Universidad Popular del Cesar

Ciencia de Datos

Introducción al Clustering Análisis de Conglomerados

Dr. José Ramón Iglesias
DSP-ASIC BUILDER GROUP
Director Semillero TRIAC
Ingenieria Electronica
Universidad Popular del Cesar

Mapa-Receta

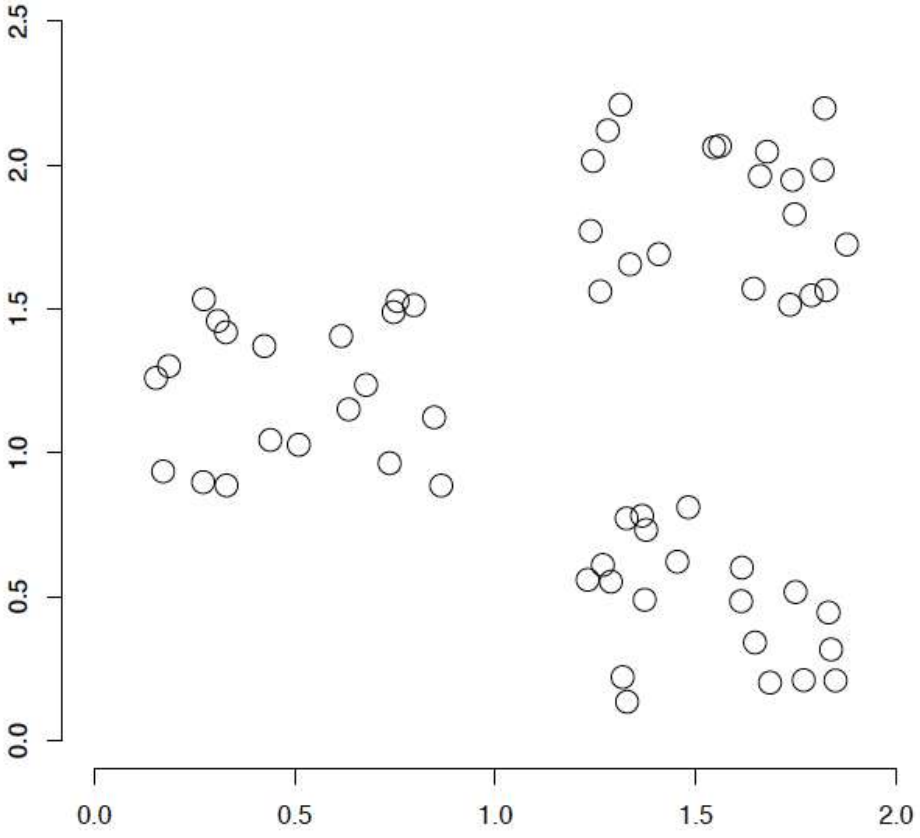
1. Intuición general de clustering
2. Conocimiento de los Datos e Información Relevante al problema
3. Importancia del conocimiento de dominio
4. Similaridad y/o Distancia entre datos
5. Algoritmos de agrupamiento
6. Evaluación de resultados: Visualización, Medidas y relevancia: utilidad o impacto

Cómo funciona clustering

Agrupar objetos semejantes, parecidos

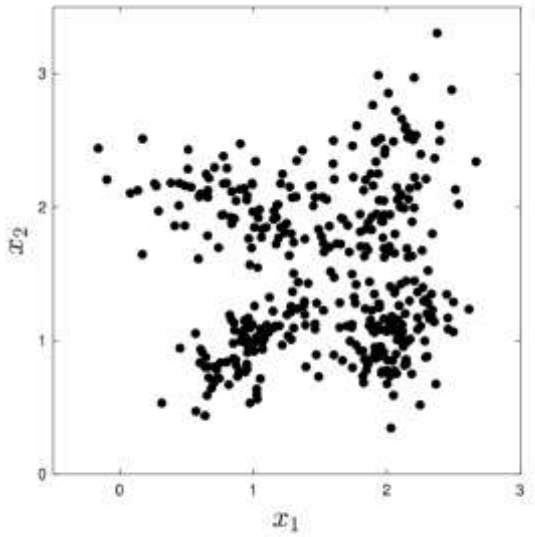
- Entrada: n objetos o individuos en un espacio m-dimensional:
  X en R^{nxm}, cada fila representa un objeto (vector con m valores)
- Salida: una **solución** con conglomerados (**clusters**) de objetos semejantes
  (semejantes ➔ cercanos en el espacio o similares)
  - Se minimiza la distancia entre los objetos de un mismo conglomerado
  - Se maximiza la distancia entre los objetos de distintos conglomerados

Dataset con clara estructura de clusters



¿Cómo sería un algoritmo para encontrar clusters en este espacio?

