Modelos de soporte no supervisado

Agglomerative Clustering

Agglomerative clustering

- El agrupamiento aglomerativo es el tipo más común de agrupación jerárquica
- Agrupa objetos en clusters en función de su similitud.
- AGNES(Agrupamiento aglomerativo). El algoritmo comienza tratando cada objeto como un cluster.
- A continuación, los pares de clusters se fusionan sucesivamente hasta que todos los clusters se han fusionado en un gran grupo que contiene todos los objetos.
- El resultado es un árbol de representación de los objetos, llamado dendrograma.

Agglomerative clustering

- Agrupamiento aglomerativo funciona de una manera "ascendente".
- Cada objeto es inicialmente considerado como un clúster de elemento único (hoja).
- En cada paso del algoritmo, los dos clusters que son los más similares se combinan en un nuevo clúster más grande (nodos).
- Este procedimiento se itera hasta que todos los puntos sean miembros de un solo gran clúster.

Agglomerative clustering

- En el caso particional, en cada paso de la iteración, el clúster más heterogéneo se divide en dos.
- El proceso se itera hasta que todos los objetos están en su propio grupo.
- Tenga en cuenta que, la agrupación aglomerativa es buena para identificar pequeños grupos.
- La agrupación particional es buena para identificar grandes agrupaciones.
- ¿Por qué razón?

Agglomerative clustering

Algoritmo en R:

1. Preparación de los datos
2. Información de (dis) similitud informática entre cada par de objetos en los datos.
3. Usar la función de vinculación (linkage) para agrupar objetos en el árbol de clúster jerárquico, en función de la información de distancia generada en el paso 1.
4. Objetos / clusters que están cerca la proximidad se vinculan utilizando la función de vinculación.
5. Determinar dónde cortar el árbol jerárquico en grupos. Esto crea un partición de los datos.

Agglomerative clustering

- Los datos deben ser una matriz numérica con:

  a) filas que representan observaciones (individuos);

  b) columnas que representan variables.

Usaremos los conjuntos de datos R base USArrests.

Agglomerative clustering

```
# Cargar los datos
    data("USArrests")
# Estandarizamos en caso de ser necesario
    df <- scale(USArrests)
# Verificamos las primeras seis filas
head(df, nrow = 6)
```

Agglomerative clustering

```
# Estimamos la matriz de disimilaridad
# df = datos estandarizados

res.dist <- dist(df, method = "euclidean")
as.matrix(res.dist)[1:6, 1:6]
```

Agglomerative clustering: linkage

- La función de vinculación toma la información de distancia, devuelta por la función dist (),
- Agrupa pares de objetos en grupos basados en su similitud.
- A continuación, estos nuevos los clústeres formados están vinculados entre sí para crear clusters más grandes.
- Este proceso es iterado hasta que todos los objetos en el conjunto de datos original estén vinculados entre sí de forma jerárquica en un árbol.

Agglomerative clustering: linkage

- Por ejemplo, dada una matriz de distancia "res.dist" generada por la función dist ()

- La función de base R hclust () se puede usar para crear el árbol jerárquico.

hclust () se puede usar de la siguiente manera:

```
res.hc <- hclust(d = res.dist, method = "ward.D2")
```

Agglomerative clustering: linkage

Donde:

d: una estructura de disimilaridad producida por la función dist ().

método: el método de aglomeración (vinculación) que se utilizará para calcular la distancia entre racimos. Los valores permitidos son "ward.D", "ward.D2", "single", "Completo", "promedio", "mcquitty", "mediana" o "centroide".

Hay muchos métodos de aglomeración de clúster (es decir, métodos de vinculación).

