Aprendizaje no supervisado

1.2. Medidas de distancia

Javier Sevilla

viu Universidad Internacional de Valencia

20.02.2020

Similitud y disimilitud

Similitud: ¿Cuánto se parecen dos elementos?

Disimilitud: ¿Cuánto se diferencian dos elementos?

Similitud y disimilitud

Disimilitud: ¿Cuánto se diferencian dos elementos?

Distancia: $\sim$ disimilitud, con una serie de condiciones:

- No negatividad:

$$d(a, b) \geq 0, \forall a, b \in \mathbb{R}$$

- Simetricidad:

$$d(a, b) = d(b, a), \forall a, b \in \mathbb{R}$$

- Identidad de los indiscernibles:

$$d(a, b) = 0 \;\Leftrightarrow a = b, \forall a, b \in \mathbb{R}$$

- Desigualdad triangular:

$$d(a, b) \leq d(a, c) + d(c, b), \forall a, b, c \in \mathbb{R}$$

Variables aleatorias:

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_v)$$

Variables aleatorias:

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_v)$$

Valores de las variables aleatorias:

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_v)$$

Similitud y disimilitud

Variables aleatorias:

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_v)$$

Valores de las variables aleatorias:

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_v)$$

- Variable continua, X: valor numérico, $x \in \mathbb{R}$
- Variable categórica, X: valor discreto, $x \in \Omega_X$
  con $\Omega_X = \{A, B, \ldots, C\}$

Similitud y disimilitud

Variables contínuas:

Una única variable

$$d(x_1, x_2) = |x_1 - x_2|$$

Similitud y disimilitud

Variables contínuas:

Varias variables

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{\sum_{j=1}^{v} (x_{1j} - x_{2j})^2} = \sqrt{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^T (\boldsymbol{x}_1 - \boldsymbol{x}_2)}$$

Distancia euclidiana

Variables contínuas:

$$d_p(\boldsymbol{x}_1, \boldsymbol{x}_2) = ||\boldsymbol{x}_1 - \boldsymbol{x}_2||_p = \left(\sum_{j=1}^{v} |x_{1j} - x_{2j}|^p \right)^{(1/p)}$$

Similitud y disimilitud

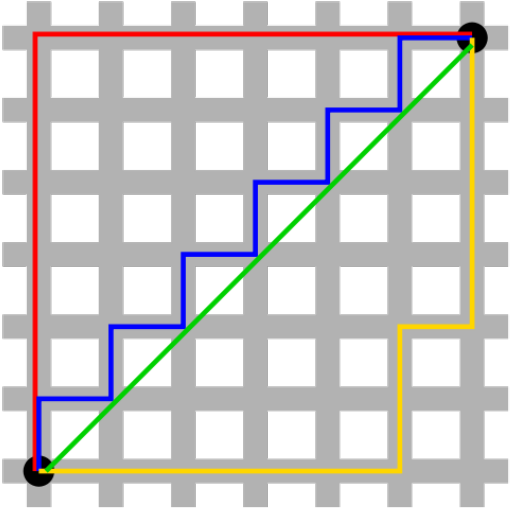
Variables contínuas:

$$d_p(\boldsymbol{x}_1, \boldsymbol{x}_2) = ||\boldsymbol{x}_1 - \boldsymbol{x}_2||_p = \left(\sum_{j=1}^{v} |x_{1j} - x_{2j}|^p \right)^{(1/p)}$$

- Manhattan ($p = 1$):

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{j=1}^{v} |x_{1j} - x_{2j}| = ||\boldsymbol{x}_1 - \boldsymbol{x}_2||_1$$

Similitud y disimilitud



Similitud y disimilitud

Variables contínuas:

$$d_p(\boldsymbol{x}_1, \boldsymbol{x}_2) = ||\boldsymbol{x}_1 - \boldsymbol{x}_2||_p = \left(\sum_{j=1}^{v} |x_{1j} - x_{2j}|^p \right)^{(1/p)}$$

- Manhattan ($p = 1$):

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{j=1}^{v} |x_{1j} - x_{2j}| = ||\boldsymbol{x}_1 - \boldsymbol{x}_2||_1$$

- Euclidiana ($p = 2$):

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{\sum_{j=1}^{v} (x_{1j} - x_{2j})^2} = ||\boldsymbol{x}_1 - \boldsymbol{x}_2||_2$$

