Machine Learning
Clustering
Evaluación

Evaluación

- ¿Cómo podemos evaluar qué tan bueno es un cluster?
- Esto nos ayudaría a:
  - Evitar encontrar cluster en datos de ruido.
  - Comparar algoritmos de clustering
  - Comparar dos conjuntos de clusters

Evaluación

- Evaluación de cluster, enfoques:
  - Determinar la **tendencia de clusters** en los datos, es decir, distinguir si realmente existen estructuras no aleatorias en los datos.
  - Evaluar si los **clusters "se ajustan"** a los datos **(no supervisado)**.
  - Dado multiples clusterizaciones de los datos, determinar cual de ellos representa "mejor" la estructura de los datos.
  - Determinar el número "correcto" de clusters.

Evaluación, tendencia de clusters

- **Tendencia de clusters**: evalúa si existe la presencia de clusters en los datos, antes de realizar el clustering.

- Enfoque más común (para datos numéricos con pocas dimensiones), es usar un test estadístico de aleatoriedad espacial.

- Estadístico de Hopkins: Toma una muestra aleatoria de p puntos desde los datos, y genera p datos aleatorios en el mismo espacio

$$H = \frac{\sum_{i=1}^{p} w_i}{\sum_{i=1}^{p} u_i + \sum_{i=1}^{p} w_i}$$

w_i: distancia del punto aleatorio i, a su vecino más cercano de los puntos **originales**.

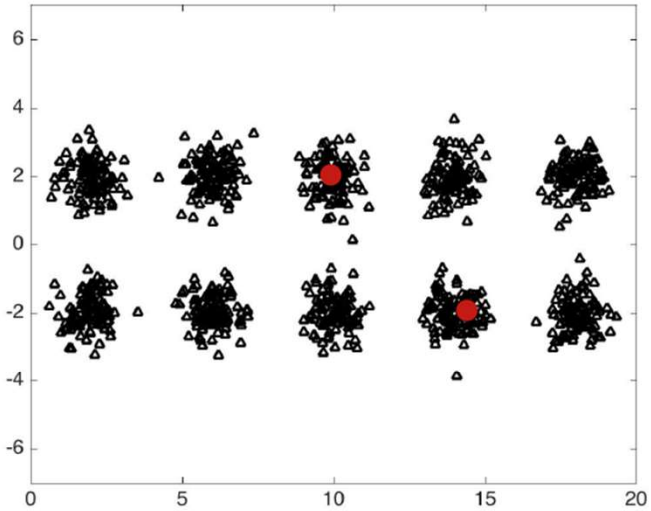
u_i: distancia del punto original i, a su vecino más cercano de los puntos **originales**.

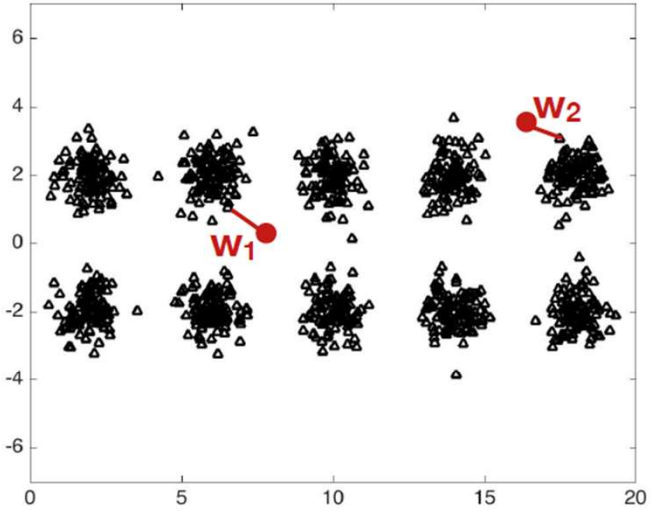
  - Valor cercano a 0.5 indica datos aleatorios.
    Valor cercano a 1.0 indica datos altamente clusterizados.
    Valor cercano a 0.0 indica datos uniformemente distribuidos.

Evaluación, tendencia de clusters, estadístico de Hopkins, ejemplo manual con p = 2

- Seleccionando puntos de los datos.
- La distancia u_1 y u_2 son casi 0.

- Seleccionando puntos aleatorios.
- La distancia w_1 y w_2 son grandes.

