

Minería de Datos

IIC2433

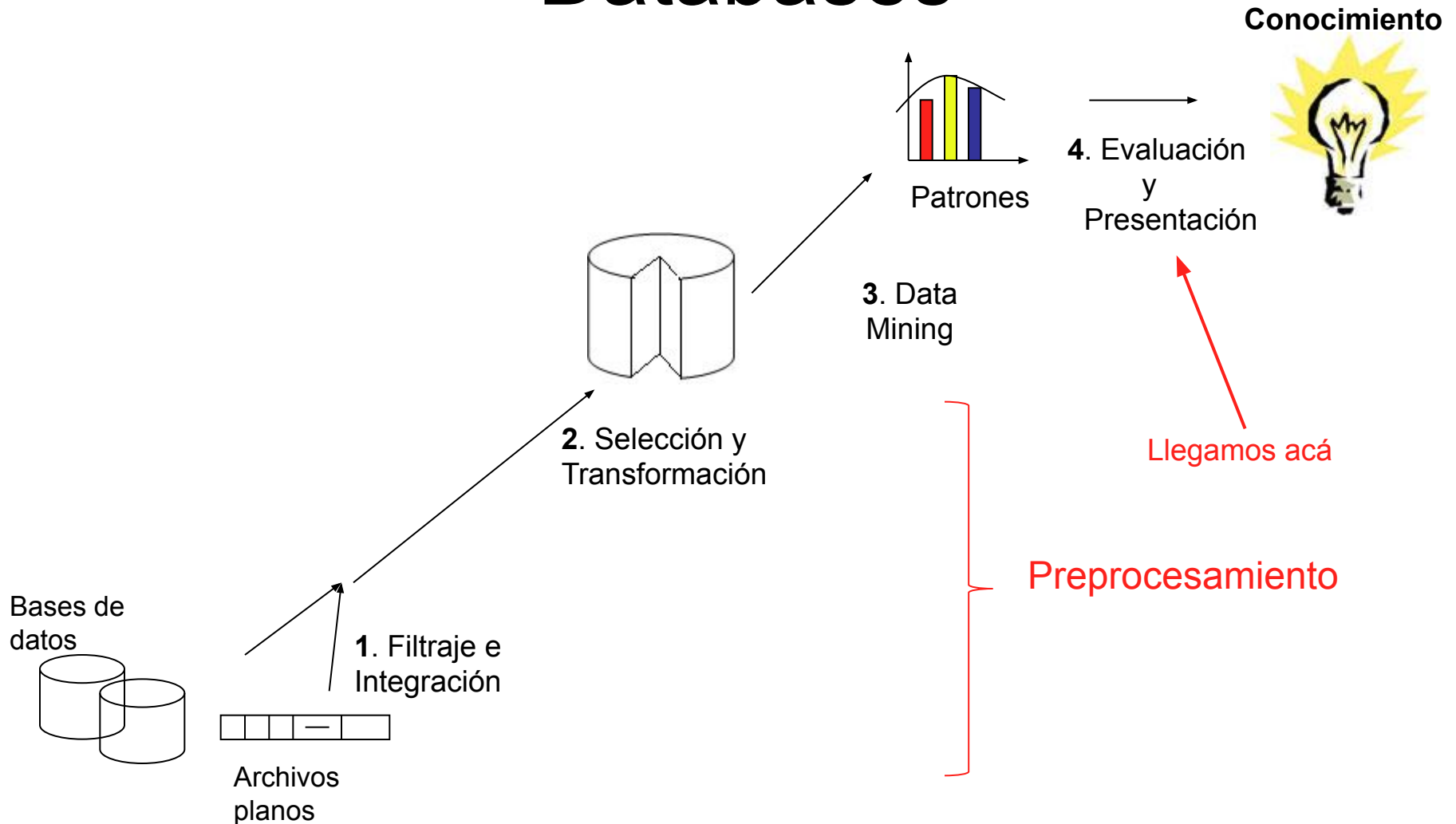
Validación de Clustering

Vicente Domínguez

¿Qué veremos esta clase?

- Como validar nuestros resultados obtenidos en el proceso de clustering clustering

Knowledge Discovery in Databases



¿Es necesario validar los clusters?

- Por lo menos en Clasificación, la validación es parte integral del proceso
- No así en Clustering...