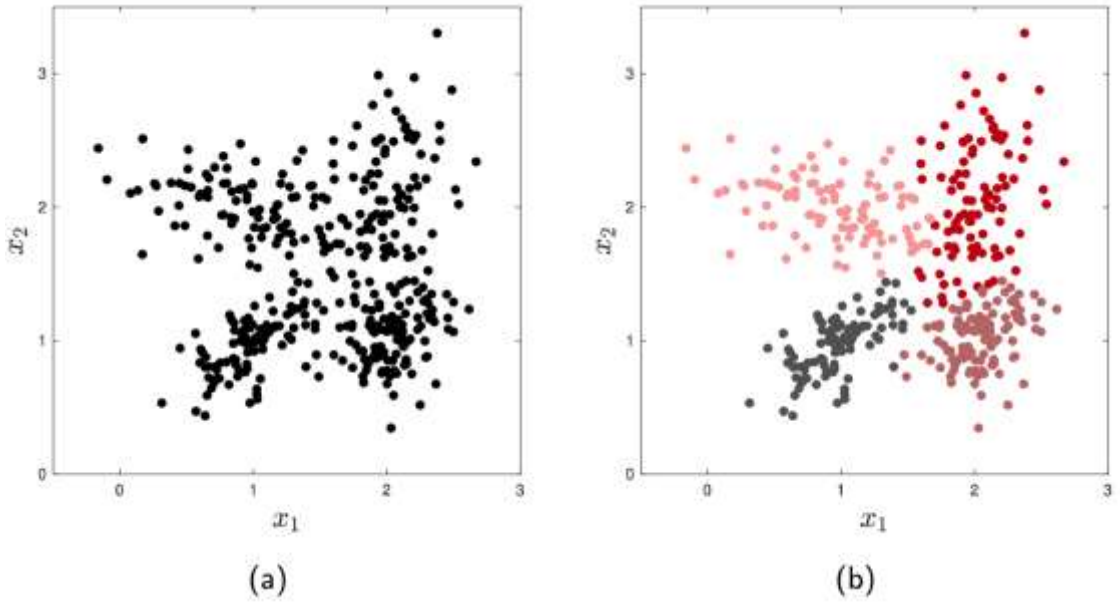
Dataset con no tan clara estructura de clusters



(a)

¿Cómo sería un algoritmo para encontrar clusters en este espacio?

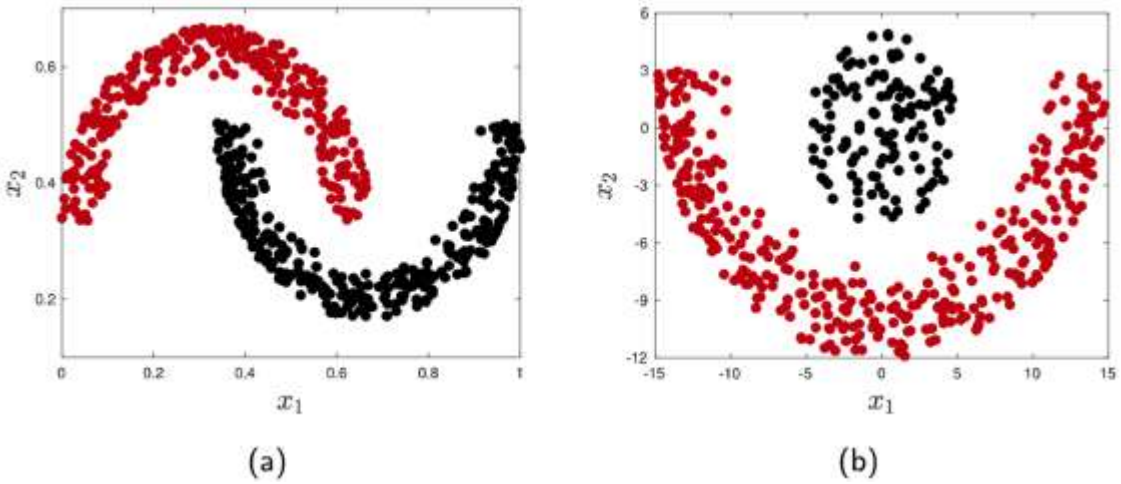
Dataset con no tan clara estructura de clusters



¿Cómo sería un algoritmo para encontrar clusters en este espacio?

