

# *Clustering*

## Estadística Multivariante

Sofía Almeida Bruno  
Daniel Bolaños Martínez  
José María Borrás Serrano  
Fernando de la Hoz Moreno  
Pedro Manuel Flores Crespo  
María Victoria Granados Pozo

20 de enero de 2020

# Clustering

- Objetivo: agrupar objetos similares.
- Dadas  $x_1, \dots, x_n$  medidas de  $p$  variables en  $n$  objetos considerados *heterogéneos*. El objetivo del análisis clúster es agrupar estos objetos en  $k$  clases *homogéneas*, donde  $k$  es también desconocido.

# Clustering

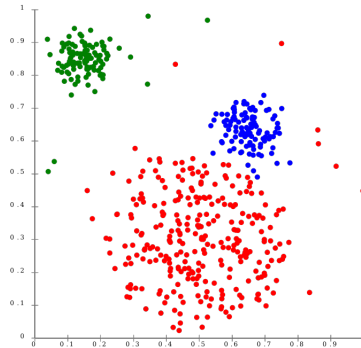


Figura: Ejemplo de *clustering*. [Chi11]

## Ejemplos de *Clustering*

- Biología: determinación de especies.
- *Marketing*: descubrimiento de grupos de clientes.



Figura: Ejemplo de *clustering*. [noa]

- Psicología: encontrar tipos de personalidad.
- Arqueología: datar objetos encontrados.
- Planificación urbana: identificar grupos de viviendas.

# *Clustering*

Para realizar un análisis clúster hay que:

- Elegir una medida de similitud.
- Elegir un algoritmo para construir los grupos.
  - ▶ Particionamiento.
  - ▶ Jerárquicos.

## Medidas de similitud

# Medidas de similitud

Consideraciones iniciales como:

- Naturaleza de las variables (discreta, continua, binaria).
- Escalas de las medidas (nominal, ordinal, intervalo).
- Conocimiento sobre el problema.

Los valores de las variables consideradas deberán ser normalizados.

## Distancias de similitud para pares de ítems

La distancia estadística entre dos observaciones  $p$ -dimensionales  $x^T = [x_1, \dots, x_p]$  e  $y^T = [y_1, \dots, y_p]$  es:

$$d(x, y) = \sqrt{(x - y)^T A (x - y)}.$$

Donde:

- $A = S^{-1}$ .
- $S$  contiene las varianzas y covarianzas de la muestra.



## Otras medidas y coeficientes de similitud

- Métrica de Minkowski:

$$d(x, y) = \left[ \sum_{i=1}^p |x_i - y_i|^m \right]^{1/m}.$$

- Métrica de Canberra (variables no negativas):

$$d(x, y) = \sum_{i=1}^p \frac{|x_i - y_i|}{(x_i + y_i)}.$$

- Coeficiente de Czekanowski (variables no negativas):

$$d(x, y) = 1 - \frac{2 \sum_{i=1}^p \min(x_i, y_i)}{\sum_{i=1}^p (x_i + y_i)}.$$

## Frecuencias de las parejas

Organizamos las frecuencias en la siguiente **tabla de contingencia**:

		Ítem $k$		Total
		1	0	
Ítem $i$	1	$a$	$b$	$a + b$
	0	$c$	$d$	$c + d$
Total		$a + c$	$b + d$	$p = a + b + c + d$

## Coeficientes de similitud para ítems *clustering*

Coeficiente	Fundamento
1 $\frac{a+d}{p}$	Las parejas 1-1 y 0-0 ponderan lo mismo.
2 $\frac{2(a+d)}{2(a+d)+b+c}$	Las parejas 1-1 y 0-0 ponderan el doble.
3 $\frac{a+d}{a+d+2(b+c)}$	Las parejas que no coinciden ponderan el doble.
4 $\frac{a}{p}$	No hay parejas 0-0 en el numerador.

## Coeficientes de similitud para ítems *clustering*

Coeficiente	Fundamento
5 $\frac{a}{a+b+c}$	No hay parejas 0-0 en el numerador ni el denominador (Las parejas 0-0 son irrelevantes).
6 $\frac{2a}{2a+b+c}$	No hay parejas 0-0 en el numerador ni el denominador. Las parejas 1-1 ponderan el doble.
7 $\frac{a}{a+2(b+c)}$	No hay parejas 0-0 en el numerador ni el denominador. Las parejas que no coinciden ponderan el doble.
8 $\frac{a}{b+c}$	Proporción de parejas que coinciden (excluyendo las 0-0) en relación a las parejas que no coinciden.

