# Aprendizaje Supervisado
# Técnicas de Clasificación y Regresión

**Enrique J. Carmona Suárez**
Dpto. Inteligencia Artificial, ETS Ingeniería Informática, UNED
ecarmona@dia.uned.es

***Second IPARCOS Workshp on Machine Learning and Applications to Physics* - 18-20 diciembre 2023**
Facultad de Ciencias Físicas, Universidad Complutense de Madrid

# Contenido

- Algoritmo *k*-vecinos más cercanos
- Clasificador Naive-Bayes
- Árboles
- Redes neuronales artificiales
- Máquinas de vectores soporte

# Contenido

- Algoritmo k-vecinos más cercanos (k-NN)
  - Clasificación y regresión
- Clasificador Naive-Bayes
- Árboles
- Redes neuronales artificiales
- Máquinas de vectores soporte

# Algoritmo k-vecinos más cercanos (k-NN)

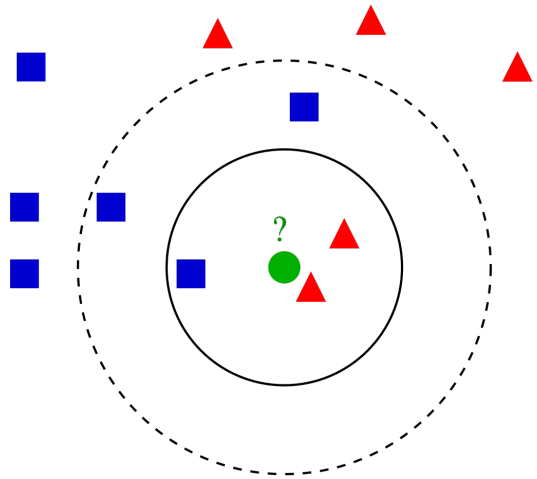
- Estima el grado de pertenencia de un elemento **x** a la clase *"C"* a partir de la información de similitud proporcionada por los k vecinos más cercanos del conjunto de datos:

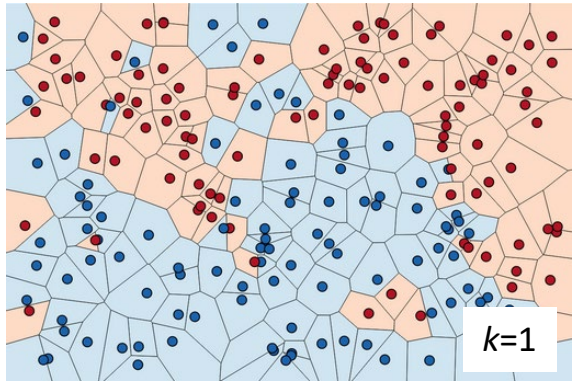
$$S(x'x') = d(x, x')$$

- No hace ninguna suposición acerca de la distribución de las variables predictoras.
- No aprende ningún modelo (*aprendizaje perezoso*)
- Cuando k↑ menos sensible al ruido, pero la clasificación de ejemplos situados entre clases puede resultar más imprecisa
- Muy sensible a atributos irrelevantes y a su rango de variación
- La similitud $S(x)$ puede ponderarse por la distancia a cada *k* vecinos:

$$w(x, x') = 1/1 + dist(x, x')$$

- El algoritmo k-NN es válido también para regresión



# Contenido

- Algoritmo k-vecinos más cercanos
- Clasificador Naive-Bayes (CNB)
  - Teorema de Bayes
  - Hipótesis de independencia estadística
  - Estimación de probabilidades
  - Ejemplo
- Árboles
- Redes neuronales artificiales
- Máquinas de vectores soporte

# Clasificador Naive-Bayes (CNB)

***Teorema de Bayes:***

$$p(C_k \mid X_i) = \frac{p(X_i \mid C_k) \cdot p(C_k)}{p(X_i)} \qquad \text{donde} \quad X_i = (x_{i1}, x_{i2}, \ldots x_{in})$$

