Profesor: Dr. Oldemar Rodríguez Rojas
Análisis de Datos 1
Fecha de Entrega: Viernes 24 de mayo a las 8 a.m.
Nota: Todas la preguntas tienen el mismo valor

Tarea Número 8

1. En este ejercicio usaremos la tabla de datos `EjemploAlgoritmosRecomendación.csv`, la cual contiene los promedios de evaluación de 100 personas que adquirieron los mismos productos o muy similares en la tienda AMAZON. La idea consiste en recomendar a un cliente los productos que ha comprado otra persona que pertenece al mismo clúster.

   a) Ejecute un Clustering Jerárquico con la distancia euclídea y la agregación del Salto Máximo, Salto Mínimo, Promedio y Ward. Guarde la tabla de datos en el archivo `AlgoritmosRecomendación2.csv` con el clúster al que pertenece cada individuo para el caso de la agregación de Ward usando 2 clústeres.

   b) "Corte" el árbol anterior usando 2 clústeres y la agregación de Ward, interprete los resultados usando interpretación usando gráficos de barras (Horizontal-Vertical) y usando gráficos tipo Radar.

   c) Si se tienen 4 clústeres usando agregación de Ward ¿Qué productos recomendaría a Teresa, a Leo y a Justin?, es decir, ¿los productos que compra cuál otro cliente? Usando distancia euclídea ¿cuál es la mejor recomendación de compra que le podemos hacer a Teresa, a Leo y a Justin?

   d) Construya un clustering jerárquico sobre las componentes principales del ACP.

2. La tabla de datos `VotosCongresoUS.csv` la cual contiene 16 votos (y=Sí, n=No, NS=No votó) dados por los congresistas de Estados Unidos respecto a 16 temáticas diferentes, además en la primera columna aparece el partido al que pertenecen (Republicano o Demócrata).

   a) Ejecute una clasificación jerárquica sobre esta tabla de datos usando la función `daisy` ya que los datos son cualitativos. Use métrica `euclidean` y método `complete` (deje el resultado en la variable `jer`). Cargue los datos con la siguiente instrucción:

```
Datos <- read.csv("VotosCongresoUS.csv",header=TRUE, sep=",", dec=".")
```

   b) Luego "corte" el árbol usando 3 clústeres y ejecute el siguiente código:

```
grupo<-cutree(jer, k = 3)
NDatos<-cbind(Datos,grupo)

cluster<-NDatos$grupo
sel.cluster1<-match(cluster,c(1),0)
Datos.Cluster1<-NDatos[sel.cluster1>0,]
dim(Datos.Cluster1)

sel.cluster2<-match(cluster,c(2),0)
Datos.Cluster2<-NDatos[sel.cluster2>0,]
```

```
dim(Datos.Cluster2)

sel.cluster3<-match(cluster,c(3),0)
Datos.Cluster3<-NDatos[sel.cluster3>0,]
dim(Datos.Cluster3)
```

Explique qué hace el código anterior. Luego ejecute el siguiente código:

```
plot(Datos$Party,col=c(4,6),las=2,main="Party",xlab="Todos los Datos")
plot(Datos.Cluster1$Party,col=c(4,6),las=2,main="Party",xlab="Cluster-1")
plot(Datos.Cluster2$Party,col=c(4,6),las=2,main="Party",xlab="Cluster-2")
plot(Datos.Cluster3$Party,col=c(4,6),las=2,main="party",xlab="Cluster-3")
```

Con ayuda de los gráficos anteriores y tomando en cuenta el tamaño de cada cluster interprete los 3 clústeres formados.

3. Realice un análisis similar al del ejercicio anterior con la tabla de datos `CompraBicicletas.csv`.

4. Dada la siguiente matriz de disimilitudes entre cuatro individuos $A1$, $A2$, $A3$ y $A4$, construya "a mano" una Jerarquía Binaria usando la agregación del Salto Máximo y del Promedio, dibuje el dendograma en ambos casos:

$$D = \begin{pmatrix} 0 & & & \\ 5 & 0 & & \\ 2 & 1 & 0 & \\ 3 & 7 & 6 & 0 \end{pmatrix}$$

Verifique los resultados con `hclust`.

5. Dada la siguiente matriz de similitudes entre cuatro individuos $A1$, $A2$, $A3$, $A4$ y $A5$, construya "a mano" una Jerarquía Binaria usando la agregación del del Salto Mínimo y del Salto Máximo, dibuje el dendograma en ambos casos:

$$S = \begin{pmatrix} 1 & & & & \\ 0.10 & 1 & & & \\ 0.42 & 0.63 & 1 & & \\ 0.54 & 0.46 & 0.41 & 1 & \\ 0.35 & 0.98 & 0.85 & 0.73 & 1 \end{pmatrix}$$

Verifique los resultados con `hclust`.

6. *a*) Demuestre que la distancia de *Chebychev* definida como:

$$d(i,j) = \max|x_{ik} - x_{jk}| \text{ para } k = 1, 2, \ldots, p$$

efectivamente es una distancia.

b) Programe una función en **R** que calcula la distancia de *Chebychev* entre dos vectores.

c) Programe una función en **R** que recibe un `DataFrame` calcula la matriz de distancias usando la distancia de *Chebychev* entre dos vectores calcula anteriormente.

d) Para la tabla de datos `EjemploAlgoritmosRecomendación.csv` ejecute un Clustering Jerárquico de *Chebychev* y la agregación Ward. Compare el resultado con el obtenido en el ejercicio 1 usando distancia *euclídea*.

7. Programe funciones en **R** que reciben un `DataFrame` con datos binarios y calculan la matriz de distancias usando las fórmulas de distancia de *Jaccard* y de *Russel-Rao* respectivamente. Verifique los resultados con al menos dos ejemplos.

8. Programe funciones en **R** para la función *Lance & Williams* para la agregación del promedio y de Ward (que es un caso particular de la fórmula general de *Jambu*). Construya un pequeño ejemplo de prueba, es decir, dada una matriz de distancias (agregaciones) encuentre el mínimo, luego en esta matriz, usando las funciones antes programadas, reconstruya la matriz para el siguiente paso del algoritmo (elimine dos filas y dos columnas para luego colocar en su lugar una fila y una columna nuevas, no en general, es hacer un caso particular).

9. [**15 puntos**] **Optativo:** Repita el ejercicio anterior con la fórmula general recursiva de *Jambu*.

Entregables: Incluya en documento autoreproducible (`HTML`) todas las instrucciones y códigos **R** utilizados en cada ejercicio, incluya los resultados de los cálculos, los gráficos generados y las respuestas a las preguntas.

oldemar rodríguez

CONSULTOR en M1NER14 DE D4T0S