Agrupamiento de imágenes

Clustering 2

Medidas de Proximidad, Disimilaridad, Similaridad, Distancias
(http://www.iiisci.org/journal/CV$/sci/pdfs/GS315JG.pdf)

- Distancias
  - métricas de Minkowski: Manhattan (L1), Euclídea (L2), …
  - distancia Canberra
  - distancia de Mahalanobis
- Ángulos
  - distancia coseno
  - Correlación de Pearson
- Variables binarias:
  - Coeficiente de coincidencias
  - Coeficiente de Jaccard
- Multiestado
- Mixtos
- ...

Matrices de Disimilaridad
Representan relaciones entre los N elementos del conjunto, entonces son matrices de N x N.

Las <u>matrices de disimilaridad</u> (**D**) deben cumplir con varias propiedades:

- La disimilaridad de dos objetos idénticos (o consigo mismo) es cero, **d(a,a) = 0**. Y en general, las disimilaridades solo pueden ser mayores o iguales que cero, **d(a,b) ≥ 0**.

- Los objetos no idénticos pueden ser distinguibles o no.
  Si **a ≠ b** entonces **d(a,b) ≥ 0**.
  Si **a = b** entonces **d(a,b) = 0**.

- Las medidas de disimilaridad son simétricas: **d(a, b) = d(b, a)**.

Matrices de Disimilaridad

Representan relaciones entre los N elementos del conjunto, entonces son matrices de N x N. Las matrices de disimilaridad (**D**) deben cumplir con varias propiedades:

- Si $\mathbf{a \neq b}$ entonces $\mathbf{d(a,b) \geq 0}$.
- Si $\mathbf{a = b}$ entonces $\mathbf{d(a,b) = 0}$ (o si son el mismo objeto entonces $\mathbf{d(a,a) = 0}$)
- $\mathbf{d(a, b) = d(b, a)}$.

  - Si además se cumple la desigualdad triangular:

    $$\mathbf{d(a,b) \leq d(a,c) + d(b,c)}$$

    entonces tenemos una **matriz de disimilaridad métrica**.

  - Cuando la medida de disimilitud no cumple con la desigualdad triangular, pero satisface la desigualdad ultramétrica:

    $$\mathbf{d(a, b) \leq \max\{d(a, c), d(c, b)\}}$$

    entonces tenemos una **matriz de disimilaridad ultramétrica**. Este es el tipo de distancias que ocurren al representar gráficamente un cluster jerárquico con un dendrograma.

Matrices de Similaridad

Una **matriz de similaridad** (S) es aquella donde $s(a,a) = 1$. En algunos casos se puede transformar una matriz de disimilaridad (D) en una de similaridad (S):

- Si el dominio de la similaridad es [0, 1]:

$$\mathbf{d(a, b) = 1 - s(a,b)}$$

- Si el dominio de S es [-1, 1] y s = -1 se corresponde con la mayor distancia normalizada:

$$\mathbf{d(a, b) = 1 - (s(a, b) + 1)/2}$$

Matrices de Proximidad, Similaridad, Disimilaridad y Distancia y Afinidad ...
¡Ojo! ¡La diferencia entre estos términos depende muchas veces del dominio!

En general,

una **matriz de disimilaridad métrica** ~ una **matriz de distancia**

una **matriz de afinidad** ~ una **matriz de similaridad**, con la particularidad de que

afinidad(a,b) ~ exp(- d(a,b)^2) ⇨ si d(a,b) = 0, afinidad(a,b) = 1
si d(a,b) >> 1, afinidad(a,b) ~ 0

Matrices de Similaridad y Disimilaridad para variables BINARIAS

Matrices de Similaridad y Disimilaridad para variables BINARIAS

Supongamos que tenemos dos objetos para los cuales se registran los valores que toman diferentes variables binarias (Verdadero/Falso, 0/1, +/-. etc.):

	var.0	var.1	var.2	var.3	var.4	var.5	var.6	var.7
item.x	1	1	1	1	0	0	0	1
item.y	1	1	0	0	1	0	0	0

Podemos definir estas cantidades:

- **a**: cantidad de variables con **1** tanto para **x** como para **y**
- **b**: cantidad de variables con **1** para **x** y **0** para **y**
- **c**: cantidad de variables con **0** para **x** y **1** para **y**
- **d**: cantidad de variables con **0** tanto para **x** como para **y**

p = a + b + c + d

Matrices de Similaridad y Disimilaridad para variables BINARIAS

Supongamos que tenemos dos objetos para los cuales se registran los valores que toman diferentes variables binarias (Verdadero/Falso, 0/1, +/-. etc.):

	var.0	var.1	var.2	var.3	var.4	var.5	var.6	var.7
item.x	1	1	1	1	0	0	0	1
item.y	1	1	0	0	1	0	0	0

Podemos definir estas cantidades:

a: cantidad de variables con **1** tanto para **x** como para **y** → **a = 2**
b: cantidad de variables con **1** para **x** y **0** para **y** → **b = 3**
c: cantidad de variables con **0** para **x** y **1** para **y** → **c = 1**
d: cantidad de variables con **0** tanto para **x** como para **y** → **d = 2**

p = a + b + c + d → **p = 8**

Matrices de Similaridad y Disimilaridad para variables BINARIAS

	var.0	var.1	var.2	var.3	var.4	var.5	var.6	var.7
item.x	1	1	1	1	0	0	0	1
item.y	1	1	0	0	1	0	0	0

Podemos definir estas cantidades:

a: x ~ 1 & y ~ 1 → **a = 2**
b: x ~ 1 & y ~ 0 → **b = 3**
c: x ~ 0 & y ~ 1 → **c = 1**
d: x ~ 0 & y ~ 0 → **d = 2**
p = a + b + c + d → **p = 8**

Para calcular la (di)similaridad entre **x** e **y** hay diferentes opciones. Dos de las más conocidas son:

- ➔ Coeficiente de coincidencias
  - (similaridad) **s(x,y) = (a + d)/p** (está acotada en [0,1]... ¡probar los límites!)
  - (disimilaridad) **d(x,y) = (b + c)/p** (está acotada en [0,1]... ¡probar los límites!)

- ➔ Coeficiente de Jaccard
  - (similaridad) **s(x,y) = a / (a + b +c)** (está acotada en [0,1]... ¡probar los límites!)
  - (disimilaridad) **d(x,y) = (b + c) / (a + b + c)** (está acotada en [0,1]... ¡probar los límites!)

Matrices de Similaridad y Disimilaridad para variables BINARIAS

	var.0	var.1	var.2	var.3	var.4	var.5	var.6	var.7
item.x	1	1	1	1	0	0	0	1
item.y	1	1	0	0	1	0	0	0

Podemos definir estas cantidades:

a: x ~ 1 & y ~ 1 → **a = 2**
b: x ~ 1 & y ~ 0 → **b = 3**
c: x ~ 0 & y ~ 1 → **c = 1**
d: x ~ 0 & y ~ 0 → **d = 2**
p = a + b + c + d → **p = 8**

El **coeficiente de coincidencias** le asigna importancia a aquellos casos a la coincidencia de valores **1-1**, **V-V**, etc. y también a aquellos donde la coincidencia es **0-0**, **F-F**, etc.

El **coeficiente de Jaccard**, en cambio, sólo considera en el numerador aquellos casos donde la coincidencia es **1-1** o **V-V**.

Dependiendo del problema y dominio de aplicación, una medida puede ser más adecuada que otra.

Matrices de Similaridad y Disimilaridad para variables BINARIAS

	var.0	var.1	var.2	var.3	var.4	var.5	var.6	var.7
item.x	1	1	1	1	0	0	0	1
item.y	1	1	0	0	1	0	0	0

Podemos definir estas cantidades:

a: x ~ 1 & y ~ 1 → **a = 2**
b: x ~ 1 & y ~ 0 → **b = 3**
c: x ~ 0 & y ~ 1 → **c = 1**
d: x ~ 0 & y ~ 0 → **d = 2**
p = a + b + c + d → **p = 8**

