INGENIERÍA INDUSTRIAL
UNIVERSIDAD DE CHILE

IN5625 - INVESTIGACIÓN DE MERCADOS

Segmentación

André Carboni

Semestre primavera 2013

acarboni@ing.uchile.cl

Estamos aquí…

SEGMENTACIÓN

Segmentación

- Recordemos que segmentación corresponde al proceso de **dividir un mercado** en grupos identificables, más o menos similares y significativos, con el propósito de ajustar el marketing-mix a la medida de las necesidades de uno o más segmentos específicos.

Bases de segmentación (consumidor)

- **Geográfica:** Dónde se encuentran.
- **Demográfica:** Edad, sexo, educación, ingresos, …
- **De uso:**
  - "Estado" del usuario: Usuario potencial, frecuente, etc.
  - Grado de lealtad
  - Forma de uso
- **Psicográfica:** "Estilo de vida" → Pensamientos, sentimientos y conductas de las personas
  - Ej: Extrovertidos, innovadores, agitadores políticos, ….
- …

→ *"Es preferible usar varios tipos de variables como bases de segmentación"*

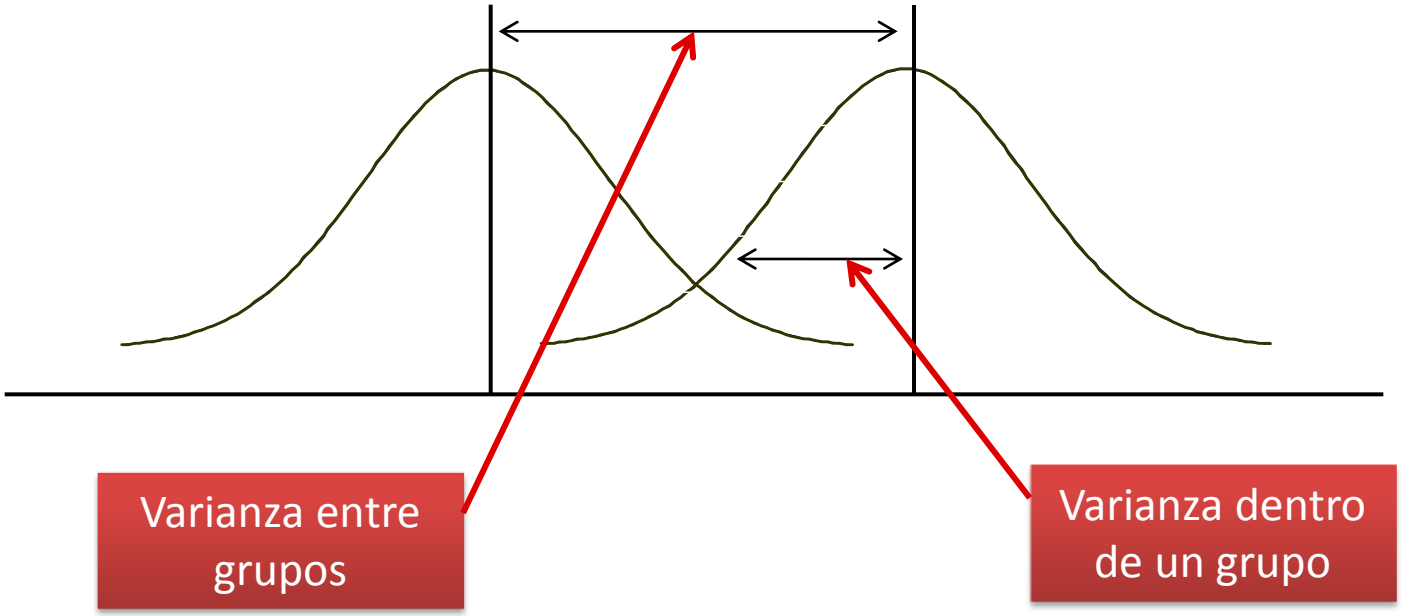
ANÁLISIS DE CONGLOMERADOS
(CLUSTER ANALYSIS)

Análisis de conglomerados

- **Definición:** conjunto de técnicas que se utilizan para clasificar los objetos o casos en grupos relativamente homogéneos llamados **conglomerados** (clusters).

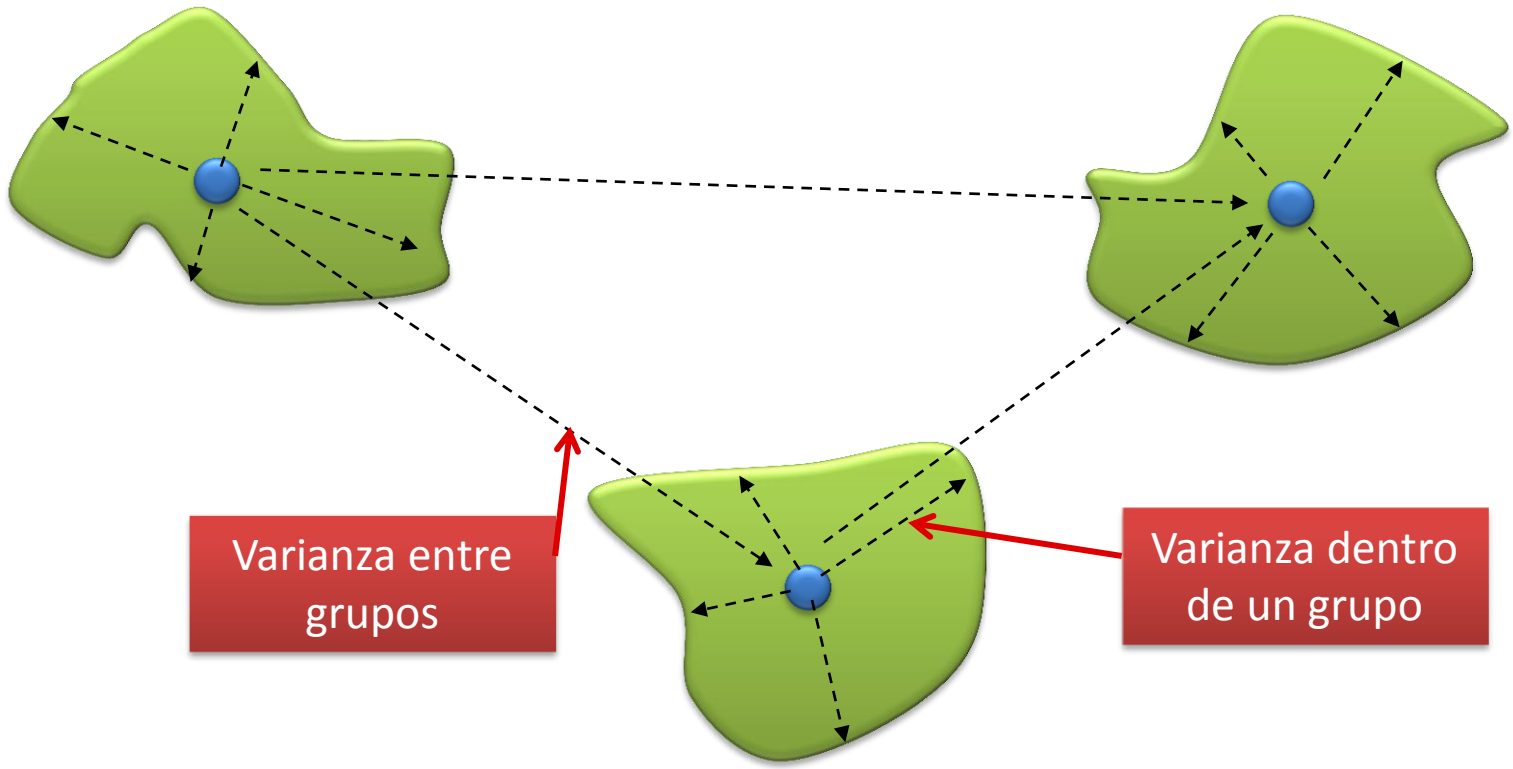
  - Analizaremos los métodos en los cuales cada objeto es asignado a un y sólo un conglomerado.
  - Se busca que los objetos dentro cada grupo sean similares entre sí y diferentes a los objetos de los otros grupos.
  - Se utiliza un principio de maximización de la varianza entre clusters mientras que minimiza la varianza dentro de un cluster

Ilustración en una dimensión



$$\max_{c \in C} \left\{ \frac{\text{varianza entre grupos}}{\text{varianza en los grupos}} \right\} \quad C = \text{Conjunto de clusters posibles}$$

Ilustración en dos dimensiones



$$\max_{c \in C} \left\{ \frac{\text{varianza entre grupos}}{\text{varianza en los grupos}} \right\} \quad C = \text{Conjunto de clusters posibles}$$

¿Estadística?

- Casi todos los procedimientos de análisis de datos proviene desde el mundo de la estadística (regresiones, correlaciones, pruebas de hipótesis).

- La mayoría de los métodos de conglomeración provienen mas bien desde el mundo de la **computación** y los **algoritmos (heurísticas),** por lo que la aplicación de análisis estadístico es mas bien tangencial

Usos del análisis de conglomerados

- **Segmentación de mercados:** encontrar grupos homogéneos a quienes se puede satisfacer con un determinado marketing mix.

- **Comprensión del comportamiento del comprador:** determinación de comportamientos de compra de los compradores.
- **Identificación de oportunidades para productos nuevos:** Determinación de las marcas que compiten con un determinado producto.
- **Selección de mercados de prueba:** Elección de una región/población sobre la que se puede hacer una prueba de mercado.
- **Reducción de datos:** Disminuir el número de objetos para facilitar la comprensión de los fenómenos descritos.