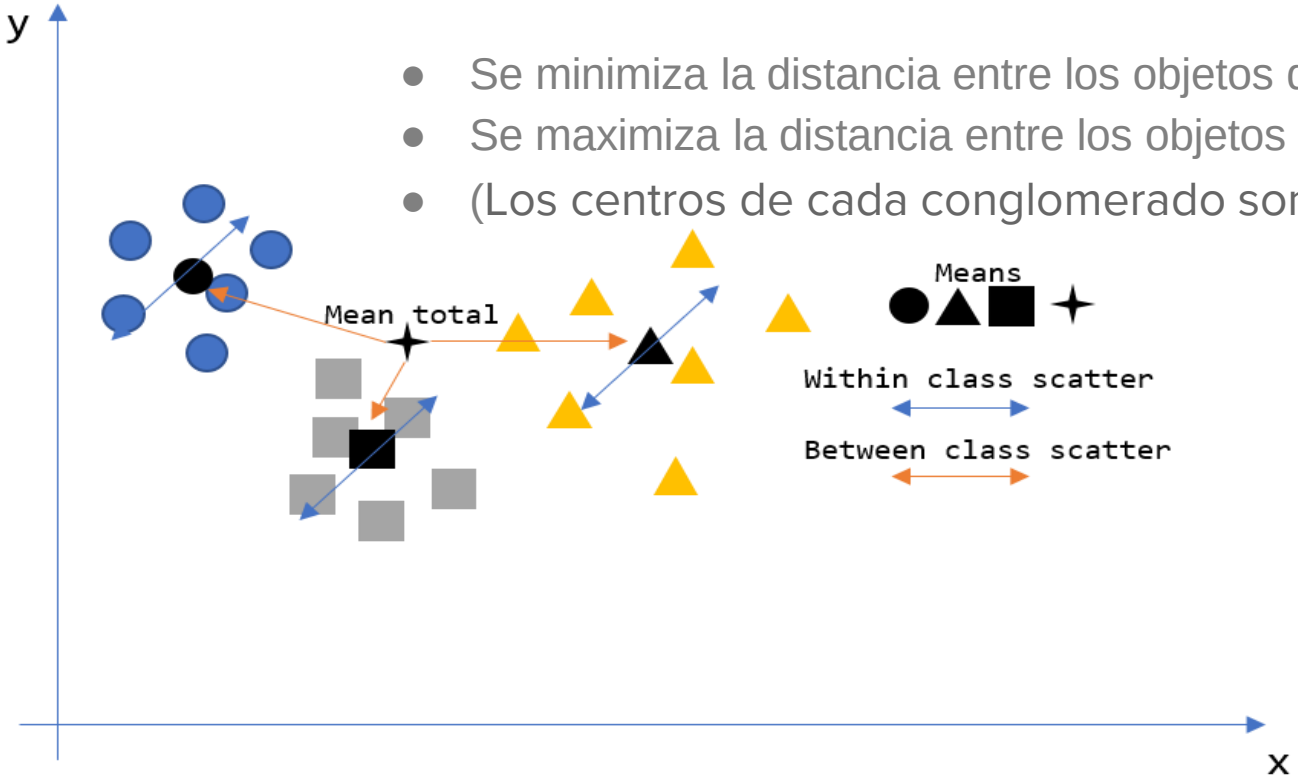
Dataset con clara estructura de clusters

¿Cómo sería un algoritmo para encontrar clusters en este espacio?



Cómo funciona clustering

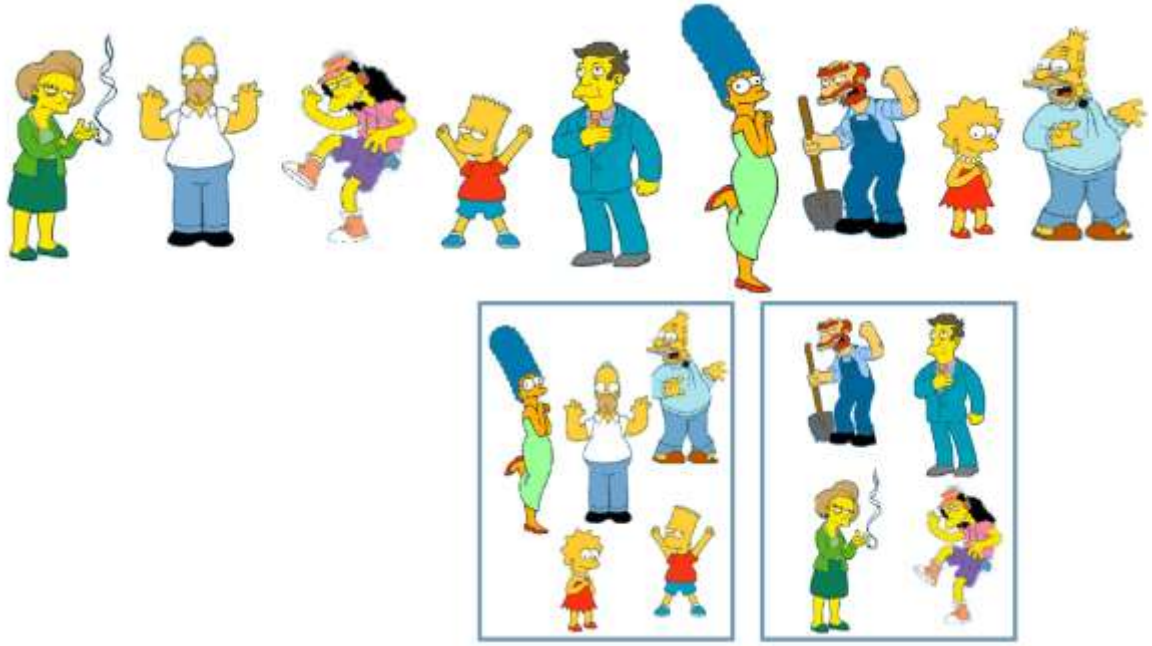
- Se minimiza la distancia entre los objetos de un mismo grupo
- Se maximiza la distancia entre los objetos de distintos clusters
- (Los centros de cada conglomerado son los **centroides)**



