Práctica 3 : Clustering

Alumno:Pablo Alonso

1)

a-

Se intenta encontrar la especie y el género de los cangrejos :

* Resultados con 2-means sin aplicar transformación logarítmica:

- Especie : Se logra clusterizar ambas especies correctamente con frecuencia relativa de 0.28
- Género : No se logra clusterizar los géneros.

* Resultados con 2-means aplicando transformación logarítmica:

- Especie : Se logra clusterizar ambas especies correctamente con frecuencia relativa de 0.35
- Género : No se logra clusterizar los géneros.

Observando los plots, se encuentra que claramente la especie es mucho más separable como bolas apretadas que el género.

Las únicas transformaciones de los datos que permiten que 2-means encuentre la especie como clusters es hacer PCA y luego un escalado de los datos. Sin embargo, se encuentra que hay una probabilidad más alta de encontrar las clases si primero se hace la transformación logarítmica . Cuando se aplican PCA y el escalado de los datos se giran los ejes en las direcciones que maximizan la varianza y luego se centran con respecto a los ejes girados (evita los problemas de distintas escalas al calcular distancias). La transformación logarítmica compacta los datos y y reduce el riesgo de caer en un mínimo local.

A continuación, se presenta las matrices de confusión para los métodos jerárquicos:

Single linkage:

Especie	1	2
B	99	1
O	100	0

Género	1	2
F	99	1
M	100	0

Average linkage:

Especie	1	2
B	99	1
O	100	0

Género	1	2
F	99	1
M	100	0

Complete linkage:

Especie	1	2
B	97	3
O	90	10

Género	1	2
F	87	13
M	100	0

Los métodos jerárquicos no logran buenos resultados encontrando los clusters que representen las clases. Single y average linkage separan outliers del resto de los datos y complete linkage encuentre conjuntos de puntos que son compactos pero que estan separados entre si. Ninguno de los métodos jerárquicos refleja la estructura subyacente de las clases.

b-

No puede hacerce log de los datos en este caso porque hay valores que son negativos.Para reducir el número de dimensiones se hace una PCA. Se observa que se explica el 99% de la varianza tomando las primeras 5 componentes principales, por lo tanto se hace clustering usando solo esas 5 variables.

No se logra hacer una perfecta separación de los datos para ninguna de las 2 clases con ningún método. Las transformaciones que se hacen son PCA y luego un escalado de los datos. Se observan mejores resultados sin escalar los datos. A continuación los resultados obtenidos de los clusterings usando solo las primeras 5 componentes principales :

2-means:

Año	1	2
2006	2	17
2007	28	2

Especie	1	2
2	10	12
10	9	18

Single linkage:

Año	1	2
2006	18	1
2007	29	1

Especie	1	2
2	21	1
10	26	1

Average linkage:

Año	1	2
2006	18	1
2007	29	1

Especie	1	2
2	21	1
10	26	1

Complete linkage:

Año	1	2
2006	18	1
2007	24	6

Especie	1	2
2	19	3
10	23	4

De los resultados para el problema lampone se interpreta que para las transformaciones hechas, la clase año es casi perfectamente separable en 2 bolas apretadas, pero la especie los datos resultan poco separables. Debido a la presencia de outliers en el problema, los métodos jerárquicos crean 2 clusters que son la separación de algunos outliers del resto de los puntos. Estos clusters no representan para nada las estructuras subyacentes que forman las clases.

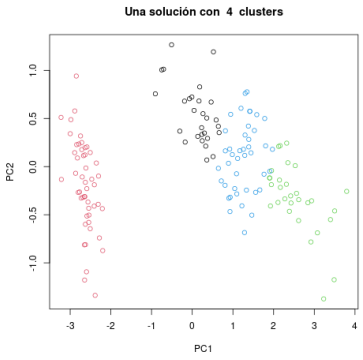
2)

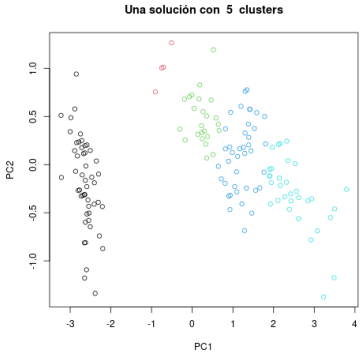
Ver código en R.

