ITBA
Instituto Tecnológico de Buenos Aires

30/AGOSTO

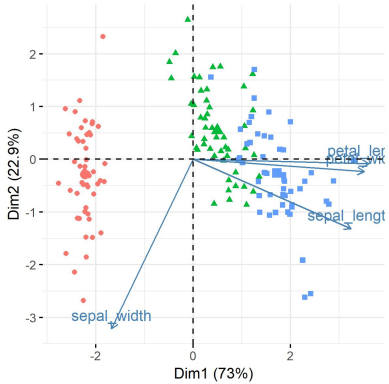
MÉTODOS DE CLUSTERING

Aprendizaje no supervisado

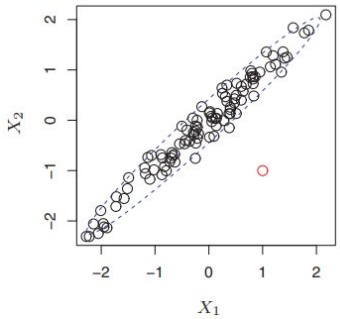
- Reglas de asociación

$$educ > 10 \wedge estado = Soltero \Rightarrow ingreso > \$30.000$$

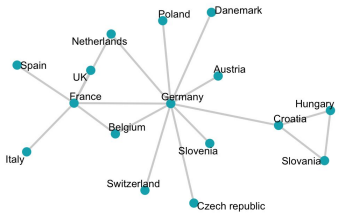
- Reducción de la dimensionalidad



- Detección de anomalías



- Grafos



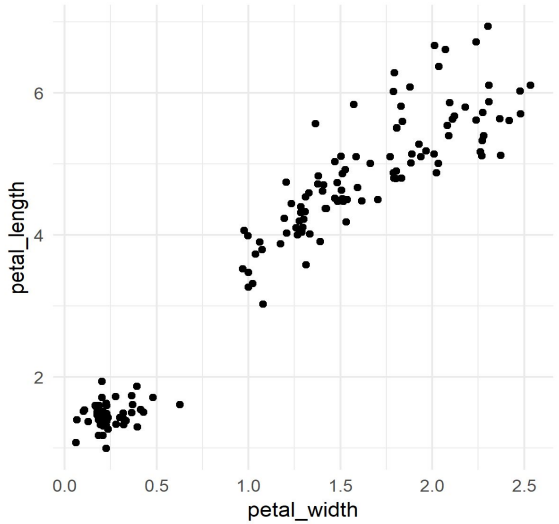
- Clustering

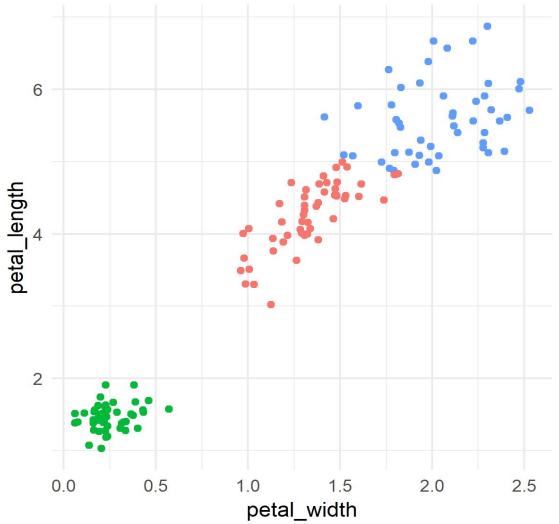
Clustering

Datos input: N observaciones - p variables

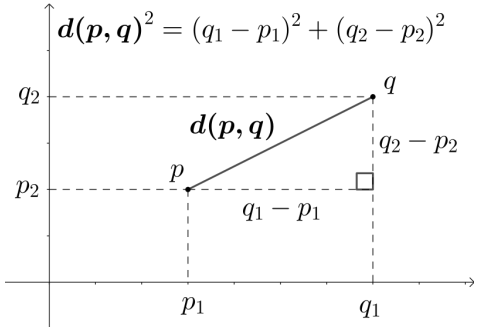
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2

- Agrupar observaciones similares (***cohesión***)
- Separar observaciones distintas (***separación***)





