Análisis Estadístico de Datos Ómicos en R/Bioconductor

Luis Valenzuela Villa,
luis.valenz.v@gmail.com

Laboratorio de Genética, Facultad de Ciencias,
Laboratorio de BioMatemática y Ómica Integrativa, Facultad de Medicina,
Universidad de Chile.

23 de Agosto, 2017

Análisis Estadístico de Datos Ómicos en R/Bioconductor

- **Coordinador**: Luis Valenzuela

- **Patrocinantes:** Carlos Jeréz,
  David Véliz.

http://cursos.ciencias.uchile.cl/ecologia/estadisticaaplicada2017/

- **Horario**: Desde el 23 de Agosto al 25 de Octubre.
  Miércoles 14:30 - 17:30hrs
  Sala de Computación, Facultad de Ciencias

Análisis Estadístico de Datos Ómicos en R/Bioconductor

- **Objetivo**: Brindar conocimientos básicos en Bioinformática, Estadística Exploratoria e Inferencial en el análisis e interpretación de las nuevas tecnologías Ómicas utilizando R/Bioconductor.

- **Metodología**: Exposiciones teóricas y pasos prácticos.

- **Evaluación**: Informe final, Análisis de Datos Ómicos Reales.

- **Programa**:

CLASE 1 (23-08-2017) - CLUSTERING Y CLASIFICACIÓN.

CLASE 2 (30-08-2017) - ANÁLISIS MULTIVARIADO DE MARCADORES MOLECULARES.

CLASE 3 (06-09-2017) – INTRODUCCIÓN TRANSCRIPTÓMICA: MICROARRAYS

12/09/2017: Introducción Inferencia Bayesiana 14:30

CLASE 4 (13-09-2017) – INTRODUCCIÓN TRANSCRIPTÓMICA: RNA-seq

CLASE 5 (27-09-2017) – VÍAS BIOLÓGICAS, ONTOLOGÍAS Y ENRIQUECIMIENTO FUNCIONAL

CLASE 6 (04-10-2017) – INMUNOPRECIPITACIÓN DE CROMATINA SEGUIDA DE SECUENCIACIÓN

CLASE 7 (11-10-2017) – PROTEOMICA

CLASE 8 (18-10-2017) – INFERENCIA REDES DE REGULACIÓN

CLASE 9 (25-10-2017) – ANÁLISIS MICROBIOTA INTESTINAL

- **Material Clases teóricas y Scripts:**

https://www.u-cursos.cl/ciencias/2017/2/CSDBIOT045/

http://cursos.ciencias.uchile.cl/ecologia/estadisticadedatosomicos/

CLUSTERING Y CLASIFICACIÓN

Luis Valenzuela Villa,
luis.valenz.v@gmail.com

Laboratorio de Genética, Facultad de Ciencias,
Laboratorio de BioMatemática y Ómica Integrativa, Facultad de Medicina,
Universidad de Chile.

23 de Agosto, 2017

CLUSTERING Y CLASIFICACIÓN

CLUSTERING

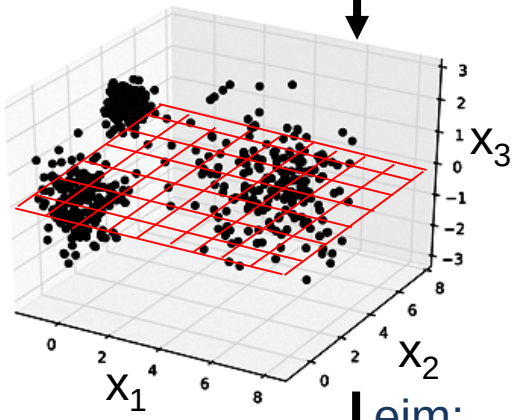
- Clustering Jerárquico:
  - Clustering Aglomerativo.
- K-means.
- Biclustering.

CLASIFICACIÓN

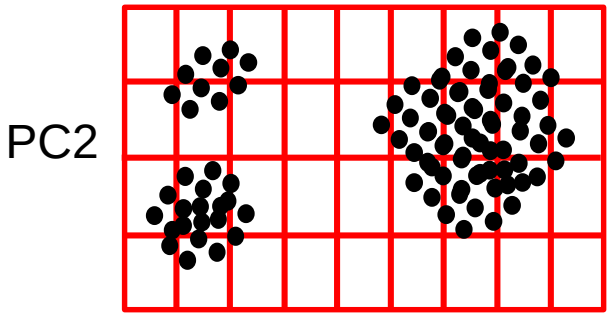
- KNN: *k-nearest neighbors.*
- L/QDA: *Linear/Quadratic discriminant analysis.*
- SVM: *Support vector machine.*
- CART: *classification and regression trees.*
- *Random Forest*.

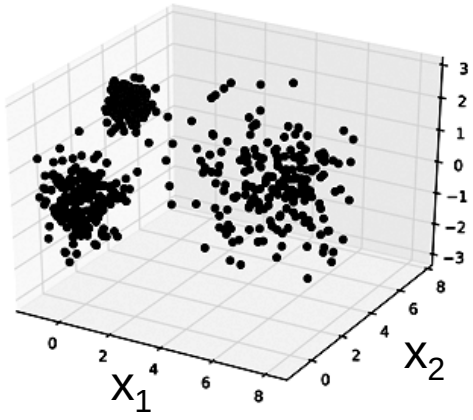
Análisis Exploratorio: dos familias de métodos no supervisados

Configuración de la nube de puntos en el mejor plano



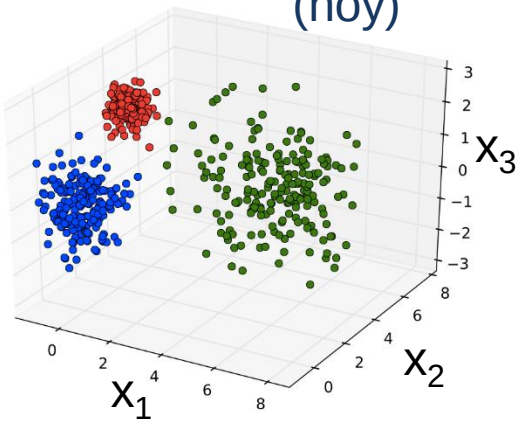
ejm:
PCA (próxima clase)





Agrupamiento en el espacio (Clustering)

ejm:
K-means, k=3
(hoy)



CLUSTERING: Idea general

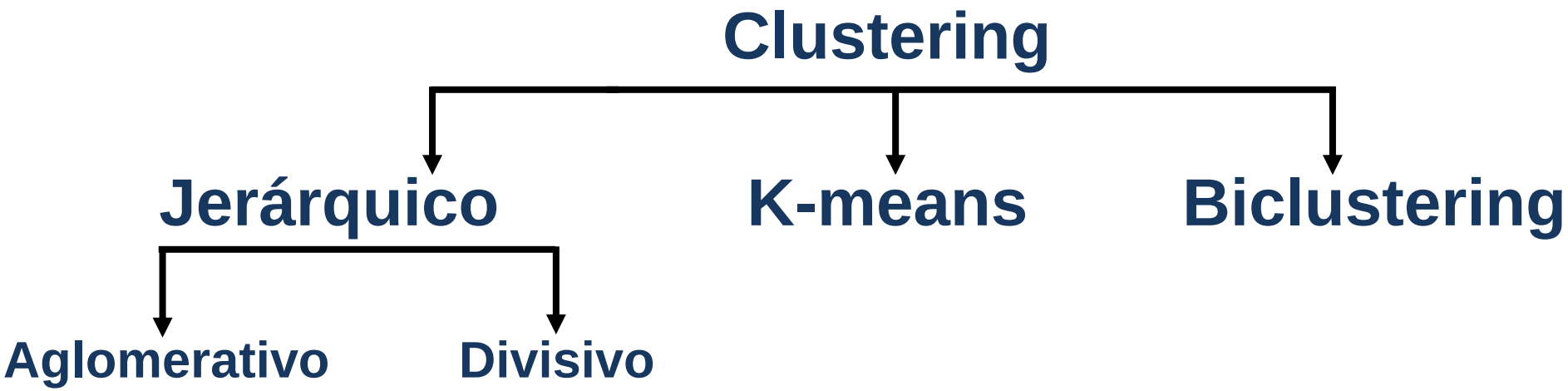
- **Métodos no supervisados, i.e. no requieren variables predictoras.**

- Clustering nos permite agrupar (o dividir) observaciones de un modo no supervisado en función de cuán similares son.

- Se realiza en base a medidas de distancias entre observaciones.

- El propósito es identificar observaciones que están lo suficientemente cerca para ser consideradas como grupo/cluster y lo suficientemente separadas del resto.

CLUSTERING: Algunas Técnicas



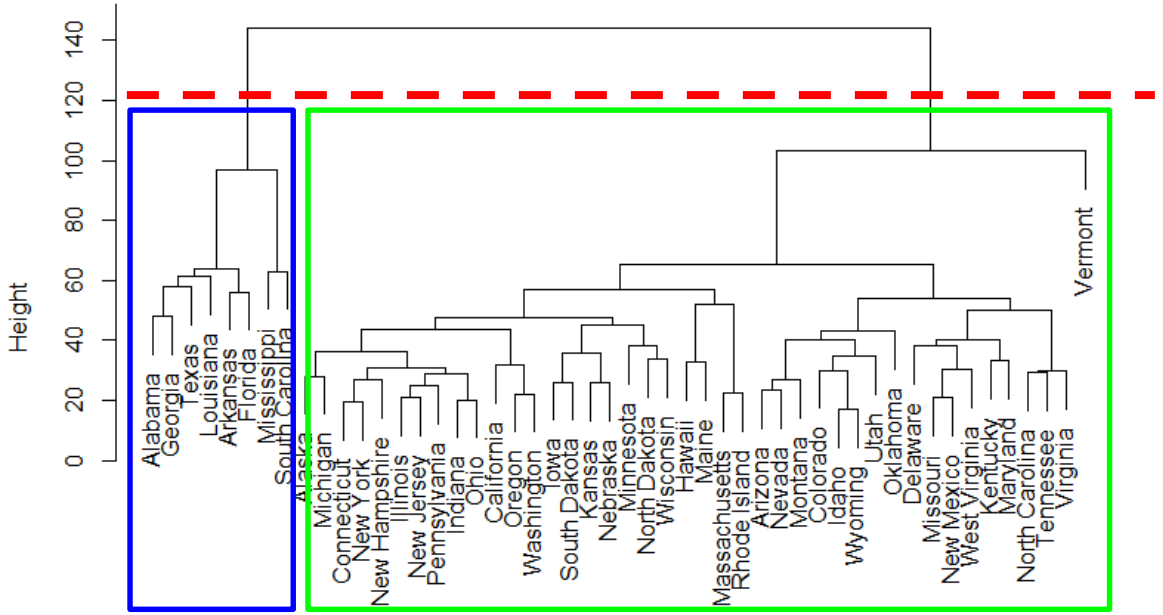
- **Aglomerativo**: Partiendo de los individuos, agrupa en cada paso los pares de observaciones/clusters más cercanos.

- **Divisivo:** Divide iterativamente la población hasta el nivel individual.

- **K-means**: Se parte con un número preestablecido, k, de clusters a formar.

Clustering: Jerárquico Aglomerativo

- Produce una jerarquía de clusters que puede ser visualizada como un dendograma.
- No asume un numero particular de clusters.
- La distancia considerada y del modo en que se evalúa la distancia entre clusters determina el resultado.