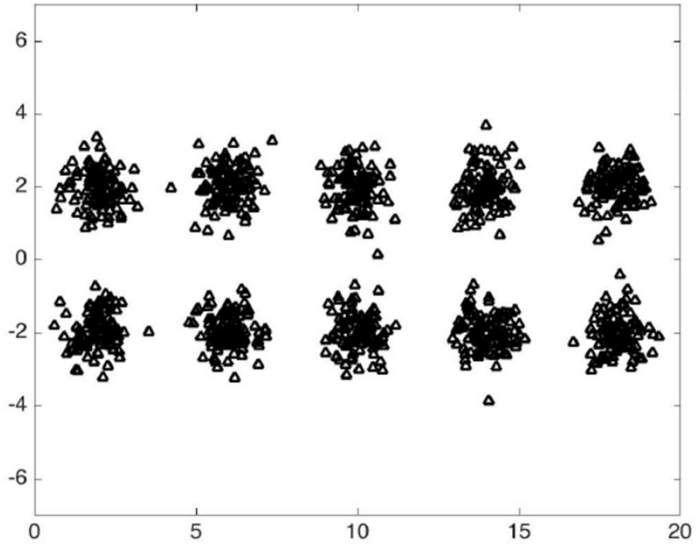




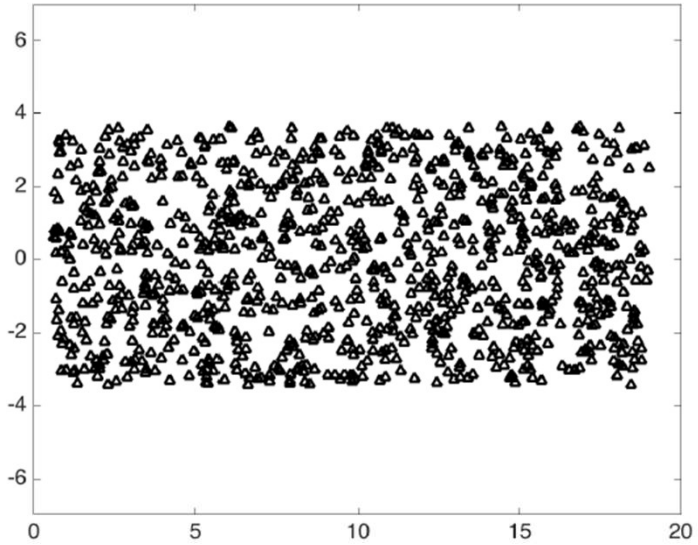
$u_1 \approx 0.00 \quad u_2 \approx 0.00$ $\quad w_1 \approx 1.80 \quad w_2 \approx 1.12$

$$H = \frac{w_1 + w_2}{u_1 + u_2 + w_1 + w_2} \approx \frac{1.80 + 1.12}{0.00 + 0.00 + 1.80 + 1.12} \approx 1.00$$