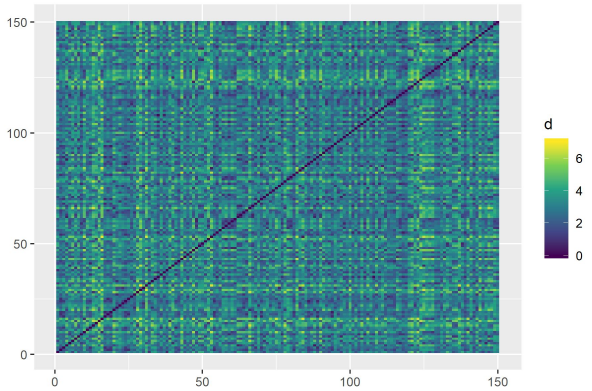
- ***Disimilitud / Distancia***



Distancia euclidiana

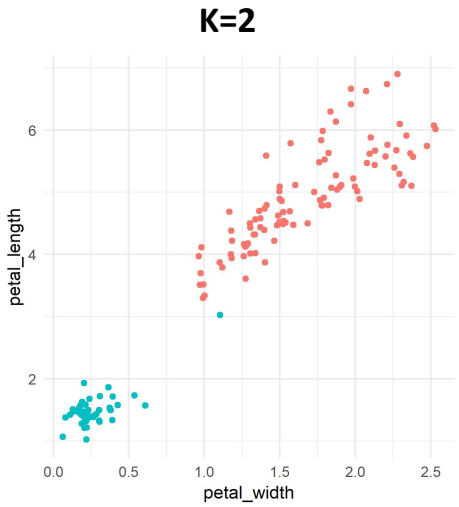
$$d_{euc}(p,q) = \sqrt{\sum_{j=1}^{k}(p_j - q_j)^2}$$

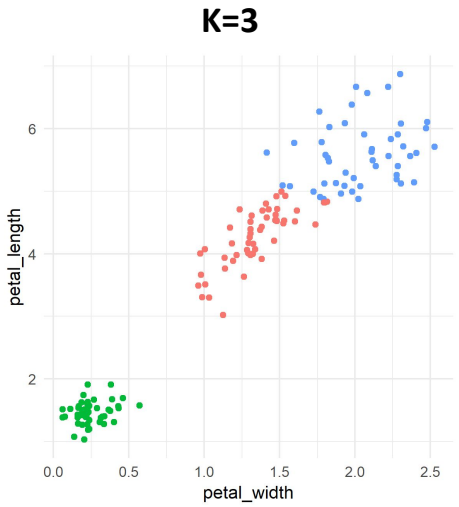
Matriz de distancias

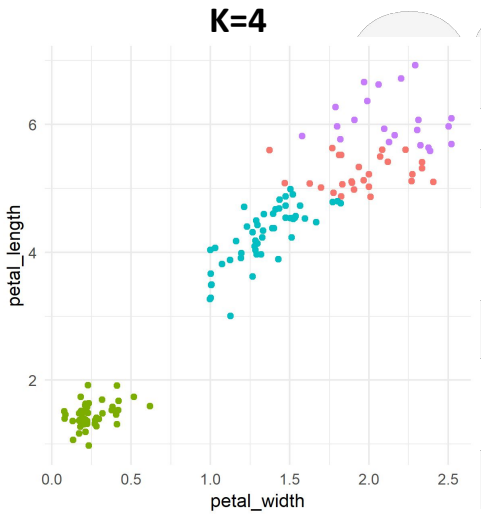


Clustering por partición

- Para *K* dado: *C1,...,Ck* clusters ***excluyentes*** y ***colectivamente exhaustivos***
- Los clusters más pequeños ***no están anidados*** dentro de los más grandes







