Semillas de trigo (**Hierarchical-clustering**)

Preparación de los datos

Importando el CSV de los datos

Antes de comenzar con el análisis de los datos, debemos cargar el archivo de **comma-separated values** (**CSV**) de éstos en **R Studio**. Para poder acceder y descargar directamente el archivo en caso de que se desee efectuar un análisis más exhaustivo de los mismos en un futuro consultar https://data.world/databeats/seeds.

```
# Biblioteca para hacer la lectura del archivo.
library(readr)
# Leyendo el CSV dándole la ruta relativa.
Data = read.csv("../../data/seeds_dataset.csv")
# Visualizando el contenido del CSV.
View(Data)
```

```
## Warning in View(Data): X cannot set locale modifiers
```

Eliminando etiquetas de clase e identificadores

Dado que haremos un análisis comparativo por medio del algoritmo de aprendizaje de máquina no supervisado de **hierarchical-clustering**, vamos a eliminar para el conjunto sobre el cuál se estudiará la estructura de las características el valor de la variable objetivo (de salida), misma que no existe en el aprendizaje no supervisado al no haber como tal respuesta correctas e incorrectas.

```
# Tenemos una copia de los datos.
datos.caracteristicas = Data
# A la copia le quitamos las columnas que no son de interés.
datos.caracteristicas$ID <- NULL
datos.caracteristicas$seedType <- NULL
# Vemos cómo quedaron los nuevos datos.
View(datos.caracteristicas)
```

```
## Warning in View(datos.caracteristicas): X cannot set locale modifiers
```

Reescalamiento de las características

Con tal de evitar que algunas características de las semillas de trigo tengan más importancia que otras —por ejemplo, el área y el perímetro tienen valores considerablemente mayores a los de los coeficientes de asimetría—, procedemos a normalizar los valores. En otras palabras, buscamos que para cada valor de las características ocurra que $-1 \leq x_i \leq 1$ aproximadamente.

Para hacer más explícita la normalización que como se hizo en el caso de **k-means** vamos a aplicar la fórmula de normalización por pasos (**as.data.frame(scale())** ya lo hace por nosotros) que sigue:

$$x_i' = \frac{x_i - \mu_i}{s_i}$$

En donde:

- $\mu_i = \frac{1}{m} \sum_{j=1}^{m} x_i^{(j)}$ es la media.

- $s_i = max(s_i) - min(s_i)$ es el rango. Se puede usar la desviación estándar.

```
# Normalizando los datos y luego visualizándolos.

# Obteniendo las medias de cada columna y guardándolas en un vector.
m <- apply(datos.caracteristicas, 2, mean)
# Obteniendo las desviaciones estándar de cada columna y guardándolas en otro vector.
s <- apply(datos.caracteristicas, 2, sd)

# Reesclando las características y luego visualizando cómo quedan.
datos.escalados <- scale(datos.caracteristicas, m, s)
View(datos.escalados)
```

```
## Warning in View(datos.escalados): X cannot set locale modifiers
```

Calculando las distancias euclidianas.

Luego se procede a calcular las distancias existentes entre las observaciones (ejemplares) a partir de todas las características que exhiben en un espacio vectorial de dimensión menor o igual a la dimensión de las variables en el que podemos representar a los individuos por puntos.

```
# Usando la función que ya proporciona R para calcular las distancias.
distancias <- dist(datos.escalados)
```

Lo que se encuentra almacenado en este momento en la variable **distancias** es una matriz cuadrada de tamaño $n \times n$ en donde n es el número de ejemplares u observaciones. En nuestro caso, es de $210 \cdot 210$. Sin embargo, las entradas de la diagonal superior son vacías (no son necesarias), pues la matriz es simétrica al ser la distancia de A a B la misma de que de B a A, i.e la relación de distancia euclidiana entre puntos en un espacio vectorial es simétrica. Más aún, como la distancia de un punto a sí mismo es cero, entonces no se considera la diagonal.

