Modelos de Soporte No Supervisado

Evaluación de las tendencias en la agrupación

Evaluación de tendencias en agrupación

- Un paso previo en la aplicación de alguna de las metodologías de clustering es la evaluación de las tendencias en agrupación
- Se busca identificar si la agrupación potencial no es meramente aleatoria
- Recuerden que al aplicar las metodologías se generarán los clusters pero no necesariamente se forman agrupaciones en los datos

Evaluación de tendencias en agrupación

Necesitamos los paquetes

factoextra

clustertend

Install.packages(c(“factoextra”, “clustertend”))

Evaluación de tendencias en agrupación

Utilizaremos dos bases de datos:

“iris” (incluida en R)

Con base en los datos “iris” generaremos una base de datos de manera aleatoria

Exploración de la base

head(iris,3)

¿Qué tratamiento se daría a la variable “Species”?

Evaluación de tendencias en agrupación

```
#Base de datos
df <- iris[, -5]

# Generación de valores aleatorios de la base de datos inicial
random_df <- apply(df, 2,
    function(x){runif(length(x), min(x), (max(x)))})
    random_df <- as.data.frame(random_df)
```

Evaluación de tendencias en agrupación

#Estandarizamos

df <- iris.scaled <- scale(df)

random_df <- scale(random_df)

Inspeccionamos visualmente los datos

¿Qué sugiere para esto?