Procedimiento de análisis de conglomerados

1. Plantear el problema

- Corresponde a la selección de las variables en las que se basa la agrupación.
- La inclusión de variables irrelevantes puede distorsionar las soluciones.
- El conjunto de variables debe describir la similitud de los objetos de acuerdo a los propósitos de la clasificación.
- La selección de variables se hace en base a investigación previa, teoría o de acuerdo a las hipótesis que se testean.

Ésta es probablemente la parte mas importante en el análisis de conglomerados y requiere el desarrollo de un buen criterio e intuición.

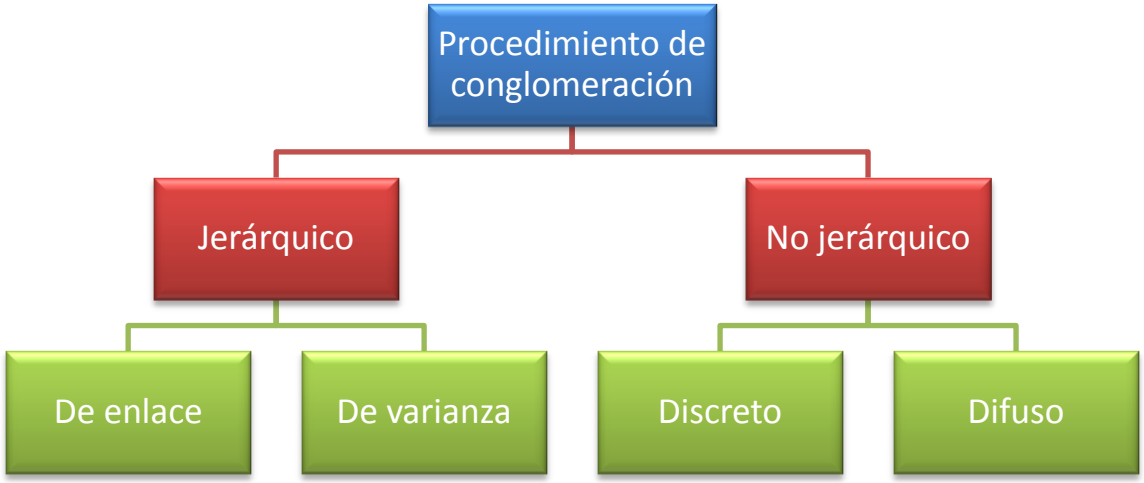
2. Elegir una medida de distancia

- Medidas típicas:
  - Distancia euclidiana: $d_{ij} = \sqrt{\sum_k \left(x_i^k - x_j^k \right)^2}$
  - Distancia Manhattan: $d_{ij} = \sum_k \left| x_i^k - x_j^k \right|$
  - Distancia de Chebychev: $d_{ij} = \max_k \left\{ \left| x_i^k - x_j^k \right| \right\}$
  - Etc.

- Si las unidades de medición de las variables seleccionadas son diferentes, conviene **estandarizar** las variables.
- Detección de outliers.
- Distintas medidas de distancia generan distintos resultados. Se recomienda usar varias medidas y comparar resultados.

3. Elegir procedimiento de conglomeración

- En la literatura se reportan una gran cantidad de métodos distintos que difieren en términos de:
  - Eficiencia computacional.
  - Medidas de similitud usadas.
  - Requerimiento de número de grupos.



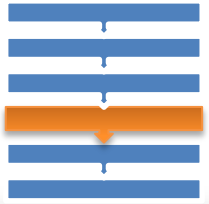
3. Elegir procedimiento de conglomeración

- **Conglomerado jerárquico:**
  - De una iteración a otra, se modifica el valor de pertenencia a grupos de un único objeto.
  - No requiere a priori fijar un número de clusters.
  - **Por aglomeración:**
    - Inicialmente todos los objetos en grupos distintos.
    - En cada iteración se agrupa el par de objetos mas similares.
  - **Por división:**
    - Inicialmente todos los objetos en un único grupo.
    - En cada iteración se separan el par de objetos mas disímiles.

- **Conglomerado no jerárquico:**
  - De una iteración a otra, se puede modificar el valor de pertenencia a grupos de todos los objetos.
  - Requiere a priori fijar un número de clusters.

4. Decidir número de conglomerados

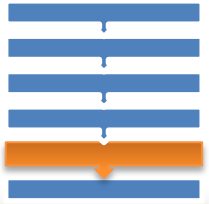
- No existe regla general. Algunos lineamientos:
  - Consideraciones teóricas, conceptuales o prácticas pueden sugerir un número de grupos.
  - En conglomerados jerárquicos, observar los cambios en las distancias en las que los grupos se combinan (ej: uso de dendograma).
  - En conglomerados no jerárquicos puede graficarse la razón (varianza entre grupos)/(número de grupos) y detectar puntos de cambios marcados.
  - Los tamaños de los grupos deben ser significativos.



5. Interpretar y describir conglomerados

- Se busca una semántica que defina a los objetos del grupo.

- Dos enfoques complementarios:
  - Análisis de los centroides de cada grupo y comparación con el de los otros grupos.
  - Análisis discriminante para determinar las variables que marcan diferencias significativas.

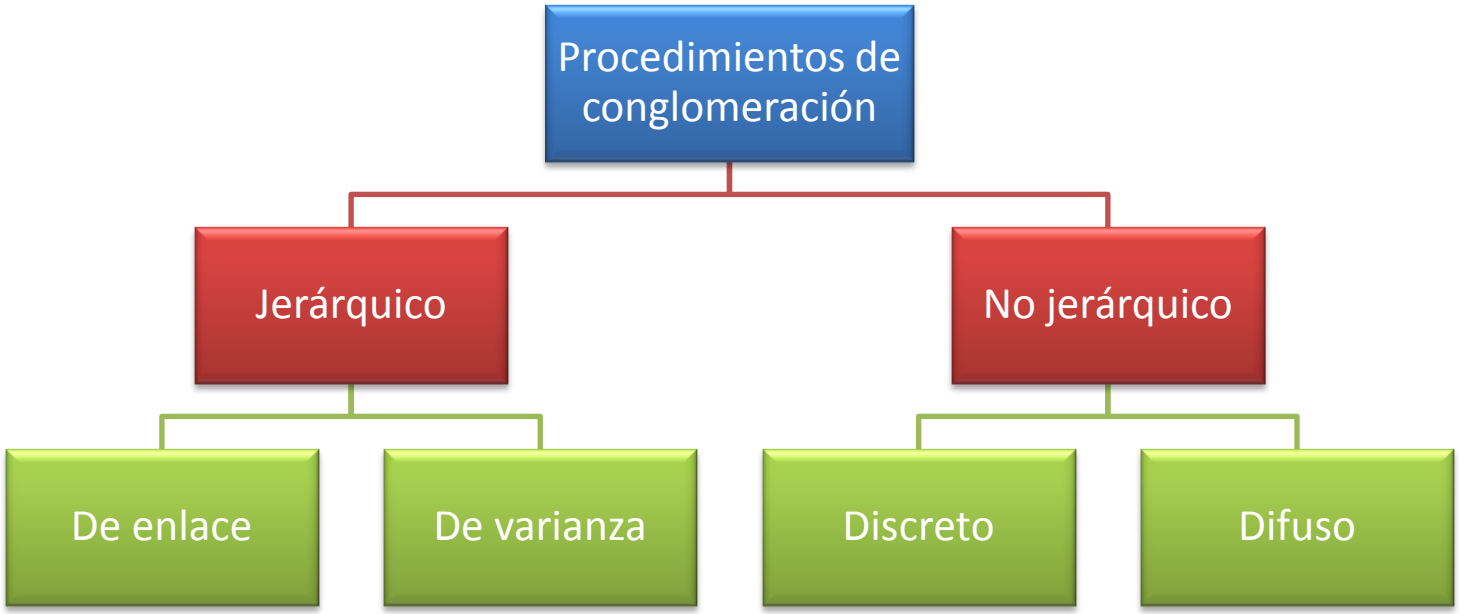


6. Evaluar validez del conglomerado

- Algunos chequeos mínimos de validez:
  - Realizar análisis de conglomerados con los mismos datos, pero con distintas medidas de distancia.
  - Analizar la estabilidad de la clasificación haciendo el análisis para varias selecciones distintas de variables.
  - Utilizar varios métodos y comparar resultados.
  - Dividir los datos en dos partes: hacer análisis de conglomerados en ambas y comparar resultados.
  - Para clasificación no jerárquica, cambiar valor de centroides iniciales y el orden de las variables.