¡Existe un gran número de medidas de (di)similaridad desarrolladas para variables binarias! Por ejemplo, *Choi, S. S., Cha, S. H., & Tappert, C. C. (2010). A survey of binary similarity and distance measures. Journal of Systemics, Cybernetics and Informatics, 8(1), 43-48.*
http://www.iiisci.org/journal/CV$/sci/pdfs/GS315JG.pdf

Medidas para variables CONTINUAS

Medidas para variables cuantitativas (continuas)

Métricas de Minkowski:

$$d(x,y) = \left(\sum_{k=1}^{p} w_k^{\lambda} |x_k - y_k|^{\lambda} \right)^{1/\lambda}$$

Distancia Euclídea (*λ*=2):

$$d(x,y) = \sqrt{\sum_{k=1}^{p} |x_k - y_k|^2}$$

Distancia Manhattan (*λ=1*):

$$d(x,y) = \sum_{k=1}^{p} |x_k - y_k|$$

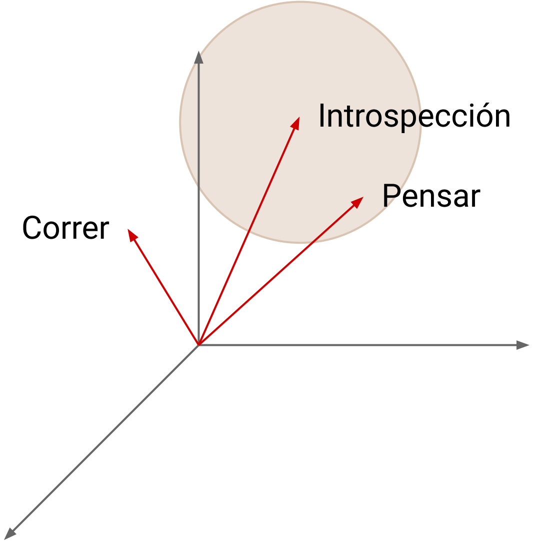
Los valores w_k son pesos que suelen aplicarse para que las variables estén estandarizadas u acotadas en [0,1].

Otras distancias menos utilizadas son la de Canberra o Mahalanobis.

Medidas para comparar los ángulos de los vectores

A veces los valores precisos que toman las variables no son tan importantes en cuanto a cómo afectarán a las distancias entre objetos. El interés, en cambio, se enfoca en la comparación de las direcciones de los vectores que definen a cada objeto en el espacio multidimensional de las variables.

El <u>coeficiente de correlación de Pearson</u> y la <u>similitud coseno</u> son dos medidas de similitud que sirven para comparar la separación angular entre objetos.



Medidas para comparar los ángulos de los vectores

Correlación de Pearson para dos objetos x e y:

$$S(x,y) = \frac{\sum_{k=1}^{p}(x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{p}(x_k - \bar{x})^2 \sum_{k=1}^{p}(y_k - \bar{y})^2}}$$

El coeficiente está acotado en el intervalo [-1,1] y está centrado, es decir que es invariante a desplazamientos.

SPOILER: Ideas parecidas vamos a usar para construir las distancias en el grafo de conexiones entre regiones cerebrales en el TP de la segunda parte de la materia.

Medidas para comparar los ángulos de los vectores

Similitud coseno:

$$s(x,y) = \cos(\phi) = \frac{a \,.\, b}{\|a\| \; \|b\|} = \frac{\sum_{k=1}^{p} x_k y_k}{\sqrt{\sum_{k=1}^{p} x_k^2 \sum_{k=1}^{p} y_k^2}}$$

donde $a \cdot b$ es el producto interno del vector a y el vector b, y $\|a\|$ o $\|b\|$ es la norma cuadrada o norma 2 de los vectores a y b respectivamente.

NOTA: Un ejemplo muy de moda últimamente son las dirección en el espacio semántico de las palabras (a partir de los *word embeddings*).

Medidas para variables MULTIESTADO

Medidas para variables categóricas multiestado y para variables ordinales

Consideramos variables categóricas multiestado a aquellas que presentan dos o más estados o categorías (más que binarias). Si estás categorías presentan algún tipo de ordenamiento, la variable es categórica ordinal, o simplemente ordinal.