Clustering: Jerárquico Aglomerativo:

- **Distancias entre observaciones:**

1. Distancia Euclidiana

$$d(f,g) = \sqrt{\sum_{i=1}^{n}(f_i - g_i)^2}$$

2. Distancia Manhattan

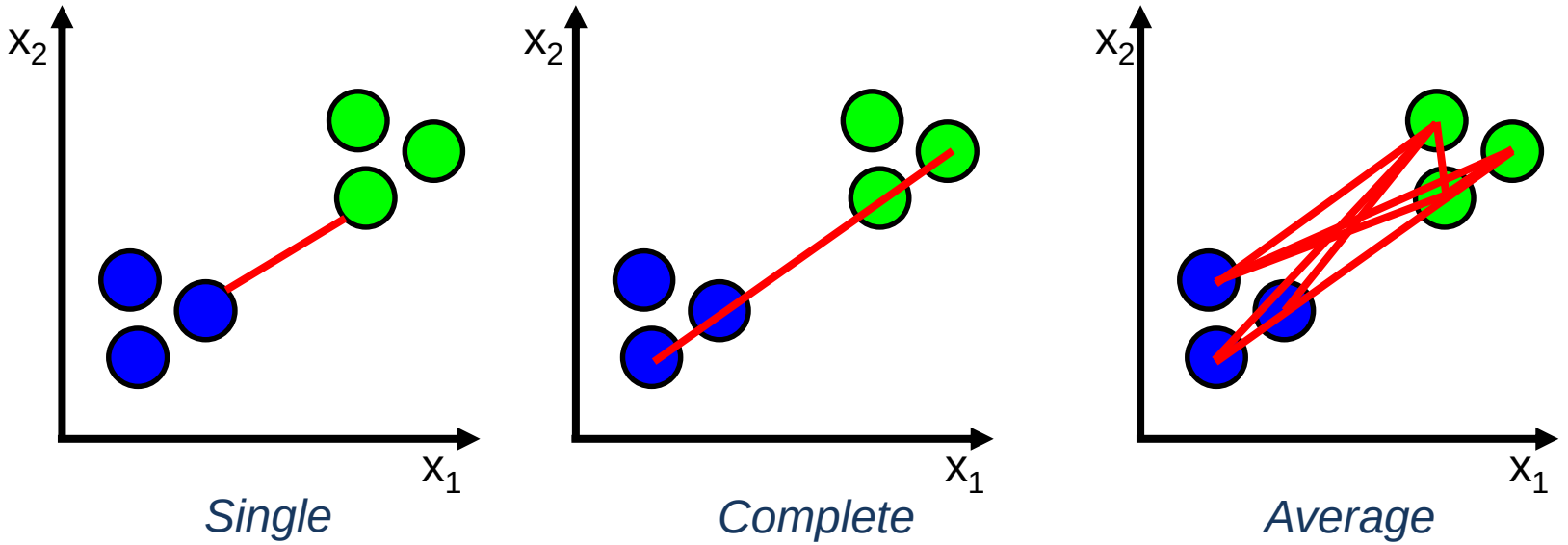
$$d(f,g) = \sum_{i=1}^{n}|f_i - gi|$$

3. Distancia Pearson

$$d(f,g) = \frac{\sum_{i=1}^{n}(f_i - \bar{f})(g_i - \bar{g})}{\sqrt{\sum_{i=1}^{n}(f_i - \bar{f})^2 \sum_{i=1}^{n}(g_i - \bar{g})^2}}$$

Clustering: Jerárquico Aglomerativo

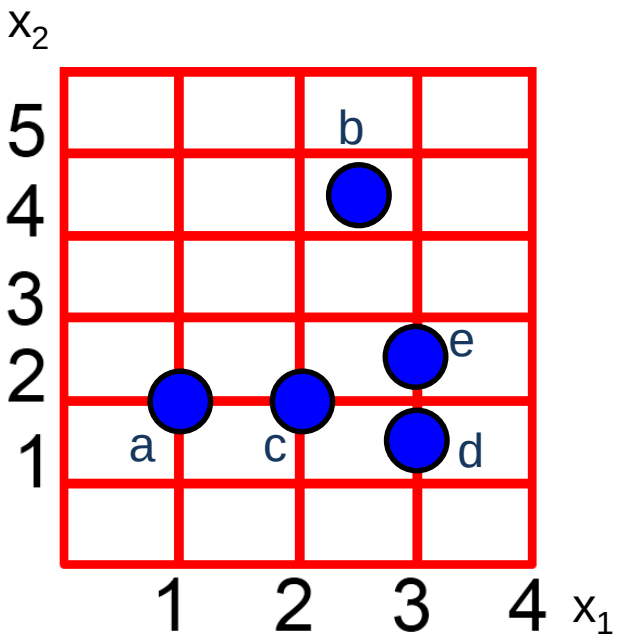
- **Distancias entre clusters:**
  1. Single linkage clustering (“nearest neighbor”).
  2. Complete linkage clustering (“farthest neighbor”).
  3. Average linkage clustering.
  4. Centroid, Ward, etc…



Clustering: Jerárquico Aglomerativo

- Descripción algoritmo:

1. Calcula las distancias entre cada par de elementos.



	a	b	c	d	e
a	0				
b	2.9	0			
c	1.0	2.5	0		
d	3.0	3.4	2.1	0	
e	3.0	2.5	2.1	1.0	0

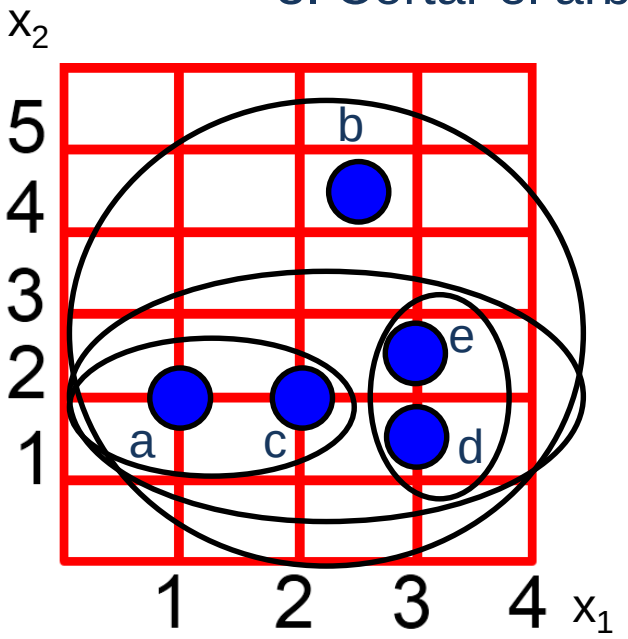
Clustering: Jerárquico Aglomerativo

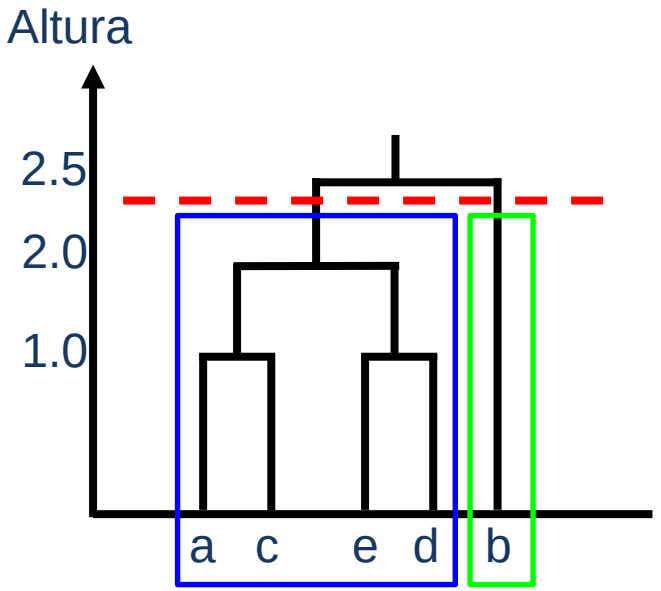
- Descripción algoritmo:

1. Calcula las distancias entre cada par de elementos.

2. Agrupa Iterativamente los puntos en un árbol jerárquico binario. Conecta los pares más cercanos y recalcula la matriz de distancia hasta obtener un solo cluster.

3. Cortar el árbol y obtener clusters.

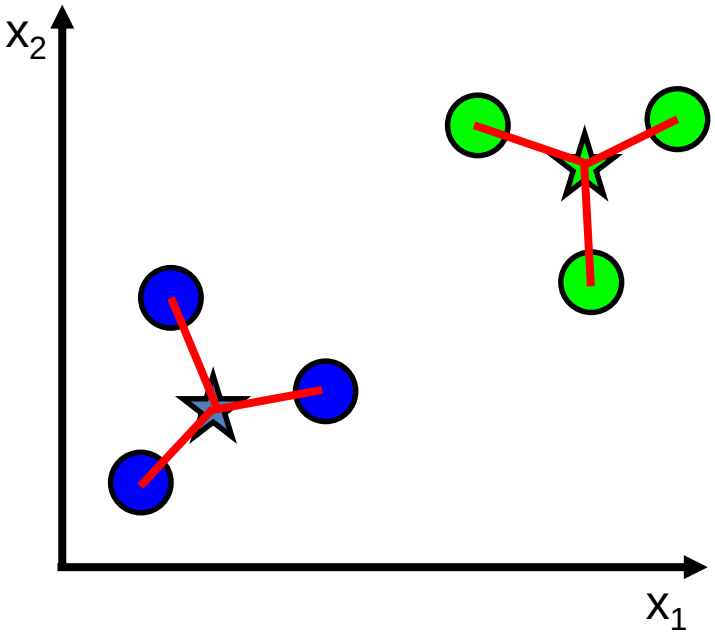




Altura: Distancia a la que el par fue conectado.

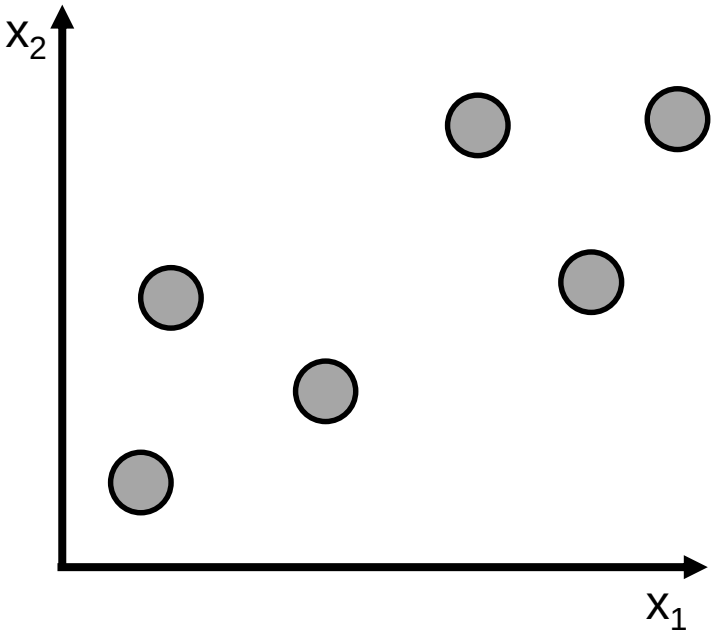
Clustering: K-means

- Se fija previamente la cantidad de clusters a formar.
- Minimiza la varianza total dentro de cada clusters.



Clustering: K-means

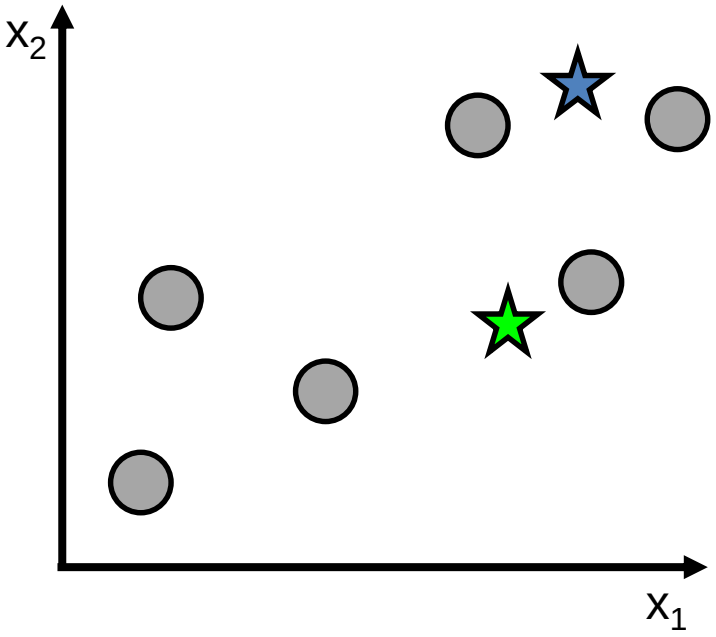
- Se fija previamente la cantidad de clusters a formar.
- Minimiza la varianza total dentro de cada clusters.



