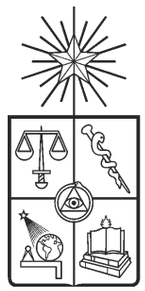


dcc
CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

Validación de Clusters

Clasificación y clustering

	Clasificación	**Clustering**
Paradigma	Aprendizaje supervisado	Aprendizaje no supervisado
Función “real”	$y = F(x)$	Modelo generador de datos
Dataset	$D_n = \{(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)\}$	$D_m = \{x_1, x_2, \ldots, x_m\}$
Se aprende:	$y = G(x)$, función candidata	…
Objetivo	$Error(F(x) - G(x)) \sim 0$	…
Evaluación	Accuracy, Precision, Recall, F1, etc	…

Evaluación de Clusters

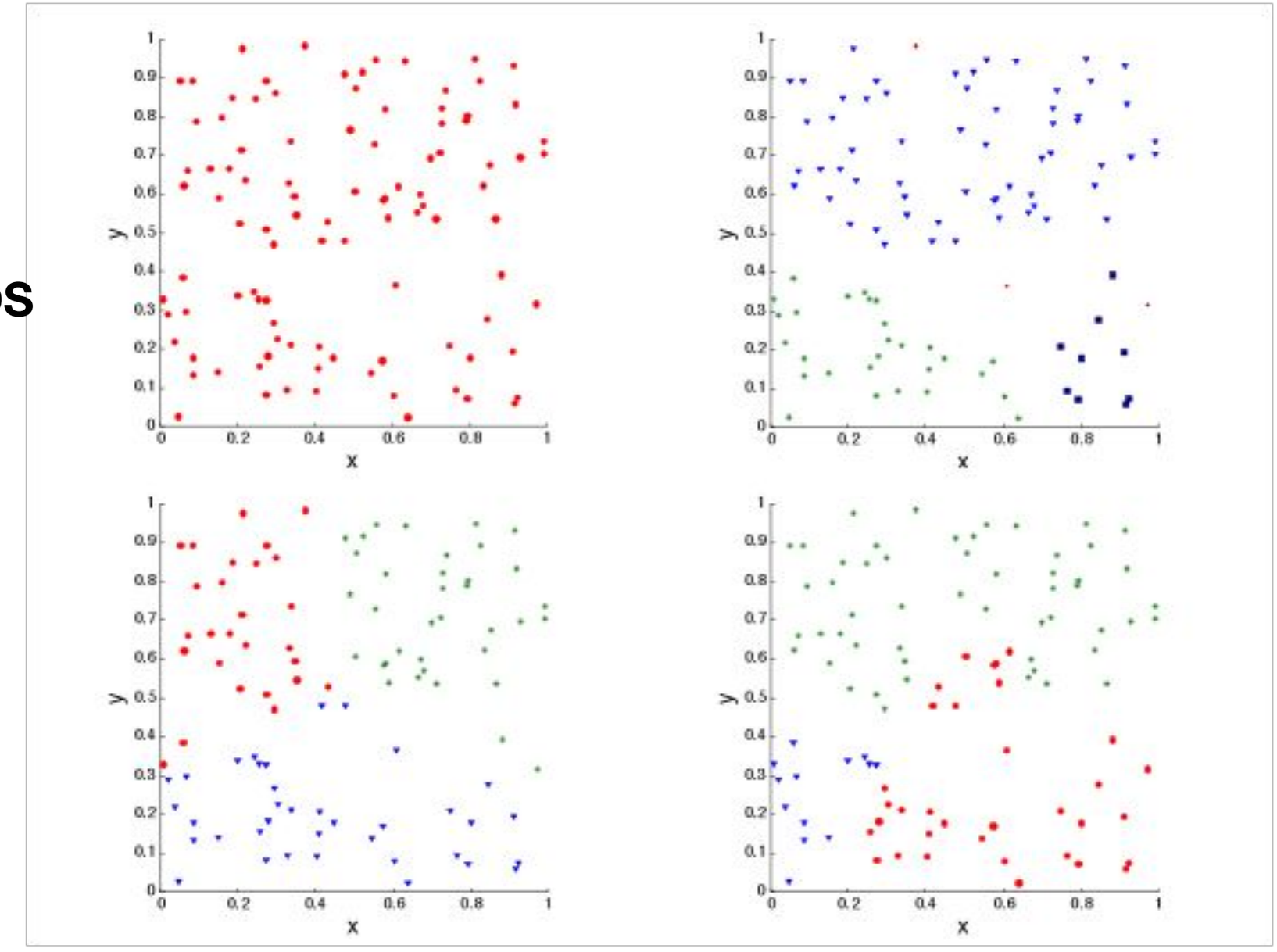
- En aprendizaje supervisado, la evaluación se hace con métricas estandarizadas (F1, Accuracy, AUC, etc) y es una parte central de la metodología para encontrar el mejor modelo.

- Para clustering la evaluación no está tan desarrollada ni comúnmente utilizada.

- A pesar de esto, la evaluación (o validación) de clusters es importante.

Clusters en Datos Aleatorios

PUNTOS ALEATORIOS



DBSCAN

K-MEANS

COMPLETE LINK

Los algoritmos de clustering puedan encuentran clusters incluso en datos aleatorios.

Motivación Central

- En la evaluación de clustering, es crucial distinguir si los clusters encontrados son más que simples agrupaciones aleatorias.

- Dado que no hay una solución esperada (ej: la clase que queremos predecir), no es claro cómo comparar la eficacia de diferentes algoritmos de clustering.

- Además de las medidas cuantitativas que veremos en esta clase, es esencial analizar los clusters de manera cualitativa.
  - Por ejemplo, examinar muestras aleatorias de cada cluster para comprender el concepto que representan.

Aspectos Importantes de la validación de clusters

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 (sobre todo para k-means).

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

Las métricas que se usan para evaluar varios aspectos sobre un clustering se dividen en 3 categorias:

1. **No Supervisadas** (o índices internos): miden la calidad sin usar información externa: SSE, cohesión, separación.

2. **Supervisadas** (o índices externos)**:** Utiliza comparación con datos externos por ejemplo clases para calcular métricas como pureza y entropía y determinar si el clustering se ajusta a esa estructura externa.

3. **Relativas:** Compara resultados de clustering o clusters, puede usar medidas anteriores. Por ejemplo comparar el SSE de dos clusterings obtenidos con K-means.

Medidas internas

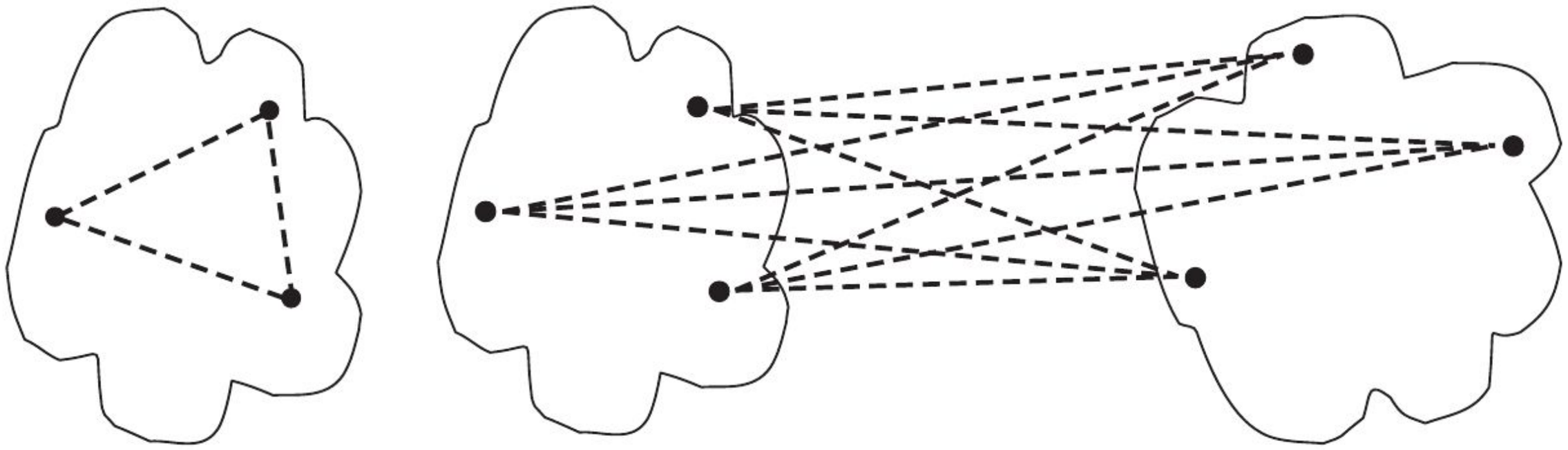
- Suelen basarse en las ideas de **cohesión** (qué tan cercanos son los puntos de un mismo cluster) y **separación** (qué tan disímiles son puntos de clusters distintos)

- Generalmente, podemos expresar la validez de un conjunto de K clusters como la validez ponderada de cada cluster:

$$overall\ validity = \sum_{i=1}^{K} w_i\ validity(C_i).$$

- Medidas para clusters basados en grafos y basados en prototipos

Medidas internas, basadas en grafos



(a) Cohesion.

(b) Separation.

