Análisis de Conglomerados (Cluster Analysis)

José A. Perusquía Cortés

Análisis Multivariado Semestre 2023-2

¿De qué va?

- Clustering es el procesos de agrupar objetos similares buscando patrones en los datos.

¿De qué va?

- Clustering es el procesos de agrupar objetos similares buscando patrones en los datos.
- Técnica de aprendizaje no supervisado, i.e., a priori no necesitamos:

¿De qué va?

- Clustering es el procesos de agrupar objetos similares buscando patrones en los datos.

- Técnica de aprendizaje no supervisado, i.e., a priori no necesitamos:
  - Conocer el número de clusters (en algunas ocasiones se puede conocer)

¿De qué va?

- Clustering es el procesos de agrupar objetos similares buscando patrones en los datos.

- Técnica de aprendizaje no supervisado, i.e., a priori no necesitamos:
  - Conocer el número de clusters (en algunas ocasiones se puede conocer)
  - Un grupo de observaciones etiquetadas (training data)

¿De qué va?

- Clustering es el procesos de agrupar objetos similares buscando patrones en los datos.

- Técnica de aprendizaje no supervisado, i.e., a priori no necesitamos:
  - Conocer el número de clusters (en algunas ocasiones se puede conocer)
  - Un grupo de observaciones etiquetadas (training data)

- Dos tipos de métodos
  - Si no conocemos el número de clusters tenemos métodos jerárquicos aglomerativos y divisivos.

¿De qué va?

- Clustering es el procesos de agrupar objetos similares buscando patrones en los datos.

- Técnica de aprendizaje no supervisado, i.e., a priori no necesitamos:
  - Conocer el número de clusters (en algunas ocasiones se puede conocer)
  - Un grupo de observaciones etiquetadas (training data)

- Dos tipos de métodos
  - Si no conocemos el número de clusters tenemos métodos jerárquicos aglomerativos y divisivos.
  - Si conocemos el número de clusters creamos particiones (cada objeto pertenece a un cluster) o métodos "fuzzy" (cada objeto puede pertenecer a varios clusters).

¿Cómo se define un cluster?

¿Cómo se define un cluster?

- Necesitamos definir una noción de cercanía
  - Matriz de distancias
  - Matriz de disimilitudes
  - Matriz de similitudes

¿Cómo se define un cluster?

- Necesitamos definir una noción de cercanía
  - Matriz de distancias
  - Matriz de disimilitudes
  - Matriz de similitudes

- En la práctica es común utilizar
  - Distancia euclidiana
  - Distancia Manhattan
  - Distancia Mahalanobis

Métodos jerárquicos aglomerativos (AGNES)

Métodos Aglomerativos

- **¿Qué necesitamos?**
  - Una matriz de proximidades (e.g. distancias, disimilitudes).
  - Medida de distancia entre clusters.

Métodos Aglomerativos

- **¿Qué necesitamos?**

  - Una matriz de proximidades (e.g. distancias, disimilitudes).
  - Medida de distancia entre clusters.

- **Idea**

  Crear un árbol de clusters empezando con n clusters de una sola observación y e irlos uniendo por cercanía.

Métodos Aglomerativos

- **¿Qué necesitamos?**

  - Una matriz de proximidades (e.g. distancias, disimilitudes).
  - Medida de distancia entre clusters.

- **Idea**

  Crear un árbol de clusters empezando con n clusters de una sola observación y e irlos uniendo por cercanía.

- **¿Cómo medir la distancia entre clusters?**

Vecino más cercano

- También conocido como single linkage method [ver: Sneath (1957), Sokal y Sneath (1963), Johnson (1967)]

Vecino más cercano

- También conocido como single linkage method [ver: Sneath (1957), Sokal y Sneath (1963), Johnson (1967)]

- Dados dos clusters C_i y C_j la distancia entre ellos es la disimilitud más pequeña entre uno sus miembros, i.e.

$$d\left(C_i, C_j\right) = \min\left\{d_{rs} : r \in C_i, s \in C_j\right\}$$

Vecino más cercano

- También conocido como single linkage method [ver: Sneath (1957), Sokal y Sneath (1963), Johnson (1967)]

- Dados dos clusters C_i y C_j la distancia entre ellos es la disimilitud más pequeña entre uno sus miembros, i.e.

$$d\left(C_i, C_j\right) = \min\left\{d_{rs} : r \in C_i, s \in C_j\right\}$$

- **Algoritmo**
  - Buscar la disimilitud más pequeña entre clusters.
  - Recalcular la matriz de disimilitudes

Vecino más lejano

- También conocido como complete linkage method [ver: Sokal y Sneath (1963), McQuitty (1964)]

Vecino más lejano

- También conocido como complete linkage method [ver: Sokal y Sneath (1963), McQuitty (1964)]

- Dados dos clusters C_i y C_j la distancia entre ellos es la disimilitud más grande entre uno sus miembros, i.e.

$$d\left(C_i, C_j\right) = \max\left\{d_{rs} : r \in C_i, s \in C_j\right\}$$

Vecino más lejano

- También conocido como complete linkage method [ver: Sokal y Sneath (1963), McQuitty (1964)]

- Dados dos clusters C_i y C_j la distancia entre ellos es la disimilitud más grande entre uno sus miembros, i.e.

$$d\left(C_i, C_j\right) = \max\left\{d_{rs} : r \in C_i, s \in C_j\right\}$$

- **Algoritmo**
  - Buscar la disimilitud más pequeña entre clusters.
  - Recalcular la matriz de disimilitudes

Centroide

- Dados dos clusters C_i y C_j se define la distancia entre ellos como la "distancia" entre sus centroides [ver: Sokal y Michener (1958), King (1966,1967)].

Centroide

- Dados dos clusters C_i y C_j se define la distancia entre ellos como la "distancia" entre sus centroides [ver: Sokal y Michener (1958), King (1966,1967)].

$$\bar{\mathbf{X}}_i = \sum_{n \in C_i} \frac{\mathbf{X}_n}{n_i} \qquad \bar{\mathbf{X}}_j = \sum_{m \in C_j} \frac{\mathbf{X}_n}{n_j} \qquad \longrightarrow \qquad d\left(C_i, C_j\right) = \delta\left(\bar{\mathbf{X}}_i, \bar{\mathbf{X}}_j\right)$$

Centroide

- Dados dos clusters C_i y C_j se define la distancia entre ellos como la "distancia" entre sus centroides [ver: Sokal y Michener (1958), King (1966,1967)].

$$\bar{\mathbf{X}}_i = \sum_{n \in C_i} \frac{\mathbf{X}_n}{n_i} \qquad \bar{\mathbf{X}}_j = \sum_{m \in C_j} \frac{\mathbf{X}_n}{n_j} \longrightarrow d\left(C_i, C_j\right) = \delta\left(\bar{\mathbf{X}}_i, \bar{\mathbf{X}}_j\right)$$

- **Algoritmo**
  - Buscar la disimilitud más pequeña entre clusters.
  - Recalcular el centroide

Centroide

- Dados dos clusters C_i y C_j se define la distancia entre ellos como la "distancia" entre sus centroides [ver: Sokal y Michener (1958), King (1966,1967)].

$$\bar{\mathbf{X}}_i = \sum_{n \in C_i} \frac{\mathbf{X}_n}{n_i} \qquad \bar{\mathbf{X}}_j = \sum_{m \in C_j} \frac{\mathbf{X}_n}{n_j} \longrightarrow d\left(C_i, C_j\right) = \delta\left(\bar{\mathbf{X}}_i, \bar{\mathbf{X}}_j\right)$$

- **Algoritmo**
  - Buscar la disimilitud más pequeña entre clusters.
  - Recalcular el centroide

$$\bar{\mathbf{X}}_{Ci \cup C_j} = \frac{n_i \bar{\mathbf{X}}_i + n_j \bar{\mathbf{X}}_j}{n_i + n_j}$$

Ward

- También conocido como incremental sum of squares method [ver: Wishart (1969a)] basado en la idea de Ward (1963).

Ward

- También conocido como incremental sum of squares method [ver: Wishart (1969a)] basado en la idea de Ward (1963).

- **Algoritmo**
  - Unir los clusters que minimicen

$$I_{C_iC_j} = \sum_{k \in C_i \cup C_j} ||\mathbf{X}_k - \bar{\mathbf{X}}||^2 - \left[\sum_{n \in C_i} ||\mathbf{X}_n - \bar{\mathbf{X}}_i||^2 + \sum_{m \in C_j} ||\mathbf{X}_m - \bar{\mathbf{X}}_j||^2 \right] = \frac{n_i n_j}{n_i + n_j} ||\bar{\mathbf{X}}_i - \bar{\mathbf{X}}_j||^2$$

Ward

- También conocido como incremental sum of squares method [ver: Wishart (1969a)] basado en la idea de Ward (1963).

- **Algoritmo**
  - Unir los clusters que minimicen

$$I_{C_iC_j} = \sum_{k \in C_i \cup C_j} ||\mathbf{X}_k - \bar{\mathbf{X}}||^2 - \left[\sum_{n \in C_i} ||\mathbf{X}_n - \bar{\mathbf{X}}_i||^2 + \sum_{m \in C_j} ||\mathbf{X}_m - \bar{\mathbf{X}}_j||^2 \right] = \frac{n_i n_j}{n_i + n_j} ||\bar{\mathbf{X}}_i - \bar{\mathbf{X}}_j||^2$$

  - En particular para dos observaciones r, s

$$I_{rs} = \frac{1}{2} ||\mathbf{X}_r - \mathbf{X}_s||^2 = \frac{1}{2} d_{rs}^2$$

Promedio

- También conocido como group average method [ver: Sokal y Michener (1958), McQuitty (1964), Lance y Williams (1966)].

Promedio

- También conocido como group average method [ver: Sokal y Michener (1958), McQuitty (1964), Lance y Williams (1966)].

- Dados dos clusters C_i y C_j la distancia entre ellos se define como el promedio de las distancias de sus miembros

$$d\left(C_i, C_j\right) = \frac{1}{n_i n_j} \sum_{n \in C_i} \sum_{m \in C_j} d_{nm}$$

Lance y Williams

- También conocido como Lance and Williams Flexible Method [ver: Lance y Williams (1967a)].

Lance y Williams

- También conocido como Lance and Williams Flexible Method [ver: Lance y Williams (1967a)].

- Dados tres clusters C_i , C_j y C_k definimos la distancia de C_k y $C_i \cup C_j$ como

$$d\left(C_k, C_i \cup C_j\right) = \alpha_1 d(C_k, C_i) + \alpha_2 d(C_k, C_j) + \beta d(C_i, C_j) + \gamma \left| d(C_k, C_i) - d(C_k, C_j) \right|$$

Lance y Williams

- También conocido como Lance and Williams Flexible Method [ver: Lance y Williams (1967a)].

- Dados tres clusters C_i , C_j y C_k definimos la distancia de C_k y $C_i \cup C_j$ como

$$d\left(C_k, C_i \cup C_j\right) = \alpha_1 d(C_k, C_i) + \alpha_2 d(C_k, C_j) + \beta d(C_i, C_j) + \gamma \left| d(C_k, C_i) - d(C_k, C_j) \right|$$

- Casos particulares: vecino más cercano, vecino más lejano, centroide, Ward y promedio.

Lance y Williams

- También conocido como Lance and Williams Flexible Method [ver: Lance y Williams (1967a)].