Evaluación de tendencias en agrupación

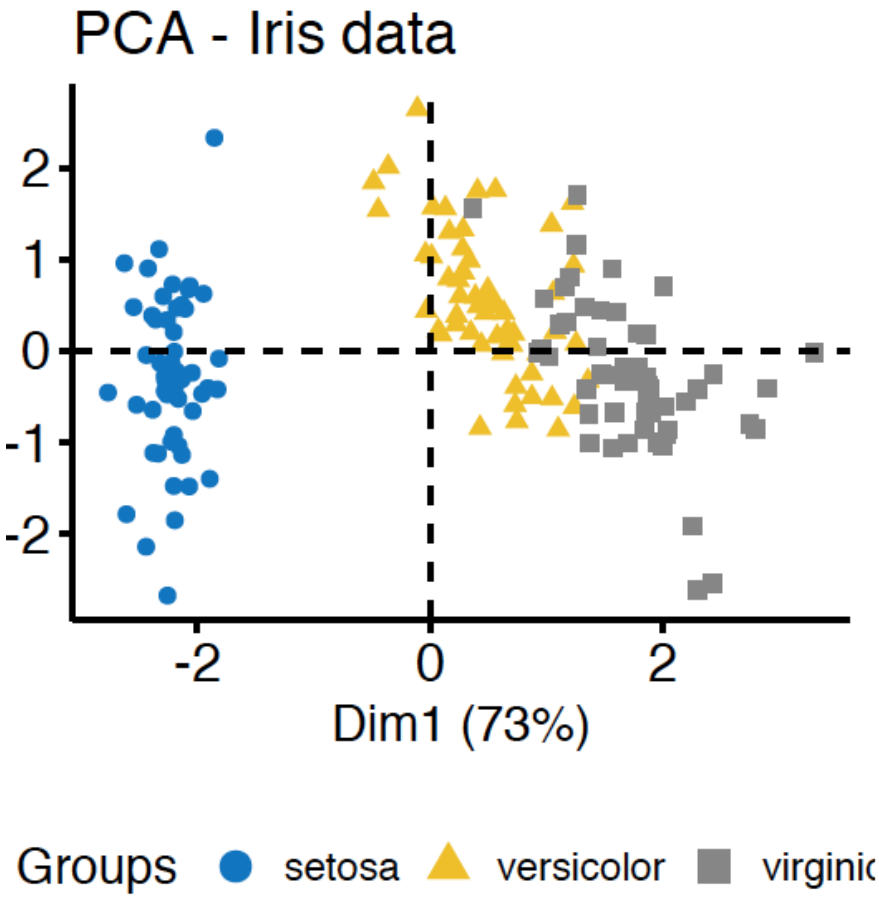
Podemos aplicar CP

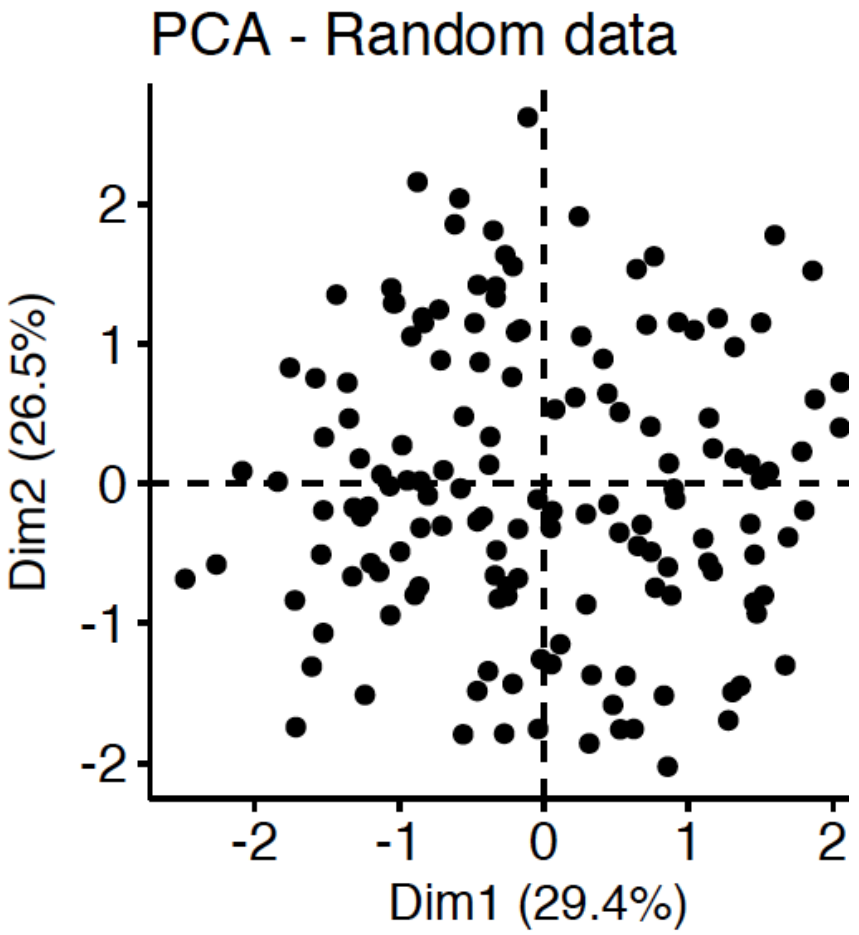
```
# Graficamos la base original

fviz_pca_ind(prcomp(df), title = "PCA - Iris data",
             habillage = iris$Species, palette = "jco",
             geom = "point", ggtheme = theme_classic(),
             legend = "bottom")

# Graficamos la base generada
fviz_pca_ind(prcomp(random_df), title = "PCA - Random data",
             geom = "point", ggtheme = theme_classic())
```

Evaluación de tendencias en agrupación





Evaluación de tendencias en agrupación

¿Por qué es importante evaluar tendencias de agrupación?