Nos importan todos los nodos y arcos del grafo

Medidas internas, basadas en grafos

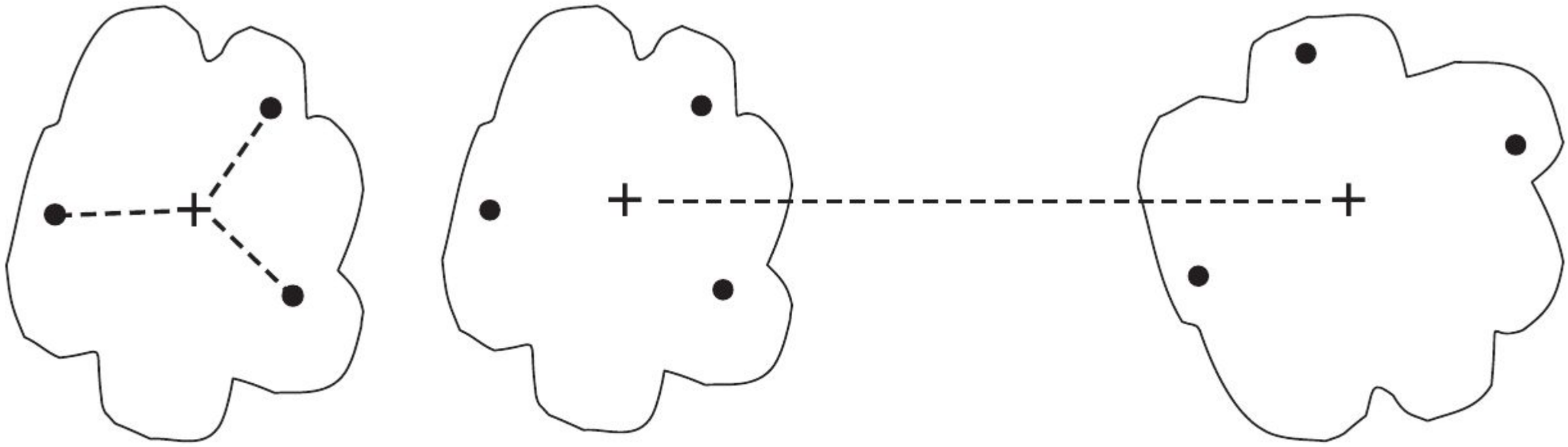
$$cohesion(C_i) = \sum_{\substack{\mathbf{x} \in C_i \\ \mathbf{y} \in C_i}} proximity(\mathbf{x}, \mathbf{y})$$

$$separation(C_i, C_j) = \sum_{\substack{\mathbf{x} \in C_i \\ \mathbf{y} \in C_j}} proximity(\mathbf{x}, \mathbf{y})$$

¿Qué es *proximity*?

- Puede ser cualquier función de distancia o similitud
  - Si usamos distancia, buscaremos un valor bajo en cohesión y un valor alto en separación
- Debe ser escogida con cuidado

Medidas internas, basadas en prototipos



(a) Cohesion.

(b) Separation.

Nos importa sólo el prototipo (generalmente, el centroide del cluster)

Medidas internas, basadas en prototipos

$$cohesion(C_i) = \sum_{\mathbf{x} \in C_i} proximity(\mathbf{x}, \mathbf{c}_i)$$

$$separation(C_i, C_j) = proximity(\mathbf{c}_i, \mathbf{c}_j)$$

$$separation(C_i) = proximity(\mathbf{c}_i, \mathbf{c})$$

- $\boldsymbol{c_i}$: centroide de cluster *i*
- $\boldsymbol{c}$: centroide de todos los datos

Medidas internas, basadas en prototipos

Si usamos distancia euclidiana (al cuadrado):

- Cohesión: es equivalente a SSE de cada cluster.

$$\text{cohesion}(C_i) = \text{SSE}\ C_i = \sum_{x \in C_i} (x - c_i)^2$$

- Separación: *between group sum of squares* (SSB). Suma de la distancia cuadrada del centroide de un cluster al centroide global de todos los datos (m_i = tamaño del cluster)

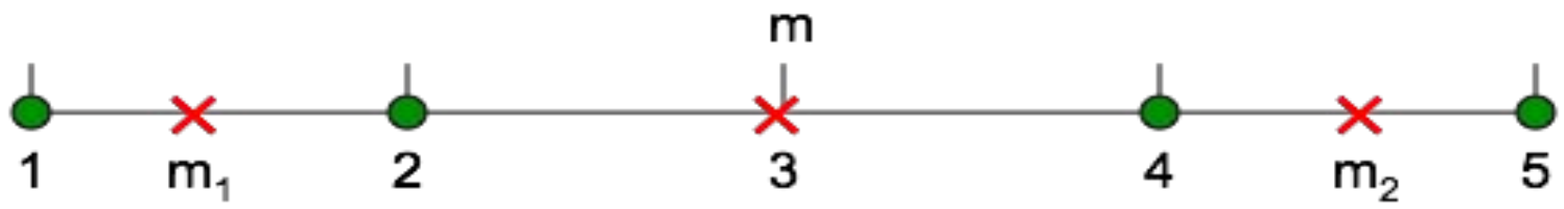
$$\text{separation}(\{C_1, C2, \ldots, C_K\}) = \text{Total BBS} = \sum_{i=1}^{K} m_i (c_i - c)^2$$

Medidas internas

Se puede demostrar que, si se usa distancia euclidiana cuadrada como medida de proximidad:

- La cohesión basada en prototipos (SSE) es directamente proporcional a la cohesión basada en grafos (distancia promedio de a pares)
- Las dos medidas de separación basadas en prototipos (separación de todos los clusters respecto al cluster central, y separación de a pares de clusters) también son equivalentes
- Y además…

Relación entre Cohesión y Separación



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:

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

- Notar que BSS + WSS es constante y equivale a la suma de cuadrados totales (TSS), que es la distancia cuadrada de cada punto con el centro total.

$$TSS = \sum_i (x_i - m)^2$$

- Entonces minimizar cohesión equivale a maximizar separación.

Distancias comunes usadas en Clustering

- Es común utilizar distancias métricas, como la distancia de Minkowski que generaliza la distancia euclidiana

- r=2, distancia Euclidiana

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

- otras: Manhattan r=1, y distancia no métrica, Jaccard, por ej.

- similitud coseno

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

Evaluando Clusters y Objetivos Individuales

- Muchas medidas de validez de cluster pueden aplicarse tanto a clusters individuales como a objetos.

- Por ejemplo podemos medir la cohesión de un cluster y si esta es baja podríamos romper el cluster en dos.

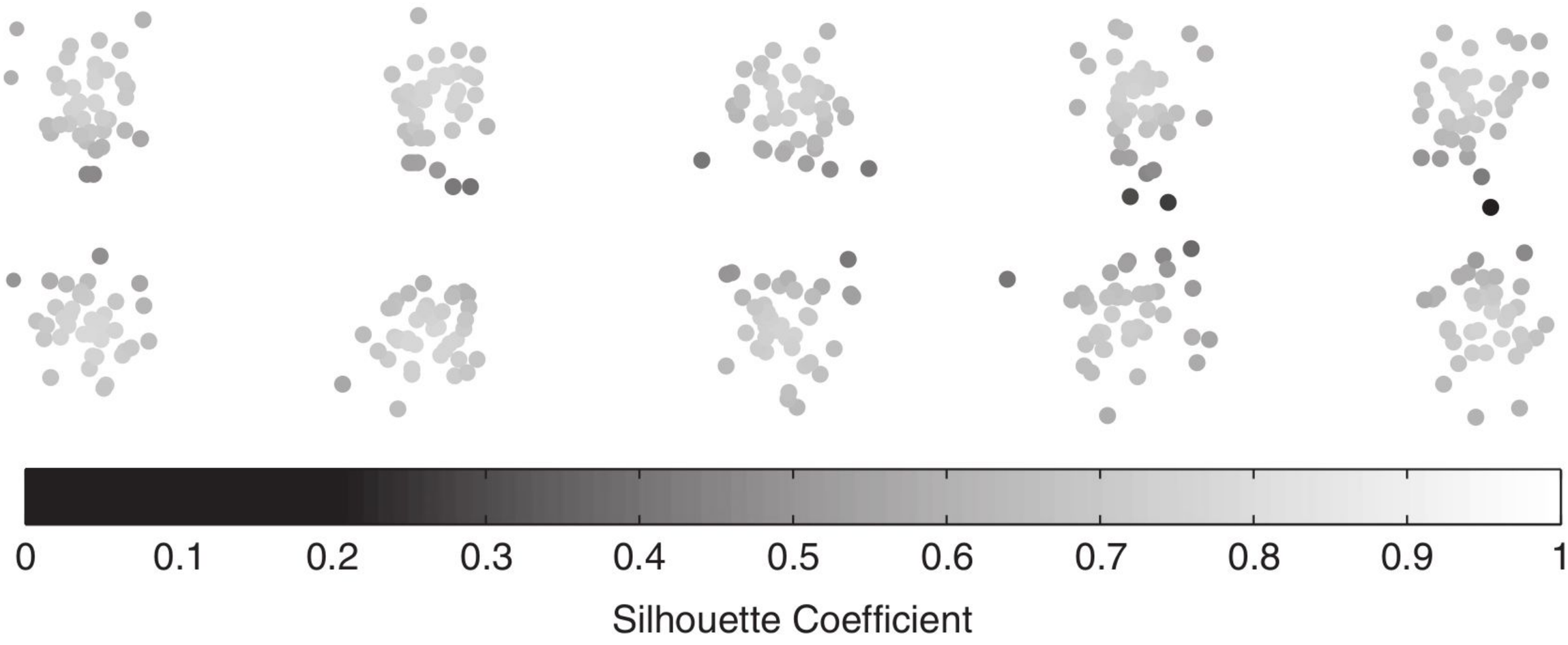
- También podríamos medir cuánto contribuye un objeto particular a la cohesión del cluster al que pertenece o a la separación con los otros clusters.

