


Laboratorio 8: Clasificación por Particiones

Fecha de entrega: ver Mediación Virtual

Entrega: individual o en parejas

Instrucciones: Resuelva los ejercicios propuestos en el Laboratorio. Guarde el script respectivo, con su nombre y súbalo en el sitio del curso. Además, guarde los resultados y súbalos también en el sitio.

En este laboratorio trabajaremos la clasificación por particiones en  usando algunos métodos clásicos, como el de k-medias y algunas de sus variantes.

- Uso del **método de k-means**:

```
clasif.KM <- kmeans(X, 3, iter.max = 25, nstart = 1, algorithm = "Forgy")
```

– Parámetros importantes:

- * El segundo parámetro (3) especifica el número de clases (en este caso los centros iniciales se escogen al azar) aunque podría especificarse los centros iniciales.
- * `iter.max` indica el número máximo de iteraciones.
- * `nstart` indica cuantas inicializaciones al azar se harán.
- * `algorithm` especifica el tipo de k-means; las opciones son: Forgy, Lloyd, MacQueen, Hartigan-Wong (por defecto).

– Las principales salidas del método son:

```
clasif.KM <- kmeans(X, 3, iter.max = 25, nstart = 1, algorithm = "Forgy")  
  
clasif.KM$cluster  
clasif.KM$centers  
clasif.KM$totss  
clasif.KM$withinss  
clasif.KM$tot.withinss  
clasif.KM$betweenss  
clasif.KM$size  
clasif.KM$iter
```

– Para hacer varias corridas:

```
clasif.KM <- kmeans(X, 3, iter.max = 25, nstart = 10, algorithm = "Forgy")
```

– Hay muchas ayudas a la visualización para los resultados de una partición, por ejemplo:

```
plot(X, col = clasif.KMmult$cluster)
```

1. Considere las siguientes tablas de datos:

- (a) Notas escolares (10 estudiantes \times 5 variables).
- (b) Peces de Amiand.
- (c) El consumo de proteínas en Europa.
- (d) Iris de Fisher.

Para cada una de ellas:

- (a) Aplique el método de k-medias 100 veces, con puntos iniciales al azar, usando 2, 3 y 4 clases en cada caso.

- (b) Grafique los óptimos locales obtenidos. Para ello puede usar un histograma, una caja de variación tipo boxplot o cualquier otra herramienta que ayude a visualizar la variabilidad de resultados que se obtienen.
 - (c) Calcule el óptimo promedio y la tasa de atracción del mejor óptimo (porcentaje de veces que ese óptimo se obtiene en su experimento).
 - (d) Reporte el tiempo empleado en hacer los cálculos.
 - (e) Para la mejor solución en cada ejemplo, reporte los resultados relevantes y use alguna ayuda para la visualización de los resultados.
2. Considere la tabla de los Iris de Fisher. Aplique las variantes Forgy, Lloyd, MacQueen y Hartigan-Wong, cada una 25 veces. Escoja la mejor solución en cada variante y compare sus resultados.
3. Experimentar con los métodos de P. Roosseuw y colaboradores (el “grupo L1”):
- Cargar `library(cluster)`
 - `pam`: “Partitioning Around Medoids”, particionamiento tipo k-means pero usando como núcleo de las clases los vectores de medianas en lugar de los vectores de medias.

Considere la tabla de los Iris de Fisher. Aplica `pam` 100 veces, para 2,3,4 clases, con inicializaciones al azar. Reporte el comportamiento de los óptimos locales y las características de mejor solución.