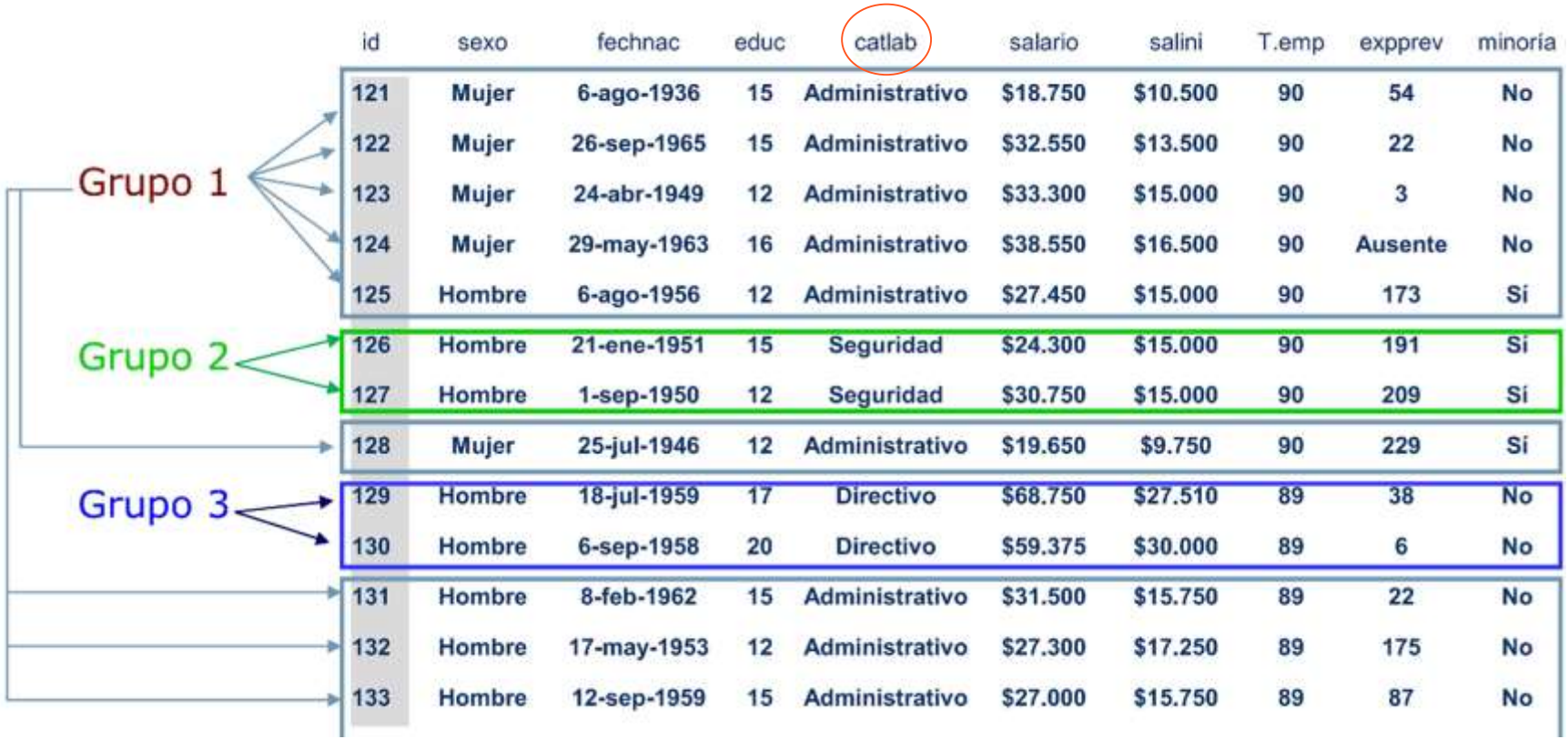
Mis Datos...

¿Cómo los agrupo?

Datos agrupados según algún criterio



Datos



id	sexo	fechnac	educ	catlab	salario	salini	T.emp	expprev	minoría
121	Mujer	6-ago-1936	15	Administrativo	$18.750	$10.500	90	54	No
122	Mujer	26-sep-1965	15	Administrativo	$32.550	$13.500	90	22	No
123	Mujer	24-abr-1949	12	Administrativo	$33.300	$15.000	90	3	No
124	Mujer	29-may-1963	16	Administrativo	$38.550	$16.500	90	Ausente	No
125	Hombre	6-ago-1956	12	Administrativo	$27.450	$15.000	90	173	Si
126	Hombre	21-ene-1951	15	Seguridad	$24.300	$15.000	90	191	Si
127	Hombre	1-sep-1950	12	Seguridad	$30.750	$15.000	90	209	Si
128	Mujer	25-jul-1946	12	Administrativo	$19.650	$9.750	90	229	Si
129	Hombre	18-jul-1959	17	Directivo	$68.750	$27.510	89	38	No
130	Hombre	6-sep-1958	20	Directivo	$59.375	$30.000	89	6	No
131	Hombre	8-feb-1962	15	Administrativo	$31.500	$15.750	89	22	No
132	Hombre	17-may-1953	12	Administrativo	$27.300	$17.250	89	175	No
133	Hombre	12-sep-1959	15	Administrativo	$27.000	$15.750	89	87	No

¿Cómo es el espacio? ¿Cómo represento mis objetos y objetivos?

- Es multi dimensional?
- Mis datos son naturalmente categóricos? ordinales? continuos?
- Tengo información que me permita decir que debería encontrar grupos compactos?
- No se nada y quiero usar clustering en forma exploratoria

¿Cómo se calculan las similaridades entre objetos en este espacio?