- Una buena medida no supervisada para evaluar objetos, clusters individuales o un conjunto de clusters, y que combina cohesión y separación, es el coeficiente de **Silhouette**.
  - Se define como un coeficiente para puntos
  - Podemos calcular el coeficiente de Silhouette de un cluster como el coeficiente promedio de sus puntos.
  - De igual modo, podemos promediar el coeficiente para todos los puntos de todos los clusters y usarlo como métrica global.

Medidas internas: Coeficiente de Silhouette

- Para un punto individual, *i*
  - Calcular $\mathbf{a_i}$= distancia promedio de *i* a los puntos de su cluster
  - Calcular $\mathbf{b_i}$ = min(distancia promedio de *i* a puntos de otro cluster)
  - $s_i = (b_i-a_i)/max(a_i,b_i)$
  - valores entre -1 y 1, mientras más cerca a 1 mejor
  - También se espera que a_i esté lo más cercano a 0 posible

Medidas internas: Coeficiente de Silhouette



El coeficiente de Silhouette cuantifica la cohesión y separación de un punto al calcular el cociente entre la distancia promedio a los puntos del mismo cluster ($\mathbf{a}_i$) y la distancia promedio al cluster más cercano ($\mathbf{b}_i$)

Comentarios generales sobre medidas de evaluación de clusters no supervisadas

- La mayoría de las medidas no supervisadas (SSE, Silhouette, Cohesión, Separación) son efectivas para clusters particionales pero no para densidad (DBSCAN).

- Con algoritmos de clustering particionales como K-means, una medida como el coeficiente de Silhouette tiende a mejorar hasta encontrar el conjunto "**natural**" de clusters, para luego empeorar si se dividen demasiado finamente.
  - La cohesión no mejora mucho, pero la separación empeora

- En general, con pocos clusters, es mejor analizar la cohesión y separación de forma independiente para cada cluster.

- Esto puede dar una visión más completa de cuán cohesivo es cada cluster y qué tan separados están un par de clusters dados.

Evaluación No Supervisada usando la Matriz de Similitud

Una matriz de similitud (o de proximidad) es una matriz de n x n con la similitud (o distancia) entre pares de puntos.

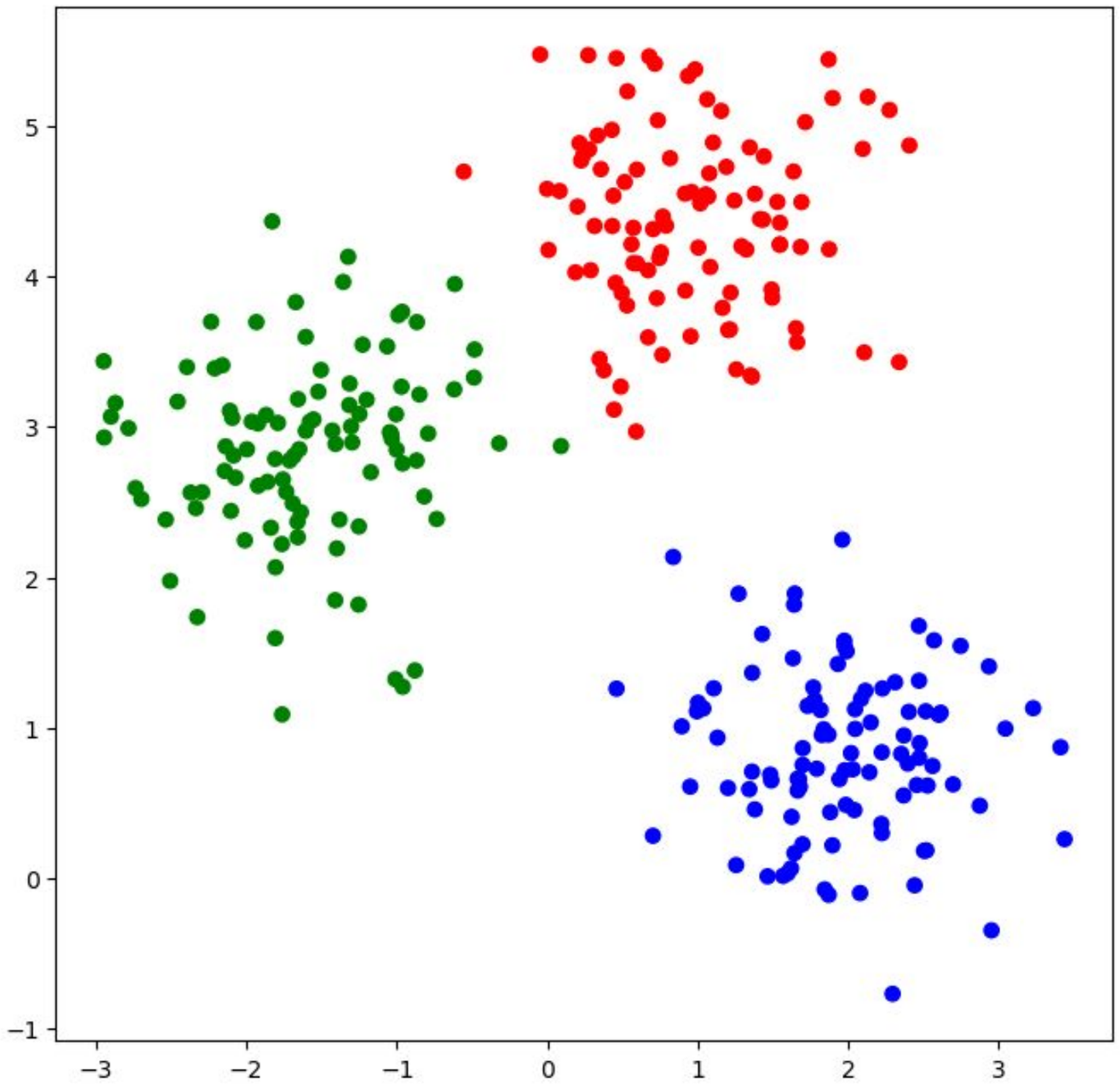
Table 1. Similarity matrix of the woody genera in the 21 dry to arid regions of the Neotropic. Jaccard index.

	chiquitania	sonora	chamela	coastdes	chilemed	perusouth	guanacaste	venezuela	chacoarg	cochab	patagonia	caatinga	tuichi	perunorth	serrarg	serrbol	prepuna	monte	puna	guajira	chaco
chiquitania	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	0,229	0,188
coastdes	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	0,298	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
tuichi	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
serrarg	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
serrbol	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	0,298	0,02	0,024	0,033	0,093	0,137	0,28	0,191	1	0,045	0,05
guajira	0,131	0,214	0,229	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

Validez usando correlación

- Comparamos 2 matrices
  - Matriz de incidencia (matriz de nxn idealizada, usando pertenencia a clusters)
    - una fila y una columna por cada punto
      - un valor es 1 si los dos puntos coinciden en el mismo cluster
      - un valor es 0 si los dos puntos no coinciden en el mismo cluster
  - Matriz de similitud (n x n usando similitud entre puntos)
- Ordenamos las matrices, de tal forma que todos los puntos que pertenecen a un mismo cluster estén juntos
- Calculamos la correlación entre ambas (simétricas, sólo compara n(n-1)/2 veces)

Validez usando correlación



- Datos artificiales, con 3 grupos claramente separados
- Corremos K-Means, y obtenemos esos clusters

Validez usando correlación

	Pt. 297	Pt. 1	Pt. 296	Pt. 295	Pt. 293	...	Pt. 22	Pt. 23	Pt. 265	Pt. 33	Pt. 2
Pt. 297	1	1	1	1	1	...	0	0	0	0	0
Pt. 1	1	1	1	1	1	...	0	0	0	0	0
Pt. 296	1	1	1	1	1	...	0	0	0	0	0
Pt. 295	1	1	1	1	1	...	0	0	0	0	0
Pt. 293	1	1	1	1	1	...	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...
Pt. 22	0	0	0	0	0	...	1	1	1	1	1
Pt. 23	0	0	0	0	0	...	1	1	1	1	1
Pt. 265	0	0	0	0	0	...	1	1	1	1	1
Pt. 33	0	0	0	0	0	...	1	1	1	1	1
Pt. 2	0	0	0	0	0	...	1	1	1	1	1

Construimos matriz de incidencia: 1 si los dos puntos pertenecen al mismo cluster, 0 en caso contrario