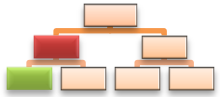
PROCEDIMIENTOS DE CONGLOMERACIÓN

Como vimos antes…

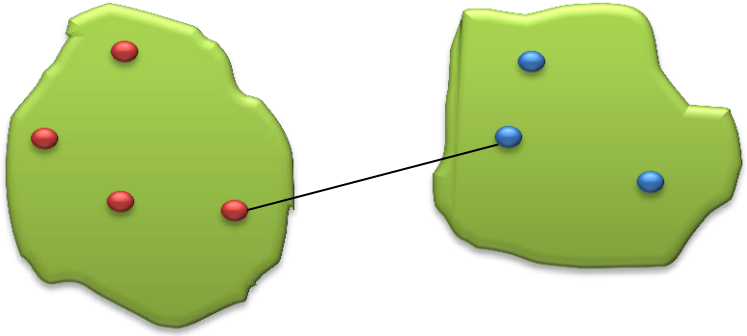


Métodos de enlace

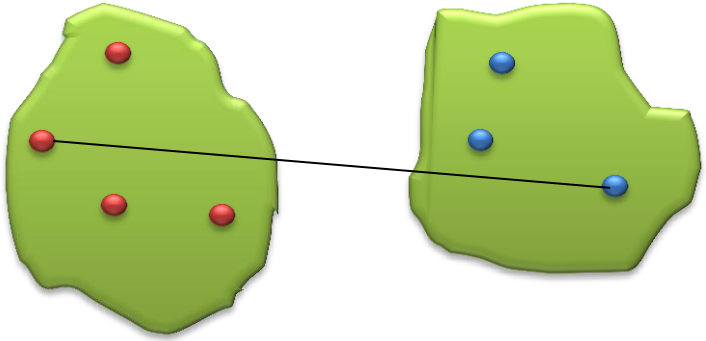
- Este procedimiento trata de identificar grupos de casos, usando un algoritmo que comienza con cada elemento en un cluster distinto.
- En cada iteración se agrupan los dos elementos cercanos.
- Las variantes de este enfoque dependen de la medida de similitud:
  - Enlace sencillo.
  - Enlace completo.
  - Enlace promedio



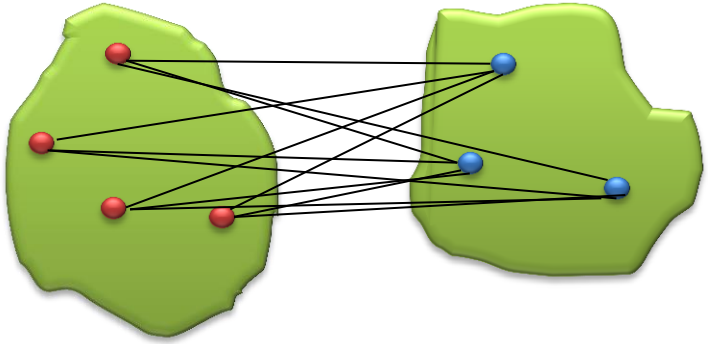
Métodos de enlace – tipos de medida



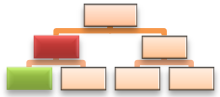
- Enlace Sencillo
- (Distancia mínima)



- Enlace completo
- (Distancia máxima)



- Enlace promedio
- (Distancia promedio)



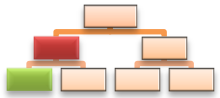
Ejemplo enlace simple

- Primero, se calculan las distancias entre los objetos.

	D1	D2	D3	D4	D5	D6	D7
D1							
D2	0,3606						
D3	0,5000	0,4243					
D4	0,9220	0,7071	0,4472				
D5	1,3416	1,0440	0,9220	0,5000			
D6	1,8385	1,5524	1,3892	0,9434	0,5099		
D7	1,7263	1,5000	1,2369	0,8062	0,5831	0,4000	

- Encontrar la distancia mínima entre pares de objetos.

	D1	D2	D3	D4	D5	D6	D7
D1							
D2	**0,3606**						
D3	0,5000	0,4243					
D4	0,9220	0,7071	0,4472				
D5	1,3416	1,0440	0,9220	0,5000			
D6	1,8385	1,5524	1,3892	0,9434	0,5099		
D7	1,7263	1,5000	1,2369	0,8062	0,5831	0,4000	

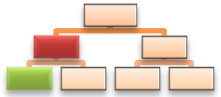


Ejemplo enlace simple

- Los clusters se unen y se usan las menores distancias como distancia mínima entre el nuevo cluster y el resto

	D1	D2	D3	D4	D5	D6	D7
D1							
D2	**0,3606**						
D3	0,5000	**0,4243**					
D4	0,9220	**0,7071**	0,4472				
D5	1,3416	**1,0440**	0,9220	0,5000			
D6	1,8385	**1,5524**	1,3892	0,9434	0,5099		
D7	1,7263	**1,5000**	1,2369	0,8062	0,5831	0,4000	

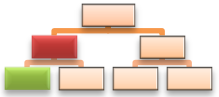
	D8	D3	D4	D5	D6	D7
D8						
D3	0,4243					
D4	0,7071	0,4472				
D5	1,0440	0,9220	0,5000			
D6	1,5524	1,3892	0,9434	0,5099		
D7	1,5000	1,2369	0,8062	0,5831	0,4000	



Ejemplo enlace simple

	D8	D3	D4	D5	D6	D7
D8						
D3	0,4243					
D4	0,7071	0,4472				
D5	1,0440	0,9220	0,5000			
D6	1,5524	1,3892	0,9434	**0,5099**		
D7	**1,5000**	**1,2369**	**0,8062**	0,5831	**0,4000**	

	D8	D3	D4	D5	D9
D8					
D3	0,4243				
D4	0,7071	0,4472			
D5	1,0440	0,9220	0,5000		
D9	1,5000	1,2369	0,8062	0,5099	



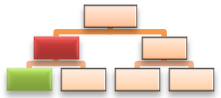
Ejemplo enlace simple

	D8	D3	D4	D5	D9
D8					
D3	**0,4243**				
D4	0,7071	**0,4472**			
D5	1,0440	**0,9220**	0,5000		
D9	1,5000	**1,2369**	0,8062	0,5099	

	D10	D4	D5	D9
D10				
D4	**0,4472**			
D5	0,9220	**0,5000**		
D9	1,2369	**0,8062**	0,5099	

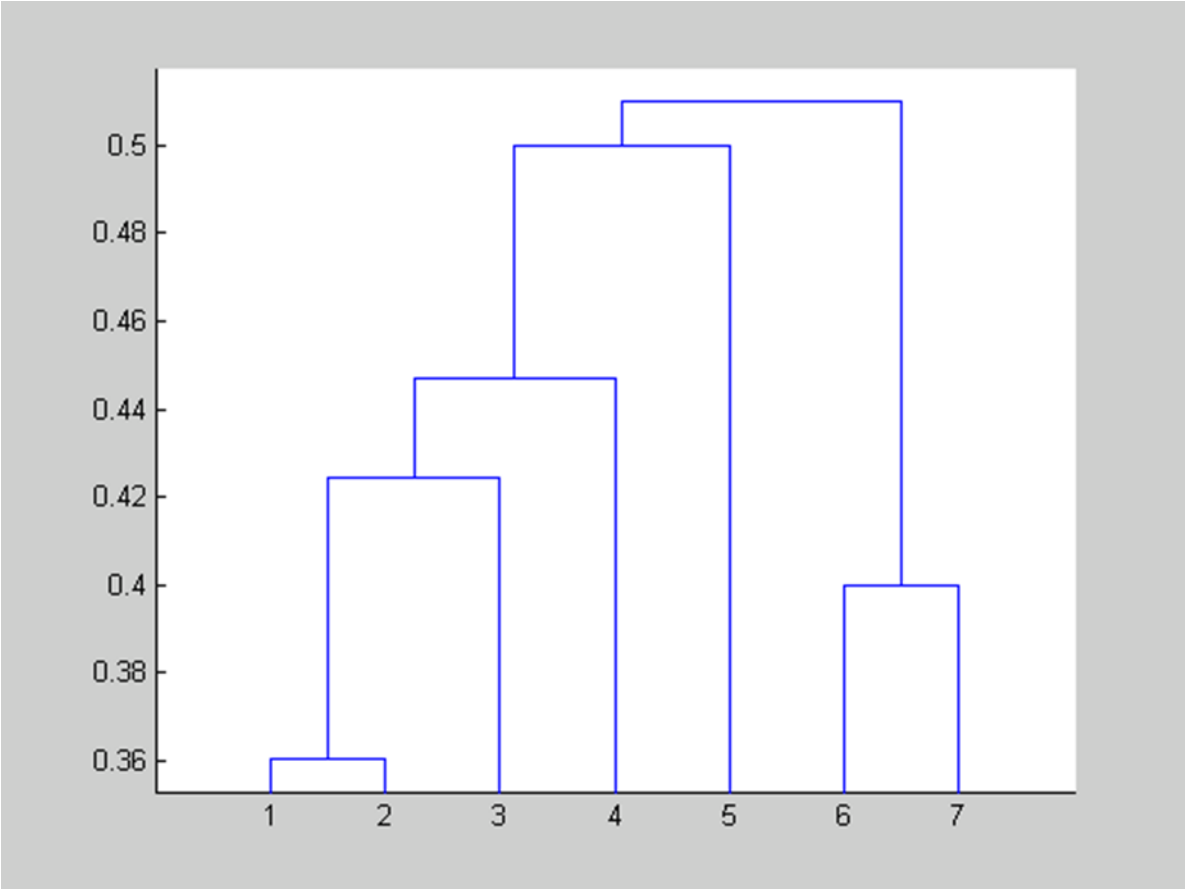
	D11	D5	D9
D11			
D5	0,5000		
D9	0,8062	**0,5099**	

	D12	D9
D12		
D9	0,5099	



Ejemplo enlace simple

- Dendrograma

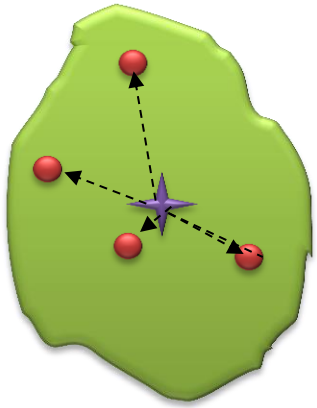


Métodos de varianza

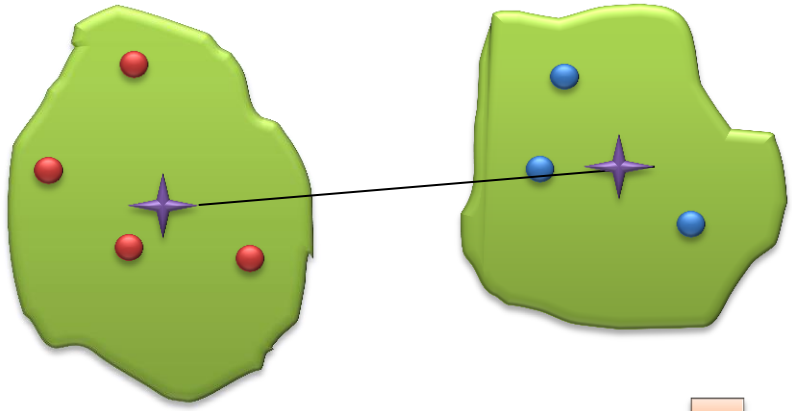
- Este procedimiento trata de identificar grupos de casos, tratando de minimizar la varianza dentro de los grupos.
- En cada iteración se agrupan los dos elementos cercanos.
- Distinguiremos al menos dos métodos que corresponden a este enfoque:
  - Método de Ward.
  - Método de centroide.