Similitud y disimilitud

Variables contínuas:

$$d_p(\boldsymbol{x}_1, \boldsymbol{x}_2) = ||\boldsymbol{x}_1 - \boldsymbol{x}_2||_p = \left(\sum_{j=1}^{v} |x_{1j} - x_{2j}|^p \right)^{(1/p)}$$

- Manhattan ($p = 1$):

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{j=1}^{v} |x_{1j} - x_{2j}| = ||\boldsymbol{x}_1 - \boldsymbol{x}_2||_1$$

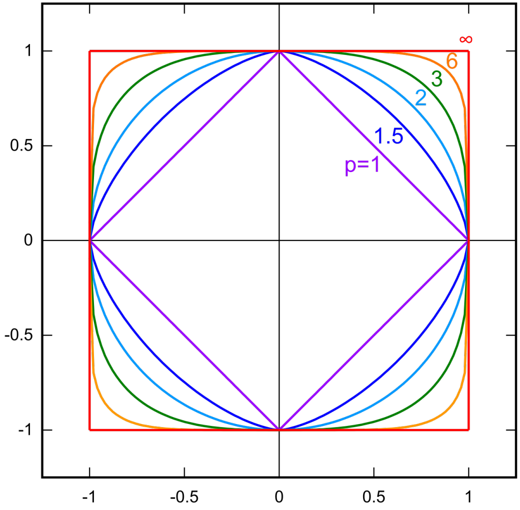
- Euclidiana ($p = 2$):

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{\sum_{j=1}^{v} (x_{1j} - x_{2j})^2} = ||\boldsymbol{x}_1 - \boldsymbol{x}_2||_2$$

- Máximo ($p = \infty$):

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \max_{j \in 1,\ldots,v} |x_{1j} - x_{2j}| = ||\boldsymbol{x}_1 - \boldsymbol{x}_2||_\infty$$

Similitud y disimilitud



Variables contínuas:

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^T \Sigma^{-1} (\boldsymbol{x}_1 - \boldsymbol{x}_2)}$$

Distancia Mahalanobis

Similitud y disimilitud

$$\Sigma = \begin{bmatrix} \mathrm{E}[(X_1-\mu_1)(X_1-\mu_1)] & \mathrm{E}[(X_1-\mu_1)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_1-\mu_1)(X_n-\mu_n)] \\ \mathrm{E}[(X_2-\mu_2)(X_1-\mu_1)] & \mathrm{E}[(X_2-\mu_2)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_2-\mu_2)(X_n-\mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n-\mu_n)(X_1-\mu_1)] & \mathrm{E}[(X_n-\mu_n)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_n-\mu_n)(X_n-\mu_n)] \end{bmatrix}$$

$$\Sigma = \mathrm{E}\left[(\boldsymbol{X} - \mathrm{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathrm{E}[\boldsymbol{X}])^{\mathrm{T}}\right]$$

$$\sigma^2 = \mathrm{var}(X) = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] = \mathrm{E}\left[(X - \mathrm{E}[X])(X - \mathrm{E}[X])\right]$$

$$\mathrm{cov}(X, Y) = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right]$$

Similitud y disimilitud

Variables contínuas:

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^T \Sigma^{-1} (\boldsymbol{x}_1 - \boldsymbol{x}_2)}$$

Distancia Mahalanobis

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{\left. \sum_{j=1}^{v} \frac{x_{1j} - x_{2j}}{\sigma_j} \right)^2} = \sqrt{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^T S^{-1} (\boldsymbol{x}_1 - \boldsymbol{x}_2)}$$

Distancia euclidiana estandarizada

Similitud y disimilitud

Variables contínuas:

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^T \Sigma^{-1} (\boldsymbol{x}_1 - \boldsymbol{x}_2)}$$

