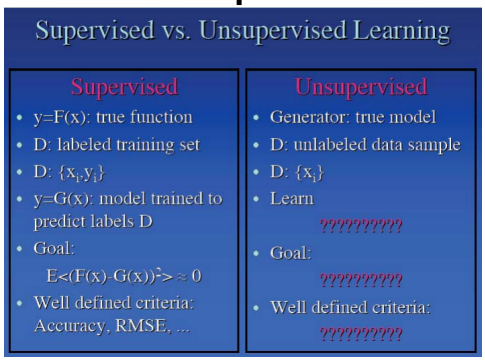
Validación de Clusters

Bárbara Poblete

¿Es necesario validar los clusters?

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

aprendizaje supervisado vs. no supervisado



source: http://www.cs.put.poznan.pl/jstefanowski/sed/DM-7clusteringnew.pdf

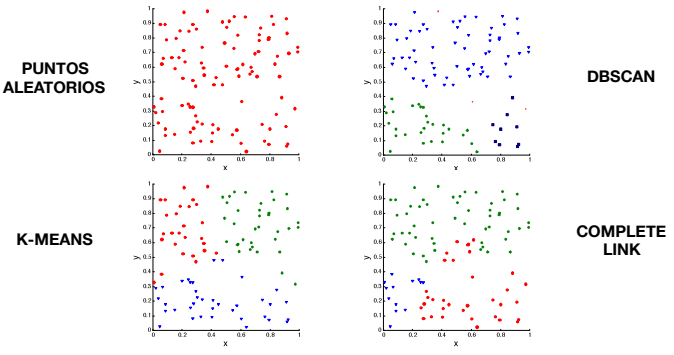
¿Cómo saber si nuestros clusters son buenos?

- No hay una repuesta absoluta
- Depende de la aplicación
- ¿Entonces, para qué evaluar?
  - ★ Clustering es muchas veces parte de un proceso exploratorio
  - ★ Evaluar parece innecesario en estos casos (pero no!)
  - ★ Cada algoritmo parece necesitar su propio tipo de evaluación
  - ★ k-means: SSE, pero no funciona para DBSCAN
  - ★ Es esencial! Porque siempre podemos encontrar clusters (hasta en datos aleatorios).

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

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

Medidas de validez

- **External Index** (Supervisado) Utiliza comparación con datos externos (pureza, entropía), para ver si el clustering se ajusta a una estructura externa.
- **Internal Index** (No-Supervisado) Sin datos externos (SSE), cohesión, separación.
- **Relative Index** (Relativo) Comprara resultados de clustering o clusters, puede usar medidas anteriores.

Concepto: Matriz de similitud (conversamente: matriz de proximidad)

Table 1. Similarity matrix of the woody genera in the 21 dry to arid regions of the Neotropic. Jaccard index.