Métodos de varianza

- **Método de Ward:**
  - Se calculan las medias de las variables dentro de cada conglomerado (centroide), Luego, se busca el par de conglomerados que minimicen el incremento de varianza inter-conglomerados al ser fusionados.



- **Método de centroide:**
  - La distancia entre dos conglomerados es la distancia entre sus centroides.



K-medias

- El método de las K-medias (o K-means) es una de las técnica de clustering más comunes.

- Es una técnica iterativa que busca localizar a cada individuo en el cluster más cercano a él.

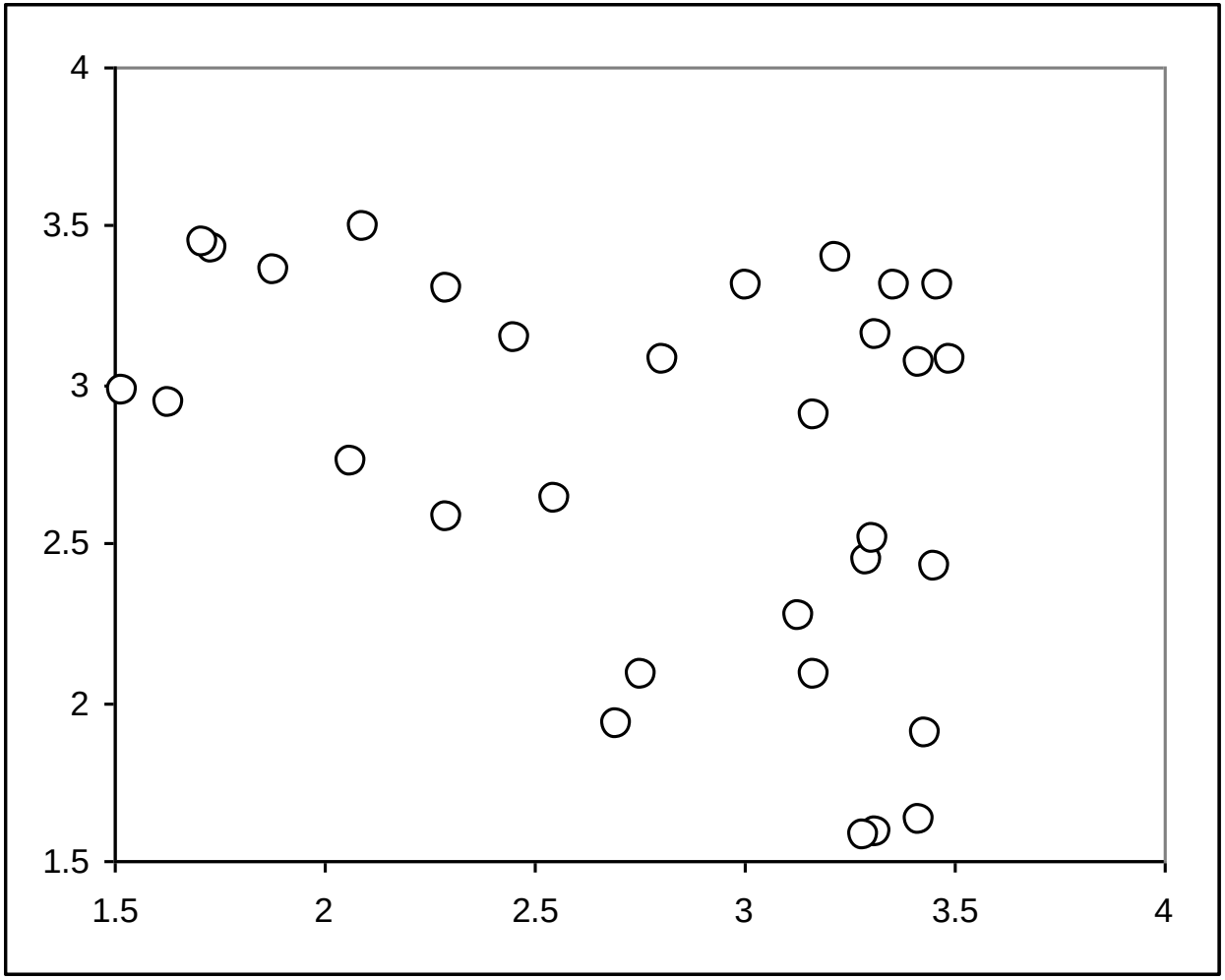
- El número de clusters es elegido por el usuario **a priori**.

K-medias - Algoritmo

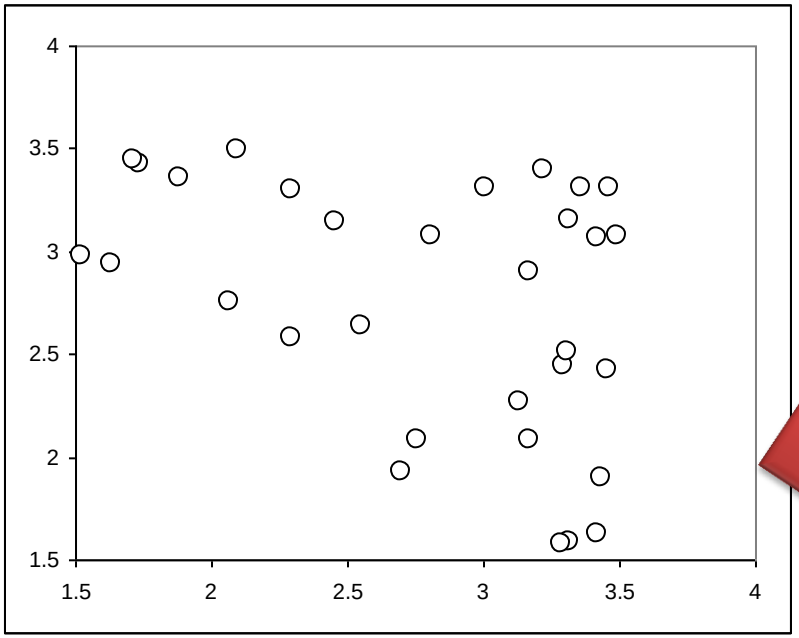
- **Inputs**
  - X: conjunto de N objetos
  - K: número de grupos
- **Outputs**
  - S1,...,Sk K conjuntos
  - Z1,...,Zk Los centros de cada grupo

- **Inicialización**
  - t=0
  - Elegir arbitrariamente Zj(t).
- **Asignación y actualización de centros**
  - Asignar Xi al grupo mas cercano para todo i=1...N.
  - Recalcular Zj j=1...K
  - t=t+1
- **Criterio de parada**
  - Si Zi(t)-Zi(t+1) < e para todo i, parar.