Aplicación del algoritmo

Aplicamos ahora el algoritmo de agrupamiento jerárquico que es **greedy** que va generando las mezclas y divisiones de los grupos. La técnica usada es aglomerativa, pues si bien cada observación empieza estando en un grupo independiente, luego se van juntando hasta que todas queden en el mismo.

En otras palabras, se parte primera en tantos grupos como observaciones se tengan, luego se selecciona una media de similitud y se agrupan los clusters con mayor simulitud. Se sigue con este proceso hasta que ocurra alguna de las siguientes condiciones:

- Se logra la conformación de un solo cluster.
- Se forma un número m de clusters que había sido dado antes y este fijado.
- Se detecta mediante medidas estadísticas que no tiene caso seguir con más iteraciones, pues es posible detectar que los datos son tan heterogéneos (hay muchos **outliers**) que no van a converger entonces en un solo cluster jamás.

También es importante mencionar que existen diferentes métodos jerárquicos de clustering aglomerativos cuya elección sobre el resto dependerá en gran parte de la experiencia con que se cuente realizando este tipo de análisis y aprendizaje no supervisado, además del propio planteamiento del problema a resolver. Estos son:

1. Distancia mínima o similitud máxima (single-linkage): La similitud entre dos grupos viene dada por la mínima distancia entre sus componentes.

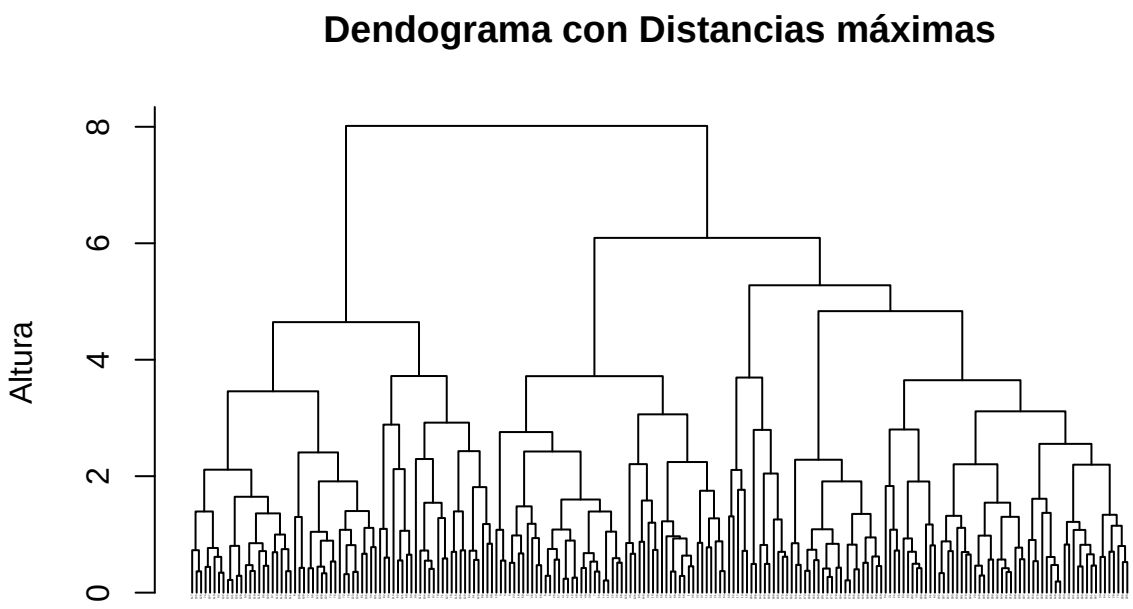
2. Distancia máxima o similitud mínima (complete-linkage): Se considera que la distancia entre dos grupos hay que medirla con respecto a sus elementos más dispares, esto es, la distancia máxima entre sus componentes.
3. Similitud ponderada (weighted arithmetic average): La distancia entre el cluster C_i y el C_j se obtiene como la media arimética de las distancias entre los componentes de los clusters en cuestión.

La representación la daremos por medio de **dendogramas** ("dibujos de árbol") para poder visualizar el agrupamiento que se fue dando paulatinamente.

Distancia máxima

Creamos ahora el dendograma luego de un clustering jerárquico aglomerativo con la técnica o estrategia de la mínima simulitud (distancia máxima).

