10/1/2019

Non-metric Similarity function:

Similarity functions which do not obey either the triangular inequality (or) symmetry come under this category. usually these similarity functions are useful for images (or) string data. They are robust to outliers (or) to extremely noisy data.

* The squared Euclidean distance is itself an example of a non-metric, but it gives the same ranking as the Euclidean distance which is a metric.

* one non-metric similarity function is the K-median distance between two vectors.

* If $X = (x_1, x_2 \cdots x_n)$ and $Y = (y_1, y_2 \cdots y_n)$ then

$$d(X,Y) = K\text{-median}\{|x_1 - y_1|, \cdots |x_n - y_n|\}$$

where the K-median operator returns the Kth val of the ordered difference vector.

Example:

If $X = (50, 3, 100, 29, 62, 140)$ and $Y = (55, 15$ $50, 70, 170)$ then

Difference vector = $\{5, 12, 20, 21, 8, 30\}$

$d(x,y)$ = K-median $\{5, 8, 12, 20, 21, 30\}$

If $K=3$ then $d(R,Y) = 12$

which property does not satisfy about this K-median distance?

Edit Distance:

Edit distance measures the distance between two strings. It is also called Levenshtein distance. The edit distance between two strings S_1 and S_2 is defined as the minimum number of point mutations required to change S_1 to S_2. A point mutation involves any one of the following operations.

1. Changing a letter
2. Inserting a letter
3. Deleting a letter.

Ex 1:

If S = TRAIN and t = BRAIN, then edit distance = 1, this requires a change of just one letter.

(2)

ii) If s = TRAIN and t = CRANE ; then edit distance = 3.

i) TRAIN and CRANE +(1)

ii) TRA and CRA + (2) because of I and N

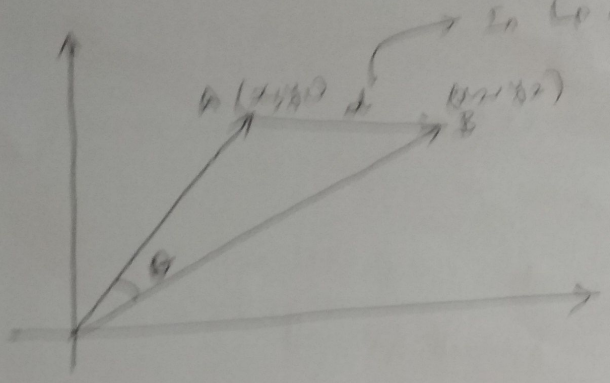
iii) T and C + (3) because of T and C.

So, we get the edit distance to be (3).

Is Edit distance metric? , yes

Prove that Edit distance metric

Non-metric similarity function



i) Cosine distance

1) Cosine looks at the angle between vectors (not taking into magnitude)

2) Euclidean distance is similar to using a ruler to actually measure the distance.

Example:

$a = [1, 2, 3]$

$b = [4, -5, 6]$

$$\frac{a \cdot b}{\|a\| \|b\|} = \frac{1 \cdot 4 + 2 \cdot -5 + 3 \cdot 6}{\sqrt{1^2+2^2+3^2}\sqrt{4^2+5^2+6^2}} = \frac{12}{\sqrt{14} \cdot \sqrt{77}}$$

Cosine Similarity in Data Mining:

Cosine Similarity is a measure to find the similarity btw two files | documents

file 1 = (0, 3, 0, 0, 2, 0, 0, 2, 0, 5)

file 2 = (1, 2, 0, 0, 1, 1, 0, 1, 0, 3)

file1 · file2 = 0×1 + 3×2 · · · · 5×3

= 25

$\|d_1\| = \sqrt{42} = 6.481$

$\|d_2\| = \sqrt{17} = 4.12$

$$\cos(d_1, d_2) = \frac{\text{file1} \cdot \text{file2}}{\|\text{file1}\| \, \|\text{file2}\|}$$

$$\cos(d_1, d_2) = \frac{25}{6.481 \times 4.12} = 0.94$$

$$D(d_1, d_2) = 1 - 0.94 = 0.06$$

def: The cosine of two non-zero vectors can be derived by using the Euclidean dot product formula:

$$A \cdot B = \|A\| \, \|B\| \cos\theta$$

(angular similarity) $\Rightarrow \cos(\theta) = \dfrac{A \cdot B}{\|A\| \cdot \|B\|} = \dfrac{\sum_{i=1}^{d} A_i B_i}{\sqrt{\sum_{i=1}^{d} A_i^2} \cdot \sqrt{\sum_{i=1}^{d} B_i^2}}$

$$S(A, B) = \frac{A \cdot B}{\|A\| \, \|B\|}$$

$$D(A, B) = 1 - S(A, B)$$

This $D(A, B)$ does not satisfy the triangular inequality. It is not a metric, However, it is symmetric, because $\cos(\theta) = \cos(-\theta)$.