## Construcción de similitudes y distancias

- Siempre se pueden construir similitudes a partir de distancias.

Fijando  $s_{ik} = \frac{1}{1+d_{ik}}$  donde  $0 < s_{ik} \leq 1$  es la similitud entre los ítems  $i$  y  $k$ , entonces  $d_{ik}$  es la distancia correspondiente.

- Las distancias se pueden construir a partir de similitudes si la matriz de similitudes es definida no negativa y la máxima similitud cumple  $s_{ii} = 1$ .

Entonces  $d_{ik} = \sqrt{2 \cdot (1 - s_{ik})}$ , cumple las propiedades de una distancia.

## Medidas de similitud para pares de variables

Las medidas de similitud para variables suelen tomar la forma de coeficientes de correlaciones muestrales.

Cuando las variables son binarias, los datos se pueden organizar en una **tabla de contingencia** que tiene la siguiente forma:

		Variable $k$		Total
		1	0	
Variable $i$	1	$a$	$b$	$a + b$
	0	$c$	$d$	$c + d$
Total		$a + c$	$b + d$	$n = a + b + c + d$

## Medidas de similitud para pares de variables

La fórmula del coeficiente de correlación producto-momento aplicada a las variables binarias de la tabla de contingencia nos da:

$$r = \frac{ad - bc}{[(a + b)(c + d)(a + c)(b + d)]^{1/2}}.$$

$r$  se puede tomar como la medida de similitud entre las dos variables.

## Ejemplo idiomas

Medimos las similitudes de 11 lenguajes en base a los primeros 10 números naturales en cada idioma.

Inglés (E)	Noruego (N)	Danés (Da)	Holandés (Du)	Alemán (G)	Francés (Fr)	Español (Sp)	Italiano (I)	Polaco (P)	Húngaro (H)	Finés (Fi)
one	en	en	een	eins	un	uno	uno	jeden	egy	yksi
two	to	to	twee	zwei	deux	dos	due	dwa	ketto	kaksi
three	tre	tre	drie	drei	trois	tres	tre	trzy	harom	kolme
four	fire	fire	vier	vier	quatre	cuatro	quattro	cztery	negy	nelja
five	fem	fem	vijf	funf	cinq	cinco	cinque	piec	ot	viisi
six	seks	seks	zes	sechs	six	seis	sei	szesc	hat	kuusi
seven	sju	syv	zeven	sieben	sept	siete	sette	siedem	het	seitseman
eight	atte	otte	acht	acht	huit	ocho	otto	osiem	nyolc	kahdeksan
nine	ni	ni	negen	neun	neuf	nueve	nove	dziewiec	kilenc	yhdeksan
ten	ti	ti	tien	zehn	dix	diez	dieci	dziesiec	tiz	kymmenen



	E	N	Da	Du	G	Fr	Sp	I	P	H	Fi
E	10										
N	8	10									
Da	8	9	10								
Du	3	5	4	10							
G	4	6	5	5	10						
Fr	4	4	4	1	3	10					
Sp	4	4	5	1	3	8	10				
I	4	4	5	1	3	9	9	10			
P	3	3	4	0	2	5	7	6	10		
H	1	2	2	2	1	0	0	0	0	10	
Fi	1	1	1	1	1	1	1	1	1	2	10

Vemos que inglés, noruego, danés, holandés y alemán parecen formar un grupo. El francés, español, italiano y polaco forman otro, mientras que el húngaro y el finés no forman parte de ninguno.

## Métodos de agrupamiento

# Métodos de agrupamiento

**Definición:** procedimiento de agrupación de una serie de vectores de acuerdo con un criterio (distancia o similitud).

## **Tipos:**

- Jerárquicos.
- No jerárquicos o particionamiento.

