UNIVERSITAT DE BARCELONA

Universitat Oberta de Catalunya

Prueba de evaluación continua 2
Estadística Multivariante

Francesc Carmona

4 de enero de 2020

El bosque de Woodyard Hammock

Vamos a trabajar el análisis de conglomerados con los datos ecológicos de Woodyard Hammock, un bosque de magnolias y hayas en el norte de Florida.
Los datos son el recuento del número de árboles de cada especie en $n = 72$ lugares. Se identificaron y contabilizaron un total de 31 especies, sin embargo, solo trabajaremos con las $p = 13$ especies más comunes que se indican a continuación:

`carcar`	*Carpinus caroliniana*	Ironwood
`corflo`	*Cornus florida*	Dogwood
`faggra`	*Fagus grandifolia*	Beech
`ileopa`	*Ilex opaca*	Holly
`liqsty`	*Liquidambar styraciflua*	Sweetgum
`maggra`	*Magnolia grandiflora*	Magnolia
`nyssyl`	*Nyssa sylvatica*	Blackgum
`ostvir`	*Ostrya virginiana*	Blue Beech
`oxyarb`	*Oxydendrum arboreum*	Sourwood
`pingla`	*Pinus glabra*	Spruce Pine
`quenig`	*Quercus nigra*	Water Oak
`quemic`	*Quercus michauxii*	Swamp Chestnut Oak
`symtin`	*Symplocus tinctoria*	Horse Sugar

Magnolia *Magnolia grandiflora*

La primera columna da un código de identificación de la especie con seis letras, la segunda columna da su nombre científico en nomenclatura binomial y la tercera columna da el nombre común (en inglés) de cada especie. Los árboles más frecuentes en este bosque son las hayas y las magnolias.

Los datos se hallan en el archivo `wood.dat` que se encuentra en la web del curso STAT 505 *Applied Multivariate Statistical Analysis* del Eberly College of Science de la Universidad de Pennsylvania[1]

Ejercicio 1 (70 pt.)

(a) Después de reducir la base de datos original a las $p = 13$ especies destacadas, nos planteamos elegir una distancia adecuada para el análisis de conglomerados. Con este tipo de datos se suele utilizar

[1] `https://newonlinecourses.science.psu.edu/stat505/lesson/14/14.1`

la disimilaridad de Bray-Curtis con la siguiente fórmula según la Wikipedia:

$$d_{BC}^{(1)}(i,j) = 1 - \frac{2C_{ij}}{S_i + S_j}$$

donde C_{ij} es la suma del menor valor para cada una de las especies en común de las dos muestras. S_i y S_j son el total de especímenes en cada una de las muestras.

También se puede calcular con la fórmula que utiliza la función `vegdist()` del paquete `vegan`:

$$d_{BC}^{(2)}(i,j) = \frac{\sum_{k=1}^{p} |x_{ik} - x_{jk}|}{\sum_{k=1}^{p} (x_{ik} + x_{jk})}$$

Calcular la disimilaridad de Bray-Curtis entre las dos primeras muestras de la base de datos con ambas fórmulas (sin utilizar la función `vegdist()`). En la primera fórmula nos puede ayudar la función `pmin()` de **R**.

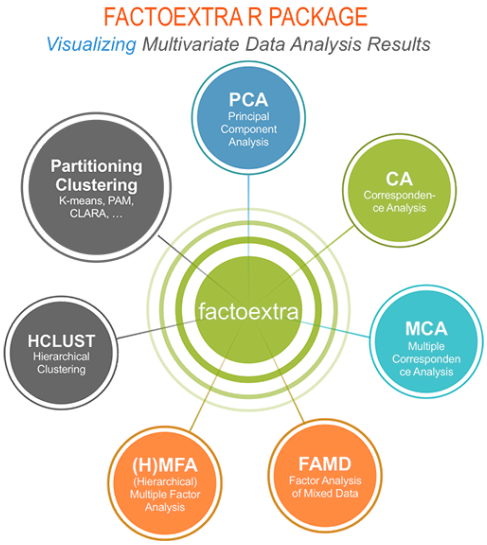
Calcular a continuación la matriz de disimilaridades completa con la función `vegdist()` y comprobamos que el valor hallado para las dos primeras muestras coincide.

(b) Sin embargo, la disimilaridad de Bray-Curtis no verifica la desigualdad triangular de forma que no es una distancia. Comprobar este hecho, tal vez con algún contraejemplo entre dos muestras o con alguna función de **R** que podamos hallar[2].