- Es un espacio Euclídeo? Métrica usual anda bien? Conviene usar otro tipo de medidas? ángulos en vez de distancias?
- No es un espacio Euclídeo? Similaridades? Matriz de afinidad?

- Entender mi espacio me ayuda a elegir un método más razonable.
- Si mi método más razonable no me da nada, quizás sea porque no hay nada para ver...

Notebook

Demo EL460_clustering_1_fifa2019.ipynb

Cuestiones cruciales

- ¿Cómo es el espacio? ¿Cómo represento mis problemas?
- ¿Cómo se calcula la distancia (o semejanza) en este espacio?
- ¿Conozco cuántos clusters quiero distinguir?
- ¿Qué distribución tienen estos clusters? ¿Gaussiana?
- ¿Busco una estructura jerárquica o plana?
- ¿Cómo veo qué hay en cada cluster?
- ¿Cómo evalúo la bondad de cada solución?

Medidas de similaridad

- A la hora de calcular la similaridad entre dos objetos
  - no hace falta usar todas las variables
  - hay que tener cuidado con las magnitudes de cada variable

- No será posible que todas las variables tengan valores similares dentro de un mismo grupo, por lo que habrá que usar una medida de similaridad global entre elementos de un mismo grupo.

Entender el contexto

- Tengo claro cuantos clusters quiero distinguir?
- Tengo información? Hice varios experimentos? tengo varias databases de días diferentes y locales diferentes?
- Interacción con el experto de dominio!
- Clustering exploratorio, debo estudiar los distintos agrupamientos para distintos números de clusters.
- Considerar distintas técnicas para encontrar el mejor modelo de agrupamiento.

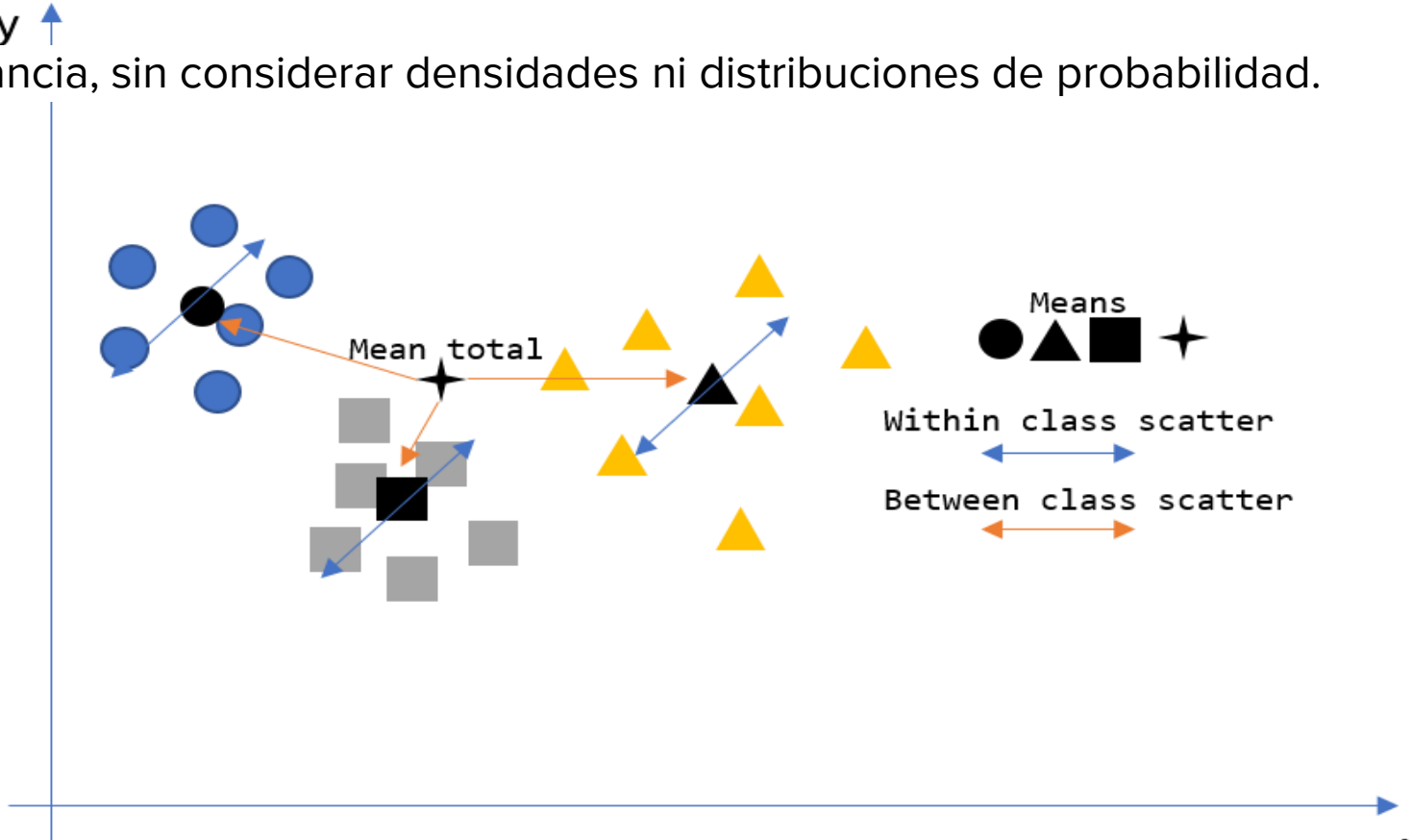
¿Busco una estructura jerárquica o plana?