Yo sólo tengo observaciones…
No sé cuántos clusters hay en los datos, la posición de centroides, etc…

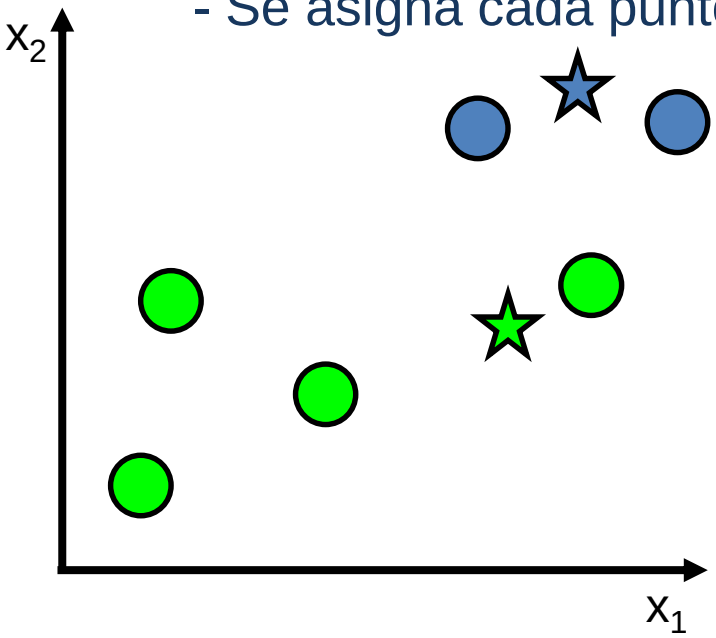
Clustering: K-means

- Descripción algoritmo:
  - Se fija previamente la cantidad de clusters a formar, ejm. k = 2.
  - Se elige al azar la posición de los k centroides.



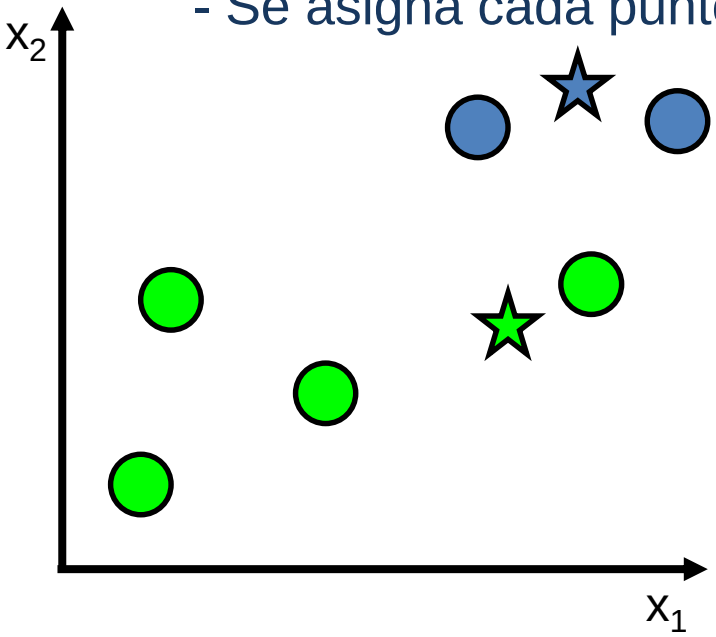
Clustering: K-means

- Descripción algoritmo:
  - Se fija previamente la cantidad de clusters a formar, ejm. k = 2.
  - Se elige al azar la posición de los k centroides.
  - Se asigna cada punto al centroide “más cercano” (distancias vistas).



Clustering: K-means

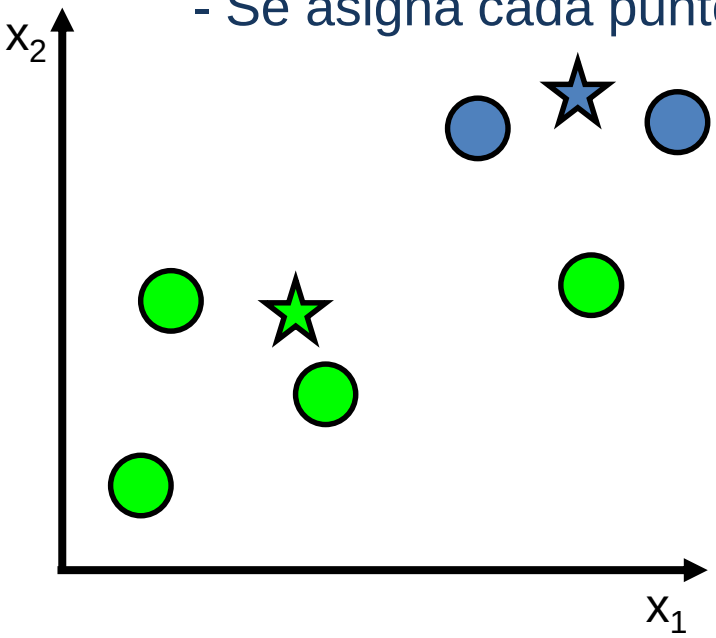
- Descripción algoritmo:
  - Se fija previamente la cantidad de clusters a formar, ejm. k = 2.
  - Se elige al azar la posición de los k centroides.
  - Se asigna cada punto al centroide “más cercano” (distancias vistas).



  - Calcula la varianza dentro del cluster.

Clustering: K-means

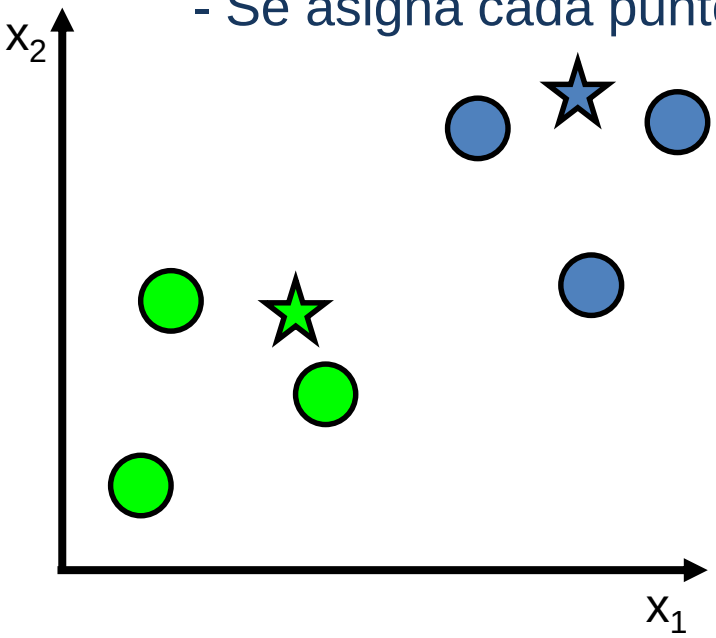
- Descripción algoritmo:
  - Se fija previamente la cantidad de clusters a formar, ejm. k = 2.
  - Se elige al azar la posición de los k centroides.
  - Se asigna cada punto al centroide “más cercano”.



- Calcula la varianza dentro del cluster.
- Reubica la posición del centroide.

Clustering: K-means

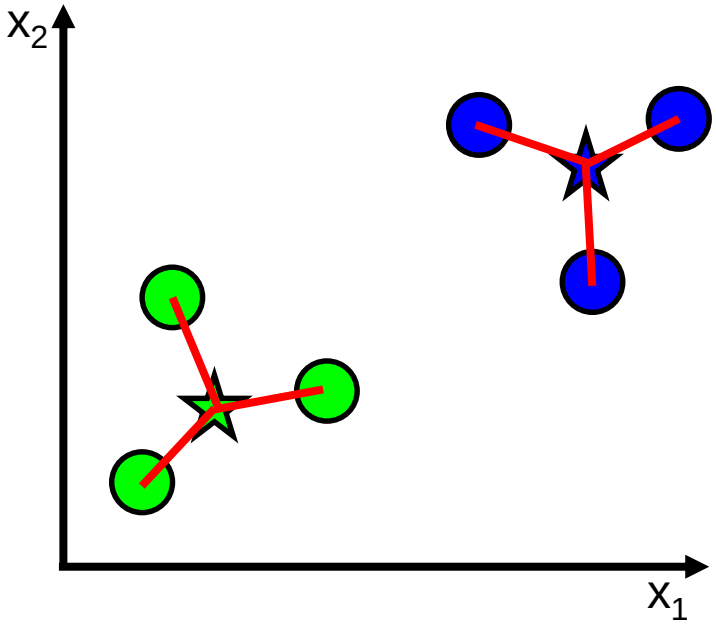
- Descripción algoritmo:
  - Se fija previamente la cantidad de clusters a formar, ejm. k = 2.
  - Se elige al azar la posición de los k centroides.
  - Se asigna cada punto al centroide “más cercano”.



- Calcula la varianza dentro del cluster.
- Reubica la posición del centroide.
- Reasigna cada punto al centroide “más cercano”.

Clustering: K-means

- Descripción algoritmo:
  - Se fija previamente la cantidad de clusters a formar, ejm. k = 2.
  - Se elige al azar la posición de los k centroides.
  - Se asigna cada punto al centroide “más cercano” (distancias vistas).



- Calcula la varianza dentro del cluster.
- Reubica la posición del centroide.
- Reasigna cada punto al centroide “más cercano”.

Loop hasta convergencia (varianza intracluster no cambie) y se detiene!

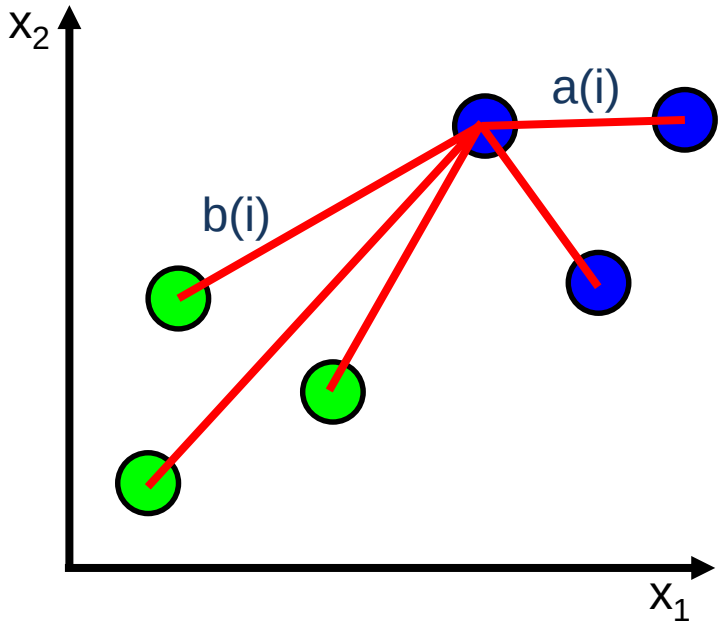
Clustering: K-means

- Cómo elegir K:

Silhouette value: medida de cuán similar es un punto respecto a puntos en el mismo cluster comparado a puntos en otros clusters.

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

$$-1 \leq s(i) \leq 1$$



a(i): distancia promedio entre el punto i respecto a puntos del mismo cluster.

b(i): la menor distancia promedio entre el punto i, y cualquier otro cluster, del cual no sea un miembro.

Clustering: K-means

- Cómo elegir K:

Silhouette value: medida de cuán similar es un punto respecto a puntos en el mismo cluster comparado a puntos en otros clusters.

$$s(i) = \frac{b(i) - a(i)}{Max(b(i), a(i))} \qquad -1 \leq s(i) \leq 1$$

Silhouette coefficient: es un promedio de los valores Silhouette de todos los puntos.

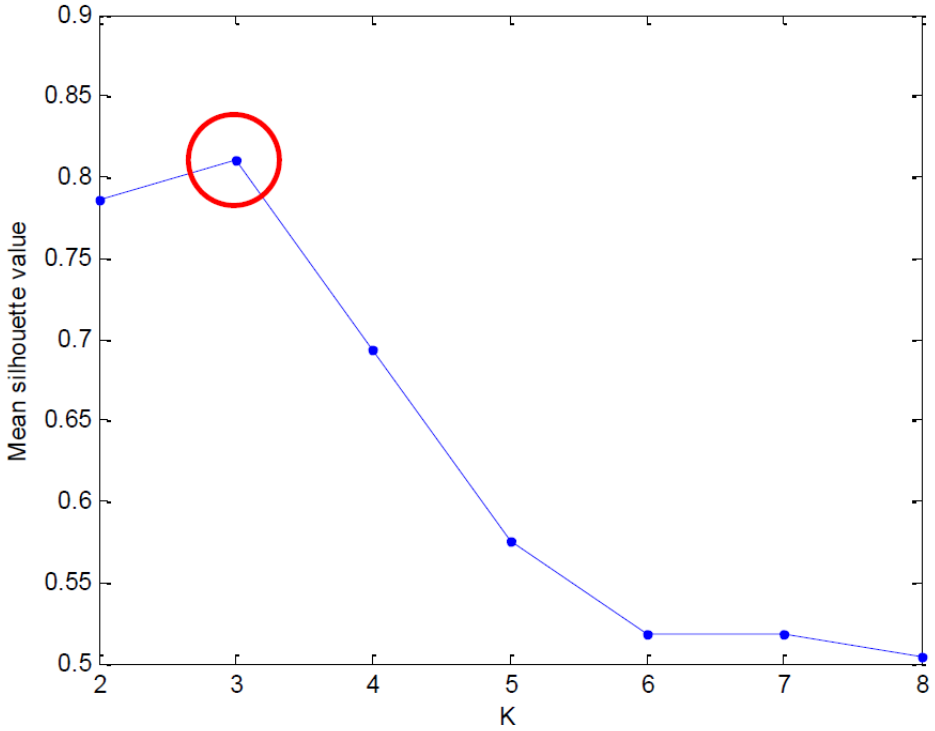
~= 1 ; el objeto está bien clasificado.

