AGRUPACIÓN
Medidas de Proximidad

Elizabeth León Guzmán

Research Group on Data Mining – MIDAS
Universidad Nacional de Colombia, Bogotá D.C., Colombia

2020

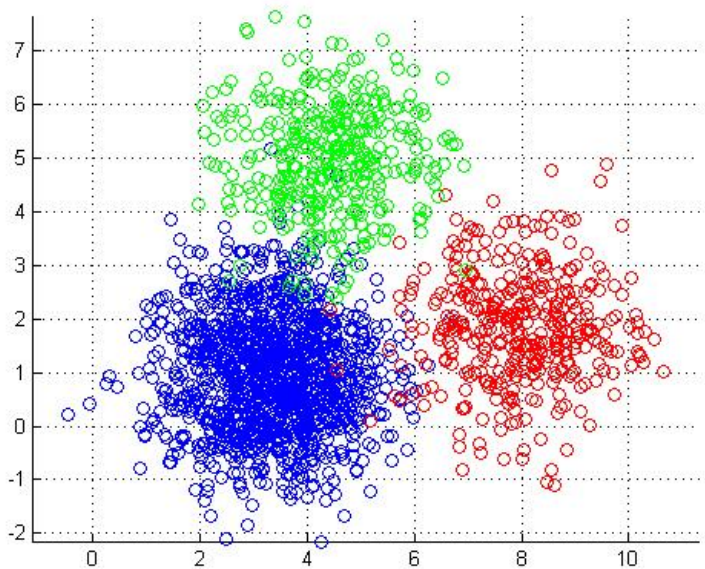
Agenda

Agenda

Definición

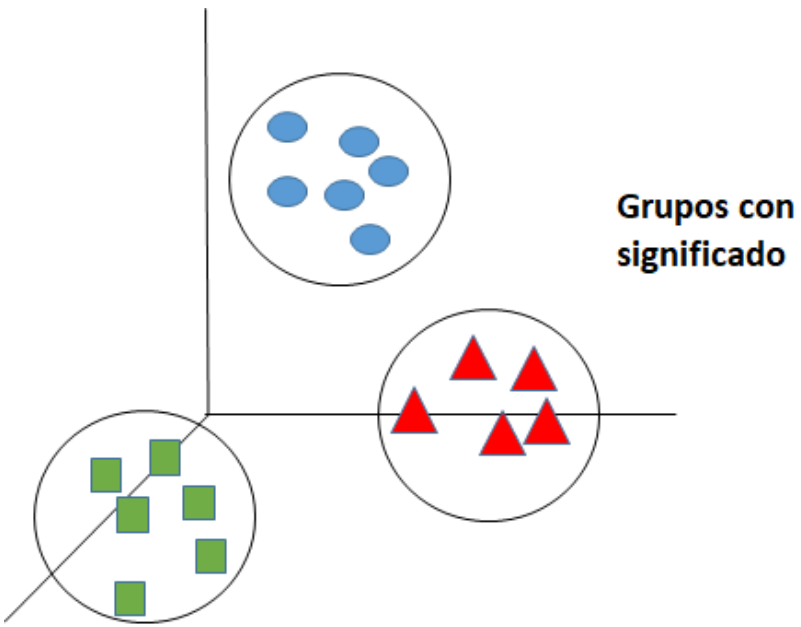
Proceso para

- Dividir los datos en grupos (clusters), de tal forma que los grupos capturan la **estructura natural** de los datos
- Dividir datos en grupos (clusters) de tal forma que datos que pertenecen al mismo grupo son **similares**, y datos que pertenecen a diferentes grupos son diferentes



Definición

- Los grupos con significado (clases) indican como las personas analizan y describen el mundo
- Los humanos tienen la habilidad de dividir los objetos en grupos (agrupación) y asignar objetos particulares a esos grupos (clasificación). Ej: los niños dividen objetos en fotografías: edificios, vehículos, gente, animales, plantas



Definición

El proceso de agrupar tiene en cuenta todas las variables implicadas.