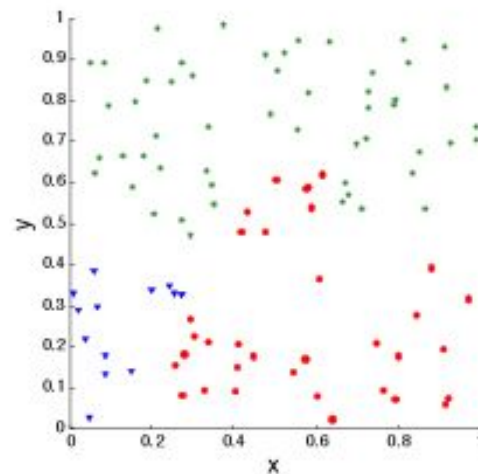
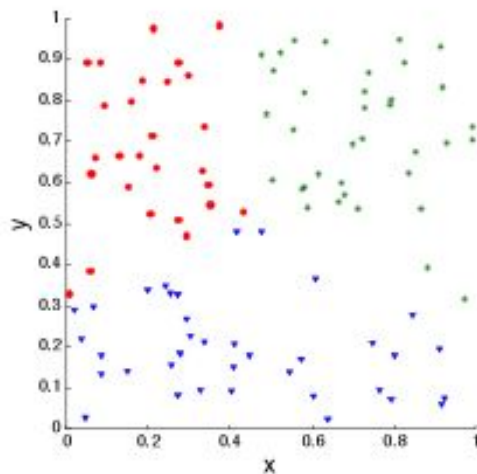
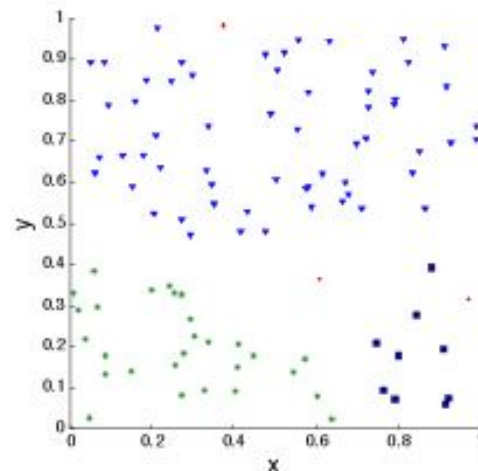
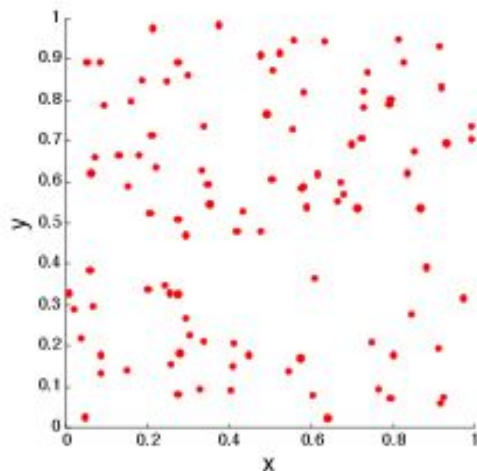
¿Cómo saber si nuestros clusters son buenos?

- No hay una respuesta absoluta
- Depende de la aplicación
- ¿Entonces, para qué evaluar?

Evaluamos para:

- Evitar encontrar patrones en el ruido
- Para comparar algoritmos de clustering diferentes
- Para comparar conjuntos de clusters diferentes
- Para comprar dos clusters

Clusters en datos aleatorios



Aspectos de la validación

- Determinar la **tendencia de agrupamiento** (clustering tendency), i.e.: si existe una estructura no-aleatoria en los datos
- Encontrar el número correcto de clusters
- Evaluar qué tan bien los resultados se ajustan a los datos (sin consultar datos externos)
- Comparar resultados con resultados externos, i.e.: clases asignadas manualmente
- Comparar dos conjuntos de clusters para saber cuál es mejor

Medidas de validez

- **External Index** (Supervisado)
- **Internal Index** (No-Supervisado)
- **Relative Index** (Relativo)