Validez usando correlación

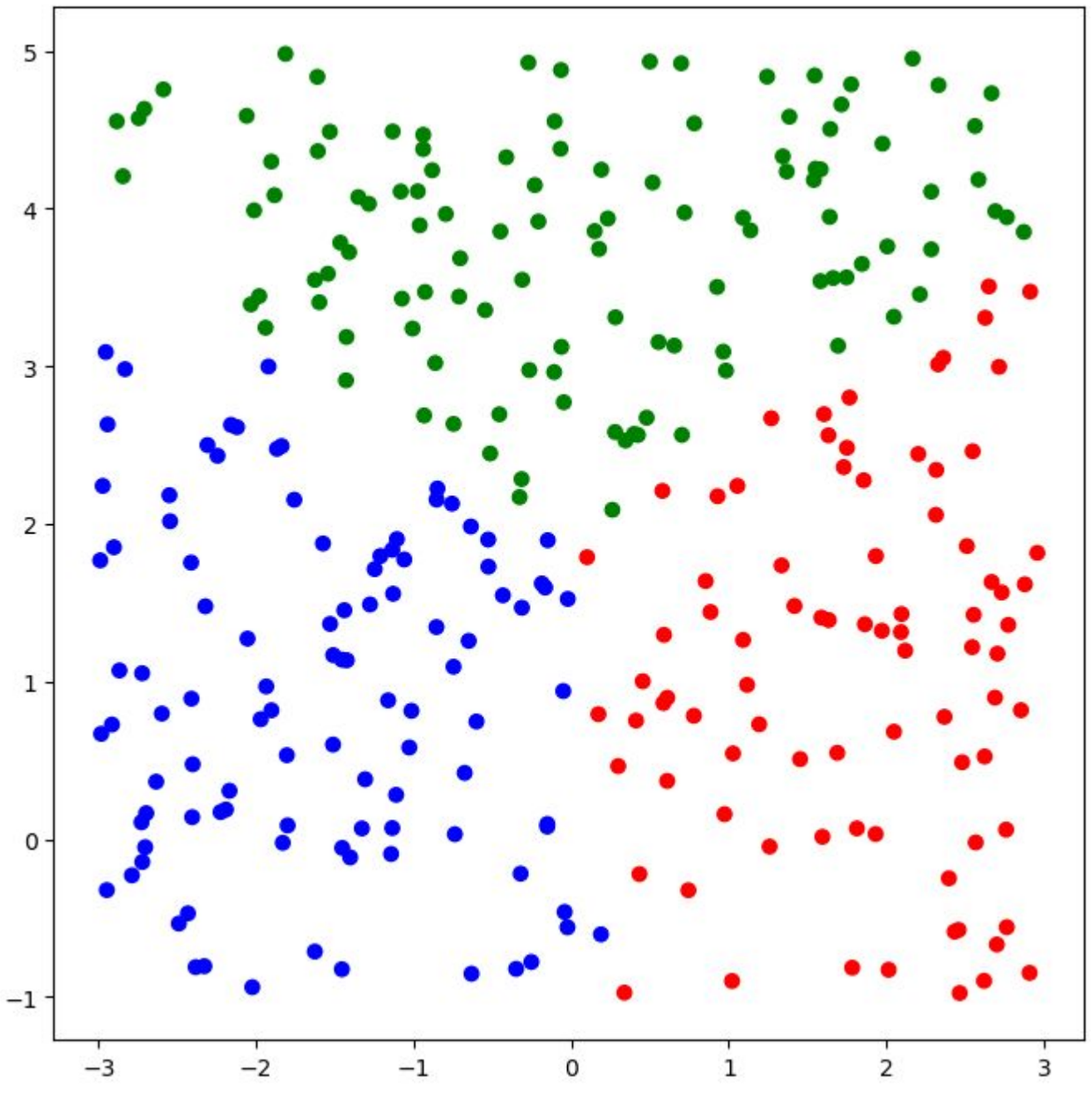
	Pt. 299	**Pt. 296**	**Pt. 292**	**Pt. 3**	**Pt. 20**	**...**	**Pt. 294**	**Pt. 2**	**Pt. 295**	**Pt. 298**	**Pt. 0**
Pt. 299	1	0.28	0.45	0.19	0.15	...	0.17	0.13	0.27	0.36	0.12
Pt. 296	0.28	1	0.6	0.38	0.32	...	0.11	0.09	0.37	0.25	0.16
Pt. 292	0.45	0.6	1	0.37	0.3	...	0.12	0.1	0.33	0.28	0.13
Pt. 3	0.19	0.38	0.37	1	0.81	...	0.05	0.04	0.14	0.11	0.06
Pt. 20	0.15	0.32	0.3	0.81	1	...	0.04	0.03	0.12	0.09	0.05
...	...	...	...	...	...	...	...	...	...	...	...
Pt. 294	0.17	0.11	0.12	0.05	0.04	...	1	0.8	0.28	0.42	0.31
Pt. 2	0.13	0.09	0.1	0.04	0.03	...	0.8	1	0.24	0.34	0.31
Pt. 295	0.27	0.37	0.33	0.14	0.12	...	0.28	0.24	1	0.53	0.41
Pt. 298	0.36	0.25	0.28	0.11	0.09	...	0.42	0.34	0.53	1	0.31
Pt. 0	0.12	0.16	0.13	0.06	0.05	...	0.31	0.31	0.41	0.31	1

Construimos la matriz con similitudes de a pares, usando el mismo ordenamiento previo. Se calculó distancia euclidiana y después se transformó a similitud usando decaimiento exponencial (exp(-x))

Validez usando correlación

- Calculamos la correlación, usando sólo los valores sobre (o bajo) la diagonal
  - n(n-1)/2 valores
- En nuestro caso, obtenemos correlación = 0.82
- Si hubiésemos usado distancia en vez de similitud, habríamos obtenido correlación negativa, pero acá lo que nos importa es la magnitud
- 0.82, ¿es un número alto o bajo?

Validez usando correlación



- Datos aleatorios, con distribución uniforme
- KMeans siempre va a generar los K clusters que le pidamos, aunque los datos sean aleatorios
- Correlación: 0.53

Enfoque visual

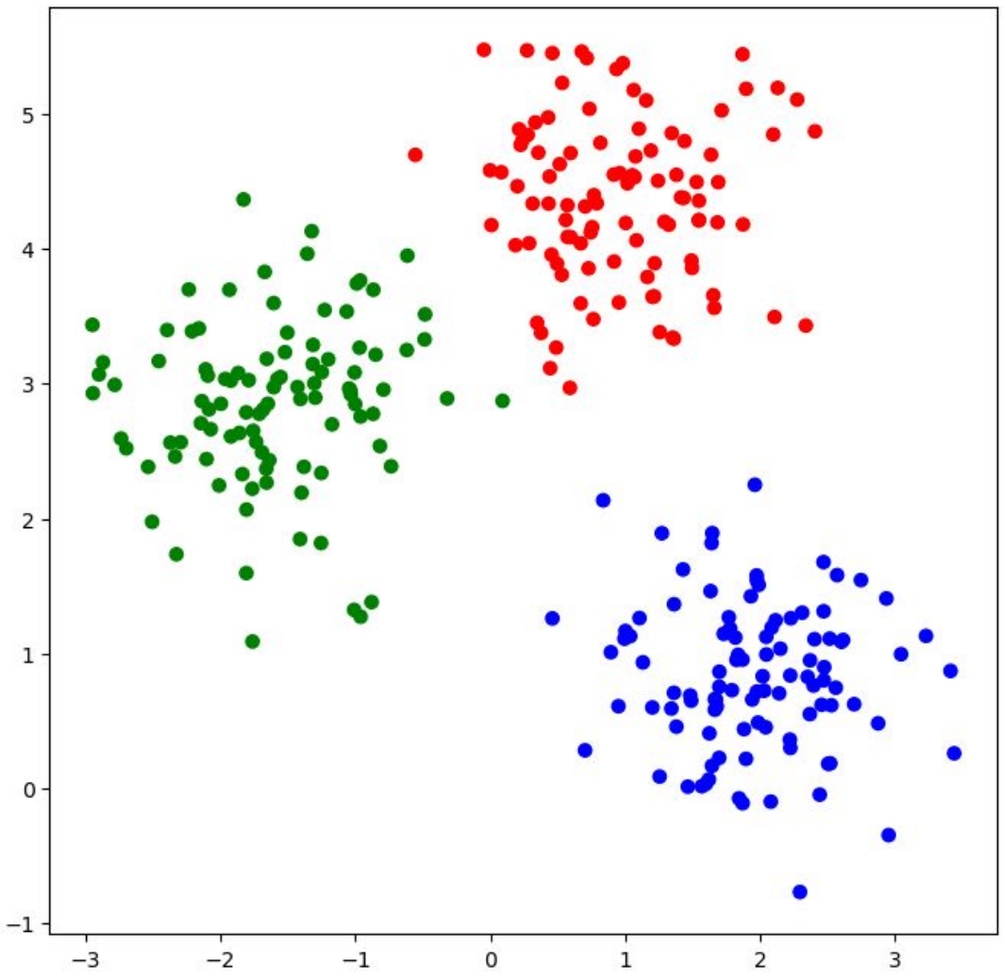
- Idea: ordenar la matriz de similitud con respecto a etiquetas de clusters e inspeccionar visualmente

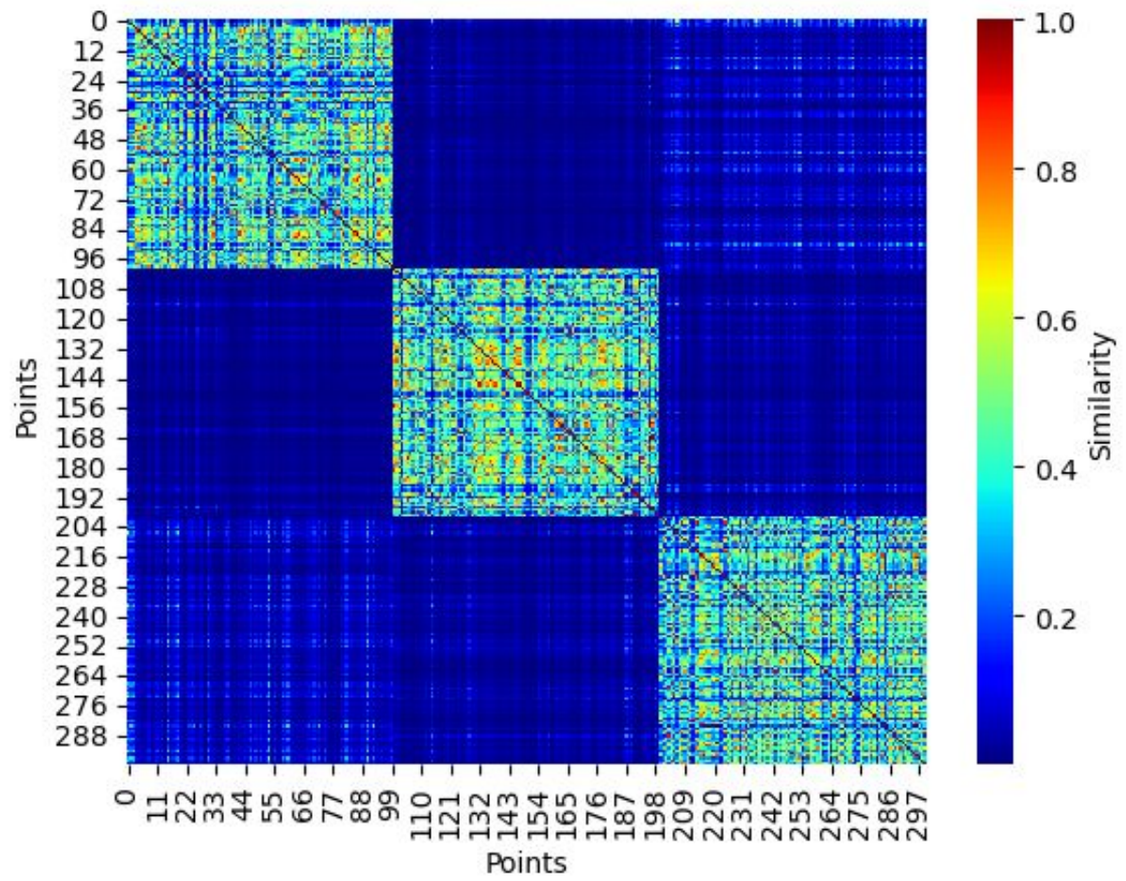
- Ordenamos las filas y columnas de la matriz de similitud de modo que todos los objetos que pertenecen al mismo cluster estén juntos.