Ejemplo

Agrupar los estudiantes de un curso de acuerdo a la similitud entre ellos.
Variables: edad, genéro, profesión, altura, peso, promedio académico, semestre, ciudad, colegio, asistencia, etc.
"Cada grupo corresponde a un perfil de estudiantes"

Aplicaciones

- **Biologia:** taxonomia (especies), analisis de información genética(grupos de genes que tienen funciones similares)
- **Psicología y Medicina:** Agrupar diferentes tipos de depresión, detectar patrones en la distribución temporal de una enfermedad
- **Recuperación de Información (Information retrieval):** Agrupar resultados de búsquedas en la web (cada grupo contiene aspectos particularres dela consulta) Ej: cine (comentarios, estrellas, teatros)

Aplicaciones

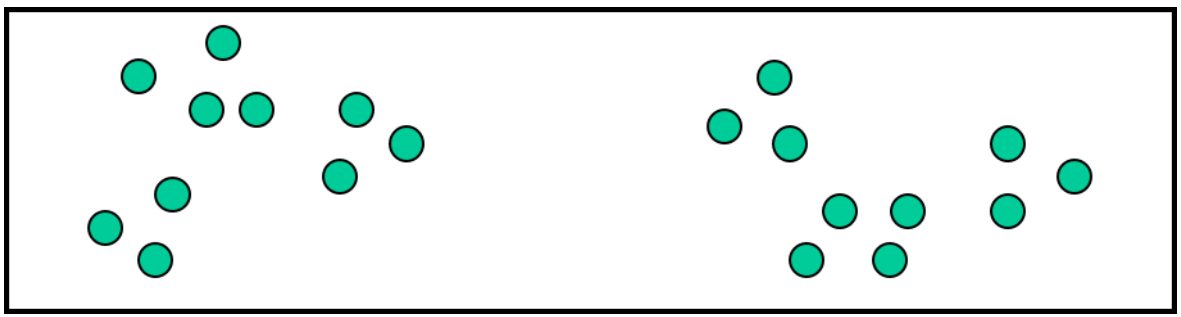
- **Clima:** Encontrando patrones en la atmosfera y oceano. Presión atmosférica de regiones polares y areas de el oceano que tienen un impacto significativo en el clima de la tierra.
- **Negocios:** Segmentar los clientes en grupos para un analisis y actividades de mercadeo

Características

- La arbitrariedad en el número de clusters es el mayor problema en clustering.
- Sistema visual del humano (espacio Euclideano).
- Grupos tienen diferentes formas, tamaños en un espacio n-dimensional
- Definición de cluster es impreciso y la mejor definición depende de la naturaleza de los datos y de los resultados deseados
- Aprendizaje NO supervisado

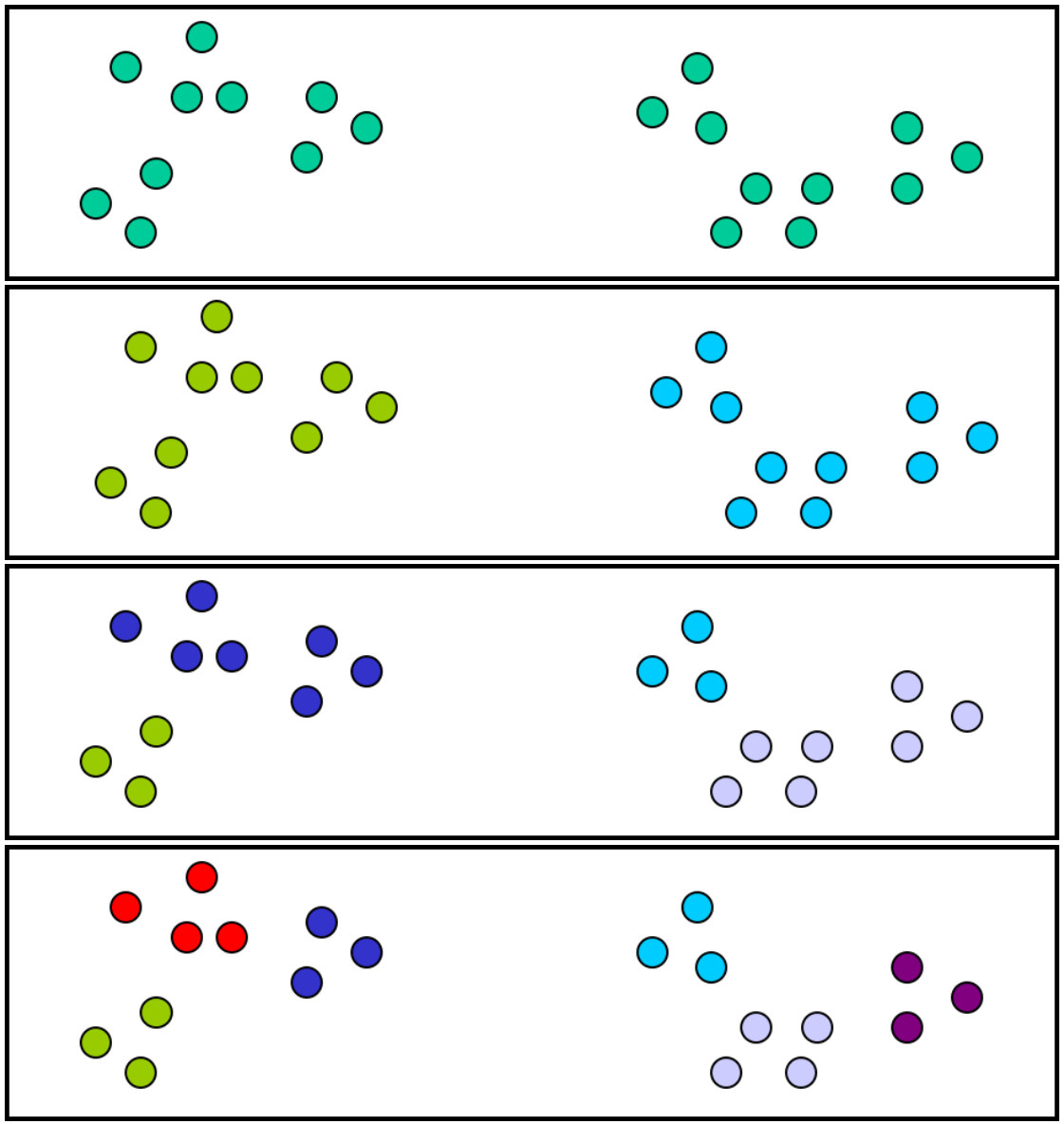
Arbitrariedad en número de "clusters"