***Clase más probable o "Máximum a Posteri" (MAP):***

$$c_{MAP} = \arg \max_{k=1,\ldots,r} p(C_k \mid X_i) = \arg \max_{k=1,\ldots,r} \frac{p(X_i \mid C_k) \cdot p(C_k)}{p(X_i)}$$

$$c_{MAP} = \arg \max_{k=1,\ldots,r} p(X_i \mid C_k) \cdot p(C_k)$$

# Hipótesis de independencia estadística

***Si asumimos la hipótesis de independencia estadística:***

$$p(X_i \mid C_1) = p(x_{i1} \mid C_1) \cdot \ldots \cdot p(x_{in} \mid C_1)$$

$$p(X_i \mid C_2) = p(x_{i1} \mid C_2) \cdot \ldots \cdot p(x_{in} \mid C_2)$$

$$\ldots$$

$$p(X_i \mid C_r) = p(x_{i1} \mid C_r) \cdot \ldots \cdot p(x_{in} \mid C_r)$$

$$c_{MAP} = \arg \max_{k=1,\ldots,r} p(X_i \mid C_k) \cdot p(C_k)$$

$$c_{MAP} = \arg \max_{k=1,\ldots,r} p(C_k) \cdot \prod_{j=1}^{n} p(x_{ij} \mid C_k)$$

# Estimación de parámetros del CNB

***Probabilidades de clase:***

$$p\left(C_k\right) = \frac{f_e\left(C = C_k\right)}{N}$$

***Probabilidades condicionadas de un atributo $A_j$ discreto:***

$$p\left(x_{ij} \mid C_k\right) = \frac{f_e\left(A_j = x_{ij} \wedge C = C_k\right)}{f_e\left(C = C_k\right)}$$

***Probabilidades condicionadas de un atributo $A_m$ continuo:***

$$p\left(x_{im} \mid C_k\right) \propto \frac{1}{\sqrt{2\pi\sigma_{mk}^2}} \exp\left(-\frac{1}{2}\frac{\left(x_{im} - \mu_{mk}\right)^2}{\sigma_{mk}^2}\right)$$

| | $p(C_k)$ |
|---|---|
| $C_1$ | $\alpha_1$ |
| ... | ... |
| $C_r$ | $\alpha_r$ |

| $p(x_{ij}\|C_k)$ | $x_{1j}$ | ... | $X_{nj}$ |
|---|---|---|---|
| $C_1$ | $\beta_{11}$ | ... | $\beta_{1n}$ |
| ... | ... | ... | ... |
| $C_r$ | $\beta_{r1}$ | ... | $\beta_{rn}$ |

| | $p(x_{im}\|C_k)$ |
|---|---|
| $C_1$ | $N_m(\mu_{1m},\sigma_{1m})$ |
| ... | ... |
| $C_r$ | $N_m(\mu_{r\,m},\sigma_{r\,m})$ |

UNED ETS de Ingeniería Informática

# Ejemplo: ¿Paciente normal o patológico?

| Paciente | Presión Arterial ($mm$ de Hg) | Leucocitos ($mill/mm^3$) | Dolor muscular | Desvanec. | Patología |
|---|---|---|---|---|---|
| 1 | 140 | 13.0 | sí | no | sí |
| 2 | 138 | 14.1 | sí | sí | sí |
| 3 | 110 | 7.5 | sí | no | no |
| 4 | 105 | 4.0 | no | no | no |
| 5 | 100 | 2.1 | no | sí | no |
| 6 | 125 | 3.1 | no | sí | no |
| 7 | 145 | 8.1 | sí | no | sí |
| 8 | 135 | 2.1 | no | no | no |
| 9 | 100 | 8.2 | no | no | no |

UNED ETS de Ingeniería Informática

# Ejemplo: ¿Paciente normal o patológico?

| Paciente | Presión Arterial ($mm$ de Hg) | Leucocitos ($mill/mm^3$) | Dolor muscular | Desvanec. | Patología |
|---|---|---|---|---|---|
| 1 | 140 | 13.0 | sí | no | sí |
| 2 | 138 | 14.1 | sí | sí | sí |
| 3 | 110 | 7.5 | sí | no | no |
| 4 | 105 | 4.0 | no | no | no |
| 5 | 100 | 2.1 | no | sí | no |
| 6 | 125 | 3.1 | no | sí | no |
| 7 | 145 | 8.1 | sí | no | sí |
| 8 | 135 | 2.1 | no | no | no |
| 9 | 100 | 8.2 | no | no | no |

***Probabilidades de clase a priori:***

| | $p(C_k),\ k = 1, 2$ |
|---|---|
| $C_1(sí)$ | $\frac{3}{9}$ |
| $C_2(no)$ | $\frac{6}{9}$ |