- Hacer un mapeo de colores por similitud en la matriz.

- Entonces una matriz de similitud ideal tiene una estructura diagonal de bloques.
  - La similitud alta dentro de los bloques de la matriz de un mismo cluster y baja para objetos de clusters distintos.

Visualizando la matriz de similitud (clusters reales)

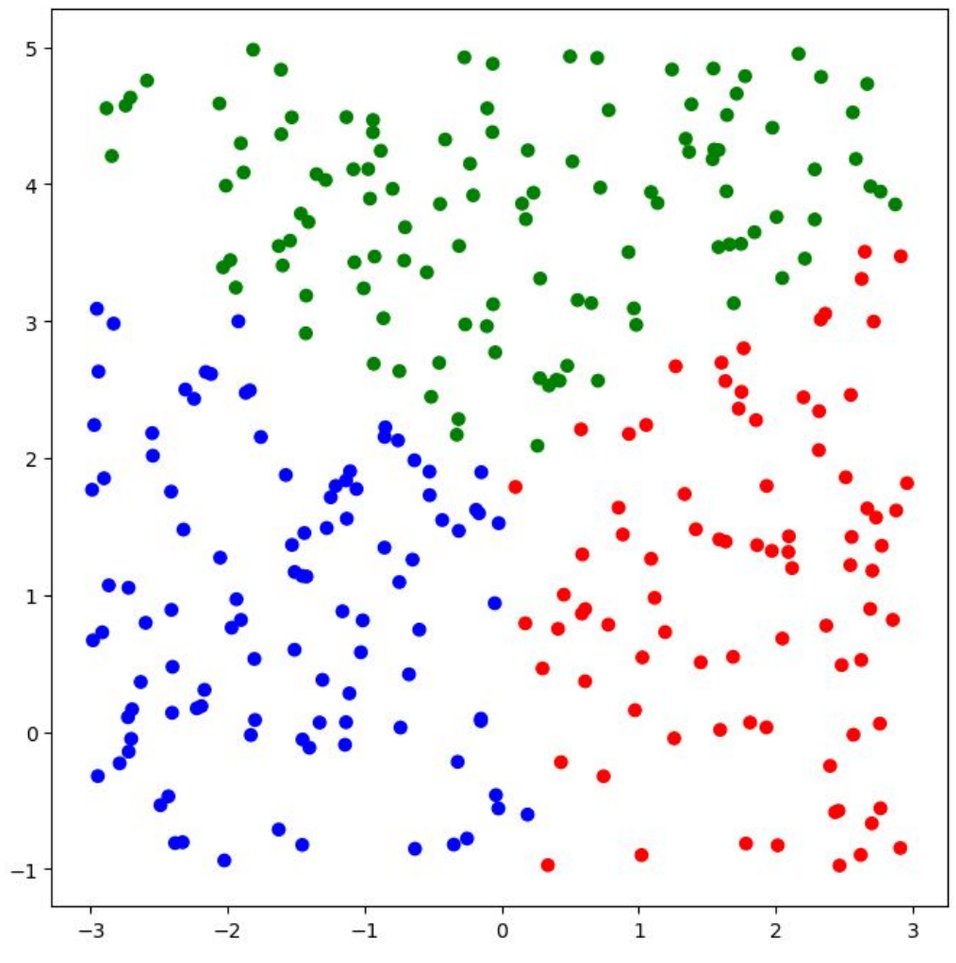


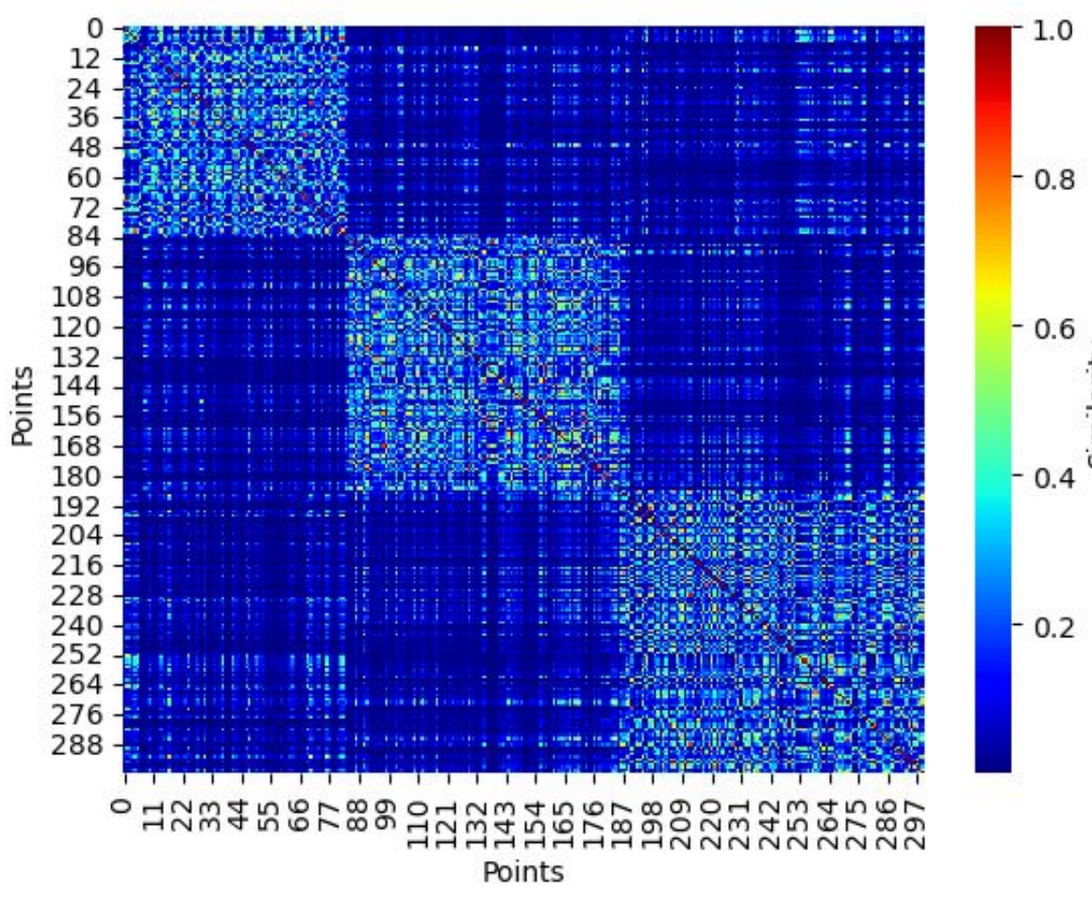


Para clusters bien separados, la estructura diagonal de bloques es clara.

Visualizando clusters sobre datos aleatorios

Para datos aleatorios las matrices no tienen una estructura diagonal de bloques (tan) clara.

