# Trabajo Práctico 3
# Clustering

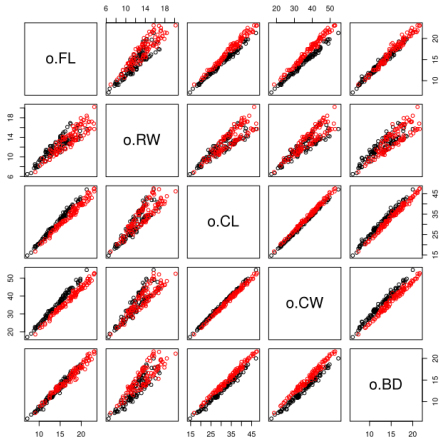
AGUSTÍN MISTA
*Universidad Nacional de Rosario*
*Tópicos de Minería de Datos*
Rosario, 21 de Noviembre de 2017

## Ejercicio 1: Ejemplos prácticos de clustering.

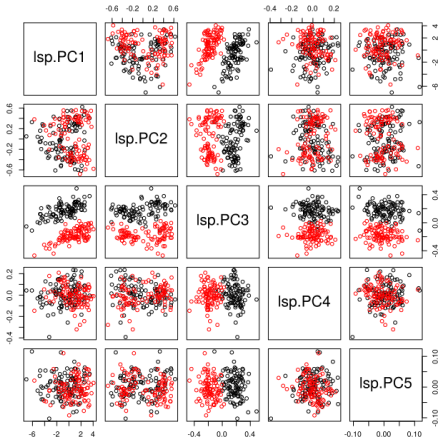
En este ejercicio analizamos la performance de dos métodos de clustering provistos por el framework R frente dos datasets del mundo real. Nos interesa encontrar clusterings de los datos que nos permitan predecir algunas de las clases de nuestros datasets. Para esto utilizamos tanto técnicas de clustering por particionado (K-means) como por división jerárquica (HClust).

**Crabs:** Este dataset agrupa distintas mediciones realizadas sobre cangrejos de la especie *Leptograpsus variegatus*. El mismo posee 200 muestras, cada una con 5 features continuas relacionadas a la morfología de los cangrejos, junto a dos features categóricas que agrupan sexo y variedad de los mismos.

A continuación se muestran dos versiones del dataset usadas para nuestro análisis (el resto resulta sumamente similar a alguna de estas dos), una sin modificaciones de ningún tipo y otra a la que se aplicó una transformación logarítmica (*log()*), centrado y escalado (*scale()*) y Principal Component Analysis (*prcomp()*). Los mismos fueron coloreados usando la feature variedad de la especie en ambos casos.

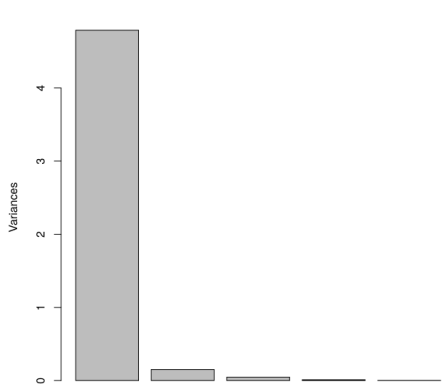



Original

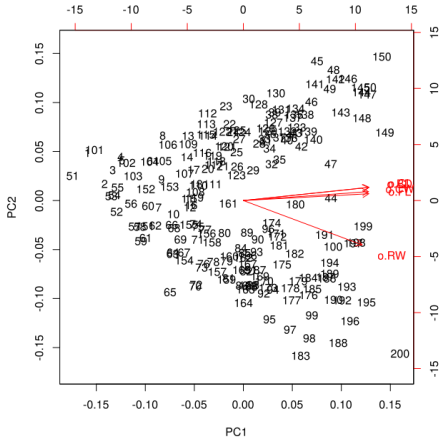



log + scale + prcomp

Además se muestra el histograma de relevancia de features arrojado por la PCA de nuestros datos, junto con el biplot de los mismos.