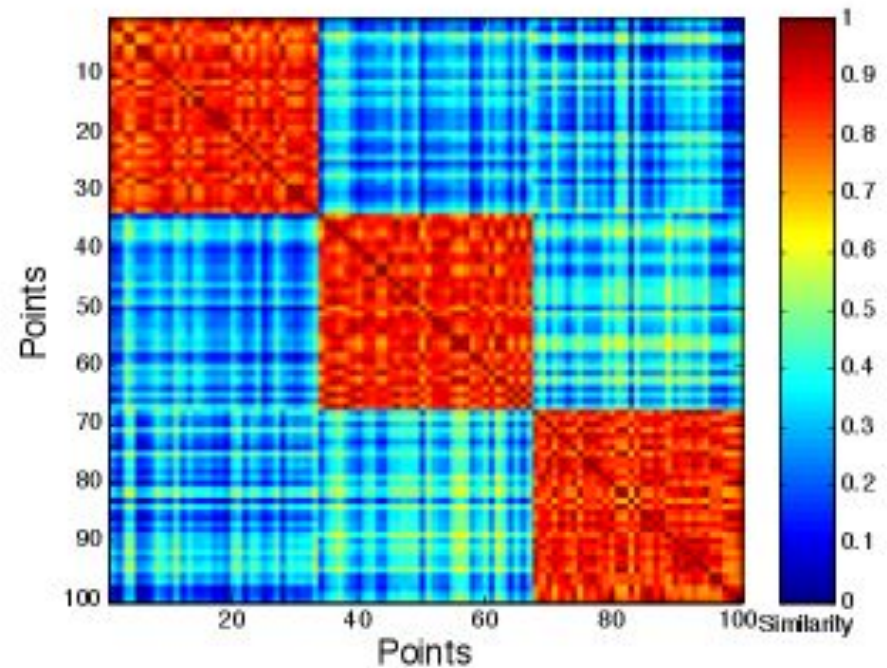
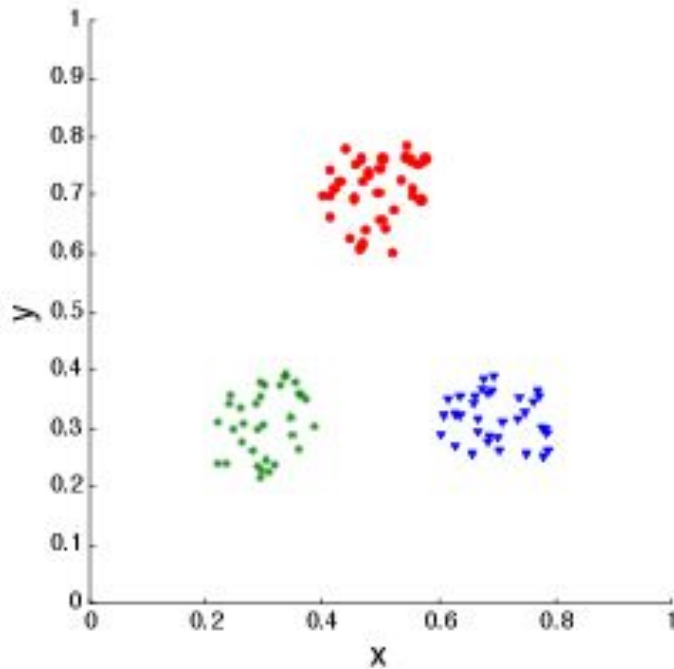
Concepto: Matriz de similitud