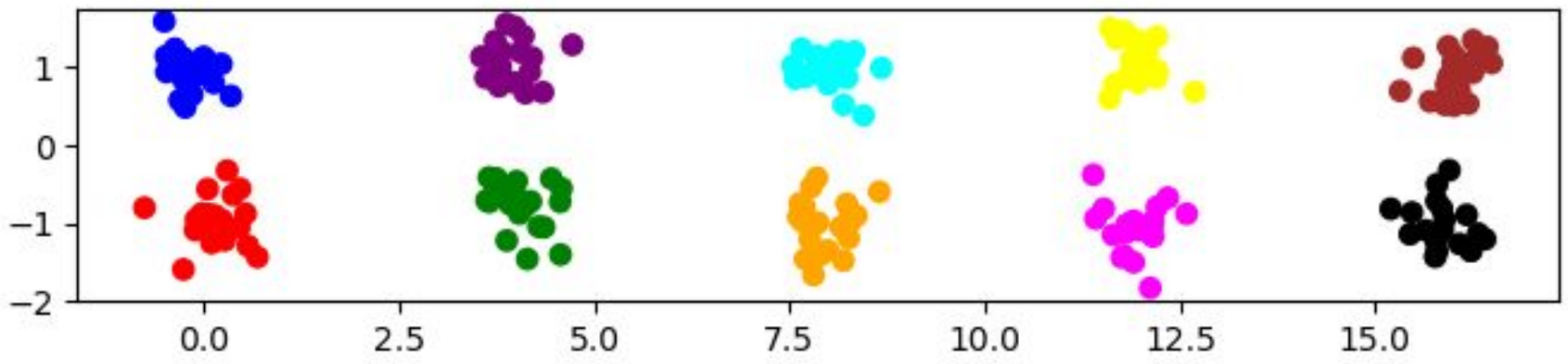


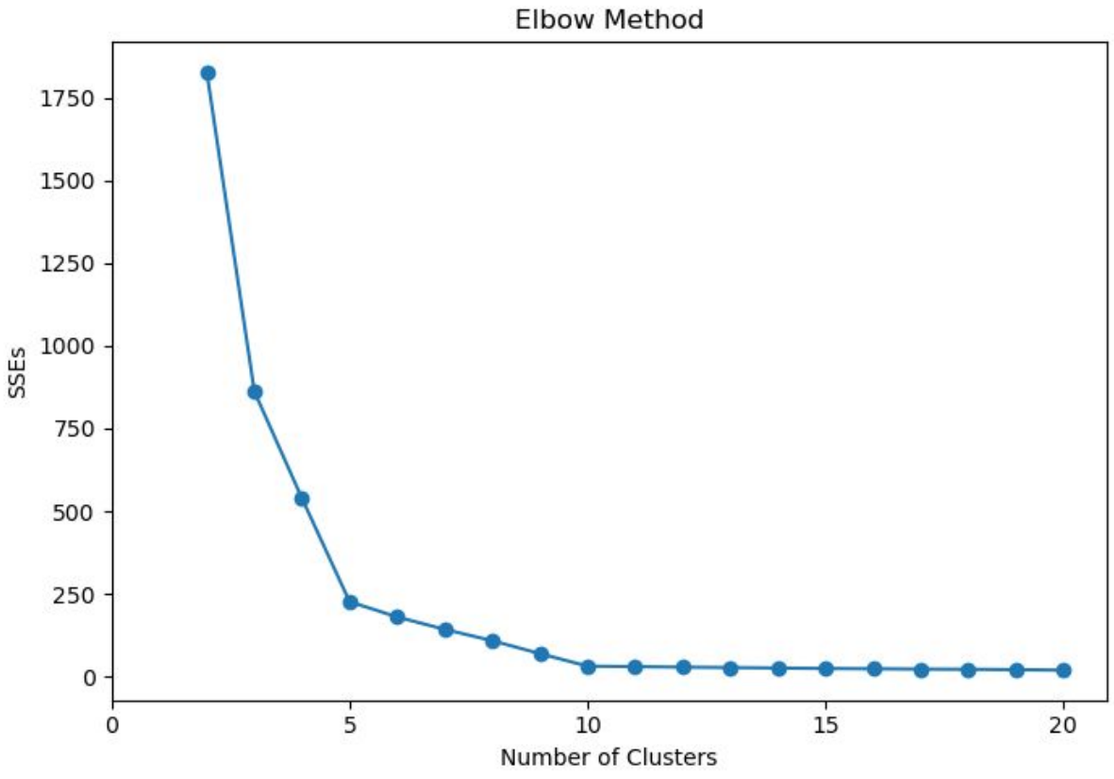


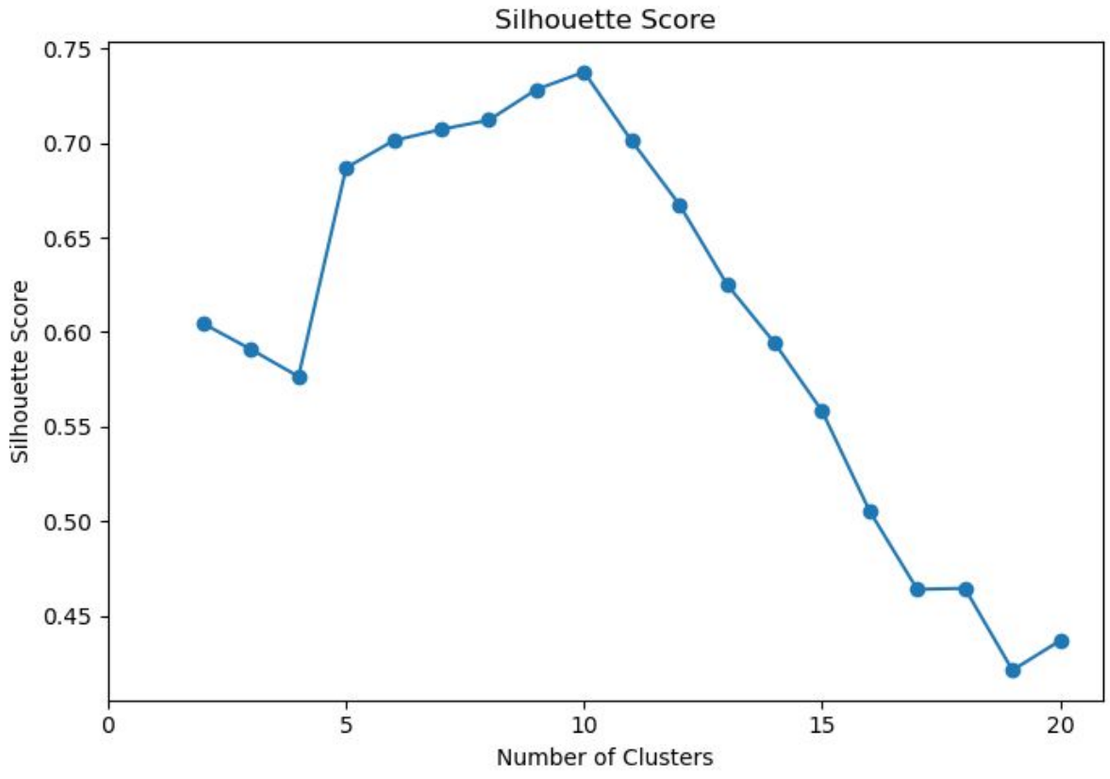
Determinando el número óptimo de clusters

- Varias medidas de evaluación de clusters no supervisados pueden utilizarse para determinar aproximadamente el número correcto o natural de clusters.
- Por ejemplo SSE y el coeficiente de Silhouette promedio para todos los puntos para K-means con distintos valores de K.