PCA Histogram




PCA Biplot

Ahora bien, nos interesa conocer si es posible "predecir" alguna de estas características categóricas mediante clustering de los datos continuos. Para esto sometimos nuestros datos a los métodos de clustering K-means y HClust, variando el criterio de similitud entre Single (S), Complete (C) y Average (A) para el caso del método basado en clustering jerárquico. Luego comparamos cuánto las distintas clusterizaciones obtenidas se asemejaban a alguna de las features categóricas presentes en nuestro dataset, calculando el porcentaje de resultados bien agrupados en cada caso. Los resultados se muestran a continuación.

| Feature | Method | Original | log | log+scale | log+scale+prcomp |
|---|---|---|---|---|---|
| Sex | K-means | 0.51 | 0.52 | 0.52 | 0.52 |
| | HClust/S | 0.51 | 0.51 | 0.51 | 0.51 |
| | HClust/C | 0.51 | 0.52 | 0.52 | 0.52 |
| | HClust/A | 0.56 | 0.51 | 0.51 | 0.51 |
| Species | K-means | 0.59 | 0.60 | 0.60 | 0.60 |
| | HClust/S | 0.51 | 0.51 | 0.51 | 0.51 |
| | HClust/C | 0.59 | 0.60 | 0.60 | 0.60 |
| | HClust/A | 0.61 | 0.57 | 0.51 | 0.51 |

Como se puede observar, ninguna de las combinaciones entre métodos y datos logra una clusterización aceptable respecto del sexo o la variedad de los cangrejos.

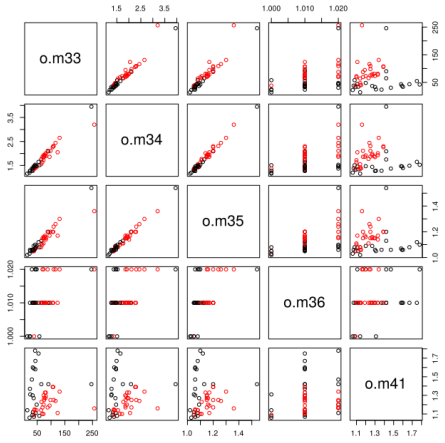
Para el caso de la feature sexo del cangrejo, los resultados son completamente aleatorios. Es decir, la clusterización no posee ninguna correlación con esta feature de nuestro dataset.

Por otro lado, para el caso de la variedad del cangrejo, algunos métodos obtienen resultados levemente mejores (sin llegar a ser satisfactorios), lo que nos haría creer que nuestras features continuas podrían tener alguna leve correlación con la variedad de los cangrejos. En particular, para el método de HClust con Single linkage, la clusterización también resulta mayormente aleatoria, lo que podría sugerir en principio que nuestro dataset posee datos dispersos con bastante solapamiento respecto de esta feature en particular.

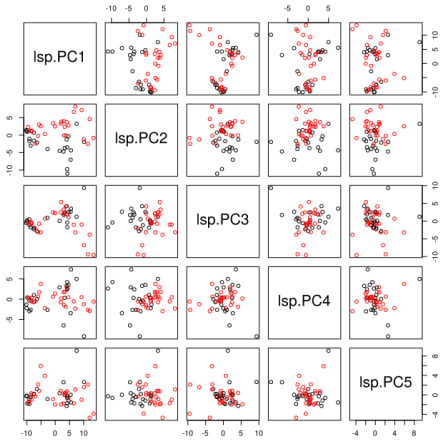
Finalmente, es interesante remarcar que este dataset no obtuvo mejoras significativas en los resultados de clustering en los casos en los que se le realizó algún tipo de preprocesamiento. Peor aún, en algunos casos el preprocesamiento tuvo efectos nocivos sobre el clustering de nuestros datos.

**Lampone:** Este dataset recoge información relacionada a la producción de arándanos durante dos temporadas distintas. Luego de eliminar columnas nulas o constantes, el mismo posee 49 muestras, cada una con 128 features, donde dos de ellas corresponden al año de medición y a la especie de arándano. El resto de las features se corresponden a distintas mediciones realizadas sobre los arándanos, y que utilizamos con el fin de recuperar información acerca de las dos primeras.