UNED ETS de Ingeniería Informática

# Ejemplo: ¿Paciente normal o patológico?

| Paciente | Presión Arterial ($mm$ de Hg) | Leucocitos ($mill/mm^3$) | Dolor muscular | Desvanec. | Patología |
|---|---|---|---|---|---|
| 1 | 140 | 13.0 | sí | no | sí |
| 2 | 138 | 14.1 | sí | sí | sí |
| 3 | 110 | 7.5 | sí | no | no |
| 4 | 105 | 4.0 | no | no | no |
| 5 | 100 | 2.1 | no | sí | no |
| 6 | 125 | 3.1 | no | sí | no |
| 7 | 145 | 8.1 | sí | no | sí |
| 8 | 135 | 2.1 | no | no | no |
| 9 | 100 | 8.2 | no | no | no |

***Probabilidades condicionadas de atributos continuos:***

| Presión Arterial (PA) | $(\mu,\ \sigma)$ | $p(PA\|C_k),\ k=1,2$ |
|---|---|---|
| $C_1(sí)$ | $(\mu_1 = 141{,}0,\ \sigma_1 = 3{,}61)$ | $0{,}1105 \cdot \exp\left[-0{,}0384\,(x-141)^2\right]$ |
| $C_2(no)$ | $(\mu_2 = 112{,}5,\ \sigma_2 = 14{,}40)$ | $0{,}0277 \cdot \exp\left[-0{,}0024\,(x-112{,}5)^2\right]$ |

| Leucocitos (L) | $(\mu,\ \sigma)$ | $p(L\|C_k),\ k=1,2$ |
|---|---|---|
| $C_1(sí)$ | $(\mu_1 = 11{,}7,\ \sigma_1 = 3{,}19)$ | $0{,}1251 \cdot \exp\left[-0{,}0491\,(x-11{,}7)^2\right]$ |
| $C_2(no)$ | $(\mu_2 = 4{,}5,\ \sigma_2 = 2{,}70)$ | $0{,}1478 \cdot \exp\left[-0{,}0686\,(x-4{,}5)^2\right]$ |

UNED ETS de Ingeniería Informática

# Ejemplo: ¿Paciente normal o patológico?

| Paciente | Presión Arterial ($mm$ de Hg) | Leucocitos ($mill/mm^3$) | Dolor muscular | Desvanec. | Patología |
|---|---|---|---|---|---|
| 1 | 140 | 13.0 | sí | no | sí |
| 2 | 138 | 14.1 | sí | sí | sí |
| 3 | 110 | 7.5 | sí | no | no |
| 4 | 105 | 4.0 | no | no | no |
| 5 | 100 | 2.1 | no | sí | no |
| 6 | 125 | 3.1 | no | sí | no |
| 7 | 145 | 8.1 | sí | no | sí |
| 8 | 135 | 2.1 | no | no | no |
| 9 | 100 | 8.2 | no | no | no |

***Probabilidades condicionadas de atributos continuos:***

| Presión Arterial (PA) | $(\mu,\ \sigma)$ | $p(PA\|C_k),\ k=1,2$ |
|---|---|---|
| $C_1(sí)$ | $(\mu_1 = 141{,}0,\ \sigma_1 = 3{,}61)$ | $0{,}1105 \cdot \exp\left[-0{,}0384\,(x-141)^2\right]$ |
| $C_2(no)$ | $(\mu_2 = 112{,}5,\ \sigma_2 = 14{,}40)$ | $0{,}0277 \cdot \exp\left[-0{,}0024\,(x-112{,}5)^2\right]$ |

| Leucocitos (L) | $(\mu,\ \sigma)$ | $p(L\|C_k),\ k=1,2$ |
|---|---|---|
| $C_1(sí)$ | $(\mu_1 = 11{,}7,\ \sigma_1 = 3{,}19)$ | $0{,}1251 \cdot \exp\left[-0{,}0491\,(x-11{,}7)^2\right]$ |
| $C_2(no)$ | $(\mu_2 = 4{,}5,\ \sigma_2 = 2{,}70)$ | $0{,}1478 \cdot \exp\left[-0{,}0686\,(x-4{,}5)^2\right]$ |