~ =0 ; el objeto están en el borde entre dos clusters.

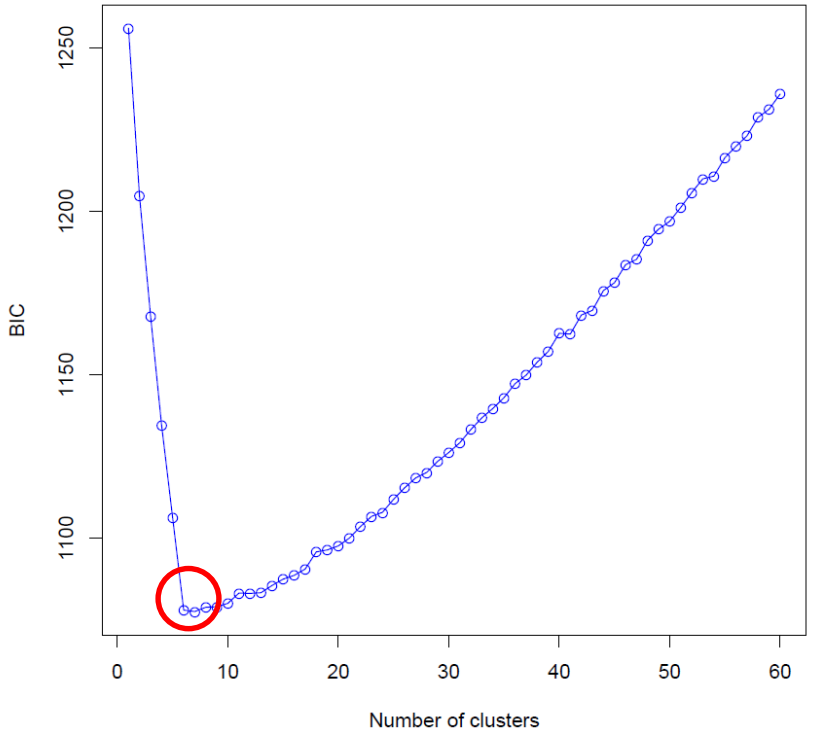
~= -1 ; el objeto está mal clasificado.

Clustering: K-means

Silhouette Coefficient



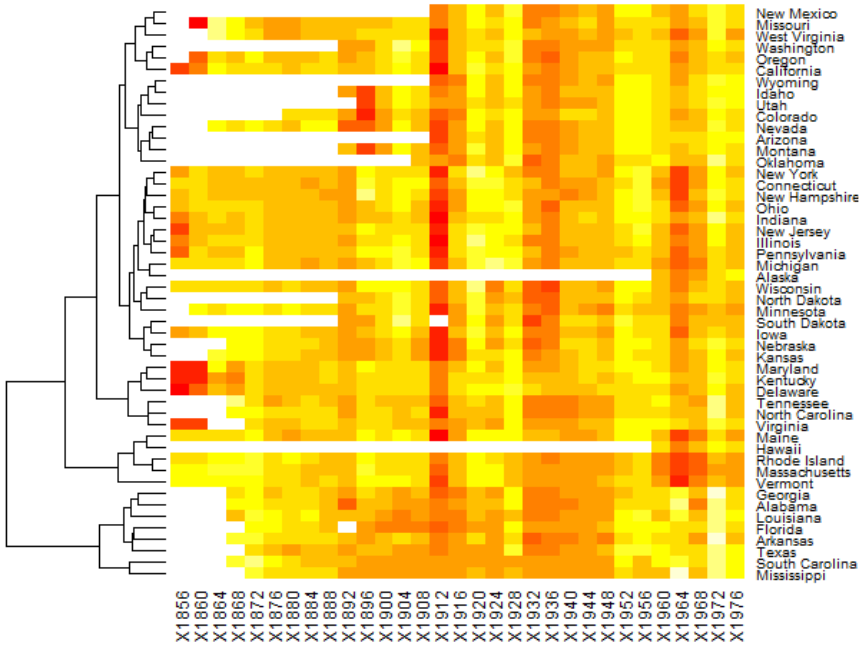
Bayesian Information Criterion (BIC)



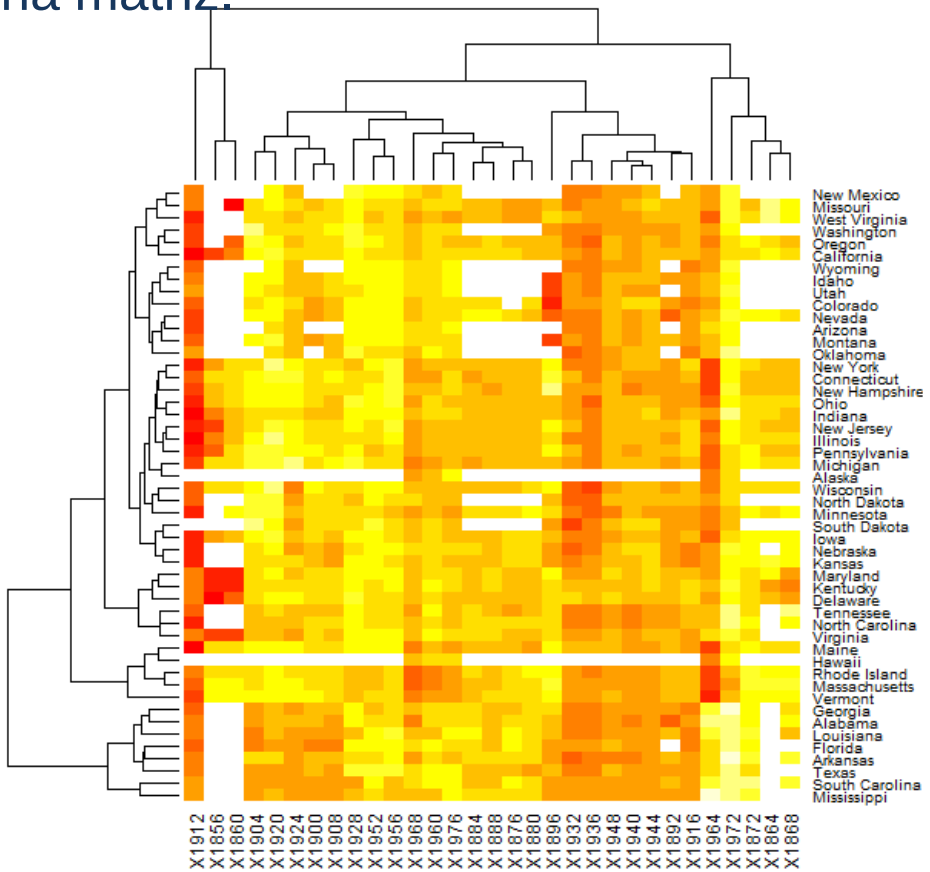
(próxima clase)

Clustering: Biclustering

- Misma idea metodológica anterior, pero ahora agrupamos simultaneamente filas y columnas de una matriz.



Agrupación por Estado



Agrupación por Estado y Año de votación

CLUSTERING Y CLASIFICACIÓN

OTROS MÉTODOS DE CLUSTERING:

- K-medioids
- K-medians.
- SOM.

Clustering: Aplicación cancer de mama

Mortalidad:

- 1º causa mundial de muerte en la mujer.
- Año 2008: en chile 2º lugar en muertes por cáncer en la mujer después del cáncer de vesícula y vías biliares.

Receptor de Estrógeno Positivo	Luminal A Luminal B Luminal HER2/neu
Receptor de Estrógeno Negativo	Basal HER2/neu Normal Like

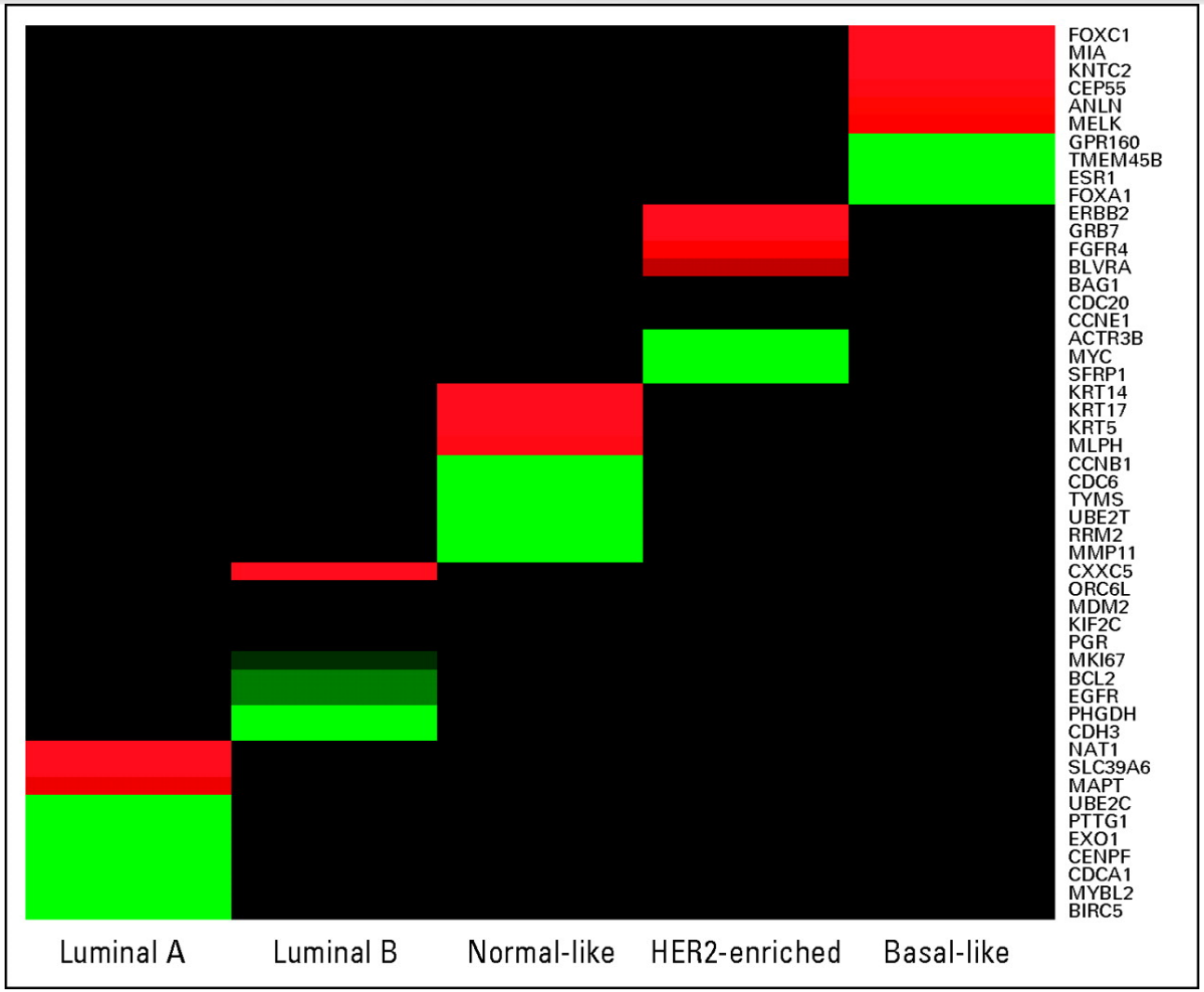
Subtipo	RE	RP	HER2/neu	Índice de proliferación Ki-67
Luminal A	+	y/o +	-	Bajo (< 14%)
Luminal B	+	y/o +	-	Alto (> 14%)
Luminal HER2/neu	-	-	+	Alto

Clustering: Aplicación cáncer de mama

Panel PAM50

MammaPrint (70 genes)

Oncotype DX (21 genes)



Parker J S et al. JCO 2009;27:1160-1167

CLUSTERING Y CLASIFICACIÓN

CLUSTERING

- Clustering Jerárquico:
  - Clustering Aglomerativo.
- K-means.
- Biclustering.

