



# Analítica y Visualización de Datos

Dr. Miguel Jesús Torres Ruiz



## Distancia de Euclidiana <sup>(1)</sup>

- **Definición**

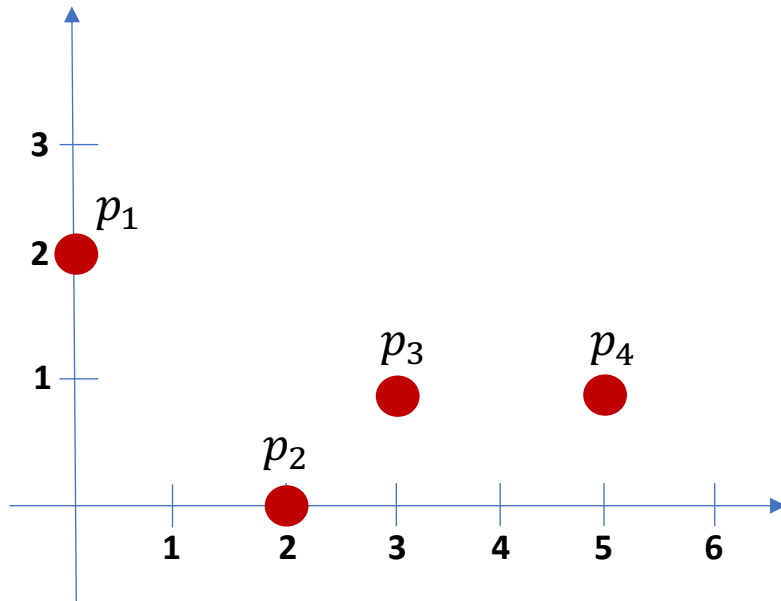
$$d_E = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$$

- **Donde:**

- $n$  es el número de dimensiones (atributos)
  - $p_k$  y  $q_k$  son respectivamente los  $k$ -ésimos atributos (componentes) u objetos de datos  $p$  y  $q$ .
  - $p_k$  y  $q_k$  son respectivamente los  $k$ -ésimos atributos (componentes) u objetos de datos  $p$  y  $q$ .
- Requiere de una estandarización, en caso de que las escalas difieran entre sí.



## Distancia de Euclidiana (2)



Punto	Altura	Peso
	$x$	$y$
$p_1$	0	2
$p_2$	2	0
$p_3$	3	1
$p_4$	5	1



	$p_1$	$p_2$	$p_3$	$p_4$
$p_1$	0	2.828	3.162	5.099
$p_2$	2.828	0	1.414	3.162
$p_3$	3.1622	1.414	0	2
$p_4$	5.099	3.162	2	0

**Matriz de Distancia**



## Distancia de Minkowski <sup>(1)</sup>

- **Definición**

- La distancia de Minkowski es una métrica definida en un espacio vectorial normado que puede considerarse como una generalización tanto de la distancia Euclidiana como de la distancia de Manhattan.
- La distancia del orden de Minkowski  $r$ , donde  $r$  es un número entero entre dos puntos.
- Por tanto, se tiene que  $P = (p_1, p_2, \dots, p_n)$  y  $Q = (q_1, q_2, \dots, q_n) \in \mathbb{R}^n$ , se define como:

$$Dist L_P = \left( \sum_{k=1}^n |p_k - q_k|^r \right)^{\frac{1}{r}}$$

- Donde:
  - $r$  es un parámetro
  - $n$  es el número de dimensiones (atributos),
  - $p_k$  y  $q_k$  son respectivamente los  $k$ -ésimos atributos (componentes) u objetos de datos  $p$  y  $q$ .



## Distancia de Minkowski <sub>(2)</sub>

- **Definición**

- Para  $r \geq 1$ , la distancia de Minkowski es una métrica como resultado de la Desigualdad de Minkowski.
- Cuando  $r < 1$ , la distancia entre  $(0,0)$  y  $(1,1)$  es  $2^{\frac{1}{r}} > 2$ , pero el punto  $(0,1)$  está a la distancia de 1 de estos dos puntos.
- Esto viola la propiedad de desigualdad del triángulo, porque  $r < 1$  no es una métrica. Sin embargo se puede obtener una métrica para estos valores, simplemente eliminando el exponente  $\frac{1}{r}$
- La distancia de Minkowski también se puede ver como un múltiplo de la potencia media de las diferencias por componentes de  $r$ .