# Métodos de agrupamiento

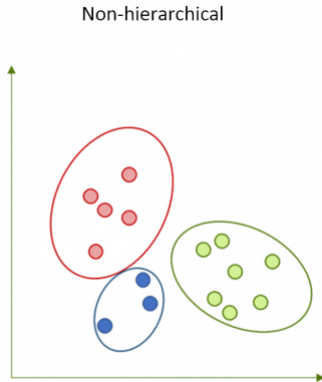
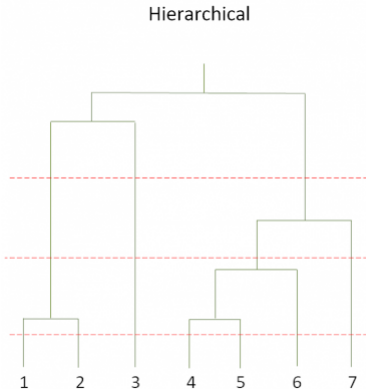


Figura: Comparación métodos jerárquico y no jerárquico.

## Métodos de agrupamiento Jerárquicos

# Métodos Jerárquicos

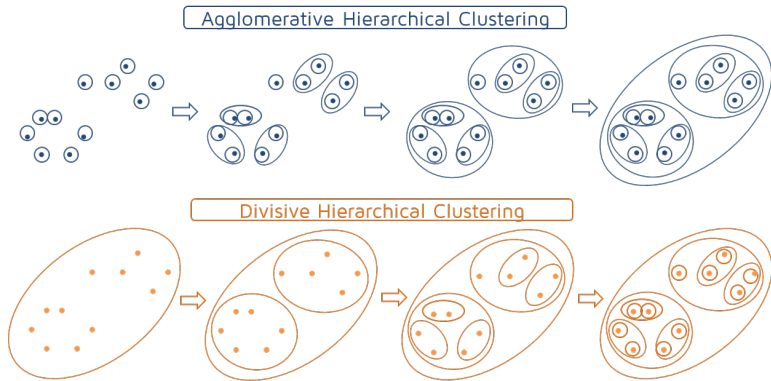
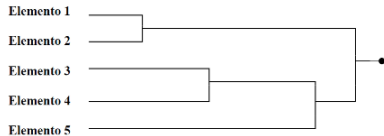
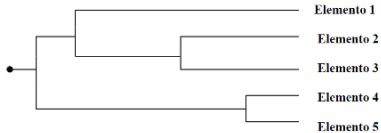


Figura: Comparación métodos aglomerativos y divisivos.

## Comparación de técnicas aglomerativas y divisivas



(a) Dendrograma aglomerativo.



(b) Dendrograma divisivo.

## Ejemplo *Single Link Method*

Distancia entre dos clústers

$$d(R, S) = \min\{d_{rs} : r \in R, s \in S\}. \quad (1)$$

Matriz de distancias

$$\mathcal{D} = [d_{rs}] = \begin{array}{c} \begin{array}{ccccc} & 1 & 2 & 3 & 4 & 5 \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{bmatrix} 0 & 2 & 4 & 7 & 9 \\ 2 & 0 & 8 & 9 & 8 \\ 4 & 8 & 0 & 3 & 7 \\ 7 & 9 & 3 & 0 & 5 \\ 9 & 8 & 7 & 5 & 0 \end{bmatrix} \end{array} \end{array}.$$



## Ejemplo *Single Link Method*

Cálculo de las nuevas distancias:

$$d_{(12)(3)} = \min\{d_{13}, d_{23}\} = \min\{4, 8\} = 4,$$

$$d_{(12)(4)} = \min\{d_{14}, d_{24}\} = \min\{7, 9\} = 7,$$

$$d_{(12)(5)} = \min\{d_{15}, d_{25}\} = \min\{9, 8\} = 8.$$

**Matriz de distancias**

$$\mathcal{D}_1 = \begin{array}{cc} & \begin{array}{cccc} (12) & 3 & 4 & 5 \end{array} \\ \begin{array}{c} (12) \\ 3 \\ 4 \\ 5 \end{array} & \begin{bmatrix} 0 & 4 & 7 & 8 \\ 4 & 0 & \mathbf{3} & 7 \\ 7 & \mathbf{3} & 0 & 5 \\ 8 & 7 & 5 & 0 \end{bmatrix} \end{array}.$$

## Ejemplo *Single Link Method*

Cálculo de las nuevas distancias:

$$\begin{aligned}d_{(34)(12)} &= \min\{d_{(3)(12)}, d_{(4)(12)}\} = \min\{4, 7\} = 4, \\d_{(34)(5)} &= \min\{d_{(3)(5)}, d_{(4)(5)}\} = \min\{7, 5\} = 5.\end{aligned}$$

