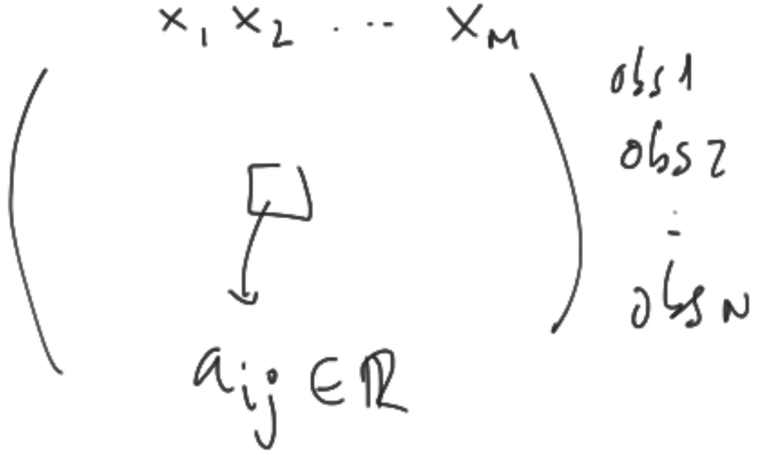
TÉCNICAS DE AGRUPAMIENTO (CLUSTERING)

DATOS NUMÉRICOS : $A_{NOBS \times MFEATURES}$ =



3 TÉCNICAS

- K-means
- aglomerativo / jerárquico

→ similitud - cercanía

- Mezclas Gaussianas (Gaussian mixture models, GMM)

↓

modelo estadístico

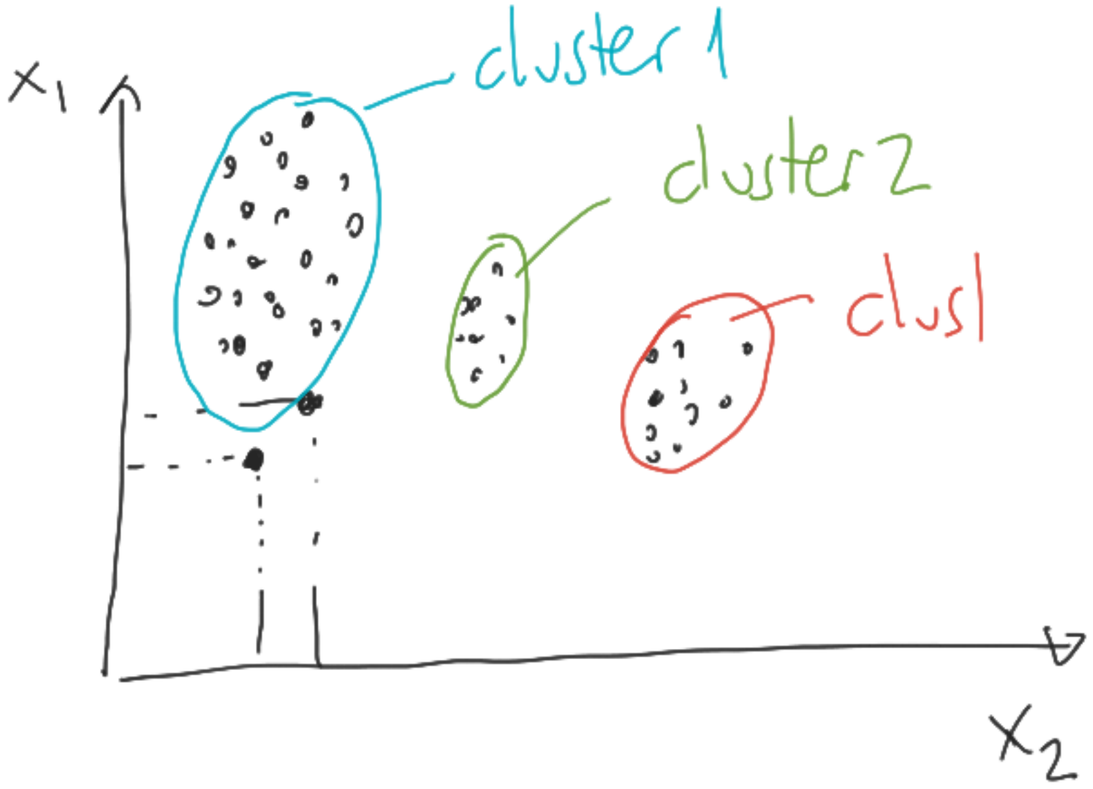
CLUSTERING — no supervisada

(no necesita etiqueta de clase

A, w_i, $i = 1 \dots N$)

$\underline{tx}$: datos 2D $\{x_1, x_2\}$

scatterplot



AI → no-supervisadas

	x_1	x_2	...	x_M
obs 1	0	0		0
obs 2	0	0		0
...				