¿Cuántos grupos (clusters) se deben encontrar?



Diferentes formas de agrupar el mismo conjunto de datos

Arbitrariedad en número de "clusters"



Medidas de Proximidad

- La medida de similitud (**semejanza**) es fundamental en la definición del cluster
- Debe ser escogida muy cuidadosamente, ya que la calidad de los resultados dependen de ella
- Se puede usar la *disimilaridad* (diferencia, **distancia**)
- Dependen de los tipos de datos

Medidas de Proximidad: similitud - diferencia

Proximidad se refiere a similitud o diferencia

- **Similitud**
  - Medida numérica de semejanza entre objetos
  - Valor alto para objetos parecidos
  - A menudo definida en el intervalo [0,1]
- **Diferencia**
  - Medida numérica de diferencia entre objetos
  - Valor bajo para objetos parecidos
  - Pueden estar en el intervalo [0,1], pero en algunos casos varia entre $[0, \infty)$
  - Usualmente es una distancia

Medidas de Proximidad: similitud - diferencia

Atributos ordinales

Atributo que mide la calidad de un producto:
$\{pobre, regular, aceptable, muybueno, excelente\}$

- El producto P_1 con una calidad *excelente* debe ser más cercano al producto P_2 que tiene calidad de *muy bueno*, que al producto P_3 que tiene calidad *Aceptable*
- Para volverlo cuantitativo, realizar mapeo a números enteros iniciando en 0 o 1, *Ejemplo:* $\{pobre = 0, regular = 1, aceptable = 2, muybueno = 3, excelente = 4\}$
  Entonces, $d(P1, P2) = 4 - 3 = 1$ y $d(P1, P3) = 4 - 2 = 2$
  si se quiere la diferencia este en el intervalo $[0, 1]$,
  $d(P1, P2) = 4 - 3/4 = 0.25$ y $d(P1, P3) = 4 - 2/4 = 0.5$
- La similitud puede ser definida como $s = 1 - d$

Medidas de Proximidad: similitud - diferencia

Se pueden realizar transformaciones para convertir una similitd en una distancia o viceversa. Algunas veces los algoritmos de minería exigen usar un espacio determinado.

Ejemplo

Las similitudes entre los objetos estan en el rango $[1, 10]$, donde 10 indica que dos objetos son completamente similares, y 1 indica que tan diferentes pueden llegar a ser. Este rabgo se puede llevar a $[0, 1]$ usando una transformación simple como $s' = (s - 1)/9$

En general, la transformación simple es *max-min* donde el máximo es 1 y el mínimo es 0:

$$s' = (s - min_s)/(max_s - min_s)$$

Medidas de Proximidad: similitud - diferencia

Si la medida de proximidad se encuentra en el intervalo $[0, \infty)$, se necesitará una transformación *nolineal* y los valores pueden no tener la misma relación en la nueva escala.

