Data Mining Home Work 1

Gurasees Singh[1]

[1]Boston University

September 23, 2017

Exercise 1

- **You are given two eggs and you have access to 100-story building. Eggs can be very hard or very fragile means it may break if dropped from the first floor, or may not even break if dropped from the 100th floor. Both eggs are identical. You need to figure out the highest floor of a 100-story building an egg can be dropped without breaking. You need to compute the minimum number of drops you need to make; you are allowed to break two eggs in the process.**

Assuming the egg is dropped form nth floor

In case egg breaks, we move on and try remaining (n-1) floors one by one from the lowest floor.

In the worst case we have to make n trials.

In cases the egg survives the drop, we move to (n-1) floors higher (as we have already used one attempt)

Therefore we drop the egg next from floor n+ (n-1)

we need to repeat the above cycle as we can have two possible outcomes either the egg breaks or it survives.

Therefore our drops can be represented as

n+(n-1)+(n-2).....

We would continue this process till the last possible floor i.e 100

n+ (n-1) + (n-2) + ... + 1 = 100
n(n+1)/2 = 100 (because, sum of first n natural number is given by n(n+1)/2)
n=13.7

We will select the next greater integer as number of trials would be an integer.

Hence, we begin dropping the egg from the 14th floor. In case it breaks we start from the 1st and try the remaining 13 floors one at a time. In case, the egg survives we jump up to the 27th floor.

We continue this till we find out solution.

Therefore, minimum number of drops you need to make is 14.

- **You are given a set V consisting of n integers. The task is to report all n products of the n distinct (n - 1)-cardinality subsets of V . Your algorithm should run in linear time and it should not use division.**

Consider v{a,b,c} of n =3 we need an algorithm to find products of carnality (n-1) i.e 2

Algorithm.

i. we create two temporary arrays

ii. copy all elements to the left of i to the first array and all elements to the right in the second array.

iii.we then find the Cartesian product of the left array and the right array. This would give us a product of (n-1 cardinality)

Exercise 2 (20 points):

Some years ago, greek video-club chain Seven had the following offer to their customers: every time a customer rented a DVD, he was given a random coupon with the title of the Academy awards (Oscars) winner movie written on it. The first person to gather the coupons with all the unique winner-movie titles won a 10-day vacation on a Caribbean island. If at that time, there were 75 unique such titles, and these titles were uniformly assigned to coupons, find the expected number of DVDs one had to rent in order to gather all of them.

I do not know the solution

Exercise 3 (20 points):

Assume two d-dimensional real vectors x and y. And denote by xi (yi) the value in the i-th coordinate of x (y). Prove or disprove the following statements:

A distance function that satisfies all the below properties is called a metric

- non-negativity, $d(x,y) \geq 0$
- isolation, $d(x,y) = 0 \Leftrightarrow x = y$
- symmetry, $d(x,y) = d(y,x)$
- triangle inequality ,$d(x,z) \leq d(x,y) + d(y,z)$

Proof:

1. Distance function $= L_1(x,y) = \sum_{i=1}^{d} |x_i - y_i|$ is a metric. (5 points)

i. non-negativity:-

Consider $D(x,y) \geq 0 \; \forall \; x, y \in R$ ——————— equation 1

$\Rightarrow \; D(x,y) = \sum_{i=1}^{d} |x_i - y_i|$

$\Rightarrow \; D(-x,-y) = \sum_{i=1}^{d} |-x_i + y_i|$

$\Rightarrow \; D(-x,-y) = \sum_{i=1}^{d} |-1|\,|x_i - y_i|$ (because $|x-y| = \; |-1| \cdot |y-x|$)

$\Rightarrow \; D(-x,-y) = \sum_{i=1}^{d} |x_i - y_i|$

Using Equation 1 we can say that $D(-x,-y) \geq 0$

Alternate Proof: The formula consists of a sum of a modulus (|x-y|)which would always be positive for all values of x and y because modulus of a number is always positive.