Variables categóricas multiestado:

- Si el número de estados es chico una solución sencilla es separar la variable multiestado original en varias variables binarias, una para cada categoría (one-hot encoding) [problema: muchos estados=muchas variables!!].

Medidas para variables categóricas multiestado y para variables ordinales

Consideramos variables categóricas multiestado a aquellas que presentan dos o más estados o categorías (más que binarias). Si estás categorías presentan algún tipo de ordenamiento, la variable es categórica ordinal, o simplemente ordinal.

Variables categóricas multiestado:

- Si el número de estados es chico una solución sencilla es separar la variable multiestado original en varias variables binarias, una para cada categoría (one-hot encoding) [problema: muchos estados=muchas variables!!].
- Otra alternativa es calcular un criterio de coincidencias:

$$S(x,y) = \frac{1}{r}\sum_{l=1}^{r} S_{xyl}$$

Donde la similaridad entre *x* e *y* son objetos multidimensionales de dimensión *r*.

$$S_{xyl}=0 \qquad si\ x_l \neq y_l$$
$$S_{xyl}=w \qquad si\ x_l = y_l$$

w es típicamente 1, aunque si se quieren premiar las coincidencias se puede aumentar.

Medidas para variables ordinales

Se puede considerar que la disimilaridad será proporcional a la cantidad de “saltos” entre estados.
Por ejemplo, suponiendo que se tienen 3 estados: A < B < C, y que la distancia entre dos estados cualquiera es 1 ⇨ podemos decir que A=0, B=1, C=2. La distancia entre dos estados *x* e *y* es:

$$d(x, y) = |x - y|^r$$

donde típicamente $r=1$. Así $d(A,B) = d(B,C) = 1$ y $d(A,C)=2$. Luego, estas medidas suelen ser normalizadas al rango [0, 1], en este caso:

$$d(x, y) = \frac{|x - y|^r}{2}$$

Para muchas variables ordinales, esta medida puede generalizarse a:

$$d(x, y) = \frac{\sum_{k=1}^{p} d_k(x, y)}{p}$$

Medidas para comparar variables de tipos mixtos

En común tener que calcular matrices de distancia a partir de un conjunto de variables con variables de diferentes tipos, binarias, categóricas, continuas. Una forma simple de combinarlas es el *coeficiente de similaridad de Gower*:

$$S(x,y) = \frac{\sum_{k=1}^{p} S_{xyk}\delta_{xyk}}{\sum_{k=1}^{p} \delta_{xyk}}$$

donde S_{xyk} es la similaridad entre x e y para la variable k (en la métrica que corresponda), y δ_{xyk} toma valores 0 o 1 según esta comparación esté presente o no.

El coeficiente de disimilaridad se toma como $1\text{-}S(x,y)$, y muchas veces la noción de distancia se extiende a $d(x,y)^2=1\text{-}S(x,y)$.

Esta distancia en el fondo se calcula tomando el promedio entre las distancias de todas las variables calculadas en cada caso como corresponda.

Medidas para comparar variables de tipos mixtos

En común tener que calcular matrices de distancia a partir de un conjunto de variables con variables de diferentes tipos, binarias, categóricas, continuas. Una forma simple de combinarlas es el *coeficiente de similaridad de Gower*:

$$S(x,y) = \frac{\sum_{k=1}^{p} S_{xyk}\delta_{xyk}}{\sum_{k=1}^{p} \delta_{xyk}}$$

donde S_{xyk} es la similaridad entre x e y para la variable k (en la métrica que corresponda), y δ_{xyk} toma valores 0 o 1 según esta comparación esté presente o no.

Típicamente,

S_{xyk} (continua) → Manhattan
S_{xyk} (binaria) → Dice (DSC = 2*a / (2*a + b +c))

Librerías

Python

sklearn: Todas métricas sobre datos continuos
https://scikit-learn.org/stable/modules/metrics.html#metrics
https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.DistanceMetric.html

scipy: También tiene sobre binarios como Jaccard
https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.pdist.html#scipy.spatial.distance.pdist

gower: En ninguno de los dos hay una implementación para datos mixtos o categóricos como Gower.
https://pypi.org/project/gower/