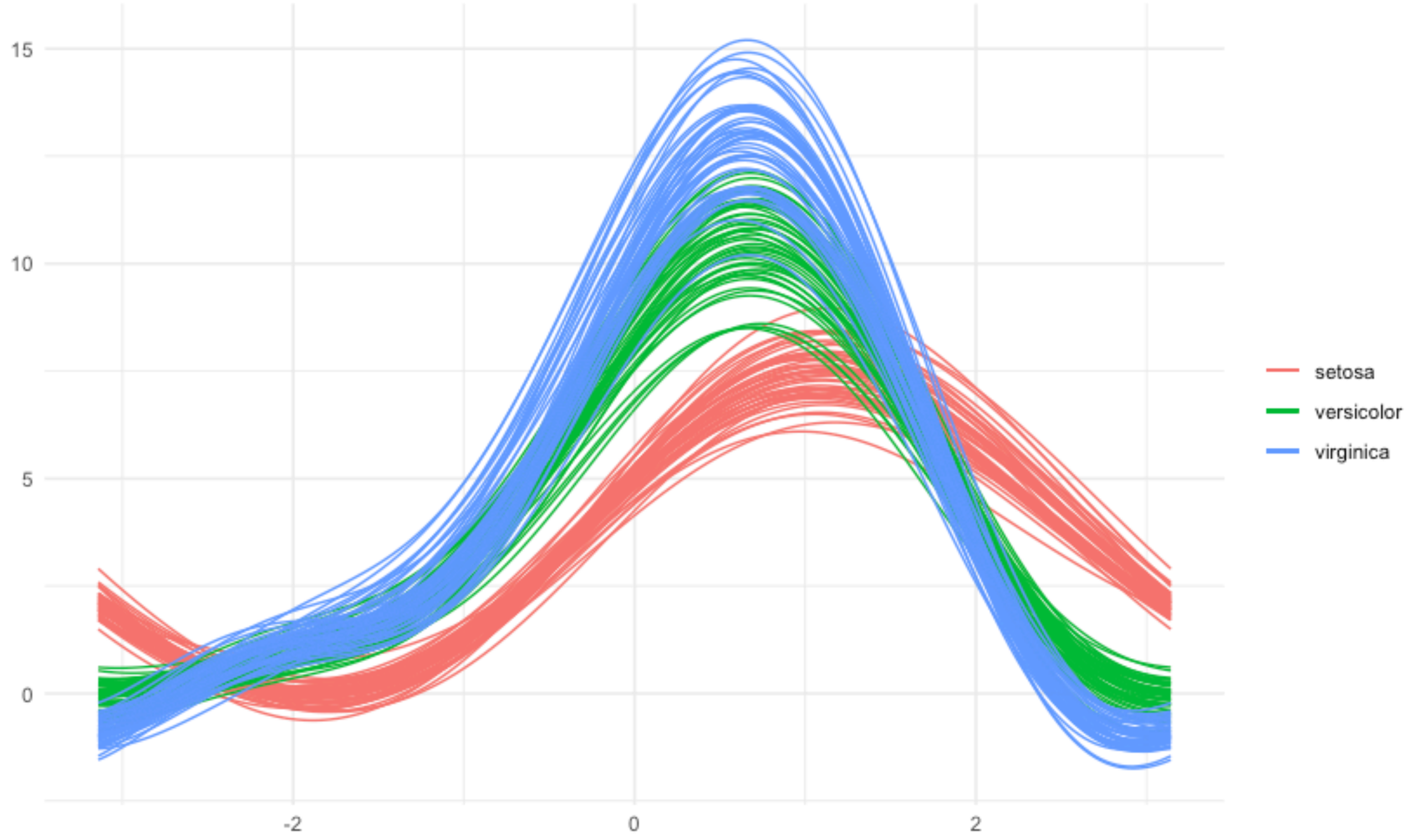
- Dados tres clusters C_i , C_j y C_k definimos la distancia de C_k y $C_i \cup C_j$ como

$$d\left(C_k, C_i \cup C_j\right) = \alpha_1 d(C_k, C_i) + \alpha_2 d(C_k, C_j) + \beta d(C_i, C_j) + \gamma \left| d(C_k, C_i) - d(C_k, C_j) \right|$$

- Casos particulares: vecino más cercano, vecino más lejano, centroide, Ward y promedio.

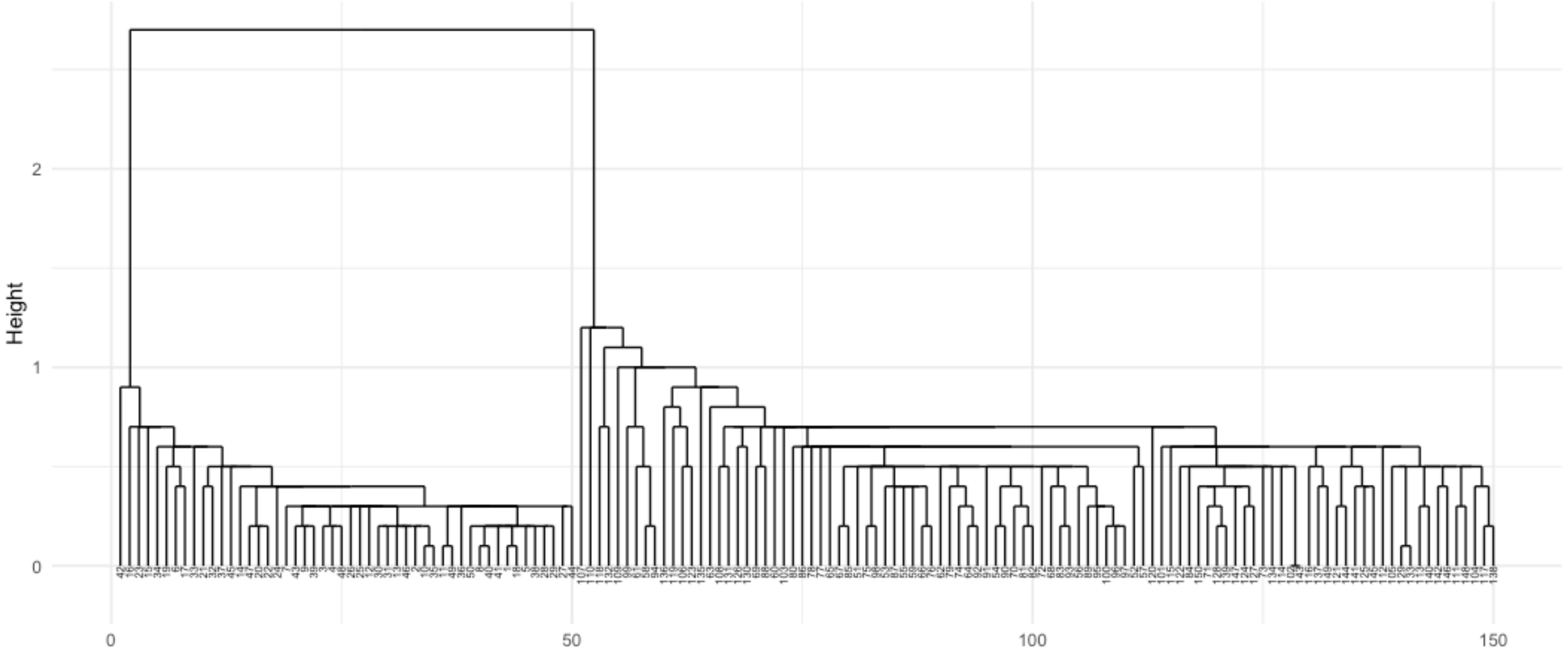
- Lance y Williams sugieren $\alpha_1 = \alpha_2$, $\beta < 1$, $\alpha_1 + \alpha_2 + \beta = 1$, $\gamma = 0$