CLASIFICACIÓN

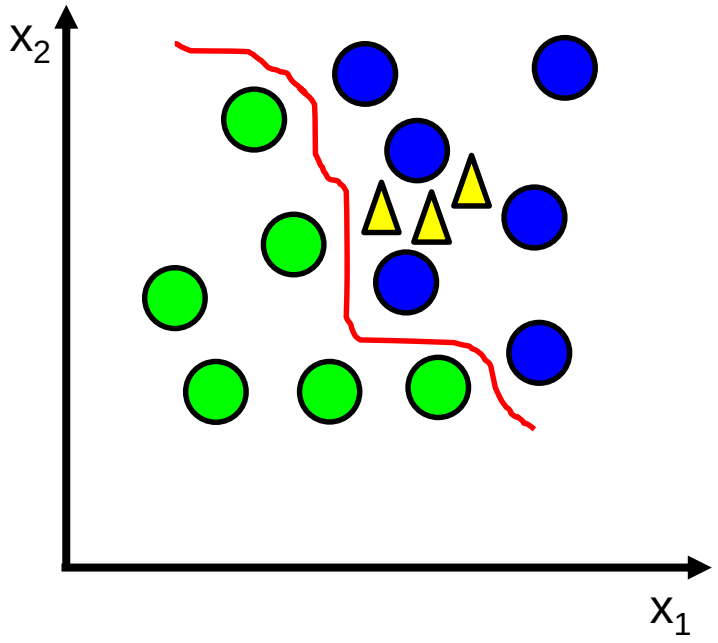
- KNN: *k-nearest neighbors.*
- L/QDA: *Linear/Quadratic discriminant analysis.*
- SVM: *Support vector machine.*
- CART: *classification and regression trees.*
- *Random Forest*.

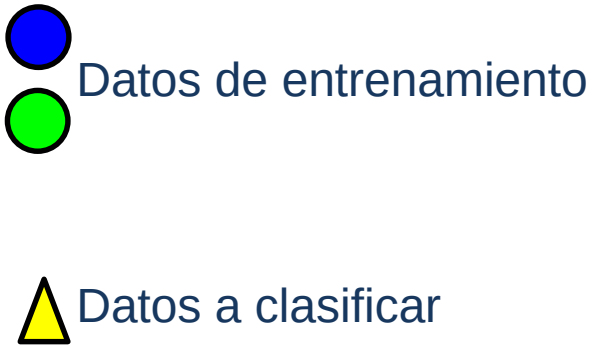
CLASIFICACIÓN: Idea general

- **Métodos supervisados; requieren variables predictoras (clases).**

Dividir el espacio de los vectores de entrada en superficies/límites de decisión.

Teniendo un vector de entrada $x \in R^D$, le asignamos una de K clases discretas C_k, k =1, …, K.





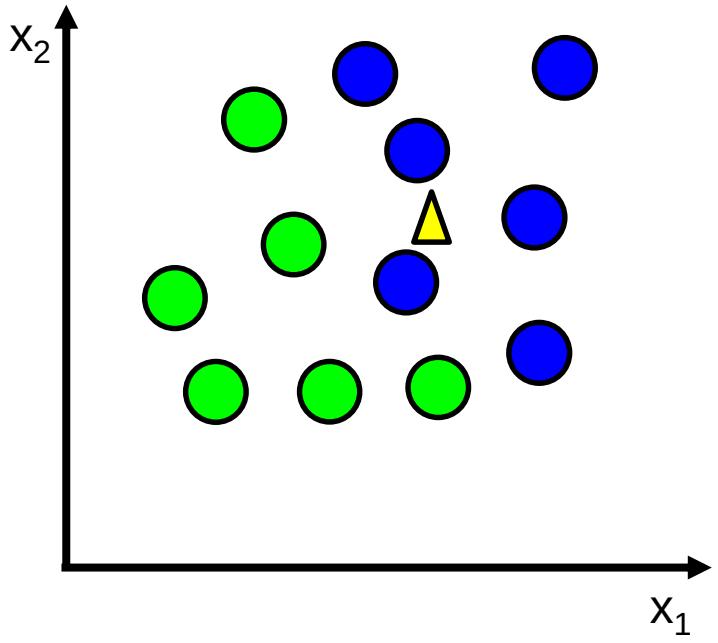
CLUSTERING Y CLASIFICACIÓN

MÉTODOS DE CLASIFICACIÓN A REVISAR

- KNN: *k-nearest neighbors.*
- L/QDA: *Linear/Quadratic discriminant analysis.*
- SVM: *Support vector machine.*
- CART: *classification and regression trees.*
- *Random Forest*.

CLASIFICACIÓN: K-Nearest Neighbors (KNN)

- Uno de los métodos más simples e intuitivos de clasificiación.
- Asume que los datos de entrenamiento son puntos en un espacio de d-dimensiones.
- La clase de una nueva observación se estima en base a lo valores más cercanos de los datos de entrenamiento.

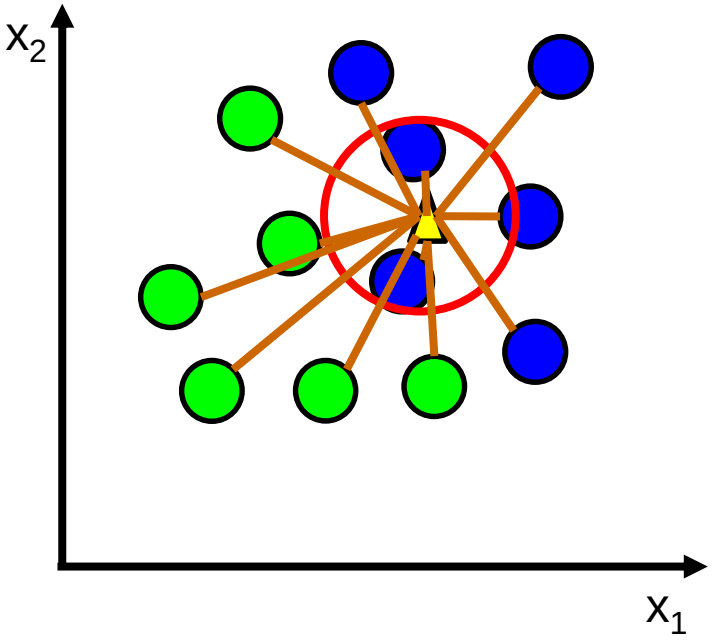


Requiere:
-Parámetro K: en cuántos vecinos fijarse.

-Función de distancia.

CLASIFICACIÓN: K-Nearest Neighbors (KNN)

- Calcula las distancia entre el punto de prueba y los presentes en los datos de entrenamiento.
- Ordena las distancias y selecciona los k vecinos más cercanos.
- Asigna la clase al punto de prueba en función de la clase mayoriataria en esos k vecinos.



Ejemplo trivial en 2D, k=2:

Al aumentar la dimensionalidad de los datos de entrenamiento la asignación de clases deja de ser obvia.

CLASIFICACIÓN: Linear (and Quadratic) discriminant analysis (L/QDA) (~1936).

- El objetivo es llevar a cabo una reducción de la dimensionalidad de los datos preservando tanto como sea possible la información que permita discriminar las clases presentes.

- Se busca una combinación lineal (o cuadráticas) de variables que caractericen o diferencien dos o más clases de objetos.

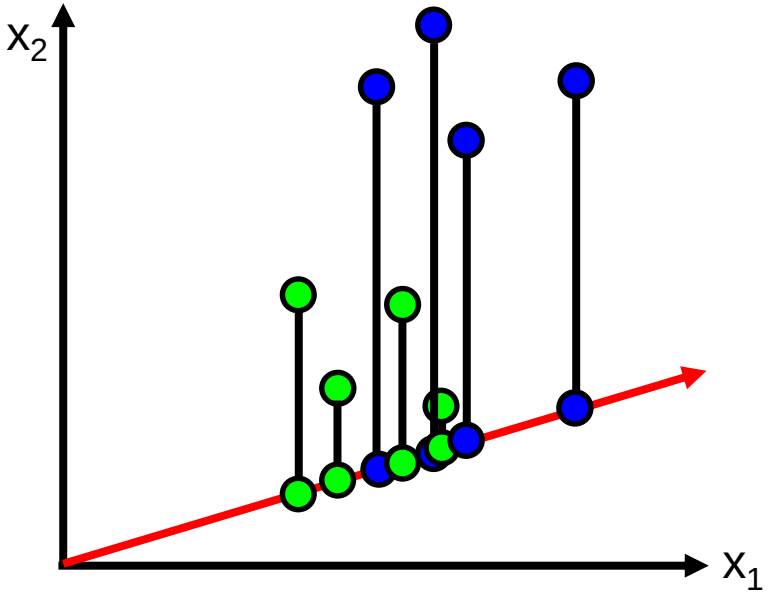
- Estas funciones discriminantes maximizan el cuociente entre la varianza entre grupos y la varianza dentro de grupos.

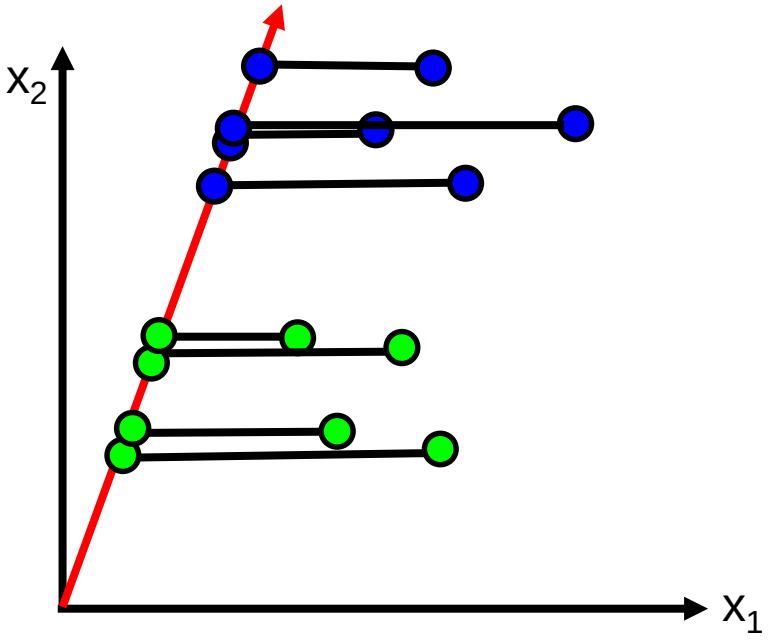
CLASIFICACIÓN: Linear Discriminant Analysis (LDA)

- Caso más simple, dos clases y dos variables:

Se busca un escalar proyectando las observaciones sobre la línea que mejor separe las clases.

De todas las posibles líneas, seleccionamos la que maximiza la separación.