ii. Isolation

$d(x,y) = 0 \; \Leftrightarrow \; x = y$

Given $D(x,y) = \sum_{i=1}^{d} |x_i - y_i|$

Considering x=y

$\Rightarrow \ D(x,y) = \sum_{i=1}^{d} |x_i - x_i|$ (because x=y)

$\Rightarrow \ D(x,y) = \sum_{i=1}^{d} |0| = 0$

iii. Symmetry

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

$D(x,y) = \sum_{i=1}^{d} |x_i - y_i|$

$D(y,x) = \sum_{i=1}^{d} |y_i - x_i|$

if $x - y \geq 0$ then $y - x \leq 0$ | if $x - y \leq 0$ then $y - x \geq 0$

$\Rightarrow |x-y| = x - y$ and $\Rightarrow |y-x| = x - y$ | $\Rightarrow |x-y| = y - x$ and $\Rightarrow |y-x| = y - x$

$\Rightarrow \sum_{i=1}^{d} |x_i - y_i| = \sum_{i=1}^{d} |y_i - x_i|$ | $\Rightarrow \sum_{i=1}^{d} |x_i - y_i| = \sum_{i=1}^{d} |y_i - x_i|$

Therefore,

$D(x,y) = D(y,x)$

iv. triangle inequality ,

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

$|x-z| \leq |x-y| + |y-z|$

Using triangle inequality for absolute values

|x-y|=|x-z+z-y|=|(x-z)+(z-y)|[?]|x-z|+|z-y|=|x-z|+|y-z| {Reference: Triangle inequality Wikipedia}

Thus $d(x,z) \leq d(x,y) + d(y,z)$

Since all 4 properties are satisfied by the distance function it is a metric.

2. . Distance function $= L_2(x,y) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$ is a metric. (5 points)

i. non-negativity:-

$D(x,y) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$

Since $(x-y)^2 \geq 0 \ \forall \ x, y \ \epsilon \ R$

$\Rightarrow \sum_{i=1}^{d} (x_i - y_i)^2 \geq 0$ (summation of all positive numbers)

$\Rightarrow \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2} \geq 0$

ii. Isolation(definiteness)

$d(x,y) = 0 \ \Leftrightarrow \ x = y$

if x=y $\Rightarrow (x-y)^2 = (x-x)^2 = 0$

$\Rightarrow \sqrt{\sum_{i=1}^{d} (x_{i-}y_i)^2} = 0 \;\; if \;\; (x = y)$

iv. Triangle Inequality

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

To Prove: $\sqrt{(x-z)^2} \leq \sqrt{(x-y)^2} + \sqrt{(y-z)^2}$

$\sqrt{(x-z)^2} = |x-z|$ and $\sqrt{(x-y)^2} + \sqrt{(y-z)^2} = |x-y| + |y-z|$

From Triangle inequality (Also proven in previews question) We know that

$|x-z| \leq |x-y| + |y-z|$

Since all 4 properties are satisfied the distance function is a metric

3. Distance function $= L_3(x,y) = \sum_{i=1}^{d} (x_{i-}y_i)^2$ is a metric. (5 points)

i. non-negativity:-

$D(x,y) = \sum_{i=1}^{d} (x_{i-}y_i)^2$

Since $(x-y)^2 \geq 0 \; \forall \; x, y \; \epsilon \; R$

therefore $D(x,y) \geq 0$ (summation of all positive terms)

ii. Isolation(definiteness)

