

# APÉNDICE J

---

## J. MEDIDAS DE PROXIMIDAD

### J.1 DEFINICIONES

En este apéndice vamos a estudiar diferentes tipos de medidas entre vectores, las componentes de estos vectores representarán medidas de ciertas propiedades o atributos de los objetos. Se pueden considerar cuatro categorías de atributos: *nominal*, *ordinal*, de *intervalo-escalado* y de *relación-escalado*.

La primera categoría incluye atributos cuyos posibles valores son códigos de estado, consideremos por ejemplo un atributo que corresponde al sexo de una especie animal, sus posibles valores pueden ser 1 para hembras y 0 para machos, una comparación cuantitativa entre esos valores carece de sentido. La siguiente categoría incluye atributos que pueden ser ordenados. Consideremos por ejemplo las notas de los estudiantes en una asignatura. Supongamos que los posibles valores son 1, 2, 3, 4, 5 y que corresponden a Suspenso, Aprobado, Notable, Sobresaliente y Matrícula de Honor. Esos valores se pueden ordenar con un cierto significado, sin embargo la diferencia entre dos valores sucesivos no tiene importancia cuantitativa.

Si para un determinado atributo, la diferencia entre dos valores es importante, mientras su relación carece de importancia, entonces se dice que es un atributo intervalo-escalado. Un ejemplo típico es la medida de la temperatura. En efecto, si las temperaturas en Madrid y Guadalajara son 20 y 10 grados Celsius, respectivamente, entonces tiene interés decir que la temperatura en Madrid es 10 grados más que en

Guadalajara, pero carece de sentido decir que Madrid es dos veces tan caliente como Guadalajara.

Finalmente, si la relación entre dos valores de un atributo específico es importante, entonces se trata de un atributo relación-escalado. Un ejemplo es el peso, puesto que es posible afirmar que una persona con 100 kilogramos de peso es doble de gruesa que otra con 50 kilogramos.

Una vez especificado el tipo de atributos, seguimos con las definiciones relativas a medidas entre vectores, y luego extendemos el concepto para incluir medidas entre subconjuntos del conjunto de datos  $X$ .

Una *medida de distinción* (MD) o de no similitud  $d$  en  $X$  es una función,

$$d: X \times X \rightarrow \Re \quad (\text{J.1})$$

donde  $\Re$  es el conjunto de números reales, tales que

$$\exists d_0 \in \Re: \quad -\infty < d_0 \leq d(\mathbf{x}, \mathbf{y}) < +\infty, \quad \forall \mathbf{x}, \mathbf{y} \in X \quad (\text{J.2})$$

$$d(\mathbf{x}, \mathbf{x}) = d_0, \quad \forall \mathbf{x} \in X \quad (\text{J.3})$$

y

$$d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x}), \quad \forall \mathbf{x}, \mathbf{y} \in X \quad (\text{J.4})$$

Si además se cumple que

$$d(\mathbf{x}, \mathbf{y}) = d_0 \quad \text{sii} \quad \mathbf{x} = \mathbf{y} \quad (\text{J.5})$$

y

$$d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}), \quad \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in X \quad (\text{J.6})$$

se dice que  $d$  es una *métrica* MD. La desigualdad (J.6) se conoce como la *desigualdad triangular*. La equivalencia (J.5) indica que el mínimo valor del nivel de distinción posible  $d_0$  entre cualesquiera dos vectores en  $X$  se consigue cuando son idénticos. Algunas veces nos referiremos al término distinción como distancia, en particular cuando no se usa en el sentido matemático estricto.

Una *medida de similitud* (MS)  $s$  en  $X$  es una función,

$$s: X \times X \rightarrow \Re \quad (\text{J.7})$$

$$\exists s_0 \in \mathfrak{R}: \quad -\infty < s(\mathbf{x}, \mathbf{y}) \leq s_0 < +\infty, \quad \forall \mathbf{x}, \mathbf{y} \in X \quad (\text{J.8})$$

$$s(\mathbf{x}, \mathbf{x}) = s_0, \quad \forall \mathbf{x} \in X \quad (\text{J.9})$$

y

$$s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x}), \quad \forall \mathbf{x}, \mathbf{y} \in X \quad (\text{J.10})$$

Si además se cumple que

$$s(\mathbf{x}, \mathbf{y}) = s_0 \quad \text{sii} \quad \mathbf{x} = \mathbf{y} \quad (\text{J.11})$$

y

$$s(\mathbf{x}, \mathbf{y})s(\mathbf{y}, \mathbf{z}) \leq [s(\mathbf{x}, \mathbf{y}) + s(\mathbf{y}, \mathbf{z})]s(\mathbf{x}, \mathbf{z}), \quad \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in X \quad (\text{J.12})$$

se dice que  $s$  es una *métrica* MS.

A modo de ejemplo consideremos una métrica muy habitual como es la distancia Euclídea  $d_2$  definida como

$$d_2(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^l (x_i - y_i)^2} \quad (\text{J.13})$$

donde  $\mathbf{x}, \mathbf{y} \in X$  y  $x_i, y_i$  son las  $i$ -ésimas componentes de  $\mathbf{x}$  e  $\mathbf{y}$  respectivamente. Ésta es una medida de distinción en  $X$ , con  $d_0 = 0$ ; esto es, la mínima distancia posible entre dos vectores de  $X$  es 0. Además la distancia de un vector desde él mismo es cero. También es fácil observar que se cumple (J.4). Además la distancia entre dos vectores toma su máximo valor  $d_0 = 0$  cuando los vectores coinciden. Finalmente, se puede demostrar la desigualdad triangular (J.6), por tanto la distancia Euclídea es una métrica de distinción. Conviene resaltar que para otras medidas los valores  $d_0$  ( $s_0$ ) pueden ser positivos o negativos.