Determinando el número óptimo de clusters



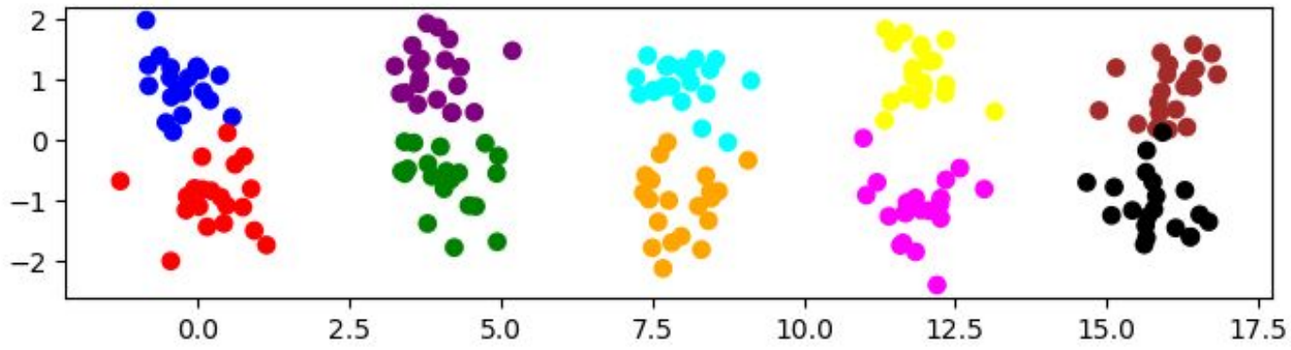


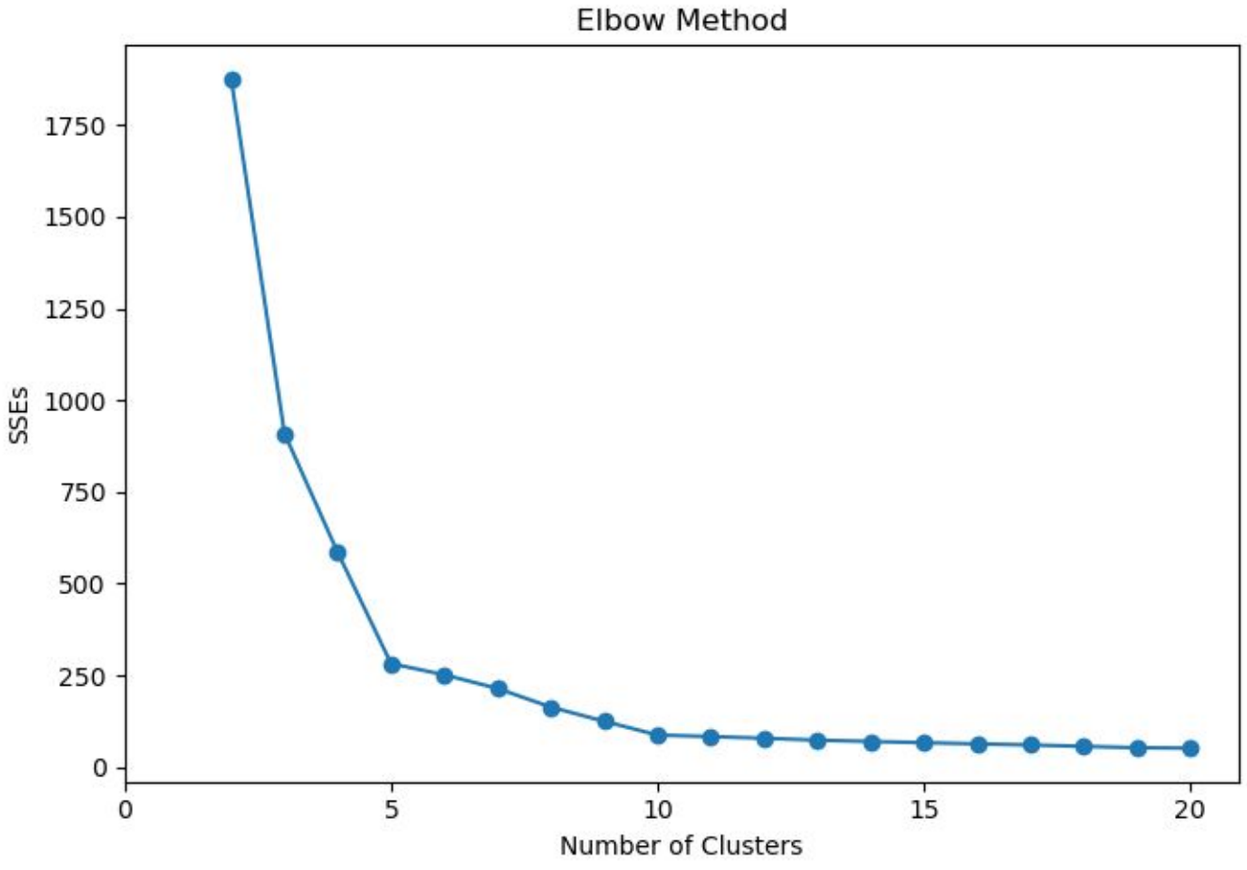


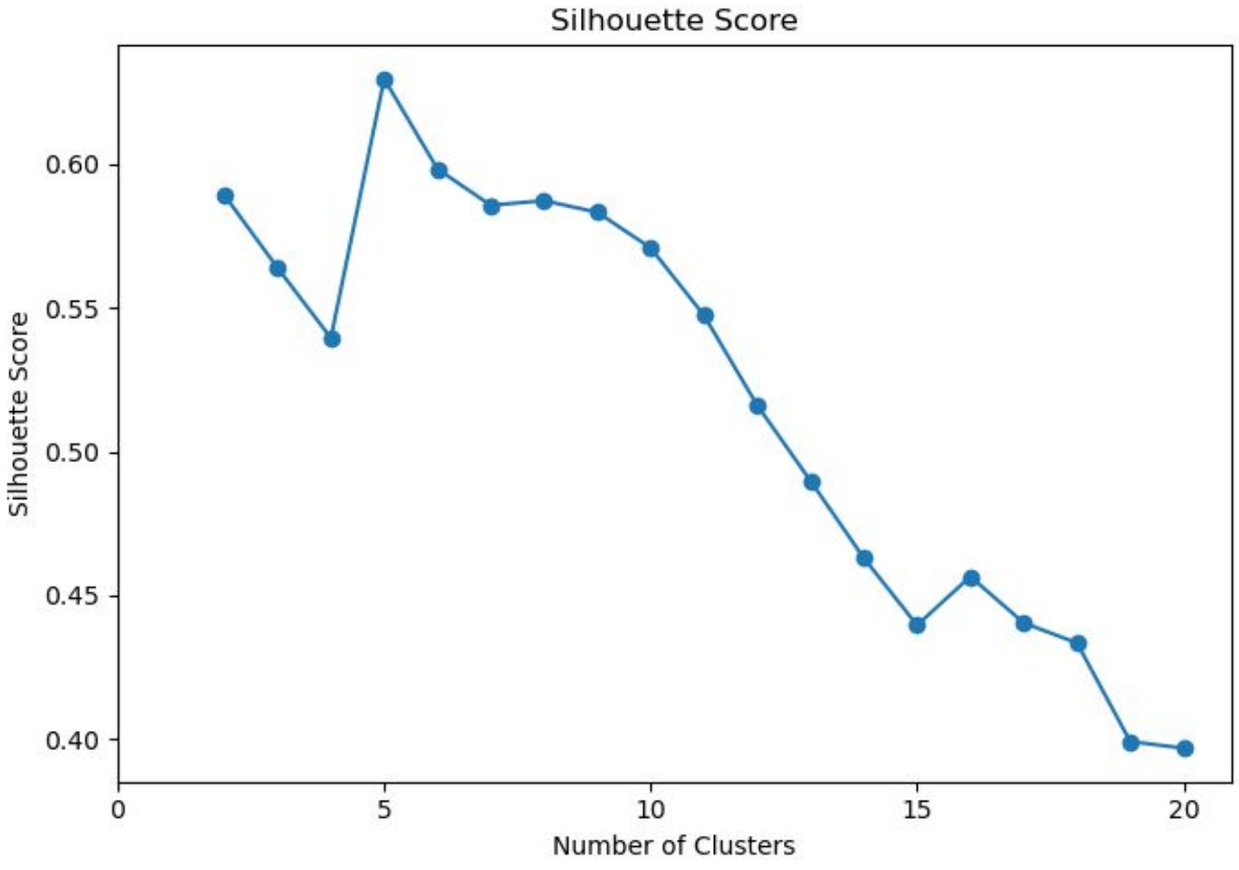
Determinando el número óptimo de clusters

- **Método del codo**: encontrar un punto de inflexión o codo en el SSE como el número óptimo de clusters (10 en el ejemplo…aunque podría ser 5 👀)
  - Importante: no buscamos el valor mínimo, buscamos un punto de inflexión
- **Método Silhouette**: encontrar el máximo del promedio del coeficiente de Silhouette para todos los puntos (10 en el ejemplo).
- Aunque útiles, estos criterios deben usarse con precaución; en cierta forma, siguen siendo un tanto subjetivos.
- Pueden no ser precisos si los clusters están muy entrelazados o solapados.
- Los datos también pueden contener clusters anidados, lo que dificulta la identificación del número de clusters.

Determinando el número óptimo de clusters (ejemplo 2)







Medidas Supervisadas o Externas

- La información externa sobre datos generalmente se presenta en forma de etiquetas de clase derivadas externamente para los ejemplos.

- Las medidas externas miden el grado de correspondencia entre las etiquetas de los clusters y las etiquetas de clase.

- ¿Por qué es esto de interés? Después de todo, si tenemos las etiquetas de clase, ¿cuál es el punto de realizar un análisis de clústeres?

Medidas Supervisadas o Externas

- Las motivaciones para dicho análisis incluyen la comparación de técnicas de clustering con la "verdad" provista por las etiquetas.

- También permite evaluar si el proceso de clasificación puede ser producido automáticamente mediante clustering.
  - Por ejemplo, clustering de noticias.

- A veces simplemente no tenemos la cantidad de etiquetas necesarias para entrenar un clasificador, pero sí algunas para validar

- A continuación presentamos dos medidas externas: pureza y entropía.

Entropía

Entropía: El grado en que cada cluster contiene objetos de una sola clase.

- Para cada cluster calculamos la probabilidad de que un elemento *i* del cluster pertenezca a la clase *j*: $p_{ij} = m_{ij} / m_i$

- m_i es el número de objetos en el cluster *i* y m_{ij} es el número de objetos de la clase *j* en el cluster *i*.

- La entropía de cada cluster i se calcula utilizando la fórmula estándar con L el número total de clases.

$$e_i = -\sum_{j=1}^{L} p_{ij} \log_2 p_{ij}$$

- La entropía total para un conjunto de clusters se calcula como la suma de las entropías de cada cluster ponderadas por el tamaño de cada cluster con K el número de clusters y m es el número total de puntos de datos.

$$e = \sum_{i=1}^{K} \frac{m_i}{m} e_i$$

Pureza

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

- Se calcula cómo la probabilidad máxima de una de las clases.

$$purity(i) = \max_j p_{ij}$$

- A nivel global, calculamos un promedio de purezas ponderado

$$\sum_{i=1}^{K} \frac{m_i}{m} purity(i)$$

Table 5.9. K-means clustering results for the *LA Times* 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

La entropía es óptima cuando es baja, mientras que la pureza lo es cuando es alta.

Evaluación de la significancia de las medidas de validez

- Las medidas de validez evalúan la calidad de los clústeres obtenidos, mediante un único indicador numérico.

- Pero no es trivial cómo interpretar esos puntajes.

- Los valores mínimos y máximos de estas medidas ofrecen alguna orientación;
  - Por ejemplo, una pureza de 0 indica una agrupación deficiente, mientras que una pureza de 1 señala una buena agrupación, especialmente cuando se alinea con etiquetas de clase o estructuras deseadas.
- Interpretaciones similares se aplican a la entropía y al SSE, donde valores más bajos indican una mejor agrupación.

Evaluación de la significancia de las medidas de validez

- En ocasiones, la falta de valores mínimos o máximos claros, junto con la influencia de la escala de los datos, puede dificultar la interpretación.

- Utilizando la significancia estadística, podemos discernir clústeres que reflejen estructuras no aleatorias, ya que estos deberían generar medidas de validez excepcionalmente altas o bajas.

- En otras palabras, cuanto más atípico sea un resultado, más probable será que refleje estructuras válidas.