Ejemplo

Consideremos la siguiente transformación: $d' = d/(1+d)$ para una medida de distancia que esta en el intervalo $[0, \infty)$:

d	0	0.5	2	10	100	1000
d'	0	0.33	0.67	0.90	0.99	0.999

Valores grandes en la escala de la distancia original son comprimidos en el rango de valores cerca de 1. Si este comportamiento es deseable o no, depende de la aplicación a realizar con los datos.

Similitud - diferencia (disimilitud) para atributos simples

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

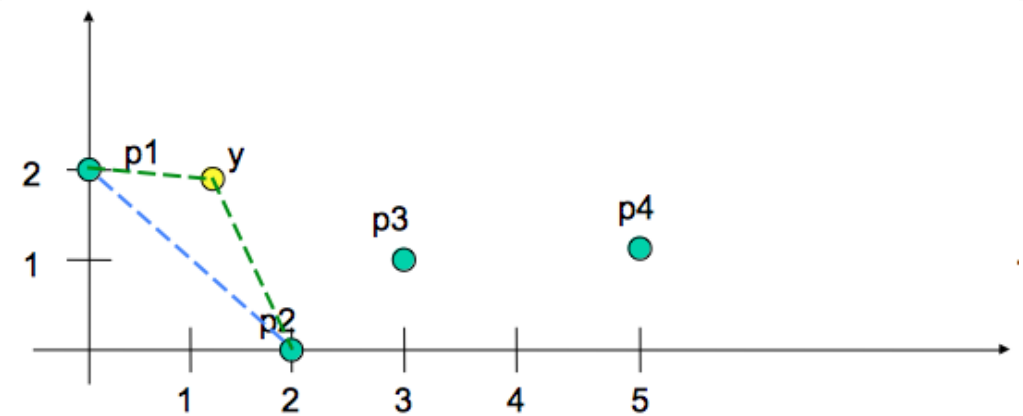
Source: Tan et al. (2005)

Medidas de Distancia

Propiedades

1. Positiva
   $d(x, y) \geq 0$
   $d(x, y) = 0$ solo si $x = y$
2. Simétrica
   $d(x, y) = d(y, x)$ para todo x y y
3. Desigualdad Triangular
   $d(x, z) <= d(x, y) + d(y, z)$ para todo punto x,y y z

Métricas



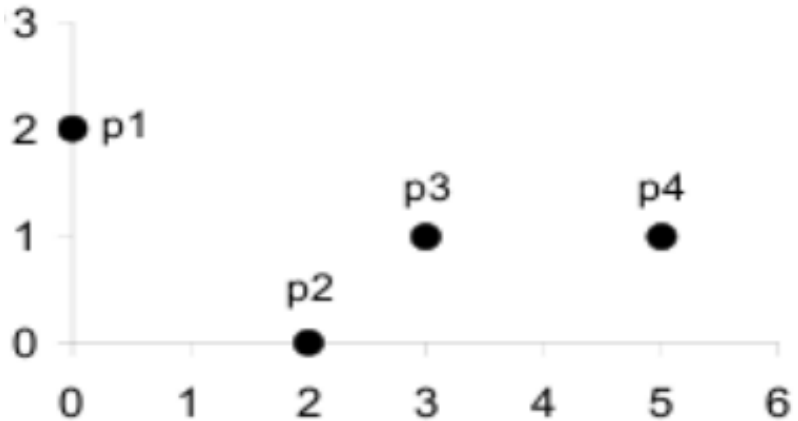
Distancia Euclideana

- Espacio Euclideano
- Longitud de la linea recta entre dos puntos

$$Dist_{Euclidean}(p,q) = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$$

- Donde n es la dimensión (número de atributos)
- p_k y q_k son los k-esimos atributos de los puntos p y q
- Normalización necesaria si las escalas de los atributos difieren.

