Homework 1
Computer Science
Fall 2016
B565

Professor Dalkilic

August 27, 2016

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.
   (b) Distance metric d over X.
   (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

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

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

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

v.

vi. $x \in B, y \in B'$ 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

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?

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, ..., B_n}, distance d over X)
                                         ▷ X is a partition and B_i are the blocks.
OUTPUT R_{≥0} v
v ← 0
for i = 1, n do
    v ← v + IntraBlockDis(B_i, d)
end for
return v
```

```
INPUT IntraBlockDis(Set X = {x_1, x_2, ..., x_n}, distance d)
OUTPUT R_{≥0} v
v ← 0
for i = 1, n − 1 do
    for j = i + 1, n do
        v ← 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 $\texttt{IntraBlockDis}(X, d) = 20$. 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

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

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.

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

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

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

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

   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."

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`.