	G1	G2	G3	G4
G1	1	0.83	0	0
G2	0.83	1	0	0
G3	0	0	1	0.32
G4	0	0	0.32	1

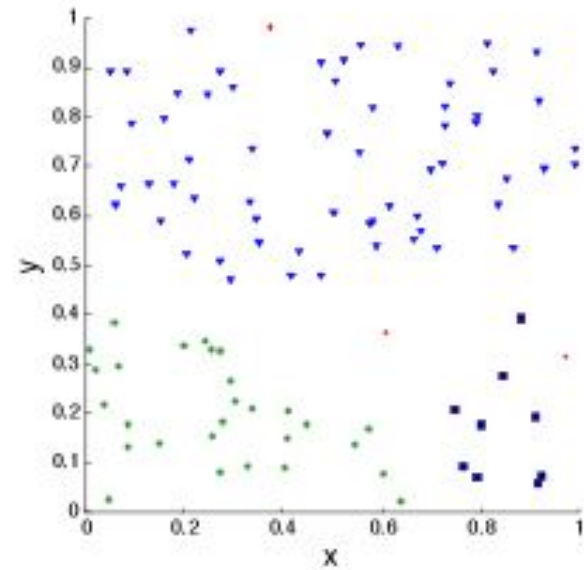
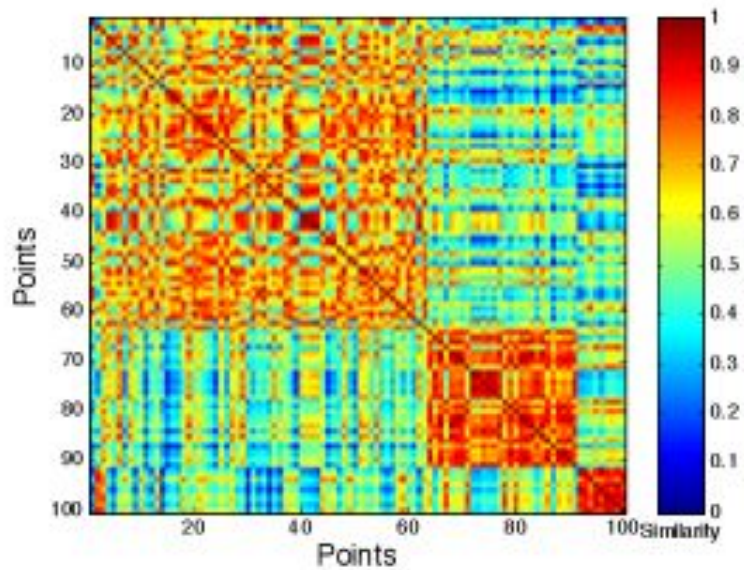
Enfoque visual

- Ordenar la matriz de similitud con respecto a etiquetas de clusters e inspeccionar visualmente

Visualizando la matriz de similitud (clusters reales)

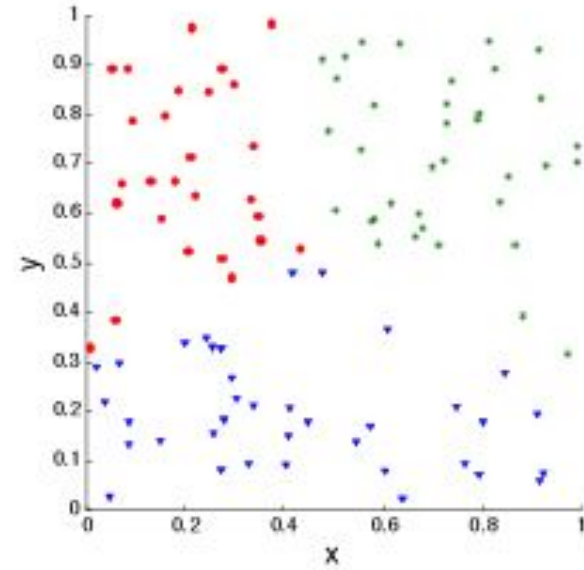
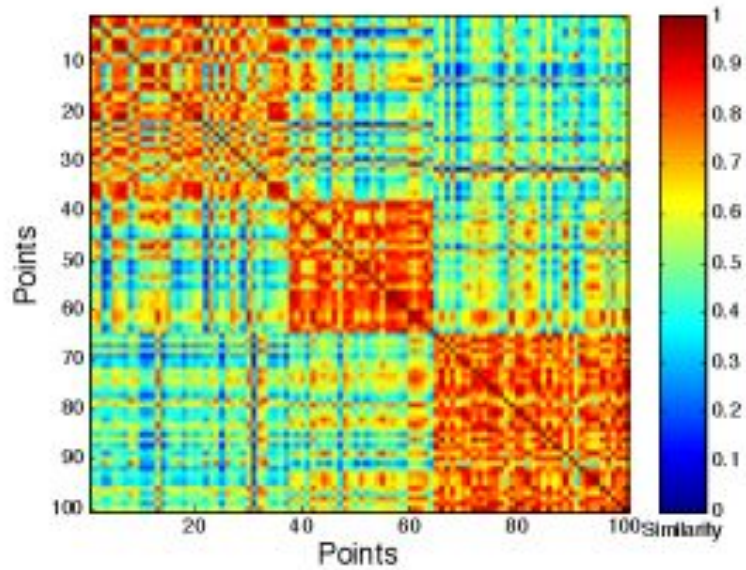