Agglomerative clustering: linkage

Métodos de aglomeración más comúnes:

- Enlace máximo o completo: la distancia entre dos clústeres se define como el valor máximo de todas las distancias por pares entre los elementos en el grupo 1 y los elementos en el grupo 2. Tiende a producir grupos más compactos.
- Enlace mínimo o único: la distancia entre dos clusters está definida como el valor mínimo de todas las distancias pairwise entre los elementos en clúster 1 y los elementos en el clúster 2. Tiende a producir clústeres largos, "sueltos".
- Enlace promedio: la distancia entre dos grupos se define como distancia promedio entre los elementos en el grupo 1 y los elementos en el grupo 2.
- Enlace centroide: la distancia entre dos clústeres se define como la distancia entre el centroide para el grupo 1 (un vector medio de variables de longitud p) y el centroide para el grupo 2.
- Método de varianza mínima de Ward: minimiza la varianza total dentro del clúster. En cada paso, el par de clusters con una distancia mínima entre ellos serán fusionados.

Agglomerative clustering: linkage

- En cada etapa del proceso de agrupamiento, los dos clusters que tienen distancia de vinculación más pequeña, están vinculados entre sí.

- Por lo general, se prefiere la vinculación completa y el método de Ward.

Agglomerative clustering: linkage

Dendrograma

- Los dendrogramas corresponden a la representación gráfica del árbol jerárquico generado por la función hclust ().
- El Dendrograma se puede producir en R usando el gráfico de función base (res.hc), donde res.hc es la salida de hclust ().
- Nosotros usaremos la función function fviz_dend () [en factoextra R paquete] para producir un dendrograma.
- Primero instalamos factoextra escribiendo: install.packages ("factoextra").

Agglomerative clustering: linkage

cex: Es la opción que nos da el tamaño de las etiquetas en las gráficas

library**("factoextra")**

fviz_dend**(res.hc, cex = 0.5)**

Agglomerative clustering: linkage

- En el dendrograma que cada hoja corresponde a un objeto.
- A medida que avanzamos en el árbol, los objetos que son similares entre sí se combinan en ramas.
- La altura de la fusión, proporcionada en el eje vertical, indica la (dis) similitud / distancia entre dos objetos / clusters.
- Cuanto mayor sea la altura de la fusión, menos similares serán los objetos.
- Esta altura se conoce como la distancia cophenetic entre los dos objetos.

Agglomerative clustering: linkage

- Tenga en cuenta que las conclusiones sobre la proximidad de dos objetos se pueden extraer solo en función de la altura donde las ramas que contienen esos dos objetos se fusionan.
- No podemos usar la proximidad de dos objetos a lo largo del eje horizontal como criterio de similitud.
- Para identificar subgrupos, podemos cortar el dendrograma a cierta altura.

Agglomerative clustering

Verificar el árbol de clúster

- Después de vincular los objetos en un conjunto de datos en un árbol de clúster jerárquico, es posible que desee evaluar que las distancias (es decir, alturas) en el árbol reflejen las distancias originales.
- Una forma de medir qué tan bien el árbol de clúster generado por la función hclust () refleja sus datos es calcular la correlación entre las distancias cophenetic y los datos de distancia originales generados por la función dist ().
- Si la agrupación es válida, la vinculación de los objetos en el árbol del clúster debe tener una fuerte correlación con el distancias entre objetos en la matriz de distancia original.
- Cuanto más cerca esté el valor del coeficiente de correlación a 1, con mayor precisión la solución de agrupamiento refleja sus datos. Se considera que los valores superiores a 0,75 son buenos.
- El método de vinculación "promedio" parece producir valores altos de esta estadística. Esto puede sea una razón por la que es tan popular.
- La función base R cophenetic () se puede usar para calcular las distancias cophenetic para agrupación jerárquica.

Agglomerative clustering

Estimación de las distancias cophenetic

res.coph <- cophenetic(res.hc)

Correlación entre la distancia cophenetic y la distancia original

cor(res.dist, res.coph)

La (di)similitud cophenetic o distancia cophenetic de dos objetos es una medida de qué tan similares deben ser esos dos objetos para agruparse en el mismo clúster. La distancia cophenetic entre dos objetos es la altura del dendrograma donde las dos ramas que incluyen los dos objetos se combinan en una sola rama. Fuera del contexto de un dendrograma, es la distancia entre los dos clusters más grandes que contienen los dos objetos individualmente cuando se fusionan en un solo cluster que contiene ambos.

Agglomerative clustering

- Ejecute la función hclust () nuevamente utilizando el método de vinculación promedio.