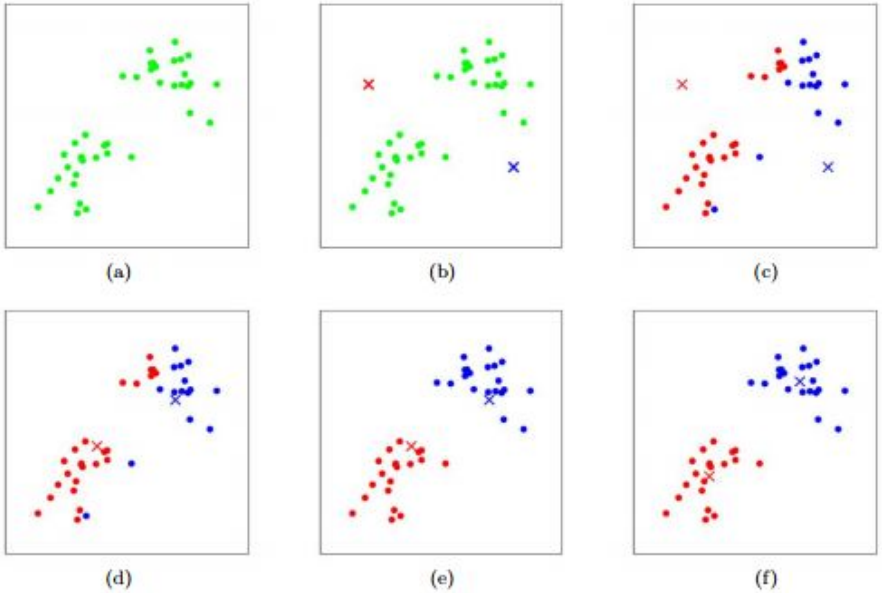
K-Medias

(0) Inicializar K centroides al azar (b)

(1) Asignar cada objeto al cluster con **centroide más cercano** (c)

(2) Computar el **centroide** de cada cluster (d)

Repetir (1) y (2) hasta que no haya cambios en las asignaciones (e) (f)



Fuente: Stanford CS221

$$\min_{C_1,\ldots,C_K} \sum_{k=1}^{K} VIC(C_k)$$

Variabilidad Intra-Cluster

VIC(C_k)

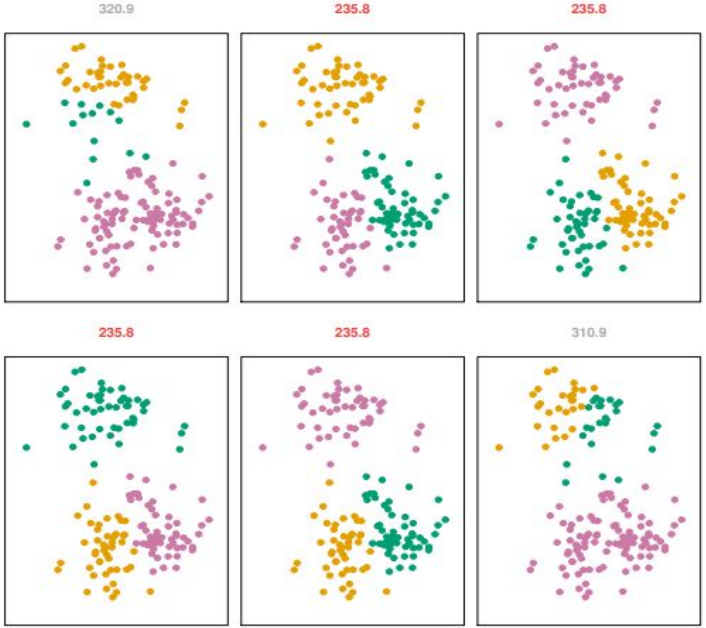
distancia euclidiana cuadrática

$$2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2 = \frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2$$

"Within cluster Sum of Squares" **(WSS)**

K-Medias

El resultado depende de la inicialización



Fuente: An Introduction to Statistical Learning (2014)