Es fácil extender las definiciones anteriores para medir la proximidad entre subconjuntos de  $X$ . En efecto, sea  $U$  un conjunto que contiene subconjuntos de  $X$ , esto es,  $D_i \subset X$ ,  $i = 1, \dots, k$ , y  $U = \{D_1, \dots, D_k\}$ . Una medida de proximidad  $\wp$  en  $U$  es una función,

$$\wp: U \times U \rightarrow \mathfrak{R} \quad (\text{J.14})$$

Las ecuaciones (J.2) a (J.6) para medidas de distinción y las (J.8) a (J.12) para medidas de similitud pueden repetirse ahora con  $D_i, D_j$  en lugar de  $\mathbf{x}$  e  $\mathbf{y}$  y  $U$  en lugar de  $X$ . Generalmente, las medidas de proximidad entre dos conjuntos  $D_i$  y  $D_j$  se definen en términos de medidas de proximidad entre elementos de  $D_i$  y  $D_j$ .

Por ejemplo, sea  $X = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$  y  $U = \{\{\mathbf{x}_1, \mathbf{x}_2\}, \{\mathbf{x}_1, \mathbf{x}_4\}, \{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}, \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}\}$ , podemos definir la siguiente función de distinción:

$$d_{min}^{ss}(D_i, D_j) = \min_{\mathbf{x} \in D_i, \mathbf{y} \in D_j} d_2(\mathbf{x}, \mathbf{y}) \quad (J.15)$$

donde  $d_2$  es la distancia Euclídea entre dos vectores. El mínimo valor posible de  $d_{min}^{ss}$  es 0. También  $d_{min}^{ss}(D_i, D_i) = 0$ , como la propiedad conmutativa también se cumple, esta función es una medida, sin embargo no es una métrica porque dos conjuntos  $D_i$  y  $D_j$  pueden tener un elemento en común y aunque sean diferentes su distancia es nula, por ejemplo  $\{\mathbf{x}_1, \mathbf{x}_2\}$  y  $\{\mathbf{x}_1, \mathbf{x}_4\}$  de  $U$  en los que  $\mathbf{x}_1$  es común.

## J.2 MEDIDAS DE PROXIMIDAD ENTRE DOS PUNTOS

### VECTORES REALES

#### A) Medidas de distinción

Las MDs más comunes entre vectores reales utilizadas en la práctica son:

A.1) Las métricas  $l_p$  promediadas, esto es

$$d_p(\mathbf{x}, \mathbf{y}) = \left( \sum_{i=1}^l w_i |x_i - y_i|^p \right)^{1/p} \quad (J.16)$$

donde  $x_i, y_i$  son las coordenadas  $i$ -ésimas de  $\mathbf{x}$  e  $\mathbf{y}$ ,  $i = 1, \dots, l$  y  $w_i \geq 0$  es el  $i$ -ésimo coeficiente de peso. Cuando los  $w_i$  toman el valor uno, obtenemos la métrica  $l_p$  no promediada. Una medida de esta categoría es la distancia Euclídea, que se introdujo anteriormente con  $p = 2$ .

La métrica  $l_2$  puede generalizarse como sigue,

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^t B (\mathbf{x} - \mathbf{y})} \quad (J.17)$$

donde  $B$  es una matriz simétrica y definida positiva (una matriz  $B$  se dice que es definida positiva si para cada vector  $\mathbf{v}$  distinto de cero, se cumple que  $\mathbf{v}^T B \mathbf{v} > 0$ ). Esto incluye la distancia de Mahalanobis como un caso especial, siendo también una métrica MD

Otras métricas derivadas a partir de (J.16) son la métrica  $l_1$  conocida como *norma de Manhattan* y la norma  $l_\infty$ , esta última definida como

$$d_\infty(\mathbf{x}, \mathbf{y}) = \max_{1 \leq i \leq l} w_i |x_i - y_i| \quad (\text{J.18})$$

Basándonos en estas MDs se pueden definir las correspondientes MSs como

$$s_p(\mathbf{x}, \mathbf{y}) = d_{\max} - d_p(\mathbf{x}, \mathbf{y}) \quad (\text{J.19})$$

donde  $d_{\max}$  denota el máximo valor de  $d_p$  entre todos los pares de elementos de  $X$ .