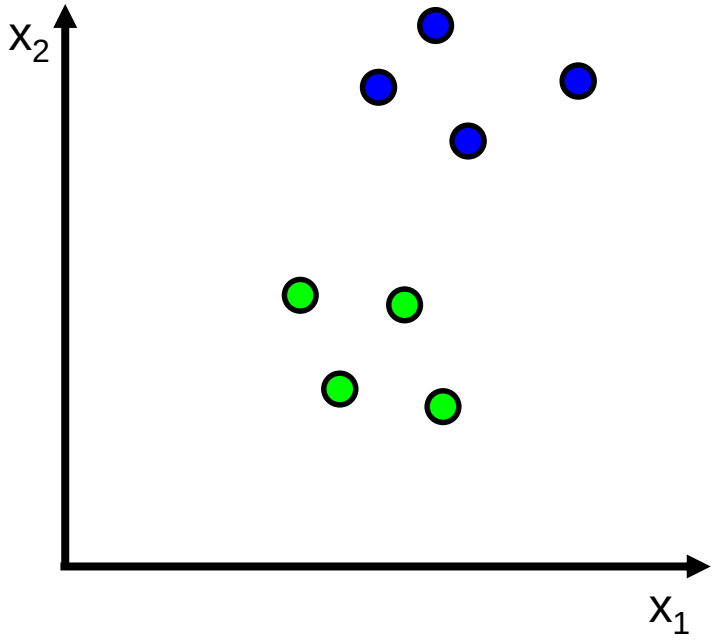
CLASIFICACIÓN: Linear (and Quadratic) discriminant analysis (L/QDA)

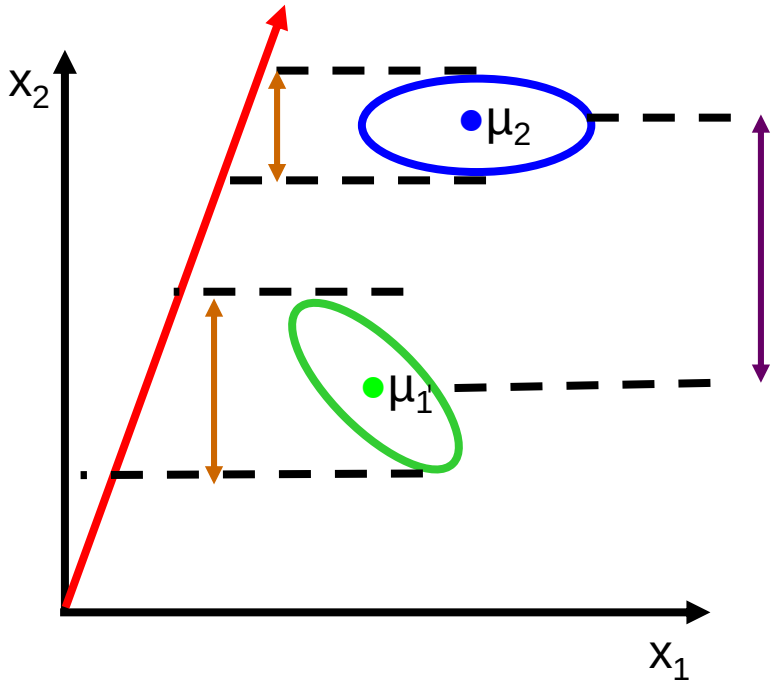
- Caso más simple, dos clases y dos variables:

$$J(w) = \frac{|\widetilde{\mu_1} - \widetilde{\mu_2}|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} = \cdots = J(w) = \frac{W^T S_B W}{W^T S_W W} \rightarrow w* = argmax_w \left\{ \frac{W^T S_B W}{W^T S_W W} \right\}$$

-Derivar.
-Igualar a 0.
-Resolver problema de valores propios.

S_B: *Between class Scatter Matrix* ~ *equivalente a varianza entre grupos*
S_W: *Within class Scatter Matrix*~ *equivalente a varianza dentro grupos*



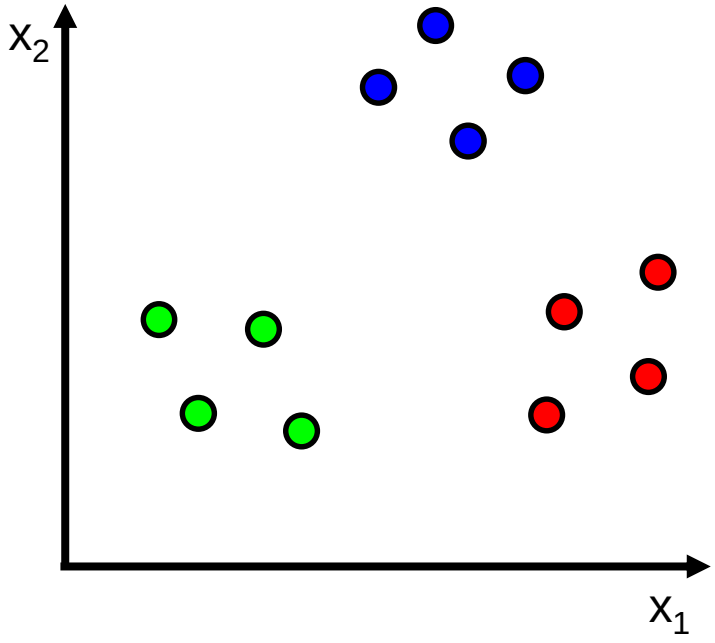


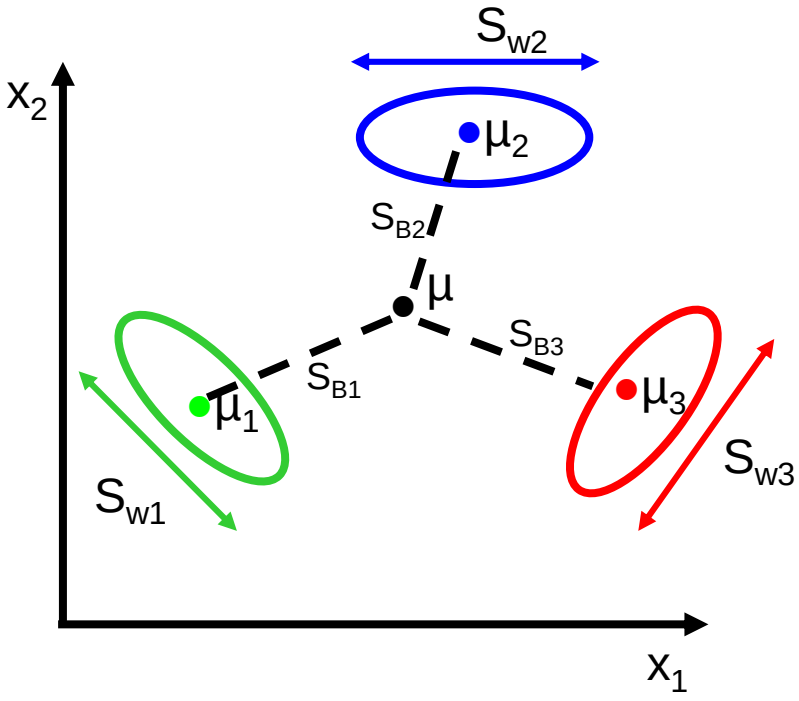
CLASIFICACIÓN: Linear (and Quadratic) discriminant analysis (L/QDA)

- Caso más simple, dos clases y dos variables:

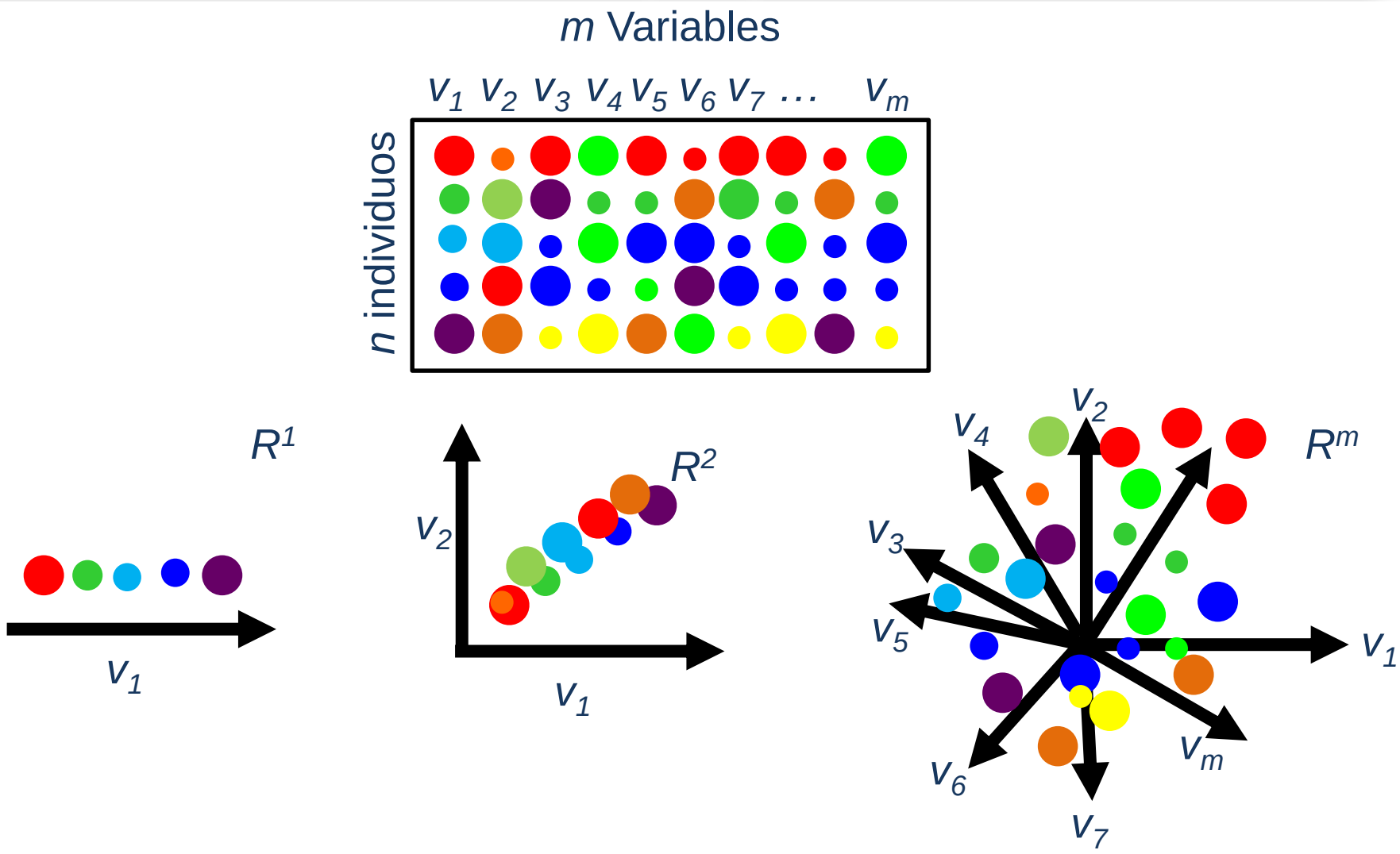
$$J(w) = \frac{W^T S_B W}{W^T S_W W} \quad \rightarrow \quad w* = [w_1|w_2|\ldots|w_{C\_1}] = \underset{w}{argmax}\left\{\frac{|W^T S_B W|}{|W^T S_W W|}\right\}$$

S_B: *Between class Scatter Matrix* ~ *equivalente a varianza entre grupos*
S_W: *Within class Scatter Matrix*~ *equivalente a varianza dentro grupos*

