Non-metric Similarity Function

- Similarity functions which do not obey either the triangle 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 itself is an example of a non-metric, but it gives the same ranking as the Euclidean

Non Metric Similarity Function

- K-median
- Cosine Distance
- KL-distance
- Bhattacharya Distance

1. K-median distance

- One non-metric similarity function is the K-median distance between two vectors.
- If $X = (x_1, x_2, x_3 \ldots, x_n)$ and
  $Y = (y_1, y_2, y_3 \ldots, y_n)$

  then

  $$D(x, y) = K - median\{|x_1 - y_1|, \ldots, |x_n$$

  where the $K - median$ operator returns the k^{th} value of the ordered difference vector.

Example

- If $X = (50, 3, 100, 29, 62, 140)$ and

  $Y = (55, 15, 80, 50, 70, 170)$

  then

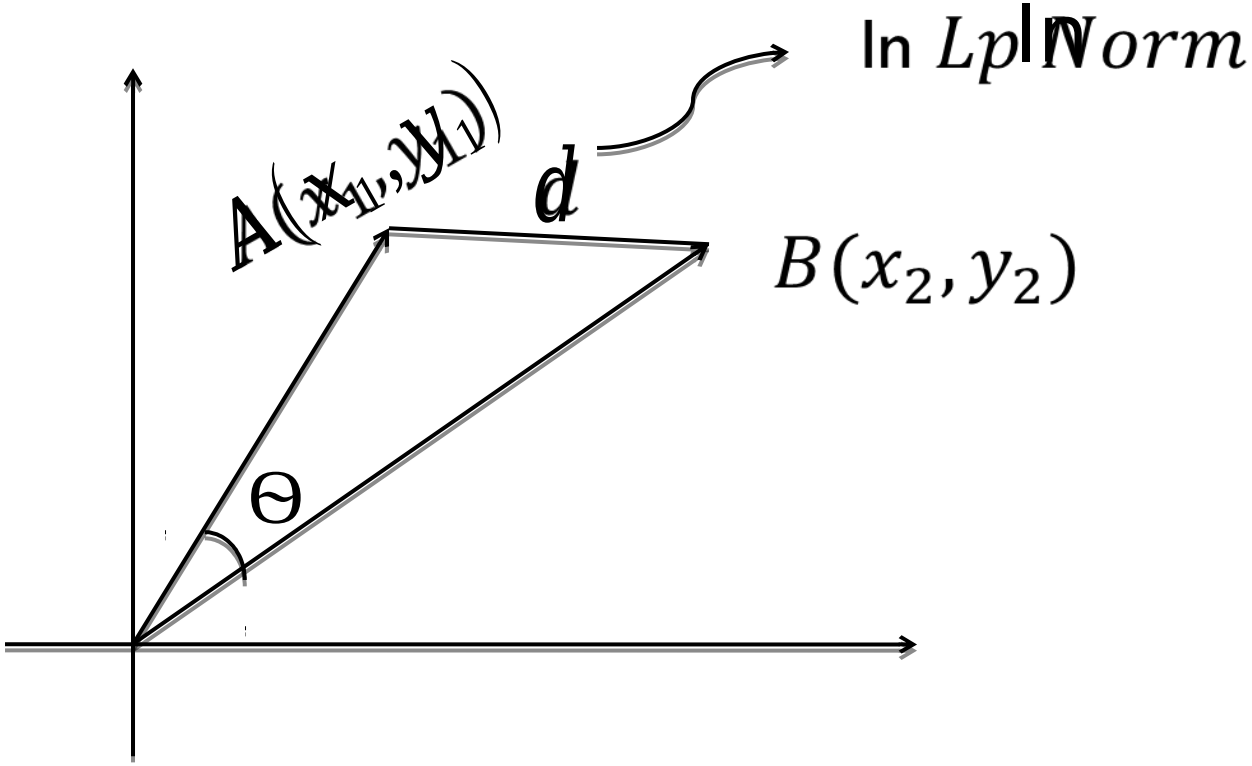
  $$Difference\ vector = \{5, 12, 20, 21, 8, 30\}$$

  $$D(X, Y) = K - median\{5, 8, 12, 20, 21, 30\}$$

- If $K = 3$ then $D(X, Y) = 12$

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

2. Cosine distance



- Cosine considers the angle between vectors (not taking magnitude into account).