**Matriz de distancias**

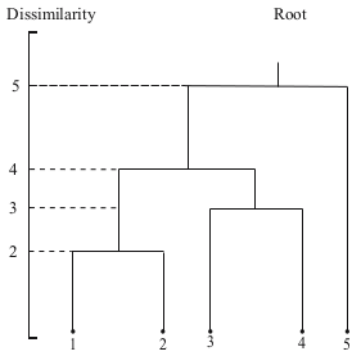
$$\mathcal{D}_2 = \begin{array}{cc} & \begin{array}{ccc} (12) & (34) & 5 \end{array} \\ \begin{array}{c} (12) \\ (34) \\ 5 \end{array} & \left[ \begin{array}{ccc} 0 & 4 & 8 \\ 4 & 0 & 5 \\ 8 & 5 & 0 \end{array} \right].\end{array}$$

## Ejemplo *Single Link Method*

Cálculo de la nueva distancia:

$$d_{(12)(34)5} = \min\{d_{(12)(5)}, d_{(34)(5)}\} = \min\{8, 5\} = 5.$$

Finalmente se obtiene el clúster  $(12345)$ .



## Métodos de agrupamiento

### Particionamiento

## *K-medias*

1. Entrada:  $\mathbf{L} = \{\mathbf{x}_i, i = 1, 2, \dots, n\}$ ,  $K$ =número de clústeres.
2. Hacer uno de los siguientes:
  - ▶ Formar una asignación aleatoria inicial de los datos en los  $K$  clústeres y, para los  $K$  clústeres, calcular su centroide,  $\bar{\mathbf{x}}_k, k = 1, 2, \dots, K$ .
  - ▶ Pre-especificar los centroides de los  $K$  clústeres,  $\bar{\mathbf{x}}_k, k = 1, 2, \dots, K$ .
3. Calcular la distancia euclídea al cuadrado para cada dato al centroide de su clúster actual:

$$ESS = \sum_{k=1}^K \sum_{c(i)=k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)^T (\mathbf{x}_i - \bar{\mathbf{x}}_k)$$

donde  $\bar{\mathbf{x}}_k$  es el centroide del  $k$ -ésimo clúster y  $c(i)$  es el clúster que contiene  $\mathbf{x}_i$ .

## *K-medias*

4. Reasignamos cada dato al clúster con el centroide más cercano de tal manera que ESS se reduce en magnitud. Actualizamos los centroides de los clústeres después de la reasignación de los datos.
5. Repetimos los pasos 3 y 4 hasta que no se produzcan más reasignaciones.

## Otros métodos de particionamiento

K-Medoides:

- Medoide: dato más representativo del clúster.
- No se utiliza la distancia euclídea.

PAM:

- Modificación de K-Medoides.

Análisis difuso:

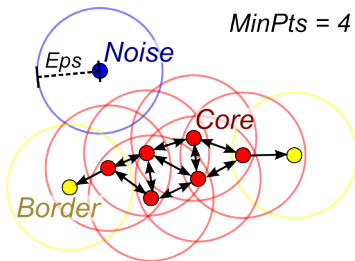
- Se asigna una probabilidad de pertenecer a cada clúster.

Mean Shift:

- Método iterativo que parte de una estimación inicial  $x$ .
- Localiza los máximos de una función de densidad.

## DBSCAN

- Se basa en la densidad de las muestras para identificar los clústeres.



- Todos los puntos de un mismo clúster están conectados entre sí.
- Si un punto  $A$  es alcanzable desde cualquier otro punto  $B$  del clúster, entonces  $A$  también forma parte del clúster.



Número de clústeres

## Método del codo

Consiste en dibujar la gráfica de las distancia a los centros de cada clúster en función del número de clústeres. Definimos:

$$SSE_k = \sum_{i=1}^{n_k} \|\mathbf{y}_i - \bar{\mathbf{y}}_k\|^2,$$

y para cada  $k$  dibujamos

$$D_k = \sum_{i=1}^k SSE_k.$$

## Método del codo

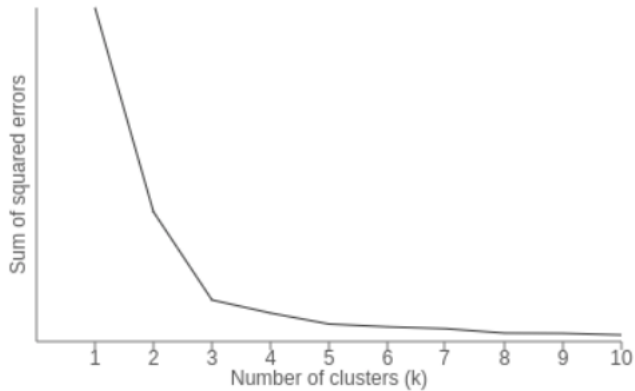


Figura: Ejemplo del método del codo [elb].