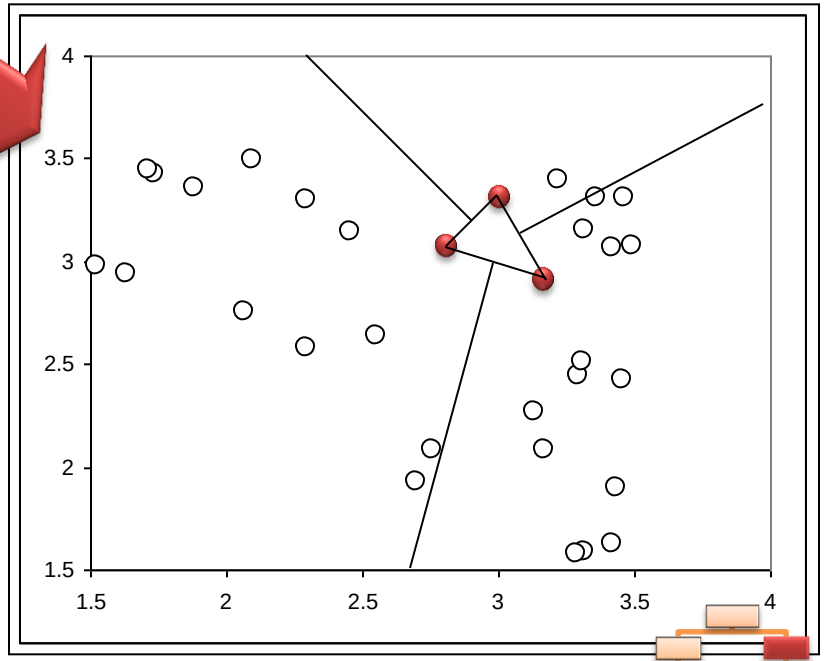
K-medias - Algoritmo



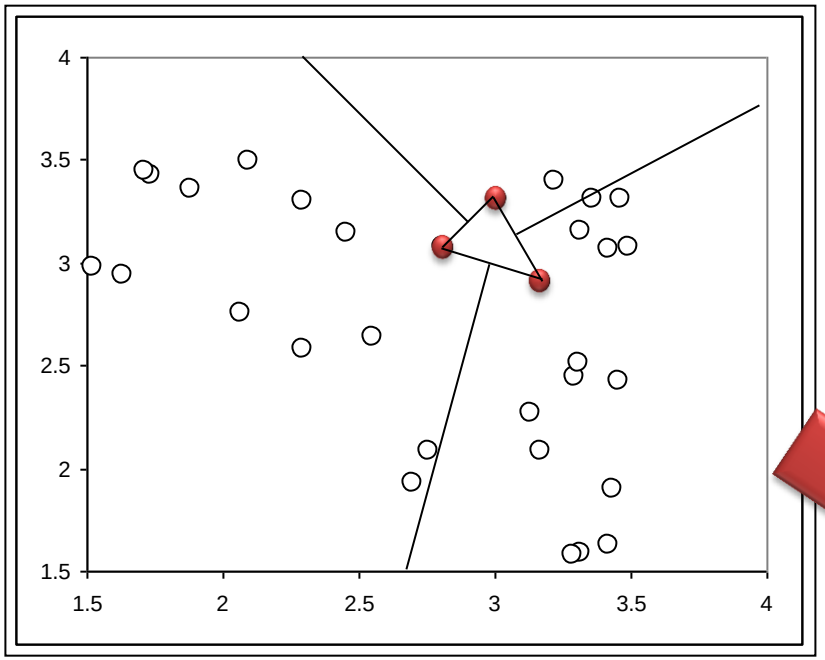
K-medias – Algoritmo Iteración 1



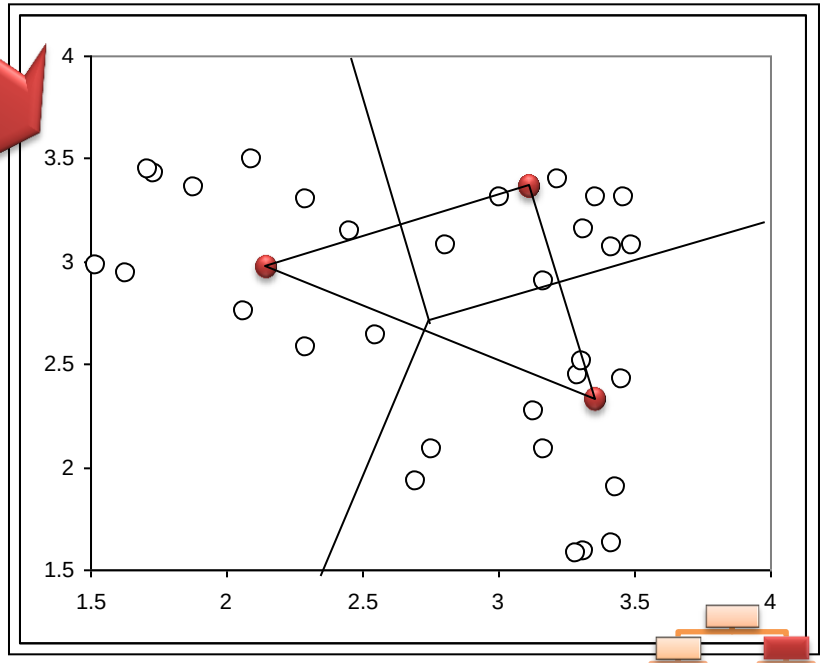
- Elegir los centros de los tres clusters aleatoriamente
- Localizar cada punto en su centro de cluster más cercano



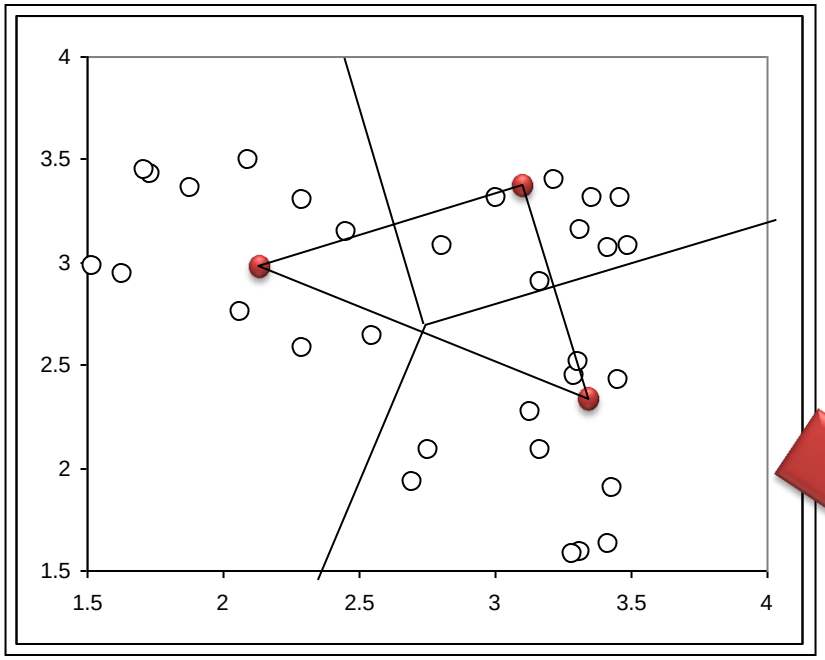
K-medias – Algoritmo Iteración 2



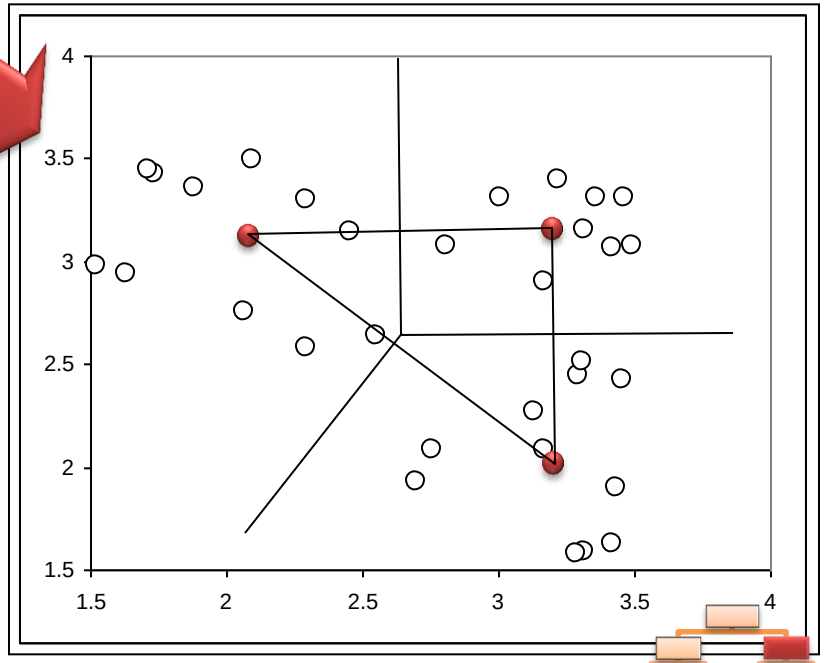
- Calcular nuevamente los centros de los clusters desde los centroides escogidos en la iteración 1.
- Localizar cada punto en el centroide que está más cerca a él.



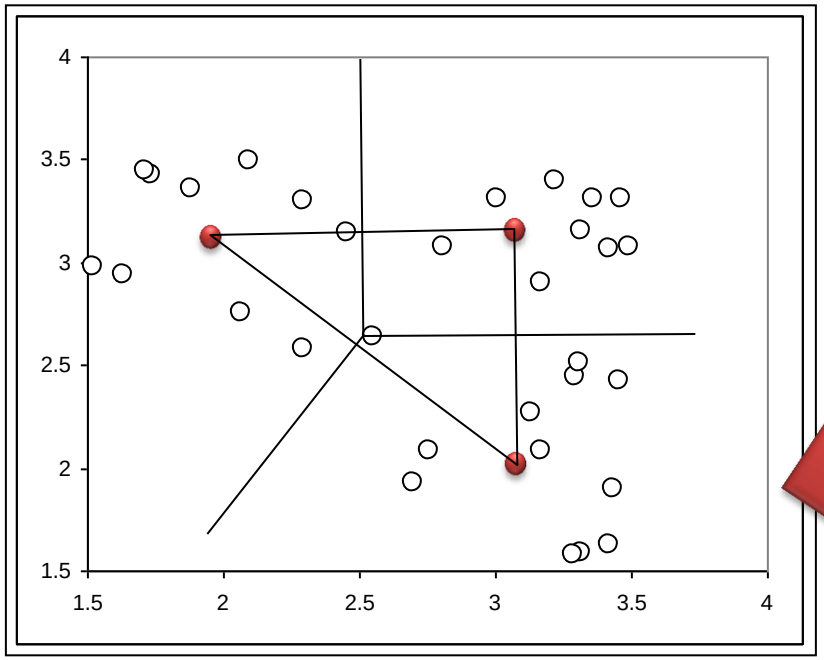
K-medias – Algoritmo Iteración 3



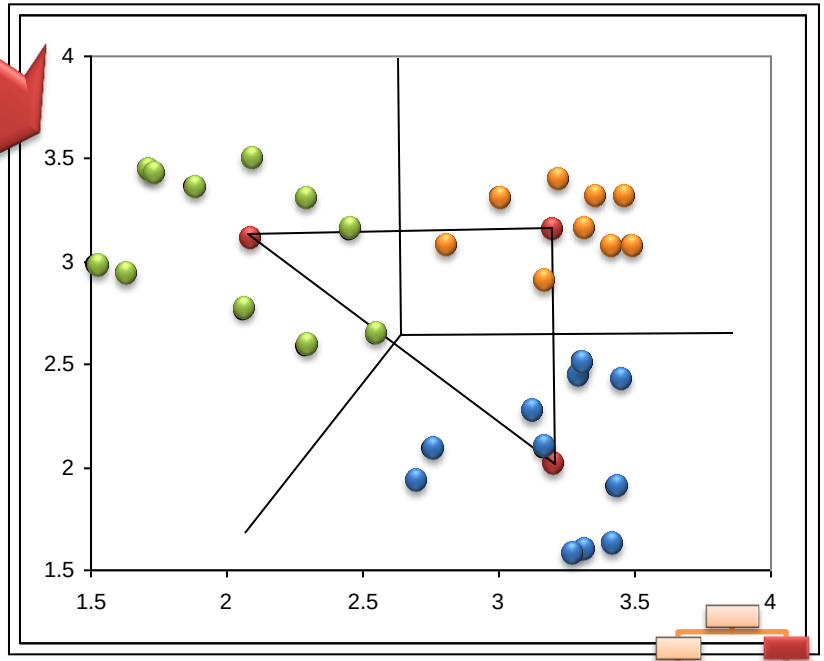
- Recalcular los centros como los centroides encontrados en la iteración 2.
- Localizar cada punto en el centroide que está más cerca a él.



