Métricas de Similitud

Programa: ————————

11/Septiembre/2025

Convenciones (2×2 de contingencia)

Dados dos vectores binarios o conjuntos X y Y, definimos:

$$
\begin{aligned}
a &= |X \cap Y| && \text{(coincidencias 1–1)} \\
b &= |X \setminus Y| && \text{(1 en } X\text{, 0 en } Y\text{)} \\
c &= |Y \setminus X| && \text{(0 en } X\text{, 1 en } Y\text{)} \\
d &= \text{coincidencias 0–0} = n - (a + b + c) \\
n &= a + b + c + d
\end{aligned}
$$

1. Tabla de contingencia (2×2)

Dados dos vectores binarios X y Y, de longitud n, construimos la tabla de contingencia:

	$Y = 1$	$Y = 0$
$X = 1$	a (coincidencias 1–1)	b (1 en X, 0 en Y)
$X = 0$	c (0 en X, 1 en Y)	d (coincidencias 0–0)

Donde:

$$
\begin{aligned}
a &= \#\{i : X_i = 1 \wedge Y_i = 1\}, \\
b &= \#\{i : X_i = 1 \wedge Y_i = 0\}, \\
c &= \#\{i : X_i = 0 \wedge Y_i = 1\}, \\
d &= \#\{i : X_i = 0 \wedge Y_i = 0\}, \\
n &= a + b + c + d.
\end{aligned}
$$

2. Ejemplo paso a paso

Sea $X = [1, 0, 1, 0, 0]$ y $Y = [1, 0, 0, 0, 0]$. Calculamos:

- Posición 1: $X = 1, Y = 1 \implies a = 1$.
- Posición 2: $X = 0, Y = 0 \implies d = 1$.
- Posición 3: $X = 1, Y = 0 \implies b = 1$.
- Posición 4: $X = 0, Y = 0 \implies d = 2$.
- Posición 5: $X = 0, Y = 0 \implies d = 3$.

Por tanto: $a = 1$, $b = 1$, $c = 0$, $d = 3$, $n = 5$.

1. Métricas *comunes*

- **Jaccard / Tanimoto (similitud)**: $S_J = \dfrac{a}{a+b+c}$.
- **Sørensen–Dice (similitud)**: $S_D = \dfrac{2a}{2a+b+c}$.
- **Simple Matching / Sokal–Michener (similitud)**: $S_{SM} = \dfrac{a+d}{n}$.
- **Hamming (distancia)**: $D_H = \dfrac{b+c}{n}$; **Hamming (similitud)**: $1 - D_H$.
- **Russell–Rao (similitud)**: $S_{RR} = \dfrac{a}{n}$.
- **Hamann (similitud)**: $S_{Ha} = \dfrac{a+d-b-c}{n}$.
- **Gower (binaria simétrica, similitud)**: $S_{Go} = \dfrac{a+d}{n}$.

2. Métricas *menos comunes*

- **Baroni–Urbani & Buser (similitud)**: $S_{BUB} = \dfrac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c}$.
- **Kulczynski (K1, similitud)**: $S_{Ku} = \dfrac{1}{2}\left(\dfrac{a}{a+b} + \dfrac{a}{a+c}\right)$.
- **Rogers–Tanimoto (similitud)**: $S_{RT} = \dfrac{a+d}{a+d+2(b+c)}$.

3. Familia Sokal & Sneath

- **SS1 (similitud)**: $S_{SS1} = \dfrac{a}{a + 2(b + c)}$.
- **SS2 (similitud)**: $S_{SS2} = \dfrac{2(a + d)}{2(a + d) + (b + c)}$.
- **SS4 (similitud, Anderberg)**: $S_{SS4} = \dfrac{1}{4}\left(\dfrac{a}{a+b} + \dfrac{a}{a+c} + \dfrac{d}{b+d} + \dfrac{d}{c+d}\right)$.
- **SS5 (similitud, Gower similarity 2)**: $S_{SS5} = \dfrac{a\,(a+b+c)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$.

Nota. SS3 La variante *SS3* tiene definiciones divergentes en la literatura. Para evitar ambigüedad, no se incluye.

4. Ejemplo numérico

Sea la tabla con $a = 5$, $b = 3$, $c = 2$, $d = 10$ ($n = 20$). En la Tabla 1 se listan los valores de cada métrica.

Métrica	Valor
Jaccard / Tanimoto	0.5000
Sørensen–Dice	0.6667
Simple Matching / Sokal–Michener	0.7500
Hamming (distancia)	0.2500
Hamming (similitud)	0.7500
Russell–Rao	0.2500
Hamann	0.5000
Gower (binaria simétrica)	0.7500
Baroni–Urbani & Buser	0.7071
Kulczynski (K1)	0.6696
Rogers–Tanimoto	0.6000
Sokal–Sneath #1	0.3333
Sokal–Sneath #2	0.8571
Sokal–Sneath #4 / Anderberg	0.7355
Sokal–Sneath #5 / Gower sim. 2	0.5350

Cuadro 1: Valores de similitud/distancia para el ejemplo ($a = 5, b = 3, c = 2, d = 10$).

Consideraciones

- Muchas similitudes se relacionan por transformaciones simples: p.ej., la distancia de Hamming es $1-$ SMC (simple matching), y Hamann es $2\,\text{SMC} - 1$.
- Para variables binarias *asimétricas* (la coausencia d no aporta), se recomienda usar Jaccard/Tanimoto o Sørensen–Dice.
- Para binarias *simétricas* (la coausencia d sí aporta), SMC / Gower binaria y Rogers–Tanimoto son opciones comunes.