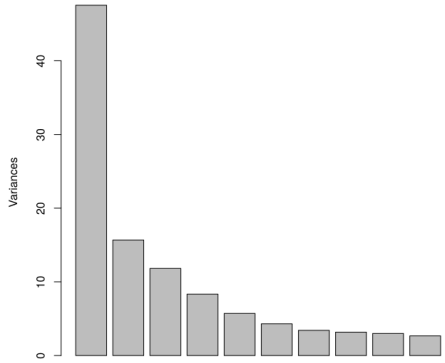
A continuación se muestran las primeras cinco features de nuestro dataset original, junto con las cinco features principales de nuestro dataset preprocesado al igual que antes mediante *log* + *scale* + *prcomp*. Como en el caso anterior, también mostramos el histograma de relevancia de features de la PCA de nuestros datos, junto con el biplot del mismo.



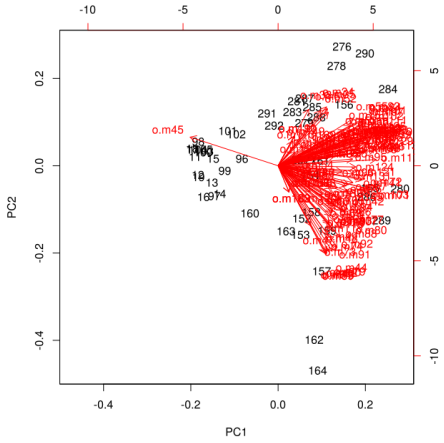

Original




log + scale + prcomp




PCA Histogram




PCA Biplot

Se realizó nuevamente el análisis del porcentaje de casos bien agrupados para los distintos métodos vistos, y con las distintas variantes de nuestro dataset, comparándolas tanto con la especie del arándano como con el año de medición de las muestras. Los resultados se muestran a continuación.

| Feature | Method | Original | log | log+scale | log+scale+prcomp |
|---|---|---|---|---|---|
| Species | K-means | 0.57 | 0.53 | 0.55 | 0.55 |
| | HClust/S | 0.55 | 0.57 | 0.57 | 0.57 |
| | HClust/C | 0.53 | 0.53 | 0.69 | 0.69 |
| | HClust/A | 0.55 | 0.57 | 0.57 | 0.57 |
| Year | K-means | 0.92 | 1.00 | 0.98 | 0.98 |
| | HClust/S | 0.61 | 0.59 | 0.59 | 0.59 |
| | HClust/C | 0.51 | 0.86 | 0.84 | 0.84 |
| | HClust/A | 0.61 | 0.96 | 0.59 | 0.59 |

En este caso, podemos ver que nuevamente se obtienen resultados pobres cuando se comparan los distintos clusterings respecto de la especie de los arándanos. Sin embargo, se obtienen muy buenos resultados cuando se comparan los clusters obtenidos respecto del año de medición de las muestras, en especial para el métodos K-means, el cual logra colocar todas las muestras tomadas en el mismo año en un cluster diferente cuando se usa el preprocesamiento indicado.

Para el caso de las clusterizaciones obtenidas mediante HClust, el resultado final es muy dependiente tanto del preprocesamiento de los datos como del método de disimilaridad elegido. Los resultados obtenidos usando Single linkage resultan pobres sin importar qué variante del dataset sea us-

ada. Esto puede deberse a que Single linkage funciona combinando clusters usando la mínima distancia entre dos puntos en cada caso, pudiendo producirse lo que se conoce como *encadenamiento*, fomentando clusters "largos" y con una gran distancia entre sus puntos extremos. Si además consideramos que nuestro dataset tiene una gran cantidad de features, podríamos sospechar que muchas de ellas no correlacionadas con el año de medición pueden tener un efecto negativo cuando se calcula la mínima distancia entre dos puntos de dos clusters distintos al momento de combinarlos.

Si en cambio observamos los resultados obtenidos mediante HClust con Complete linkage podemos observar como se obtienen resultados mucho mejores que en la variante anterior, debido posiblemente a que nuestros datos son bastante compactos respecto del año de medición.

