UNIT II

Distances & Nearest Neighbors.

Definition:

Let x be any given set.

A distance $d: x \times x \rightarrow R^+$ is a bivariate operator (A takes two arguments $a \in X, b \in X$) that maps $R^+ = [0, \infty]$. is a metric is

(i) $d(a,b) \geq 0$. (non-negativity)

(ii) $d(a,b) = 0$ if and only if $a = b$ (identity)

$R^+ = [0 \infty]$

iii) $d(a,b) = d(b,a)$ [symmetry]

(iv) $d(a,b) \leq d(a,c) + d(c,b)$ [Triangle inequality

NOTE 1: Psedometric.

A distance that statifies (i), (iii) & (iv) (i.e) (i) $d(a,b) \geq 0$, (iii) $d(a,b) = d(b,a)$ (iv) $d(a,b) \leq d(a,c) + d(c,b)$ is called psedometric

NOTE 2:

A distance that statisfied (i), (ii), (iv) (i.e), $d(a,b) \geq 0$ (ii) $d(a,b) = 0$ if and only if $a=b$.

(iv) $d(a,b) \leq d(a,c) + d(c,b)$ is called Quasi metric

Definition of Lp - distance:

consider two vectors $a = (a_1, a_2, \ldots a_d)$ and $b = (b_1, b_2, b_3 \ldots b_d) \in R^d$.

Lp distance is defined as

$$d_p(a,b) = \| a-b \|_p = \left(\sum_{i=1}^{d} |a_i - b_i|^p \right)^{1/p}$$

Definition L_2 - distance (or) Euclidean distance

L_2 distance is defined as (p=2)

$$d_2(a,b) = \|a-b\|_2 = \sum_{i=1}^{d} \left(|a_i - b_i|^2 \right)^{1/2}$$

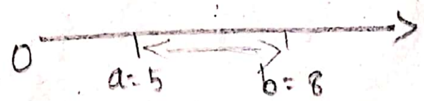
~~L_2~~ L_2 distance is called euclidean distance As also, it is distance b/w any two point along a line.

Definition : L_1 distance (or) Manhotten distance

L_1 is distance is defined as

$$d_1(a,b) = \|a-b\|_1 = \left(\sum_{i=1} |a_i - b_i| \right)$$

L_1 distance is called manhotten distance. It is lenght on the each co-ordinate axis.



$$d_1(5,8) = |5-8| = |-3| = 3$$

Scanned by CamScanner

20.7.23

2m

RANK - NULLITY THEOREM:

STATEMENTS:

Let V and W be two vector spaces. $T: V \rightarrow W$ be a linear transformation.

The dimension of N(T) is called Nullity of T. Dimension of R(T) is called range of T.

$$T(\alpha+\beta) = T(\alpha) + T(\beta)$$

$$T(k\alpha) = kT(\alpha)$$

N(T) is a set which contains non-zero vectors.

Nullity of T = dimension of Null space of T

Rank of T = dimension of range of T

The number of the non-zero vectors is the dimension of N(T) and it is the Nullity of T.

Number of element in the basis of R(T) is the dimension of R(T)

4M Rank - Nullity Theorem:

Statement:

Let V and W be vector spaces and $T: V \rightarrow W$ be a linear transformation.

If V is finite dimensional vector space then

$\dim V =$ Rank of T + Nullity of T

Case study:

Verify dimension theorem for the linear transformation defined by $T: R^2 \to R^3$

$T(a_1, a_2) = (a_1 + a_2, 0, 2a_1 - a_2)$

Solution:

$V = R^2$, $W = R^3$ are vector spaces

$T =$ Linear transformation

$T(a_1, a_2) = (a_1 + a_2, 0, 2a_1 - a_2)$

$\boxed{\dim(V) = \dim(R^2) = 2}$

To Find $R(T)$

Standrad basis for R^2

$\{(1,0); (0,1)\}$

Standrad basis for R^3.