Ejemplo: Iris

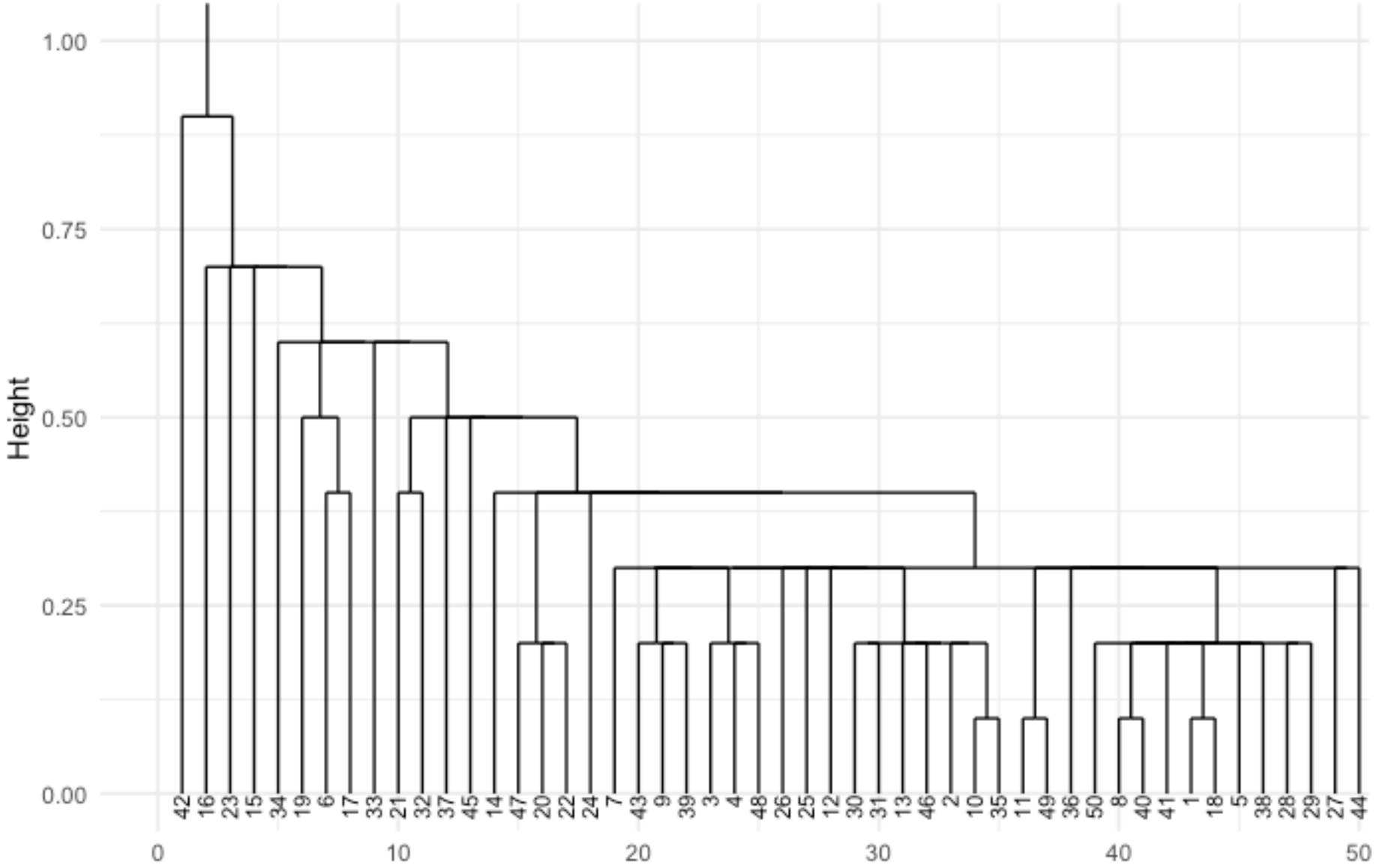


Ejemplo: Iris

- Vecino más cercano

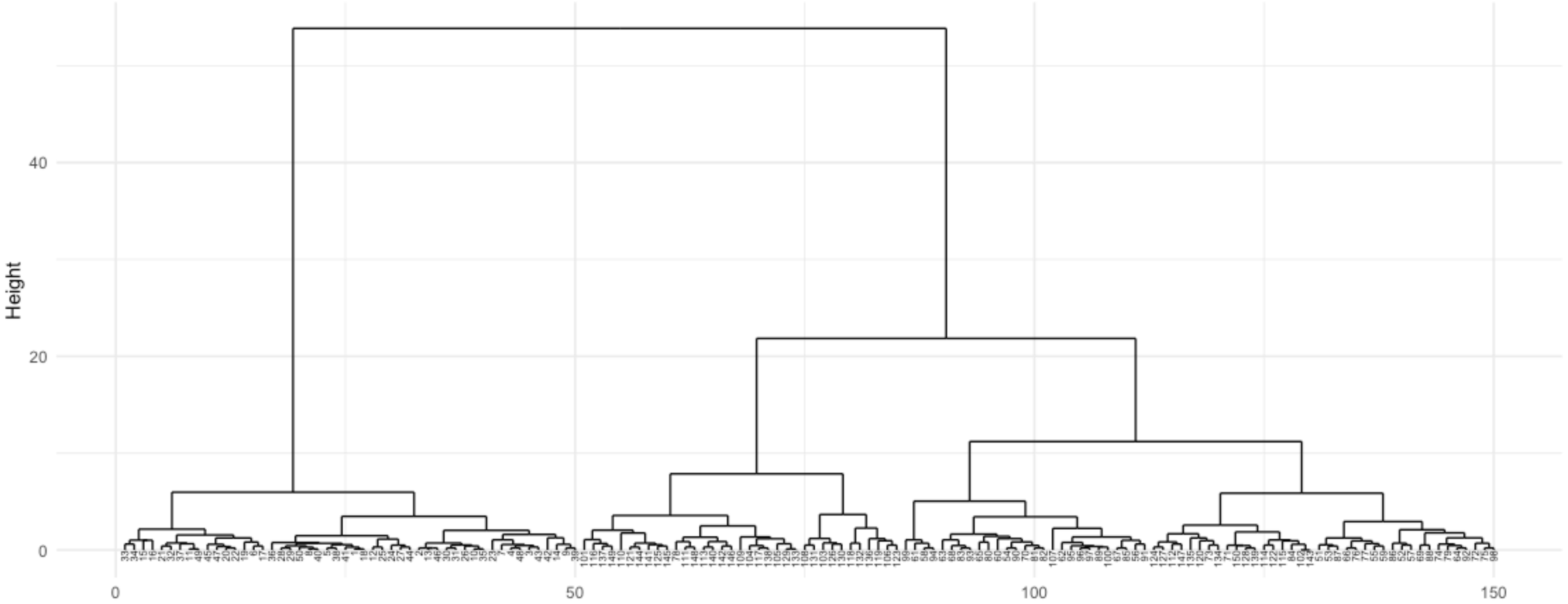


Ejemplo: Iris

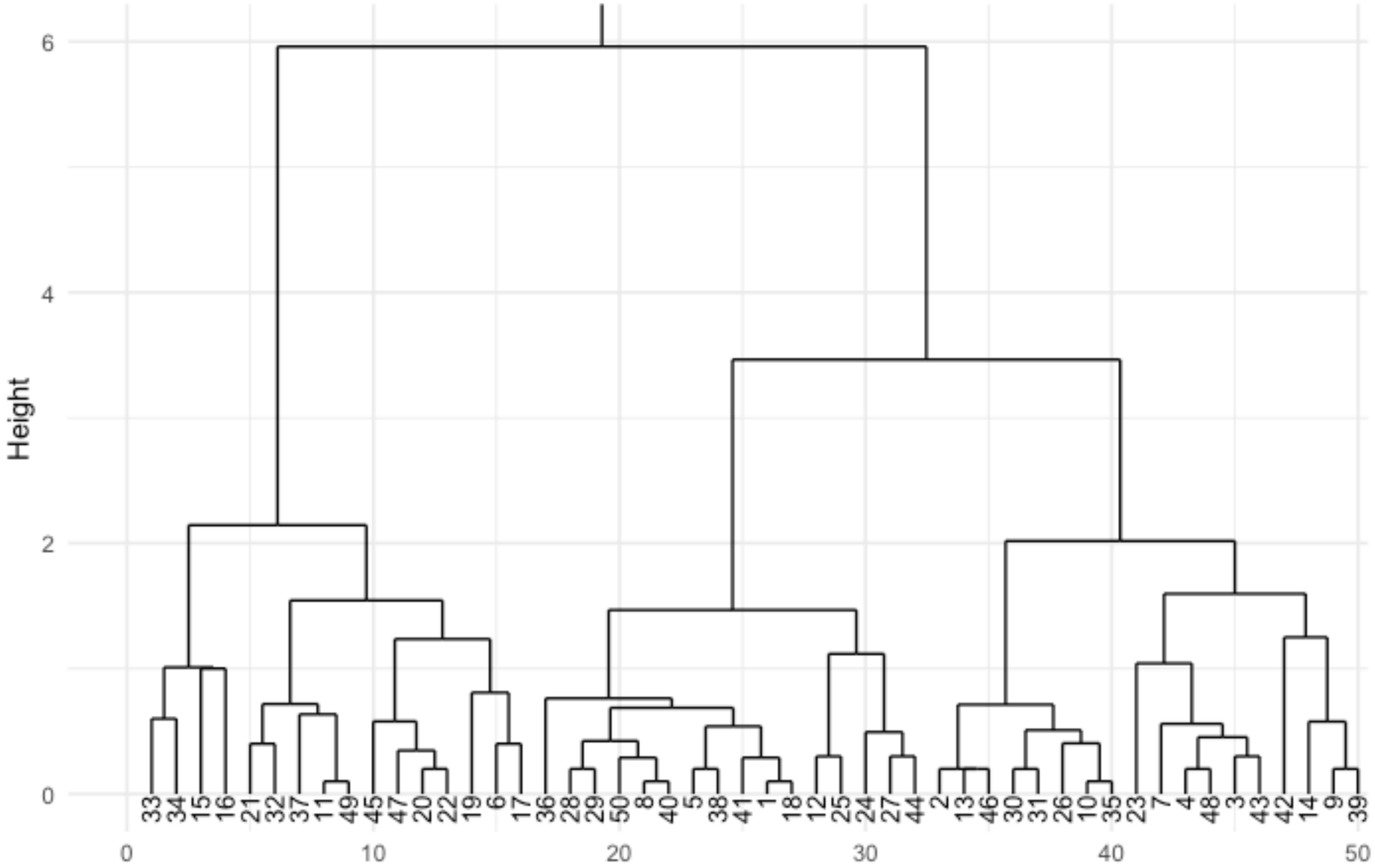


Ejemplo: Iris

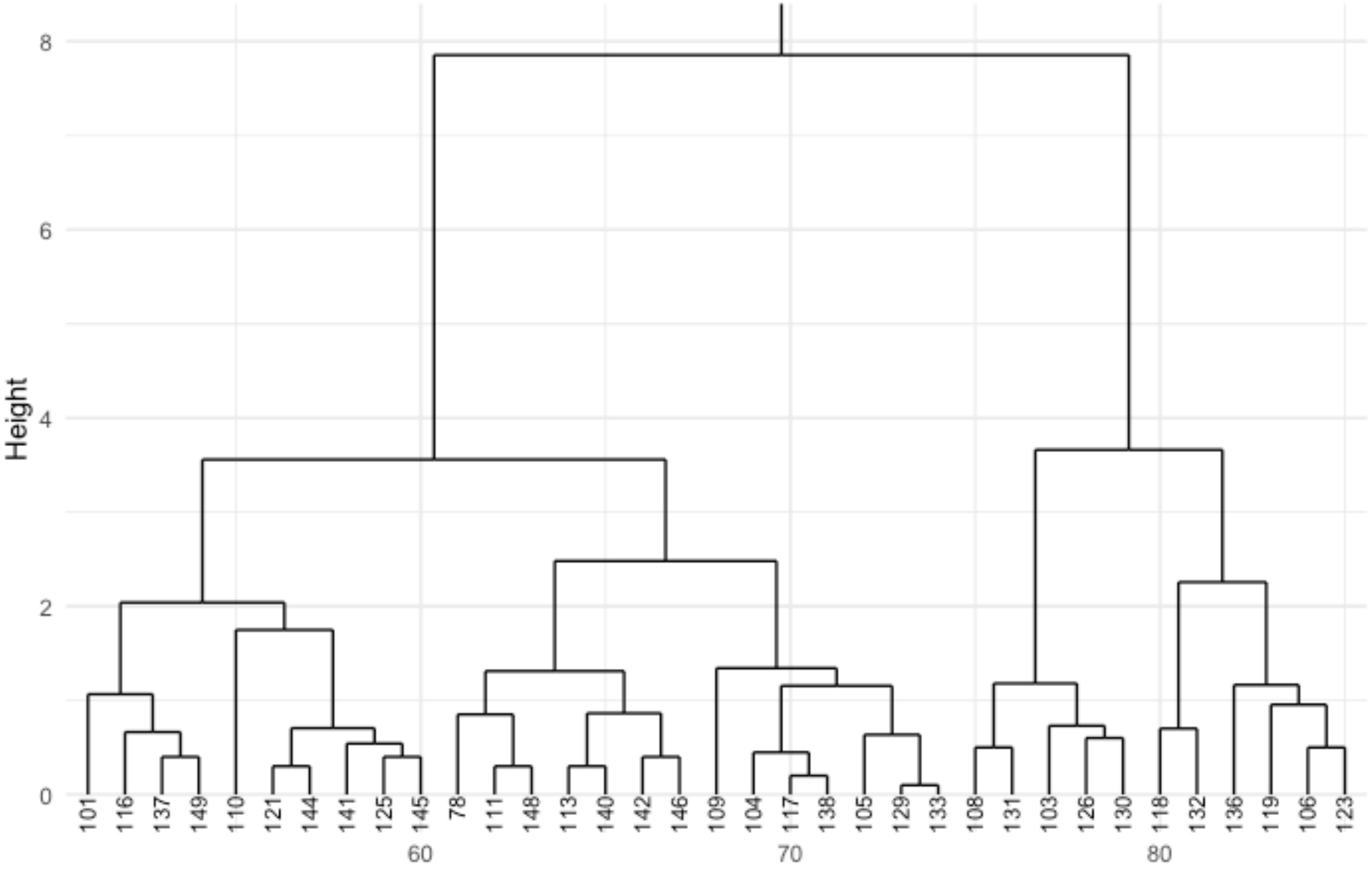
- Ward



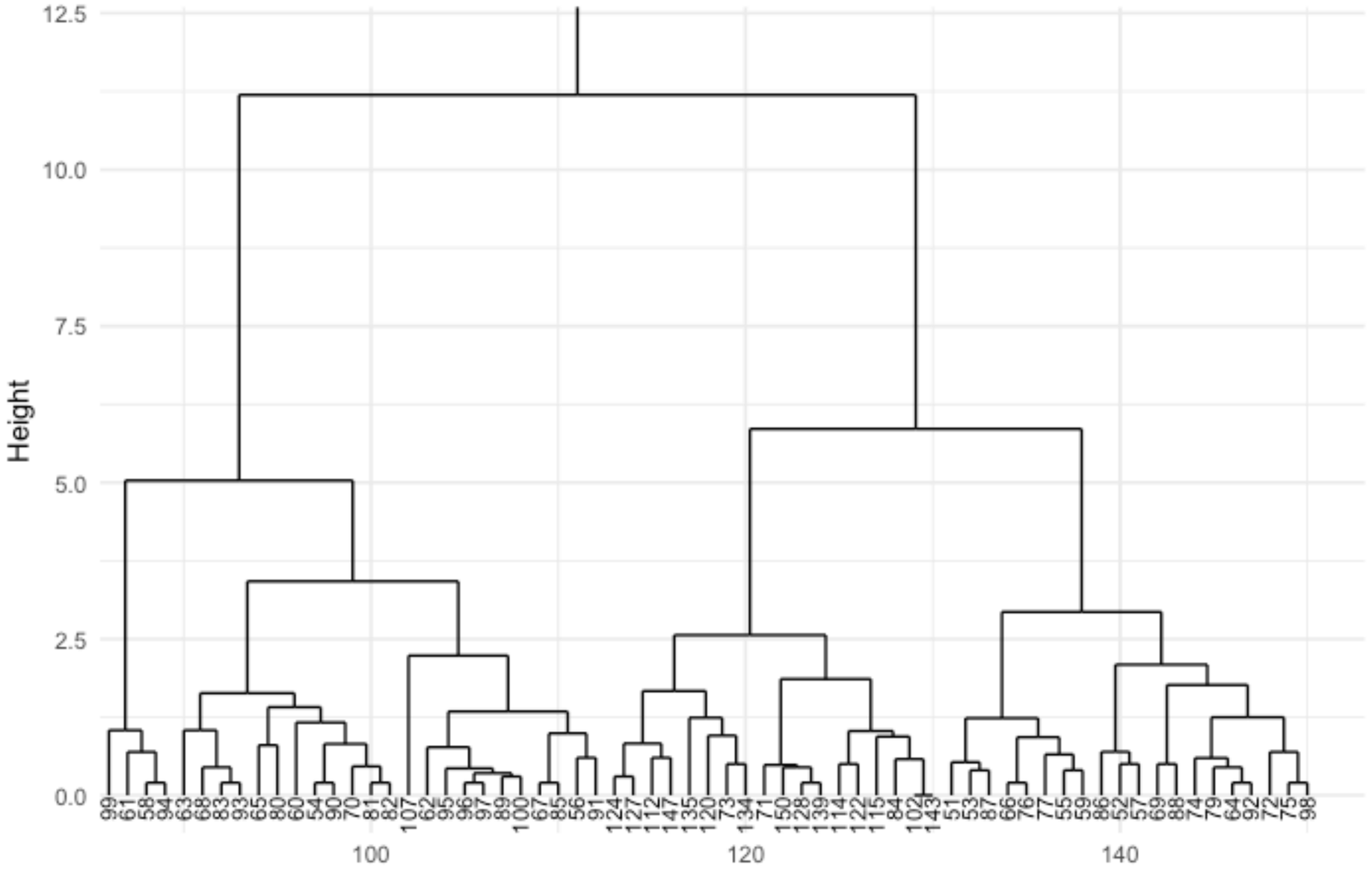
Ejemplo: Iris



Ejemplo: Iris



Ejemplo: Iris



Métodos jerárquicos divisivos (DIANA)

Métodos Divisivos

- **Idea**

  Empezar con un cluster de tamaño n e irlo dividiendo.