- Si mis clusters están anidados, tengo una estructura muy fuerte que explica los datos

- Si mis clusters son estructuras cercanas, las une sin remedio

Estructura plana: K-means por ejemplo

- Usa distancia, sin considerar densidades ni distribuciones de probabilidad.

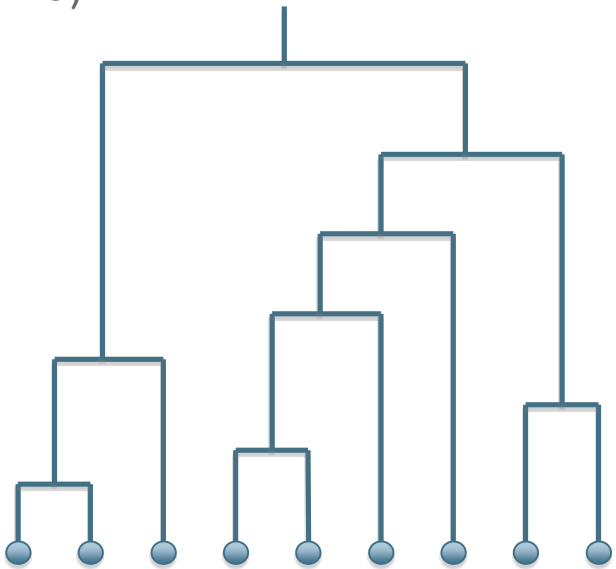


Clustering jerárquico

Algoritmos jerárquicos que generan una

taxonomía jerárquica de clusters (dendrograma)

- Interpretación más rica
- Más difícil de interpretar

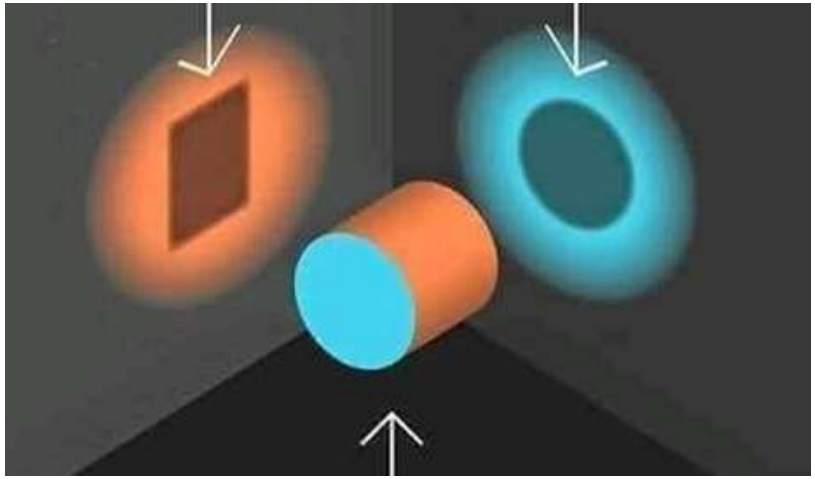


Semejanzas, Distancias y Afinidades

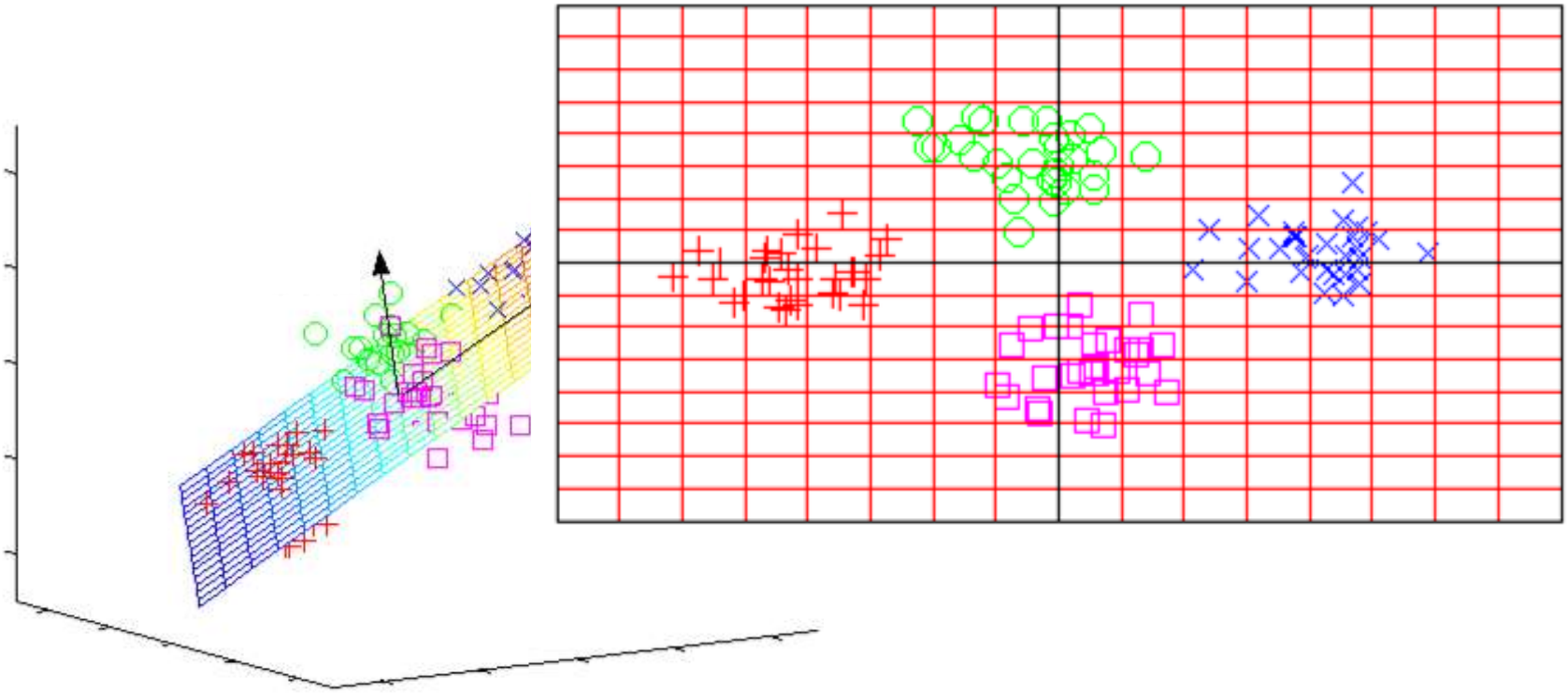
- La semejanza debería acercarse a las causas latentes
  - Entre documentos: semántica
  - Entre clientes: motivación para las compras
  - Entre imágenes: objetos físicos que representan
  - Entre propiedades inmobiliarias: elementos que otorgan valor
- Idealmente, debería calcularse de forma independiente para cada dimensión