$\{(1,0,0), (0,1,0), (0,0,1)\}$

Let basis be $B = \{(1,0); (0,1)\}$

$R(T) = \text{spam}(T(\beta))$

$T(a_1, a_2) = (a_1 + a_2, 0, 2a_1 - a_2)$

$T(1,0) = (1+0, 0, 2(1)-0)$

$= (1, 0, 2)$

$T(0,1) = (0+1, 0, 2(0)-1)$

$= (1, 0, \cancel{0}\ {-1})$

$a(1,0,2) + b(1,0,-1) = (0,0,0)$

$\Rightarrow (a, 0, 2a) + (b, 0, -b) = (0,0,0)$

$\Rightarrow (a+b, 0+0, 2a-b) = (0,0,0)$

$\Rightarrow (a+b, 0, 2a-b) = (0,0,0)$

$\Rightarrow (a+b) = 0$ and $2a-b=0$

Solving $a+b=0$ and $2a-b=0$ we get

$a+b=0$ $\quad$ $2a-b=0$

$a=-b$ $\quad$ $2a-(-a)=0$

$b=-a$ $\quad$ $3a=0$

$a=0$

$a=0,\quad 0-b = \boxed{b=0}$

$\therefore (1,0,2)$ and $(1,0,-1)$ are linearly independent.

$\therefore \{(1,0,2)\ (1,0,-1)\}$ is a basis for

R dim $R(T) = 2$

Rank of $T = 2$

Rank - Nullity Theorem

dim V = Rank of T = Nullity of T

To find Nullity of T

Null space of T $\quad x \xrightarrow{T} T(x) = \bar{0}$

definition:

$$N(T) = \{x \in V : T(x) = \bar{0}\}$$

$$N(T) = \{x = (a_1, a_2) \in \mathbb{R}^2 : T(a_1, a_2) = (0, 0, 0)\}$$

But, $T(a_1, a_2) = (a_1 + a_2, 0, 2a_1 - a_2)$

$$\therefore N(T) = \{x = (a_1, a_2) \in \mathbb{R}^2 : (a_1 + a_2, 0, 2a_1 - a_2) = (0, 0, 0)\}$$

$\Rightarrow a_1 + a_2 = 0 \longrightarrow (1)$

$2a_1 - a_2 = 0 \longrightarrow (2)$

Solving (1) & (2)

Adding (1) & (2)

$3a_1 = 0$

$a_1 = 0$

Sub $a_1 = 0$ in (1)

$0 + a_2 = 0$

$a_2 = 0$

$\therefore N(T) = \{(0, 0)\} = \{\bar{0}\}$

dim $N(T) = 0$

$\therefore$ Nullity of $T = 0$

Sub. dim $V = 2$, Rank of $T = 2$, Nullity of $T = 0$ in equation (I)

$$2 = 2 + 0 \qquad 2 = 2$$

$$\dim V = \text{Rank of } T + \text{Nullity of } T$$

2. If $T: R^3 \to R^2$ be a linear transformation defined by $T(x_1, x_2, x_3) = (x_1 - x_2, x_1 + x_3)$ find the rank of T and nullity of T.

(X) Doubt 6M

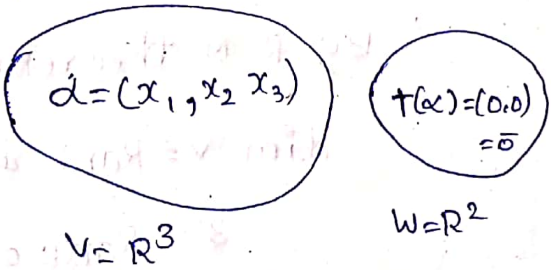
Soln:

To find nullity of T

Let $\alpha \in R^3 \Rightarrow \alpha(x_1, x_2, x_3)$

Suppose $\alpha \in N(T)$, by definition

$T(\alpha) = \bar{0}$



$= (x_1 - x_2, x_1 + x_3) = (0, 0)$

$= \cancel{x_1}\; x_1 - x_2 = 0, \; x_1 + x_3 = 0$

$x_1 = x_2$ and $x_1 = -x_3$

$x_2 = -x_3$

Scanned by CamScanner

Let $x_1 = k$, $k \in R$ (Real no.) in ①

$x_1 = x_2 = k = x_2 \Rightarrow x_2 = k$

$\therefore x_3 = -x_2 = -k$
$x_3 = -k$ $\Big\} \longrightarrow$ ①

$\therefore \alpha = (x_1, x_2, x_3) = (k, k, -k)$ $k \in R$

$k(1, 1, -1)$ $k \in R$

$\therefore \alpha \in N(T) = \alpha = k(1, 1, -1)$, $k \in R$,

Every vector of $N(T)$ can be spanned by a single vector $\{1, 1, -1\}$

$\therefore \dim N(T) = 1 \Rightarrow$ Nullity of $T = 1$.

Since $V = R^3$, we have $\dim V = 3$

By R-N theorem, we have

$\dim V =$ Rank of T + Nullity of T

$3 =$ Rank of $T + 1$

Rank of $T = 3 - 1 = 2$

Rank of $T = 2$

Nullity of $T = 1$

26/7/23

1) L_1 distance

$$d_1(a,b) = \|a_i - b_i\| = \sum_{i=1}^{d} |a_i - b_i|$$

2) L_2 distance

$$d_2(a,b) = \|a_i - b_i\|_2 = \sqrt{\sum_{i=1}^{d} |a_i - b_i|^2}$$

3) L_0 distance

$$d_0(a,b) = \|a-b\|_0 = d - \sum^{d} \mathbb{1}(a=b) \quad \text{where } \mathbb{1}(a=b) =$$

$$\begin{cases} 1 \text{ if } a=b \\ 0 \text{ if } a \neq b \end{cases}$$

4) L_∞ distance

$$d_\infty(a,b) = \|a-b\|_\infty = \max_{i=1,2,d} |a_i - b_i|$$

Q Consider two vectors : $a = (1, 2, -4, 3, -6)$
$b = (1, 2, 5, -2, 3) \in R^5$

calculate $d_1(a,b)$ $d_2(a,b)$, $d_\infty(a,b)$ $d_0(a,b)$ & sort the answers.

soln:

dimension, $d = 5$

given $a = (1, 2, -4, 3, -6)$
$= (a_1, a_2, a_3, a_4, a_5)$

$a_1 = 1$, $a_2 = 2$ $a_3 = -4$ $a_4 = 3$ $a_5 = -6$

$b = (1, 2, 5, -2, 3)$

$b_1 = 1$, $b_2 = 2$, $b_3 = 5$, $b_4 = -2$, $b_5 = 3$

(i) $d(a,b) = \sum_{i=1}^{J} |a_i - b_i|$

$= |a_1 - b_1| + |a_2 - b_2| + |a_3 - b_3| + |a_4 - b_4| + |a_5 - b_5|$

$= |1-1| + |2-2| + |-4-5| + |3-(-2)| + |-6-3|$

$= 0 + 0 + |-9| + |5| + |-9|$

$= 0 + 0 + 9 + 5 + 9 = 23$

$d_1(a,b) = 23$

(ii) $d_2(a,b) = \sqrt{\sum_{i=1}^{d} |a_i - b_i|^2}$

$= \sqrt{\sum_{i=1}^{5} (a_i - b_i)^2}$

$= \sqrt{(a_1-b_2)^2 + (a_2-b_2)^2 + (a_3-b_3)^2 + (a_4-b_4)^2 + (a_5-b_5)^2}$

$= \sqrt{(1-1)^2 + (2-2)^2 + (-4-5) + (3-(-2))^2 + (-6-3)^2}$

$= \sqrt{(-1)^2 + (2-2)^2 + (-4-5)^2 + (3-(-2))^2 + (-6-3)^2}$

$= \sqrt{0+0+81+25+81}$

$= \sqrt{187} = 13.67$

$d_2(a,b) = 13.67$

(iii) $d_0(a,b) = d - \sum_{i=1}^{d} \pi(a=b)$

when $\mathbb{1}(a=b) = \begin{cases} 1 \text{ if } a=b \\ 0 \text{ if } a \neq b \end{cases}$