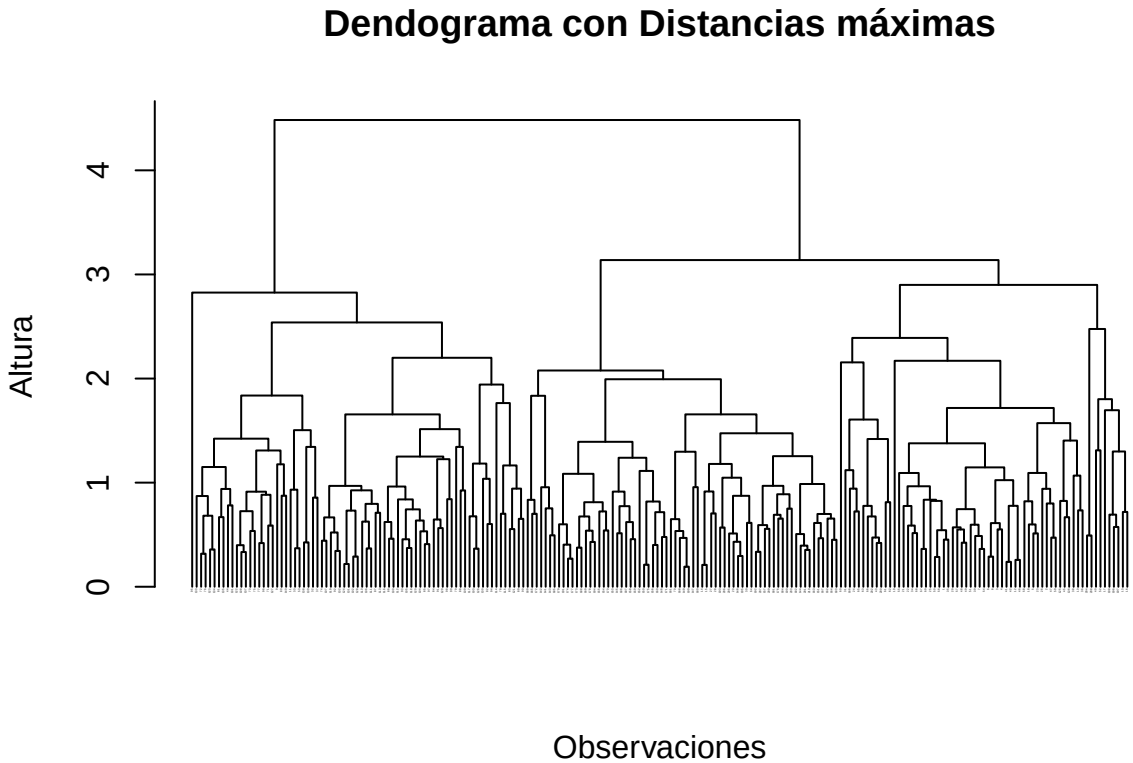
```
# R proporciona y la función para llevar a cabo el algoritmo aglomerativo jerárquico.
hc.c <- hclust(distancias)
# Graficando luego de etiquetar los ejes y reducir el tamaño de las etiquetas de observaciones.
plot(hc.c, xlab="Observaciones",ylab="Altura",main="Dendograma con Distancias máximas",sub="",cex = 0.1
```



Tenemos en este caso muchas observaciones (210), pero podríamos tener un peor panorama. La complejidad del algoritmo de agrupamiento aglomerativo es del orden en tiempo $\mathcal{O}(n^3)$ con n el número de datos. En este caso, se necesitaron ocho iteraciones para que todas observaciones confluyeran en un solo cluster.

Similitud ponderada

```
# R proporciona y la función para llevar a cabo el algoritmo aglomerativo jerárquico.
hc.a <- hclust(distancias, method = "average")
# Graficando luego de etiquetar los ejes y reducir el tamaño de las etiquetas de observaciones.
plot(hc.a,xlab="Observaciones",ylab="Altura",main="Dendograma con Distancias máximas",sub="",cex = 0.1,
```



Podemos ver que la creación de los clusters usando la estrategia de distnacias ponderadas fue similar, pero terminó antes el agrupamiento.