Métodos Divisivos

- **Idea**

  Empezar con un cluster de tamaño n e irlo dividiendo.

- **Ventajas** (Williams y Lance, 1977)
  - El proceso empieza con el contenido máximo de información.
  - La división no tiene que continuar hasta tener n clusters.

Métodos Divisivos

- **Idea**

  Empezar con un cluster de tamaño n e irlo dividiendo.

- **Ventajas** (Williams y Lance, 1977)
  - El proceso empieza con el contenido máximo de información.
  - La división no tiene que continuar hasta tener n clusters.

- **Restricción**
  - $2^{n-1} - 1$ formas de separar n objetos en 2 grupos ... imposible analizar todos los casos

Mono-variables

- **Objetivo**

  La división se basa en una sola variable

Mono-variables

- **Objetivo**

  La división se basa en una sola variable

- **Problemas**
  - Sensible a outliers
  - Difícil de adaptar con una mezcla de variables cuantitativas y cualitativas
  - Errores frecuentes de clasificación

Mono-variables

- Si todas las variables son dicotómicas:

  - Dividimos las observaciones en dos grupos dependiendo si tienen el atributo A

Mono-variables

- Si todas las variables son dicotómicas:

  - Dividimos las observaciones en dos grupos dependiendo si tienen el atributo A

- **¿Cómo elegimos a A?**

Mono-variables

- Si todas las variables son dicotómicas:
  - Dividimos las observaciones en dos grupos dependiendo si tienen el atributo A

- **¿Cómo elegimos a A?**
  - Maximice alguna medida de distancia entre dos grupos e.g.: estadístico χ^2

- Si todas las variables son dicotómicas:
  - Dividimos las observaciones en dos grupos dependiendo si tienen el atributo A

- **¿Cómo elegimos a A?**
  - Maximice alguna medida de distancia entre dos grupos e.g.: estadístico χ^2

Var. r\ Var. s	Presente	Ausente	Total
Presente	a	b	a+b
Ausente	c	d	c+d
Total	a+c	b+d	a+b+c+d

$$\chi^2_{rs} = \frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

- Si todas las variables son dicotómicas:
  - Dividimos las observaciones en dos grupos dependiendo si tienen el atributo A

- **¿Cómo elegimos a A?**
  - Maximice alguna medida de distancia entre dos grupos e.g.: estadístico χ^2

Var. r\ Var. s	Presente	Ausente	Total
Presente	a	b	a+b
Ausente	c	d	c+d
Total	a+c	b+d	a+b+c+d

$$\chi^2_{rs} = \frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

Elegimos A como la que maximice: $\sum_{j:j\neq A} \chi^2_{jA}$

- Para variables cuantitativas:
  - Dividir de tal forma que se minimice la suma de cuadrados dentro del grupo (within-group) o maximizar la suma de cuadrados entre grupos (between-group)

$$B = n_1 \left(\bar{x}_1 - \bar{x}\right)^2 + n_2 \left(\bar{x}_1 - \bar{x}\right)^2 = n_1 \bar{x}_1^2 + n_2 \bar{x}_2^2 - n\bar{x}^2$$

- Para variables cuantitativas:
  - Dividir de tal forma que se minimice la suma de cuadrados dentro del grupo (within-group) o maximizar la suma de cuadrados entre grupos (between-group)

$$B = n_1 \left(\bar{x}_1 - \bar{x}\right)^2 + n_2 \left(\bar{x}_1 - \bar{x}\right)^2 = n_1 \bar{x}_1^2 + n_2 \bar{x}_2^2 - n\bar{x}^2$$

- Para no considerar todas las posibles divisiones se sugiere:
  - Ordenar los datos y hacer la división en la r donde se maximice

$$R = r\bar{x}_1^2 + (n - r)\bar{x}_2^2$$

Splinter

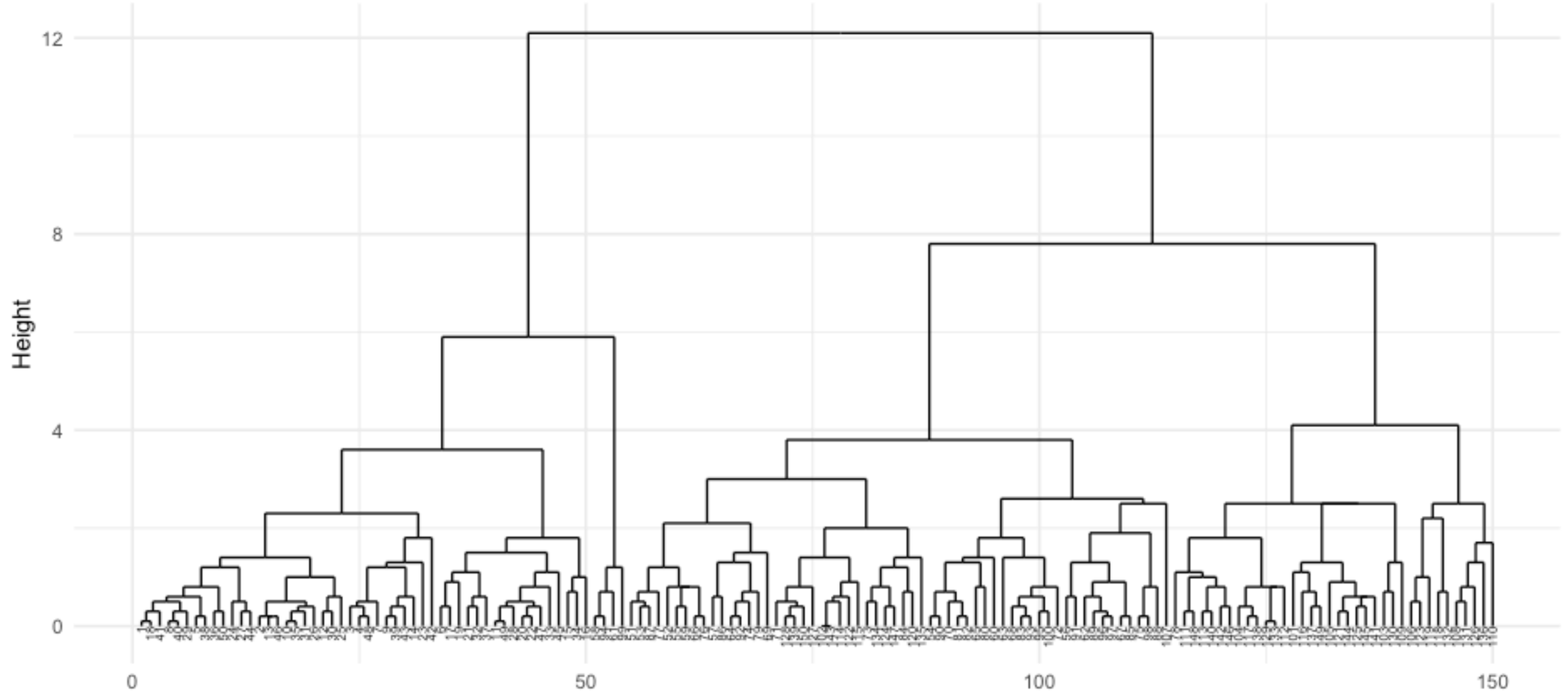
- **Objetivo**

  La división se basa elegir en cada paso a la observación más disimilar en promedio

- **Algoritmo**
  - Seleccionar el cluster más grande
  - Buscar la observación más disimilar en promedio
  - Empezar un nuevo grupo con esta observación
  - Reagrupar las observaciones dependiendo su disimilitud entre los miembros del grupo antiguo y el nuevo