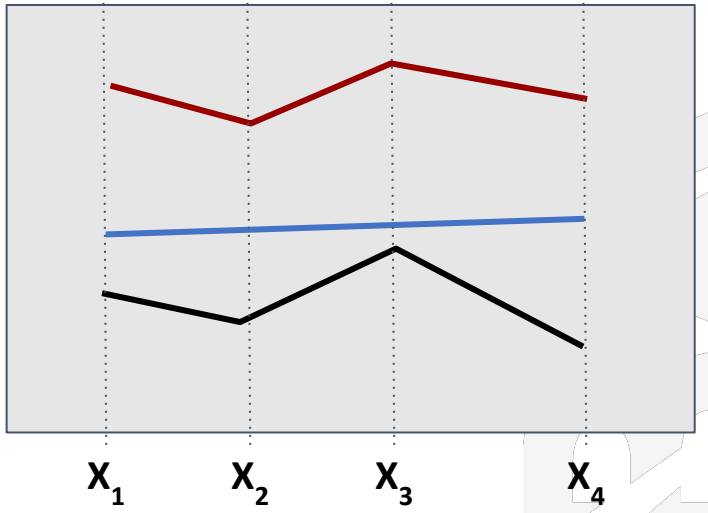
Distancia / (di)similitud

Variables cuantitativas

$$d_{euc}(x,y) = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2}$$ **Euclidiana**

$$d_{man}(x,y) = \sum_{j=1}^{p}|(x_j - y_j)|$$ **Manhattan**

$$d_{cor}(x,y) = 1 - \frac{\sum_{j=1}^{p}(x_j - \overline{x})(y_j - \overline{y})}{2\sqrt{\sum_{j=1}^{n}(x_j - \overline{x})^2 \sum_{j=1}^{p}(y_j - \overline{y})^2}} + \frac{1}{2}$$ **Correlación**



Es fundamental analizar si las variables deben ser **normalizadas**

Distancia / (di)similitud

*Variables binarias**

- Coeficiente de coincidencias $\frac{n_{1,1} + n_{0,0}}{n_{1,1} + n_{1,0} + n_{0,1} + n_{0,0}}$
- Coeficiente de Jaccard $\frac{n_{1,1}}{n_{1,1} + n_{1,0} + n_{0,1}}$

Dados dos objetos i, j se computa la cantidad de de atributos para cada posible combinación

	$j = 1$	$j = 0$
$i = 1$	$n_{1,1}$	$n_{1,0}$
$i = 0$	$n_{0,1}$	$n_{0,0}$

Por ejemplo, $n_{1,1}$ indica la cantidad de atributos en los que ambos tienen 1

Fuente: Principles of Data Mining (2001)

Variables de tipo mixto

- Distancia de Gower

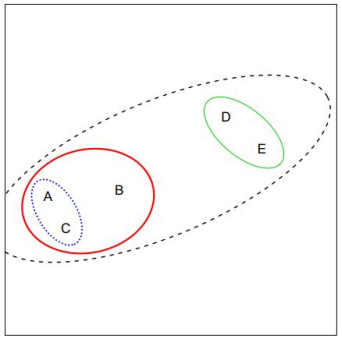
*Ver *A Survey of Binary Similarity and Distance Measures* (Choi et al, 2010)

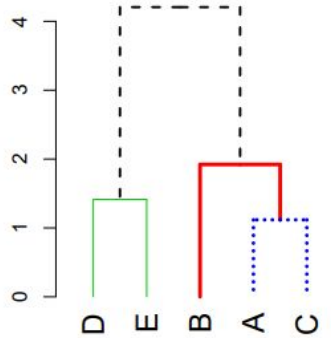
Clustering jerárquico (aglomerativo)

0) *Inicialización: cada objeto es un cluster*

1) **Fusionar los dos clusters *más similares*** en un solo cluster

Repetir (1) hasta tener un solo cluster





Dendrograma

Fuente: An Introduction to Statistical Learning

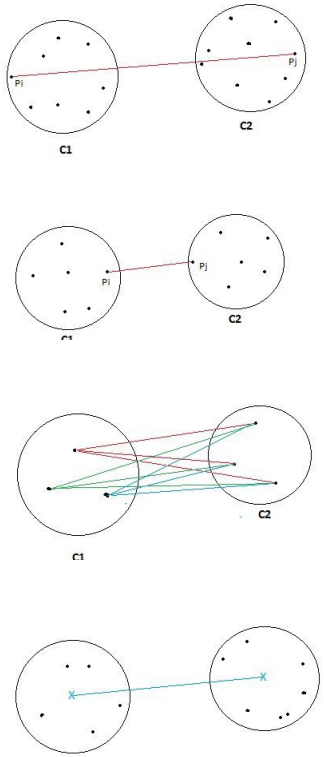
¿Qué es *similar*?