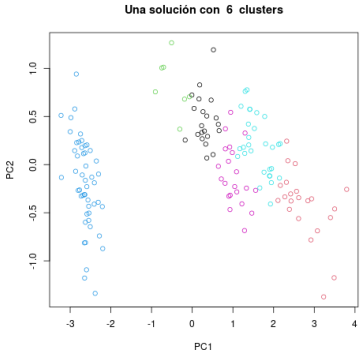
3)

A continuación, gráficos con las clusterizaciones determinadas por GAP:

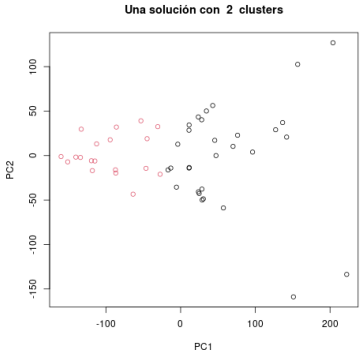
-Iris devuelve como resultado 4,5 y 6:

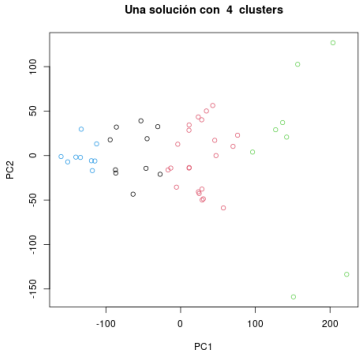




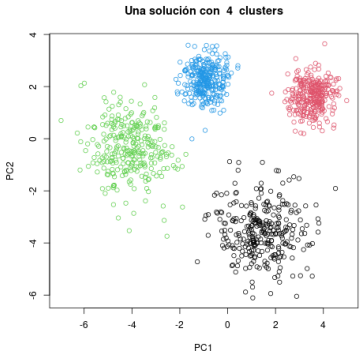


- Lampone devuelve como resultado 2 y 4:

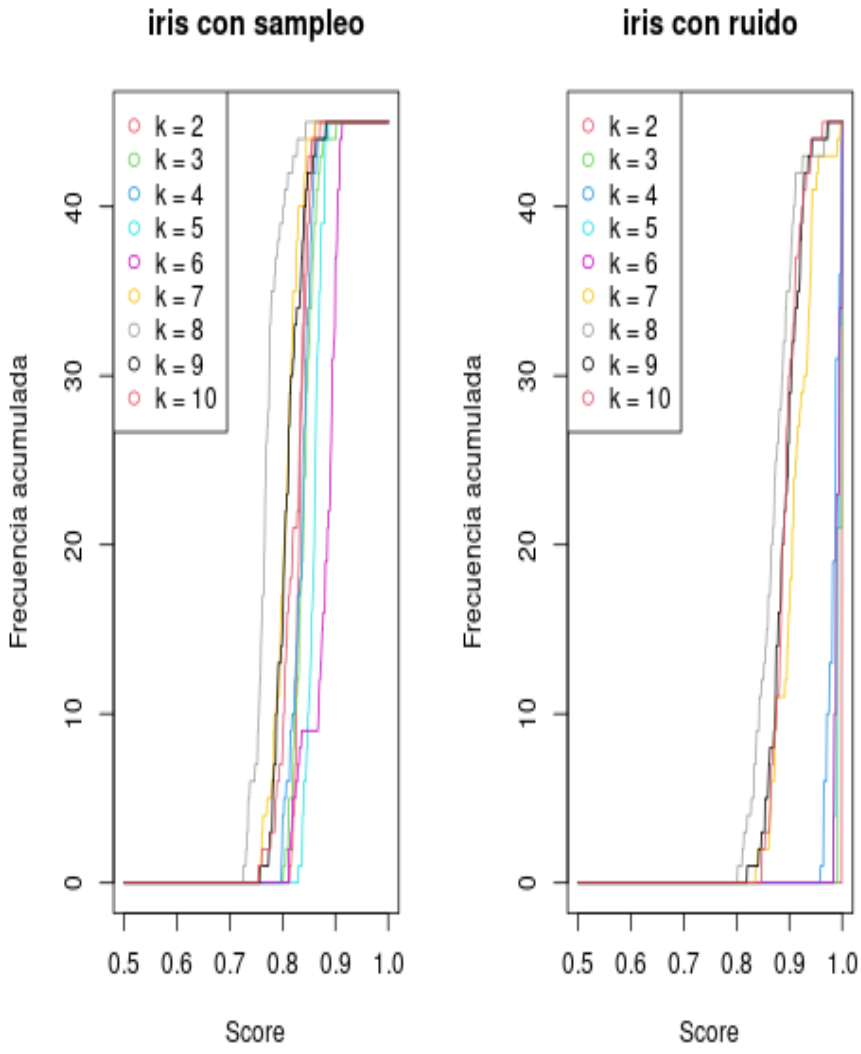




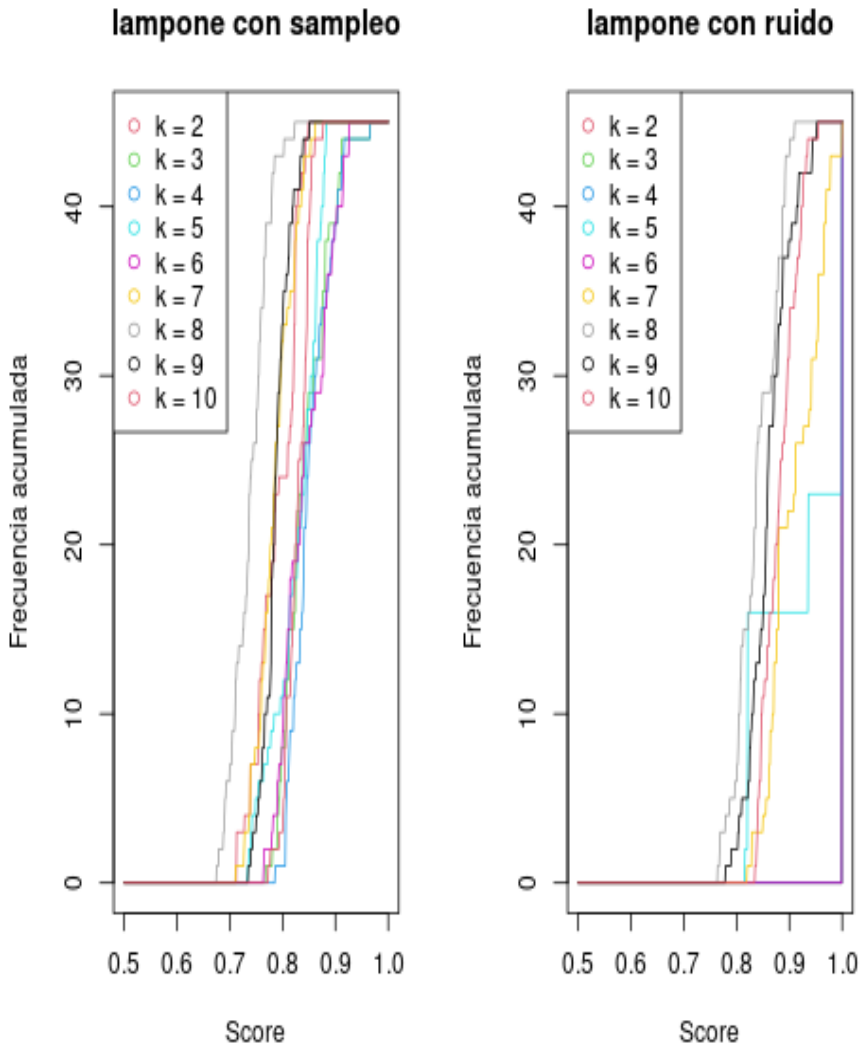
- Para gausianas devuelve que la única solución es 4:



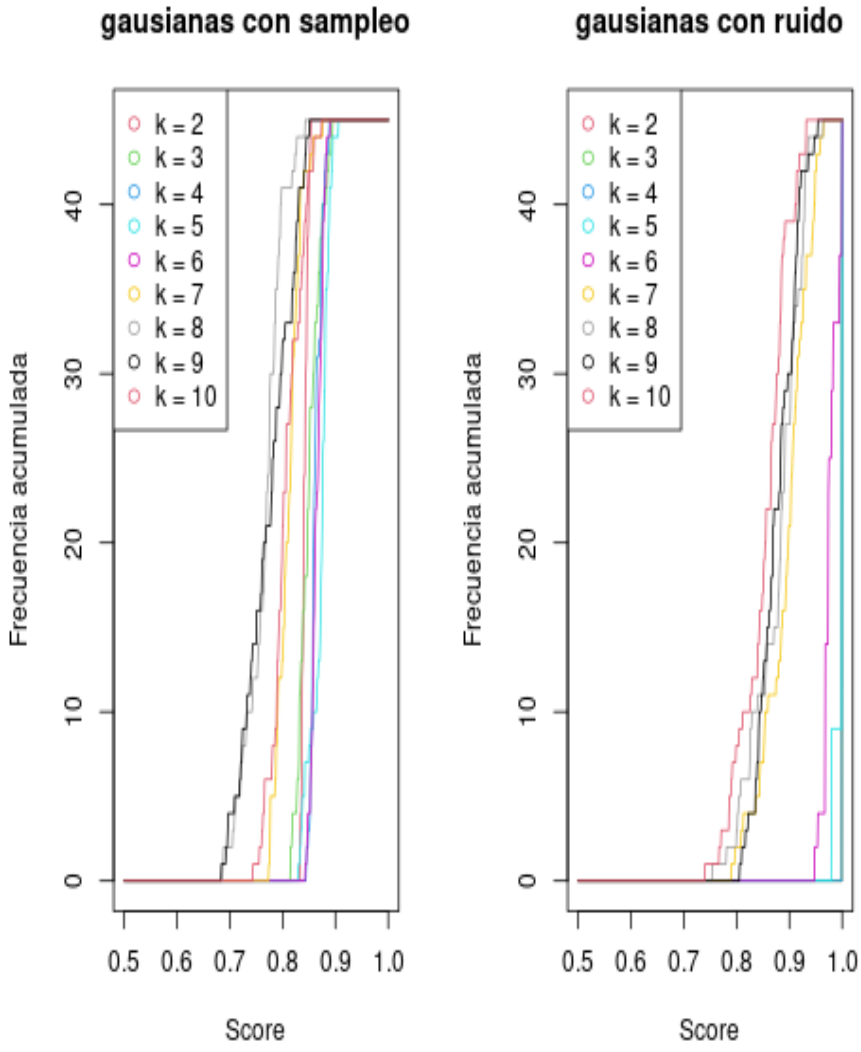
Para estabilidad se usaron las técnicas de sampleo y agregar ruido. A continuación los gráficos obtenidos:



Se observa una diferencia importante entre las soluciones de los datasets sampleados. En los datasets con ruido se logran descartar los mayores valores de k , la distribución para 3 está bastante pegada a 1 y parece ser la mejor solución .



Las soluciones con sampleo difieren bastante entre sí. Para esta técnica la solución más estable parece ser 4 y para los datasets con ruido lo más estable parece ser 6.

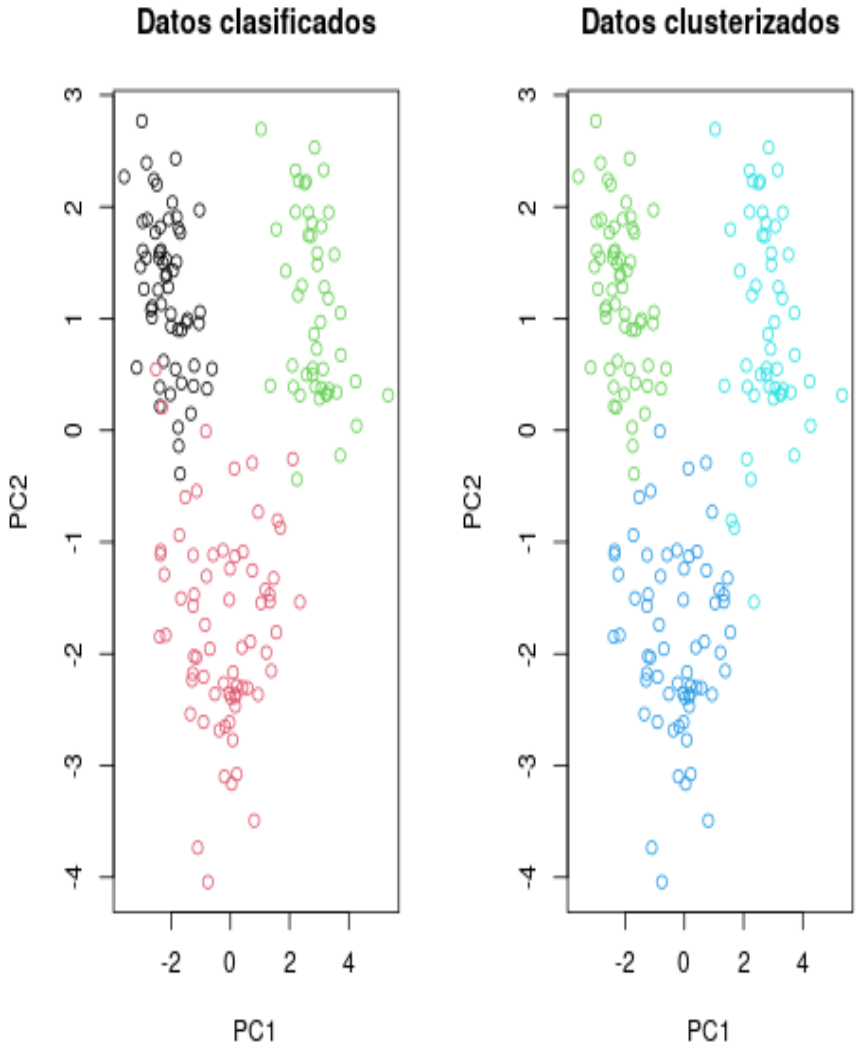


Aca también ocurre que las soluciones con sampleo difieren mucho. Para esta técnica la mejor solución parece estar entre 4,5 y 6. Para las soluciones con ruido el mejor k es 4.