- clustering
- reducción dimensionalidad (PCA)

→ supervisadas

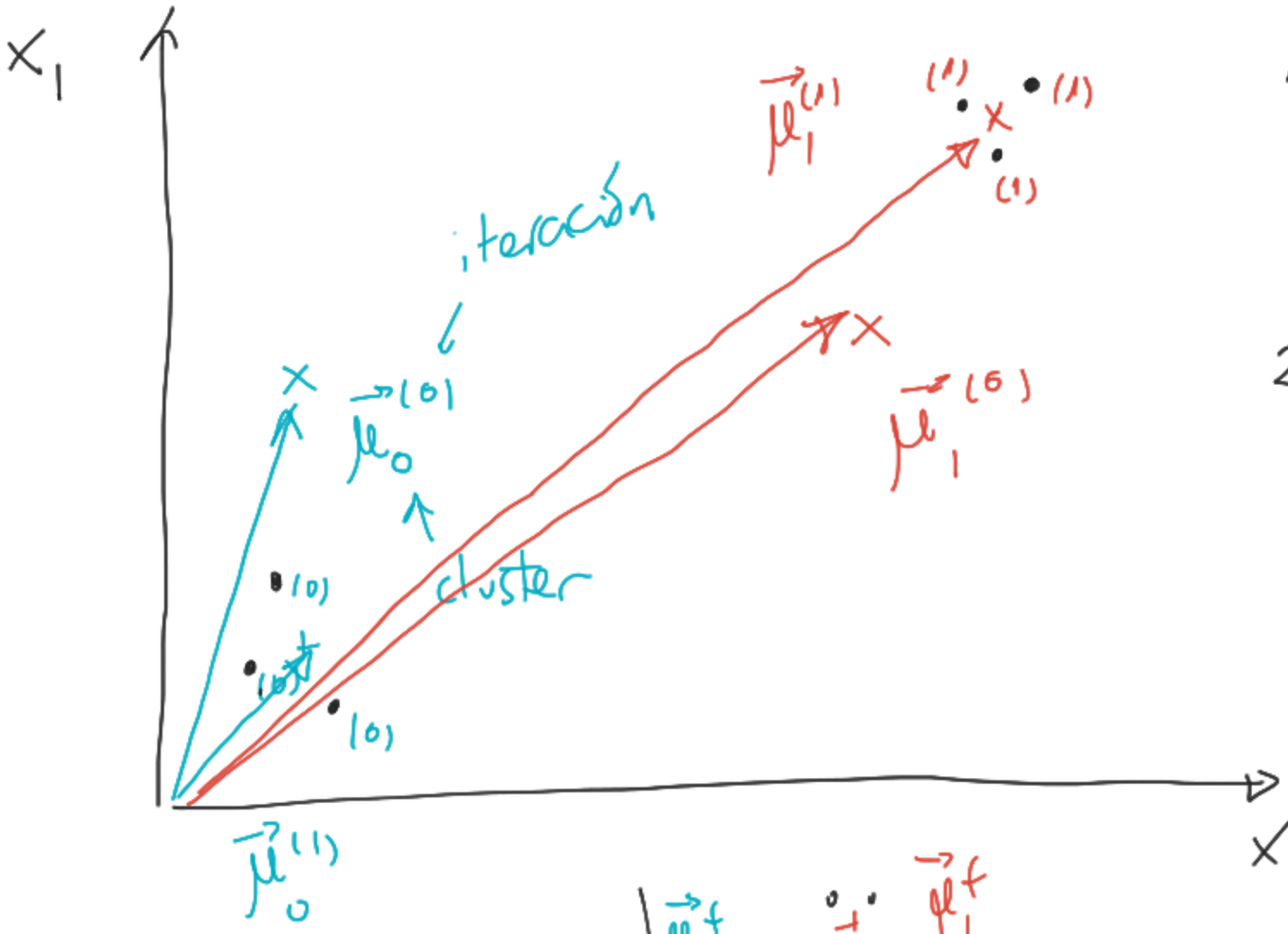
- clasificación
- regresión

	atributos				etiqueta clase
	x_1	x_2	...	x_M	w
obs 1	0	0		0	+ 'coche'
obs 2	0	0		0	+ 'moto'
...					