K-medias – Algoritmo Iteración 4 y final



- Recalcular los centros como los centroides de los clusters desde la iteración 3.
- Nada cambió!!
- Ok, está listo.



Ejemplo - Clínicas

- Se conduce investigación de mercados y se llega a la siguiente tabla de frecuencias:

•	Actual, Moderna	Confianza	Trayectoria	Especialista	Prestigio	Atención	Reconocim.	Innovadora	Variedad	Profesional	Infrestruct.	Insumos de calidad	Precio convenien.	Tecnologia avanzada	Ubicación adecuada
Alemana	63	47	57	58	61	42	51	52	53	54	64	55	24	57	38
Central	2	3	3	12	3	4	4	4	8	7	6	6	4	5	4
Clínica UC	18	25	25	32	24	20	22	18	25	26	24	25	17	24	17
Cordillera	2	3	2	13	2	5	3	4	7	8	7	5	3	4	3
Dávila	5	7	7	15	6	11	8	5	11	11	10	8	9	5	5
Hosp. UC	24	31	35	43	32	28	35	25	38	36	32	32	24	28	26
Hosp. UCH	14	19	23	32	19	16	23	18	26	27	20	23	19	20	12
Indisa	9	7	9	18	9	8	6	7	12	12	11	10	8	6	6
Las Condes	49	30	34	44	39	28	36	40	41	38	45	40	13	41	21
Las Lilas	7	6	7	16	6	10	6	5	11	10	10	8	5	6	8
Las Nieves	4	3	2	13	3	4	3	4	7	6	7	6	1	4	2
Santa Maria	32	29	31	35	26	24	25	22	31	27	33	32	23	27	23
Tabancura	6	6	4	15	6	8	6	7	10	9	9	9	4	7	5
Vitacura	3	3	2	13	3	5	4	4	8	7	6	6	2	5	4

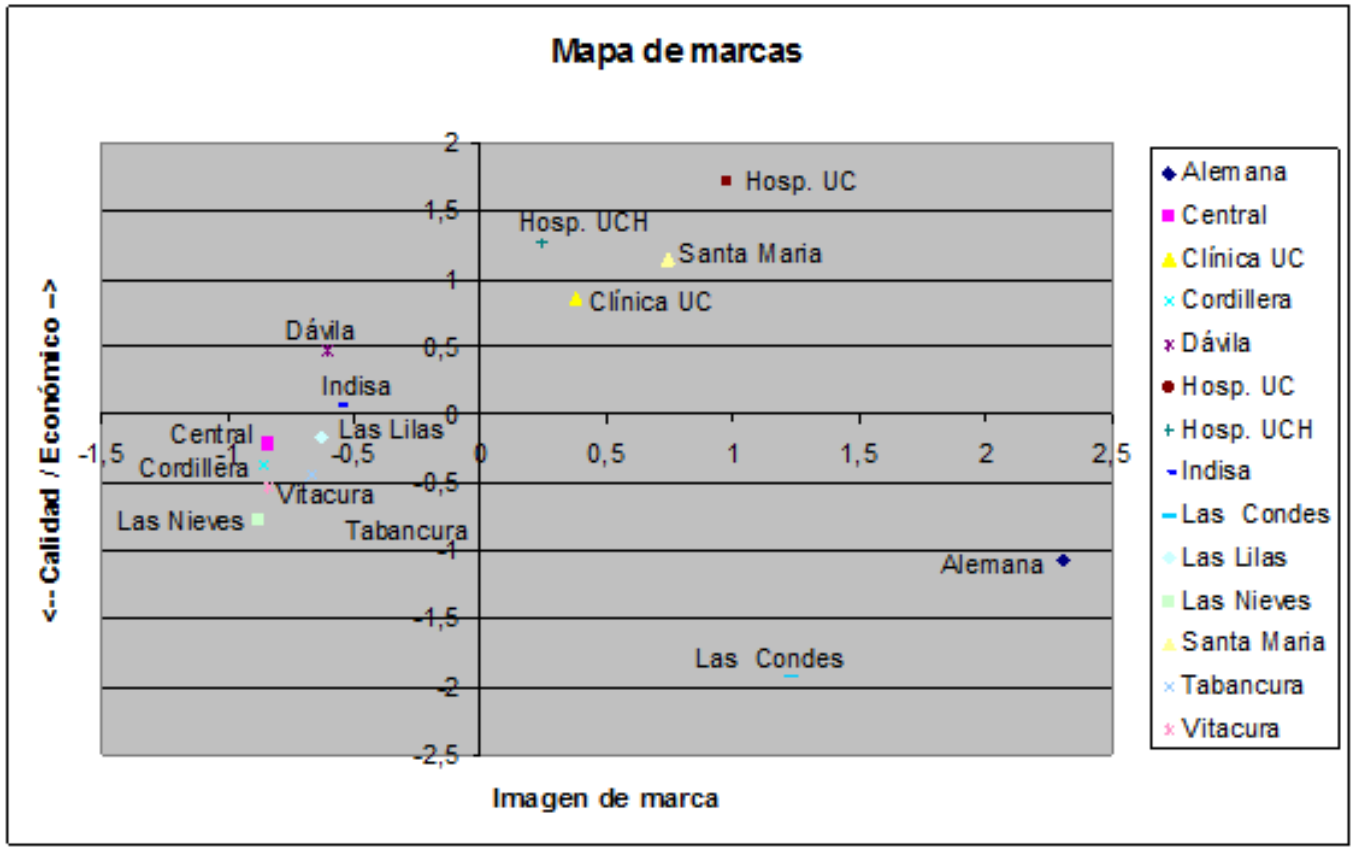
Podemos hacer K-medias directamente, pero será difícil de interpretar. Conviene "combinar" factores.

ACP ➔

	FACTOR 1	FACTOR 2
Alemana	2,307	-1,078
Central	-0,840	-0,227
Clínica UC	0,381	0,853
Cordillera	-0,857	-0,367
Dávila	-0,609	0,476
Hosp. UC	0,977	1,715
Hosp. UCH	0,242	1,267
Indisa	-0,558	0,067
Las Condes	1,236	-1,927
Las Lilas	-0,628	-0,162
Las Nieves	-0,878	-0,782
Santa Maria	0,739	1,144
Tabancura	-0,672	-0,443
Vitacura	-0,840	-0,536

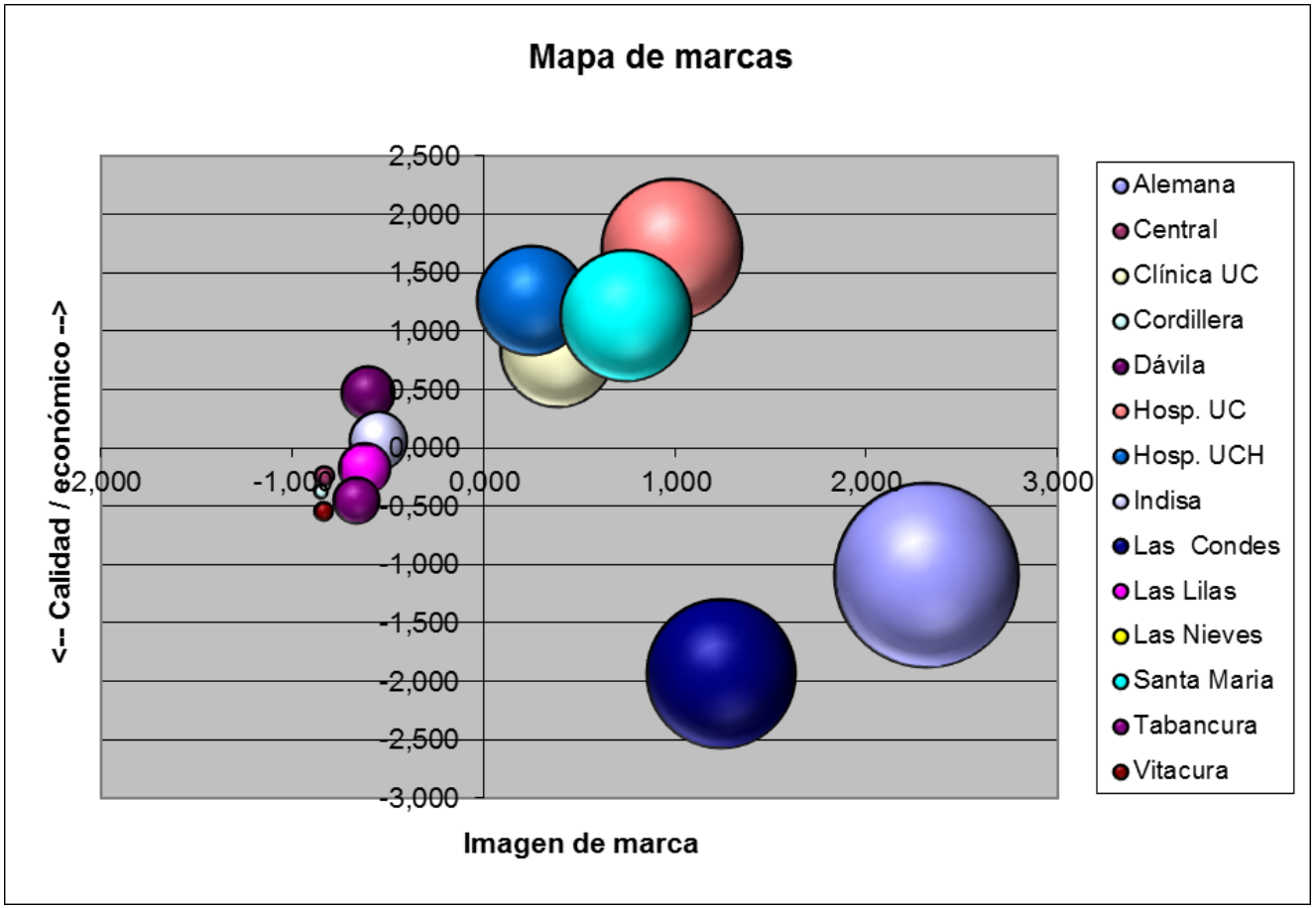
Ejemplo - Clínicas

- Mapa de marcas:



Ejemplo - Clínicas

- Mapa de marcas 2 – “Agregando valor al trabajo”: Radio de las esferas proporcional a la cantidad de menciones.



