INTELIGÊNCIA COMPUTACIONAL

MEDIDAS DE SIMILARIDADE

FELIPE TORRES

SIMILARIDADE

Quanto a maçã amarela e vermelha são diferentes da verde ?

SIMILARIDADE

Essas maçãs são diferentes ? Quanto ?

MEDIDAS DE SIMILARIDADE

- Para comparar dois conjuntos de valores, os algoritmos utilizam as medidas de similaridade.

- Tabela de contigência para dados binários:

- Medida de distância para variáveis binárias simétricas:

- Medida de distância para variáveis binárias assimétricas :

- Coeficiente de Jaccard (similaridade): similaridade entre conjuntos

DADOS BINÁRIOS - JACCARD

- Para comparar dois conjuntos de valores, os algoritmos utilizam as medidas de similaridade.
- Tabela de contigência para dados binários:

Instância *i* \ Instância *j*	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

- Medida de distância para variáveis binárias simétricas:

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

- Medida de distância para variáveis binárias assimétricas :

$$d(i,j) = \frac{r+s}{q+r+s}$$

- Coeficiente de Jaccard (similaridade): similaridade entre conjuntos

$$sim_{Jaccard}(i,j) = \frac{q}{q+r+s}$$

JACCARD – EXEMPLOS

- Atributos binários assimétricos
- Y e P são 1, N é 0

Name	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	Y	N	P	N	N	N
Mary	Y	N	P	N	P	N
Jim	Y	P	N	N	N	N

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

JACCARD– EXEMPLOS

$Grupo\ 1 = \{A, B, C, D\}$
$Grupo\ 2 = \{E, F, G, A, C\}$

|Grupo 1| = 4 e |Grupo2| = 5

União Grupo 1 U Grupo 2 = {A,B,C,D,E,F,G}
Interjeição Grupo 1 e Grupo 2 = {A,C}

Jaccard Similarity J (A,B) = | Intersection (A,B) | / | Union (A,B) |

= 2 / 7

= 0.286

MINKOWSKI

- Distância de Minkowski:

$$d(i,j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

onde i = (xi1, xi2, ..., xip) e j = (xj1, xj2, ..., xjp) são duas instâncias de dados de p dimensões, e h é a ordem (definindo a distância chamada norma L-h)

MANHATTAN

- Casos especiais da distância de Minkowski
  - h = 1: Manhattan (city block, L1 norm)

$$D(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|$$

  - Já a Distância Manhattan tem uma definição mais simples na qual é apenas a **soma das diferenças entre x e y em cada dimensão**.

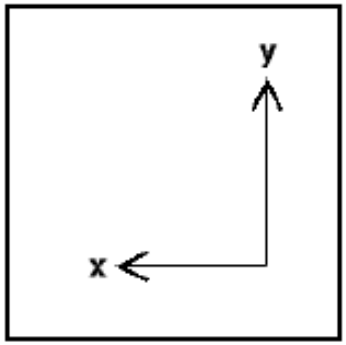
EUCLIDIANA

- Casos especiais da distância de Minkowski
  - h = 2: (L2 norm) Euclidiana

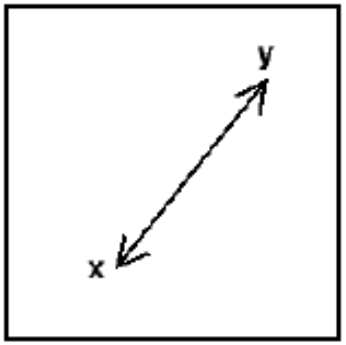
$$D(i,j) = \sqrt{(\left|x_{i1} - x_{j1}\right|^2 + \left|x_{i1} - x_{j1}\right|^2 + \cdots + \left|x_{ip} - x_{jp}\right|^2)}$$

  - A Distância Euclidiana é definida **como a soma da raiz quadrada da diferença entre x e y em suas respectivas dimensões**.

COMPARAÇÃO DE DISTÂNCIAS



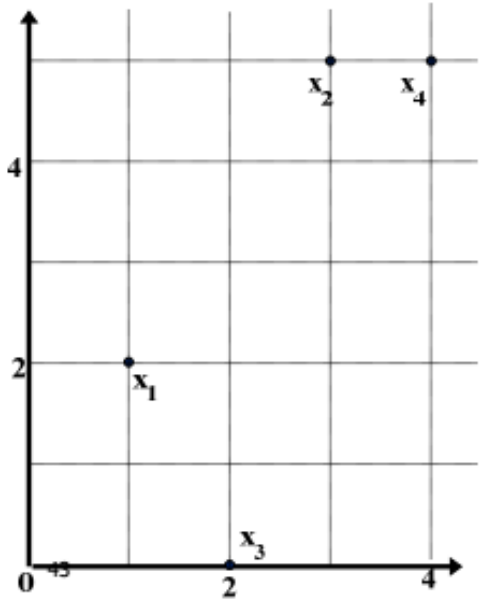
Manhattan



Euclidean

DISTÂNCIAS - EXEMPLOS

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

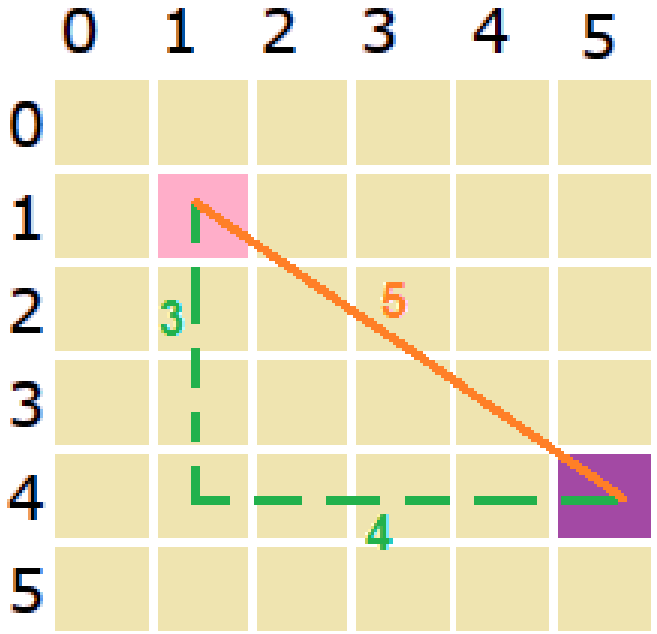
L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Para datasets com grande dimensão o Manhattan trabalha melhor que a distância Euclidiana.

DISTÂNCIAS - EXEMPLOS



Point P1 = (1,1)

Point P2 = (5,4)

Euclidean distance = $\sqrt{(5-1)^2 + (4-1)^2} = 5$

Manhattan distance = $|5-1| + |4-1| = 7$

MATRIZ DE SIMILARIDADE – BLOSUM62

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Você usava medidas de similaridade e não sabia...

INTELIGÊNCIA COMPUTACIONAL

MEDIDAS DE SIMILARIDADE

FELIPE TORRES