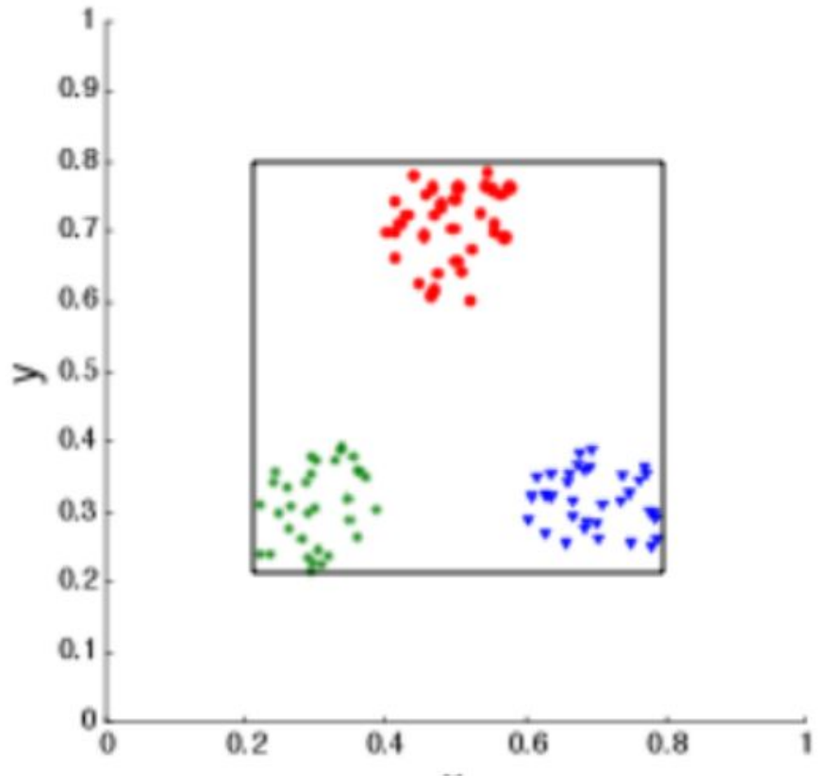
- Por lo tanto, podemos contrastar los índices resultantes de datos aleatorios con los de nuestros datos para determinar su significancia.

- Valores poco probables indican resultados significativos desde un punto de vista estadístico.

Ejemplo con SSE

- Para ilustrar cómo funciona esto, presentamos un ejemplo basado en K-means y el SSE.

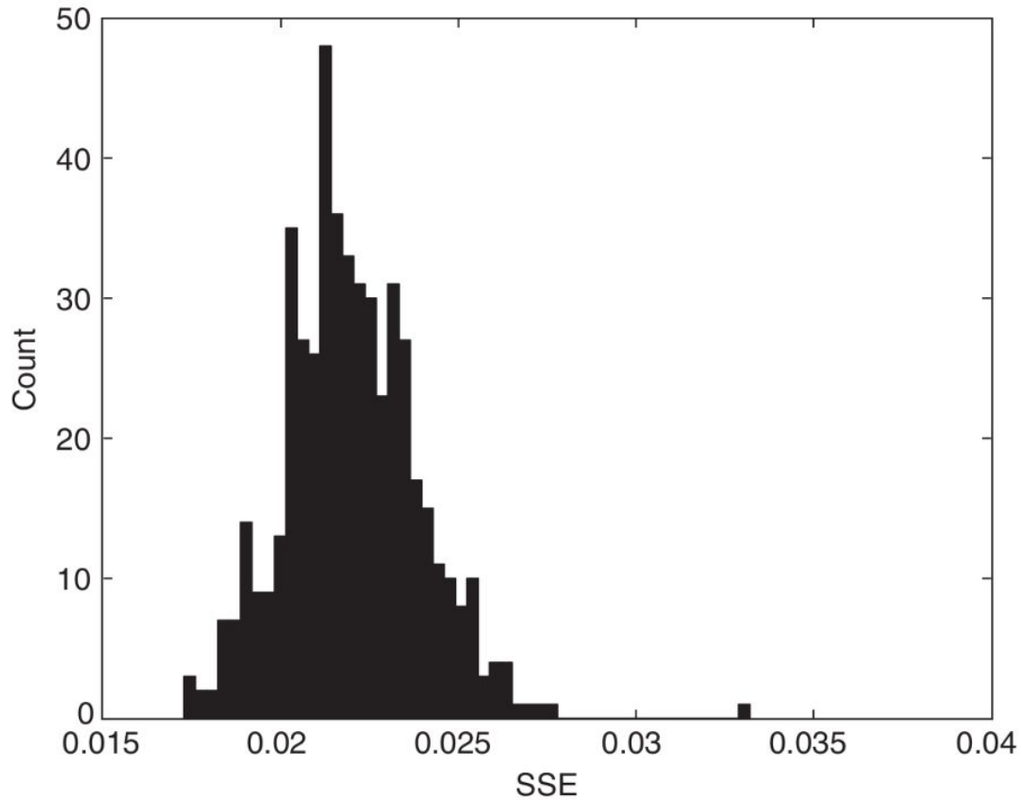
- Supongamos que queremos medir qué tan buenos son los clústeres bien separados de la figura con respecto a datos aleatorios.



Ejemplo con SSE

- Generamos 500 conjuntos aleatorios de 100 puntos que tienen el mismo rango que los puntos en los tres clústeres.

- En cada conjunto de datos, encontramos tres clústeres usando K-means y acumulamos la distribución de valores de SSE para estos agrupamientos.



Ejemplo con SSE

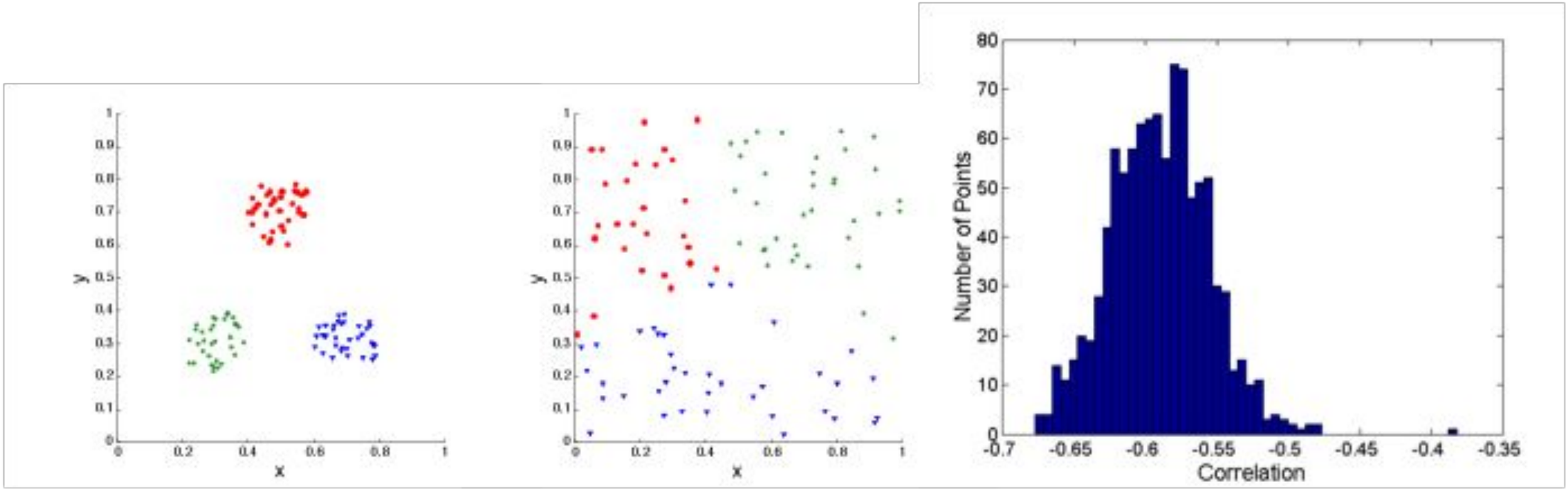
- Utilizando esta distribución de los valores de SSE, podemos estimar la probabilidad del valor de SSE para los clústeres originales.

- El SSE más bajo mostrado en el histograma es 0.0173, mientras que para los tres clústeres originales, el SSE es 0.0050.

- Por lo tanto, podríamos afirmar de manera conservadora que hay menos de un 1% de probabilidad de que un clustering como el obtenido en los datos originales pueda ocurrir por azar.

Otro ejemplo: Correlación

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



Corr = -0.9235

Corr = -0.5810

Validación con Expertos

- La validación con expertos implica evaluar los clústeres para determinar si producen el resultado esperado.

- Esto se puede lograr mediante la revisión de especialistas en el campo (ej: calzan los grupos de clientes encontrados con perfiles conocidos?).

- También se puede generar una clasificación de validación mediante la cual los expertos asignan etiquetas a los clústeres según su conocimiento y criterio.

Conclusiones

- En general, clustering se suele usar como una técnica exploratoria de datos.

- El objetivo no es proporcionar una respuesta definitiva, sino más bien ofrecer una visión de la estructura subyacente de los datos.

- En esta situación, las medidas de validez de clusters que hemos estudiado pueden ser útiles

Conclusiones

- Aún así, evaluar clusters no es fácil: depende de para qué se quiere usar el conjunto de clusters, y cuál es su tipo (particional, jerárquico, basado en densidad, etc).

- Idealmente, se deben utilizar tanto medidas externas como internas para evaluar y validar los clústeres

Bonus

¿Hay algo más que clustering en aprendizaje no supervisado? 👀

Referencias

1. Halkidi, Maria, Yannis Batistakis, and Michalis Vazirgiannis. "On clustering validation techniques." Journal of intelligent information systems 17 (2001): 107-145.
2. Tan, P. N., Steinbach, M., & Kumar, V. (2016). Introduction to data mining. Pearson Education India.

dcc

CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE CHILE

www.dcc.uchile.cl

/ DCCUCHILE