## Estadístico $R^2$

Para  $n$  clústeres la suma total de las distancias al cuadrado es  $T = \sum_{i=1}^n \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2$ . Así, para  $k$  clústeres definimos  $R^2$  como

$$R_k^2 = \frac{T - \sum_k SSE_k}{T}.$$

Para  $n$  clústeres  $SSE_k = 0$  por lo que  $R^2 = 1$ . Una gran disminución en  $R_k^2$  representaría un mal agrupamiento.

También podríamos tener en cuenta el cambio en  $R^2$  al unir los clústeres  $R$  y  $S$  como  $SR^2 = R_k^2 - R_{k-1}^2$ .

## Varianza agrupada

Para un solo clúster

$$s^2 = \sum_{i=1}^n \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2 / p(n-1).$$

Para el clúster  $C_k$

$$s^2 = \sum_{i=1}^{n_k} \|\mathbf{y}_i - \bar{\mathbf{y}}_k\|^2 / p(n_k - 1).$$

Valores grandes de la varianza agrupada indica que los clústeres no son homogéneos. Por lo tanto, si tiende a cero para algún  $k < n$  indica la formación de un clúster homogéneo.

## Pseudo estadísticos

El pseudo estadístico  $F$  se define como

$$F_k^* = \frac{(T - \sum_k SSE_k)/(k - 1)}{\sum_k SSE_k/(n - k)}.$$

El pseudo estadístico  $t^2$  se define como

$$\text{pseudo } t^2 = \frac{[SSE_t - (SSE_r + SSE_s)](n_R + n_S - 2)}{SSE_r + SSE_s}.$$

## *Silhouette method*

Definimos el índice:

$$s(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}}, \quad \forall i = 1, \dots, n$$

donde

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, j \neq i} d(i, j)$$

y

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j).$$

Se escoge el  $k$  que maximice el valor medio de  $s(i)$ .

## *Silhouette method*

k	Silhouette coeff.
2	0.7049787496083262
3	0.5882004012129721
4	0.6505186632729437
5	0.5745566973301872
6	0.43902711183132426

Cuadro: Ejemplo *silhouette method* [sil].

Vemos que se obtienen los mejores resultados con 2 o 4 clústeres.



## Gap method

El  $k$  elegido será aquel que maximice el valor de:

$$\text{Gap}(k) = E_n^* \{\log(W_k)\} - \log(W_k).$$

En la fórmula anterior  $E_n^*$  denota la media de una muestra de tamaño  $n$  y

$$W_k = \sum_{R=1}^k \frac{1}{2n_R} \sum_{ij \in C_R} d(i, j).$$

## Gap method

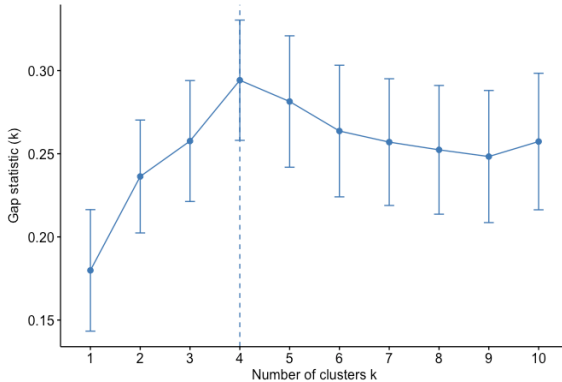


Figura: Ejemplo del método de la brecha [gap].

# Conclusiones

- Busca relaciones entre objetos.
- Depende del conjunto de datos, variables seleccionadas, medida de proximidad y método de agrupamiento.
- Métodos jerárquicos: exploratorios, métodos no jerárquicos: confirmatorios.
- Problema: validación de la solución.

## Caso Práctico: *Clustering* en Python

## Flor de Iris

Estudiaremos el conjunto de datos iris de **Fisher**.

Contiene 50 muestras de cada una de tres especies de flor **Iris**. Para cada muestra, se recogen las medidas de: largo y ancho del sépalo y largo y ancho del pétalo, en centímetros.