$a = (1, 2, -4, 3, -6)$

$b = (1, 2, 5, -2, 3)$

$\sum_{i=1}^{5} \mathbb{1}(a=b) = 1+1+0+0+0$

$= 2$

$$\sum_{i=1}^{d_1} \mathbb{I}(a_i b) = 2$$

$$d_0(a,b) = d - \sum_{i=1}^{d} \mathbb{I}(a = b)$$

$$= 5 - 2$$

$$\therefore d_0(a,b) = 3$$

(iv) $d_\infty(a,b) = \|a-b\|_\infty = \max_{i=1,2,\dots 5} |a_i - b_i|$

$$= \max(|a_1-b_1| + |a_2-b_2| + |a_3-b_3| + |a_4-b_4| + |a_5-b_5|)$$

$$= \max(|1-1| + |2-2| + |-4-5| + |3-(-2)| + |-16-3|)$$

$$= \max(0, 0, 9, 5, 9)$$

$$d_\infty(a,b) = 9$$

Sorting the answers

$d_1(a,b) = 23$

$d_2(a,b) = 13.67$

$d_0(a,b) = 3$

$d_\infty(a,b) = 9$

(lowest to highest)

$3 < 9, < 13.67 < 23$

$d_0(a,b) < d_\infty(a,b) < d_2(a,b) < d_1(a,b)$

Scanned by CamScanner

27/7/23 Set operation:

1) If $A=\{1,3,5\}$, $B=\{2,3,4\}$ find

(i) $A\cap B$ (ii) $A\cup B$ (iii) $A\setminus B$ (iv) $B\setminus A$ (v) $A\Delta B$

vi) $\bar{A}$

Soln: Given $A=\{1,3,5\}$ $B=\{2,3,4\}$

(i) $A\cap B=\{3\}$

(ii) $A\cup B=\{1,2,3,4,5\}$

(iii) $A\setminus B=\{1,5\}$

(iv) $B\setminus A=\{2,4\}$

(v) symmetric diffrence:

$$A\Delta B=(A\setminus B)\cup(B\setminus A)$$
$$=\{1,2,4,5\}$$

(vi) $U=\{1,2,3,4,5\}$

$$\bar{A}=\{2,4\}$$

2) Verify triangle inequality of Jacard distance for any three sets A, B & C

we have to prove that

$$d_J(A,B)\leq d_J(A,C)+d_J(C,B)$$

C.

Consider

$$d_J(A,C) + d_J(C,B) = \frac{|A \setminus C|}{|A|} + \frac{|B \setminus C|}{|B|} \geq \frac{|A \setminus C| + |C \setminus B|}{|A \cup B|}$$

$$\geq \frac{|A \Delta B|}{|A \cup B|}$$

Hence $d_J(A,B) \leq d_J(A,C) + d_J(C,B)$

Hence Triangular inequality is verified

Q3) Find the cosine distance between points.

6M

$D_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$

$D_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$

Soln:

$$\cos(D_1, D_2) = \frac{|d_1 d_2|}{\sqrt{\Sigma d_1^2}\ \sqrt{\Sigma d_2^2}}$$ (cosine angular distance)

$$= \frac{5\times3 + 0\times0 + 3\times2 + 0\times0 + 2\times1 + 0\times1 + 0\times0 + 2\times1 + 0\times0 + 0\times1}{\sqrt{5^2+0^2+3^2+0^2+2^2+0^2+0^2}\ \sqrt{3^2+0^2+2^2+0^2+1^2+1^2+0^2+1^2+0^2+1^2}}$$

$$= \frac{15+6+2+2}{\sqrt{25+9+4+4}\ \sqrt{9+4+1+1+1+1}}$$

$$= \frac{25}{\sqrt{42}\sqrt{7}} = \frac{25}{5.48\times4.12} = \frac{25}{26.71} = 0.9359$$

$$\cos(D_1, D_2) = 0.94$$

4) Find the Jacord distance between sets

$A = \{1, 2, 5, 4\}$ $B = \{2, 3, 5, 7\}$

Soln :