Ejemplo - Clínicas

- Usando K-medias con 3 conglomerados:

Centros de los conglomerados finales

	Conglomerado		
	1	2	3
REGR factor score 1 for analysis 1	1,77127	-,73530	,58496
REGR factor score 2 for analysis 1	-1,50277	-,24678	1,24495

Número de casos en cada conglomerado

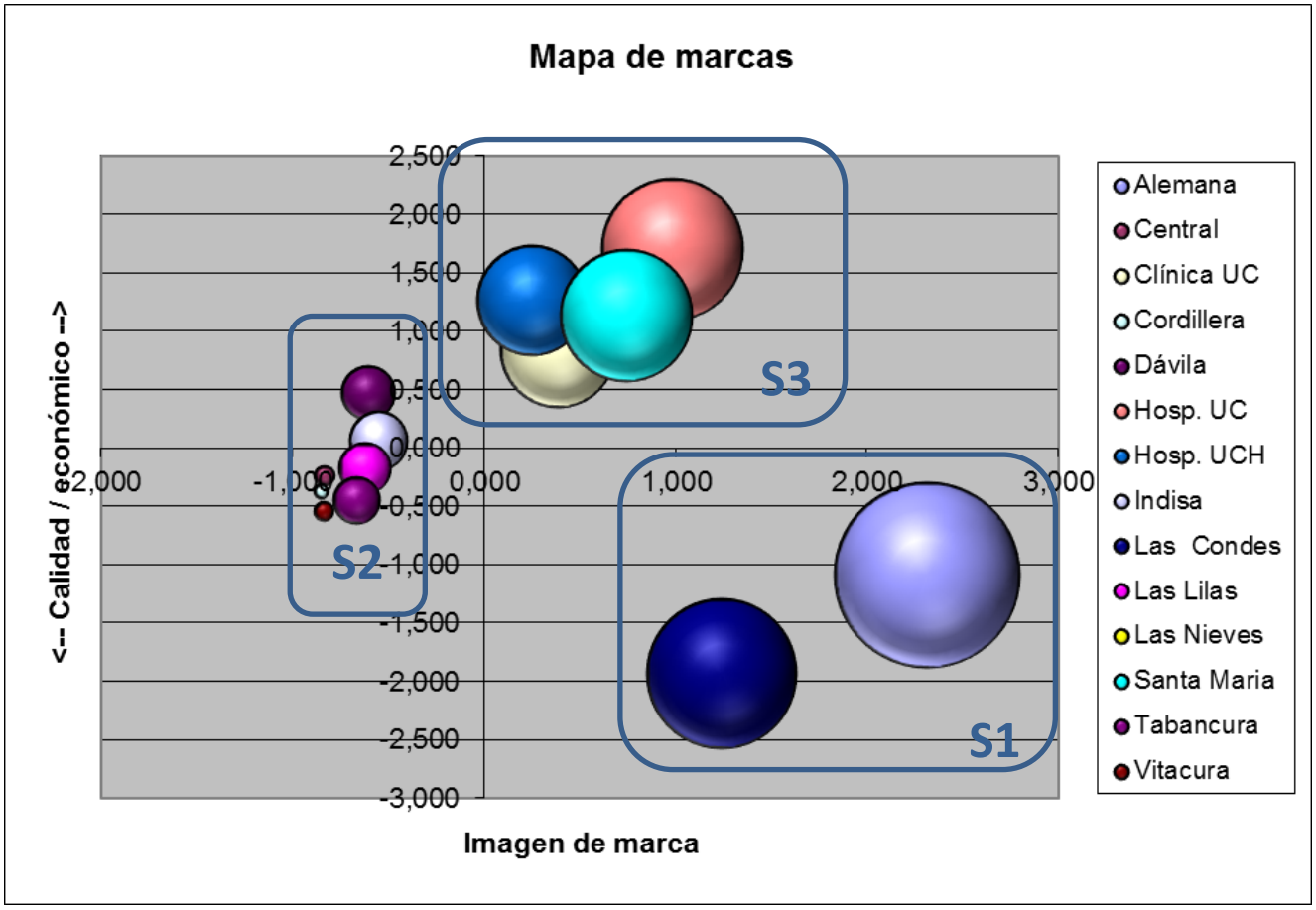
Conglomerado	1	2,000
	2	8,000
	3	4,000
Válidos		14,000
Perdidos		,000

Pertenencia a los conglomerados

Número de caso	CLINICA	Conglomerado	Distancia
1	Alemana	1	,683
2	Central	2	,107
3	Clínica	3	,441
4	Cordille	2	,171
5	Dávila	2	,734
6	Hosp. UC	3	,612
7	Hosp. UCH	3	,344
8	Indisa	2	,360
9	Las Con	1	,683
10	Las Lila	2	,137
11	Las Niev	2	,554
12	Santa Ma	3	,184
13	Tabancur	2	,206
14	Vitacura	2	,307

Ejemplo - Clínicas

- Mapa de marcas:



Ejemplo - Clínicas

Segmento 1: “De calidad y prestigio”

- Clínica Alemana
- Clínica Las Condes

Segmento 2: “Las del montón” (poco diferenciadas)

- Central
- Cordillera
- Dávila
- Indisa
- Las Lilas
- Las nieves
- Tabancura
- Vitacura

Segmento 3: “Reconocidas más económicas”

- Clínica UC
- Hospital UC
- Hospital Universidad de Chile
- Clínica Santa María

Ejemplo - Clínicas

- Usando K-medias con 4 conglomerados:

Centros de los conglomerados finales

	Conglomerado			
	1	2	3	4
REGR factor score 1 for analysis 1	2,30699	,58496	1,23555	-,73530
REGR factor score 2 for analysis 1	-1,07844	1,24495	-1,92710	-,24678

Número de casos en cada conglomerado

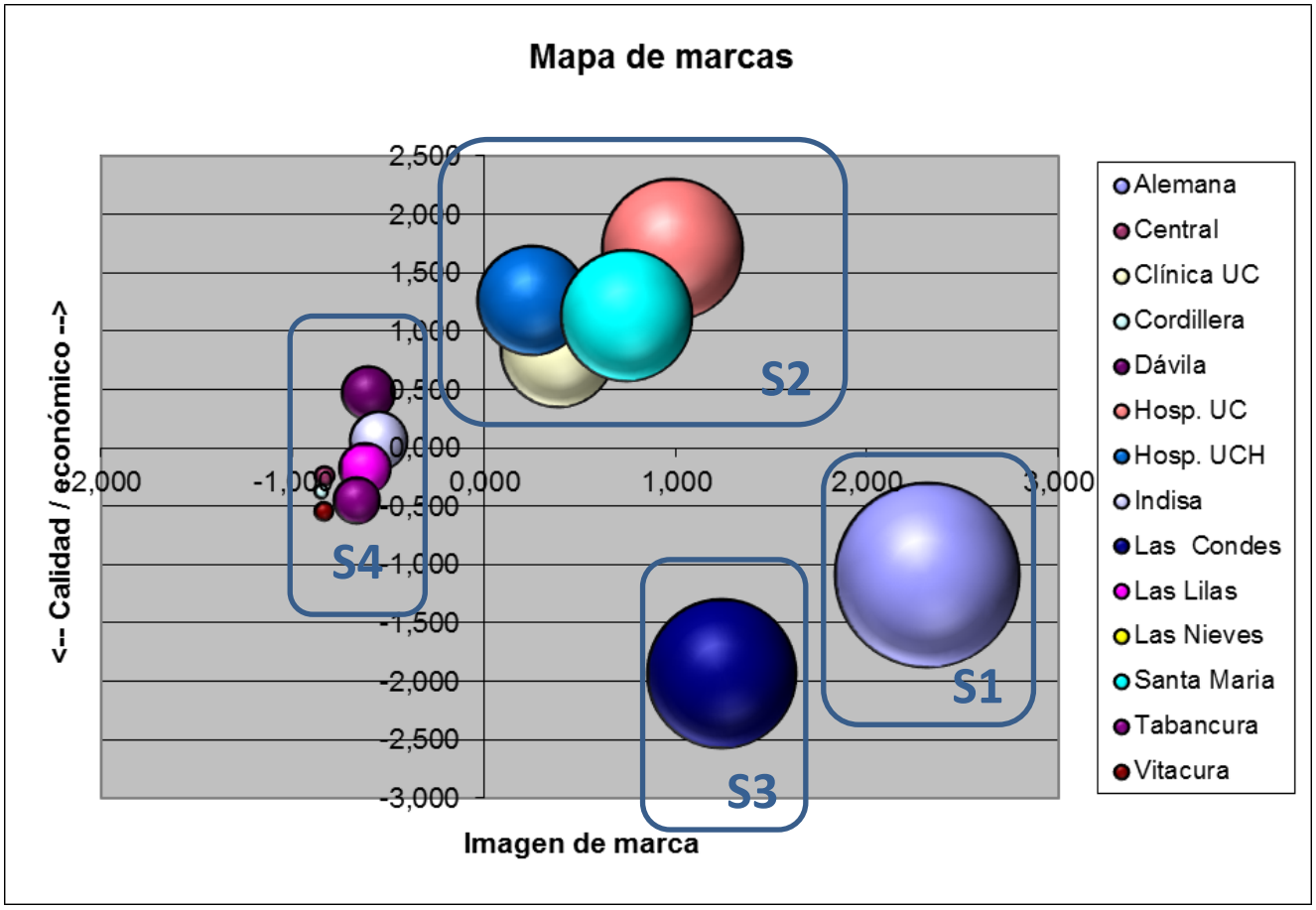
Conglomerado	1	1,000
	2	4,000
	3	1,000
	4	8,000
Válidos		14,000
Perdidos		,000