4)

El dataset se saco de https://archive.ics.uci.edu/ml/datasets/wine. Hay 3 viñedos y la idea es encontrar a que viñedo pertenece cada vino. Para empezar se hace un análisis de los datos. Dado que los valores son todos positivos se hace

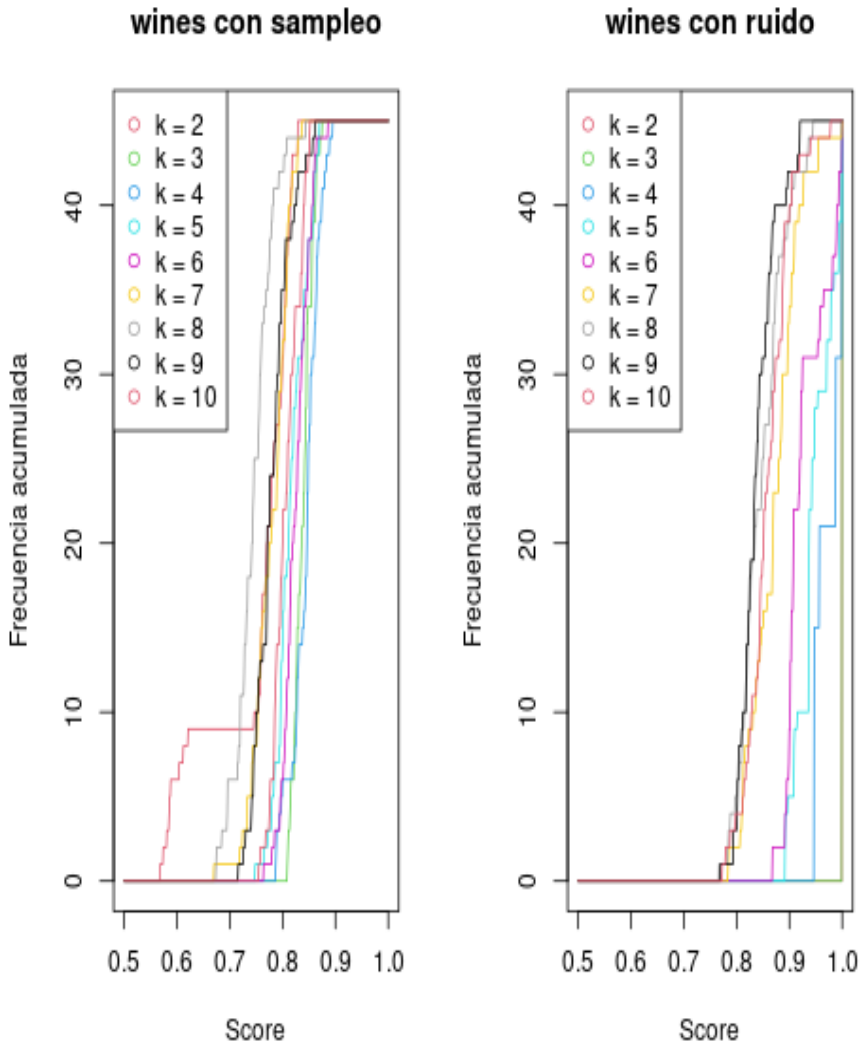
una transformación logarítmica para acercar los outliers, luego un escalado para llevar los datos a mismas escalas y PCA para eliminar el ruido. A continuación se muestran los datos con sus respectivas clasificaciones y el resultado de 3-means:



El acierto de la clusterización con respecto a las clases es de 95.48 % y la matriz de confusión es la siguiente:

	Clusters		
Viñedo	1	2	3
1	58	0	0
2	4	63	4
3	0	0	48

Se aplica estabilidad para encontrar el k más estable:



Para sampleo, 3 tiene menor varianza mientras que 4 toma valores más cercanos a 1. Para ruido, los valores más estables son 2 y 3 pero nos quedamos

con 3.