## Ejercicio 2: Implementación de algoritmos.

Ambos algoritmos se encuentran en el archivo `ej2.R`. Los mismos se encuentran parametrizados por:

- Dataset de entrada (data matrix o data frame).
- Método de clustering ("kmeans","hclust.s","hclust.c", "hclust.a").
- Máximo número de clusters a considerar (K).
- Cantidad de datasets de referencia (B).
- Porcentaje de datos en cada dataset de referencia (sólo `stability`).

**Ejemplo de uso:**

```
> gapStatistic(iris[,1:4], method="kmeans", K=10, B=25)
[1] 4

> scores <- stability(iris[,1:4], method="hclust.a",
                      K=10, B=25, ratio=0.9)
> plotScores(scores) # muestra scores acumulados para cada K
```

## Ejercicio 3: Selección del número óptimo de clusters.

En este ejercicio pusimos a prueba nuestros algoritmos previamente presentados ante los datasets *Cuatro gaussianas*, *Iris* y *Lampone*.

### Gap Statistic

A continuación se muestra el valor medio de 25 ejecuciones del algoritmo de Gap Statistic para cada dataset (más algunas versiones preprocesadas) y

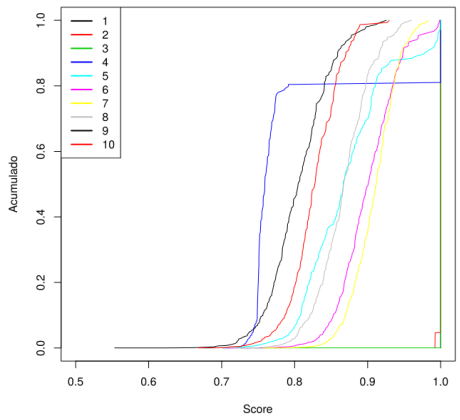
cada método de clustering. Los resultados de cada corrida fueron obtenidos usando `K=10` y `B=10`.

| Dataset | K-means | HClust/S | HClust/C | HClust/A |
|---|---|---|---|---|
| Cuatro gaussianas | 4 | 3 | 4 | 2 |
| Iris | 5 | 3 | 3 | 3 |
| Iris + log + scale | 2 | 2 | 3 | 2 |
| Lampone | 2 | 2 | 3 | 2 |
| Lampone + log + scale | 3 | 1 | 2 | 1 |

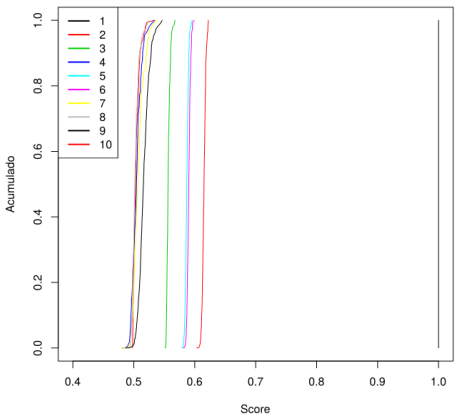
A partir de la tabla anterior podemos extraer algunos datos interesantes. En principio, y como era de esperarse, sólo los métodos K-means y HClust con Complete linkage lograron obtener el número óptimo de clusters para el dataset de las cuatro gaussianas, puesto que ambos benefician el agrupamiento en clusters compactos. En segundo lugar, K-means sobreestimó o subestimó (dependiendo del preprocesamiento) la cantidad ideal de clusters para el dataset Iris, mientras que sin preprocesamiento, todas las variantes de HClust arrojan el resultado óptimo. Por otro lado, HClust con Single linkage tiende a subestimar la cantidad óptima de clusters en cada dataset respecto de su homónimo usando Complete linkage, ya que tiende a buscar alta conectividad en los clusters resultantes. También es interesante notar que los resultados parecen ser sumamente dependientes del preprocesamiento que se efectúe sobre los datasets, lo cual se nota especialmente en el dataset Lampone. Finalmente, podría ser adecuado concluir que la elección del nivel de preprocesamiento y el método de clustering no es trivial, y puede ser facilitada conociendo algunas propiedades del dataset en cuestión.