UNED ETS de Ingeniería Informática

# Ejemplo: ¿Paciente normal o patológico?

| Paciente | Presión Arterial ($mm$ de Hg) | Leucocitos ($mill/mm^3$) | Dolor muscular | Desvanec. | Patología |
|---|---|---|---|---|---|
| 1 | 140 | 13.0 | sí | no | sí |
| 2 | 138 | 14.1 | sí | sí | sí |
| 3 | 110 | 7.5 | sí | no | no |
| 4 | 105 | 4.0 | no | no | no |
| 5 | 100 | 2.1 | no | sí | no |
| 6 | 125 | 3.1 | no | sí | no |
| 7 | 145 | 8.1 | sí | no | sí |
| 8 | 135 | 2.1 | no | no | no |
| 9 | 100 | 8.2 | no | no | no |

***Probabilidades condicionadas de atributos nominales:***

| Dolor Muscular (DM) | $p(DM = no \mid C_k),\ k = 1,2$ | $p(DM = sí \mid C_k),\ k = 1,2$ |
|---|---|---|
| $C_1(sí)$ | $\frac{0}{3} = 0 \rightarrow 0{,}05$ | $\frac{3}{3} = 1 \rightarrow 0{,}95$ |
| $C_2(no)$ | $\frac{5}{6}$ | $\frac{1}{6}$ |

| Desvanecimientos (D) | $p(D = no \mid C_k),\ k = 1,2$ | $p(D = sí \mid C_k),\ k = 1,2$ |
|---|---|---|
| $C_1(sí)$ | $\frac{2}{3}$ | $\frac{1}{3}$ |
| $C_2(no)$ | $\frac{4}{6}$ | $\frac{2}{6}$ |

UNED ETS de Ingeniería Informática

# Ejemplo: ¿Paciente normal o patológico?

| Paciente | Presión Arterial ($mm$ de Hg) | Leucocitos ($mill/mm^3$) | Dolor muscular | Desvanec. | Patología |
|---|---|---|---|---|---|
| 1 | 140 | 13.0 | sí | no | sí |
| 2 | 138 | 14.1 | sí | sí | sí |
| 3 | 110 | 7.5 | sí | no | no |
| 4 | 105 | 4.0 | no | no | no |
| 5 | 100 | 2.1 | no | sí | no |
| 6 | 125 | 3.1 | no | sí | no |
| 7 | 145 | 8.1 | sí | no | sí |
| 8 | 135 | 2.1 | no | no | no |
| 9 | 100 | 8.2 | no | no | no |

***Probabilidades condicionadas de atributos nominales:***

| Dolor Muscular (DM) | $p(DM = no\|C_k),\ k = 1,2$ | $p(DM = sí\|C_k),\ k = 1,2$ |
|---|---|---|
| $C_1(sí)$ | $\frac{0}{3} = 0 \rightarrow 0{,}05$ | $\frac{3}{3} = 1 \rightarrow 0{,}95$ |
| $C_2(no)$ | $\frac{5}{6}$ | $\frac{1}{6}$ |

| Desvanecimientos (D) | $p(D = no\|C_k),\ k = 1,2$ | $p(D = sí\|C_k),\ k = 1,2$ |
|---|---|---|
| $C_1(sí)$ | $\frac{2}{3}$ | $\frac{1}{3}$ |
| $C_2(no)$ | $\frac{4}{6}$ | $\frac{2}{6}$ |

# Ejemplo: ¿Paciente normal o patológico?

| | $p(C_k),\ k=1,2$ |
|---|---|
| $C_1(sí)$ | $\frac{3}{9}$ |
| $C_2(no)$ | $\frac{6}{9}$ |

***¿El paciente X=(PA:150, L:15, DM:sí, D=sí) es patológico o normal?***

| Presión Arterial (PA) | $(\mu,\ \sigma)$ | $p(PA\|C_k),\ k=1,2$ |
|---|---|---|
| $C_1(sí)$ | $(\mu_1 = 141{,}0,\ \sigma_1 = 3{,}61)$ | $0{,}1105 \cdot \exp\left[-0{,}0384\,(x-141)^2\right]$ |
| $C_2(no)$ | $(\mu_2 = 112{,}5,\ \sigma_2 = 14{,}40)$ | $0{,}0277 \cdot \exp\left[-0{,}0024\,(x-112{,}5)^2\right]$ |

| Leucocitos (L) | $(\mu,\ \sigma)$ | $p(L\|C_k),\ k=1,2$ |
|---|---|---|
| $C_1(sí)$ | $(\mu_1 = 11{,}7,\ \sigma_1 = 3{,}19)$ | $0{,}1251 \cdot \exp\left[-0{,}0491\,(x-11{,}7)^2\right]$ |
| $C_2(no)$ | $(\mu_2 = 4{,}5,\ \sigma_2 = 2{,}70)$ | $0{,}1478 \cdot \exp\left[-0{,}0686\,(x-4{,}5)^2\right]$ |