Figura: Iris Setosa.



Figura: Iris virginica.



Figura: Iris versicolor.

## Métricas utilizadas

Se utilizarán las siguientes métricas para medir la bondad de los algoritmos:

- **Calinski-Harabaz:** Nos indica si estamos usando un buen número de clústeres para un algoritmo en concreto.
- **Silhouette:** Cuanto mayor sea su valor, más similar será un objeto respecto a su grupo y más diferente a los de otros clúster. Toma valores entre -1 y +1.

## Tabla comparativa de los algoritmos

Nombre	Nº clústeres	CH	SH	Tiempo (s)	Clústeres
K-Means	3	359.845074	0.504769	0.016456	0: 61 (40.67 %) 1: 50 (33.33 %) 2: 39 (26.00 %)
DBSCAN	4	94.991819	0.306404	0.002353	0: 45 (30.00 %) 1: 39 (26.00 %) -1: 36 (24.00 %) 2: 30 (20.00 %)
AggCluster	3	349.254185	0.504800	0.019058	0: 67 (44.67 %) 1: 50 (33.33 %) 2: 33 (22.00 %)
MeanShift	3	290.470683	0.476961	0.289073	0: 81 (54.00 %) 1: 50 (33.33 %) 2: 19 (12.67 %)

## K-Means

Para el algoritmo **K-Means** se ha utilizado el siguiente código en python y se ha obtenido la siguiente agrupación de las muestras:

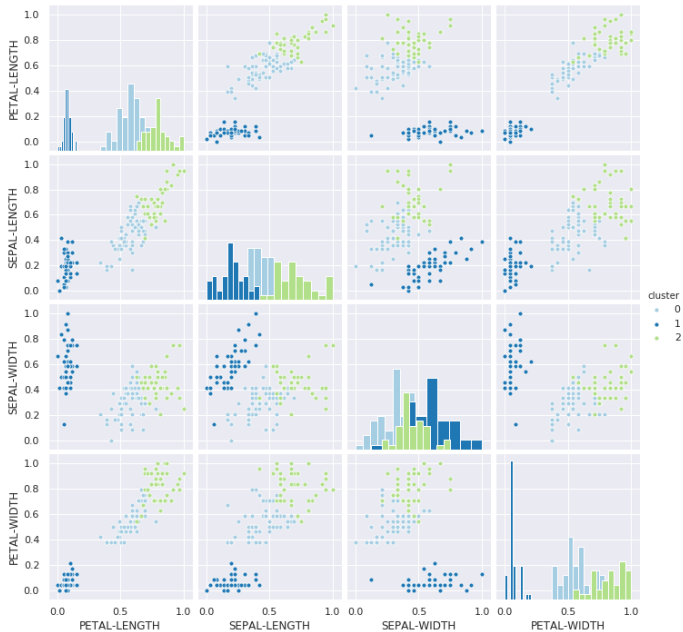
```
KMeans(init='k-means++', n_clusters=3,  
       n_init=5, random_state=12345)
```

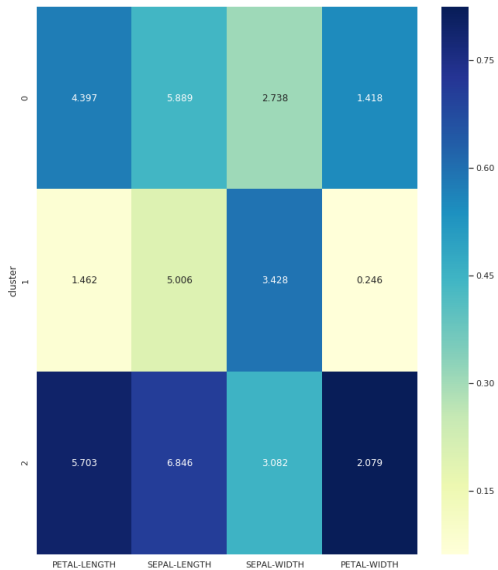
```
cluster 0: 61 (40.67%)
```

```
cluster 1: 50 (33.33%)
```

```
cluster 2: 39 (26.00%)
```