## Distancia de Minkowski <sub>(3)</sub>

- **Definición**

- La familia de distancias de Minkowski  $Dist L_P = \left( \sum_{k=1}^n |p_k - q_k|^r \right)^{\frac{1}{r}}$ 
  - Costo de evaluación  $O(n)$
  - Cumplen con las propiedades métricas
  - Cuando  $r = 1$ : Se conoce como City Block (Manhattan, [Taxicab](#), norma  $L_1$ )
    - Un ejemplo común de esto es la distancia de *Hamming*, que es solo la cantidad de bits que son diferentes entre dos vectores binarios.
  - Cuando  $r = 2$ : Distancia Euclidiana (norma  $L_2$ )
  - Cuando  $r = \infty$ : Distancia Chebyshev (chessboard, norma  $L_{max}$ , norma  $L_\infty$ , Distancia Máxima, Distancia Suprema)
    - Esta es la diferencia máxima entre cualquier componente de los vectores.
      - Por ejemplo:  $L_\infty$  de  $(1, 0, 2)$  y  $(6, 0, 3) = ??? = 5$
  - No confundir  $r$  con  $n$ , es decir, todas estas distancias están definidas para todos los números de dimensiones.

$$L_1(P, Q) = \sum_{k=1}^n |p_k - q_k|$$

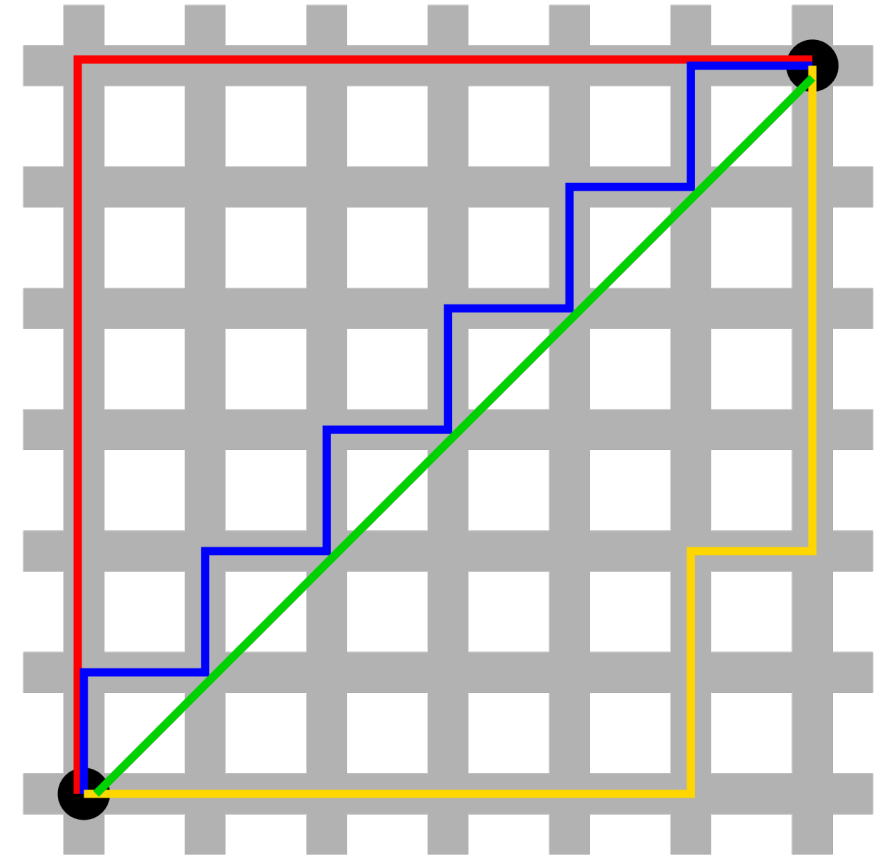
$$L_2(P, Q) = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$$