Jacord distance

Formula :

$$d_s(A,B) = 1 - J_s(A-B)$$

$$= 1 - \frac{|A \cap B|}{|A \cup B|}$$

$A \cap B = \{2, 5\}$

$|A \cap B| = 2$

$A \cup B = \{1, 2, 3, 4, 5, 7\}$

$|A \cup B| = 6$

$$d_s(A,B) = 1 - \frac{2}{6} = 1 - 1/3 = 1 - 0.33 = 0.67$$

$$d_s[A,B] = 0.67$$

Scanned by CamScanner

UNIT I

(X) 4m 6M (X) 10M

Consider 2 vectors

$U = (0.5, 0.4, 0.4, 0.5, 0.1, 0.4, 0.1)$

$V = (-1, -2, 1, -2, 3, 1, -5)$

(i) Check U or V is a unit vector

(ii) calculate dot product $\langle U, V \rangle$

(iii) Are U and V orthogonal

Soln:

$u = (0.5, 0.4, 0.5, 0.1, 0.4, 0.1)$

Dimension $d = 7$

$$\|u\| = \sqrt{\sum_{i=1}^{d} u_i^2}$$

$$= \sqrt{\sum_{i=1}^{7} u_i^2}$$

$$= \sqrt{u_1^2 + u_2^2 + \dots u_7^2}$$

$$= \sqrt{(0.5)^2 + (0.4)^2 + (0.4)^2 + (0.5)^2 + (0.1)^2 + (0.4)^2 + (0.1)^2}$$

$$= \sqrt{0.25 + 0.16 + 0.16 + 0.25 + 0.01 + 0.16 + 0.01}$$

$$= \sqrt{1.00}$$

$\|u\| = 1$

Since $\|u\| = 1$, u is a unit vector

$|||^{ly}$ $\|v\| = \sqrt{\sum_{i=1}^{d} v_i^2}$

$v = (-1, -2, 1, -2, 3, 1, -5)$

$$= \sqrt{(-1)^2 + (-2)^2 + (1)^2 + (-2)^2 + 3^2 + 1^2 + (-5)^2}$$

Scanned by CamScanner

$= \sqrt{1+4+1+4+9+1+25}$

$= \sqrt{45}$

$\|v\| = 6.70$

$\therefore \|v\| = 6.70 \neq 1 = \|u\|$

So v is vector unit vector

(ii) Dot product (or) Inner product

$\langle U, V \rangle = U_1V_1 + U_2V_2 + U_3V_3 + \dots U_dV_d$

$= 0.5(-1) + 0.4(-2) + 0.4(1) + 0.5(-2) + 0.1 \times 3$
$+ 0.4 \times 1 + 0.1(5)$

$= -0.5 - 0.8 + 0.4 - 1 + 0.3 + 0.4 - 0.5$

$= -1.3 + 0.7 - 1 - 0.1$

$= -2.3 + 0.6$

$= -1.7$

(iii) Since $\langle U, V \rangle = -1.7 \neq 0$ the vector U and V are not orthogonal.

2) Consider following 3 vectors in R^9.

$v = (1, 2, 5, 2, -3, 1, 2, 6, 2)$

$u = (-4, -3, -2, 2, 1, -3, 4, 1, -2)$

$\omega = (3, 3, -3, -1, 6, -1, 2, -5, -7)$

Report the following.

(i) $\langle v, w \rangle$

(ii) Any pair of vectors orthogonal, and if so which ones?