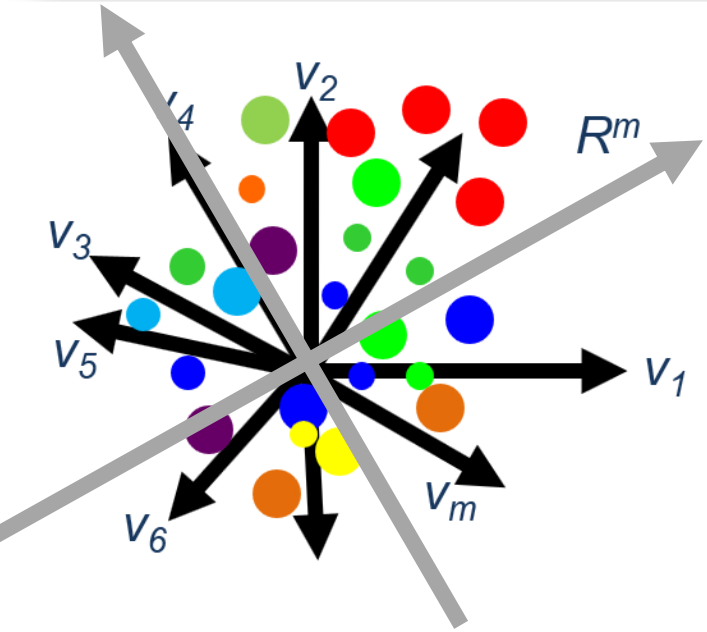


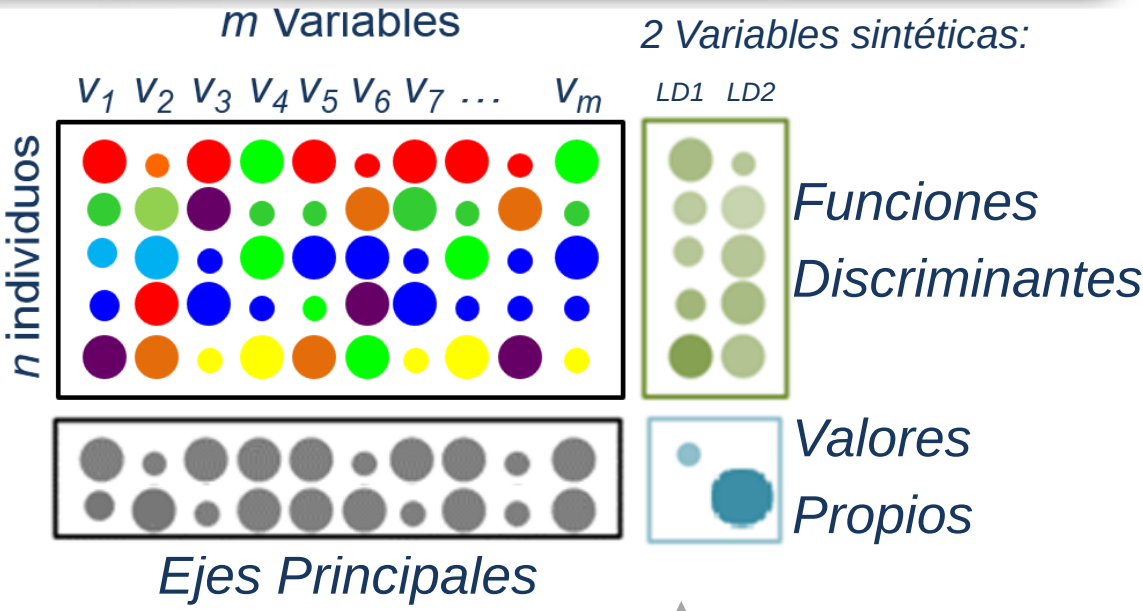


CLASIFICACIÓN: Reducción de la dimensionalidad



CLASIFICACIÓN: Reducción de la dimensionalidad

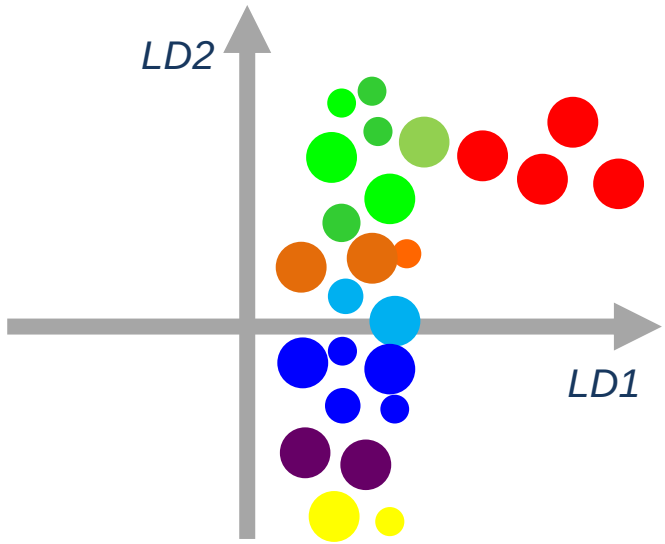




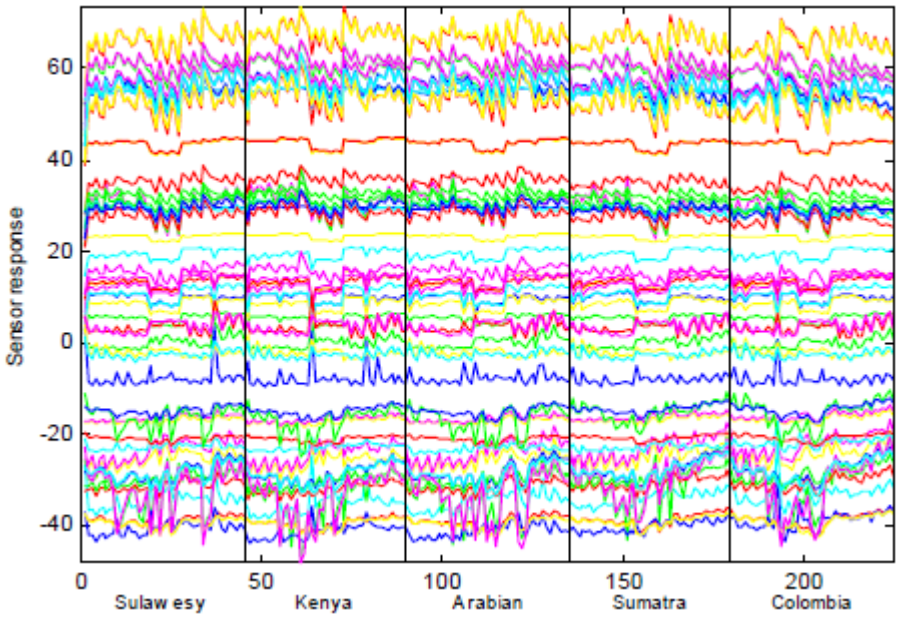
$$LDi = a_{1i}v_1 + a_{2i}v_2 + \ldots + a_{mi}v_m$$

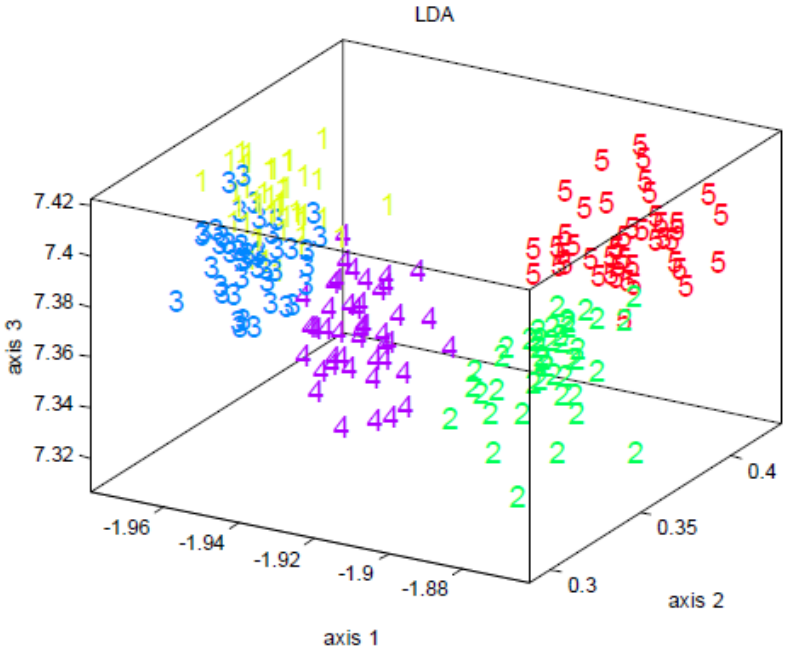
$$Cov(LDj;LDk) = 0$$

$$J(w) = \frac{W^T S_B W}{W^T S_W W}$$



CLASIFICACIÓN: Linear (and Quadratic) discriminant analysis (L/QDA).





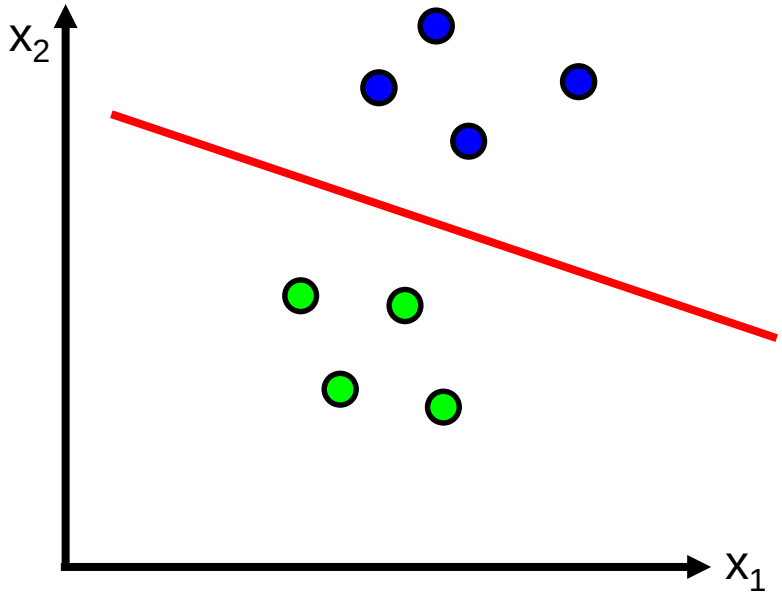
Con las coordenadas del LDA podemos asignar clase a nuevos datos.

LDA tiene limitaciones, por que existen otras alternativas.

CLASIFICACIÓN: Support Vector Machine (SVM) (~1990).

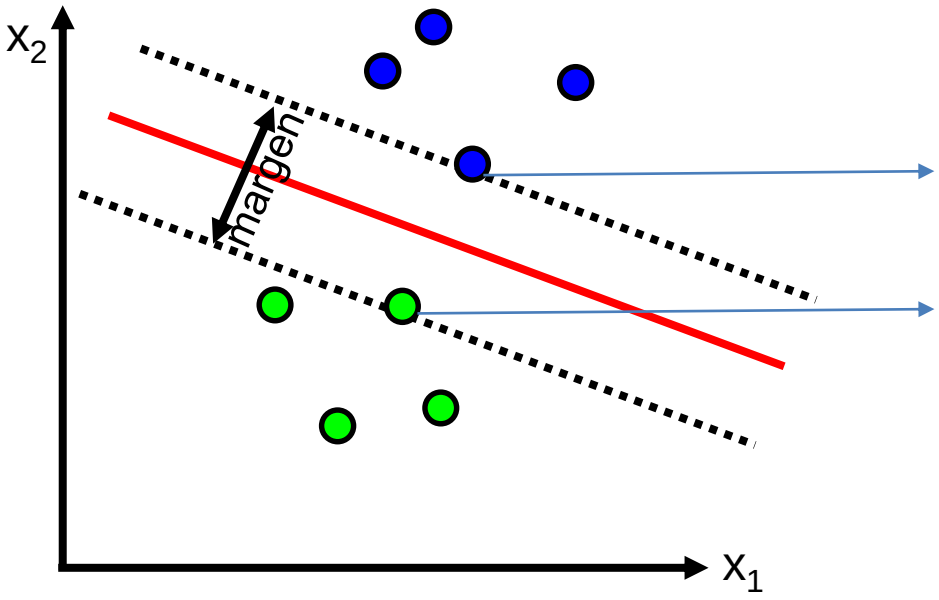
Método de aprendizaje supervisado.

- Determinan hiperplano óptimo para patrones linealmente separables.
- Extensión a patrones no linelmente separables vía transfomaciones de los datos originales ubicándolos en un nuevo espacio - Funciones Kernel.



CLASIFICACIÓN: Support Vector Machine (SVM) (~1990).

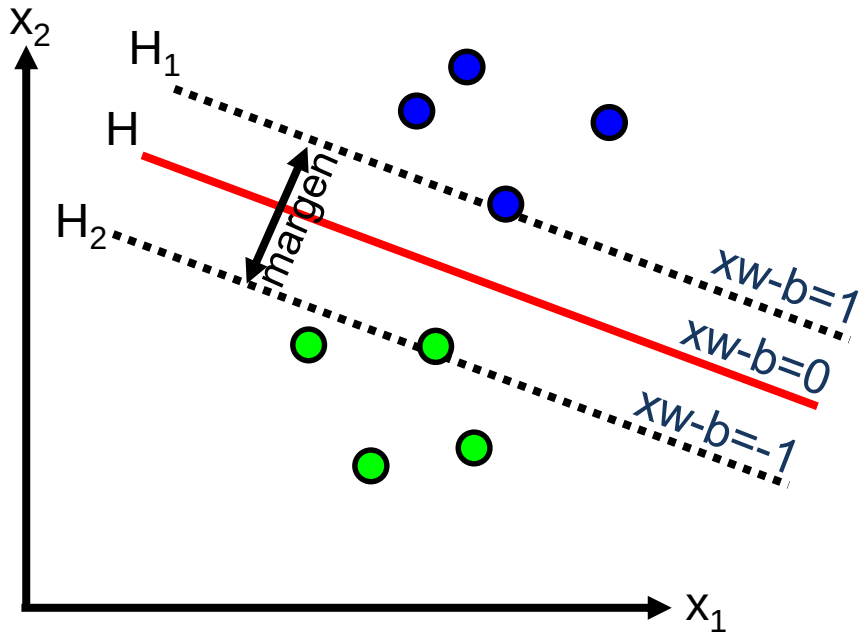
- SVMs maximizan el margen alrededor del plano separador.
- La función de decision está completamente especificada por un subconjunto de la muestra de entrenamiento, los vectores de soporte.



