Algorithmic Methods of Data Mining - Assignment 1

Maksad Donayorov

September 28, 2018

1. **Problem:** *Prove or disprove: d_{cos} is a metric.*
   Metric definition suggest that d is metric if:
   1. $d(x, y) \geq 0 \qquad non - negativity$
   2. $d(x, y) = 0 \; x = y \qquad identity\ of\ indiscernibles$
   3. $d(x, y) = d(y, x) \qquad symmetry$
   4. $d(x, z) \leq d(x, y) + d(y, z) \qquad triangle\ inequality$

   Consequently, we have to either prove of disprove d_{cos} in order to say if it is metric or not. Let's consider a vector $\vec{x} = (1, 1)$ and $\vec{y} = (2, 2)$. In this case $\vec{x} \neq \vec{y}$ but our function d_{cos} computes 0:
   $d_{cos}(x, y) = 1 - \frac{1+1}{2} = 1 - 1 = 0$ and that leads to a conclusion that *identity of indiscernibles* does not hold for this function. Thus, it **is not a metric**.

2. **Problem:** *Let $d : X \times X \to R_+$ be a metric on X.*

   2.1. *Show that D(x, y) is also a metric on X*:
   To prove this, we have to look back to the definition of metric stated in the section 1 and validate each point.

   1. Non negativity:
   Since $d : X \times X \to R_+$ is a metric on X, $d(x, y)$ is > 0 by definition. Consequently, $\frac{d(x,y)}{d(x,y)+1} > 0$. So, we can conclud that **non negativity holds**.

   2. Identity of indiscernibles:
   Assuming that $d(x, y) = 0$ we will have a result: $\frac{d(x,y)}{d(x,y)+1} = \frac{0}{0+1} = 0$. So, **identity of indiscernibles holds**

   3. Symmetry
   Based on this definition $\frac{d(x,y)}{d(x,y)+1}$ should be equal to $\frac{d(y,x)}{d(y,x)+1}$.
   $\frac{d(x,y)}{d(x,y)+1} = \frac{d(y,x)}{d(y,x)+1}$
   $[d(y, x) + 1] \times d(x, y) = [d(x, y) + 1] \times d(y, x)$ which is equal to
   $d(x, y)d(y, x) + d(x, y) = d(x, y)d(y, x) + d(y, x)$.
   So we can say that **symmetry holds**

   4. Triangle inequality
   For simplicity let's change $d(,)$ to a notation so that this $d(x, z) \leq d(x, y) + d(y, z)$

inequality becomes easier to write. Assuming $J = d(x,z), K = (x,y), L = (y,z)$ we have to prove that $J \leq K + L$. Following similar steps that we did in 3, we can rewrite it to $\frac{J}{J+1} \leq \frac{K}{K+1} + \frac{L}{L+1}$. Solving this arithmetically (which I'm not going to write here) we can infer that:
$JKL + JK + JL + J \leq 2JKL + 2KL + JK + JL + K + L$ does not change the initial inequality and because of that the **triangle inequality holds**

With the final step we can conclude that $D(x,y)$ is also a metric and D is still a metric even if we change 1 to k, as 1 could be any number that is greater than 0.

2.2. *What is the role of k? How can the choice of k affect the new metric?*
k impacts on the dissimilarity of $D(x,y)$, the more it's value increases the more different $D(x,y)$ will be.

2.3. *What is a possible application of the new metric?*
It can measure the dissimilarity or similarity of $D(x,y)$.

3. **Problem:**
This can be solved with the same approach as the previous examples by following the 4 steps of identifying metric.

3.1. Non negativity:
Since $d(x,y)$ is distance function, by definition it cannot by less that 0. Following this logic we can infer that $D_h(A,B) = max(d(a,b))$ and that is $\geqslant$ than 0. Consequently, **non negativity True**

3.2. Identity of indiscernible:
Again, following the definition we can say that if $d(A,B) = 0$ then two sets are the same. This gives us two conditions: when $A = B$ and when $A \neq B$.
The firs case ($A = B$) is easy. We can say that if two sets are equal then the minimum distance is 0.
In case of $A \neq B$ we can infer that if two sets are not equal, then there is always one point in one of the set whose minimum distance to the other set cannot be 0.
Consequently, **identity of indiscernible holds**.

3.3. Symmetry:
Referring to the definition we write that $D_h(A,B) = D_h(B,A)$. To prove that they are hold symmetry property, we can think of taking their max such as:
$max\{d_h(A,B), d_h(B,A)\} = maxd_h(B,A), d_h(A,B)$
Since the order does not matter when thinking about max, we can say that **the symmetrical property holds**.

3.4. Triangle inequality:
To prove $d(A,B) \leqslant d(A,C) + d(C,B)$ let's again think about the max, such that:
$maxd(A,B), d(B,A) \leqslant maxd(A,C), d(C,A) + maxd(C,B), d(B,C)$
from that we can say that: $d(a,B) \leqslant d(a,C) + d(C,B)$
$d(A,B) = max_{a \in A} d(a,B) \leqslant d(A,C) + d(C,B)$
$d(A,B) = maxd(A,B), d(B,A) \leqslant d(A,C) + d(C,B)$
Thus, we can say that the **triangle inequality holds**.

4. **Problem:**
I don't know

5. **Problem:**
Let's assume we have 3 vectors $\vec{x} = [1, 2, 3]$, $\vec{y} = [6, 7, 8]$ and $\vec{z} = [0, 3, 4]$. For better visualization let's represent them as:

$$d \begin{Bmatrix} 0 & 6 & 0 \\ 2 & 7 & 3 \\ 1 & 8 & 4 \end{Bmatrix} \qquad \begin{matrix} diff & indices \\ 6 & i, j \\ 5 & i, j \\ 7 & i, j \end{matrix}$$

$$\text{---------}$$
$$n$$

As you might have noticed the above table conceptualizes the solution. All wee need to do is to loop through d and compute min and max for n elements.

```
for i in d:
    for j in n:
        min_v = ...
        max_v = ...
        diff = max_v-min_v
        vector_min_max_with_indices = ...
        if new_min < min_v or new_max > max_v:
            dif = recalculate_new_diff(...)
            vector_min_max_with_indices.append(...)
```

The complexity of this algorithm will be: $d \times n \times k_{constant} + d \times c_{constant} = O(nd)$.

6. **Problem:**
I don't know