¿Cómo veo qué hay en cada cluster? Datos numéricos, muchas variables

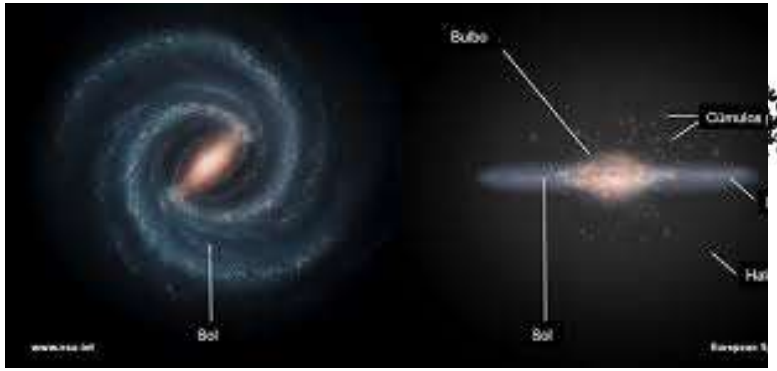
- Visualización es una pesadilla. Rápido de correr, lento de analizar!!!



Embeddings



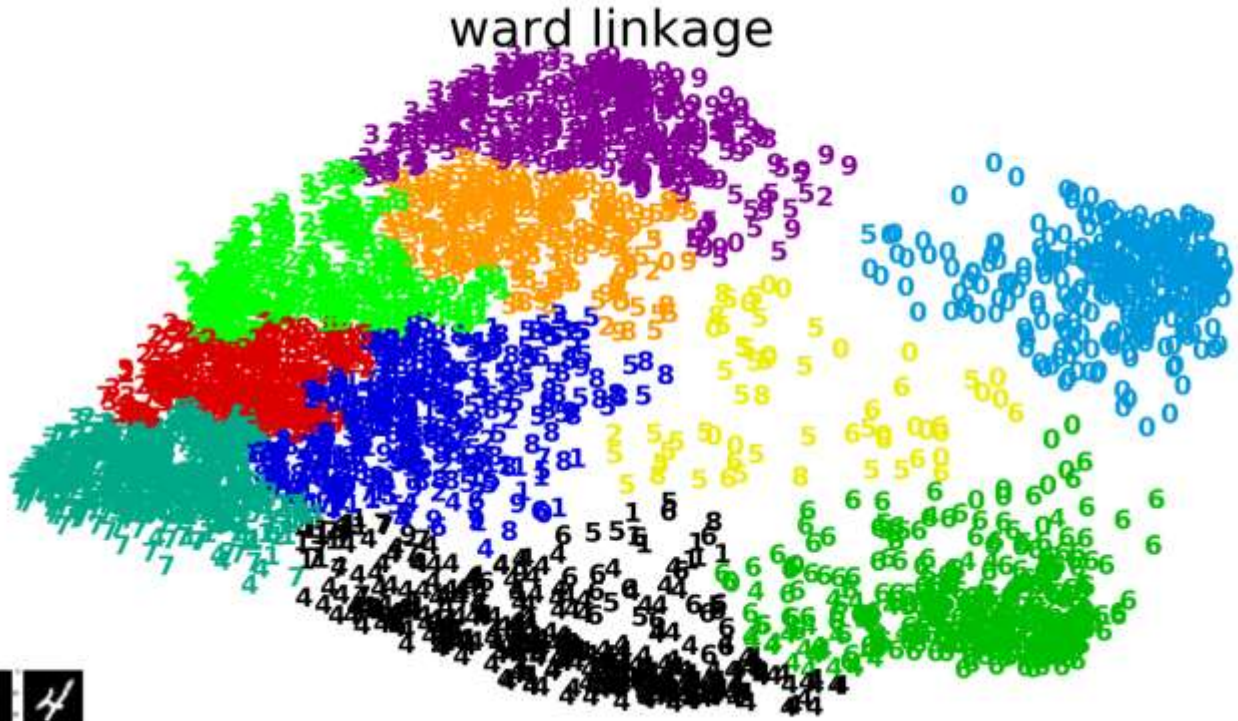
¿Cómo veo qué hay en cada cluster? Datos numéricos, muchas variables