A.2) Algunas métricas adicionales son las siguientes (Spath 1980)

$$d_G(\mathbf{x}, \mathbf{y}) = -\log \left( 1 - \frac{1}{l} \sum_{j=1}^l \frac{|x_j - y_j|}{b_j - a_j} \right) \quad (\text{J.20})$$

donde  $b_j$  y  $a_j$  son, respectivamente, los valores máximo y mínimo entre las  $j$ -ésimas componentes de los  $N$  vectores de  $X$ . Obsérvese que los valores de  $d_G$  no solamente dependen de  $\mathbf{x}$  y  $\mathbf{y}$  sino también del total de  $X$ . Entonces si  $d_G(\mathbf{x}, \mathbf{y})$  es la distancia entre dos vectores  $\mathbf{x}$  e  $\mathbf{y}$  que pertenecen a un conjunto  $X$  y  $d'_G(\mathbf{x}, \mathbf{y})$  es la distancia entre los mismos dos vectores cuando pertenecen a un conjunto diferente  $X'$ , entonces en general  $d_G(\mathbf{x}, \mathbf{y}) \neq d'_G(\mathbf{x}, \mathbf{y})$ .

Otra MD es (Spath 1980)

$$d_Q(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{l} \sum_{j=1}^l \left( \frac{x_j - y_j}{x_j + y_j} \right)^2} \quad (\text{J.21})$$

## B) Medidas de similitud

Las medidas de similitud para vectores con componentes reales más utilizadas en la práctica son:

B.1) El *producto escalar*. Se define como  $s_{\text{escalar}}(\mathbf{x}, \mathbf{y}) = \mathbf{x}^t \mathbf{y} = \sum_{i=1}^l x_i y_i$ . En muchos casos el producto escalar se utiliza cuando los vectores  $\mathbf{x}$  e  $\mathbf{y}$  están normalizados, de forma que su módulo sea  $a$ . En estos casos, los límites inferior y superior de  $s_{\text{escalar}}$  son  $+a^2$  y  $-a^2$ , respectivamente, y  $s_{\text{escalar}}$  depende exclusivamente del ángulo entre  $\mathbf{x}$  e  $\mathbf{y}$ . La correspondiente medida de distinción para el producto escalar es  $d_{\text{escalar}}(\mathbf{x}, \mathbf{y}) = b_{\text{max}} - s_{\text{escalar}}$ , donde  $b_{\text{max}}$  denota el máximo valor de  $s_{\text{escalar}}$  entre todos los pares de elementos de  $X$ .

B.2) Otra MS muy utilizada es la *medida de Tanimoto*, que también se conoce como la *distancia de Tanimoto*, que se puede utilizar tanto para vectores con componentes discretas como reales. Se define como,

$$s_T(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^t \mathbf{y}}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \mathbf{x}^t \mathbf{y}} \quad (\text{J.22})$$

Añadiendo y sustrayendo el término  $\mathbf{x}^t \mathbf{y}$  en el denominador de (J.22) y tras algunas manipulaciones algebraicas, obtenemos

$$s_T(\mathbf{x}, \mathbf{y}) = \frac{1}{1 + \frac{(\mathbf{x} - \mathbf{y})^t (\mathbf{x} - \mathbf{y})}{\mathbf{x}^t \mathbf{y}}} \quad (\text{J.23})$$

Esto significa que la medida de Tanimoto entre  $\mathbf{x}$  e  $\mathbf{y}$  es inversamente proporcional a la distancia Euclídea al cuadrado entre  $\mathbf{x}$  e  $\mathbf{y}$  dividida por su producto escalar. Intuitivamente, esto significa que puesto que el producto escalar se puede considerar como una medida de la correlación entre  $\mathbf{x}$  e  $\mathbf{y}$ ,  $s_T(\mathbf{x}, \mathbf{y})$  es inversamente proporcional a la distancia Euclídea al cuadrado entre  $\mathbf{x}$  e  $\mathbf{y}$ , dividida por su correlación.

En el supuesto de que los vectores hayan sido normalizados para tener el mismo módulo  $a$ , la última ecuación resulta ser,

$$s_T(\mathbf{x}, \mathbf{y}) = \frac{1}{-1 + \frac{2a^2}{\mathbf{x}^t \mathbf{y}}} \quad (\text{J.24})$$

En este caso,  $s_T$  es inversamente proporcional a  $a^2 / \mathbf{x}^t \mathbf{y}$ . Por tanto, cuanto mayor sea la correlación entre  $\mathbf{x}$  e  $\mathbf{y}$  mayor es el valor de  $s_T$ .

B.3) Finalmente otra medida de similitud que se ha probado satisfactoriamente en algunas aplicaciones (Fu y col. 1993) es la siguiente

$$s_c(\mathbf{x}, \mathbf{y}) = 1 - \frac{d_2(\mathbf{x}, \mathbf{y})}{\|\mathbf{x}\| + \|\mathbf{y}\|} \quad (\text{J.25})$$

$s_c(\mathbf{x}, \mathbf{y})$  toma su valor máximo (1) cuando  $\mathbf{x} = \mathbf{y}$  mientras que su valor mínimo (0) lo toma cuando  $\mathbf{x} = -\mathbf{y}$

## VECTORES DISCRETOS

Consideramos ahora los vectores  $\mathbf{x}$  cuyas coordenadas pertenecen al conjunto finito  $F = \{0, 1, \dots, k-1\}$ , donde  $k$  es un entero positivo. Es claro que hay exactamente  $k^l$  vectores  $\mathbf{x} \in F^l$ . Consideremos  $\mathbf{x}, \mathbf{y} \in F^l$  y sea  $A$  la matriz de dimensión  $k \times k$  tal que

$$A(\mathbf{x}, \mathbf{y}) = [a_{ij}] \quad i, j = 0, 1, \dots, k-1 \quad (\text{J.26})$$

donde el elemento  $a_{ij}$  es el número de lugares donde el primer vector tiene el símbolo  $i$  y el correspondiente elemento del segundo vector tiene el símbolo  $j$ ,  $i, j \in F$ . Esta matriz se conoce como una *tabla de contingencia*. Por ejemplo, si  $l = 6$ ,  $k = 3$  y  $\mathbf{x} = [0, 1, 2, 1, 2, 1]^t$ ,  $\mathbf{y} = [1, 0, 2, 1, 0, 1]^t$ , entonces la matriz  $A(\mathbf{x}, \mathbf{y})$  es igual a,

$$A(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix} \quad (\text{J.27})$$

