Profesor: Dr. Oldemar Rodríguez Rojas
Análisis de Datos I
Fecha de Entrega: Viernes 31 de mayo - 8 a.m.
Todas las preguntas tienen el mismo valor

Tarea Número 9

1. Explique el método Gap Statistic propuesto por *R. Tibshirani, G. Walther, and T. Hastie (Standford University, 2001)*, en el aula virtual está el artículo completo. Otro método es el The Average Silhouette propuesto por Kaufman and Rousseeuw en 1987-1990, en el aula virtual está el artículo completo, dé también una explicación de este método.

2. En este ejercicio vamos a utilizar el conjunto de datos `weatherAUS.csv` que contiene aproximadamente 5 años de observaciones meteorológicas diarias de diferentes lugares en Australia. La tabla contiene 53934 filas (individuos) y 11 columnas (variables), las cuales se explican a continuación.

   - `MinTemp`: La temperatura mínima en grados centígrados.
   - `MaxTemp`: La temperatura máxima en grados centígrados.
   - `Rainfall`: La cantidad de lluvia registrada para el día en mm.
   - `Evaporation`: La denominada bandeja de evaporacion (mm) clase A.
   - `Sunshine`: El número de horas de sol brillante en el día.
   - `WindGustDir`: La dirección de la ráfaga de viento más fuerte en las 24 horas hasta la medianoche.
   - `WindGustSpeed`: La velocidad (km/h) de la ráfaga de viento más fuerte en las 24 horas hasta la medianoche.
   - `WindSpeed9am`: Velocidad del viento (km/h) promediada durante 10 minutos antes de las 9 a.m.
   - `Humidity9am`: Humedad (porcentaje) a las 9 a.m.
   - `Pressure9am`: Presión atmosférica (hpa) reducida al nivel medio del mar a las 9 a.m.
   - `Temp9am`: Temperatura (grados C) a las 9 a.m.

   Efectúe un análisis de k-medias realizando los siguientes pasos:

   a) Cargue la tabla de datos y ejecute un `str(...)`, `summary(...)` y un `dim(...)`, verifique la correcta lectura de los datos.

   b) Elimine las filas con `NA` usando el comando `na.omit(...)`. ¿Cuántas filas de eliminaron?

   c) Elimine de la tabla de datos la variable `WindGustDir`. ¿Por qué se debe eliminar? ¿Qué otra alternativa se tiene en lugar de eliminarla?

   d) Observe que si ejecutamos el método clustering jerárquico `hclust(...)` con esta tabla de datos este nunca termina. ¿Por qué sucede esto?

   e) Ejecute un $k-$medias con $k = 3$ con los parámetros por defecto.

   f) Dé una interpretación de los resultados del punto anterior usando un gráfico de barras.

g) Ejecute un $k-$medias con $k = 3$ con los parámetros por defecto, pero antes estandarice los datos.

h) Dé una interpretación de los resultados del punto anterior usando un gráfico de barras. ¿Hay alguna diferencia respecto a la interpretación del punto *f)*?

i) Ejecute un $k-$medias con $k = 3$ con `iter.max=1000` y `nstart=50`.

j) Dé una interpretación de los resultados del punto anterior usando un gráfico tipo radar.

k) Observe que si ejecutamos el método $k-$medoides con $k = 3$ con `nstart=50` con esta tabla de datos este método nunca termina o tarda demasiado. ¿Por qué sucede esto?

l) Ejecute un $k-$medoides con $k = 3$ con `nstart=50`. Para esto tome una muestra de 5 % de los datos, esto se puede lograr con el siguiente código:

```
numero.filas  <- nrow(datos.estandarizados)
muestra       <- sample(1:numero.filas,numero.filas*0.05)
datos.muestra <- datos.estandarizados[muestra,]
modelo_kmd    <- pam(datos.muestra, 3, nstart = 50)
```

m) Aplique el método `dbscan` con todos los datos ¿Termina el método? Sino termina ejecútelo en la muestra del 5 %.

n) Dé una interpretación de los resultados del punto anterior usando usando un gráfico tipo radar. ¿Qué diferencias nota respecto a k-medias?

ñ) Construya el Codo de Jambu usando `iter.max=100` y `nstart=5`, ¿cuántos conglomerados (clústeres) sugiere el codo? Utilice también el método `silhouette` de la función `fviz_nbclust`, ¿cuántos conglomerados (clústeres) sugiere este método? Para este ejercicio puede utilizar una muestra del 20 % en caso de limitaciones computacionales.

3. Programe en **R** funciones para calcular la inercia total, la inercia inter-clases y la inercia intra-clases, estas funciones deben recibir al menos la tabla de datos, la cantidad de clústeres y a qué cluster pertenece cada individuo de la tabla de datos. Luego para el ejemplo Estudiantes visto en clase (puede usar el archivo `NotasEscolaresExcelKMeans.csv` que está en el aula virtual) verifique el **Teorema de Fisher** utilizando solamente 2 clústeres usando las funciones que usted programó y las respectivos resultados que proporciona **R** con los comandos:

```
grupos <- kmeans(datos, 2, iter.max = 100)
grupos$totss  # Inercia Total
grupos$tot.withinss  # Inercia Intra-clases
grupos$betweenss  # inercia Inter-clases
```

4. El objetivo de este ejercicio es probar nuevamente el **Teorema de Fisher**. Sea $C_1, C_2, \ldots, C_K$ una partición del conjunto de n individuos Ω y sean $g_1, g_2, \ldots, g_K$ sus repectivos centros de gravedad.

a) Pruebe que si g es el centro de gravedad de Ω entonces el *Término Cruzado* es igual a cero, es decir, pruebe que:

$$\sum_{i=1}^{K} \sum_{x \in C_i} (x - g_i)(g - g_i) = 0.$$

b) Pruebe que:

$$\frac{1}{2|C_i|}\sum_{x\in C_i}\sum_{y\in C_i}(x-y)^2=\sum_{x\in C_i}(x-g_i)^2.$$

c) Pruebe que si se tiene que $|C_i|=\frac{n}{K}$ donde n es la cantidad total de individuos, entonces:

$$\frac{1}{2K}\sum_{i=1}^{K}\sum_{j=1}^{K}|C_i|(g_j-g_i)^2=\sum_{i=1}^{K}|C_i|(g-g_i)^2.$$

d) Pruebe que:

$$\sum_{i=1}^{K}\sum_{x\in C_i}(x-g)^2=\sum_{i=1}^{K}\sum_{x\in C_i}(x-g_i)^2+\sum_{i=1}^{K}|C_i|(g-g_i)^2.$$

e) ¿Porqué de la identidad anterior se deduce el **Teorema de Fisher**?

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. El ejercicio 5 lo pueden hacer a mano.

oldemar rodríguez

CONSULTOR en M1NER14 DE D4T0S