...n There is a way to convert into a metric if the vectors are always positive:

$$\text{Angular distance} = \frac{2 \cdot \cos^{-1}(\text{Cosine Similarity})}{\pi}$$

$$D(\hat{A}, \hat{B}) = \frac{2 \cdot \cos^{-1} S(A, B)}{\pi}$$

$$S(A, B) = 1 - D(\hat{A}, \hat{B})$$

Angular Similarity = 1 − angular distance

Example for Triangular inequality:

If X, Y and Z are three vectors in a 2-d space such that the angle between X & Y is 45 and that between Y and Z is 45, then:

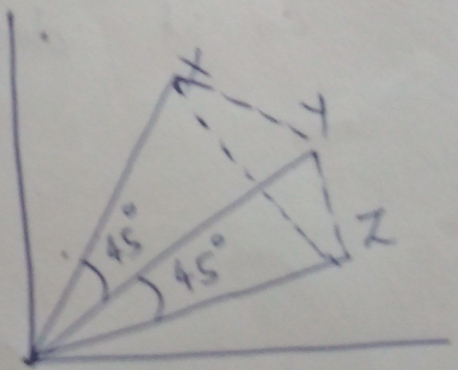
$$d(X, Z) = 1 - S(X, Z)$$

$$= 1 - \cos(45 + 45)$$

$$= 1 - \cos(90)$$

$$= 1 - 0$$

$$= 1 \quad \text{——— Eqn (1)}$$



whereas

$$d(x,y) + d(y,z) = (1 - \cos 45) + (1 - \cos 45)$$

$$= 1 - \left(\frac{\sqrt{2}}{2}\right) + 1 - \frac{\sqrt{2}}{2}$$

$$= 2 - \frac{2\sqrt{2}}{2}$$

$$= 2 - \sqrt{2} \quad \text{——— Eqn ②}$$

$\theta \rightarrow$	0	30	45	60	90
$\cos\theta$ ↓	1	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$	0

from Eqn① & Eqn ②;

it is proved that cosine distance, is not a metric.

$$1 \neq 2 - \sqrt{2}$$

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

Hence, cosine distance is not a metric

(*) as it does not satisfy triangular inequality

16b) KL - distance: Kullback leibler (also called relative entropy)

→ It is a measure of how one probability distribution is different from refrence probability distribution.

→ It is a asymmetric measure and thus does not qualify as a statistical metric.

→ A non-metric which is non-symmetric is the kullback leibler distance.

→ It is the natural distance function from a "true" probability distribution P, to a target probability distribution q.

→ For a discrete probability distribution I is $P = \{P_1, P_2, \ldots P_n\}$ and $q = \{q_1, q_2 \ldots q_n\}$, then the KL distance is defined as

$$KL(P,q) = \sum P_i \log_2 \left(\frac{P_i}{q_i} \right)$$

→ for continous P.d, The sum is replaced by an integral.

Example

K	0	1	2	
Distribution P(K)	0.36	0.48	0.16	Binomial distribution
Distribution Q(K)	0.333	0.333	0.333	Uniform distribution

$$D_{KL}(P,Q) = 0.36 \ln\left(\frac{0.36}{0.333}\right) + 0.48 \ln\left(\frac{0.48}{0.333}\right) + 0.16 \ln\left(\frac{0.16}{0.332}\right)$$

$$= 0.0852 \text{ nat} \quad \text{(natural unit of information (or entropy))}$$

$$D_{KL}(Q,P) = 0.333 \ln\left(\frac{0.333}{0.36}\right) + 0.333 \ln\left(\frac{0.333}{0.48}\right) + 0.33 \ln\left(\frac{0.333}{0.16}\right)$$

$$= 0.0974$$

Also, It does not obey triangular inequality.

How do you check?

Hence, This distance is not a metric.

$$(6.36 \times 0.033) + (0.48 \times 0.158) + (0.16 \times -0.318)$$

$$0.01188 + 0.07584 + (-0.814) = -0.6388$$