	chiquitmia	sonora	chamela	cosatdes	chilemed	perusouth	guanacaste	venezuela	chacoarg	cochab	patagonia	caatinga	raichi	perunorth	semarg	serbol	prepuna	monte	puna	guajira	chaco
chiquitmia	1	0,078	0,273	0,06	0,05	0,103	0,193	0,215	0,12	0,111	0,005	0,251	0,352	0,218	0,191	0,152	0,063	0,077	0,016	0,131	0,235
sonora	0,078	1	0,148	0,19	0,123	0,183	0,042	0,058	0,175	0,176	0,093	0,102	0,092	0,111	0,183	0,166	0,182	0,192	0,136	0,214	0,181
chamela	0,273	0,148	1	0,066	0,034	0,142	0,223	0,23	0,092	0,13	0	0,207	0,267	0,221	0,185	0,176	0,08	0,072	0,022	0,239	0,188
cosatdes	0,06	0,19	0,066	1	0,33	0,237	0,033	0,037	0,155	0,295	0,159	0,075	0,077	0,086	0,209	0,207	0,296	0,296	0,188	0,132	0,128
chilemed	0,05	0,123	0,034	0,33	1	0,189	0,029	0,033	0,115	0,227	0,231	0,061	0,058	0,099	0,168	0,186	0,276	0,252	0,217	0,085	0,108
perusouth	0,103	0,183	0,142	0,237	0,189	1	0,052	0,077	0,167	0,358	0,086	0,139	0,131	0,3	0,317	0,42	0,342	0,236	0,135	0,184	0,193
guanacaste	0,193	0,042	0,223	0,033	0,029	0,052	1	0,322	0,074	0,033	0	0,195	0,153	0,142	0,095	0,067	0,027	0,038	0	0,147	0,134
venezuela	0,215	0,058	0,23	0,037	0,033	0,077	0,322	1	0,099	0,056	0	0,269	0,19	0,189	0,117	0,084	0,038	0,041	0,008	0,157	0,15
chacoarg	0,12	0,175	0,092	0,155	0,115	0,167	0,074	0,099	1	0,165	0,065	0,181	0,111	0,152	0,315	0,21	0,172	0,34	0,054	0,197	0,56
cochab	0,111	0,176	0,13	0,295	0,227	0,358	0,033	0,056	0,165	1	0,107	0,116	0,141	0,17	0,399	0,465	0,457	0,289	0,181	0,161	0,195
patagonia	0,005	0,093	0	0,159	0,231	0,086	0	0	0,065	0,107	1	0	0,014	0,019	0,058	0,077	0,155	0,191	0,296	0,027	0,043
caatinga	0,251	0,102	0,207	0,075	0,061	0,139	0,195	0,269	0,181	0,116	0	1	0,175	0,213	0,216	0,18	0,088	0,102	0,02	0,22	0,231
raichi	0,352	0,092	0,267	0,077	0,058	0,131	0,153	0,19	0,111	0,141	0,014	0,175	1	0,262	0,228	0,184	0,087	0,102	0,024	0,156	0,184
perunorth	0,218	0,111	0,221	0,086	0,099	0,3	0,142	0,189	0,152	0,17	0,019	0,213	0,262	1	0,262	0,269	0,143	0,121	0,033	0,198	0,224
semarg	0,191	0,183	0,185	0,209	0,168	0,317	0,095	0,117	0,315	0,399	0,058	0,216	0,228	0,262	1	0,586	0,282	0,289	0,093	0,212	0,372
serbol	0,152	0,166	0,176	0,207	0,186	0,42	0,067	0,084	0,21	0,465	0,077	0,18	0,184	0,269	0,586	1	0,387	0,249	0,137	0,182	0,304
prepuna	0,063	0,182	0,08	0,296	0,276	0,342	0,027	0,038	0,172	0,457	0,155	0,088	0,087	0,143	0,282	0,387	1	0,411	0,28	0,15	0,186
monte	0,077	0,192	0,072	0,296	0,252	0,236	0,038	0,041	0,34	0,289	0,191	0,102	0,102	0,121	0,289	0,249	0,411	1	0,191	0,147	0,281
puna	0,016	0,136	0,022	0,188	0,217	0,135	0	0,008	0,054	0,181	0,296	0,02	0,024	0,033	0,093	0,137	0,28	0,191	1	0,045	0,05
guajira	0,131	0,214	0,239	0,132	0,085	0,184	0,147	0,157	0,197	0,161	0,027	0,22	0,156	0,198	0,212	0,182	0,15	0,147	0,045	1	0,253
chaco	0,235	0,181	0,188	0,128	0,108	0,193	0,134	0,15	0,56	0,195	0,043	0,231	0,184	0,224	0,372	0,304	0,186	0,281	0,05	0,253	1

Distancias comunes

- Es común utilizar distancias métricas, como la distancia de Minkowski
- r=2, distancia Euclideana

$$p_{ij} = \left(\sum_{k=1}^{d} \left| x_{ik} - x_{jk} \right|^{r} \right)^{1/r}$$