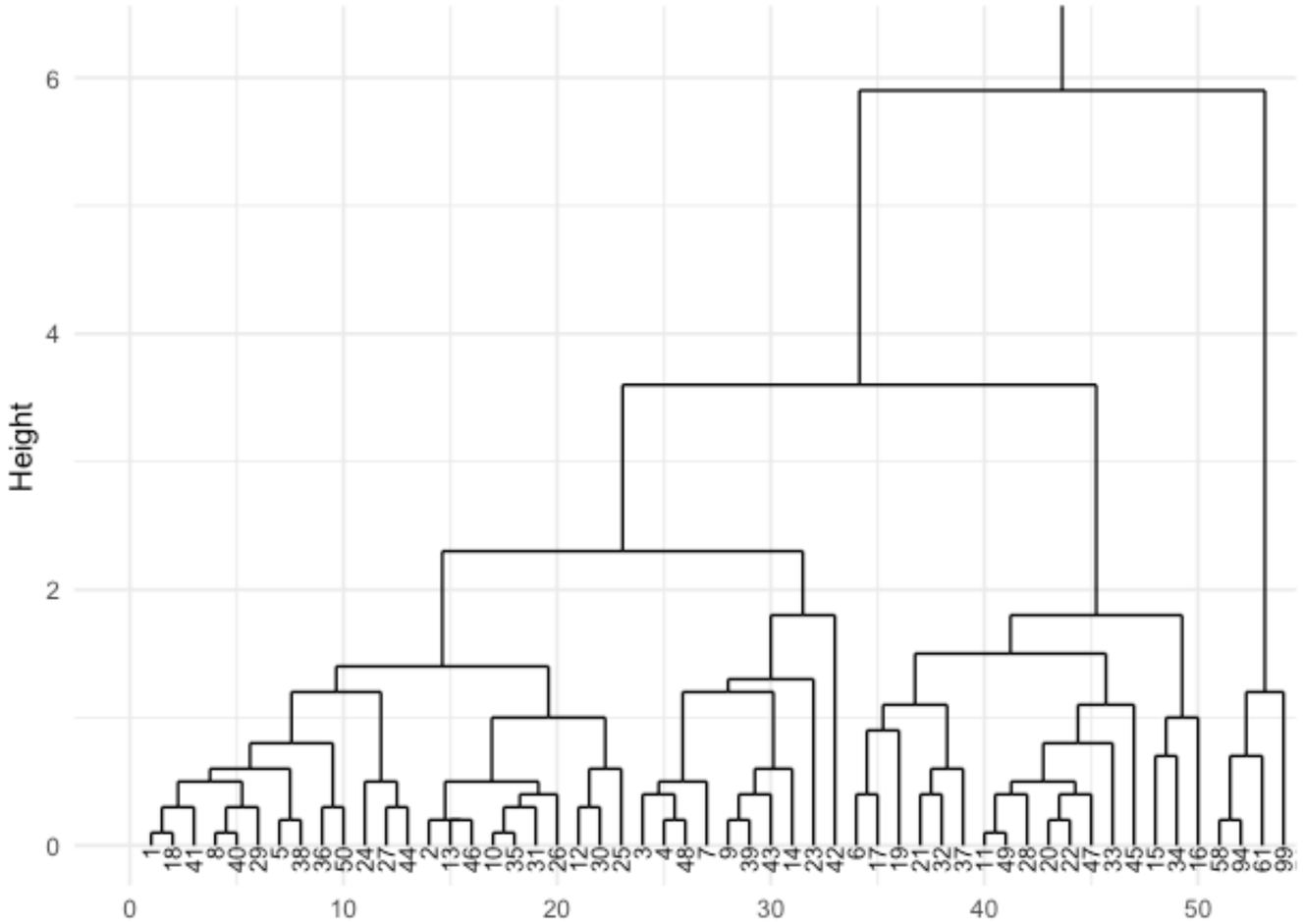
Ejemplo: Iris

Dendograma

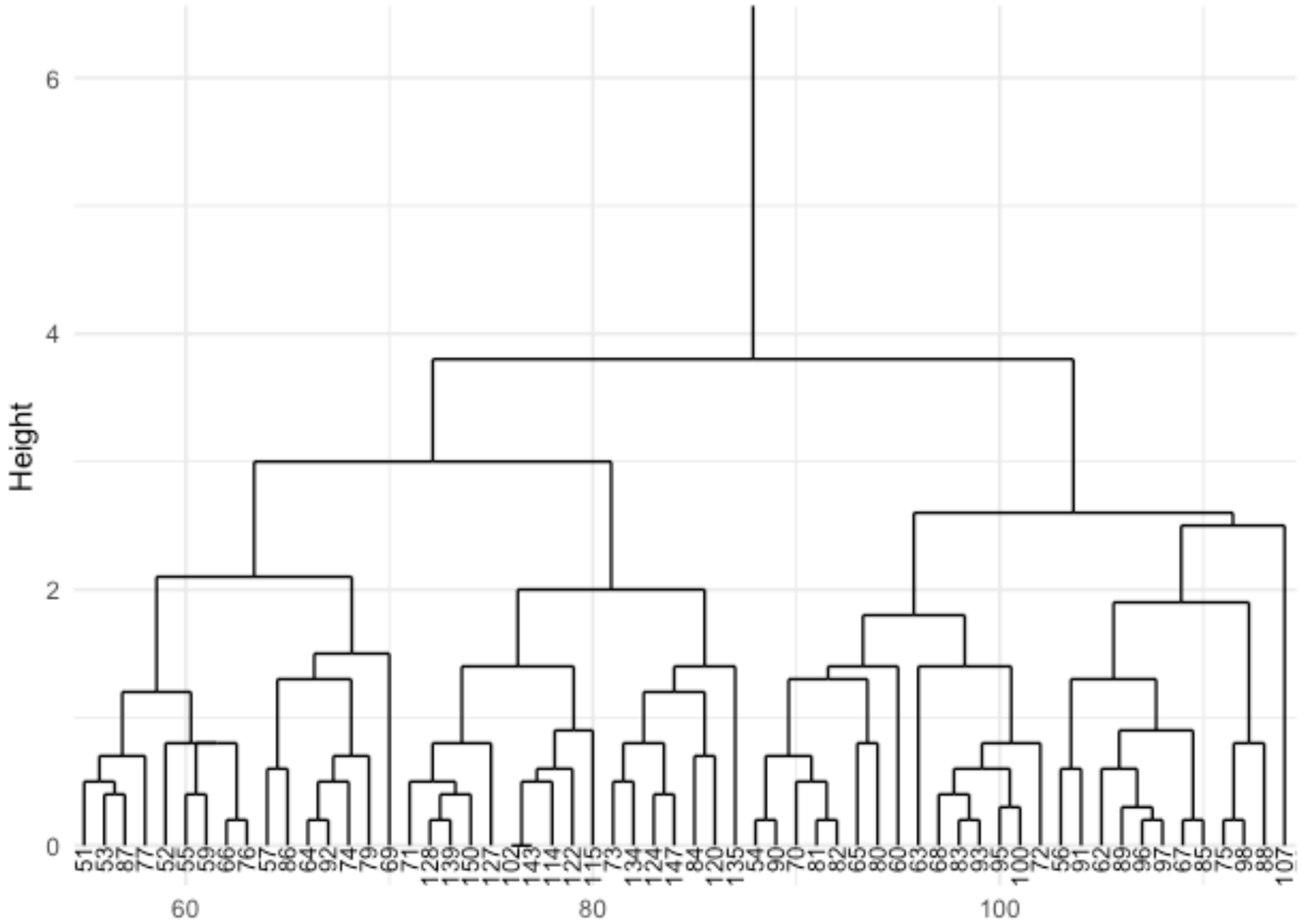


Ejemplo: Iris

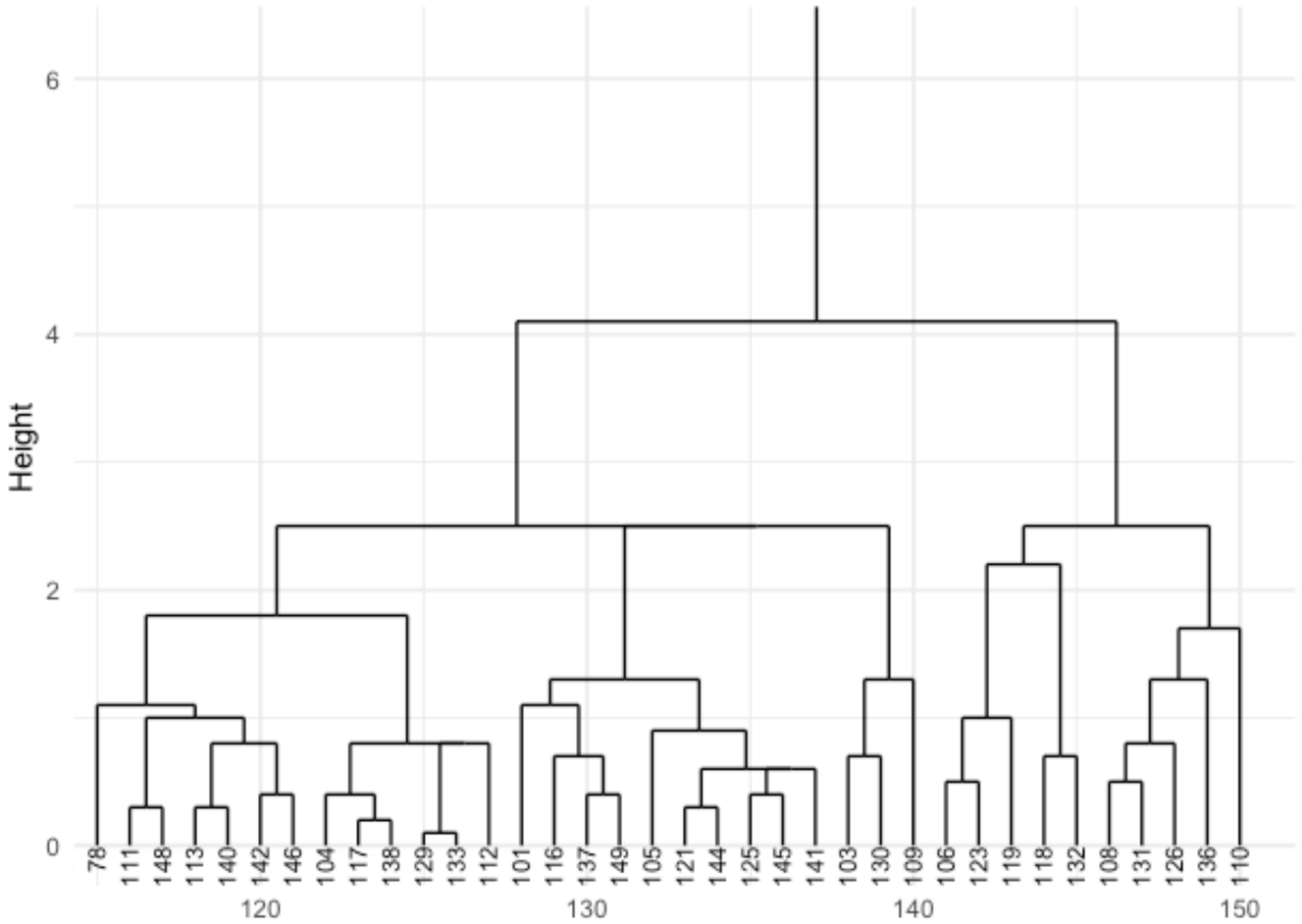
Primer grupo



Segundo grupo



Tercer grupo



Métodos de particiones

¿De qué va?

- Dado un número de clusters k se busca agrupar las n observaciones en estos clusters optimizando algún criterio.

¿De qué va?

- Dado un número de clusters k se busca agrupar las n observaciones en estos clusters optimizando algún criterio.

- **Dificultad**

  El número de formas de separar n objetos en k grupos está dado por el número de Sterling del segundo tipo:

$$S(n,k) = \left\{ \begin{matrix} n \\ k \end{matrix} \right\}$$

¿De qué va?

- Dado un número de clusters k se busca agrupar las n observaciones en estos clusters optimizando algún criterio.

- **Dificultad**

  El número de formas de separar n objetos en k grupos está dado por el número de Sterling del segundo tipo:

$$S(n,k) = \left\{ \begin{matrix} n \\ k \end{matrix} \right\}$$

- Por ejemplo: $S(16,8) = 2{,}141{,}764{,}053$ (Imposible de considerar todas las particiones)

¿ Cómo escoger k ?

- Usar método aglomerativo

¿ Cómo escoger k ?

- Usar método aglomerativo (no es ideal)

¿ Cómo escoger k ?

- Usar método aglomerativo (no es ideal)
- Usar algún modelos que permita reasignar las observaciones.

¿ Cómo escoger k ?

- Usar método aglomerativo (no es ideal)
- Usar algún modelos que permita reasignar las observaciones.

- **Algoritmo**

- Seleccionar k observaciones como los centroides de los clusters.

¿ Cómo escoger k ?

- Usar método aglomerativo (no es ideal)
- Usar algún modelos que permita reasignar las observaciones.

- **Algoritmo**

- Seleccionar k observaciones como los centroides de los clusters.
- Asignar el resto de las observaciones al cluster más cercano.

¿ Cómo escoger k ?

- Usar método aglomerativo (no es ideal)
- Usar algún modelos que permita reasignar las observaciones.

- **Algoritmo**

- Seleccionar k observaciones como los centroides de los clusters.
- Asignar el resto de las observaciones al cluster más cercano.
- Actualizar el centroide a cada paso (e.g. k-means) o hasta el final.

¿ Cómo escoger k ?

- Usar método aglomerativo (no es ideal)
- Usar algún modelos que permita reasignar las observaciones.

- **Algoritmo**

- Seleccionar k observaciones como los centroides de los clusters.
- Asignar el resto de las observaciones al cluster más cercano.
- Actualizar el centroide a cada paso (e.g. k-means) o hasta el final.
- Buscar objetos mal alocados y reasignar.

¿ Cómo escoger k ?

- Usar método aglomerativo (no es ideal)
- Usar algún modelos que permita reasignar las observaciones.

- **Algoritmo**

- Seleccionar k observaciones como los centroides de los clusters.
- Asignar el resto de las observaciones al cluster más cercano.
- Actualizar el centroide a cada paso (e.g. k-means) o hasta el final.
- Buscar objetos mal alocados y reasignar.
- Repetir hasta optimizar el criterio.