Siendo fácil verificar que

$$\sum_{i=0}^{k-1} \sum_{j=0}^{k-1} a_{ij} = l \quad (\text{J.28})$$

Muchas medidas de proximidad entre dos vectores con componentes discretas pueden expresarse como combinaciones de los elementos de la matriz  $A(\mathbf{x}, \mathbf{y})$ .

### A) Medidas de distinción

#### A.1) Distancia de Hamming (Gersho y Gray 1992)

Se define como el número de lugares en los que difieren los dos vectores  $\mathbf{x}, \mathbf{y}$ . Utilizando la matriz  $A$ , se puede definir la distancia de Hamming  $d_H(\mathbf{x}, \mathbf{y})$  como

$$d_H(\mathbf{x}, \mathbf{y}) = \sum_{i=0}^{k-1} \sum_{j=0, j \neq i}^{k-1} a_{ij} \quad (\text{J.29})$$

es decir, se trata de la suma de todos los elementos fuera de la diagonal principal de  $A$ , que indican las posiciones donde  $\mathbf{x}$  e  $\mathbf{y}$  difieren.

En el caso especial en el que  $k = 2$ , los vectores  $\mathbf{x} \in F^l$  toman valores binarios y la distancia de Hamming es,

$$d_H(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^l (x_i + y_i - 2x_i y_i) = \sum_{i=1}^l (x_i - y_i)^2 \quad (\text{J.30})$$

En el caso en el que  $\mathbf{x} \in F_1^l$ , donde  $F_1 = \{-1, 1\}$ ,  $\mathbf{x}$  se dice que es un vector bipolar y la distancia de Hamming está dada por

$$d_H(\mathbf{x}, \mathbf{y}) = 0.5 \left( l - \sum_{i=1}^l x_i y_i \right) \quad (\text{J.31})$$

Obviamente, la medida correspondiente de similitud de  $d_H$  es  $s_H(\mathbf{x}, \mathbf{y}) = b_{\max} - d_H(\mathbf{x}, \mathbf{y})$ .

## A.2) Distancia $l_1$ .

Se define como en el caso de vectores con componentes continuas, esto es,

$$d_1(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^l |x_i - y_i| \quad (\text{J.32})$$

La distancia  $l_1$  y la distancia de Hamming coinciden cuando se consideran vectores binarios.

## B) Medidas de similitud

Una medida de similitud para vectores discretos es la medida de *Tanimoto*, que se inspira en la comparación de conjuntos. Si  $X$  e  $Y$  son dos conjuntos y  $n_X$ ,  $n_Y$ ,  $n_{X \cap Y}$  son las cardinalidades (número de elementos) de  $X$ ,  $Y$ , y  $X \cap Y$  respectivamente, la medida de Tanimoto entre dos conjuntos  $X$  e  $Y$  se define como

$$\frac{n_{X \cap Y}}{n_X + n_Y - n_{X \cap Y}} = \frac{n_{X \cap Y}}{n_{X \cup Y}} \quad (\text{J.33})$$



En otras palabras, la medida de Tanimoto entre dos conjuntos es la razón entre el número de elementos que tienen en común y el número de todos los elementos diferentes.

Veamos ahora la medida de Tanimoto entre dos vectores discretos  $\mathbf{x}$  e  $\mathbf{y}$ . La medida tiene en cuenta todos los pares de las correspondientes coordenadas de  $\mathbf{x}$  e  $\mathbf{y}$ , excepto aquellos cuyas coordenadas  $(x_i, y_i)$  son ambas 0. Ahora se define  $n_x = \sum_{i=1}^{k-1} \sum_{j=0}^{k-1} a_{ij}$  y  $n_y = \sum_{i=0}^{k-1} \sum_{j=1}^{k-1} a_{ij}$ , donde  $a_{ij}$  son los elementos de la matriz  $A(\mathbf{x}, \mathbf{y})$ . Dicho con palabras,  $n_x$  y  $n_y$  indican el número de coordenadas distintas de cero de  $\mathbf{x}$  e  $\mathbf{y}$ , respectivamente. Por tanto, la medida de Tanimoto se define como,

$$s_T(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{k-1} a_{ii}}{n_x + n_y - \sum_{i=1}^{k-1} \sum_{j=1}^{k-1} a_{ij}} \quad (\text{J.34})$$

En el caso especial en el que  $k = 2$ , esta ecuación resulta,

$$s_T(\mathbf{x}, \mathbf{y}) = \frac{a_{11}}{a_{11} + a_{01} + a_{10}} \quad (\text{J.35})$$