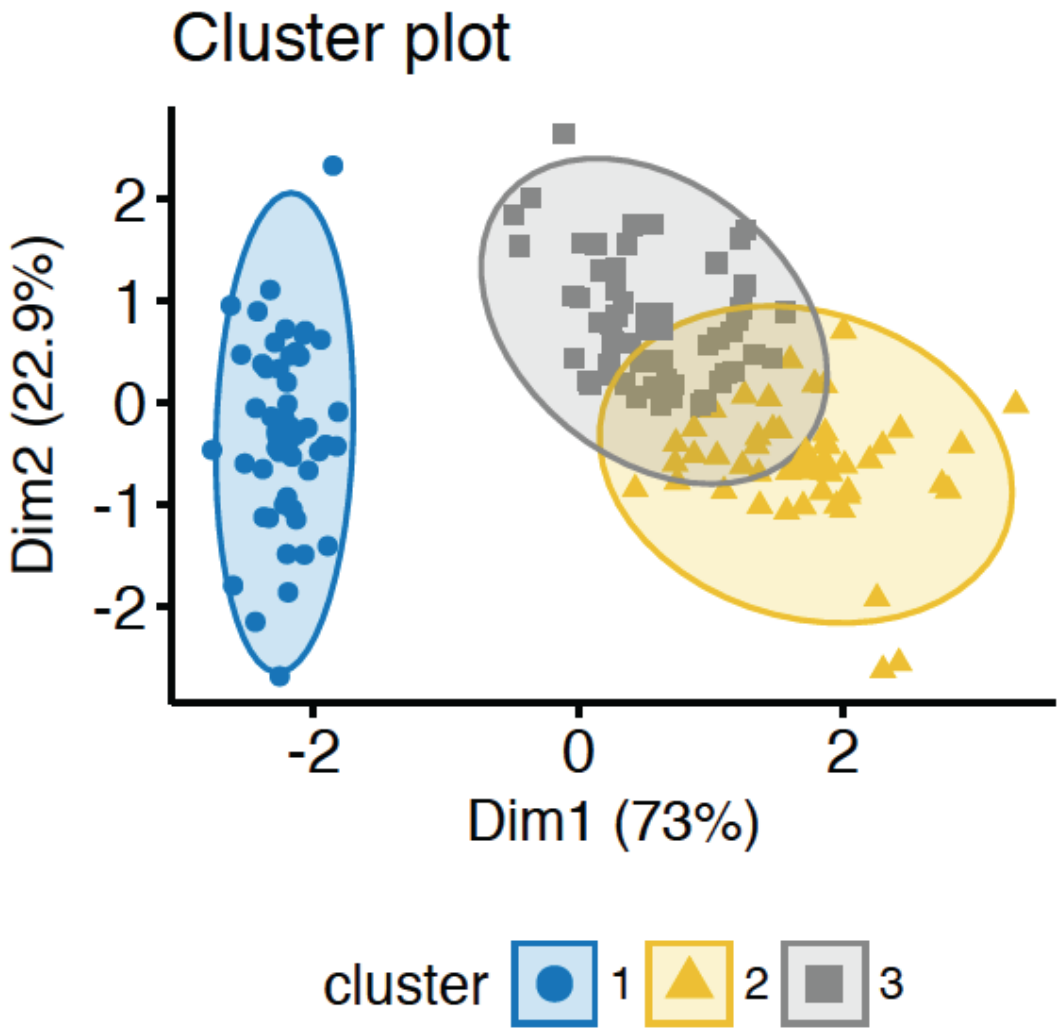
- Para ilustrar la utilidad de este proceso, generaremos clustering por k-medias y jerárquico aglomerativo con las dos bases de datos
- Utilizaremos las funciones vistas previamente

Evaluación de tendencias en agrupación

K-medias

```
set.seed(123)
# K-means base iris
km.res1 <- kmeans(df, 3)
fviz_cluster(list(data = df, cluster =
km.res1$cluster),
    ellipse.type = "norm", geom = "point", stand =
    FALSE, palette = "jco", ggtheme =
    theme_classic())
```