Evaluación, tendencia de clusters, estadístico de Hopkins, ejemplo



H=0.8191
p=50



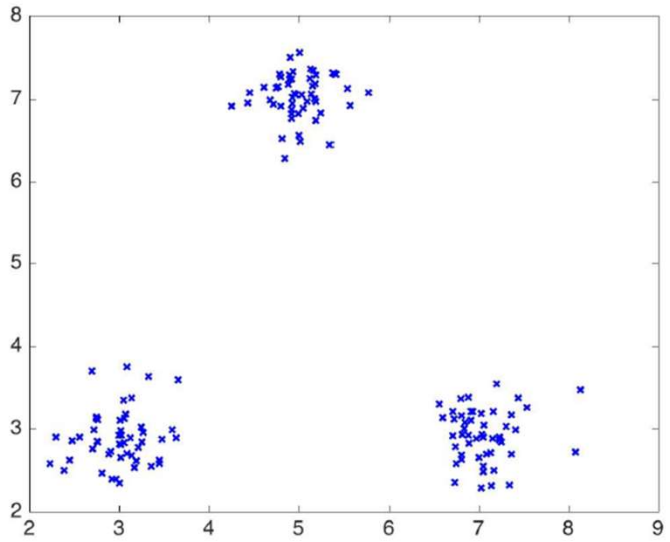
H=0.4975
p=50

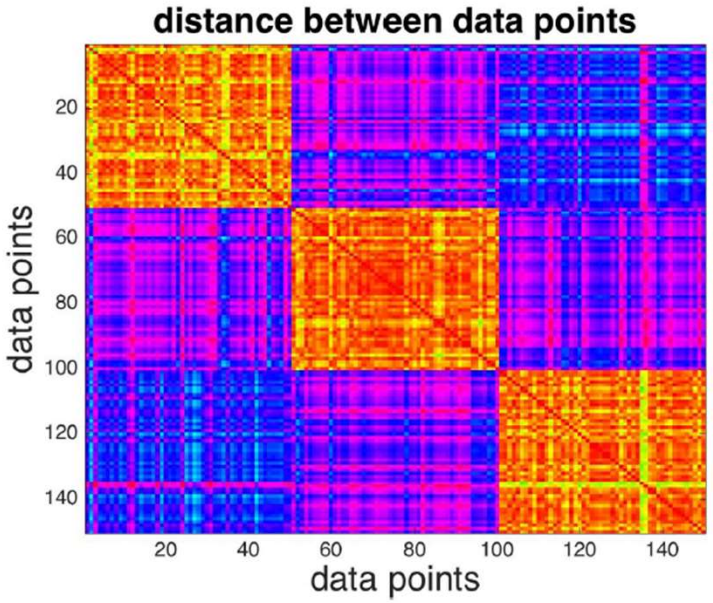
Evaluación, evaluación no supervisada

- **Evaluación no supervisada:** mide el ajuste de los cluster en los datos que no tienen etiquetas/clases.
- Existen diversos enfoques en la evaluación no supervisada:
  - Inspección visual de la matriz de proximidad
  - Medidas internas: **Cohesión**, **Separación**, y **Coeficiente de Silhouette**.

Evaluación, evaluación no supervisada, visual

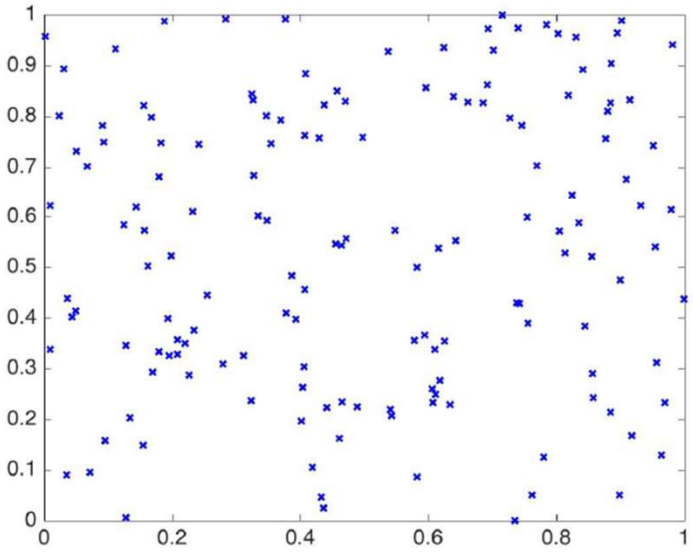
- Calcular la matriz de proximidad (distancia) entre los puntos.
- Ordenar la matrix de proximidad según la distancia entre los puntos.
- Inspeccionar en forma visual (un buen clustering exhibe claros patrones de bloques)

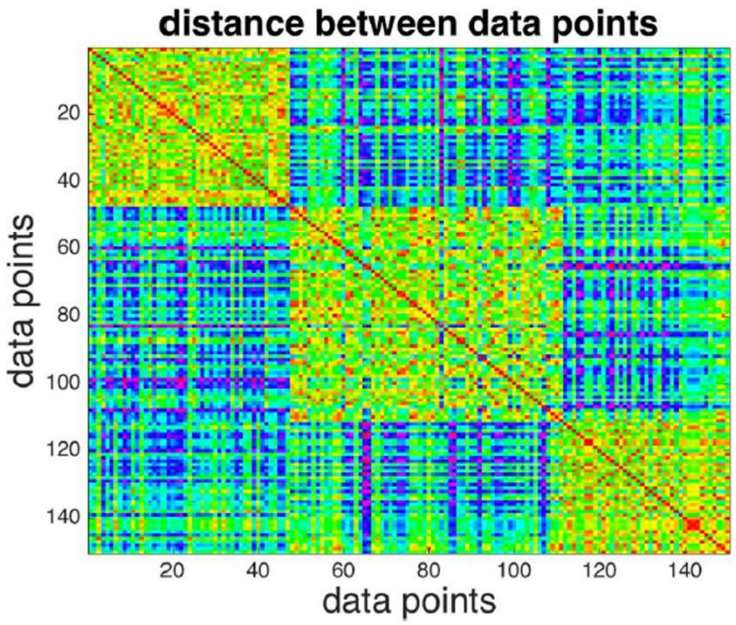




Evaluación, evaluación no supervisada, visual

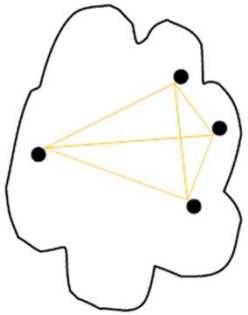
- Calcular la matriz de proximidad (distancia) entre los puntos.
- Ordenar la matrix de proximidad según la distancia entre los puntos.
- Inspeccionar en forma visual (un buen clustering exhibe claros patrones de bloques)





