0.$
   Assume $a = 0$
   $d(a,a) = 0$ by definition.
   Assume $a > 0$ w.l.o.g.
   $d(a,a) = \left| \frac{a-a|}{\frac{1}{1+|a|}} \right| = 0$

$(\forall\, x, y)\ d(x,y) = d(y,x)$

Assume $a \leq b$. Then $\max\{a, b\} = b$.

Since $d(a,b) = \left|(b-a)/\frac{1}{1+|a|} * \frac{1}{1+|b|}\right|$ and $d(b,a) = |(b-a)/1+|b||$, 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.

$(b-a)/1+|b| + (c-b)/1+|c| \geq (c-a)/c$

$(b-a)/1+|b| \geq (b-a)/1+|c|$

$c \geq b$

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.

8. In no more than a paragraph, summarize the paper, "On the Surprising Behavior of Distance Metrics in High Dimensional Space."
   The paper talks about the metrics used for calculating the distance between points in higher dimensional spaces. The authors provide proof to show that as the dimension increases, commonly used metrics like L_1 norm, L_2 norm so on are not as effective as they are thought to be for identifying proximity of points in high dimensions. They propose a metric called fractional distance metric, where the k is allowed to be smaller than 1.This distance metric helps in understanding the behavior of points in high dimensions and to find the distance between them in much more meaningful ways.

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:

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

   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`.
   Solution : As we see, the volume of the hypersphere increases as the dimensions increase but as the dimension reaches beyond 5 we see that the volume of the hypersphere starts to fall. It appears that at higher dimensions the volume of the hypersphere approaches zero.
   The authors of the paper use the volume of the hypersphere as proof to show the behavior of distance metric's where k ¡ 1. They discuss this example to give the reasoning behind why Euclidean distance metrics does not perform nearly as well as the fractional distance metrics for high dimensional data.