Se pueden definir otras funciones de similitud entre  $\mathbf{x}, \mathbf{y} \in F^l$  utilizando los elementos de  $A(\mathbf{x}, \mathbf{y})$ . Algunas de ellas consideran sólo el número de lugares donde los dos vectores coinciden y el correspondiente valor es distinto de cero, mientras otras consideran todos los lugares donde los dos vectores coinciden. A la primera categoría pertenecen,

$$\frac{\sum_{i=1}^{k-1} a_{ii}}{l} \quad \text{y} \quad \frac{\sum_{i=1}^{k-1} a_{ii}}{l - a_{00}} \quad (\text{J.36})$$

Una medida representativa de la segunda categoría es,

$$\frac{\sum_{i=0}^{k-1} a_{ii}}{l} \quad (\text{J.37})$$

## MEDIDAS DE SIMILITUD DINÁMICAS

Las medidas de proximidad se aplican a vectores con la misma dimensión. No obstante, en ciertas aplicaciones, tales como las medidas de similitud en los métodos estructurales de reconocimiento de patrones propuestas en la sección 15.2.2 no se da

esta circunstancia. Con tal propósito se utilizan medidas de similitud dinámicas, tales como la distancia de *Levenshtein*.

## VECTORES MIXTOS

Un caso interesante surge en la práctica cuando los valores del vector de características no son todos reales o todos discretos.

En este caso se puede utilizar por ejemplo la distancia  $l_l$ .

Sin embargo, otro método consiste en convertir los valores reales a valores discretos. Si una característica  $x_i$  toma valores en el intervalo  $[a, b]$ , podemos dividir este intervalo en  $k$  subintervalos. Si el valor de  $x_i$  cae en el  $r$ -ésimo subintervalo, se le asignará el valor  $r - 1$ . Una vez hecha la discretización se puede utilizar cualquier medida para vectores discretos.

Una función de similitud que trata con vectores mixtos es la propuesta por Gower (1971). Consideremos dos vectores mixtos  $\mathbf{x}_i$  y  $\mathbf{x}_j$  de dimensión  $l$ . Entonces, la función de similitud entre ambos se define como,

$$s(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{q=1}^l s_q(\mathbf{x}_i, \mathbf{x}_j)}{\sum_{q=1}^l w_q} \quad (\text{J.38})$$

donde  $s(\mathbf{x}_i, \mathbf{x}_j)$  es la similitud entre las coordenadas  $q$ -ésimas de  $\mathbf{x}_i$  y  $\mathbf{x}_j$  y  $w_q$  es un factor de peso correspondiente a la coordenada  $q$ -ésima. Específicamente, si al menos una de las coordenadas  $q$ -ésimas de  $\mathbf{x}_i$  y  $\mathbf{x}_j$  no está definida, entonces  $w_q = 0$ . También si la coordenada  $q$ -ésima es una variable binaria y es cero para ambos vectores, entonces  $w_q = 0$ . En el resto de los casos,  $w_q$  es 1. Finalmente, si todos los  $w_q$  son nulos entonces  $s(\mathbf{x}_i, \mathbf{x}_j)$  está indefinida. Si las  $q$ -ésimas coordenadas de los dos vectores son binarias entonces,

$$s_q(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 1, & \text{si } x_{iq} = x_{jq} = 1 \\ 0, & \text{en otro caso} \end{cases} \quad (\text{J.39})$$

Si las  $q$ -ésimas coordenadas de los dos vectores corresponden a variables nominales u ordinales (ver sección J.1), entonces  $s_q(\mathbf{x}_i, \mathbf{x}_j) = 1$  si  $x_{iq}$  y  $x_{jq}$  tienen los mismos valores, de otro modo  $s_q(\mathbf{x}_i, \mathbf{x}_j) = 0$ . Finalmente, si las  $q$ -ésimas coordenadas corresponden a variables intervalo o relación escaladas, entonces,

$$s_q(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{|x_{iq} - x_{jq}|}{r_q} \tag{J.40}$$

donde  $r_q$  es la longitud del intervalo donde caen los valores de las coordenadas  $q$ -ésimas. Se deduce fácilmente a partir de (J.40) que cuando  $x_{iq}$  y  $x_{jq}$  coinciden,  $s_q$  toma su valor máximo, es decir 1, y cuando el valor absoluto de la diferencia es igual a  $r_q$  entonces toma su valor mínimo.

Por ejemplo, consideremos los siguientes cuatro vectores de dimensión 5, cada uno representando una empresa. Las tres primeras coordenadas son las ventas en miles de millones de pesetas de los tres últimos años, la cuarta si hay o no una actividad en el extranjero y la quinta corresponde al número de empleados de la empresa. La última propiedad es ordinal y toma los valores 0 (número pequeño de empleados), 1 (número medio de empleados), y 2 (número grande de empleados). Los cuatro vectores son los siguientes:

Empresa	ventas 1	ventas 2	ventas 3	Activ. externa	Empleados
1( $\mathbf{x}_1$ )	1.2	1.5	1.9	0	1
2( $\mathbf{x}_1$ )	0.3	0.4	0.6	0	0
3( $\mathbf{x}_1$ )	10	13	15	1	2
4( $\mathbf{x}_1$ )	6	6	7	1	1