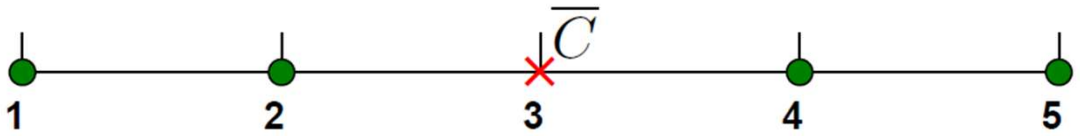
Evaluación, evaluación no supervisada, medidas internas, cohesión

- **Cohesión**: Mide que tan cercanos están los objetos dentro de cada cluster.
- **Suma de errores cuadrados (SSE)** es la suma de la distancia al cuadrado de un punto al centroide de su cluster.



cohesión

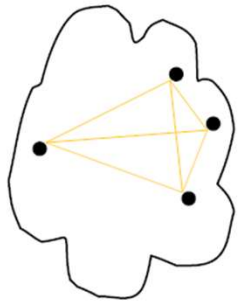
$$SSE_{total} = \sum_{k=1}^{K} \sum_{\mathbf{x} \in C_k} (\mathbf{x} - \overline{C}_i)^2$$



$$\text{K=1} \Rightarrow SSE_{total} = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

Evaluación, evaluación no supervisada, medidas internas, cohesión

- **Cohesión**: Mide que tan cercanos están los objetos dentro de cada cluster.
- **Suma de errores cuadrados (SSE)** es la suma de la distancia al cuadrado de un punto al centroide de su cluster.



cohesión

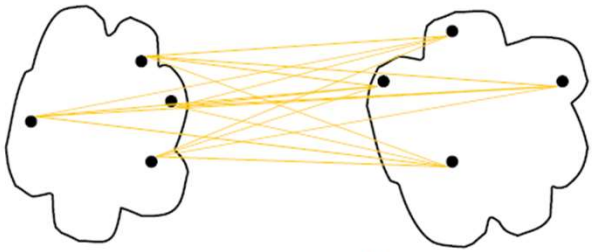
$$SSE_{total} = \sum_{k=1}^{K} \sum_{\mathbf{x} \in C_k} (\mathbf{x} - \overline{C}_i)^2$$

$$\text{K=2} \Rightarrow SSE_{total} = (1 - 1.5)^2 + (2 - 1.5)^2 + (4 - 4.5)^2 + (5 - 4.5)^2 = 1$$

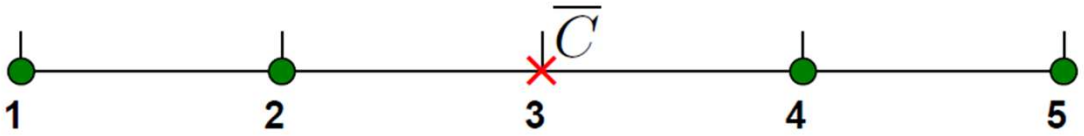
Evaluación, evaluación no supervisada, medidas internas, separación

- **Separación**: Mide que tan distinto son los clusters con respecto a los otros.
- **Suma cuadrada entre grupos (SSB)** es la suma de la distancia al cuadrado de un centroide a la media de todos los datos.

$$SSB_{total} = \sum_{k=1}^{K} |C_i|(\overline{C}_i - \overline{\mathbf{X}})^2$$



separación

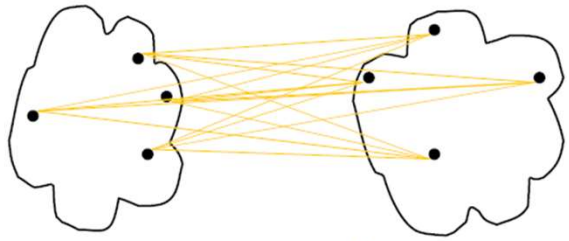


$$\text{K=1} \Rightarrow SSB_{total} = 4 * (3-3)^2 = 0$$

Evaluación, evaluación no supervisada, medidas internas, separación

- **Separación**: Mide que tan distinto son los clusters con respecto a los otros.
- **Suma cuadrada entre grupos (SSB)** es la suma de la distancia al cuadrado de un centroide a la media de todos los datos.