$$L_{max}(P, Q) = \max_{1 \leq k \leq n} \{|p_k - q_k|\}$$



# Distancia de Minkowski <sub>(4)</sub>

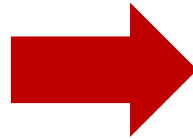
**Taxicab**





## Distancia de Minkowski <sup>(5)</sup>

Punto	$x$	$y$
$p_1$	0	2
$p_2$	2	0
$p_3$	3	1
$p_4$	5	1



### Matriz de Distancia

$L_1$	$p_1$	$p_2$	$p_3$	$p_4$
$p_1$	0	4	4	6
$p_2$	4	0	2	4
$p_3$	4	2	0	2
$p_4$	6	4	2	0

$L_2$	$p_1$	$p_2$	$p_3$	$p_4$
$p_1$	0	2.828	3.162	5.099
$p_2$	2.828	0	1.414	3.162
$p_3$	3.162	1.414	0	2
$p_4$	5.099	3.162	2	0


$L_\infty$	$p_1$	$p_2$	$p_3$	$p_4$
$p_1$	0	2	3	5
$p_2$	2	0	1	3
$p_3$	3	1	0	2
$p_4$	5	3	2	0



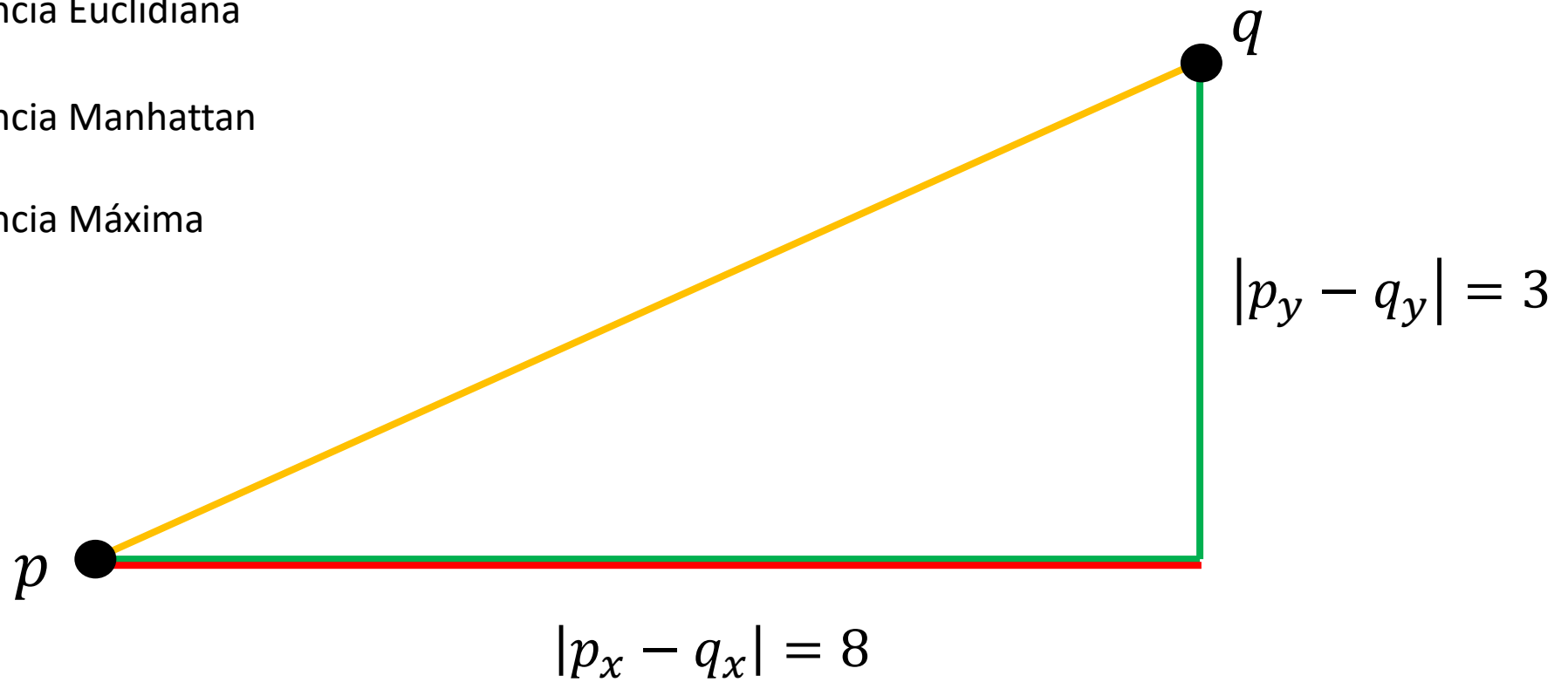


## Distancia de Minkowski <sub>(6)</sub>

 Distancia Euclidiana