Resultados

Comparativa de ténicas aglomerativas

Conviene entonces hacer un análisis cuantitativo de qué técnica de clustering jerárquico nos conviene más, sabiendo ya de antemano que en este caso y para este problema nos conviene tener tres clusters (pues hay tres clases distintas de semillas de trigo) como ya discutimos en la parte de **k-means** tras graficar las incercias inter-cluster.

```
# viendo cómo se agrupan en 3 cluster con cada técnica.
distancias_maximas <- cutree(hc.c, 3)
similitud_ponderada <- cutree(hc.a, 3)
# Tabla comparativa.
table(distancias_maximas, similitud_ponderada)
```

```
##                   similitud_ponderada
## distancias_maximas  1  2  3
##                  1 43  7  2
##                  2  0 68  0
##                  3 22  0 68
```

Esto nos indica que hubo:

- Para el caso de las similitudes ponderadas:
  - Un total de $43 + 0 + 22 = 65$ observaciones en el primer cluster.
  - Un total de $7 + 68 + 0 = 75$ observaciones en el segundo cluster.
  - El resto de las observaciones, que fueron $2 + 0 + 68 = 70$ fueron agrupadas en el tercer cluster.

- Para el caso de clustering con distancias máximas:
  - Un total de $43 + 7 + 2 = 52$ obervaciones en el primer cluster.
  - Un total de $0 + 68 + 0 = 68$ observaciones en el segundo cluster.
  - Un total de $22 + 0 + 68 = 90$ observaciones en el tercer cluster.

De acuerdo con estos resultados de clasificación se puede realizar un análisis de cómo se están comportando estas dos técnicas de clustering alglomerativo jerárquicas. Los resultados fueron:

- 48 observaciones fueron consideradas para el primer cluster en ambas técnicas, 68 para el segundo y otras 68 para el tercer cluster.
- 7 observaciones fueron consideras por la técnica de similitud ponderada como del segundo cluster, pero puestas por la técnica de distancias máximas en el primer cluster.
- Tan sólo dos observaciones que fueron colocadas por la similitud ponderda en el tercer cluster fueron consideras como del primero por distancias máximas.
- 22 de las observaciones fueron catalogadas como del primer cluster por similitud ponderada fueron consideradas a su vez como del tercer cluster por la técnica de distancias máximas. Este fue el desajuste más notable entre ambas técnicas.

Debido al desajuste, nos lleva a concluir que es complicado diferenciar entre el primer y tercer cluster, siendo que con respecto al primero y segundo hubo un desajuste menor.

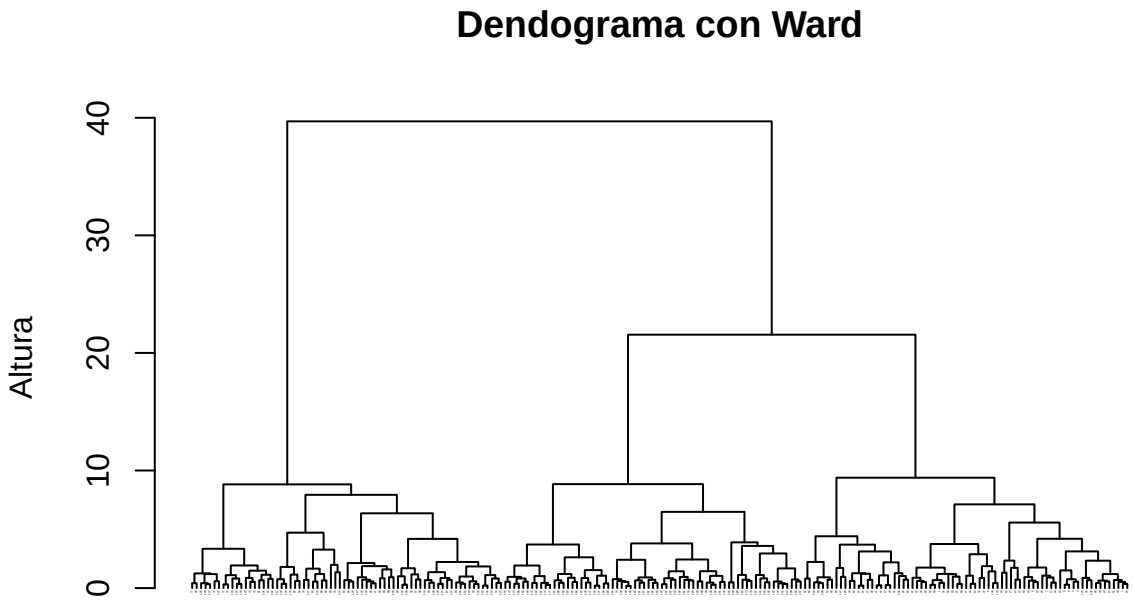
Una técnica de clustering alglomerativo más