$$SSB_{total} = \sum_{k=1}^{K} |C_i|(\overline{C}_i - \overline{\mathbf{X}})^2$$

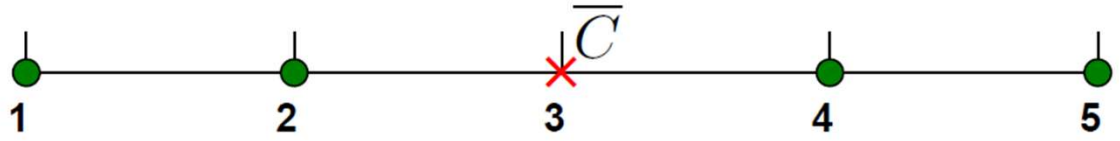


separación

$$\text{K=2} \Rightarrow SSB_{total} = 2 * (1.5 - 3)^2 + 2 * (4.5 - 3)^2 = 9$$

Evaluación, evaluación no supervisada, medidas internas, cohesión y separación

- **Cohesión y separación**: La suma de SSE_{total} y SSB_{total} es igual a la suma de la distancia al cuadrado de todos los puntos con respecto a la media.
- Entonces minimizar cohesión es equivalente a maximizar separación.



$$\text{K=1} \Rightarrow SSE_{total} = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

$$\text{K=1} \Rightarrow SSB_{total} = 4 * (3-3)^2 = 0$$

Evaluación, evaluación no supervisada, medidas internas, cohesión y separación

- **Cohesión y separación**: La suma de SSE_{total} y SSB_{total} es igual a la suma de la distancia al cuadrado de todos los puntos con respecto a la media.
- Entonces minimizar cohesión es equivalente a maximizar separación.

$$\text{K=2} \Rightarrow SSE_{total} = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

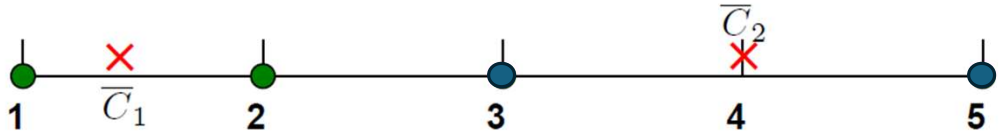
$$\text{K=2} \Rightarrow SSB_{total} = 2 * (1.5-3)^2 + 2 * (4.5-3)^2 = 9$$

Evaluación, evaluación no supervisada, medidas internas, coeficiente de silhouette (silueta)

- El **coeficiente de Silhouette** combina cohesión y separación. Normalmente, varia entre [-1,1], con valores cercanos a 1 indicando una mejor clusterización.
- Para cada punto i:
  - Calcular a_i, la distancia promedio del punto i a los puntos del mismo cluster.
  - Calcular b_{ij}, la distancia promedio del punto i a todos los puntos del cluster j.
  - Calcular b_i, el mínimo b_{ij} tal que i no pertenezca al cluster j.
  - El coeficiente de Silhouette para el punto i es $S_i = (b_i - a_i) / \max(a_i, b_i)$
- Un valor negativo implica que el punto i es más cercano a otro cluster, que a su propio cluster. Si a_i es cercano a 0 (baja cohesión), entonces S_i es cercano a 1.
- El coeficiente de silhouette de un cluster es el promedio de los coeficientes de silhouette de los puntos pertenecientes al cluster.
- El coeficiente de silhouette general es el promedio de los coeficientes de silhouette de todos los puntos.