Métrica de distancia

Euclidiana, Manhattan, correlación, etc.

Criterio de Linkage

- **Complete** (disimilitud máxima)
- **Single** (disimilitud mínima)
- **Average** (disimilitud promedio)
- **Centroid** (disimilitud entre centroides)
- **Ward** (incremento de VIC)

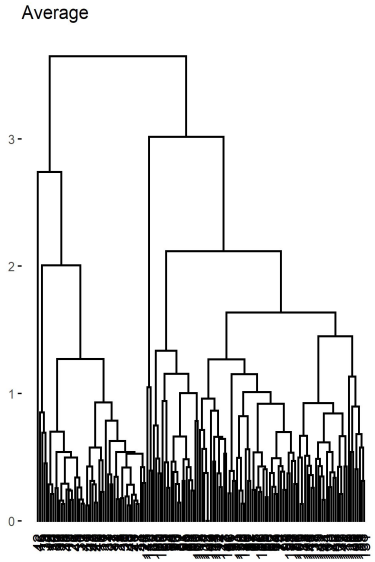


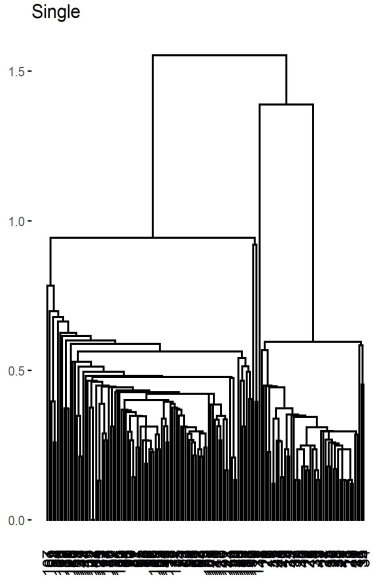
Fuente: towardsdatascience.com

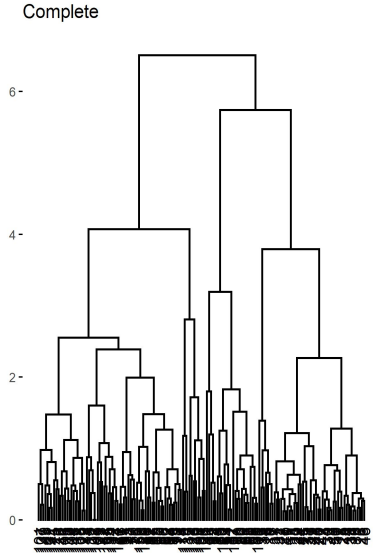
¿Qué método elegir?

