Clustering 3

Métodos de validación: ¿Por qué es importante validar?

- Siempre vamos a encontrar grupos PERO ¿Son buenos grupos?
- Determinar el mejor número de clusters (k).
- Comparar métodos de Clustering (sobre los mismos datos).

Métodos de validación: ¿Por qué es importante validar? Validación Externa vs Validación Interna

- Siempre vamos a encontrar grupos PERO ¿Son buenos grupos?
  - Antes de empezar a buscar es importante verificar si existe una ***tendencia al clustering***.

- Determinar el mejor número de clusters (k).

- Comparar métodos de Clustering (sobre los mismos datos).

Métodos de validación: ¿Por qué es importante validar? Validación Externa vs Validación Interna

- Siempre vamos a encontrar grupos PERO ¿Son buenos grupos?
  - Antes de empezar a buscar es importante verificar si existe una ***tendencia al clustering***.

- Determinar el mejor número de clusters (k).
  - Determinar cuán buenos son los grupos (***externa***/***interna***).
  - Determinar cuán buena es la separación entre ellos (***externa***/***interna***).
  - Constatar el número de clusters con las etiquetas *a priori* (si existen) (***externa***).

- Comparar métodos de Clustering (sobre los mismos datos).
  - Determinar cuán buenos son los grupos (***externa***/***interna***).
  - Determinar cuán buena es la separación entre ellos (***externa***/***interna***).
  - Constatar la pertenencia a los grupos con las etiquetas *a priori* (si existen) (***externa***).

Métodos de validación: ¿Por qué es importante validar? Validación Externa vs Validación Interna

- Siempre vamos a encontrar grupos PERO ¿Son buenos grupos?
  - Antes de empezar a buscar es importante verificar si existe una ***tendencia al clustering***.

- Determinar el mejor número de clusters (k).
  - Determinar cuán buenos son los grupos (***externa***/***interna***).
  - Determinar cuán buena es la separación entre ellos (***externa***/***interna***).
  - Constatar el número de clusters con las etiquetas *a priori* (si existen) (***externa***).

- Comparar métodos de Clustering (sobre los mismos datos).
  - Determinar cuán buenos son los grupos (***externa***/***interna***).
  - Determinar cuán buena es la separación entre ellos (***externa***/***interna***).
  - Constatar la pertenencia a los grupos con las etiquetas *a priori* (si existen) (***externa***).

Existen diferentes medidas de validación y no existe un criterio único para determinar cuál es la mejor.

No hay una medida única que se pueda usar para todos los métodos de clustering.

Criterios general: Cohesión y Separación

- ***Cohesión***: es una medida de las proximidades de los miembros de un clúster con respecto al prototipo.
- ***Separación***: es la proximidad entre miembros de diferentes clústeres o entre prototipos de grupos y el prototipo general.

c_i = centroide o medoide (prototipo)

Suma de los Errores al Cuadrado (dentro de un cluster) (SSE)

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist(c_i, x)^2$$

Suma Total de los Errores al Cuadrado (TSE)

$$TSE = \text{SSE} + \text{SSB}$$

Suma de Cuadrados de Separación (SSB)

Tendencia al Clustering

Estadísitico (coeficiente) de Hopkins

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

w_i = distancia de un elemento i al azar al vecino más cercano

u_i = distancia de un punto i agregado al azar al vecino más cercano

Validación externa

1. Matriz de confusión

Clusters Labels	0	1	2	3	4
a	1	25	1	0	13
b	7	7	2	0	24
c	34	0	2	0	4
d	0	0	39	0	1
e	0	0	0	40	0

Validación externa

***2. Medida Normalizada de van Dongen**: Medida mejorada de la pureza (que mide cuánto se aleja de tener sólo tengo un elemento por fila|columna)*

Clusters / Labels	0	1	2	3	4
a	1	25	1	0	13
b	7	7	2	0	24
c	34	0	2	0	4
d	0	0	39	0	1
e	0	0	0	40	0

$$VD_n = \frac{(2n - \sum_i \max_j n_{ij} - \sum_j \max_i n_{ij})}{(2n - \max_i n_{i.} - \max_j n_{.j})}$$

Validación externa

***3. Índice Rand e Índice Rand Ajustado o Normalizado (Adjusted Rand Index, ARI)**:*

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

a = número de pares de elementos que aparecen juntos en un clúster y además pertenecen a la misma clase.

b = número de pares de elementos que pertenecen a clases diferentes y además están en clústeres diferentes.

c = número de pares de elementos que comparten la clase, pero se ubican en diferentes clústeres.

d = número de elementos que pertenecen a clases diferentes, sin embargo se agrupan en el mismo clúster.

n = número total de elementos.