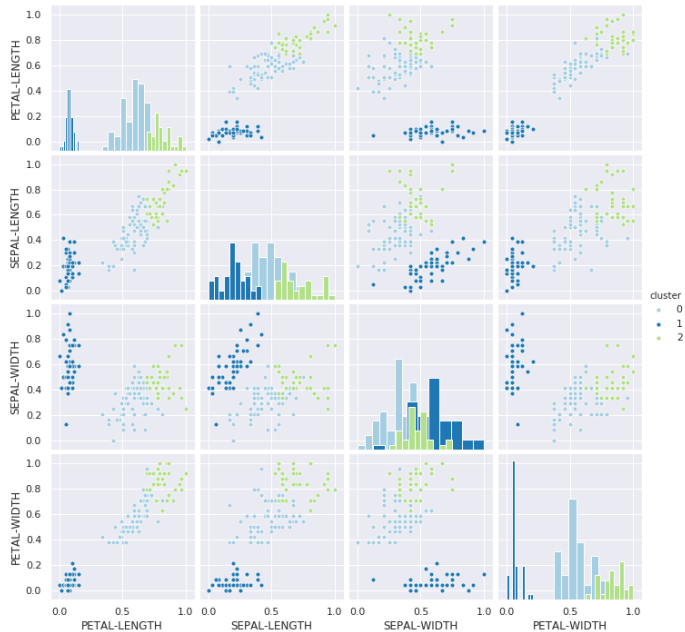


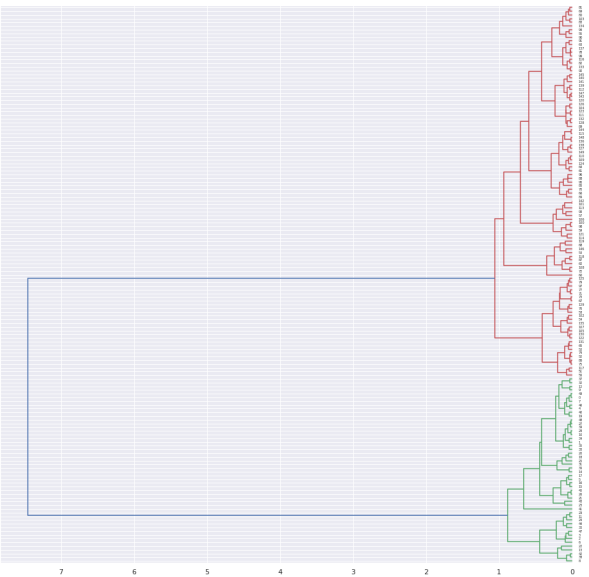
## Agrupamiento Jerárquico

Para el algoritmo **Agglomerative Clustering**, se ha utilizado el siguiente código en python y se ha obtenido la siguiente agrupación de las muestras:

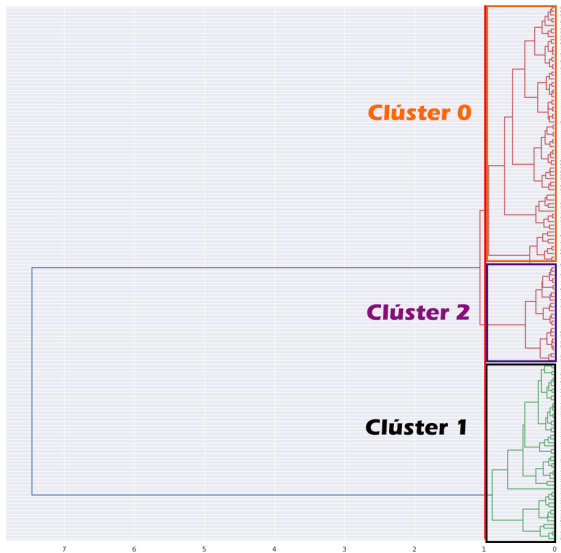
```
AgglomerativeClustering(n_clusters=3,  
                        linkage="ward", affinity='euclidean')
```

```
cluster 0: 67 (44.67%)  
cluster 1: 50 (33.33%)  
cluster 2: 33 (22.00%)
```





0: 67 (44.67 %), 1: 50 (33.33 %), 2: 33 (22.00 %)