Coeficiente de correlación cofenético

- Comparación entre **distancias reales** y **distancias cofenéticas**

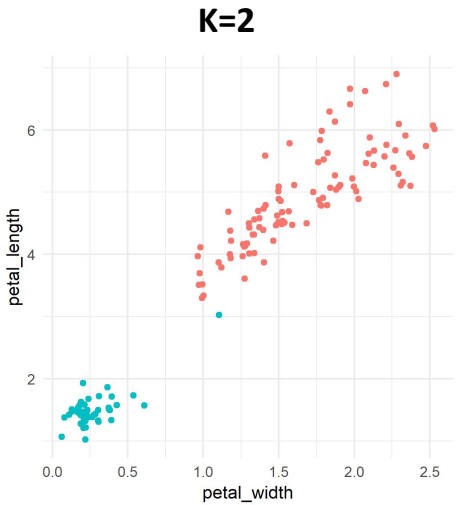


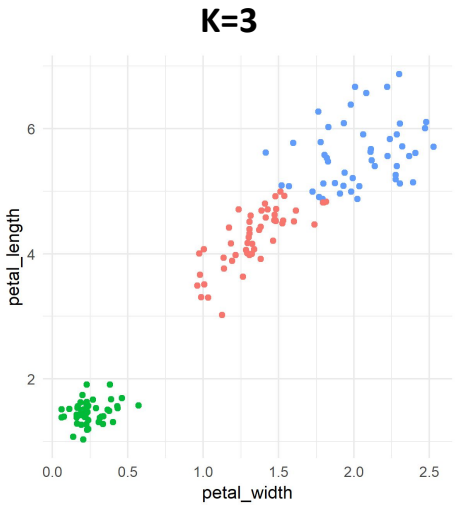


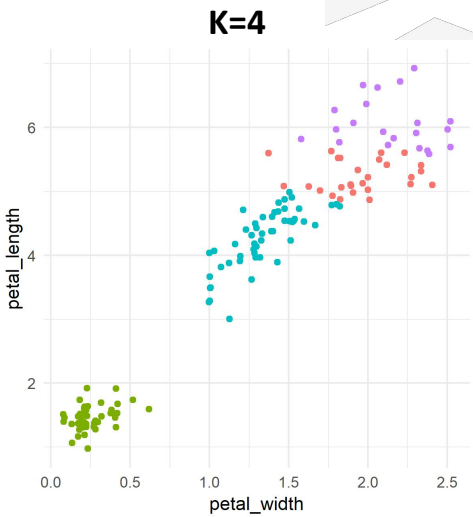


	Cophenetic Corr.
complete	0.7514592
average	0.8543606
single	0.8300050

¿Cómo elegir K?

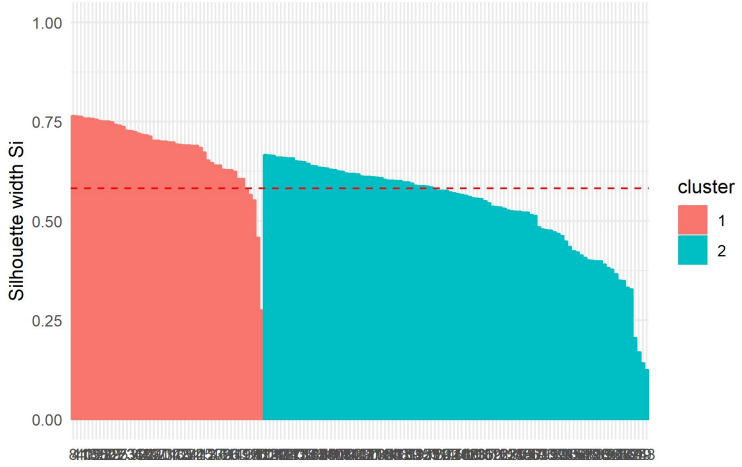






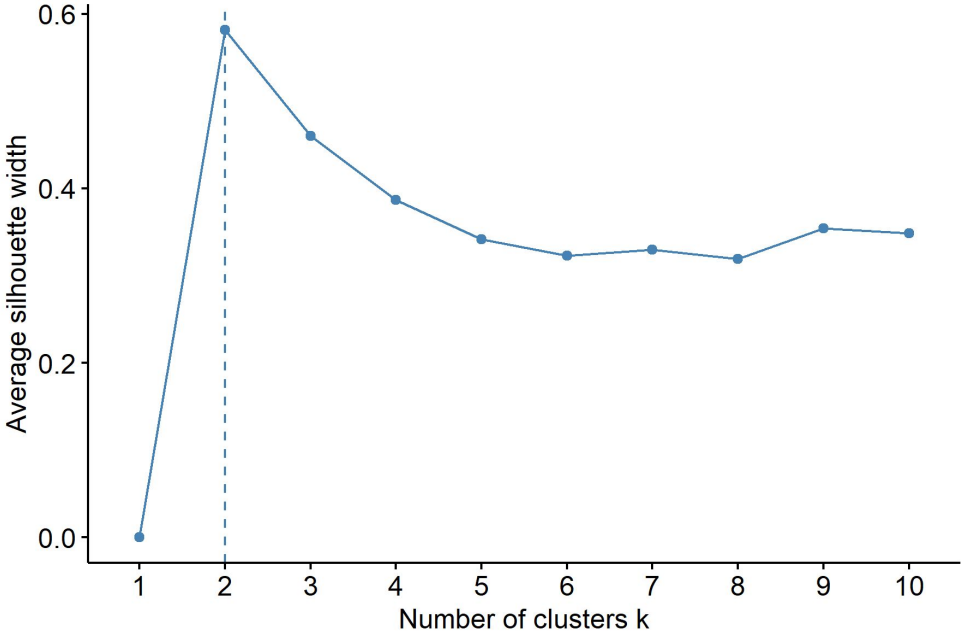
¿Cómo elegir K?