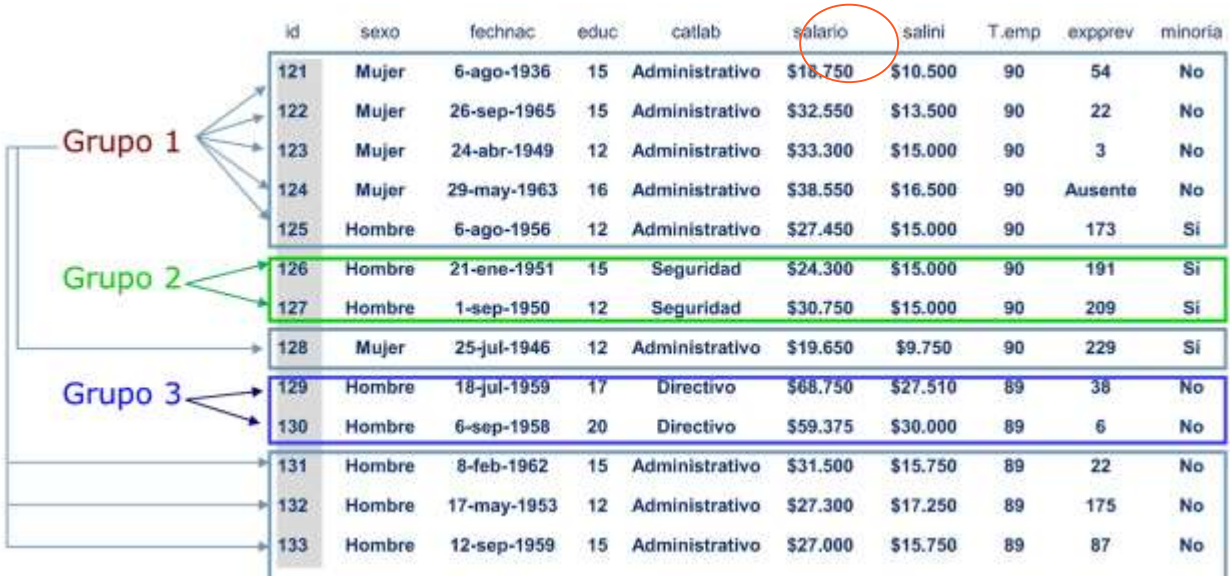
- Proyecciones en espacios de menor dimensión ayudan a visualizar los resultados.
  - Proyecciones, transformaciones
  - Principal component analysis (PCA),
  - t-distributed Stochastic Neighbor Embedding (t-SNE)

¿Cómo veo qué hay en cada cluster? Datos numéricos, muchas variables

Spectral embedding



Datos categóricos o mixtos



id	sexo	fechnac	educ	catlab	salario	salini	T.emp	expprev	minoría
121	Mujer	6-ago-1936	15	Administrativo	$18.750	$10.500	90	54	No
122	Mujer	26-sep-1965	15	Administrativo	$32.550	$13.500	90	22	No
123	Mujer	24-abr-1949	12	Administrativo	$33.300	$15.000	90	3	No
124	Mujer	29-may-1963	16	Administrativo	$38.550	$16.500	90	Ausente	No
125	Hombre	6-ago-1956	12	Administrativo	$27.450	$15.000	90	173	Sí
126	Hombre	21-ene-1951	15	Seguridad	$24.300	$15.000	90	191	Sí
127	Hombre	1-sep-1950	12	Seguridad	$30.750	$15.000	90	209	Sí
128	Mujer	25-jul-1946	12	Administrativo	$19.650	$9.750	90	229	Sí
129	Hombre	18-jul-1959	17	Directivo	$68.750	$27.510	89	38	No
130	Hombre	6-sep-1958	20	Directivo	$59.375	$30.000	89	6	No
131	Hombre	8-feb-1962	15	Administrativo	$31.500	$15.750	89	22	No
132	Hombre	17-may-1953	12	Administrativo	$27.300	$17.250	89	175	No
133	Hombre	12-sep-1959	15	Administrativo	$27.000	$15.750	89	87	No

Análisis Descriptivo de cada grupo encontrado

Tabla de contingencia grupos vs. alguna categórica

Cómo evalúo la bondad de cada solución

- Experto de Dominio, utilidad, relevancia
- Comparación de métodos
  - Rand measure
  - Mutual Information score
  - Contingency Matrix
  - Silhouette
- Para un método fijo,
  - Elbow method
  - BIC, AIC