Algoritmos de Minería de Datos

Wilmer Gonzalez

6213 - Minería de datos
Facultad de ciencias
Universidad Central de Venezuela

17 de julio de 2024

Contenido

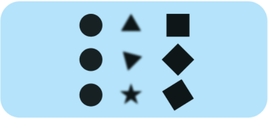
"Notoriamente no hay clasificación del universo que no sea arbitraria y conjetural."[a]

[a] El idioma analítico de John Wilkins - Jorge Luis Borges

Medidas de distancia

Medidas de distancia (Han, Kamber y Pei, 2012)

¿Cómo podemos comprender mejor las relaciones entre las observaciones presentes en los datos?



Medidas de distancia

¿Cómo podemos comprender mejor las relaciones entre las observaciones presentes en los datos?

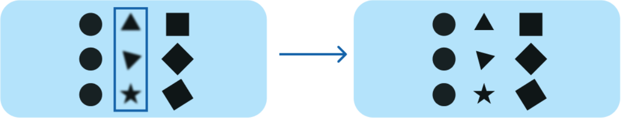


Figura: Preprocesamiento de datos

Medidas de distancia

¿Cómo podemos comprender mejor las relaciones entre las observaciones presentes en los datos?

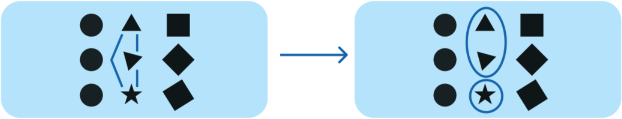
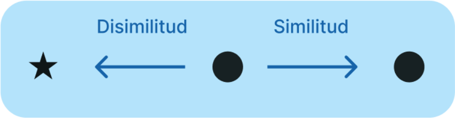


Figura: Medidas de distancia

Medidas de distancia

Objetivo Cuantificar qué tan similares/disimilares son un par de observaciones. $dist \in [0, 1]; sim = 1 - dist$



Medidas de distancia

Datos nominales[1]

Para un grupo de variables de tipo nominales, la distancia entre un par de observaciones puede escribirse, sea:

- p la cantidad de variables
- m la cantidad de co-ocurrencias

$$d(i,j) = \frac{p-m}{p}$$

[1] También pueden representarse cómo un arreglo de variables binarias

Medidas de distancia

Datos binarios

Para un grupo de variables de tipo binario, la distancia entre un par de observaciones puede escribirse, sea p la cantidad de variables:

	Observación i	
Observación j	1	0
1	q	r
0	s	t

Distancia simétrica

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

Distancia asimétrica

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

Similitud asimétrica (Jaccard coef.)

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

Medidas de distancia

Datos numéricos

Para un grupo de variables de tipo nominales, la distancia entre un par de observaciones puede escribirse, sea:

- $x_{i,p}$ el valor de la variable p-ésima de la observación i.
- $x_{j,p}$ el valor de la variable p-ésima de la observación j.

Distancia Minkowski

$$d(i,j) = \sqrt[h]{\sum_{k=1}^{p} |x_{ik} - x_{jk}|^h}$$

Medidas de distancia

Datos numéricos

Distancia Minkowski

$$d(i,j) = \sqrt[h]{\sum_{k=1}^{p} |x_{ik} - x_{jk}|^h}$$

Distancia Manhattan (h=1)

$$d(i,j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}|$$

Distancia Euclideana (h=2)

$$d(i,j) = \sqrt[2]{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

Medidas de distancia

Datos numéricos

Distancia Chebyshev

$$d(i,j) = lim_{h \to \infty} (\sum_{k=1}^{p} |x_{ik} - x_{jk}|^h)^{1/h} = max_k^p |x_{ik} - x_{jk}|$$

Medidas de distancia

Datos numéricos

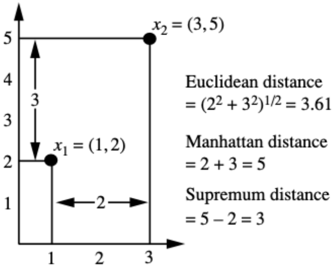


Figura: Ejemplos de distancias numéricas

Medidas de distancia

Datos ordinales

Normalización de rangos

$$z_{ik} = \frac{r_{ik} - 1}{M_k - 1}$$

Luego de este paso se pueden aplicar cualquiera de las distancias numéricas.

Medidas de distancia

Datos especializados (texto) [2]

Similitud del coseno

Sea $||x|| = \sqrt{x_1^2 + \cdots + x_p^2}$

$$sim(x, y) = \frac{x \cdot y}{||x|| ||y||}$$

[2]Ver más (Steck, Ekanadham y Kallus, 2024)

Referencias

Han, Jiawei, Micheline Kamber y Jian Pei (2012). *Data mining concepts and techniques, third edition*.

Steck, Harald, Chaitanya Ekanadham y Nathan Kallus (2024). "Is Cosine-Similarity of Embeddings Really About Similarity?" En.

¡Gracias!

github.com/ucvia/dm