Vectores de soporte:

muestras más cercanas al hiperplano.

CLASIFICACIÓN: Support Vector Machine (SVM) (~1990).



Definimos el hiperplano H tal que:

$x_i w+b \geq 1$ cuando $y_i = 1$

$x_i w+b \leq 1$ cuando $y_i = -1$

Recordar que distancia entre un punto (x_0, y_0) a una recta $ax+by+c=0$ es:

$$\frac{ax_0 + by_0 + c}{\sqrt{a^2 + b^2}}$$

La distancia entre H y H1 es:

$$\frac{|wx + b|}{\|w\|} = \frac{1}{\|w\|}$$

La distancia entre H1 y H2 es:

$$\frac{2}{\|w\|}$$

CLASIFICACIÓN: Support Vector Machine (SVM) (~1990).

La distancia entre H1 y H2 es: $\frac{2}{\|w\|}$

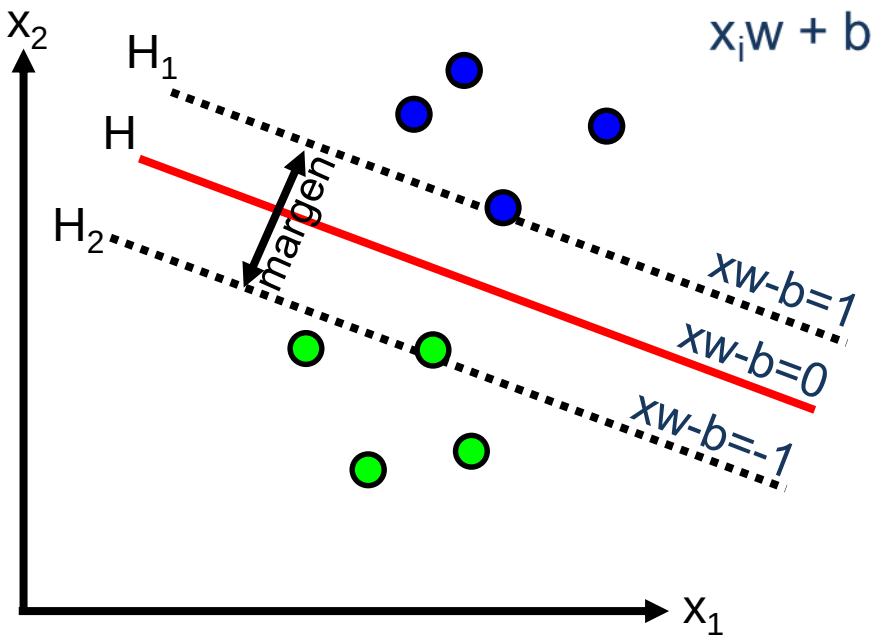
Para maximizar el margen, debemos minimizar $\|w\|$

Con la condici[on de que no hayan puntos entre H1 y H2.

$x_i w + b \geq 1$ cuando $y_i = 1$

$x_i w + b \leq 1$ cuando $y_i = -1$

Combinadas en $y_i(x_i w) \geq 1$



Este es un problema de optimización con restricciones:

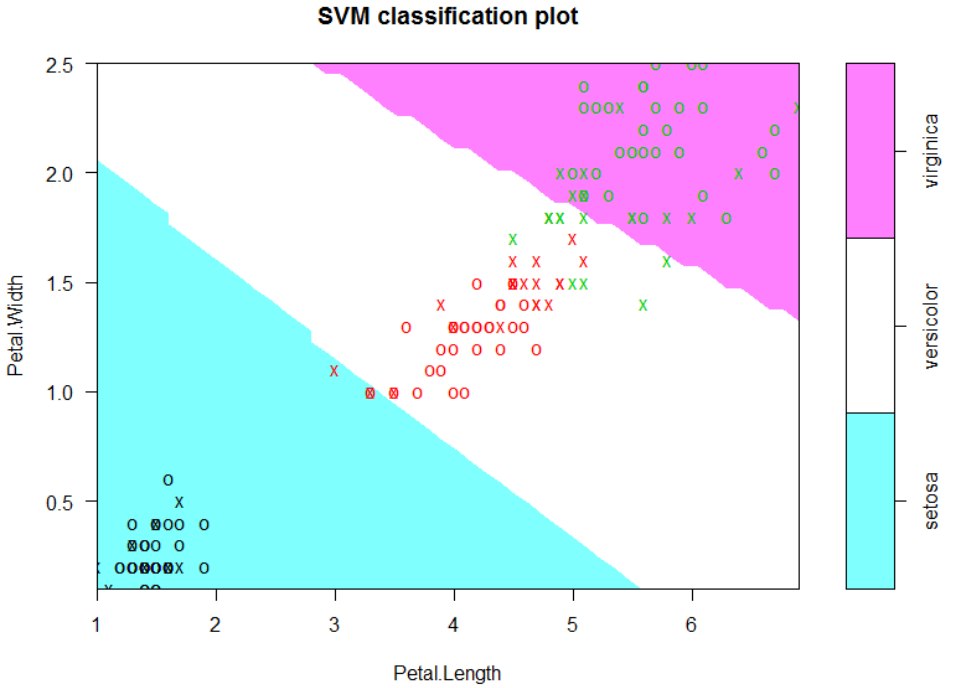
minf(x) donde f: $(1/2)\|w\|$

g: yi(xiw) – b = 1

Se puede resolver usando el método de multiplicadores de Lagrange.

CLASIFICACIÓN: Support Vector Machine (SVM) (~1990).

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa



CLASIFICACIÓN: Classification And Regression Trees (CART) (~1980).

- Algoritmo de partición recursivo que busca variables de decisión que permitan dividir el conjunto de datos en dos subconjuntos más pequeños.

- En cada paso se identifica una pregunta que divide nodos en dos ramas cuyas homogenidades sean máximas.

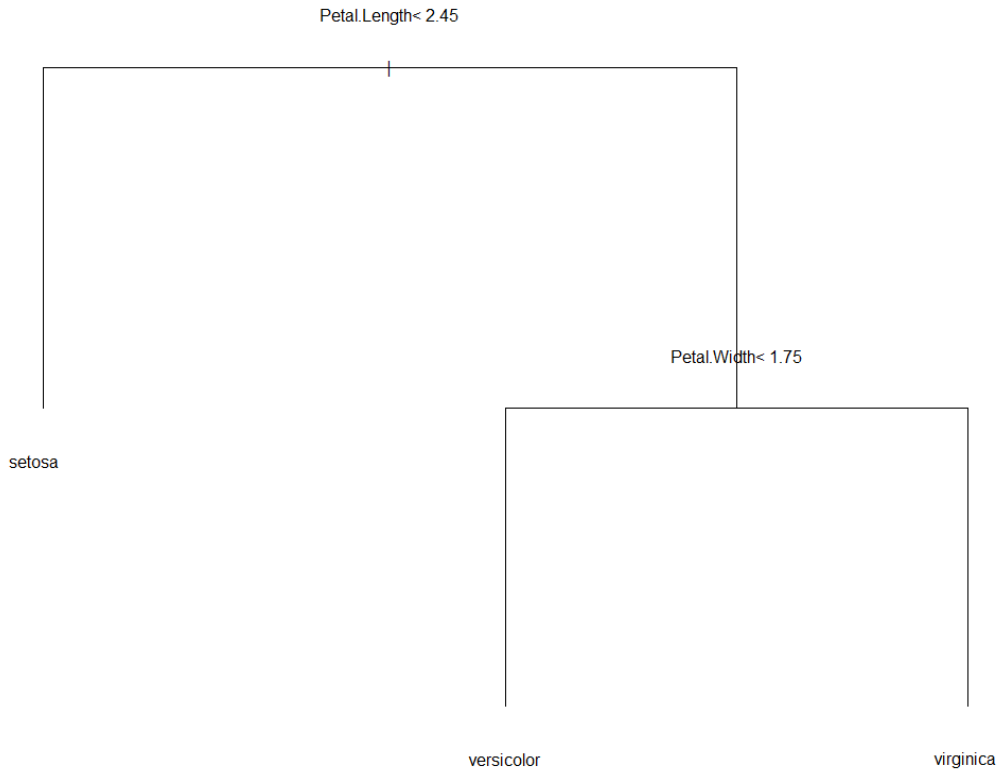
- La homogeneidad se evalúa mediante medidas como Entropía, Índice de Gini, o Ganancia de Información.

- Fáciles de interpretar, sirven para datos categóricos y continuos.

- Los árboles pueden ser podados a cierto nivel para tener mayor coherencia.

CLASIFICACIÓN: Classification And Regression Trees (CART) (~1980).

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
53	6.9	3.1	4.9	1.5	versicolor
54	5.5	2.3	4.0	1.3	versicolor
55	6.5	2.8	4.6	1.5	versicolor
56	5.7	2.8	4.5	1.3	versicolor
57	6.3	3.3	4.7	1.6	versicolor
103	7.1	3.0	5.9	2.1	virginica
104	6.3	2.9	5.6	1.8	virginica
105	6.5	3.0	5.8	2.2	virginica
106	7.6	3.0	6.6	2.1	virginica
107	4.9	2.5	4.5	1.7	virginica
108	7.3	2.9	6.3	1.8	virginica
109	6.7	2.5	5.8	1.8	virginica



CLASIFICACIÓN: Classification And Regression Trees (CART) (~1980).

- Índice de Gini: $$Gini(n) = 1 - \sum_{x \in X} P(n,x)^2$$

P(n, x): Probabilidad de seleccionar la clase x en el nodo n.

Gini(n) es la probabilidad de no seleccionar dos elementos de la clase x en el mismo nodo.

Mientras menor sea, mayor es la pureza del Split.

- Gini Split

Dado un split Sn = {S1, …, Sk} del nodo n, donde |n| es la cardinalidad de elementos en el nodo n

$$GiniSplit(n,S) = \sum_{x\ S} \frac{|s|}{|n|} Gini(s)$$

Seleccionamos para ramificar, aquel split con menor GiniSplit.

CLASIFICACIÓN: Classification And Regression Trees (CART) (~1980).

- Entropía: mide la impureza de los datos en el nodo n.

$$Entropía(n) = -\sum_{x \epsilon X} P(n,x)\, log(P(n,x))$$

P(n, x): Probabilidad de seleccionar la clase x en el nodo n.

Mientras menor sea, mayor es la pureza del Split.

Entropía = 1 : la misma cantidad de objetos de cada clase.

Entropía = 0 : Todos los objetos son de una clase

- Ganancia de Información: Reducción en la entropía casada por conocer el valor de un atributo.

$$Ganancia(n,S) = Entropía(n) - \sum_{s \epsilon S} \frac{|s|}{|n|} Entropía(s)$$

Dado un split Sn = {S1, …, Sk} del nodo n, se elige el que presente la mayor ganancia de información.

CLASIFICACIÓN: Random Forest.

- Esencialmente el algoritmo implica la generación de múltiples árboles de decisión, tomando al azar un subconjunto de la muestra para generar las preguntas que llevan a los splits.

- Por lo tanto, en vez de obtener un modelo generado por un árbol, obtenemos un modelo de decision construido en base a un bosque.

CLUSTERING Y CLASIFICACIÓN

CLUSTERING

- Clustering Jerárquico:
  - Clustering Aglomerativo.
- K-means.
- Biclustering.

CLASIFICACIÓN

- KNN: *k-nearest neighbors.*
- L/QDA: *Linear/Quadratic discriminant analysis.*
- SVM: *Support vector machine.*
- CART: *classification and regression trees.*
- *Random Forest*.

PRÁCTICO CLUSTERING Y CLASIFICACIÓN

1. INTRODUCCIÓN A R/BIOCONDUCTOR

2. CLUSTERING

- Clustering Jerárquico:
  - Clustering Aglomerativo.
- K-means.
- Biclustering.

3. CLASIFICACIÓN

- KNN: *k-nearest neighbors.*
- L/QDA: *Linear/Quadratic discriminant analysis.*
- SVM: *Support vector machine.*
- CART: *classification and regression trees.*
- *Random Forest*.