Para las tres primeras coordenadas que son relación-escaladas tenemos,  $r_1 = 9.7$ ,  $r_2 = 12.6$  y  $r_3 = 14.4$ . La similitud entre los dos primeros vectores es,

$$s_1(\mathbf{x}_1, \mathbf{x}_2) = 1 - |1.2 - 0.3|/9.7 = 0.9072; \quad s_2(\mathbf{x}_1, \mathbf{x}_2) = 1 - |1.5 - 0.4|/12.6 = 0.9127; \\ s_3(\mathbf{x}_1, \mathbf{x}_2) = 1 - |1.9 - 0.6|/14.4 = 0.9097; \quad s_4(\mathbf{x}_1, \mathbf{x}_2) = 0; \quad s_5(\mathbf{x}_1, \mathbf{x}_2) = 0$$

También  $w_4$  es cero, mientras el resto de factores es 1. Utilizando (J.38) obtenemos,  $s(\mathbf{x}_1, \mathbf{x}_2) = 0.6824$ . Del mismo modo se procedería para obtener  $s(\mathbf{x}_1, \mathbf{x}_3) = 0.0541$ ,  $s(\mathbf{x}_1, \mathbf{x}_4) = 0.5588$ ,  $s(\mathbf{x}_2, \mathbf{x}_3) = 0$ ,  $s(\mathbf{x}_2, \mathbf{x}_4) = 0.3047$ ,  $s(\mathbf{x}_3, \mathbf{x}_4) = 0.4953$ .

## AUSENCIA DE DATOS

Un problema común que surge en aplicaciones de la vida real es la ausencia de datos, esto significa que para algunos vectores de atributos no se conocen todas sus componentes, por ejemplo como consecuencia del fallo de algún dispositivo. A continuación proporcionamos algunas técnicas para manejar esta situación (Dixon 1979, Jain y Dubes 1988):

- 1) Descartar todos los vectores a los que les faltan atributos, esto sólo es viable cuando el número de vectores con ausencia de datos es pequeño comparado con el total de vectores.
- 2) Para la  $i$ -ésima componente encontrar el valor medio con todos los vectores disponibles y sustituir este valor en los vectores a los que les falte esta componente.
- 3) Para todos los pares de componentes  $x_i$  e  $y_i$  de los vectores  $\mathbf{x}$  e  $\mathbf{y}$  definir  $b_i$  como,

$$b_i = \begin{cases} 0, & \text{si tanto } x_i \text{ como } y_i \text{ no faltan} \\ 1, & \text{en caso contrario} \end{cases} \quad (\text{J.41})$$

Entonces, la proximidad entre  $\mathbf{x}$  e  $\mathbf{y}$  se define como

$$\wp(\mathbf{x}, \mathbf{y}) = \frac{l}{l - \sum_{i=1}^l b_i} \sum_{\forall i: b_i=0} \phi(x_i, y_i) \quad (\text{J.42})$$

donde  $\phi(x_i, y_i)$  indica la proximidad entre los dos escalares  $x_i$  e  $y_i$ . Una buena elección resulta ser  $\phi(x_i, y_i) = |x_i - y_i|$ .

## MEDIDAS DIFUSAS

Vamos a considerar vectores  $\mathbf{x}$  e  $\mathbf{y}$  cuyas componentes  $x_i$  e  $y_i$  son reales y pertenecen al intervalo  $[0,1]$ ,  $i = 1, \dots, l$ . Estas componentes expresan el grado de pertenencia a un determinado evento. Por ejemplo, en la teoría del color sabemos que cada componente RGB toma valores en el intervalo  $[0,1]$  cuando dichos valores se restringen a dicho intervalo. Si suponemos por ejemplo que  $G = B = 0$ , podemos decir que si  $R = 0.8$  el color es fundamentalmente rojo y si es  $R = 0.2$  el color es un poco rojo, es decir en función de dicho valor afirmamos que es más o menos rojo. Por contra, si  $R = 1$ , diríamos que el color es rojo puro y si  $R = 0$  el color no es rojo, en este último caso estamos ante un caso de lógica binaria donde  $R$  toma los valores 0 ó 1, mientras que en

el primer caso hablaríamos de lógica difusa. La lógica binaria es una caso particular de la lógica difusa.

Vamos a definir la similitud entre dos variables reales en  $[0,1]$ , como una generalización de la equivalencia entre dos variables binarias (Zadeh 1973). La equivalencia de dos variables binarias  $a$  y  $b$  viene dada por la siguiente relación:

$$(a \equiv b) = (\bar{a} \wedge \bar{b}) \vee (a \wedge b) \quad (\text{J.43})$$

donde  $\wedge$  y  $\vee$  son los operadores lógicos “AND” y “OR” y la barra es la negación. En efecto, si  $a = b = 0(1)$ , el primer (segundo) argumento del operador OR es 1. En el otro caso si  $a = 0(1)$  y  $b = 0(1)$ , entonces ninguno de los argumentos del operador OR llega a ser 1.