(iii) $\|u\|_2$ (iv) $\|w\|_\infty$

(i) $\langle v, w \rangle = 3+6-15-2-18-1+4-30-14$

$= -6-20+3-44$

$= -67$

$$\|u_2\| = \sqrt{\sum_{i=1}^{d} u_i^2} = \sqrt{\sum_{i=1}^{9} u_i^2}$$

$$= \sqrt{u_1^2 + u_2^2 + u_3^2 + \dots u_9^2}$$

$$= \sqrt{(-4)^2 + (-3)^2 + (-2)^2 + (2)^2 + (1)^2 + (-3)^2 + (4)^2 + (1)^2 + (-2)^2}$$

$$= \sqrt{16 + 9 + 4 + 4 + 1 + 9 + 16 + 1 + 4}$$

$$= \sqrt{64} = 8$$

$$\boxed{\|v\|_\infty = \max_{i=1,2,\dots d} |v_i|}$$

Apply $\|w\|_\infty = \max_{i=1,2,\dots 9} |w_i|$

$= \max(|w_1|, |w_2|, |w_3| \dots |w_9|$

$= \max(|3|, |3|, |-3|, |-1|, |6|, |-1|, |2|, |-5|, |-7|)$

$= \max(3, 3, 3, 1, 6, 1, 2, 5, 7)$

$\therefore \|w\|_\infty = 7$

1) construct character - 2 - grams for the document

D_2 : G_2 = {{[Sam I], [I am]},

Soln :

{[sa], [am], [mi], [Ia], [am]}

= {[sa], [am], [mi], [Ia]}

2) Jarcad distance between two sets A & B

$$d_j(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

2) compute Jarcad distance Document among the document.

6M

D_1, D_2, D_3, D_4

D_1 : G_1 = {[I am], [am, sam]}

D_2 : G_2 = {[Sam I], [I am]}

D_3 : G_3 = {[I, do], [do not], [not like], [like Jelly] [Jelly and], [an ham]}

soln:-

$D_1 \cap D_2$ = {[I am]} = $|D_1 \cap D_2|$ = 1

$D_1 \cup D_2$ = {[I am], [am, sam], [Sam I]} = 3

(4) D - Dimension Space.

D = 11

for each co-ordinate list the corresponding word are

(1 am, 2 and, 3 do, 4 ham, 5 I, 6 jelly, 7 like, 8 not, 9 sam, 10 them, 11 zebra)

D_1 : I am sam

D_2 : Sam I am

D_3 : I do not like Jelly and ham

D_4 : I do not like them sam I am

For each of the above document from the representative vector.

Soln :-

$$V_1 = \begin{array}{ccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ \end{array}$$
$$\begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

$$V_2 = \begin{array}{ccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ \end{array}$$
$$\begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

$$V_3 = \begin{array}{ccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ \end{array}$$
$$\begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

$$V_4 = \begin{array}{ccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ \end{array}$$
$$\begin{pmatrix} 1 & 0 & 2 & 0 & 2 & 0 & 1 & 2 & 1 & 1 & 0 \end{pmatrix}$$

$$d_2(a, b) = \|a - b\| = \|a - b\|_2 = \sqrt{\sum_{i=1}^{d} (a_i - b_i)^2}$$

$d_2(v_1, v_2)$

$d_2(v_1, v_3)$

$d_2(v_1, v_4)$

$d_2(V_2, V_3)$

$d_2(V_2, V_4)$

$$d_2(V_3\, V_4) = \sqrt{\sum_{i=1}^{11} (a_i - b_i)^2}$$

$$= \sqrt{(0-1)^2 + (1-0)^2 + (1-2)^2 + (1-0)^2 + (1-2)^2 + (1-0)^2 + (-1-1)^2 + (1-2)^2 + (0-1)^2 + (0-0)^2 + (0-1)^2}$$

$$= \sqrt{98} = 3.,$$