(1) **Silhouette** promedio máximo



$$s(i) = \frac{\min_{j \neq l(i)} \bar{d}(i, C_j) - \bar{d}(i, C_{l(i)})}{\max(\bar{d}(i, C_{l(i)}), \min_{j \neq l(i)} \bar{d}(i, C_j))}$$

b ***a*** ***a*** ***b***



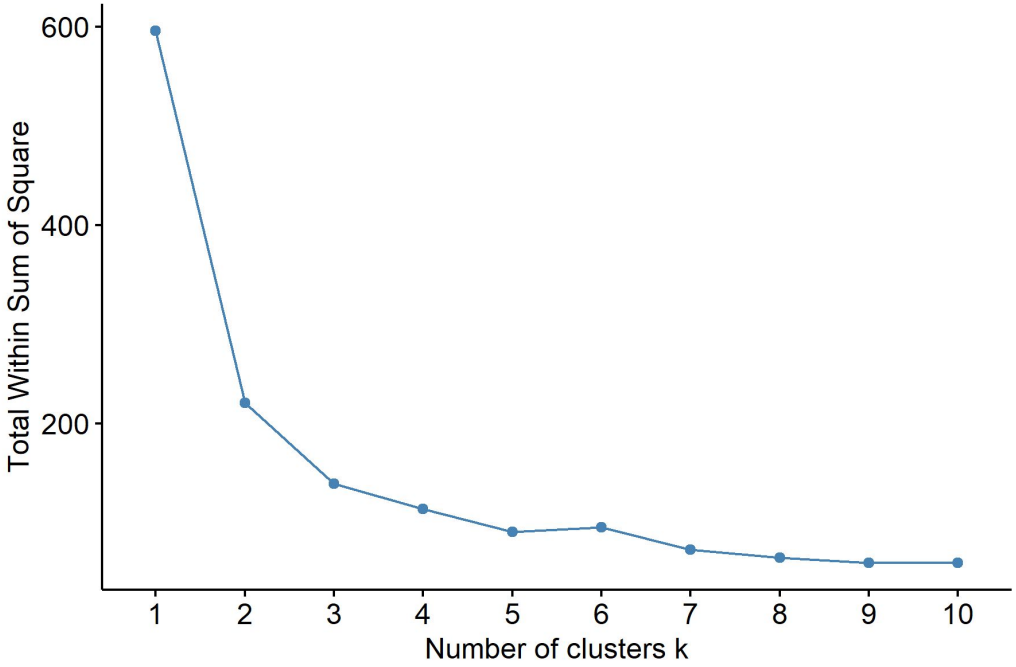
a : distancia promedio de *i* a objetos de su cluster
b: mínima distancia promedio de *i* a otro cluster

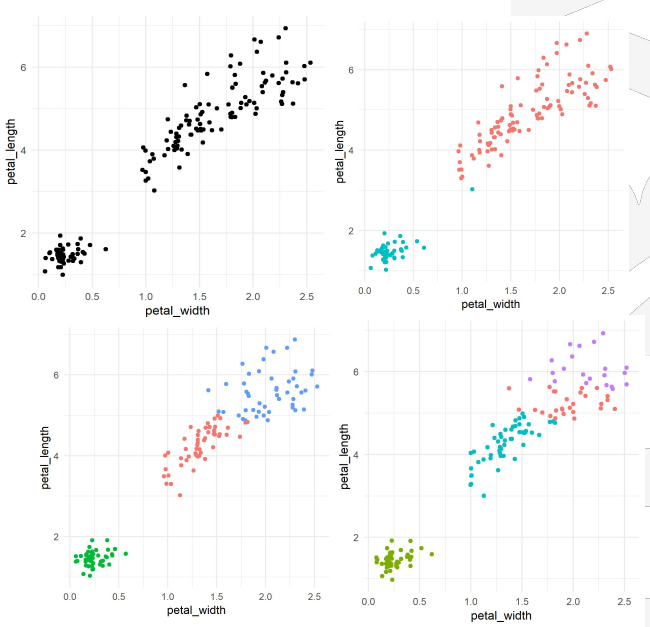
¿Cómo elegir K?