Evaluación de tendencias en agrupación



Evaluación de tendencias en agrupación

K-medias Base simulada

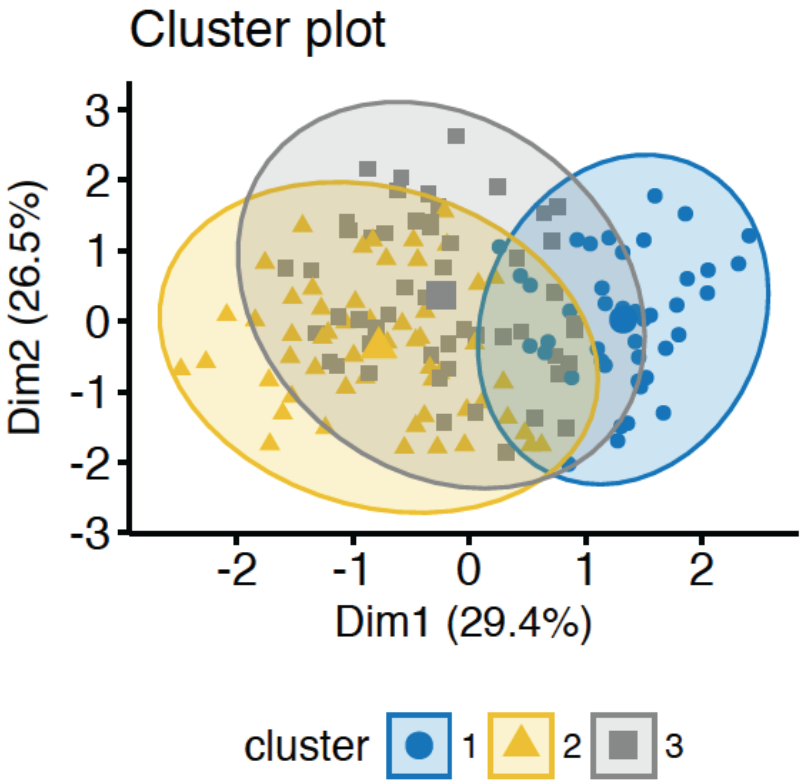
```
set.seed(123)
# K-means base iris
km.res1 <- kmeans(df, 3)
fviz_cluster(list(data = df, cluster =
km.res1$cluster),
    ellipse.type = "norm", geom = "point", stand =
    FALSE, palette = "jco", ggtheme =
    theme_classic())
```

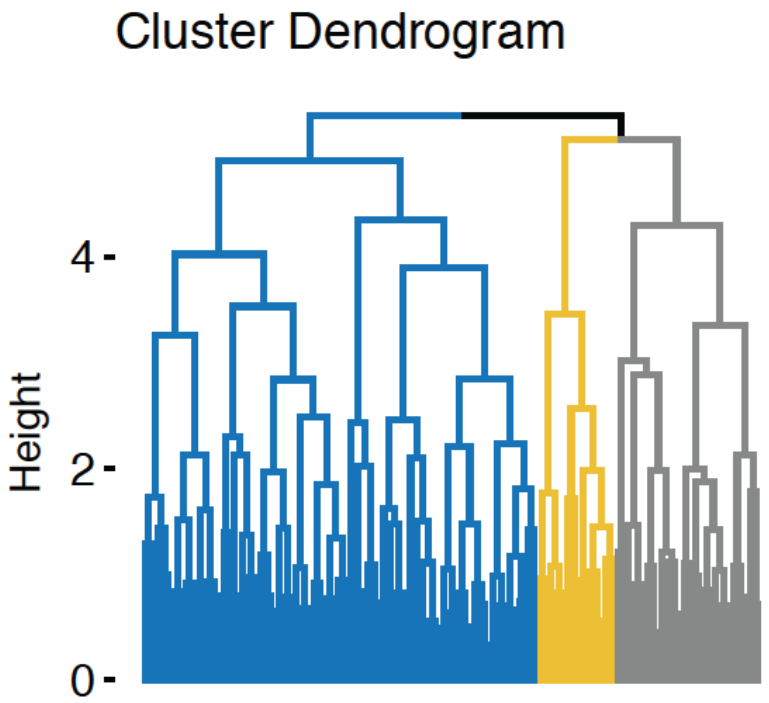
Evaluación de tendencias en agrupación

Jerárquico aglomerativo en la base simulada

```
fviz_dend(hclust(dist(random_df)), k = 3,
    k_colors = "jco",
    as.ggplot = TRUE, show_labels = FALSE)
```

Evaluación de tendencias en agrupación





Evaluación de tendencias en agrupación

- Se puede ver que el algoritmo k-means y la agrupación jerárquica imponen una clasificación en el conjunto de datos aleatoriamente distribuidos uniformemente, incluso si no hay clusters significativos presentes en él.