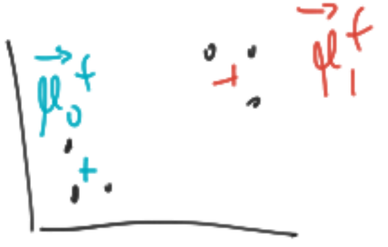
iris

pl	pw	sl	sw	class
1.3	2.4	...		{0, 1, 2}

K-means → algoritmo iterativo basado en distancias entre observaciones

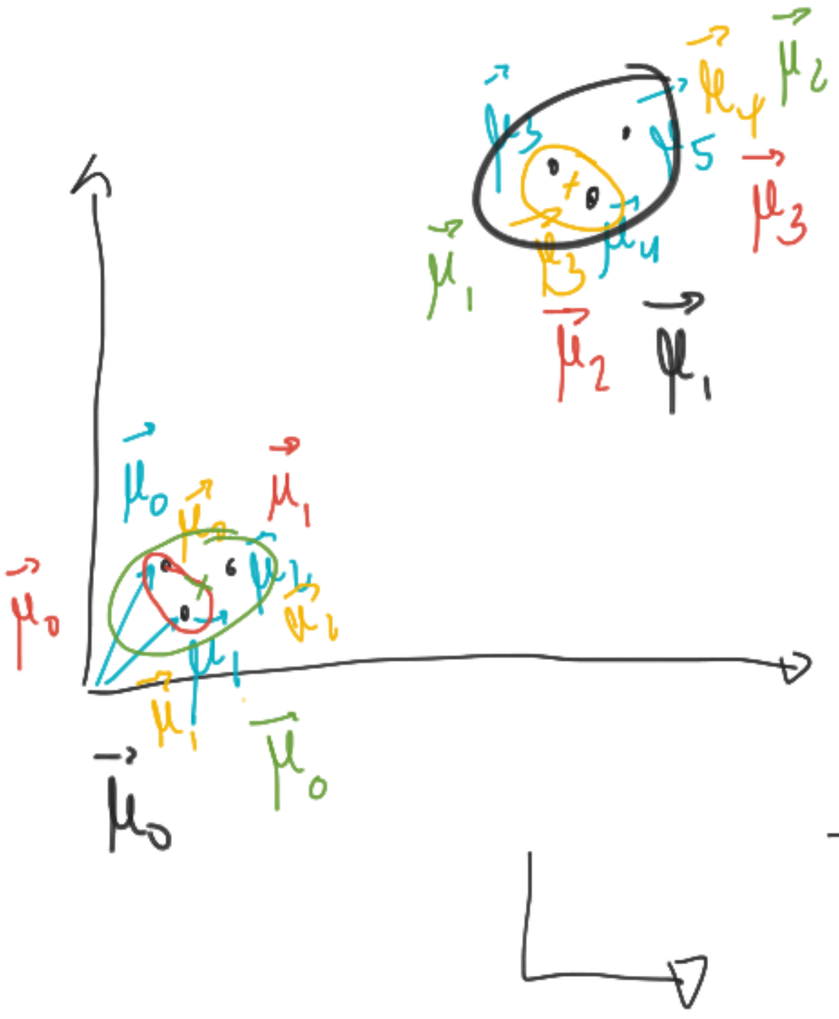


resultado:



1. Establecer el # clústers que se desea identificar
   NCLUSTERS = 2
2. Valores iniciales de los centroides (aleatorio)
   $\vec{\mu}_0^{(0)}$, $\vec{\mu}_1^{(0)}$
3. Calcular distancia entre cada observación y los centroides y asociarla al más cercano
4. Recalcular los centroides
5. Repetir pasos 3 y 4 hasta que los cent.

aglomerativo jerárquico



→ 1. Indicar #clusters ($N = 2$)