Una observación interesante es que el operador AND(OR) entre dos variables binarias puede verse como el operador  $\min(\max)$  entre ellos. También la negación de una variable binaria  $a$  puede escribirse como  $1 - a$ . En el contexto de lógica difusa y basándonos en esta observación los operadores AND, OR y negación se reemplazan por  $\min$ ,  $\max$  y  $1 - a$  (Klir y Yuan 1995). Esto sugiere que el grado de similitud entre dos variables reales  $x_i$  e  $y_i$  en  $[0,1]$  puede definirse como,

$$s(x_i, y_i) = \max(\min(1 - x_i, 1 - y_i), \min(x_i, y_i)) \quad (\text{J.44})$$

Basándonos en este concepto se puede generalizar la similitud para el caso de vectores  $\mathbf{x}$  e  $\mathbf{y}$  de dimensión  $l$  como sigue,

$$s_F^r(\mathbf{x}, \mathbf{y}) = \left( \sum_{i=1}^l s(x_i, y_i)^r \right)^{1/r} \quad (\text{J.45})$$

Resulta fácil verificar que los valores máximo y mínimo de  $s_F$  son  $l^{1/r}$  y  $0.5l^{1/r}$ . A medida que  $r \rightarrow +\infty$ , se obtiene que  $s_F(\mathbf{x}, \mathbf{y}) = \max_{1 \leq i \leq l} s(x_i, y_i)$ . También sucede que cuando  $r = 1$  entonces,  $s_F(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^l s(x_i, y_i)$ .

### J.3 MEDIDAS DE PROXIMIDAD ENTRE UN PUNTO Y UN CONJUNTO

En muchos problemas de agrupamientos, un vector se asigna a un grupo  $\mathbf{R}$  teniendo en cuenta la proximidad entre  $\mathbf{x}$  y  $\mathbf{R}$ ,  $\wp(\mathbf{x}, \mathbf{R})$ . Existen dos orientaciones para

la definición de  $\wp(\mathbf{x}, \mathbf{R})$ . La primera es que todos los puntos de  $\mathbf{R}$  contribuyen a la definición de  $\wp(\mathbf{x}, \mathbf{R})$  y la segunda es que sólo contribuyen los puntos representativos.

Ejemplos típicos del primer caso son:

*Función de proximidad máxima:*

$$\wp_{\max}(\mathbf{x}, \mathbf{R}) = \max_{y \in \mathbf{R}} \wp(\mathbf{x}, y) \quad (\text{J.46})$$

*Función de proximidad mínima:*

$$\wp_{\min}(\mathbf{x}, \mathbf{R}) = \min_{y \in \mathbf{R}} \wp(\mathbf{x}, y) \quad (\text{J.47})$$

*Función de proximidad media:*

$$\wp_{\text{med}}(\mathbf{x}, \mathbf{R}) = \frac{1}{n_{\mathbf{R}}} \sum_{y \in \mathbf{R}} \wp(\mathbf{x}, y) \quad (\text{J.48})$$

donde  $n_{\mathbf{R}}$  es la cardinalidad de  $\mathbf{R}$ .

En las definiciones anteriores,  $\wp(\mathbf{x}, y)$  puede ser cualquier medida de similitud entre dos puntos  $\mathbf{x}$  e  $y$ .

Ejemplos típicos del segundo caso incluyen:

La proximidad entre  $\mathbf{x}$  y  $\mathbf{R}$  se mide como la proximidad entre  $\mathbf{x}$  y un representante de  $\mathbf{R}$ . Se han considerado muchos representantes de  $\mathbf{R}$ , entre ellos los más habituales son los puntos para agrupaciones compactas, hiperplanos para agrupaciones lineales e hiperesferas para agrupaciones esféricas, ver figura J.1

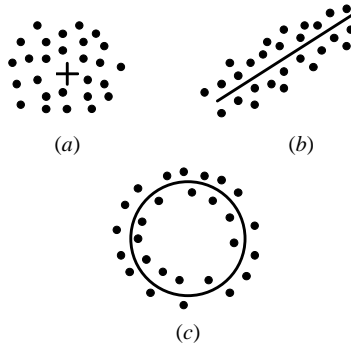


Figura J.1 (a) Agrupación compacta; (b) hiperplana; (c) esférica

## Puntos representativos

Elecciones típicas de puntos representativos de un grupo son:

El *punto medio o vector medio*

$$\mathbf{m}_p = \frac{1}{n_{\mathbf{R}}} \sum_{\mathbf{y} \in \mathbf{R}} \mathbf{y} \quad (\text{J.49})$$

El *centro medio*  $\mathbf{m}_{\mathbf{R}} \in \mathbf{R}$

$$\sum_{\mathbf{y} \in \mathbf{R}} d(\mathbf{m}_{\mathbf{R}}, \mathbf{y}) \leq \sum_{\mathbf{y} \in \mathbf{R}} d(\mathbf{z}, \mathbf{y}), \quad \forall \mathbf{z} \in \mathbf{R} \quad (\text{J.50})$$

donde  $d$  es una medida de no similitud entre dos puntos. Cuando se utilizan medidas de similitud se invierte el orden de la desigualdad.

El *centro mediano*  $\mathbf{m}_m \in \mathbf{R}$

$$\text{med}(d(\mathbf{m}_m, \mathbf{y}) \mid \mathbf{y} \in \mathbf{R}) \leq \text{med}(d(\mathbf{z}, \mathbf{y}) \mid \mathbf{y} \in \mathbf{R}) \quad \forall \mathbf{z} \in \mathbf{R} \quad (\text{J.51})$$

donde  $d$  es una medida de no similitud entre dos puntos. Aquí  $\text{med}(T)$  es la mediana estadística de  $T$ .