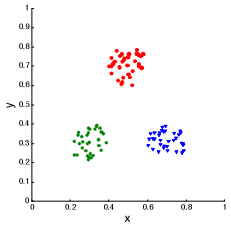
- otras: Manhattan, y distancia no metrica, Jaccard, por ej.
- similitud coseno

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}},$$

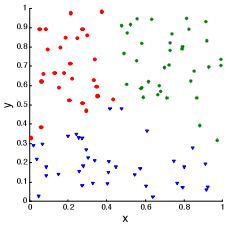
Validez usando correlación

- Comparamos 2 matrices
  - Matriz de proximidad (n x n usando la distancia entre puntos)
  - Matriz de incidencia (n x n matriz idealizada usando pertenencia a clusters)
- Calculamos la correlación entre ambas (simétricas, sólo compara n(n-1)/2 veces)

Correlación usando k-means



Corr = -0.9235

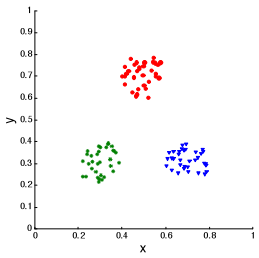


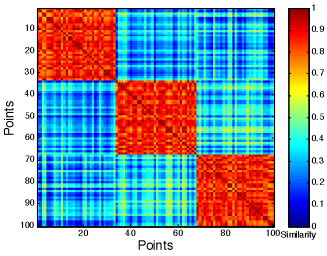
Corr = -0.5810

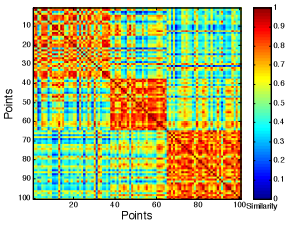
Enfoque visual

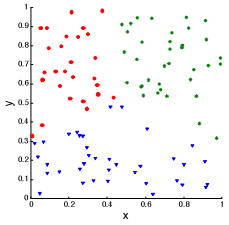
- Ordenar la matriz de similitud con respecto a etiquetas de clusters e inspeccionar visualmente

Visualizando la matriz de similitud (clusters reales)



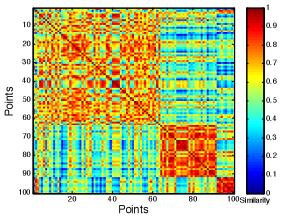


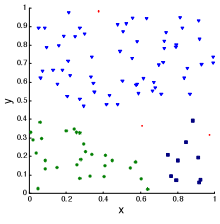




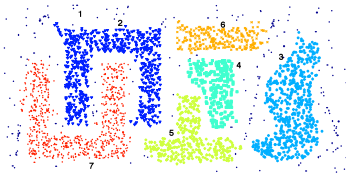
K-means

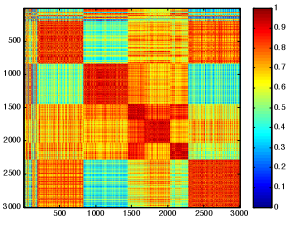
Visualizando clusters sobre datos aleatorios





DBSCAN

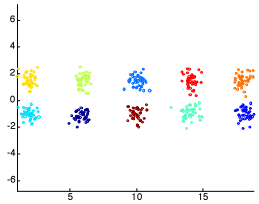


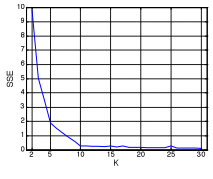


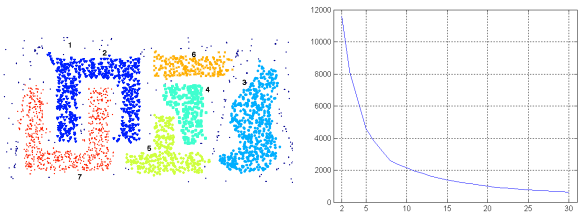
DBSCAN

Medidas internas: SSE

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







SSE of clusters found using K-means

Curva SSE para un dataset más complicado

Metodología para validar clusters

- Necesidad de contar con una metodología para interpretar cualquier medida (¿qué es bueno? ¿qué no?)
- Usamos la estadística para crear una metodología