Pertenencia a los conglomerados

Número de caso	CLINICA	Conglomerado	Distancia
1	Alemana	1	,000
2	Central	4	,107
3	Clínica	2	,441
4	Cordille	4	,171
5	Dávila	4	,734
6	Hosp. UC	2	,612
7	Hosp. UCH	2	,344
8	Indisa	4	,360
9	Las Con	3	,000
10	Las Lila	4	,137
11	Las Niev	4	,554
12	Santa Ma	2	,184
13	Tabancur	4	,206
14	Vitacura	4	,307

Ejemplo - Clínicas

- Mapa de marcas:



Ejemplo - Clínicas

Segmento 1: “De calidad y prestigio”

- Clínica Alemana

Segmento 2: “Reconocidas y más económicas”

- Clínica UC
- Hospital UC
- Hospital Universidad de Chile
- Clínica Santa María

Segmento 3: “Calidad superior”

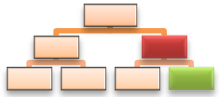
- Clínica Las Condes

Segmento 4: “Las del montón”

- Central
- Cordillera
- Dávila
- Indisa
- Las Lilas
- Las nieves
- Tabancura
- Vitacura

Métodos difusos - Preliminar

- En problemas de conglomerados con muchas variables, analizados desde la perspectiva tradicional:
  - Un objeto en el centro del cluster, ¿tiene el mismo comportamiento que otro de la periferia?.
  - **Problema de cercanía**: dos objetos cercanos a una misma frontera son detectados como teniendo un comportamiento muy distinto.
  - **Problema de lejanía:** dos objetos lejanos entre si son detectados como teniendo un comportamiento identico.
  - En la realidad, pareciera que los objetos no tienen un comportamiento tan discreto.

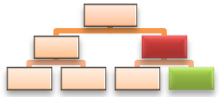


Métodos difusos – Lógica difusa

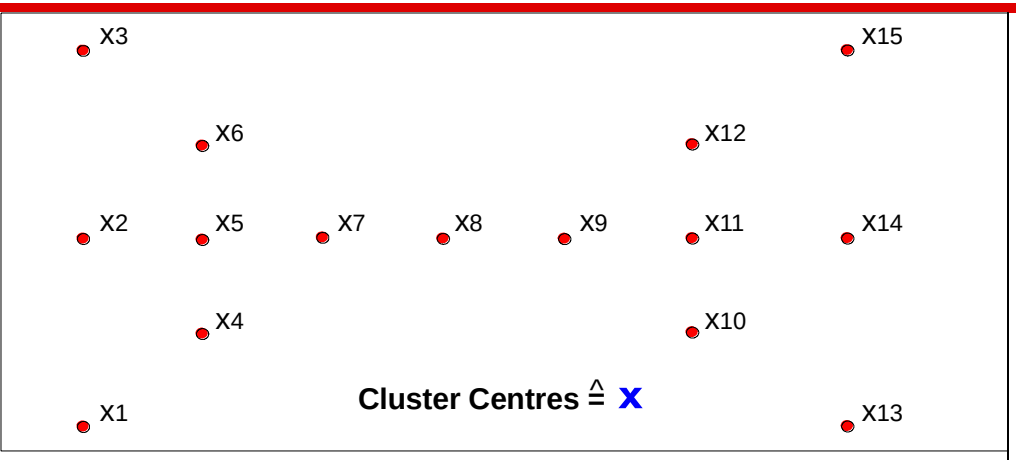
- **Lógica:** Las proposiciones no tienen un valor de verdad verdadero o falso sino que tienen **grados de verdad**.
- **Conjuntos**: Los elementos tienen un grado de pertenencia a los distintos conjuntos.
  - Formalmente, sea X un conjunto tradicional. Un conjunto difuso Y se define como:

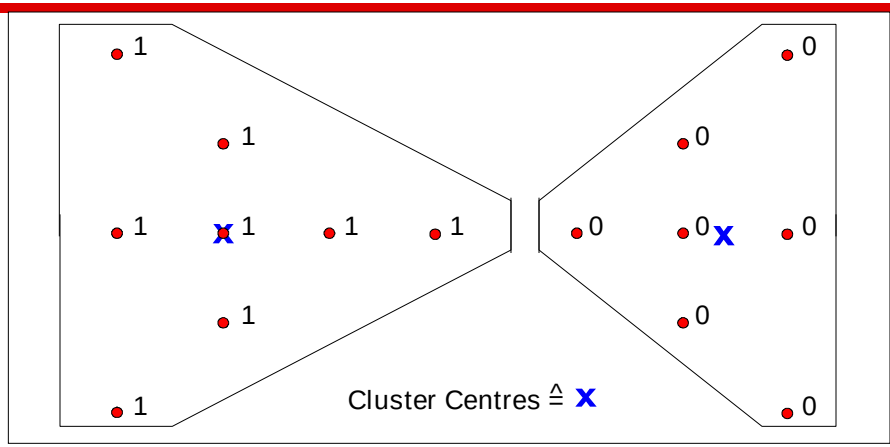
$$Y = \{(x, u_Y(x)); x \in X\}$$

  - Donde $u_Y(x)$ es el grado de pertenencia de el elemento x al conjunto difuso Y.

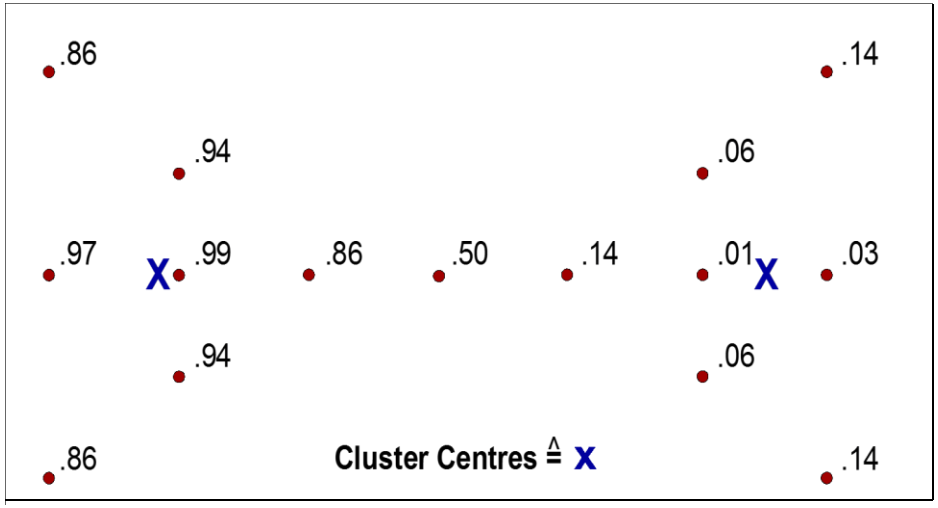


Métodos difusos – Ejemplo lógica difusa

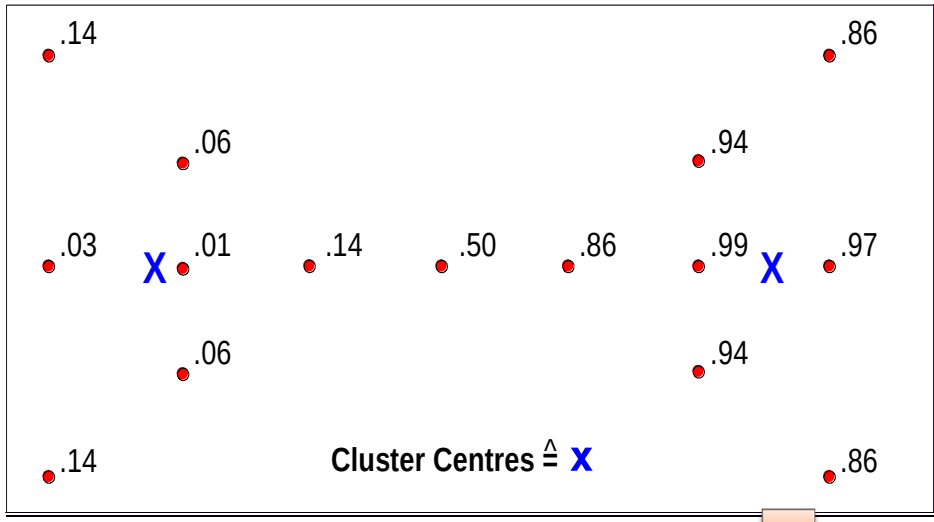




Grupos “estándar”



Grupo difuso 1



Grupo difuso 2

Métodos difusos – Fuzzy C-means

- La idea es muy similar al algoritmo de K medias, pero la idea es en cada iteración asignar un grado de pertenencia de acuerdo a las distancias a los centroides.

- Los centroides se calculan ponderando el valor de los atributos por los grados de pertenencia de los objetos.