## Hiperplanos representativos

En este caso no se puede utilizar un punto representativo y es necesario recurrir a los hiperplanos (Duda y Hart 1973).

La ecuación general de un hiperplano  $H$  es

$$\sum_{j=1}^l a_j x_j + a_0 = \mathbf{a}^t \mathbf{x} + a_0 = 0 \quad (\text{J.52})$$

donde  $\mathbf{x} = [x_1, \dots, x_l]^t$  y  $\mathbf{a} = [a_1, \dots, a_l]^t$  es el vector de coeficientes de  $H$ . La distancia de un punto  $\mathbf{x}$  desde  $H$  se define como,

$$d(\mathbf{x}, H) = \min_{\mathbf{z} \in H} d(\mathbf{x}, \mathbf{z}) \quad (\text{J.53})$$

En el caso de la distancia Euclídea entre dos puntos y utilizando argumentos geométricos simples, ver figura J.2(a) obtenemos,

$$d(\mathbf{x}, H) = \frac{|\mathbf{a}'\mathbf{x} + a_0|}{\|\mathbf{a}\|} \quad (\text{J.54})$$

donde  $\|\mathbf{a}\| = \sqrt{\sum_{j=1}^l a_j^2}$

### Hiperesferas representativas

Cuando las agrupaciones se manifiestan en forma esférica recurrimos a las hiperesferas (Duda y col. 2000).

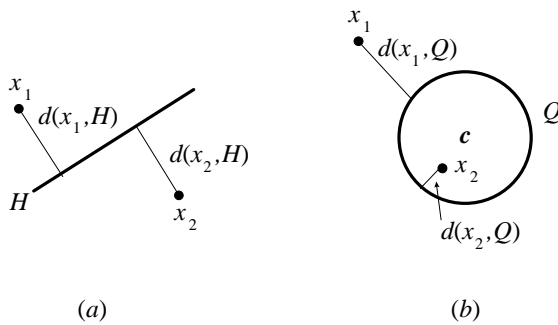
La ecuación general de una hiperesfera  $Q$  es

$$(\mathbf{x} - \mathbf{c})^t (\mathbf{x} - \mathbf{c}) = r^2 \quad (\text{J.55})$$

donde  $\mathbf{c}$  es el centro de la hiperesfera y  $r$  su radio. La distancia de un punto  $\mathbf{x}$  a  $Q$  se define como,

$$d(\mathbf{x}, Q) = \min_{z \in Q} d(\mathbf{x}, z) \quad (\text{J.56})$$

En muchos casos de interés, la distancia Euclídea se utiliza en esta definición. La figura J.2(b) da una idea de esta definición. En la literatura se han utilizado otras distancias no geométricas (Dave y Bhaswan 1992, Krishnapuram y col. 1995 o Frigui y Krishnapuram 1996)



*Figura J.2 (a) Distancia entre un punto y un hiperplano; (b) Distancia entre un punto y una hiperesfera*



## J.4 MEDIDAS DE PROXIMIDAD ENTRE DOS CONJUNTOS

Muchas de las funciones de proximidad utilizadas para la comparación de conjuntos  $\wp^{cc}$  están basadas en medidas de proximidad  $\wp$  entre vectores (Duda y Hart 1973). Dados  $D_i$  y  $D_j$  dos conjuntos de vectores, las funciones de proximidad más comunes son:

La *función de proximidad máxima*

$$\wp_{max}^{cc}(D_i, D_j) = \max_{x \in D_i, y \in D_j} \wp(x, y) \quad (J.57)$$

La *función de proximidad mínima*

$$\wp_{min}^{cc}(D_i, D_j) = \min_{x \in D_i, y \in D_j} \wp(x, y) \quad (J.58)$$

La *función de proximidad promediada*

$$\wp_{prom}^{cc}(D_i, D_j) = \frac{1}{n_{D_i} n_{D_j}} \sum_{x \in D_i} \sum_{y \in D_j} \wp(x, y) \quad (J.59)$$

donde  $n_{D_i}$  y  $n_{D_j}$  son las cardinalidades de  $D_i$  y  $D_j$  respectivamente.

La *función de proximidad media*

$$\wp_{media}^{cc}(D_i, D_j) = \wp(\mathbf{m}_{D_i}, \mathbf{m}_{D_j}) \quad (J.60)$$

donde  $\mathbf{m}_{D_i}$  y  $\mathbf{m}_{D_j}$  son representantes de  $D_i$  y  $D_j$  respectivamente.

Otra función de proximidad basada en la *función de proximidad media* es,

$$\wp_e^{cc}(D_i, D_j) = \sqrt{\frac{n_{D_i} n_{D_j}}{n_{D_i} + n_{D_j}}} \wp(\mathbf{m}_{D_i}, \mathbf{m}_{D_j}) \quad (J.61)$$

donde  $\mathbf{m}_{D_i}$  y  $\mathbf{m}_{D_j}$  se definen como en el caso anterior.

## J.5 COMENTARIOS FINALES

El desarrollo de este apéndice está basado fundamentalmente en Theodoridis y Koutroumbas (1999), además de las referencias mencionadas a lo largo del mismo pueden ser de interés las dos siguientes: Chaudhuri y col. (1992) o más recientemente Santini y Jain (1999). En esta última se introduce una novedad en el sentido de abandonar los axiomas de distancia, sirve para encontrar correspondencias en bases de datos.