- Esta es la razón por la cual los métodos de evaluación de tendencias de agrupamiento debe usarse para evaluar la validez del análisis de agrupamiento.

- Es decir, si un el conjunto de datos dado contiene clusters significativos.

Evaluación de tendencias en agrupación

Describiremos dos métodos para evaluar la tendencia de agrupamiento:

- i) una estadística (estadística de Hopkins)

- ii) un método visual (Evaluación visual del algoritmo de Tendencia de clúster)

Estadística de Hopkins

La estadística de Hopkins se utiliza para evaluar la tendencia a la agrupación de un conjunto de datos

Midiendo la probabilidad de que un conjunto de datos determinado sea generado por datos de una distribución uniforme

Es una prueba de la aleatoriedad espacial de los datos.

Estadística de Hopkins

Algoritmo: Sea D una base de datos

1. Genere una muestra de n puntos (p1, ..., pn) de D.
2. Para cada punto pi en D, encuentre su vecino más cercano pj; luego calcule la distancia entre pi y pj y denotarlo como xi = dist (pi, pj)
3. Generar un conjunto de datos simulados (randomD) aleatorios extraídos de una distribución uniforme con n puntos (q1, ..., qn)
4. Para cada punto qi en randomD, encuentre su vecino más cercano qj en D; entonces calcular la distancia entre qi y qj y denotarlo yi = dist (qi, qj)
4. Calcule la estadística de Hopkins (H) como la suma de la distancia del vecino más cercano en el conjunto de datos aleatorios dividido por la suma de las distancias en el conjunto de datos reales y simulados.

Estadística de Hopkins

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

Un valor de H alrededor de 0.5 significa que:

$$\sum_{i=1}^{n} y_i \qquad \sum_{i=1}^{n} x_i$$

están cerca uno del otro, y por lo tanto los datos D están distribuidos uniformemente.

Las hipótesis nula y alternativa se definen de la siguiente manera:

Hipótesis nula: el conjunto de datos D está distribuido uniformemente (es decir, no existen clusters significativos)

Hipótesis alternativa: el conjunto de datos D no está uniformemente distribuido (es decir, contiene clusters significativos)

Si el valor de la estadístico de Hopkins es cercano a cero, entonces podemos rechazar la hipótesis nula y concluir que el conjunto de datos D es significativamente susceptible de aplicar alguna metodología de cluster.

Estadística de Hopkins

Estimación del estadístico de Hopkins

hopkins**(data, n)**

data: matriz de datos

n: el número de puntos a seleccionar de los datos

Estadística de Hopkins

Ejemplo:

```
library(clustertend)
# Estimar el estadístico de Hopkins base original
set.seed(123)
hopkins(df, n = nrow(df)-1)
# Estimar el estadístico de Hopkins base simulada
set.seed(123)
hopkins(random_df, n = nrow(random_df)-1)
```

VAT (visual assesment of cluster tendency)

Algoritmo:

1. Calcule la matriz de disimilitud (DM) entre los objetos en el conjunto de datos usando la medida de distancia euclidiana
2. Reordenar la DM para que objetos similares estén cerca el uno del otro. Este proceso crea una matriz de disimilitud ordenada (ODM)
3. La ODM se muestra como una imagen de desemejanza ordenada (ODI), que es el resultado visual del VAT