Criterio de Clustering

Varios criterios han sido propuestos basados en la identidad

$$\mathbf{T} = \mathbf{W} + \mathbf{B} = \sum_{i=1}^{n} \left(\mathbf{x}_i - \bar{\mathbf{x}}\right) \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^T$$

Criterio de Clustering

Varios criterios han sido propuestos basados en la identidad

$$\mathbf{T} = \mathbf{W} + \mathbf{B} = \sum_{i=1}^{n} \left(\mathbf{x}_i - \bar{\mathbf{x}}\right) \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^T$$

- Donde:
  - $\mathbf{W}$ es la matriz de variación dentro del cluster (within-cluster).
  - $\mathbf{B}$ es la matriz de variación entre clusters (between-cluster).

Criterio de Clustering

Varios criterios han sido propuestos basados en la identidad

$$\mathbf{T} = \mathbf{W} + \mathbf{B} = \sum_{i=1}^{n} \left(\mathbf{x}_i - \bar{\mathbf{x}}\right) \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^T$$

- Donde:
  - $\mathbf{W}$ es la matriz de variación dentro del cluster (within-cluster).
  - $\mathbf{B}$ es la matriz de variación entre clusters (between-cluster).

- Se busca minimizar (alguna función) de $\mathbf{W}$ o maximizar (alguna función) $\mathbf{B}$

Criterios

- Minimizar $\text{tr}(\mathbf{W})$
  - Popular por su simplicidad y costo computacional.
  - Invariante ante transformaciones ortogonales pero no ante todas las transformaciones singulares no lineales (i.e. diferente soluciones para los datos y los datos estandarizados).

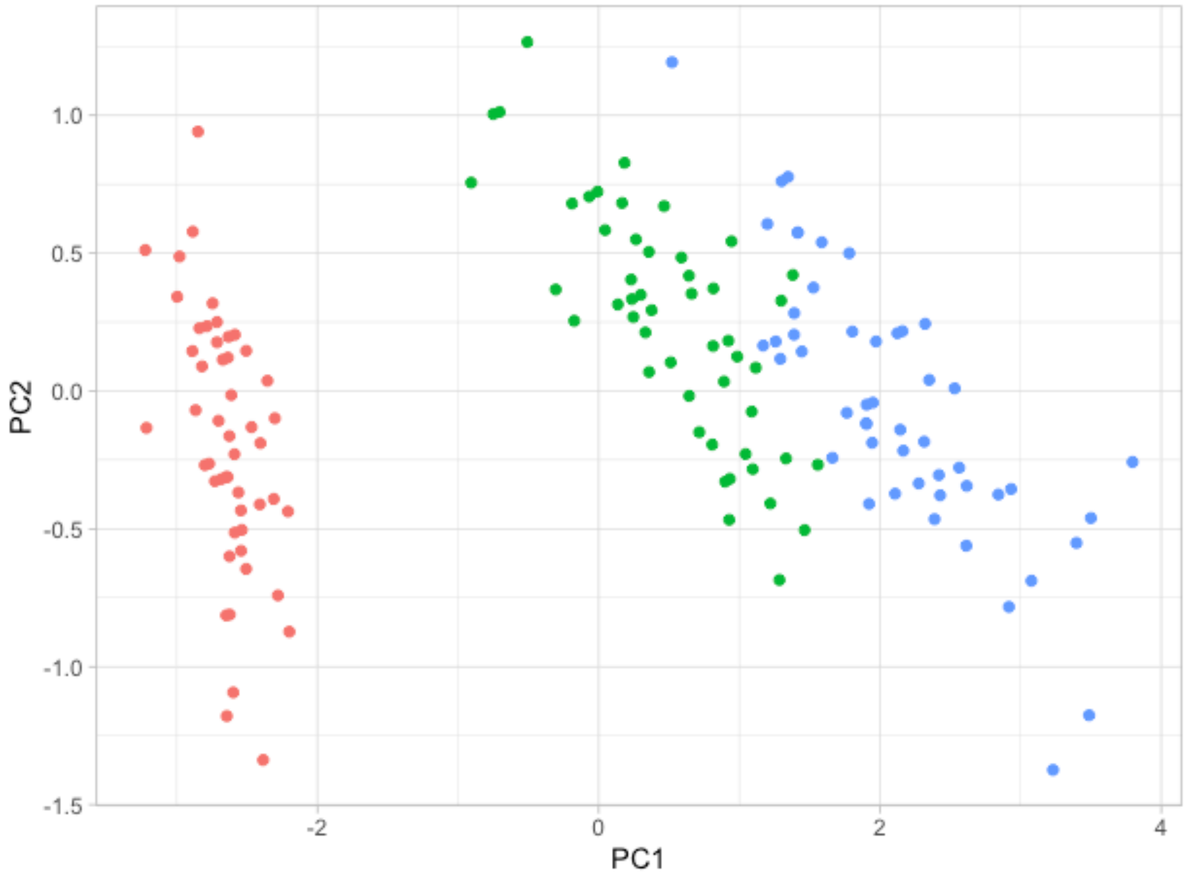
- Maximizar $\text{tr}\,(\mathbf{BW})^{-1}$
  - No es muy confiable ya que no corrige errores en los grupos (Maronna & Jacovkis, 1974)

Criterios

- Minimizar $|\mathbf{W}|$
  - Invariante ante transformaciones no singulares.
  - Mayor sensibilidad a la estructura de los datos (Friedman & Rubin, 1967).
  - Puede verse influenciado por una variable que permita crear clusters bien definidos (Marriott, 1971).

