FACTORIAL OF A NON-NEGATIVE INTEGER

$0! = 1$

$n! = n \cdot (n-1)!$ $\quad n = 1, 2, 3, \ldots$ $\square$

EXAMPLES:

$0! = 1$

$1! = 1$

$2! = 2 \cdot 1 = 2$

$3! = 3 \cdot 2 \cdot 1 = 6$

$4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$ $\square$

BINOMIAL COEFFICIENTS

DEFINITION: $\binom{n}{r} = \dfrac{n!}{r!(n-r)!}$ $\quad 0 \leq r \leq n$

$\binom{n}{r} \triangleq {}^{n}C_{r} \triangleq {}^{n}_{r}C$ $\square$

EXAMPLES

$\binom{5}{0} = 1$; $\binom{5}{1} = 5$; $\binom{5}{2} = 10$; $\binom{5}{3} = 10$; $\binom{5}{4} = 5$

$\binom{5}{5} = 1$

$\binom{10}{3} = \dfrac{10!}{3!7!} = \dfrac{10 \cdot 9 \cdot 8}{3!} = 120$ $\square$

NUMBER OF CLUSTERS, GIVEN A FIXED NUMBER OF DATA POINTS

LET $\overset{n}{\cancel{m}}$ = NUMBER OF DATA POINTS ; $n \in \mathbb{P}$

m = NUMBER OF DATA CLUSTERS ; $m \in \mathbb{P}$

$m \leq n$

$S(n, m)$ = # OF DIFFERENT CLUSTERS

= STIRLING NUMBER OF THE SECOND KIND

$$S(n,m) = \frac{1}{m!} \sum_{i=0}^{m} (-1)^{m-i} \binom{m}{i} i^n \quad ; \quad 1 \leq m \leq n$$

□

EXAMPLES: $S(5,3) = 25$; $S(3,2) = 3$

$S(4,3) = 6$; $S(3,1) = 1$ □

EXAMPLE : COMPUTE EXPLICITLY $S(5,3)$

$$S(5,3) = \frac{1}{3!} \sum_{i=1}^{3} (-1)^{3-i} \binom{3}{i} i^5$$

$$= \frac{1}{6} \left[(-1)^{3-1} \binom{3}{1} 1^5 + (-1)^{3-2} \binom{3}{2} 2^5 + (-1)^{3-3} \binom{3}{3} 3^5 \right]$$

$$= \frac{1}{6} \left[3 - 3(32) + 3^5 \right]$$

$$= \frac{3}{6} \left[1 - 32 + 81 \right] = \frac{3}{6} \cdot 50 = 25$$

□

SIMILARITY MEASURES BETWEEN POINTS

SOME POPULAR SIMILARITY MEASURES BETWEEN POINTS ARE:

1. COSINE SIMILARITY MEASURE
2. CORRELATION SIMILARITY MEASURE
3. EUCLIDEAN SIMILARITY MEASURE

THESE MEASURES ARE BEST ILLUSTRATED VIA EXAMPLES.

EXAMPLE 1: LET $x = (1,1,1,1)$; $y = (2,2,2,2)$

(a) COSINE SIMILARITY:

DOT PRODUCT $\|\cdot\|$ = EUCLIDEAN NORM / DISTANCE

$$\cos(x,y) = \frac{x \cdot y}{\|x\|\,\|y\|}$$

NOTE THAT IF: $x = (x_1, x_2, x_3, x_4)$

$$y = (y_1, y_2, y_3, y_4)$$

$$x \cdot y = x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4$$

IN THIS EXAMPLE: $x \cdot y = (1 \cdot 2)4 = 8$

$\|x\|^2 = x_1^2 + x_2^2 + x_3^2 + x_4^2 = 1^2 + 1^2 + 1^2 + 1^2 = 4$; $\|x\| = 2$