Distancia Mahalanobis

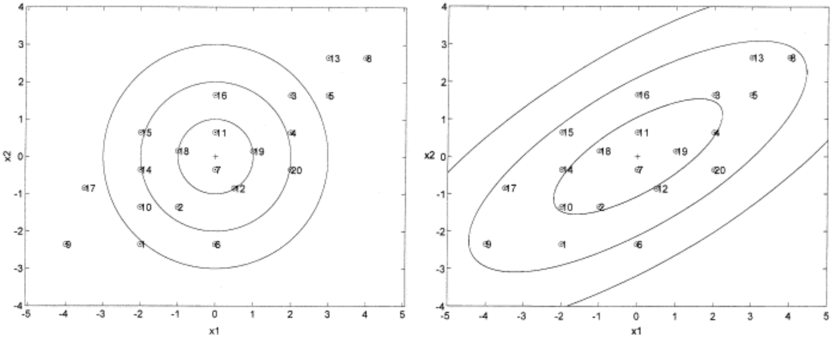
$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{\sum_{j=1}^{v} \left(\frac{x_{1j} - x_{2j}}{\sigma_j} \right)^2} = \sqrt{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^T S^{-1} (\boldsymbol{x}_1 - \boldsymbol{x}_2)}$$

Distancia euclidiana estandarizada

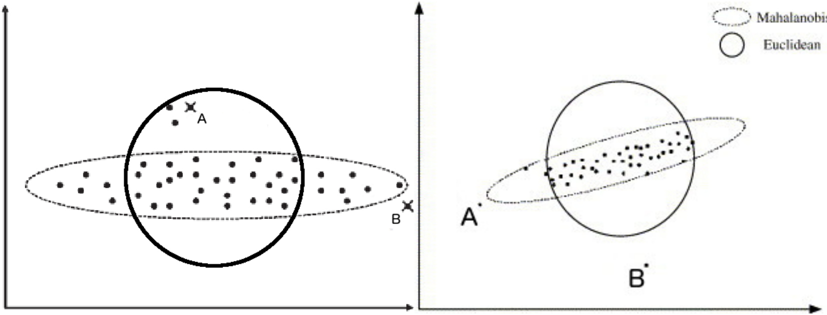
$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{\sum_{j=1}^{v} (x_{1j} - x_{2j})^2} = \sqrt{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^T (\boldsymbol{x}_1 - \boldsymbol{x}_2)}$$

Distancia euclidiana

Similitud y disimilitud



Similitud y disimilitud



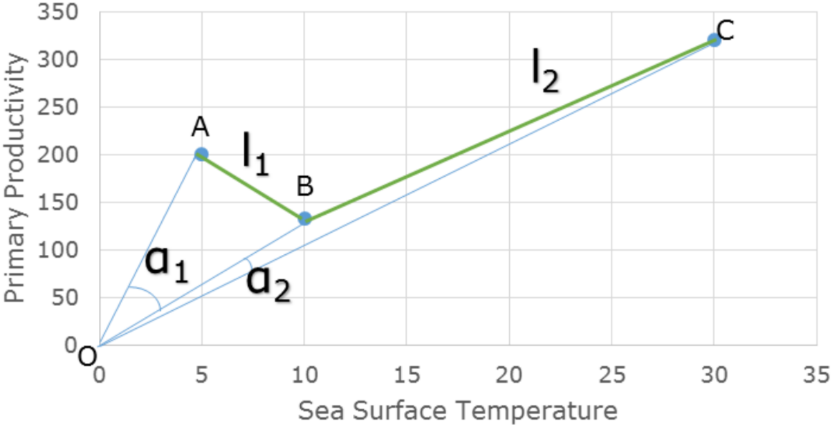
Variable continuas:

$$s(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{\boldsymbol{x}_1, \boldsymbol{x}_2}{||\boldsymbol{x}_1|| \cdot ||\boldsymbol{x}_2||} = \frac{\sum_{j=1}^{v} x_{1j} \cdot x_{2j}}{\sqrt{\sum_{j=1}^{v} x_{1j}^2}\sqrt{\sum_{j=1}^{v} x_{2j}^2}}$$

Similitud coseno

Variable continuas:

Similitud coseno



Similitud y disimilitud

Variable binarias:

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = |x_{1j} = x_{2j}|_{j \in \{1,\ldots,v\}}$$

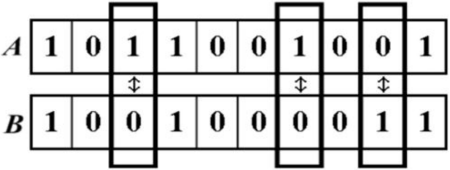
Distancia de Hamming

$$s(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{|x_{1j} = 1 \wedge x_{2j} = 1|_{j \in \{1,\ldots,v\}}}{|x_{1j} = 1 \vee x_{2j} = 1|_{j \in \{1,\ldots,v\}}}$$