Técnica de Ward

El método de clustering aglomerativo de Ward fue introducida en el año de 1983 tras una publicación de J.H. Ward y se reduce al error de la suma de cuadrados que ha sido generalizada de muchas maneras.

A continuación vamos a usarla con el cuadrado de las distancias obtenidas antes y comprarla con las dos ténicas usadas antes.

```
# Este método (dado por la función de R) ya eleva las distancias al cuadrado.
hc.w <- hclust(distancias, method = "ward.D2")
plot(hc.w, xlab="Observaciones",ylab="Altura",main="Dendograma con Ward",sub="",cex = 0.1,hang = -1)
```



Notemos que le tomó más tiempo encontrar el cluster que agrupa a todas las observaciones, pero los clusters se ven miu bien delimitados.

Comparativa con Distancias Máximas

```
# Obtenemos cómo argupa en tres clusters con Ward.
ward <- cutree(hc.w, 3)
# Tabla comparativa con distancias máximas.
table(distancias_maximas, ward)
```

```
##                   ward
## distancias_maximas  1  2  3
##                  1 47  5  0
##                  2  3 65  0
##                  3 23  0 67
```

De acuerdo con la tabla comparativa:

- Hubo un total de 47 observaciones que ambas ténicas identificaron como del cluster número uno, 65 del segundo y 67 del tercero.
- 5 observaciones que la ténica de Ward agrupó en el segundo cluster fueron agrupadas por la técnica de mínima similitud en el primer cluster.
- 23 observaciones que fueron agrupadas por Ward en el primer cluster, fueron agrupadas por distancias máximas en el tercer cluster.

Esto nos indica que si bien el primer y tercer cluster siguen causando el mayor desajuste, la diferencia entre la ténica de Ward y la de distancias máximas nos arrojaron una comparativa similar a la efectuada anteriormente entre distancias máximas y similitud ponderada, pero en este caso pudimos agrupar un número ligeramente superior de observaciones dentro del primer cluster en ambos casos.

Comparativa con Similitud Ponderada

```
# Tabla comparativa con similitud ponderada.
table(similitud_ponderada, ward)
```

```
##                     ward
## similitud_ponderada  1  2  3
##                   1 57  3  5
##                   2  8 67  0
##                   3  8  0 62
```

Se obtuvieron los siguientes resultados en la comparativa:

- Hubo 57 observaciones clasificadas por ambas ténicas como del primer cluster, 67 dentro del segundo y otras 62 dentro dle tercer cluster.
- 3 observaciones fueron agrupadas por la técnica de Ward en el segundo cluster y por similitud ponderada en el primero.
- 5 observaciones fueron colocadas por Ward en el tercer cluster y por similitud ponderada en el primero.
- 8 observaciones fueron consideradas por Ward dentro del primer vluster y por similitud ponderada dentro del segundo cluster.
- Hubo otro desajuste entre ambas técnicas de ocho observaciones que fueron consideradas por Ward dentro del primer cluster y por similitud ponderada dentro del tercero.

En etse caso sí hubo un resultado comparativo distinto, pues hubo un menor desajuste entre ambas ténicas en cuando a la diferenciación de las observaciones consideradas para el primero y tercer clusters.