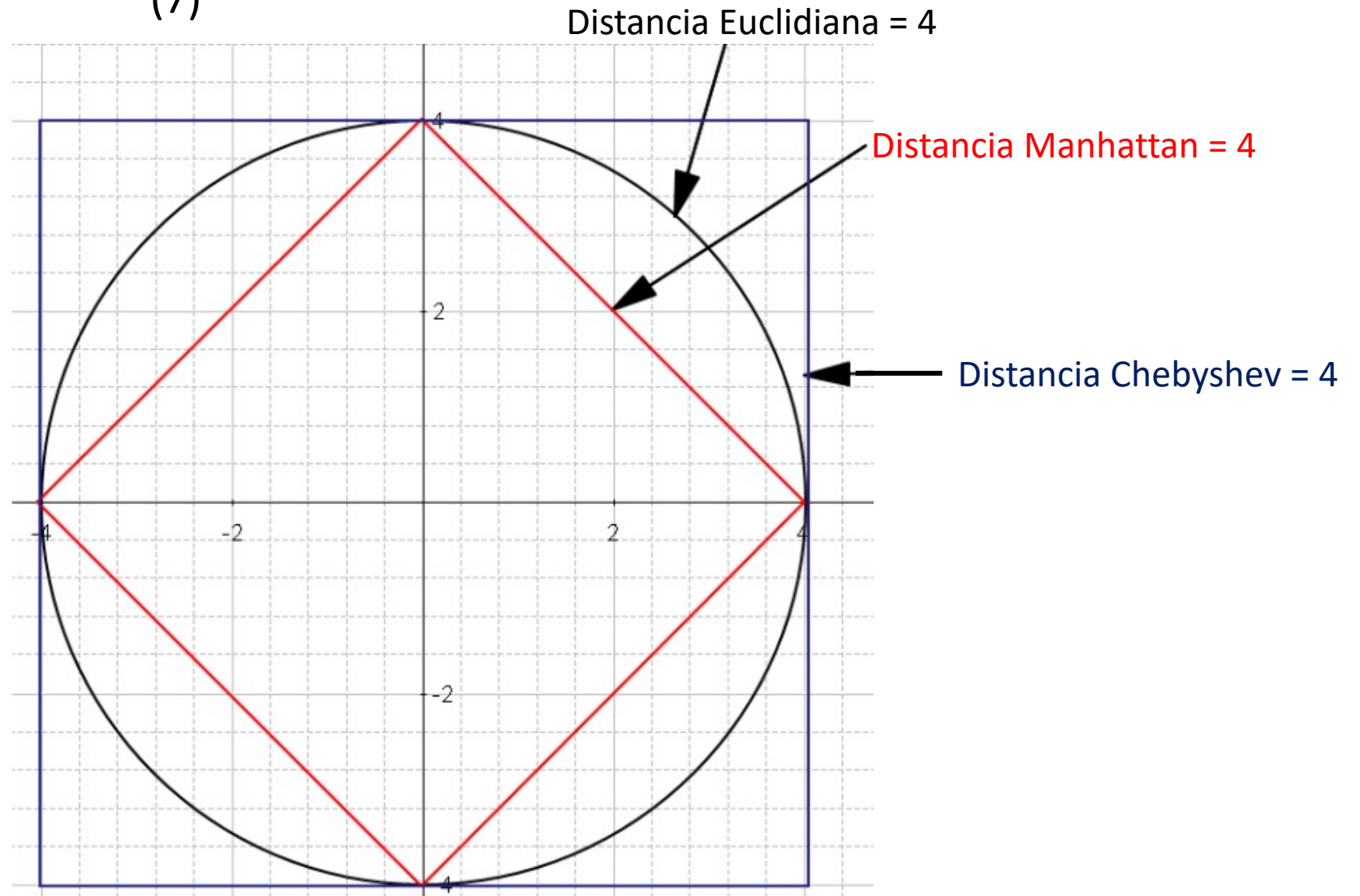
 Distancia Manhattan

 Distancia Máxima





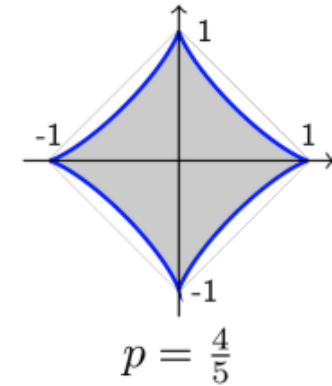
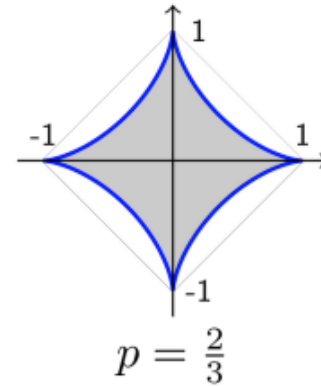
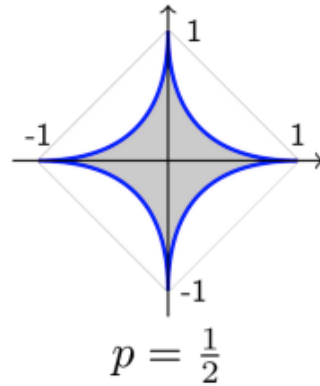
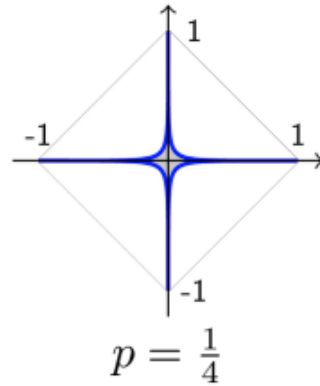
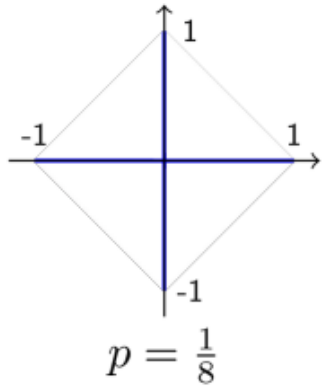
## Distancia de Minkowski <sup>(7)</sup>