Similitud de Jaccard

Similitud y disimilitud

Variable binarias:

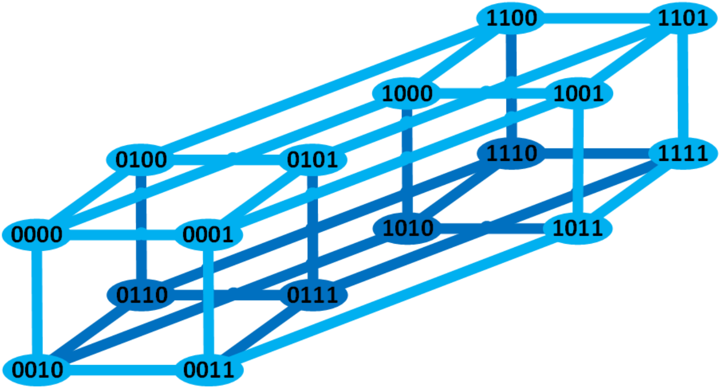


Distancia de Hamming

Similitud y disimilitud

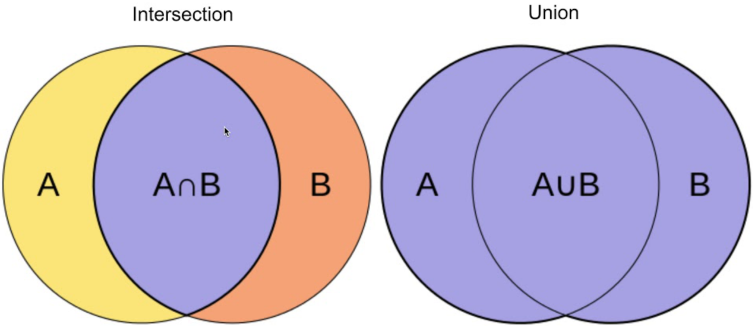
Variable binarias:

0 0 0 0
0 0 0 1
0 0 1 0
0 0 1 1
0 1 0 0
0 1 0 1
0 1 1 0
0 1 1 1
1 0 0 0
1 0 0 1
1 0 1 0
1 0 1 1
1 1 0 0
1 1 0 1
1 1 1 0
1 1 1 1



Distancia de Hamming

Variable binarias:



Similitud de Jaccard

Similitud y disimilitud

Variable categórica:

$$d_j(x_{1j}, x_{2j}) = \begin{cases} 1, & \text{si } x_{1j} \neq x_{2j} \\ 0, & \text{si } x_{1j} = x_{2j} \end{cases}$$

Combinar medidas por variable:

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{j=1}^{v} d_j(x_{1j}, x_{2j})$$

Similitud y disimilitud

Combinar medidas por variable:

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{j=1}^{v} d_j(x_{1j}, x_{2j})$$

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{j=1}^{v} w_j \cdot d_j(x_{1j}, x_{2j}), \text{ con } \sum_{j=1}^{v} w_j = 1$$

Propuesta de Hastie et al. (2008):

$$w_j = 1/\hat{d}_j, \text{ con } \hat{d}_j = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{i'=1}^{n} d_j(x_{ij}, x_{i'j})$$

Similitud y disimilitud

Combinar medidas por variable:

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{j=1}^{v} d_j(x_{1j}, x_{2j})$$

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{j=1}^{v} w_j \cdot d_j(x_{1j}, x_{2j}), \text{ con } \sum_{j=1}^{v} w_j = 1$$

Propuesta de Hastie et al. (2008):

$$w_j = 1/\hat{d}_j, \text{ con } \hat{d}_j = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{i'=1}^{n} d_j(x_{ij}, x_{i'j})$$

Si $d_j(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2$ para todo j, entonces: $w_j = 1/(2\text{var}_j)$

Similitud y disimilitud

Transformar matriz de ejemplos D $(n \times v)$ en...

matriz de distancias, M $(n \times n)$, tal que:

$$M_{ij} = d(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

y ésta, a su vez, en una matriz de similitudes, S $(n \times n)$:

$$S_{ij} = \exp(-M_{ij}^2/c)$$

Gracias

Universidad Internacional de Valencia