Evaluación, evaluación no supervisada, medidas internas, coeficiente de silhouette, ejemplo

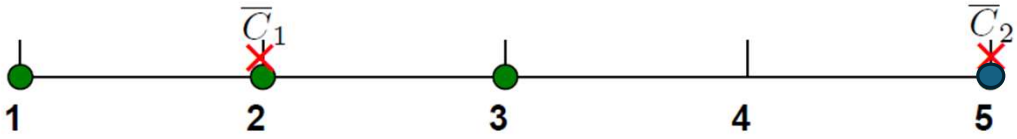
- a_i => la distancia promedio del punto i a los puntos del mismo cluster.
  b_{ij} => la distancia promedio del punto i a todos los puntos del cluster j.
  b_i => el mínimo b_{ij} tal que i no pertenezca al cluster j.
  El coeficiente de Silhouette para el punto i es $S_i = (b_i - a_i) / \max(a_i, b_i)$



	a_i	b_{i1}	b_{i2}	b_i	s_i
1	1.0	—	3.0	3.0	2/3
2	1.0	—	2.0	2.0	1/2
3	2.0	1.5	—	1.5	-0.5/2
5	2.0	3.5	—	3.5	1.5/2

Evaluación, evaluación no supervisada, medidas internas, coeficiente de silhouette, ejemplo

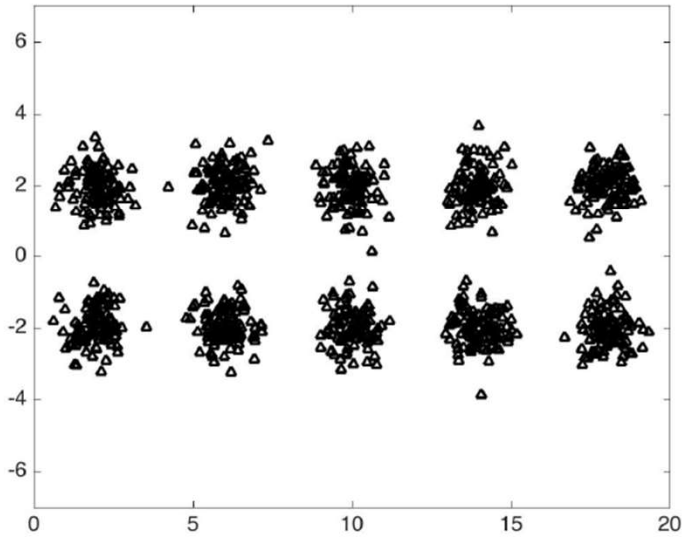
- a_i => la distancia promedio del punto i a los puntos del mismo cluster.
  b_{ij} => la distancia promedio del punto i a todos los puntos del cluster j.
  b_i => el mínimo b_{ij} tal que i no pertenezca al cluster j.
  El coeficiente de Silhouette para el punto i es $S_i = (b_i - a_i) / \max(a_i, b_i)$

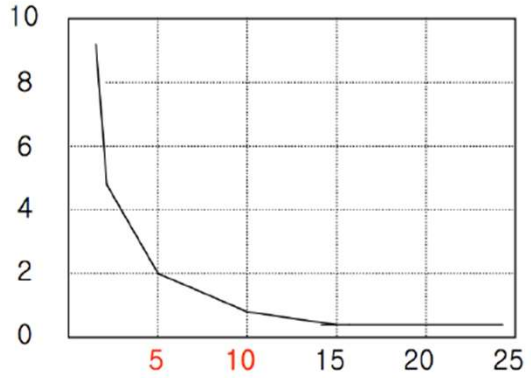


	a_i	b_{i1}	b_{i2}	b_i	s_i
1	1.5	—	4.0	4.0	2.5/4
2	1.0	—	3.0	3.0	2/3
3	1.5	—	2.0	2.0	0.5/2
5	0.0	3.0	—	3.0	1.0

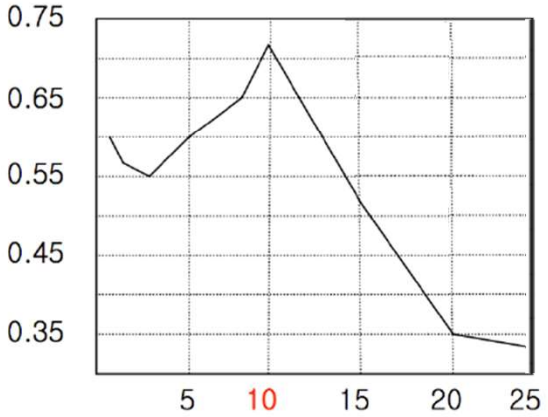
Evaluación, determinar K

- Para determinar el mejor valor de K, hay que evaluar alguna medida específica (Silhouette, SSE_{total}, BIC), sobre un rango de K clusters, y mirar por un peak, "bajada", mínimo, o codo en la medida de evaluación.





SSE



Silhouette

Evaluación, determinar K, ejemplo

- WCD sugiere entre 3 y 5 clusters.
- Silhouette sugiere de 2 a 4 clusters.
- Se podría analizar 3 y 4 clusters.