Validación externa

***3. Índice Rand e Índice Rand Ajustado o Normalizado (Adjusted Rand Index, ARI)**:*

$$R = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

ARI = (R - E(R)) / (max(R) - E(R))

E(R) = valor esperado de R si se distribuyen al azar.

max(R) = valor máximo posible de R para los datos.

Validación Interna

***1. Coeficiente de Silhoutte**:*

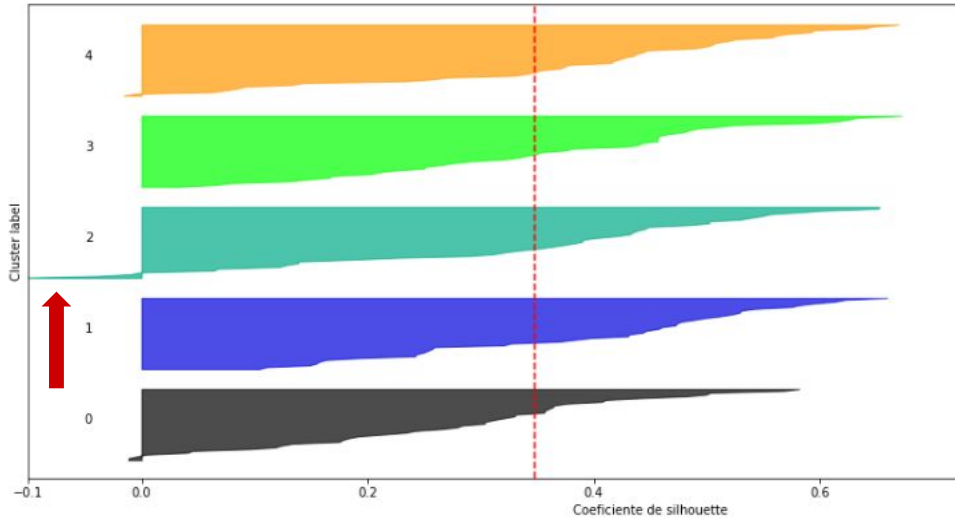
$$s_i = \frac{(b_i - a_i)}{max(a_i, b_i)}$$

1. Para cada elemento ***i*** se calcula su distancia promedio a todos los otros elemento de su clúster (***a_i***).
2. Para el elemento ***i*** y todos los otros clústeres que no lo contienen, se calcula las distancias promedio a todos los elementos de cada clúster.
3. Se buscar el mínimo de esas distancias promedio a cada clúster (***b_i***).
4. Se calcula el coeficiente Silhouette (***s_i***) del elemento i.
5. Luego se puede calcular el promedio para cada cluster o el promedio global.

Validación Interna

***1. Coeficiente de Silhoutte**:*

Puede ocurrir que algunos clusters tengan peor coeficiente o algunos elementos dentro del cluster. Esto puede ser indicativo de que quizás es mejor cambiar el valor de k.



Validación Interna

2. (Jerárquico) Coeficiente de Correlación Cofenético: Mide la correlación entre la matriz de distancia que dio origen al agrupamiento y los distancias extraídas del árbol (Altura del nodo que une por primera vez dos elementos).

3. (Jerárquico) Partición del árbol por distancia o por número de clusters*:* y luego se pueden aplicar las medidas de validación interna o externa igual que otros métodos.

4. Bootstrapping: Sirve para evaluar la estabilidad de los clusters, y así determinar cuáles son “reales” y cuáles no. Vamos a volver a estos métodos más adelante.