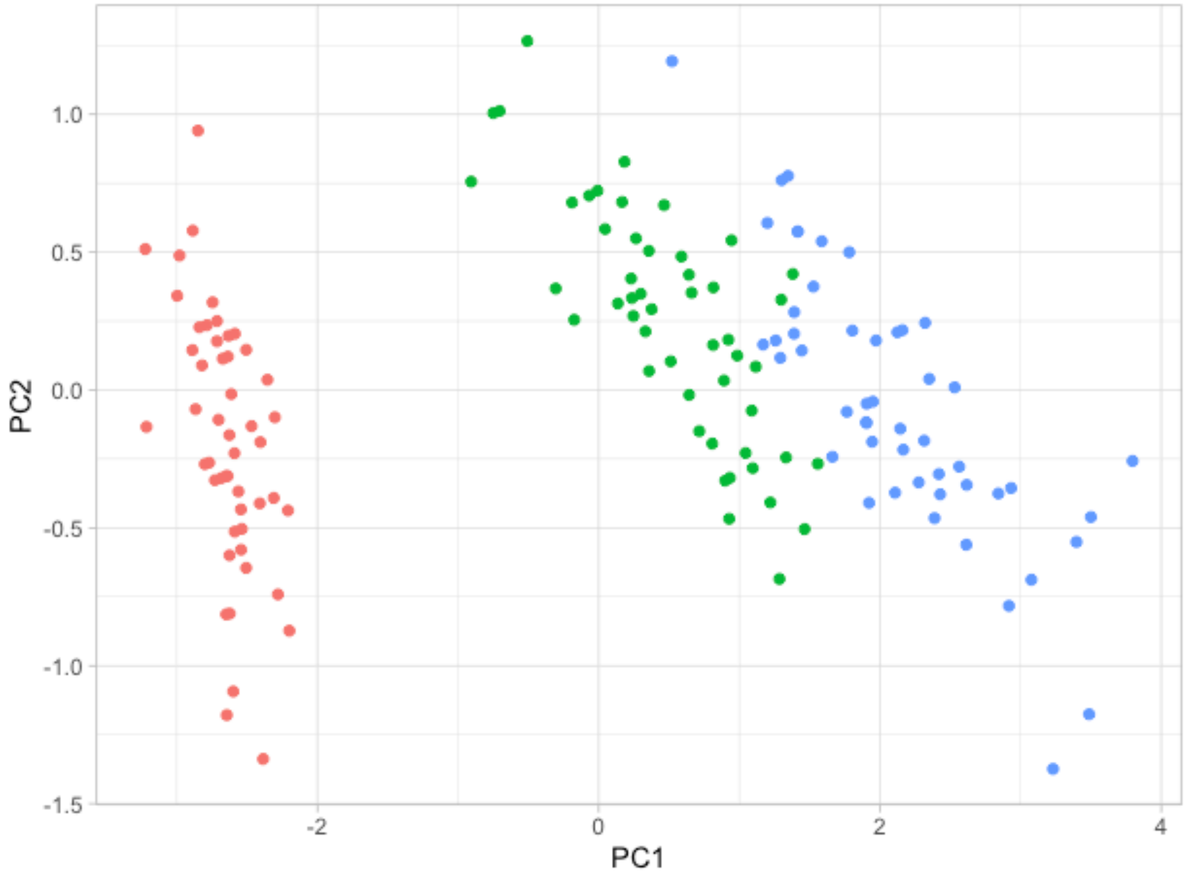
Ejemplo k-means

Etiquetas verdaderas

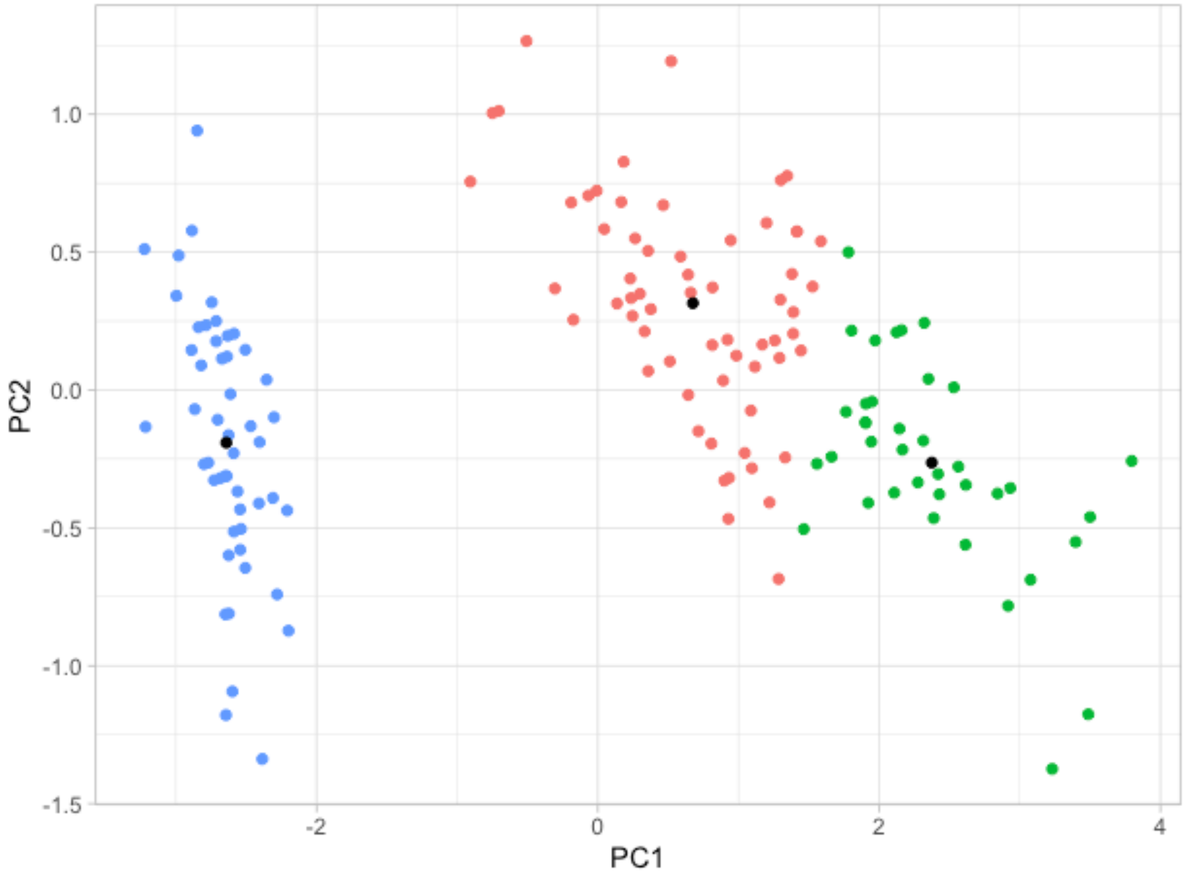


Ejemplo k-means

Etiquetas verdaderas

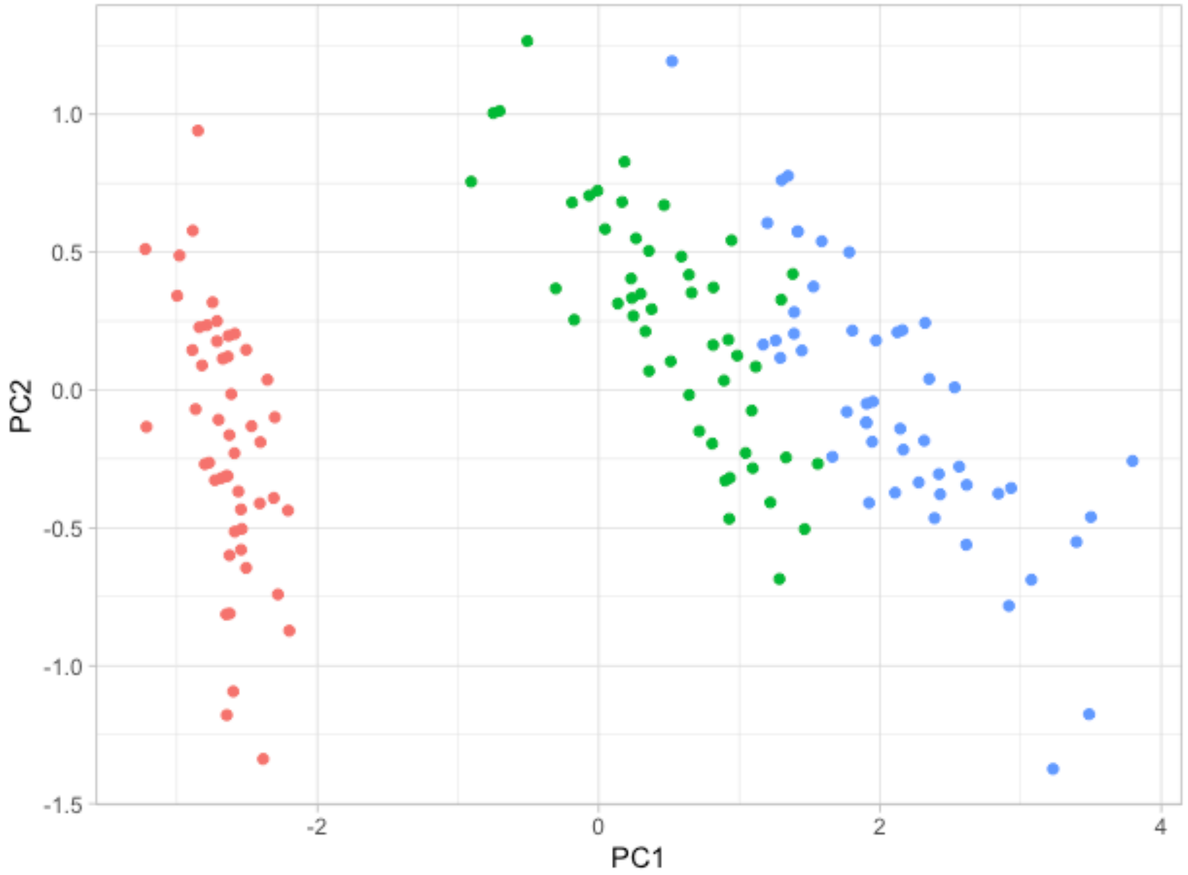


Etiquetas k-means

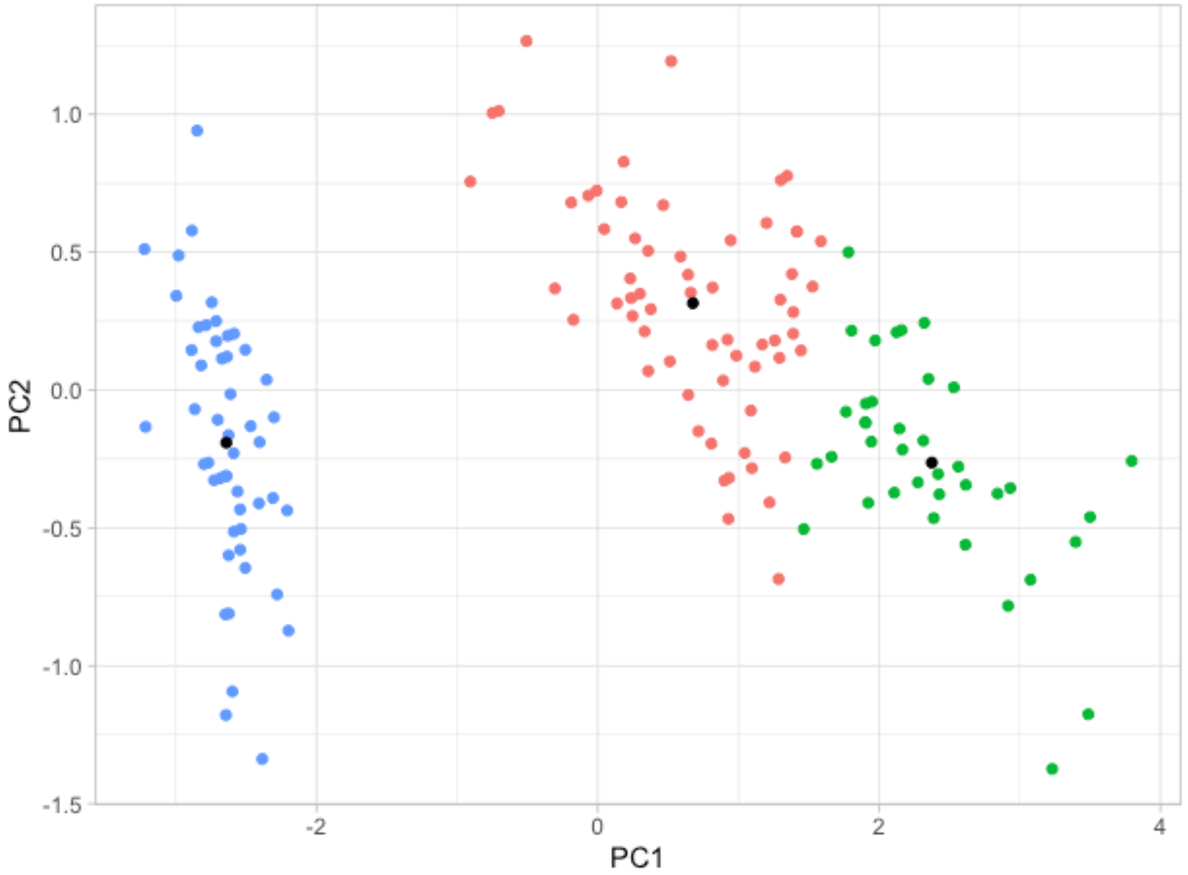


Ejemplo k-means

Etiquetas verdaderas



Etiquetas k-means



134 aciertos

Alternativas

- Seleccionar otra métrica (e.g. Manhattan).

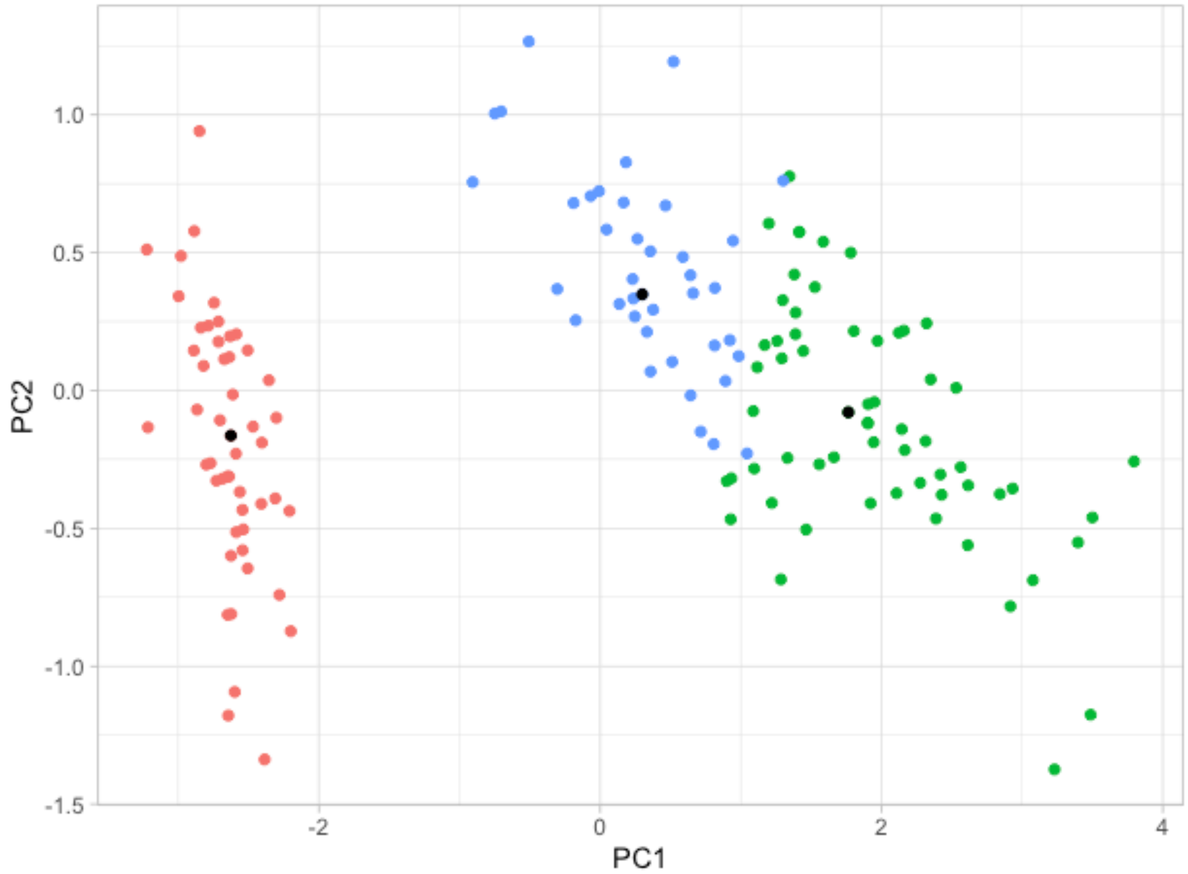
Alternativas

- Seleccionar otra métrica (e.g. Manhattan).
- Usar el medoide en lugar de la medias (algoritmo pam() en R)

Alternativas

- Seleccionar otra métrica (e.g. Manhattan).
- Usar el medoide en lugar de la medias (algoritmo pam() en R)

135 aciertos



Silhouette

- Medida de similitud de un objeto y el cluster al que pertenece comparado con el resto (Rousseeuw, 1987).

Silhouette

- Medida de similitud de un objeto y el cluster al que pertenece comparado con el resto (Rousseeuw, 1987).

- Construcción para el i-ésimo objeto en el cluster A:

Silhouette

- Medida de similitud de un objeto y el cluster al que pertenece comparado con el resto (Rousseeuw, 1987).

- Construcción para el i-ésimo objeto en el cluster A:
  - Obtener la disimilitud promedio de su cluster $a(i)$

Silhouette

- Medida de similitud de un objeto y el cluster al que pertenece comparado con el resto (Rousseeuw, 1987).

- Construcción para el i-ésimo objeto en el cluster A:
  - Obtener la disimilitud promedio de su cluster $a(i)$
  - Obtener el mínimo de las disimilitudes promedio de los otros clusters, i.e.,

$b(i) = \min_{C \neq A} \{d(i, C)\}$ (dicho cluster es la segunda mejor opción)

Silhouette

- Medida de similitud de un objeto y el cluster al que pertenece comparado con el resto (Rousseeuw, 1987).