Para corregir este problema podemos transformar la disimilaridad de Bray-Curtis calculando la raíz cuadrada de sus valores. Comprobar que dicha raíz cuadrada verifica la desigualdad triangular.

(c) Muchos procedimientos de Análisis multivariante requieren que la distancia utilizada sea euclídea. Comprobar que la raíz cuadrada de la disimilaridad de Bray-Curtis es una distancia euclídea.

(d) Evaluar la tendencia al agrupamiento con el estadístico de Hopkins que podéis calcular con la función `get_clust_tendency()` del paquete `factoextra`[3].



La función `fviz_dist()` del mismo paquete permite realizar una imagen de la disimilaridad propuesta en el apartado anterior (*ordered dissimilarity image*, ODI) para evaluar gráficamente la

[2] No hay que programar. Se puede hallar una función en un cierto paquete y otra que mejora la anterior.
[3] Un estadístico recíproco se puede calcular con la función `hopkins()` del paquete `clustertend`.

tendencia al agrupamiento. También la misma función `get_clust_tendency()` lo hace con una imagen menos vistosa. ¿Cuantos conglomerados se intuyen?

Para saber más sobre el análisis de la tendencia al agrupamiento, la fórmula del estadístico de Hopkins y un ejemplo se puede consultar la página

`https://www.datanovia.com/en/lessons/assessing-clustering-tendency/`

Nota: Un valor alto del estadístico de Hopkins nos permitirá realizar el análisis de conglomerados con la confianza de que los grupos estarán bastante bien formados. Un valor bajo nos dice que la formación de los conglomerados será mucho más difícil y no tan clara.

(e) Realizar un análisis de conglomerados jerárquico con el método de Ward[4] de la distancia propuesta en el apartado (c). Dibujar el dendograma resultante.

Nota: Justamente el método de Ward exige que la distancia sea euclídea, pero no necesariamente la distancia euclídea.

Dibujar también un *heatmap* de este análisis con la función `heatmap()`. Para ello, hay que elegir bien los parámetros `distfun=` y `hclustfun=` y una escala de colores. ¿Para qué sirve el *heatmap*?

Nota: No nos interesa reordenar las variables o columnas y tampoco un dendograma sobre ellas.

(f) El siguiente paso es estudiar por algún criterio el número óptimo de conglomerados para el análisis jerárquico. Con la distancia del apartado (c) en particular, lo más sencillo es utilizar el criterio de las siluetas.

Sin embargo, el criterio de las siluetas y otros métodos como el de la suma de cuadrados dentro de los grupos o WSS o también el estadístico GAP se pueden aplicar con la función `fviz_nbclust()` del paquete `factoextra`. Habrá que especificar que se trata del método de clasificación jerárquico (que por defecto utiliza el método de Ward) y también la distancia raíz cuadrada de la disimilaridad de Bray-Curtis.

(g) Estudiar con la misma distancia el número óptimo de conglomerados con el método PAM.

(h) De los apartados anteriores se deduce que no hay un número claro de conglomerados. Un ecólogo experto propone que sean 6, ya que su experiencia así lo ha visto en otros bosques similares. Dibujar el dendograma del apartado (e) con esta partición.

Con el objetivo de caracterizar los conglomerados, calcular una tabla con las medianas de las frecuencias de cada especie en cada uno de los seis conglomerados del dendograma anterior. Descartar de esta tabla las especies con una mediana inferior o igual a 4 en todos los conglomerados.

Representar la tabla anterior con un análisis de correspondencias y un gráfico asimétrico.

[4] Se puede utilizar la función `hclust()` o la función `agnes()`.

Ejercicio 2 (30 pt.)

Es posible utilizar la regresión logística para estudiar la discriminación entre dos grupos dado un conjunto de variables explicativas. Para más detalles podéis consultar el apartado 8.10 del libro de Manly[1].

(a) Realizar un análisis discriminante con ayuda de la regresión logística con los datos de los 49 gorriones hembra después de una tormenta.

Reproducir con todo detalle la tabla 8.6 de la página 153 del libro de Manly.

(b) Contrastar con un test χ^2 la significación de la regresión y explicar su resultado.

(c) Realizar un gráfico como el de la figura 8.2 de la página 156 del libro de Manly pero con los datos de este ejercicio y valorar el resultado.

(d) Calcular la tabla de confusión de la clasificación obtenida con la regresión logística.

Referencias

[1] Bryan F.J. Manly and Jorge A. Navarro Alberto (2017), *Multivariate Statistical Methods: A Primer*, Fourth Edition. Chapman and Hall/CRC. Springer-Verlag.