Distancia Euclideana



punto	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

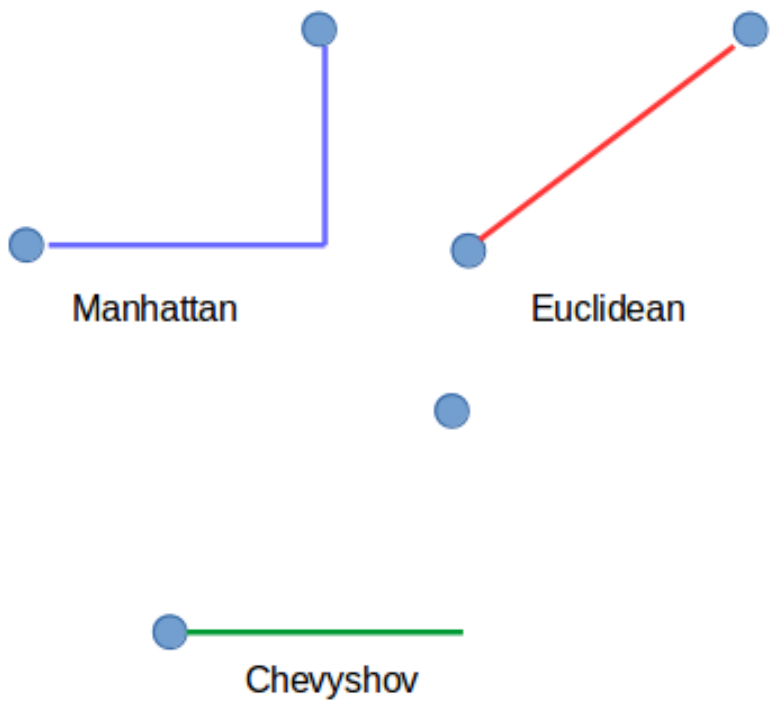
Matriz de Distancias

Distancia Minskonsky

Generalización de la distancia Euclideana, mediante el paramétro r

$$Dist_{Minskonsky}(p,q) = \sqrt[r]{\sum_{k=1}^{n} |p_k - q_k|^r}$$

- $r = 1$ Distancia **Manhattan o "City Block"**.
- $r = 2$ Distancia **Euclideana**
- $r \to \propto$ Distancia **Chevychov** "supremo". $L_\propto$ o L_{max}. La máxima diferencia entre los atributos



Distancia Minskonsky

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

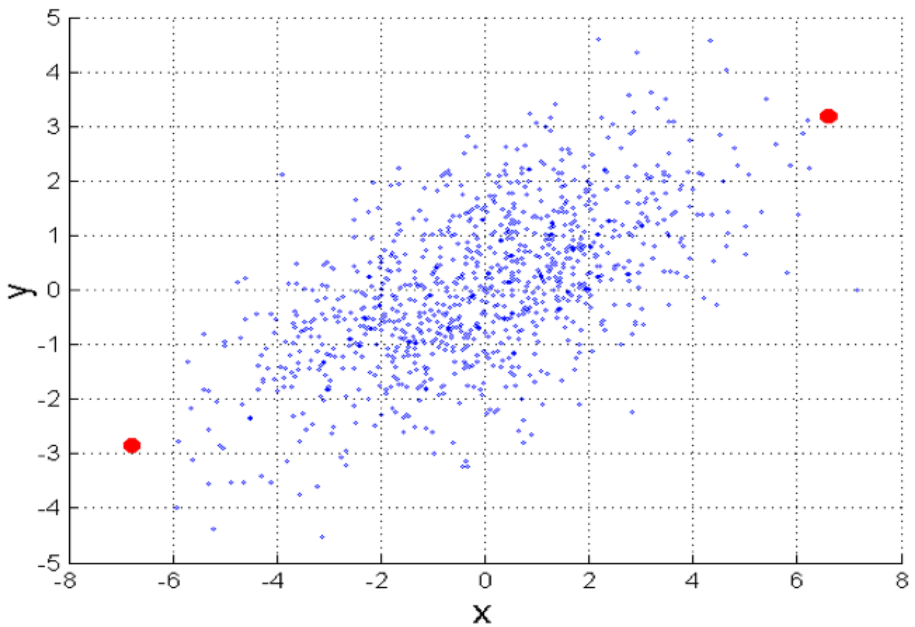
L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distancia Mahalanobis