2. Inicialización: tantos centroides como observaciones

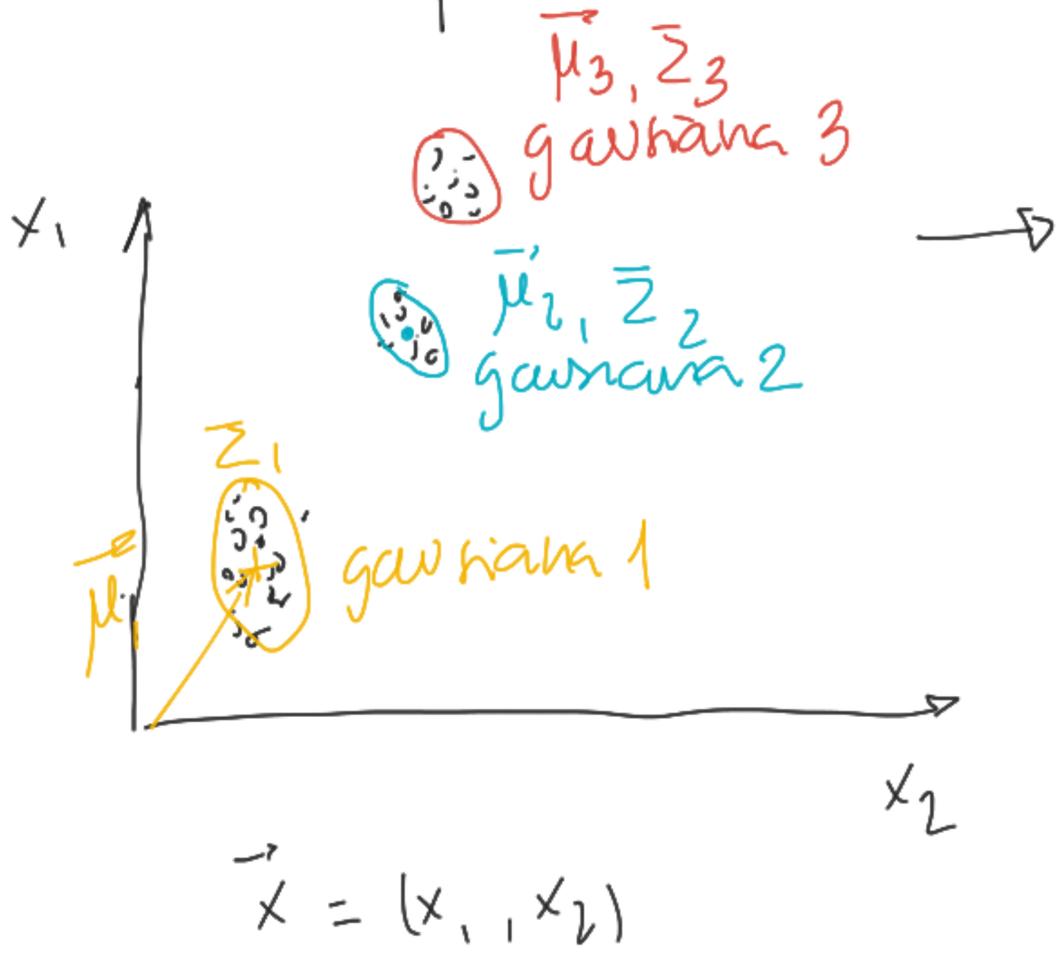
3. Agrupar centroides por distancia

4. Repetir hasta que sólo queden $N = 2$

DENDROGRAMA !

Mezclas gaussianas (GMM)

Modelo probabilístico distribución observaciones
inconveniente: requiere elevado # observaciones



→ Suma ponderada de Gaussianas

$$p(\vec{x}) = \sum_{i=1}^{3} \pi_i \cdot N(\vec{\mu}_i, \bar{\Sigma}_i) =$$

pesos

$$= \pi_1 \cdot N(\vec{\mu}_1, \bar{\Sigma}_1) + \pi_2 \cdot N(\vec{\mu}_2, \bar{\Sigma}_2) +$$

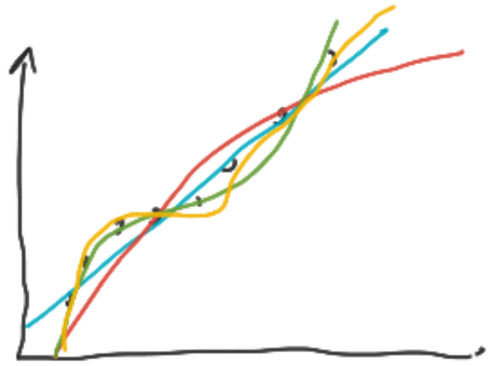
$$\pi_3 \cdot N(\vec{\mu}_3, \bar{\Sigma}_3)$$