- Construcción para el i-ésimo objeto en el cluster A:
  - Obtener la disimilitud promedio de su cluster $a(i)$
  - Obtener el mínimo de las disimilitudes promedio de los otros clusters, i.e.,

    $b(i) = \min_{C \neq A} \{d(i, C)\}$ (dicho cluster es la segunda mejor opción)

  - Definimos la silhouette como:

$$s(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}}$$

Observaciones

- $-1 \leq s(i) \leq 1$ por lo que:

Observaciones

- $-1 \leq s(i) \leq 1$ por lo que:
  - Si $s(i) > 0$

Observaciones

- $-1 \leq s(i) \leq 1$ por lo que:
  - Si $s(i) > 0$ el objeto está bien clasificado

Observaciones

- $-1 \leq s(i) \leq 1$ por lo que:
  - Si $s(i) > 0$ el objeto está bien clasificado
  - Si $s(i) = 0$

Observaciones

- $-1 \leq s(i) \leq 1$ por lo que:
  - Si $s(i) > 0$ el objeto está bien clasificado
  - Si $s(i) = 0$ el objeto está a la misma distancia de A y de B

Observaciones

- $-1 \leq s(i) \leq 1$ por lo que:
  - Si $s(i) > 0$ el objeto está bien clasificado
  - Si $s(i) = 0$ el objeto está a la misma distancia de A y de B
  - Si $s(i) < 0$

Observaciones

- $-1 \leq s(i) \leq 1$ por lo que:
  - Si $s(i) > 0$ el objeto está bien clasificado
  - Si $s(i) = 0$ el objeto está a la misma distancia de A y de B
  - Si $s(i) < 0$ el objeto está mal clasificado

Observaciones

- $-1 \leq s(i) \leq 1$ por lo que:
  - Si $s(i) > 0$ el objeto está bien clasificado
  - Si $s(i) = 0$ el objeto está a la misma distancia de A y de B
  - Si $s(i) < 0$ el objeto está mal clasificado

- Podemos crear una gráfica poniendo los silhouettes ordenados por cada cluster, en **R** usar la función silhouette()

Observaciones

- $-1 \leq s(i) \leq 1$ por lo que:
  - Si $s(i) > 0$ el objeto está bien clasificado
  - Si $s(i) = 0$ el objeto está a la misma distancia de A y de B
  - Si $s(i) < 0$ el objeto está mal clasificado

- Podemos crear una gráfica poniendo los silhouettes ordenados por cada cluster, en **R** usar la función silhouette()

- Proporciona una forma de medir que tanta estructura hemos descubierto usando el promedio de las silhouettes $\bar{s}(k)$ y una posible forma de elegir k con el silhouette coefficient

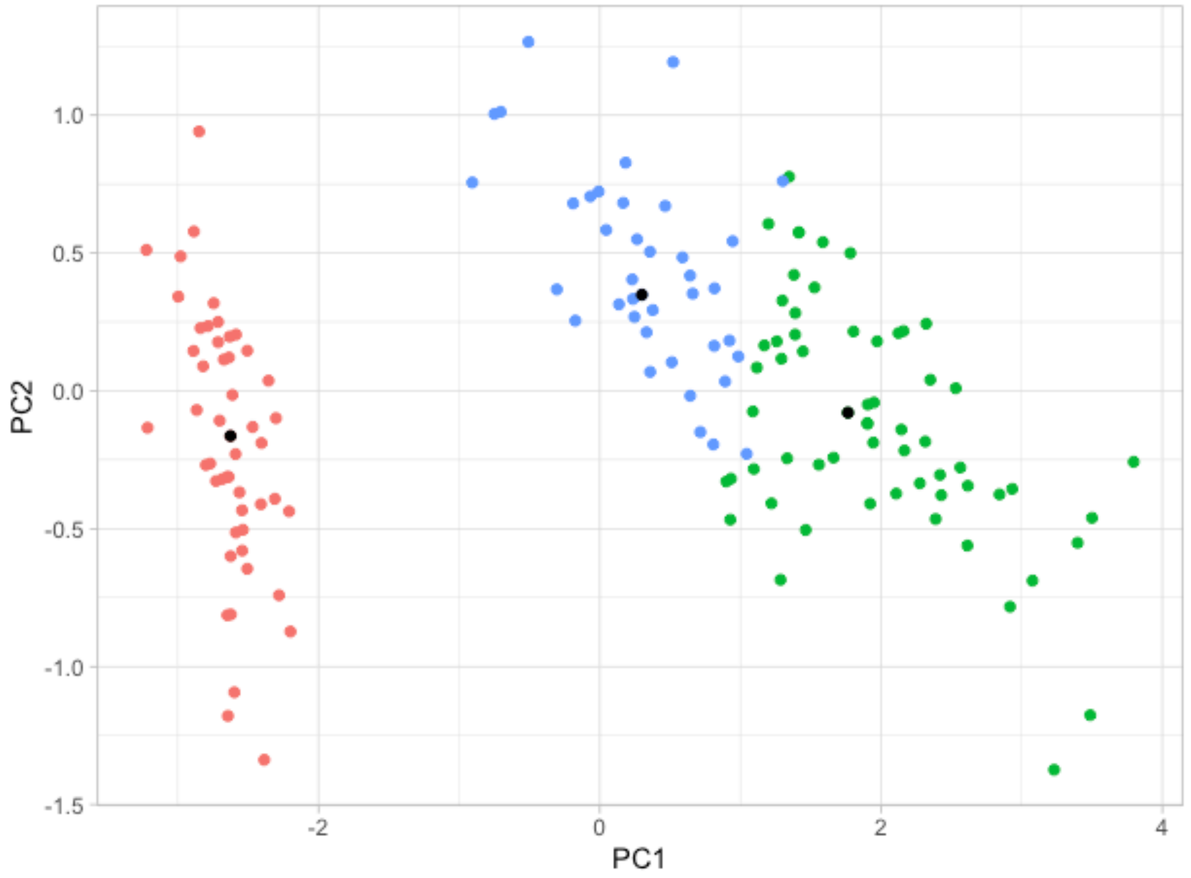
$$SC = \max_k \{\bar{s}(k)\}$$

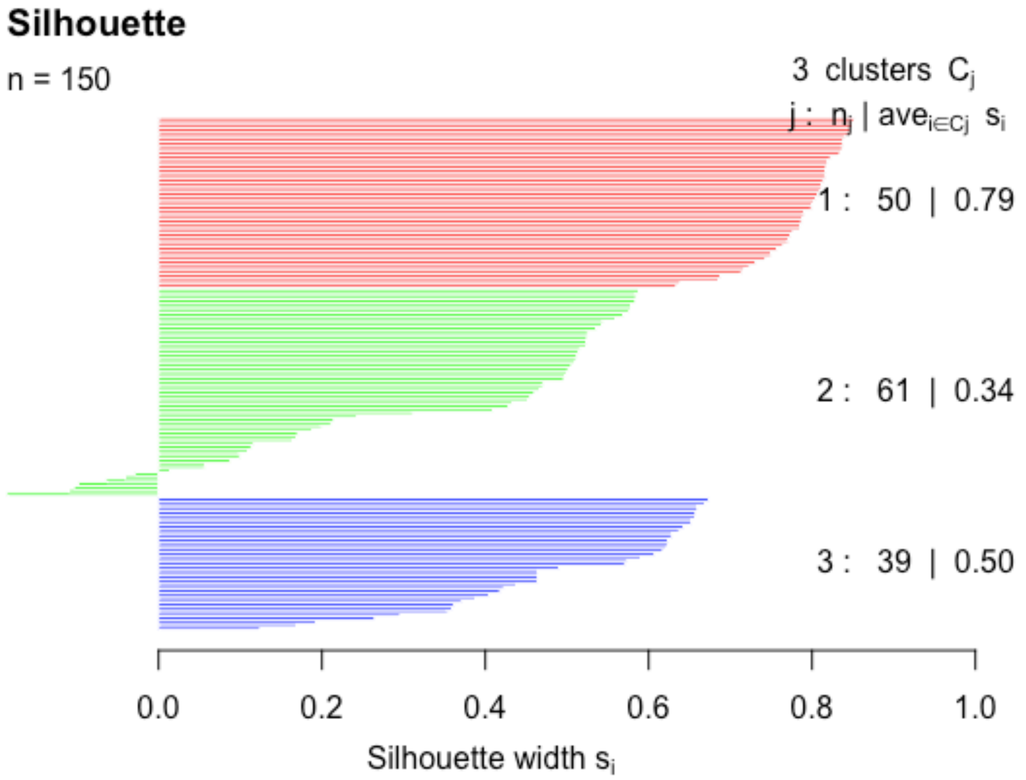
(Posible) Interpretación

- Kaufman (1990) proporciona la siguiente tabla basado en su experiencia:
  - Si $SC \in (.70,1]$ se ha encontrado una fuerte estructura de clustering
  - Si $SC \in (.5,.7]$ se ha encontrado una estructura razonable
  - Si $SC \in (.25,.5]$ la estructura es débil y se debe considerar otro método
  - Si $SC \leq .25$ no se encontró una estructura sustancial

Ejemplo Iris

- Usando el algoritmo k-medoids





Average silhouette width : 0.53

Ejemplo Iris

- Encontramos que las observaciones mal clasificadas son:

Observación	Cluster Asignado	Cluster Alternativo	Silhouette
114	2	3	-0.028
122	2	3	-0.042
73	2	3	-0.063
52	2	3	-0.098
55	2	3	-0.103
66	2	3	-0.109
76	2	3	-0.185

Métodos de Multi-Pertenencia

Motivación

En ocasiones es más significativo permitir que las observaciones pertenezcan a varios grupos.

Motivación

En ocasiones es más significativo permitir que las observaciones pertenezcan a varios grupos.

Idea

Encontrar un coeficiente de pertenencia para cada objeto, $u_{im} \in [0,1]$, (membership coefficient) para cada cluster de tal forma que $\sum_{m=1}^{k} u_{im} = 1$

Algoritmo Fanny

- Algoritmo iterativo propuesto por Kaufman en 1990 (en **R** usamos fanny())

Algoritmo Fanny

- Algoritmo iterativo propuesto por Kaufman en 1990 (en **R** usamos fanny())

- Se busca minimizar la función:

$$\sum_{m=1}^{k} \frac{\sum_{i,j=1}^{n} u_{im}^2 u_{jm}^2 d(i,j)}{2 \sum_{j=1}^{n} u_{jm}^2}$$

Algoritmo Fanny

- Algoritmo iterativo propuesto por Kaufman en 1990 (en **R** usamos fanny())

- Se busca minimizar la función:

$$\sum_{m=1}^{k} \frac{\sum_{i,j=1}^{n} u_{im}^2 u_{jm}^2 d(i,j)}{2\sum_{j=1}^{n} u_{jm}^2}$$

- Para medir el tipo de clustering (suave o duro) usamos el coeficiente de Dunn (1976):

$$F_k = \sum_{i=1}^{n}\sum_{m=1}^{k} \frac{u_{im}^2}{n}$$

Algoritmo Fanny

- Algoritmo iterativo propuesto por Kaufman en 1990 (en **R** usamos fanny())

- Se busca minimizar la función:

$$\sum_{m=1}^{k} \frac{\sum_{i,j=1}^{n} u_{im}^2 u_{jm}^2 d(i,j)}{2\sum_{j=1}^{n} u_{jm}^2}$$

- Para medir el tipo de clustering (suave o duro) usamos el coeficiente de Dunn (1976):

$$F_k = \sum_{i=1}^{n}\sum_{m=1}^{k} \frac{u_{im}^2}{n}$$

- El mínimo de F_k se alcanza cuando hay máxima difusión (complete fuzziness) y el máximo cuando se crea una partición.

Algoritmo Fanny

Observación	Grupo 1	Grupo 2	Grupo 3
102	0.32	0.22	0.45
143	0.08	0.16	0.75
57	0.09	0.65	0.25
71	0.07	0.44	0.47
139	0.06	0.72	0.20
84	0.05	0.69	0.25
114	0.05	0.75	0.18
122	0.11	0.61	0.27
73	0.14	0.28	0.56
52	0.09	0.63	0.27
55	0.07	0.65	0.27
66	0.11	0.62	0.26
76	0.05	0.69	0.25