$$DistMahalanobis(p,q) = (p-q)\sum^{-1}(p-q)^T$$

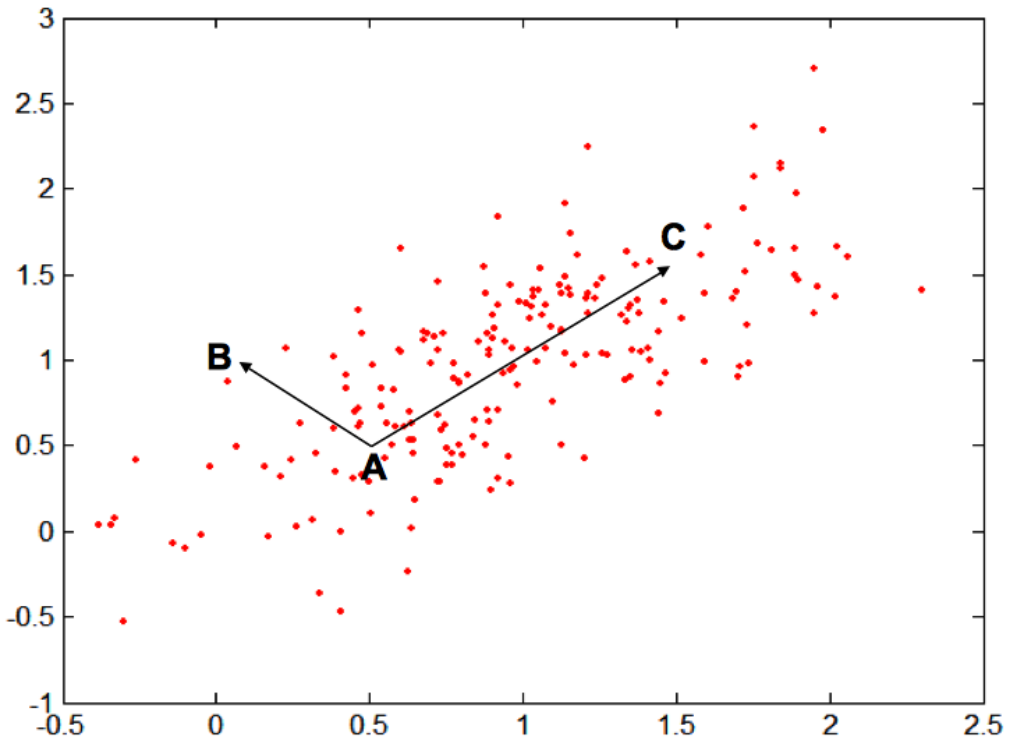
Donde $\sum$ es la matriz de covarianza del conjunto X de n dimensiones. $Cov(j,k) = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X_j})(X_{ik} - \overline{X_k})$



Source: Tan et al. (2005)

Para puntos rojos, la distancia Euclideana es 14.7, la distancia Mahalanobis es 6.

Distancia Mahalanobis



Source: Tan et al. (2005)

Matriz de covarianza

$$\sum = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)
B: (0, 1)
C: (1.5, 1.5)
Mahal(A,B) = 5
Mahal(A,C) = 4

Distancia Hamming

Distancia entre cadenas de **bits**

Número de bits que son diferentes entre dos objetos.

$$Dist_{Hamming}(p, q) = (b + c)$$

	1	0
1	a	b
0	c	d

Tabla de contingencia

Distancia Hamming

Distancia entre cadenas de **bits**

Número de bits que son diferentes entre dos objetos.

$$Dist_{Hamming}(p, q) = (b + c)$$

	1	0
1	a	b
0	c	d

Tabla de contingencia

Medidas de Similitud

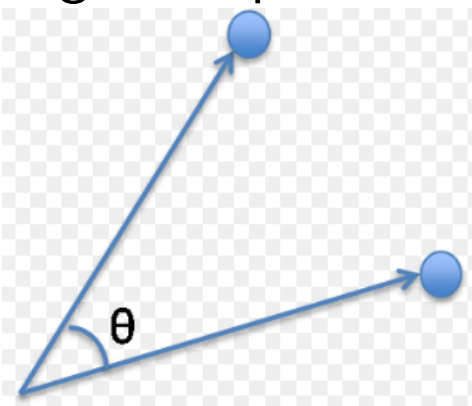
Mide la cercanía (que tan parecidos son dos objetos). Inversa a la "distancia".
$s(p, q)$ es la similitud entre puntos (objetos de datos) p y q.

Características

1. Máxima similitud: $s(p, q) = 1$ solo si $p = q$.
2. Simetria: $s(p, q) = s(q, p)$ para todo p y q.

Similitud de Coseno

Los objetos se consideran "Vectores". La Similitud se mide por el ángulo θ que los separa usando el $coseno\ \theta$.