Modelo GMM N Gaussianas en d dimensiones

Gaussiana	parámetros gaussianas	pesos
1	$\vec{\mu}_1 (1 \times d)$, $\bar{\Sigma}_1 (d \times d)$	π_1
2	$\vec{\mu}_2 (1 \times d)$, $\bar{\Sigma}_2$	π_2
⋮	⋮	⋮
N	$\vec{\mu}_N$, $\bar{\Sigma}_N$	π_N

ESTIMAR TODOS LOS PARÁMETROS!

How many clusters? → SELECCIÓN MODELOS



GMM + Índice de parsimonia

ERROR vs COMPLEJIDAD
(Residuo) (# parámetros)

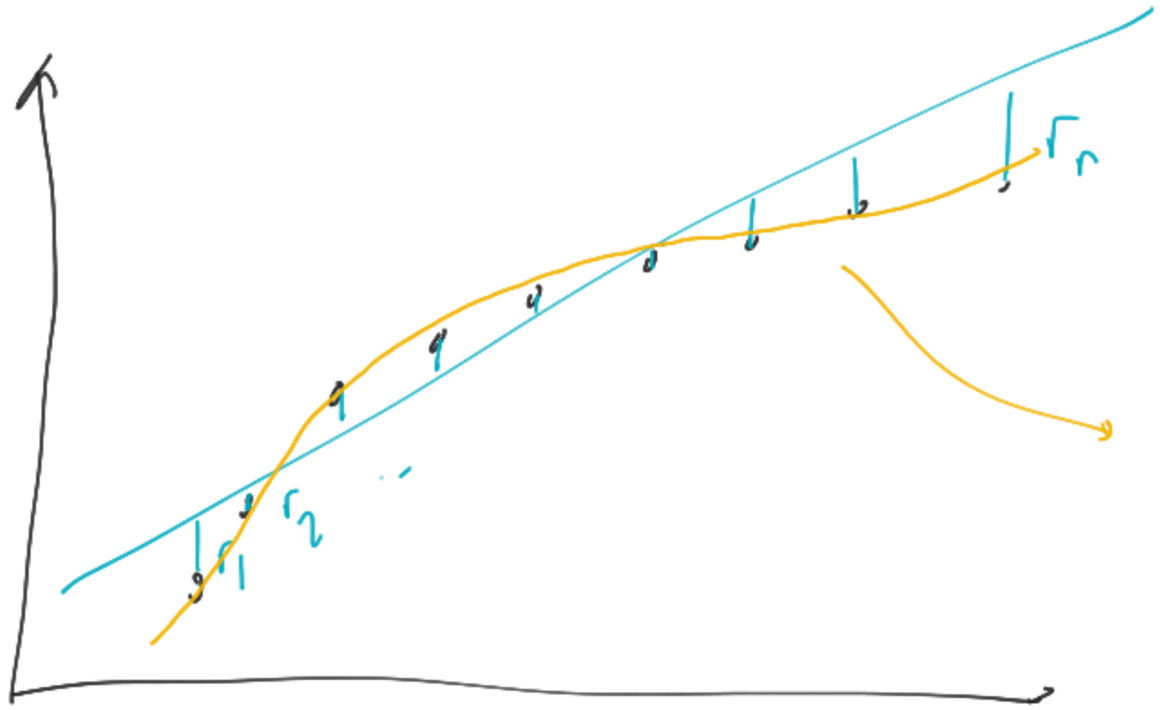
$$BIC = -ERROR + COMPLEJIDAD$$

	BIC
GMM 1 gaussiana	700
GMM 2 gaussianas	600
GMM 3 gaussianas	500 → menor BIC modelo
GMM 4 gaussianas	550

$y_{model} - y_{real}$

parsimonia

óptimo → 3 clusters



$ax+b \rightarrow (a,b)$

$\sum r_i^2 \rightarrow$ error

$ax^4 + bx^3 + cx^2 + dx^1 + e$

(a,b,c,d,e)