- Euclidean distance is similar to using a ruler to actually measure the distance.
- E.g.

$$a = [1,2,3]$$

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

$$\frac{a.b}{||a||\,||b||} = \frac{1.4+2.-5+3.6}{\sqrt{1^2+2^2+3^2}\sqrt{4^2+5^2+6^2}} = \frac{12}{\sqrt{14}\sqrt{77}}$$

Cosine similarity in data mining

- Cosine similarity is a measure to find the similarity between two files/documents.

$$file_1 = (0, 3, 0, 0, 2, 0, 0, 2, 0, 5)$$
$$file_2 = (1, 2, 0, 0, 1, 1, 0, 1, 0, 3)$$
$$file_1 . file_2 = 0 \times 1 + 3 \times 2 + \cdots .. + 5 \times 3$$
$$= 25$$
$$||d_1|| = \sqrt{42} = 6.481$$
$$||d_2|| = \sqrt{17} = 4.12$$

$$\cos(d_1, d_2) = \frac{file_1 . file_2}{||file_1|| ||file_2||}$$

$$\cos(d_1, d_2) = \frac{25}{6.481x4.12}$$

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

$$= 0.06$$

Cosine Distance: Definition

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

$$A.B = ||A||||B|| \cos\theta$$

Angular Similarity

$$\cos\theta = \frac{A.B}{||A||\,||B||}$$

$$= \frac{\sum_{i=1}^{d} A_i B_i}{\sqrt{\sum_{i=1}^{d} A_i^2}\sqrt{\sum_{i=1}^{d} B_i^2}}$$

- Angular Similarity

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

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

- This $D(A,B)$ does not satisfy the *triangular inequality* and hence it is not a metric
- However, it is symmetric, because

$$\cos\theta = \cos(-\theta)$$

- There is a way to convert into a metric.
- If the vectors are always positive:

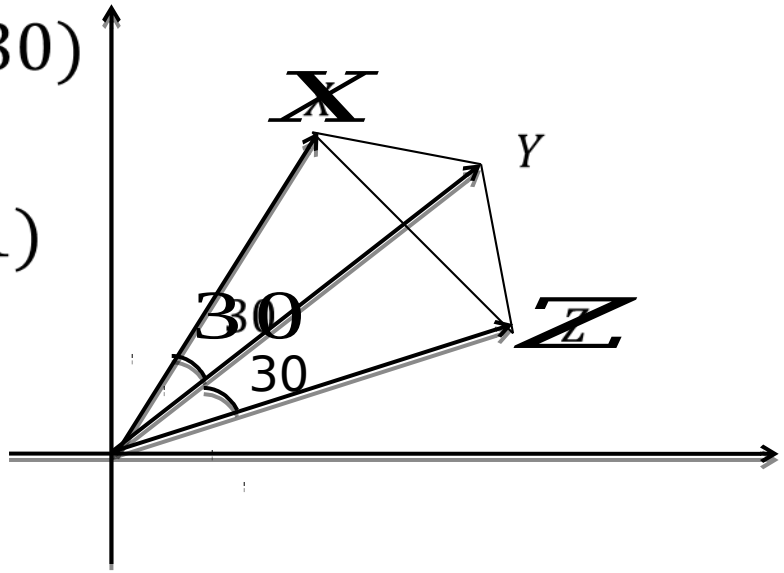
$$Angular\ Distance = \frac{2\cos^{-1}(cosine\ similarity)}{\pi}$$

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

Example for Triangular Inequality

- If X, Y and Z are the three vectors in a $2-d$ space such that the angle between $X\ and\ Y$ is $30°$ and that between $Y\ and\ Z$ is $30°$, then :

$$D(X,Z) = 1 - S(X,Z)$$
$$= 1 - \cos(30+30)$$
$$= 1 - \cos 60$$
$$= \frac{1}{2} \text{ --------} Eqn(1)$$



- $D(X,Y) + D(Y,Z) = (1 - \cos 30) + (1 - \cos 30)$

$$= \left(1 - \frac{\sqrt{3}}{2}\right) + \left(1 - \frac{\sqrt{3}}{2}\right)$$

- From and ; $= 2\left(1 - \frac{\sqrt{3}}{2}\right)$------$Eqn(2)$
- From $Eqn1$ and $Eqn2$;

$$\frac{1}{2} \nleq 2 - \sqrt{3}$$

$$D(X,Z) \nleq DX,Y) + D(Y,Z)$$

- Hence, cosine distance is not a metric, as it does not satisfy triangular inequality.