Visualizando clusters sobre datos aleatorios

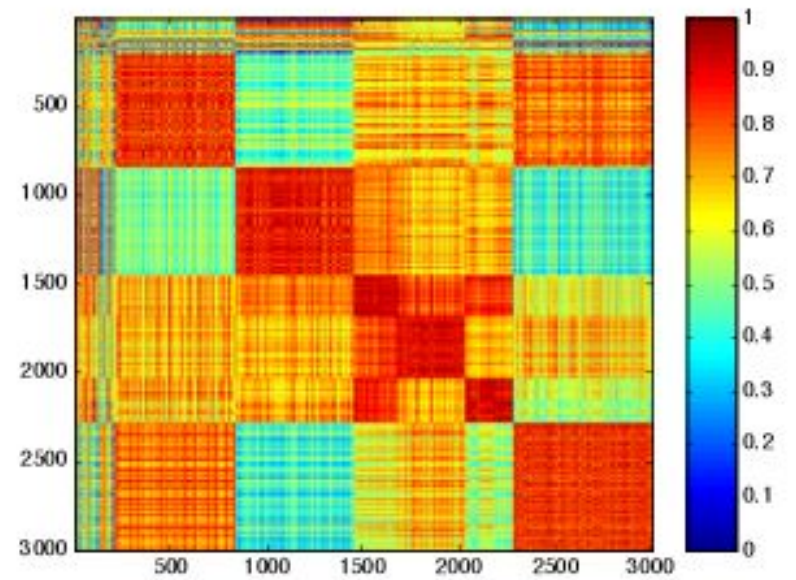
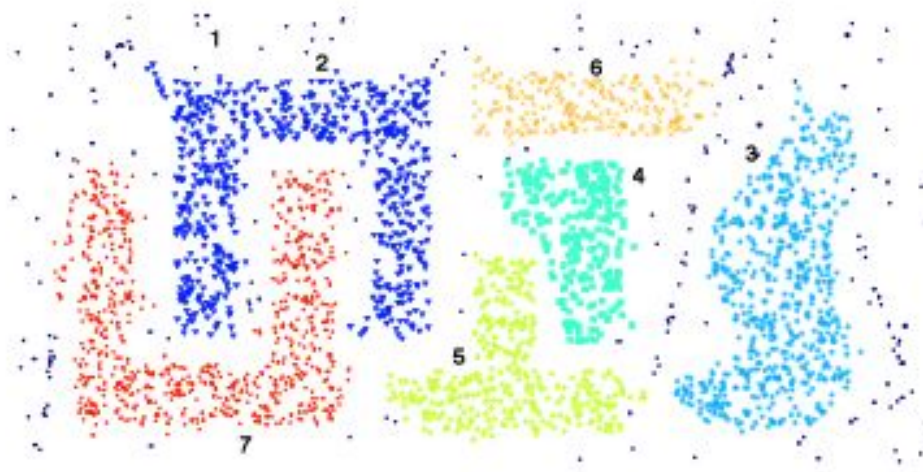