(2) Punto de quiebre en **VIC total**

"Within cluster Sum of Squares"
(WSS)

$$2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2 = \frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2$$





Validación

- Estadístico de Hopkins*

$$H = \frac{\sum_{i=1}^{s} x_i}{\sum_{i=1}^{s} x_i + \sum_{i=1}^{s} z_i}$$

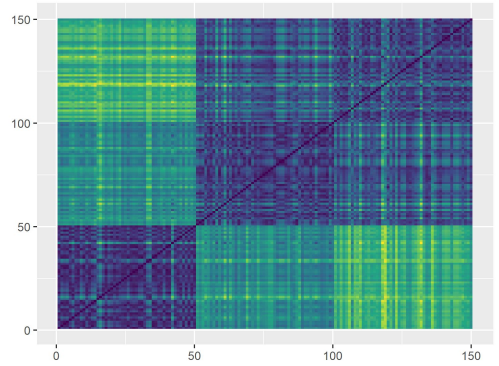
- Dataset X
- Muestra de *s* objetos:

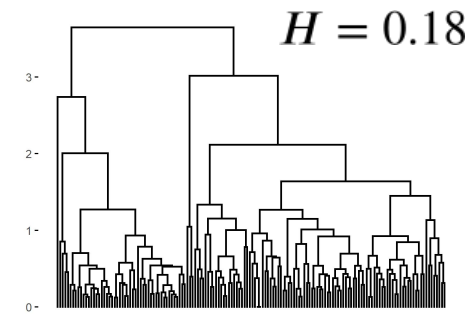
x_i : distancia entre objeto *i* y vecino más cercano en X

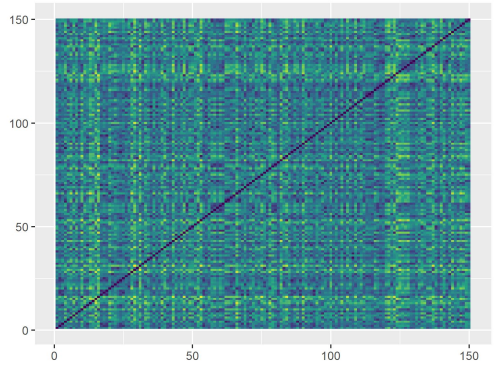
- Muestra de *s* objetos de un dataset con distribución uniforme:

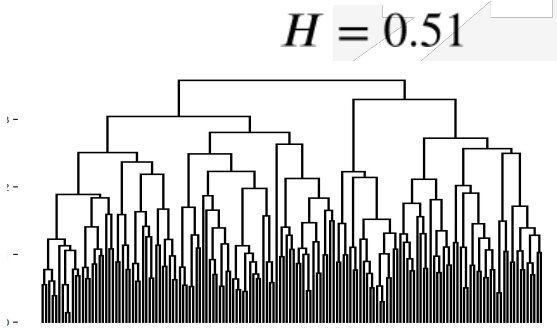
z_i : distancia entre objeto *i* y vecino más cercano en X

*En algunas implementaciones se usa *z* en el numerador — esto invierte la interpretación pero no altera las conclusiones









Conclusión

Selección de variables, **normalización** de variables, métrica de **distancia**, **algoritmo** de clustering, criterio de **linkage**, etc....

"... **decisions can have a strong impact on the results** obtained. In practice, we **try several different choices**, and look for the one with the most useful or interpretable solution. With these methods, **there is no single right answer**—any solution that exposes some interesting aspects of the data should be considered"

"Most importantly, we must be careful about how the results of a clustering analysis are reported. These **results should not be taken as the absolute truth about a data set**. Rather, they should constitute a starting point for the development of a scientific hypothesis and further study"

Introduction to Statistical Learning

Lecturas recomendadas

- *Introduction to Statistical Learning* Cap. 12.4
- *The Elements of Statistical Learning* Cap. 14.3
- *Probabilistic Machine Learning: An Introduction* Cap. 21