Metodología para validar clusters

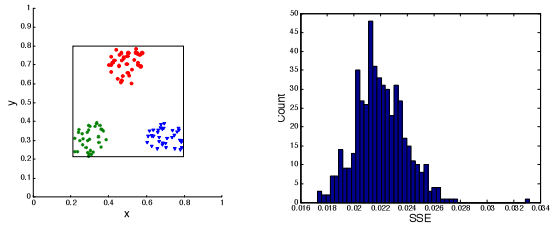
- Mientras más **atípico** es un resultado, más probable que sea reflejo de estructuras válidas
- Podemos comparar índices que resultan de datos aleatorios, con los de nuestros datos
- valores poco probables indican resultados válidos

Metodología para validar clusters

- Al comparar resultados de dos clustering (dos cluster sets), no es muy necesario usar una metodología
- Pero en este caso la pregunta es si la diferencia es **significativa** (estadísticamente - repetible y en magnitud)

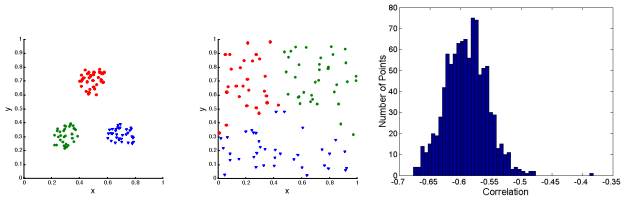
Metodología: Ejemplo SSE

- 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



Otro ejemplo: Correlación

- Correlación entre matrices de incidencia y proximidad para 2 sets de datos



Corr = -0.9235 Corr = -0.5810

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

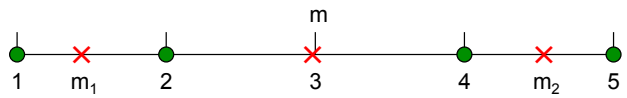
EJ. (SSE) Cohesión y Separación

- Cohesión se mide como **within cluster sum of squares (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:

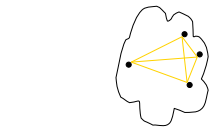
$$WSS = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$
$$BSS = 4 \times (3-3)^2 = 0$$
$$Total = 10 + 0 = 10$$

K=2 clusters:

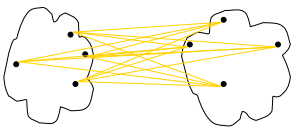
$$WSS = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$
$$BSS = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$
$$Total = 1 + 9 = 10$$

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 internas: Coeficiente de Silhouette

- Combina ideas de cohesion y separación, pero para puntos individuales, como también para clusters y clusterings (estos últimos son promedios)
- Para un punto individual, i
  - Calcular **a** = distancia promedio de *i* a los puntos de su cluster
  - Calcular **b** = min(distancia promedio de *i* a puntos de otro cluster)
  - s = (b-a)/max(a,b)
  - valores entre -1 y 1, mientras más cerca a 1 mejor (y a más cerca de 0

Pureza y Entropía

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

Medidas externas: Pureza y entropía

Table 5.9. K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

entropy For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij}, the 'probability' that a member of cluster j belongs to class i as follows: $p_{ij} = m_{ij}/m_j$, where m_j is the number of values in cluster j and m_{ij} is the number of values of class i in cluster j. Then using this class distribution, the entropy of each cluster j is calculated using the standard formula $e_j = \sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$, where the L is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e., $e = \sum_{i=1}^{K} \frac{m_i}{m} e_j$, where m_j is the size of cluster j, K is the number of clusters, and m is the total number of data points.

purity Using the terminology derived for entropy, the purity of cluster j, is given by $purity_j = \max p_{ij}$ and the overall purity of a clustering by $purity = \sum_{i=1}^{K} \frac{m_i}{m} purity_j$.

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

Comentario final

- 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