DBSCAN

Visualizando clusters sobre datos aleatorios



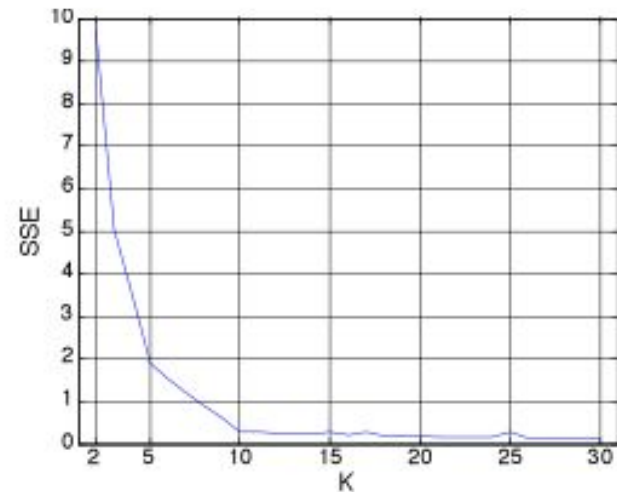
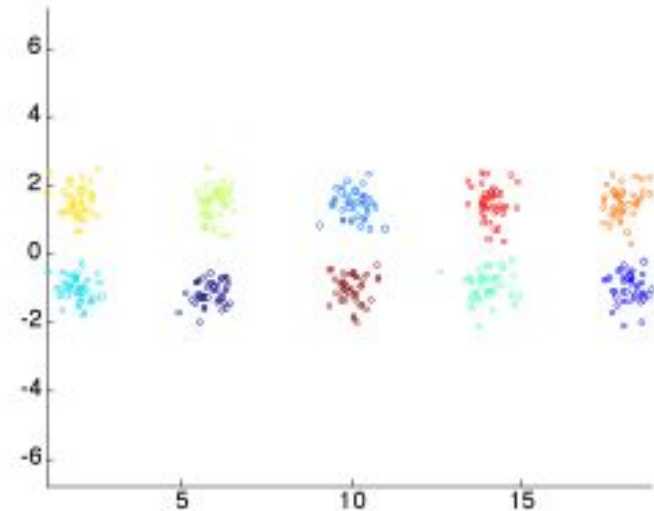
K-means



DBSCAN

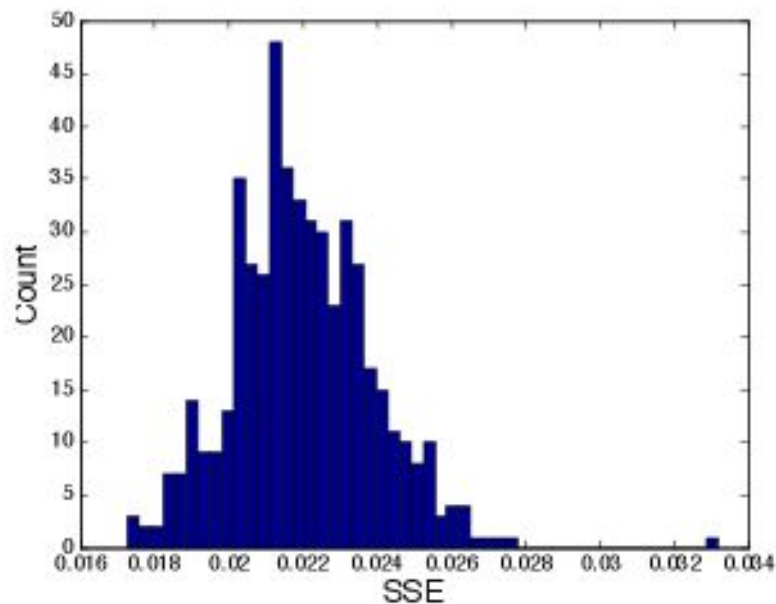
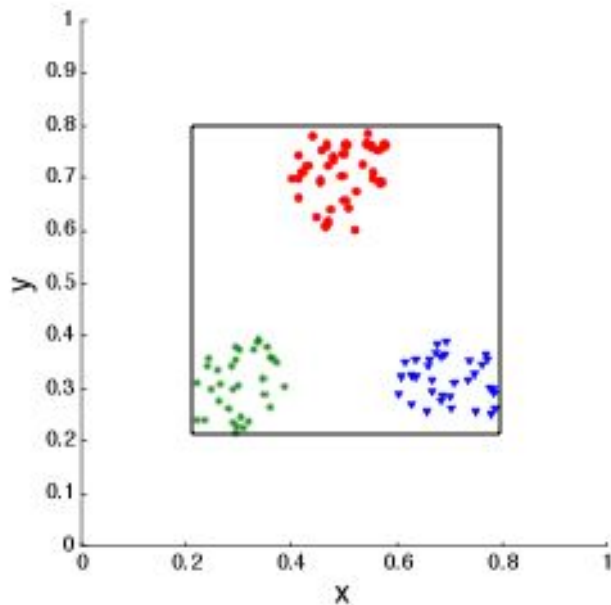
Medidas internas: SSE

- Clusters en figuras más complicadas no están bien separados
- Índice interno: SSE
- Permite comparar 2 clusters, o 2 soluciones de clustering
- Permite estimar el número de clusters