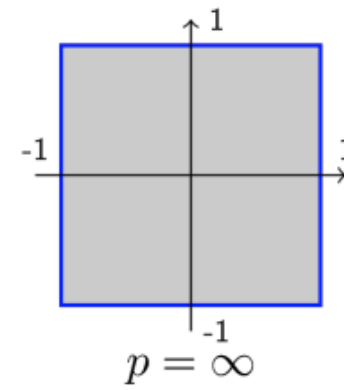
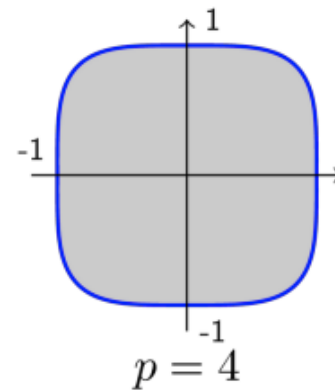
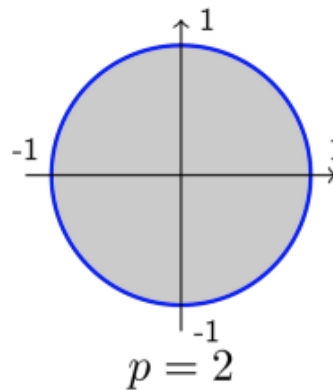
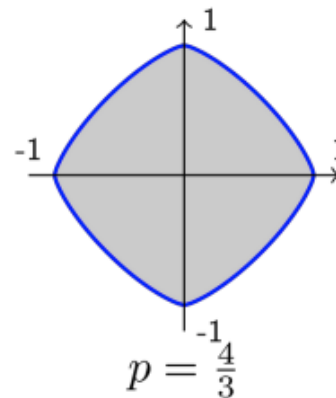
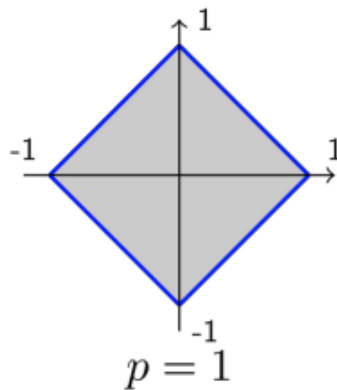


## Distancia de Minkowski (8)

$$C_p = \{(x, y) | (|x|^p + |y|^p)^{1/p} \leq 1\}$$



$p < 1$ : Conjuntos no convexos



$p \geq 1$ : Conjuntos convexos



# Distancia de Hamming <sup>(1)</sup>

- **Definición**

- La distancia de Hamming se define como:  $D_H(x, y) = \sum_{i=1}^p \rho(x^{(i)}, y^{(i)})$  y considerando la métrica discreta:

$$\rho(x, y) = \begin{cases} 0 & \text{si } x = y \\ 1 & \text{Otro caso} \end{cases}$$

- Entonces, la distancia de Hamming produce la cantidad de valores de características que no coinciden.
  - Para características binarias, la distancia de Hamming es igual a la distancia de Manhattan.
- Sin embargo, la distancia de Hamming no está asociada con una norma porque la condición  $\|\alpha \cdot x\| = |\alpha| \cdot \|x\| \forall \alpha \in \mathbb{R}, x \in \mathbb{R}^p$  no se cumple.
- Las variantes de la distancia de Hamming usan funciones modificadas  $\rho$  para especificar similitudes entre características individuales.
  - Por ejemplo, si las características son páginas web (escala nominal), entonces  $\rho$  podría ser menor para pares de páginas con contenido similar y mayor para pares de páginas con contenido bastante diferente.



## Distancia de Hamming <sup>(2)</sup>

- **Definición**

- La distancia de Hamming se utiliza en procesamiento de señales y telecomunicaciones.
  - Contar el número de bits corruptos en la transmisión de un mensaje de una longitud determinada.
- Permite cuantificar la diferencia entre dos secuencias de símbolos.
- Es una distancia en el sentido matemático, con dos secuencias de símbolos de la misma longitud y asocia el número de posiciones donde difieren las dos secuencias.
- Para comparar secuencias de longitudes variables o cadenas de caracteres que pueden sufrir no solo sustituciones, sino también inserciones o borrados, se utiliza la **distancia de Levenshtein**.

- **Ejemplo:**

- $\alpha = (0\ 0\ 0\ 1\ 1\ 1\ 1)$ ;  $\beta = (1\ 1\ 0\ 1\ 0\ 1\ 1) \therefore d_H = 1 + 1 + 0 + 0 + 1 + 0 + 0 = 3$
- La distancia entre  $\alpha$  y  $\beta$  es igual a 3 porque **3 bits difieren**.
- La distancia de Hamming entre **(1 0 1 1 0 1)** y **(1 0 0 1 0 0 1)** es **2**
- La distancia de Hamming entre **(2 14 3 8 96)** y **(2 23 3 7 96)** es **3**
- La distancia de Hamming entre "**r a m e r**" y "**c a s e s**" es **3**