$\|y\|^2 = (2^2)4 = 16$; $\|y\| = 4$

$$\cos(x,y) = \frac{8}{2 \cdot 4} = \boxed{1}$$

(b) CORRELATION SIMILARITY MEASURE

$\bar{x} = \frac{1}{4}(x_1 + x_2 + x_3 + x_4)$; $\bar{y} = \frac{1}{4}(y_1 + y_2 + y_3 + y_4)$

$S_x^2 = \frac{1}{4} \sum_{i=1}^{4} (x_i - \bar{x})^2$; $S_y^2 = \frac{1}{4} \sum_{i=1}^{4} (y_i - \bar{y})^2$

$S_{xy} = \frac{1}{4} \sum_{i=1}^{4} (x_i - \bar{x})(y_i - \bar{y})$; $\mathrm{CORR}(x,y) = \dfrac{S_{xy}}{S_x S_y}$

$$s_x^2 = \frac{1}{4}\{(1-1)^2+(1-1)^2+(1-1)^2+(1-1)^2\} = 0$$

SIMILARLY $s_y^2 = 0$

$$s_{xy} = \frac{1}{4}\{(1-1)(2-2)+(1-1)(2-2)+(1-1)(2-2)+(1-1)(2-2)\} = 0$$

$$\text{CORR}(x,y) = \frac{0}{0} = \boxed{\text{UNDEFINED}}$$

(c) EUCLIDEAN SIMILARITY

$$\| x-y\| = \{(1-2)^2+(1-2)^2+(1-2)^2+(1-2)^2\}^{1/2} = \boxed{2}$$ □

EXAMPLE 2: LET $x = (0,1,0,1)$; $y = (1,0,1,0)$

(a) COSINE SIMILARITY

$$\cos(x,y) = \frac{x \cdot y}{\|x\| \, \|y\|}$$

$x \cdot y = 0$; $\|x\| = \sqrt{2}$; $\|y\| = \sqrt{2}$; $\cos(x,y) = \frac{0}{2} = \boxed{0}$

(b) CORRELATION SIMILARITY MEASURE

$\bar{x} = \frac{1}{2}$; $\bar{y} = \frac{1}{2}$

$$s_x^2 = \frac{1}{4}\{(0-0.5)^2+(1-0.5)^2+(0-0.5)^2+(1-0.5)^2\} = \frac{1}{4}$$

$$s_y^2 = \frac{1}{4}\{4 \cdot \frac{1}{4}\} = \frac{1}{4}$$

$$s_{xy} = \frac{1}{4}\{(0-0.5)(1-0.5)+(1-0.5)(0-0.5)+(0-0.5)(1-0.5)+(1-0.5)(0-0.5)\}$$
$$= -\frac{1}{4}$$

$$\text{CORR}(x,y) = s_{xy}/(s_x s_y) = \boxed{-1}$$

(c) EUCLIDEAN SIMILARITY MEASURE

$$\| x-y\| = \{(0-1)^2+(1-0)^2+(0-1)^2+(1-0)^2\}^{1/2} = \boxed{2}$$ □

CANONICAL MEASURES OF DISSIMILARITY BETWEEN TWO CLUSTERS C_i AND C_j

DENOTE THE DISSIMILARITY MEASURE BETWEEN TWO CLUSTERS C_i AND C_j BY $D(C_i, C_j)$; WHERE $i \neq j$,
$d(x, y)$ = SIMILARITY MEASURE BETWEEN POINTS x AND y

1. MINIMUM MEASURE

$$\hat{D}_{MIN}(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y)$$

2. MAXIMUM MEASURE

$$\hat{D}_{MAX}(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x, y)$$

3. MEAN MEASURE

ASSUME THAT THE MEANS OF CLUSTERS C_i AND C_j ARE PROPERLY DEFINED. LET THESE BE a AND b RESPECTIVELY.

$$\hat{D}_{MEAN}(C_i, C_j) = |a - b|$$

4. AVERAGE MEASURE

$$\hat{D}_{AVG}(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{\substack{x \in C_i \\ y \in C_j}} d(x, y)$$, WHERE

$$C_i \neq \phi \ \ \& \ \ C_j \neq \phi$$