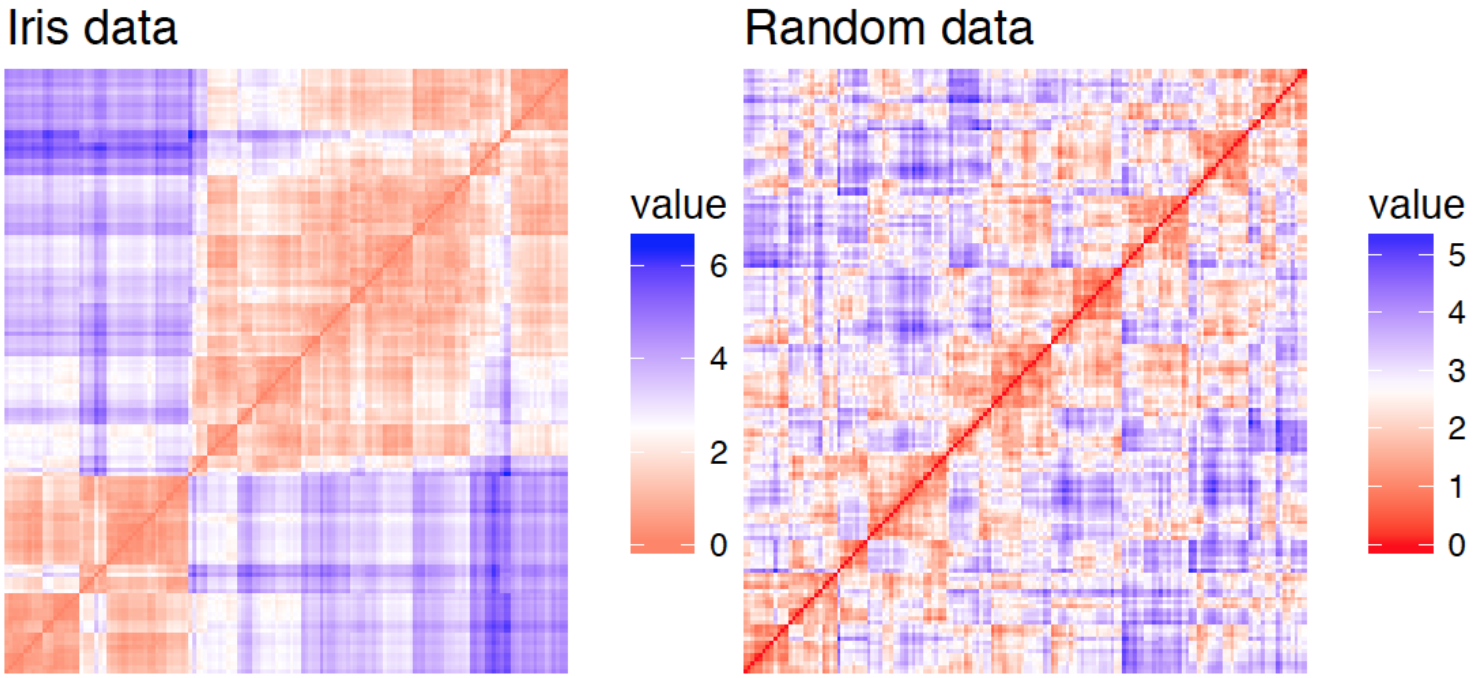
VAT (visual assesment of cluster tendency)

Ejemplo:

```
fviz_dist(dist(df), show_labels = FALSE)+
    labs(title = "Iris data")

fviz_dist(dist(random_df), show_labels = FALSE)
    +labs(title = "Random data")
```

VAT (visual assesment of cluster tendency)



El nivel de color es proporcional al valor de la disimilaridad entre las observaciones:
Rojo si dist (xi, xj) = 0 y azul si dist (xi, xj) = 1.
Objetos que pertenecen al mismo grupo se muestran en orden consecutivo.
El VAT detecta la tendencia de agrupación en una forma visual contando el número de bloques oscuros de forma cuadrada a lo largo de la diagonal en una imagen de VAT.

Resumen

Se describió cómo evaluar la tendencia de agrupación usando el estadístico de Hopkins y un método visual.

Después de mostrar que los datos son agrupables, el siguiente paso es determinar la cantidad de cluster óptimos en los datos.