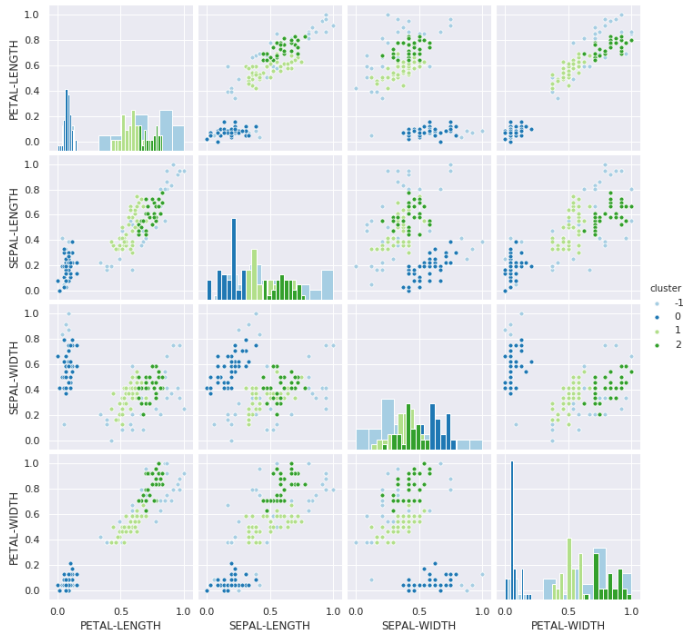


## DBSCAN

Para **DBSCAN**, se ha utilizado el siguiente código en python y se ha obtenido la siguiente agrupación de las muestras, donde el clúster -1 representa las muestras formadas por ruido:

```
DBSCAN(eps=0.12, min_samples=5)
```

```
cluster 0: 45 (30.00%)  
cluster 1: 39 (26.00%)  
ruido -1: 36 (24.00%)  
cluster 2: 30 (20.00%)
```





## Mean Shift

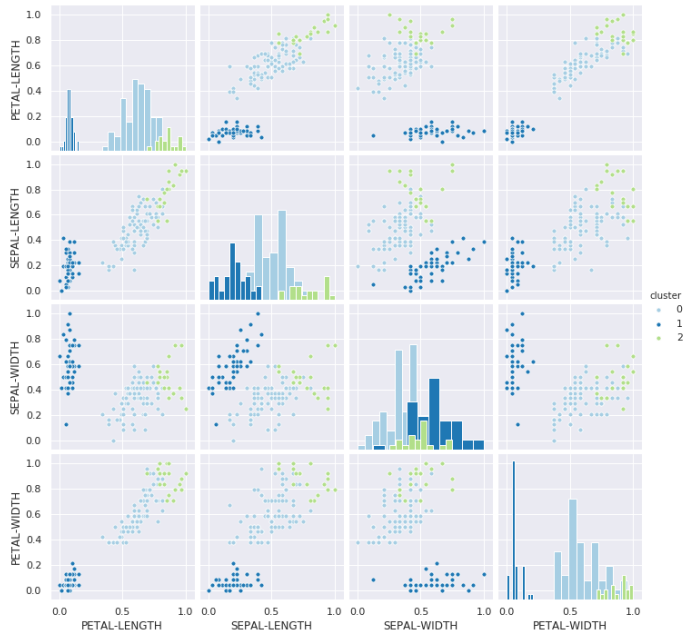
Para **Mean Shift**, se ha utilizado el siguiente código en python y se ha obtenido la siguiente agrupación de las muestras:

```
MeanShift(bandwidth=estimate_bandwidth(X_normal,  
                                         quantile=0.67, n_samples=400))
```





```
cluster 0: 81 (54.00%)
```

```
cluster 1: 50 (33.33%)
```

```
cluster 2: 19 (12.67%)
```



## Referencias I

-  Chire, *Cluster analysis with optics on a density-based data set.*, <https://commons.wikimedia.org/wiki/File:OPTICS-Gaussian-data.svg>, October 2011.
-  *Using the elbow method to determine the optimal number of clusters for k-means clustering*, <https://bl.ocks.org/rpgove/0060ff3b656618e9136b>, note = "Último acceso: 28/12/2019".
-  *K-means cluster analysis*, [https://uc-r.github.io/kmeans\\_clustering](https://uc-r.github.io/kmeans_clustering), note = "Último acceso: 28/12/2019".
-  *Understanding data mining clustering methods.*

## Referencias II



*Selecting the number of clusters with silhouette analysis on kmeans clustering,*

[https://scikit-learn.org/stable/auto\\_examples/cluster/plot\\_kmeans\\_silhouette\\_analysis.html](https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html), Último acceso: 28/12/2019.