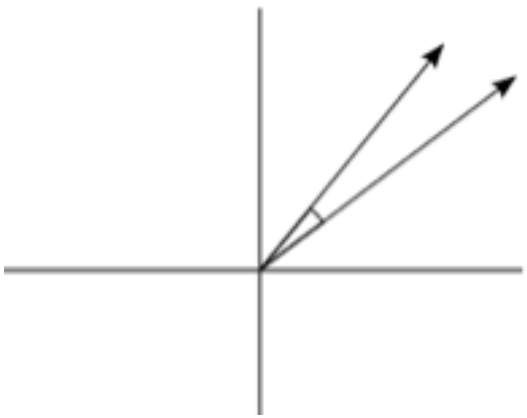
$$Sim_{Coseno}(p,q) = \frac{\sum_{i=1}^{n}(p_i * q_i)}{\sum_{i=1}^{n} p_i^2 * \sum_{i=1}^{n} q_i^2}$$

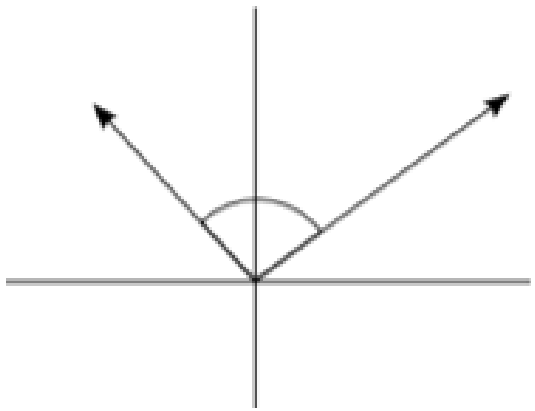
Medida del ángulo

1. Si $\theta = 0$, la similaridad es 1, p y q son el mismo excepto por magnitud.
2. Si $\theta = 90$, la similaridad es 0 (perpendiculares)

Similitud de Coseno

La Similitud se mide por el ángulo θ que los separa usando el $coseno\ \theta$.





Se usa en $textmining$ cuando los documentos u oraciones representados en VSM (Vector Space Model), usando cuadrante positivo.

Similitud de Coseno

Ejemplo

$$p = 3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0$$
$$q = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2$$

$$Sim_{Coseno}(p,q) = \frac{p \cdot q}{\|p\|\|q\|}$$

Producto punto de los vectors sobre la multiplicación de las normas de los vectores

$$p \cdot q = 3*1+2*0+0*0+5*0+0*0+0*0+0*0+2*1+0*0+0*2 = 5$$
$$\|p\| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$$
$$\|q\| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.245$$
$$Sim_{Coseno}(p,q) = 5/(6.481 * 2.245) = \mathbf{0.3150}$$

Similitud de Correlación

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{covariance}(\mathbf{x}, \mathbf{y})}{\text{standard\_deviation}(\mathbf{x}) * \text{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x \; s_y}, \quad (2.11$$

where we are using the following standard statistical notation and definitions

$$\text{covariance}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y}) \quad (2.12$$

$$\text{standard\_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2}$$

$$\text{standard\_deviation}(\mathbf{y}) = s_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \overline{y})^2}$$

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k \text{ is the mean of } \mathbf{x}$$

$$\overline{y} = \frac{1}{n} \sum_{k=1}^{n} y_k \text{ is the mean of } \mathbf{y}$$

Similitudes Binarias

Vectores con solo valores binarios: 0 y 1

Matriz de contingencia

	1	0
1	a	b
0	c	d

a = número de dimensiones donde p es 1 y q es 1
b = número de dimensiones donde p es 1 y q es 0
c = número de dimensiones donde p es 0 y q es 1
d = número de dimensiones donde p es 0 y q es 0

"Simple Matching Coeficient"- SMC

Número de dimensiones con igual valor/ número de dimensiones

$$Sim_{SMC} = \frac{a+d}{a+b+c+d}$$

Ejemplo

$$p = 1\ 0\ 1\ 0\ 0\ 1\ 1\ 1\ 1\ 0$$
$$q = 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 1\ 0$$

$$Sim_{SMC}(p,q) = \frac{3+2}{3+3+2+2} = \frac{5}{10} = 0,5$$

Coeficiente de Jaccard

El cero es considerado $ausencias$, no se tienen en cuenta las coincidencias de ceros (d)
Número de dimensiones con coincidencias en $unos$ (a)/ (número de dimensiones menos coincidencias en $ceros$ (d))

$$Sim_{Jaccard} = \frac{a}{a+b+c}$$

Ejemplo

$$p = 1\ 0\ 1\ 0\ 0\ 1\ 1\ 1\ 1\ 0$$
$$q = 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 1\ 0$$