### Stability

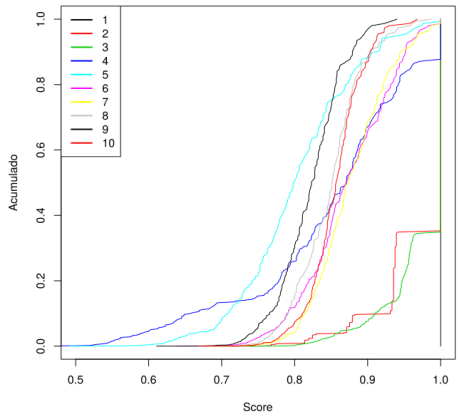
A continuación se muestran los gráficos de scores acumulados calculados mediante el algoritmo de *stability* y los métodos K-means y HClust/A, usando `K=10`, `B=25` y `ratio=0.8` sobre los mismos datasets que en el caso anterior.



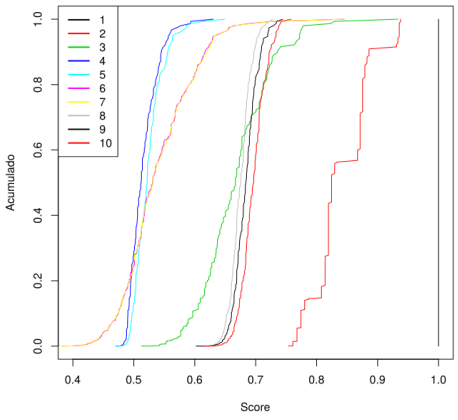

Cuatro gaussianas @ K-means



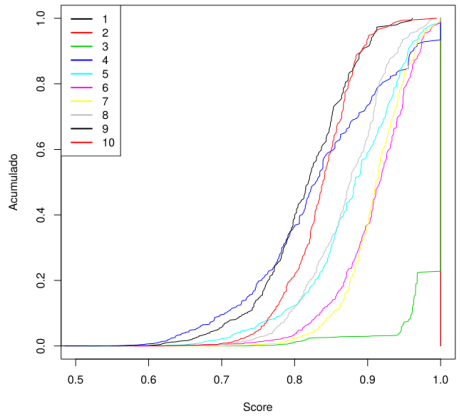

Cuatro gaussianas @ HClust/A



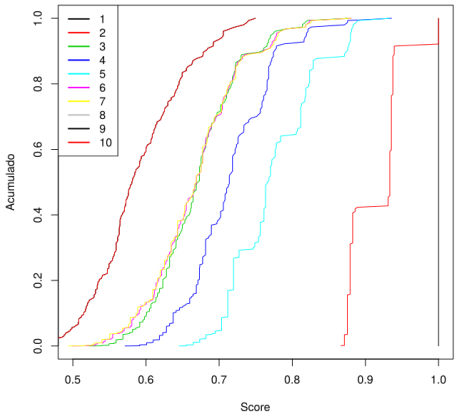

Iris + log + scale @ K-means




Iris + log + scale @ HClust/A




Lampone + log + scale @ K-means




Lampone + log + scale @ HClust/A

Se puede observar como los casos en los que se utilizó HClust/A como método de clustering no obtuvieron resultados particularmente interesantes, puesto que solo se logró considerar como estable aquellos casos donde todos los datos pertenecen al mismo y único cluster—para el caso de los datasets Iris y Lampone quizá podría considerarse estable `K=2`, aunque esto queda a criterio del data scientist.

Para el caso de K-means, los datos obtenidos son en general más útiles. En primer lugar, en el dataset de las cuatro gaussianas se observa claramente que los valores estables de `K` son 1, 2 y 3, por lo que elegiríamos a priori 3 clusters como número óptimo—nótese que el pico alrededor del score 0.75 para `K=4` no nos da demasiadas garantías de ser un valor estable si no conociéramos la distribución real de este dataset. Finalmente, para los datasets Iris y Lampone con preprocesamiento logarítmico y escalado, se observa de manera bastante clara que los valores estables de `K` también son 1, 2 y 3, por lo que en un primer intento también elegiríamos a 3 como nuestro número óptimo de clusters en ambos datasets.