Metodología: Ejemplo SSE

- Comparar $SSE = 0.005$ contra 3 clusters de datos aleatorios
- Histograma muestra distribución SSE para 500 sets de datos aleatorios (100 puntos), en el mismo rango



Medidas internas: Cohesión y separación

- **Cohesión de clusters:** mide qué tan cercanos son los objetos en un cluster (ej: SSE)
- **Separación de clusters:** mide qué tan diferente o bien separado es un cluster de otros

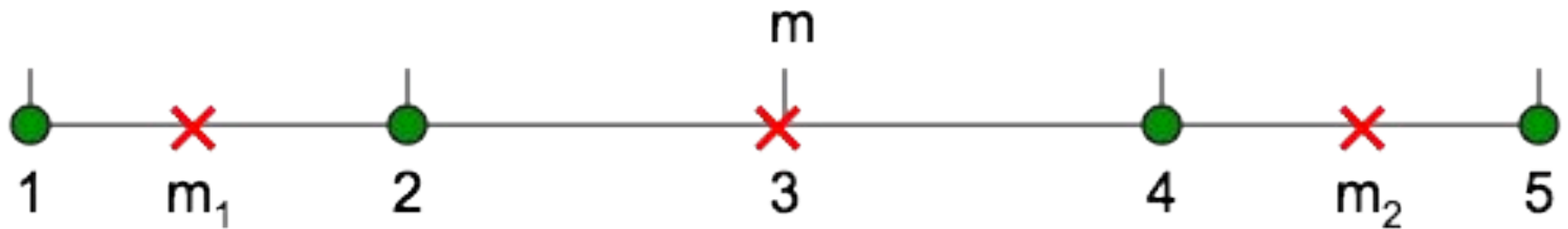
EJ. (SSE) Cohesión y Separación

- Cohesión se mide como **within cluster sum of squares (WSS o SSE)**

$$WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$$

- Separación se mide como **between cluster sum of squares (BSS)**

$$BSS = \sum_i |C_i| (m - m_i)^2$$



K=1 cluster:

$$WSS = (1 - 3)^2 + (2 - 3)^2 + (4 - 3)^2 + (5 - 3)^2 = 10$$

$$BSS = 4 \times (3 - 3)^2 = 0$$

$$Total = 10 + 0 = 10$$

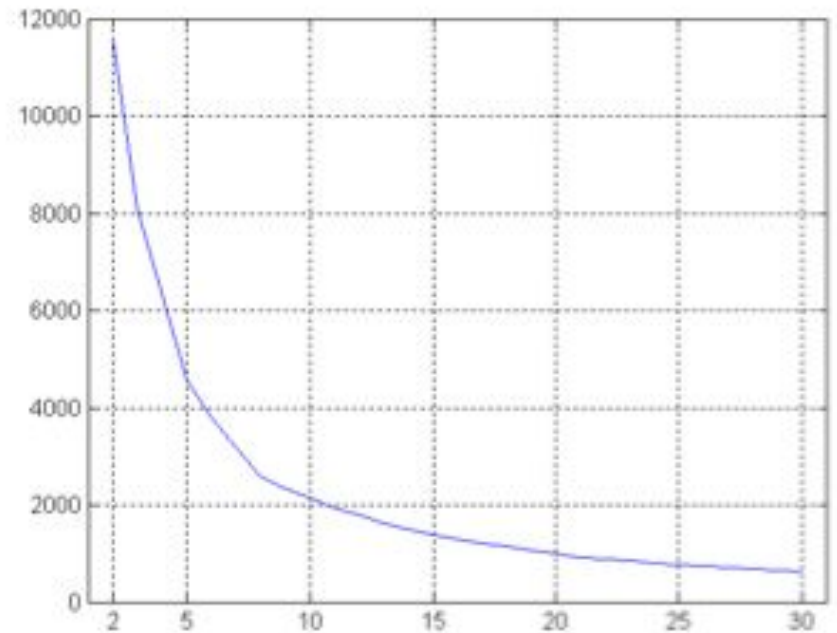
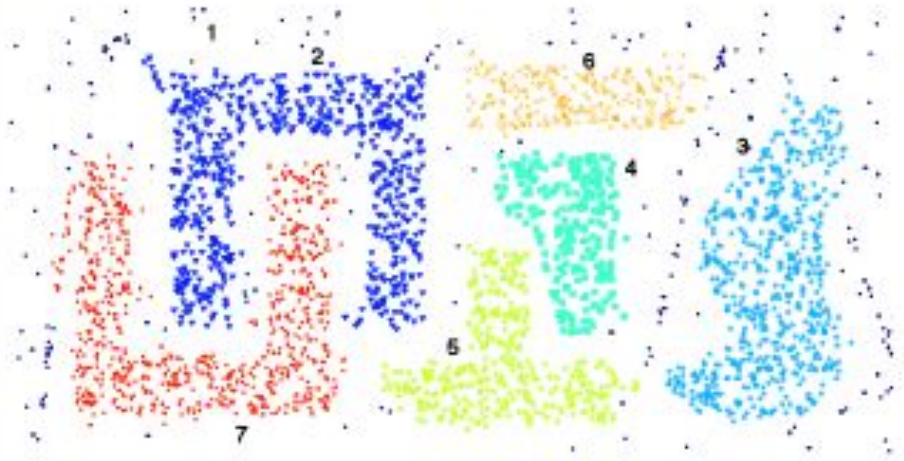
K=2 clusters:

$$WSS = (1 - 1.5)^2 + (2 - 1.5)^2 + (4 - 4.5)^2 + (5 - 4.5)^2 = 1$$

$$BSS = 2 \times (3 - 1.5)^2 + 2 \times (4.5 - 3)^2 = 9$$

$$Total = 1 + 9 = 10$$

Curva SSE para un dataset más complicado



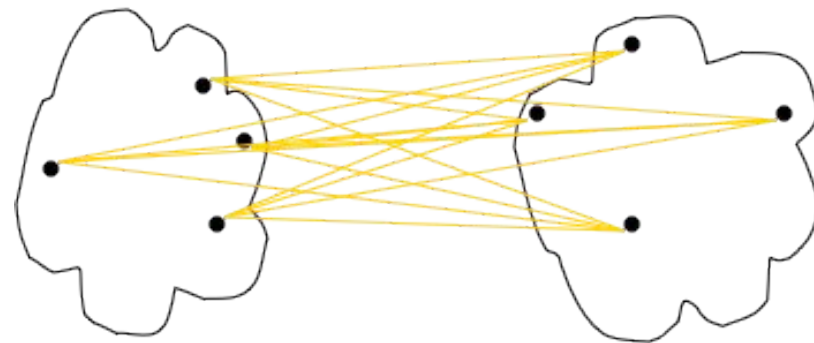
SSE of clusters found using K-means

Medidas internas: Cohesión y separación

- Enfoque basado en grafos de proximidad
- Cohesión: suma de los pesos de todos los arcos en un cluster
- Separación: suma de los pesos entre nodos del cluster y de otros clusters



cohesion



separation

Medidas Externas: Pureza y Entropía

- **Pureza:** Nivel en que un cluster contiene elementos de una sólo clase (se usa la clase predominante)
- **Entropia:** Cantidad de clases diferentes que contiene un cluster

Medidas Externas: Pureza y Entropía

Table 1: Entropy and Purity in CLUTO

	Entropy	Purity
Single Cluster	$E(S_r) = -\frac{1}{\log q} \sum_{i=1}^q \frac{n_r^i}{n_r} \log \frac{n_r^i}{n_r}$	$P(S_r) = \frac{1}{n_r} \max_i (n_r^i)$
Overall	$Entropy = \sum_{r=1}^k \frac{n_r}{n} E(S_r)$	$Purity = \sum_{r=1}^k \frac{n_r}{n} P(S_r)$

S_r is a cluster, n_r is the size of the cluster, q is the number of classes, n_r^i is the number of concepts from the i th class that were assigned to the r th cluster, and n is the number of concepts and k is the number of clusters.

Validación con Expertos

- Se pueden evaluar los clusters para ver si producen el resultado esperado y comparar con otras soluciones
- Se puede generar una clasificación de validación

Comentarios finales

- La etapa de validación es la parte más difícil y frustrante del análisis de clusters
- Sin embargo es necesario
- Idealmente se deben combinar medidas externas e internas