Data Mining I - Homework 1 : Distance Functions on vectors

Exercise 2a

(i) When $n = 1$,

let $x = 3$, $y = 6$ and $z = 5$

$$d(x,z) = (3-5)^2 = 4$$
$$d(y,z) = (6-5)^2 = 1$$
$$d(x,y) = (3-6)^2 = 9$$

$\because d(x,y) > d(x,z) + d(y,z)$,

$d(x,y)$ violates the condition of the triangle inequality, hence is NOT a metric.

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

When $n = 1$,

if $x = 0$ and $y \neq 0$,

$$d(x,y) = 0 \cdot y \cdot (-y)^2$$
$$= 0$$

For any value of y_i, if $x_i = 0$, $d(x,y) = 0$.
This violates the condition that

$d(x,y) = 0$ only if $x = y$.

$\therefore d(x,y)$ is NOT a metric.

Exercise 2a (cont'd)

(iii) $d(x,y) = \sum_{i=1}^{n} w_i |x_i - y_i|, \ w_i > 0 \ \forall i$

Proof ① :

$\because w_i > 0 \ \forall i$ and $|x_i - y_i| \geq 0$

$\therefore d(x,y) \geq 0$

$\therefore$ The condition that $d(x,y)$ is non-negative is satisfied

Proof ② :

$\because w_i > 0 \ \forall i$

For all values of i,

$d(x,y) = 0$ only if $|x_i - y_i| = 0$

where $x_i = y_i$

Proof ③ :

$\because |x_i - y_i| = |y_i - x_i|$

$\therefore$ For all values of i,

$d(x_i, y_i) = d(y_i, x_i)$

$\therefore d(x,y)$ is symmetrical.

Proof ④ :

When $n = 1$

let $w_i = 1$ and $z_i = 0$

$\sum_{i=1}^{n} w_i |x_i - z_i| + \sum_{i=1}^{n} w_i |z_i - y_i|$

$= |x_i| + |y_i|$

$\because |x_i| + |y_i| \geq |x_i - y_i|$

$\therefore d(x,y)$ satisfies the triangle inequality.

$\Rightarrow$ function (iii) is a metric.

Exercise 2a (cont'd)

(iv) When $n=2$,

let $x(0.2, 0.8)$, $y(0.7, 0.3)$

$$d(x, y) = 0.2 \log\left(\frac{0.2}{0.7}\right) + 0.8 \log\left(\frac{0.8}{0.3}\right)$$
$$= 0.23196$$

$$d(y, x) = 0.7 \log\left(\frac{0.7}{0.2}\right) + 0.3 \log\left(\frac{0.3}{0.8}\right)$$
$$= 0.25305$$
$$\neq d(x, y)$$

$\therefore$ $d(x, y)$ violates the condition that $d(x, y)$ and $d(y, x)$ have to be symmetrical, hence it's not a metric.

v) $d(x,y) = \frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{x_i + y_i}$

let us have vectors $x(0.1, 0.9)$, $y(0.7, 0.3)$, $z(0.6, 0.4)$

$$d(x,y) = \frac{(x_1 - y_1)^2}{x_1 + y_1} + \frac{(x_2 - y_2)^2}{x_2 + y_2} = 0.45 + 0.3 = 0.75$$

$$d(x,z) = \frac{(x_1 - z_1)^2}{x_1 + z_1} + \frac{(x_2 - z_2)^2}{x_2 + z_2} = 0.357 + 0.192 = 0.549$$

$$d(y,z) = \frac{(y_1 - z_1)^2}{y_1 + z_1} + \frac{(y_2 - z_2)^2}{y_2 + z_2} = 0.00769 + 0.0142 = 0.0218$$

Exercise 2a (cont'd)

(v) cont'd

$\because \quad 0.75 > 0.549 + 0.0218$

$d(x,y) > d(x,z) + d(y,z)$

$\therefore$ $d(x,y)$ violates the triangular inequality, hence it is NOT a metric.

(vi) When $n = 1$,

$$d(x,y) = \frac{2}{\pi} \arccos\left(\frac{x_i y_i}{x_i y_i}\right)$$

$$= \frac{2}{\pi} \arccos(1)$$

$\because \arccos(1) = 0$

$\therefore$ For $\forall x, \forall y$, $d(x,y) = 0$

$\therefore$ $d(x,y) = 0$ not only when $x = y$, hence it violates the condition where $d(x,y) = 0$ only when $x = y$.

Therefore $d(x,y)$ is NOT a metric.