$$Sim_{Jaccard}(p,q) = \frac{3}{3+3+2} = \frac{3}{8} = 0,375$$

Coeficiente de Rao

Número de dimensiones con coincidencias en $unos$ (a)/ número de dimensiones

$$Sim_{Rao} = \frac{a}{a+b+c+d}$$

Ejemplo

$$p = 1\ 0\ 1\ 0\ 0\ 1\ 1\ 1\ 1\ 0$$
$$q = 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 1\ 0$$

$$Sim_{Rao}(p,q) = \frac{3}{3+3+2+2} = \frac{3}{10} = 0,3$$

Ejercicio Similitudes Binarias

Ejercicio

$$p = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0$$
$$q = 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1$$

$$Sim_{SMC}(p, q) = ?$$
$$Sim_{Jaccard}(p, q) = ?$$
$$Sim_{Rao}(p, q) = ?$$

¿Qué concluye?

Ejercicio

Ejercicio

$$p = 0\ 0\ 1\ 1\ 0\ 1\ 0\ 1$$
$$q = 0\ 1\ 1\ 0\ 0\ 1\ 0\ 0$$

$$Sim_{SMC}(p, q) = ?$$
$$Sim_{Jaccard}(p, q) = ?$$
$$Sim_{Rao}(p, q) = ?$$
$$Sim_{Cosine}(p, q) = ?$$

Coeficiente de Jaccard Extendido: Tanimoto

Extendido a valores continuos

$$Sim_{Tanimoto} = \frac{p \cdot q}{\|p\|^2 + \|q\|^2 - p \cdot q}$$

Similaridad de Gower

Para datos mixtos: variables binarias, cualitativas y cuantitativas. La proximidad entre individuos que presenten esta combinación de datos puede ser medida usando similaridad de Gower (1971).

$$Sim_{Gower}(\text{p,q}) = \frac{1}{T}\sum_{t=1}^{T}\delta_{pqt}$$

La similitud entre p y q, es el promedio de las similitudes por cada variable (T es el número de variables o dimensiones).

En caso de...

- variables binarias o cualitativas:
  - $\delta_{pqt} = 1$ Si $X_{pt} = X_{qt}$
  - $\delta_{pqt} = 0$ Si $X_{pt} \neq X_{qt}$
- variables cuantitativas: $\delta_{pqt} = 1 - \frac{|X_{pt} - X_{qt}|}{r_t}$
  r_t es el rango (diferencia entre el máximo y el mínimo) de la variable t

Combinación de similitudes

Cuando los atributos son de diferentes tipos:

- Computar la similitud entre cada atributo separadamente, y luego combinar las similitudes aplicando un métdodo que de un valor entre 0 y 1. Por ejemplo, el promedio de todas las similitudes individuales.
- No funciona bien si los atributos son asimétricos. En este caso se pueden omitir esos atributos en el cálculo de la similitud.
- Si existen atributos más importantes que otros, la ecuación de la similitud puede ser modificada adicionando *pesos de contribución* a cada atributo.

Combinación de similitudes

Algoritmo para calcular similitud de dos objetos heterogeneos

for $k = 1$ to n **do**

computar $s_k(x, y)$ en rango $[0, 1]$

definir $\delta_k = \begin{cases} 0 & \text{si el atributo k es asimétrico y} \\ & \text{ambos objetos tienen valor de 0, o} \\ & \text{si uno de los atributos tiene valor perdido} \\ 1 & \text{otro caso} \end{cases}$

end for

computar similitud completa entre los dos objetos usando:

$$S(x, y) = \frac{\sum_{k=1}^{n} \delta_k s_k(x, y)}{\sum_{k=1}^{n} \delta_k}$$

Combinación de similitudes

Cálculo de similitud de dos objetos con *pesos*

$$S(x,y) = \frac{\sum_{k=1}^{n} w_k \delta_k s_k(x,y)}{\sum_{k=1}^{n} \delta_k}$$

Selección de Medida de Proximidad

- La medida de proximidad debe permitir el manejo de tipo datos. La distancia Euclideana es muy usada para tipos de datos continuos en espacios densos.
- La proximidad en datos continuos es por lo general expresado en términos de diferencias, por lo general se usan métricas de distancia. Si los atributos tienen diferentes escalas y/o diferente importancia se maneja con normalizaciones, transformaciones y pesos.
- para datos dispersos (por lo general atributos asimetricos), se usan medisa de similitud que ignoren los "matchs" entre ceros (la similitud se ventra en los atributos que comparten). Se usan similitudes de Coseno y Jaccard.

eleonguz@unal.edu.co

www.midas.unal.edu.co

References I

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining, (first edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.