$d(x,y) = 0 \;\Leftrightarrow\; x = y$

if x=y $\Rightarrow (x-y)^2 = (x-x)^2 = 0$

d(x,x)=0

Hence $d(x,y) = 0 \;\Leftrightarrow\; x = y$

iii. Symmetry

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

$Given \; D(x,y) = \sum_{i=1}^{d} (x_i - y_i)^2$

Since $(x-y)^2 = (y-x)^2 \; \forall \; x, y \; \epsilon \; R$

Therefore $\sum_{i=1}^{d} (x_i - y_i)^2 = \sum_{i=1}^{d} (y_i - x_i)^2$

Hence $d(x,y) = d(y,x)$

iv. Triangle Inequality

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

To Prove: $(x-z)^2 \leq (x-y)^2 + (y-z)^2$

consider x =4 y=2 and z =-2

$LHS = (4+2)^2 = 36$

$RHS = (4-2)^2 + (2+2)^2 = 4 + 16 = 20$

Since LHS is not less than or equal to RHS triangle inequality fails.

Therefore $D(x,y) = \sum_{i=1}^{d} (x_{i-}y_i)^2$ **is not a metric**

Exercise 4 (20 points): Consider a set of n data points x1, . . . , xn.

***** *Assume a random number generator R() that generates values in the interval (0, 1]. Let distance function dR between two points xi and xj with xi 6= xj be dR(xi , xj) = R(). Also dR(xi , xi) = 0 for every xi . Prove or disprove that dR is a metric. (10 points)*

To prove a distance function to be a metric we need to prove the 4 properties discussed in previous question.

i. Symmetry:

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

Since $d_R(x_i, x_j) = R()$ (*Here Function R returns a random number between* 0 *and* 1)

$d_R(x_i, x_j)$ *would not be equal to* $d_R(x_j, x_i)$ because R() would be different each time.

Since the distance function does not satisfy the property of Symmetry it is not a metric

- ***Construct a graph G = (V, E), where a point xi is represented by node vi [?] V . For every pair of nodes vi , vj , there exists an undirected edge in G with weight wij = dR(xi , xj). We define the distance function between two nodes vi and vj , denoted by dG(vi , vj), be the weight of the shortest path between vi and vj in graph G. Prove or disprove that dG is a metric. (15 points)***

To prove a distance function to be a metric we need to prove the 4 properties discussed in previous question.

i. non-negativity:- $d_G(x_i,x_j)$ gives the shortest path between two nodes x_i *and* x_j .Therefore it would be always positive because shortest path can never be negative.

ii.Isolation $d(x_i, y_j) = 0 \Leftrightarrow x_i = y_j$

if $x_i = x_j$, we would get the shortest distance between a node and itself which will always be zero .

iii. Symmetry

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

The shortest distance between two nodes of an un-directed graph would always be same irrespective of the order of the terms. Therefore $d(x, y) = d(y, x)$

iv. Triangle Inequality

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

$d(x, z)$ gives the shortest path between x and z and therefore it will always be shorter or equal to their path via another node y.

Therefore $d(x, z) \leq d(x, y) + d(y, z)$

Since all 4 properties have been satisfied the distance function is a metric

Excercise 5

i.Prove or disprove that the edit distance function as defined above is a metric.

To prove a distance function to be a metric we need to prove the 4 properties discussed in previous question(q2)

i. non-negativity:- .

Edit distance gives the costof delete,insert or substitute operations which depends on the number of operations required to transform string x into string y and therefore it will always be positive.

ii. Isolation:

$d(x, y) = 0 \Leftrightarrow x = y$

if x =y the number of operations required to convert x to y would be zero and hence the edit distance would be zero.

iii. Symmetry

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

Since d(x,y)gives the cost of operations to convert x to y it would be equal to cost of operations to convert y to x as both would involve the same operations.

eg.

x=apple and y=bapple

to convert x to y we need to substitute a with b and to convert y to x we need to substitute b with a. Cost of both the operations would be the same

iv. Triangle inequality

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

As edit distance is the cost of conversion of strings

the cost to convert x into z would be less than converting x into and y into z

Since all 4 properties have been satisfied the distance function is a metric

2. (10 points:) Find two instantiations of the edit-distance function that are metrics. An instantiation of the edit distance function is defined by a specific way of allocating costs to operations such as deletions, insertions and substitutions.

I do not know the solution