- Luego, estime la disimilaridad cophenetic () para evaluar la solución de agrupamiento.

Agglomerative clustering

```
res.hc2 <- hclust(res.dist, method = "average")

cor(res.dist, cophenetic(res.hc2))
```

Agglomerative clustering

Cortar el dendrograma en diferentes grupos:

- Uno de los problemas con la agrupación jerárquica es que no nos dice cuantas agrupaciones hoy, o dónde cortar el dendrograma para formar grupos.
- Puede cortar el árbol jerárquico a una altura determinada para dividir sus datos en racimos.
- El comando cutree de la función base R () se puede usar para cortar un árbol, generado por la función hclust (), en varios grupos especificando el número deseado de grupos o la altura de corte.
- Devuelve un vector que contiene el número de clúster de cada observación.

Agglomerative clustering

```
# Cortar el árbol en cuarto grupos

grp <- cutree(res.hc, k = 4)
head(grp, n = 4)

## Alabama Alaska Arizona Arkansas
## 1 2 2 3
# Números de miembros en cada cluster

table(grp)

## grp
## 1 2 3 4
# Obtener los nombres de los miembros del clúster 1
rownames(df)[grp == 1]
```

Agglomerative clustering

El resultado de los cortes se puede visualizar fácilmente utilizando la función fviz_dend () [en factoextra]:

Cortar en cuatro grupos y colorear por grupos

fviz_dend**(res.hc, k = 4,** # Cortar en cuatro grupos

cex = 0.5, # Tamaño de la etiqueta

k_colors = c**("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"), color_labels_by_k = TRUE,** # Color de las etiquetas por grupo

rect = TRUE # Agregar un rectángulo alrededor de los grupos

)

Agglomerative clustering

- Usando la función fviz_cluster () [en factoextra], también podemos visualizar el resultado en un diagrama de dispersión.

- Las observaciones están representadas por puntos, utilizando componentes principales.

- Y dibujando un marco alrededor de cada grupo.

Agglomerative clustering

- fviz_cluster**(**list**(data = df, cluster = grp),**

  palette = c**("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),**

  ellipse.type = "convex", # Elipse de concentración

  repel = TRUE, # Evitar que se sobrepongan las etiquetas

  show.clust.cent = FALSE, ggtheme = theme_minimal**())**

Agglomerative clustering

- Paquete Cluster R
- El clúster de paquete R facilita la realización del análisis de clúster en R.
- Proporciona la función agnes () y diana () para calcular el agrupamiento aglomerativo y divisivo, respectivamente.
- Estas funciones realizan todos los pasos necesarios. No necesitas para ejecutar la función scale (), dist () y hclust () por separado.
- Las funciones se pueden ejecutar de la siguiente manera:

Agglomerative clustering

```
library("cluster")

# Cluster Jerárquico Aglomerativo
    res.agnes <- agnes(x = USArrests, # Matríz de datos
    stand = TRUE, # Estandarización de datos
    metric = "euclidean", # Métrica de distancia
    method = "ward" # Método de Linkage
    )

# Cluster particional
    res.diana <- diana(x = USArrests, # Matriz de datos
    stand = TRUE, # Estandarización de datos
    metric = "euclidean" # Métrica de distancia
    )
```

Agglomerative clustering: linkage

- Después de ejecutar agnes () y diana (), puede usar la función fviz_dend () [in factoextra] para visualizar el resultado:

fviz_dend**(res.agnes, cex = 0.6, k = 4)**

Agglomerative clustering: linkage

- La agrupación jerárquica es un método de análisis de conglomerados, que produce un árbol basado representación (es decir, dendrograma) de un dato.
- Los objetos en el dendrograma están vinculados juntos con base a su similitud.
- Para realizar el análisis de clúster jerárquico en R, el primer paso es calcular la matriz de distancia utilizando la función dist ().
- A continuación, se utiliza el resultado de este cálculo por la función hclust () para producir el árbol jerárquico.
- Finalmente, puedes usar el function fviz_dend () [en factoextra R paquete] para trazar fácilmente un dendrograma.
- También es posible cortar el árbol a una altura dada para dividir los datos en múltiples grupos (función R cutree ()).