Exercise 2a (cont'd)

vii) $$d(x,y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$$

According to the conditions of $d(x,y)$:

Proof ① : $\because$ The only two values that $d(x,y)$ can equate to is 0 and 1.

$\therefore$ $d(x,y) \geqslant 0$

Proof ② : $\because$ Only when $x = y$, $d(x,y) = 0$, otherwise $d(x,y) = 1$

$\therefore$ The condition than $d(x,y) = 0$ only when $x = y$ is satisfied.

Proof ③ : $\because$ $d(x,y)$ only assess the equality / inequality of x and y, irrespective of order

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

$\therefore$ The function $d(x,y)$ is symmetric.

Exercise 2a (cont'd)

(vii) (cont'd)

Proof (4) :

When $n=1$,

Let there be 3 vectors : x, y, z.

5 possible combinations of ~~inequality~~ equality/inequality between x, y and z

1. $x=y$, $y=z$, $x=z$
2. $x\neq y$, $y=z$, $x\neq z$
3. $x=y$, $y\neq z$, $x\neq z$
4. $x\neq y$, $y\neq z$, $x\neq z$
5. $x=y$, $y=z$, $x\neq z$

For each of these combinations, we assess if the triangle inequality holds :

	$d(x,y)$	$\leq$	$d(x,z)$	$+$	$d(y,z)$
1.	0	$\leq$	0	+	0
2.	1	$\leq$	0	+	1
3.	0	$\leq$	1	+	1
4	1	$\leq$	1	+	1
5.	0	$\leq$	0	+	1

$\therefore$ $d(x,y)$ satisfies the triangle inequality.

$\Rightarrow$ Function (vii) is a metric

Exercise 2b

Minkowski distance is given by
$\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}$, where $p \in R^+$

i) $d(ax, ay)$ for the Minkowski distance =

$$\left(\sum_{i=1}^{n} |a \cdot x_i - a \cdot y_i|^p\right)^{\frac{1}{p}}$$

$$= \left(\sum_{i=1}^{n} |a^p| \cdot |x_i - y_i|^p\right)^{\frac{1}{p}}$$

$$= |a| \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

$$= |a| \, d(x, y)$$

ii) $d(x+z, y+z)$ for the Minkowski distance =

$$\left(\sum_{i=1}^{n} |(x_i + z_i) - (y_i + z_i)|^p\right)^{\frac{1}{p}}$$

$$= \left(\sum_{i=1}^{n} |x_i + z_i - y_i - z_i|^p\right)^{\frac{1}{p}}$$

$$= \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

$$= d(x, y)$$

Exercise 2c

$$d(ax, ay) = |a|\, d(x, y)$$

For function (vii), $d(ax, ay) = \begin{cases} 0 & \text{if } ax = ay \\ 1 & \text{if } ax \neq ay \end{cases}$

$\because$ a cancel out as a common factor between ax and ay,

which gives $d(ax, ay) = d(x, y)$,

$\therefore$ homogeneity does not apply to function (vii).

Exercise 2d

When $n=2$, let vectors $x\,(1,1)$, $y\,(1,2)$ and constant $z=1$

$x\,(1,1)$, $x+z\,(2,2)$
$y\,(1,2)$ $\quad$ $y+z\,(2,3)$

For function (vi),

$$d(x,y) = \frac{2}{\pi} \arccos\left(\frac{x_1 y_1 + x_2 y_2}{\sqrt{x_1^2 + x_2^2} \cdot \sqrt{y_1^2 + y_2^2}}\right)$$

$$= \frac{2}{\pi} \arccos\left(\frac{1+2}{\sqrt{2} \times \sqrt{5}}\right) = \frac{2}{\pi} \arccos\left(\frac{3}{\sqrt{10}}\right) = 11.7\ (3\text{s.f.})$$

$$d(x+z, y+z) = \frac{2}{\pi} \arccos\left(\frac{(x_1+z)\cdot(y_1+z) + (x_2+z)\cdot(y_2+z)}{\sqrt{(x_1+z)^2 + (x_2+z)^2} \times \sqrt{(y_1+z)^2 + (y_2+z)^2}}\right)$$

$$= \frac{2}{\pi} \arccos\left(\frac{4+6}{\sqrt{8} \times \sqrt{13}}\right) = \frac{2}{\pi} \arccos\left(\frac{10}{\sqrt{104}}\right)$$

$$= 7.20\ (3\text{s.f.})$$

$\because d(x,y) \neq d(x+z, y+z)$, $\therefore$ Translation invariance does not apply to function (vi).