3. KL-Distance

- Kullback leibler distance is also called relative entropy.
- It is a measure of how a probability distribution is different from reference probability distribution.
- It is an asymmetric measure and 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$(P.D)$,

  if $p = \{p1, p2, \ldots\ldots, pn\}$ and

  $q = \{q1, q2, \ldots\ldots, qn\}$,

  then the KL distance is defined as:

$$D_{KL}(p, q) = \sum p_i \log_2 \frac{p_i}{q_i}$$

- For continuous $P.D$, the sum is replaced by an integral.

Example

X	0	1	2	X	0	1	2	X	0	1	2	X	0	1	2
Distribution $P(X)$	0.36	0.48	0.16	Distribution $P(X)$	0.36	0.48	0.16	Distribution $P(X)$	0.36	0.48	0.16	Distribution $P(X)$	0.36	0.48	0.16
Distribution $Q(X)$	0.333	0.333	0.333	Distribution $Q(X)$	0.333	0.333	0.333	Distribution $Q(X)$	0.333	0.333	0.333	Distribution $Q(X)$	0.333	0.333	0.333
X	0	1	2	X	0	1	2	X	0	1	2	X	0	1	2
Distribution $P(X)$	0.36	0.48	0.16	Distribution $P(X)$	0.36	0.48	0.16	Distribution $P(X)$	0.36	0.48	0.16	Distribution $P(X)$	0.36	0.48	0.16
Distribution $Q(X)$	0.333	0.333	0.333	Distribution $Q(X)$	0.333	0.333	0.333	Distribution $Q(X)$	0.333	0.333	0.333	Distribution $Q(X)$	0.333	0.333	0.333
X	0	1	2	X	0	1	2	X	0	1	2	X	0	1	2
Distribution $P(X)$	0.36	0.48	0.16	Distribution $P(X)$	0.36	0.48	0.16	Distribution $P(X)$	0.36	0.48	0.16	Distribution $P(X)$	0.36	0.48	0.16
Distribution $Q(X)$	0.333	0.333	0.333	Distribution $Q(X)$	0.333	0.333	0.333	Distribution $Q(X)$	0.333	0.333	0.333	Distribution $Q(X)$	0.333	0.333	0.333

The distribution $P(X)$ is a binomial distribution and $Q(X)$ is a uniform distribution

- $D(P,Q) = 0.36 \ln\left(\frac{0.36}{0.333}\right) + 0.48 \ln\left(\frac{0.48}{0.333}\right) + 0.16 \ln\left(\frac{0.16}{0.333}\right)$

  $= 0.0852$

- $D(Q,P) = 0.333 \ln\left(\frac{0.333}{0.36}\right) + 0.48 \ln\left(\frac{0.333}{0.48}\right) + 0.16 \ln\left(\frac{0.333}{0.48}\right)$

  $= 0.0974$

- $D(P,Q) \neq D(Q,P)$ hence KL distance is not a metric.

4. Bhattacharya Distance

- One non-metric similarity function is the K-median distance between two vectors.
- If $X = (x_1, x_2, x_3 \ldots, x_n)$ and
  $Y = (y_1, y_2, y_3 \ldots, y_n)$

  then

  $$D(x, y) = K - median\{|x_1 - y_1|, \ldots, |x_n$$

  where the $K - median$ operator returns the k^{th} value of the ordered difference vector.

- Now, find the dot product of x' and y'

$$x'.\, y' = |x'||y'| \cos\theta \text{ ---- } -Eqn\ 2$$

- Substituting the values from $Eqn\ 1$ to $Eqn\ 2$,

$$\sqrt{x_1 y_1} + \sqrt{x_2 y_2} + \ldots\ldots \sqrt{x_n y_n} =$$

$$\sqrt{x_1 + x_2 + \cdots x_n}\ \sqrt{y_1 + y_2 + \cdots y_n} cos\theta$$

$$\text{---- } -Eqn\ 3$$

- As x and y denotes probability distributions;

$$x_1 + x_2 + \cdots x_n = 1\ and$$

$$y_1 + y_2 + \cdots y_n = 1$$

- From the above condition,

  $Eqn\ 3$ becomes,

$$\sqrt{x_1 y_1} + \sqrt{x_2 y_2} + \ldots\ldots \sqrt{x_n y_n} = 1 \cos\theta$$

$$\therefore \cos\theta = \sum_{i=1}^{n} \sqrt{x_i y_i}$$

- Bhattacharya Coefficient

$$B(x, y) = \sum_{i=1}^{n} \sqrt{x_i y_i}$$

- Bhattacharya Distance

$$D(x, y) = -\ln B(x, y)$$