| Dolor Muscular (DM) | $p(DM = no\|C_k),\ k=1,2$ | $p(DM = sí\|C_k),\ k=1,2$ |
|---|---|---|
| $C_1(sí)$ | $\frac{0}{3} = 0 \rightarrow 0{,}05$ | $\frac{3}{3} = 1 \rightarrow 0{,}95$ |
| $C_2(no)$ | $\frac{5}{6}$ | $\frac{1}{6}$ |

| Desvanecimientos (D) | $p(D = no\|C_k),\ k=1,2$ | $p(D = sí\|C_k),\ k=1,2$ |
|---|---|---|
| $C_1(sí)$ | $\frac{2}{3}$ | $\frac{1}{3}$ |
| $C_2(no)$ | $\frac{4}{6}$ | $\frac{2}{6}$ |

$$p(C_{sí} \mid X) \propto p(X \mid C_{sí}) \cdot p(C_{sí}) = p(PA_{150} \mid C_{sí}) \cdot p(L_{15} \mid C_{sí}) \cdot p(DM_{sí} \mid C_{sí}) \cdot p(D_{sí} \mid C_{sí}) \cdot p(C_{sí}) =$$

$$0{,}0049 \times 0{,}0733 \times 0{,}95 \times \frac{1}{3} \times \frac{3}{9} = 3{,}79 \times 10^{-5}$$

$$p(C_{no} \mid X) \propto p(X \mid C_{no}) \cdot p(C_{no}) = p(PA_{150} \mid C_{no}) \cdot p(L_{15} \mid C_{no}) \cdot p(DM_{sí} \mid C_{no}) \cdot p(D_{sí} \mid C_{no}) \cdot p(C_{no}) =$$

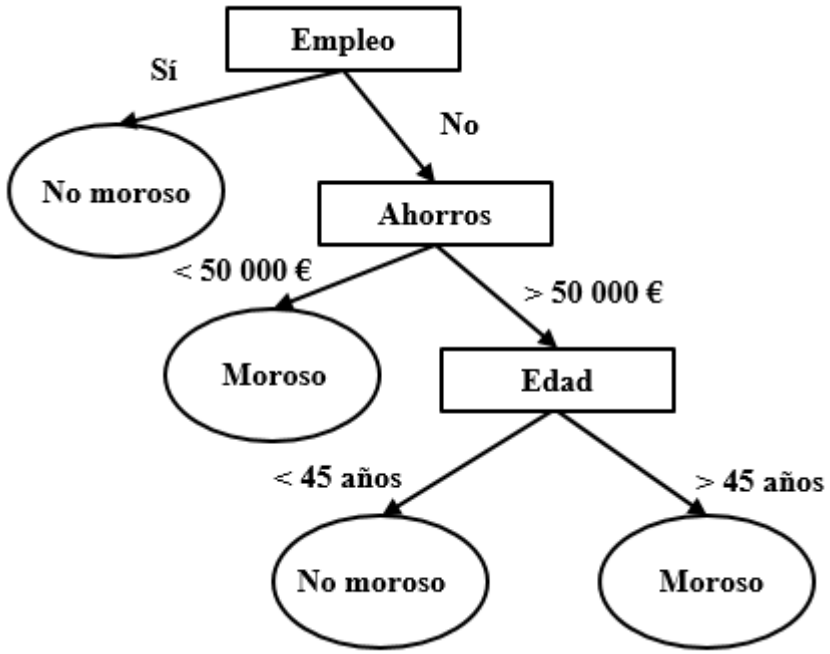
$$0{,}0009 \times 0{,}00008 \times \frac{1}{6} \times \frac{2}{6} \times \frac{6}{9} = 2{,}66 \times 10^{-9}$$

# Contenido

- Algoritmo k-vecinos más cercanos
- Clasificador Naive-Bayes
- Árboles
  - De decisión (clasificación)
  - De regresión y de modelos (regresión)
- Redes neuronales artificiales
- Máquinas de vectores soporte

# Ejemplo de árbol de decisión

***Ejemplo de Árbol de Decisión usado por un banco para predecir si un cliente es moroso:***




Cómo aprender un AD
- AD trivial
- AD mínimo
- AD inductivo "*top-down*"

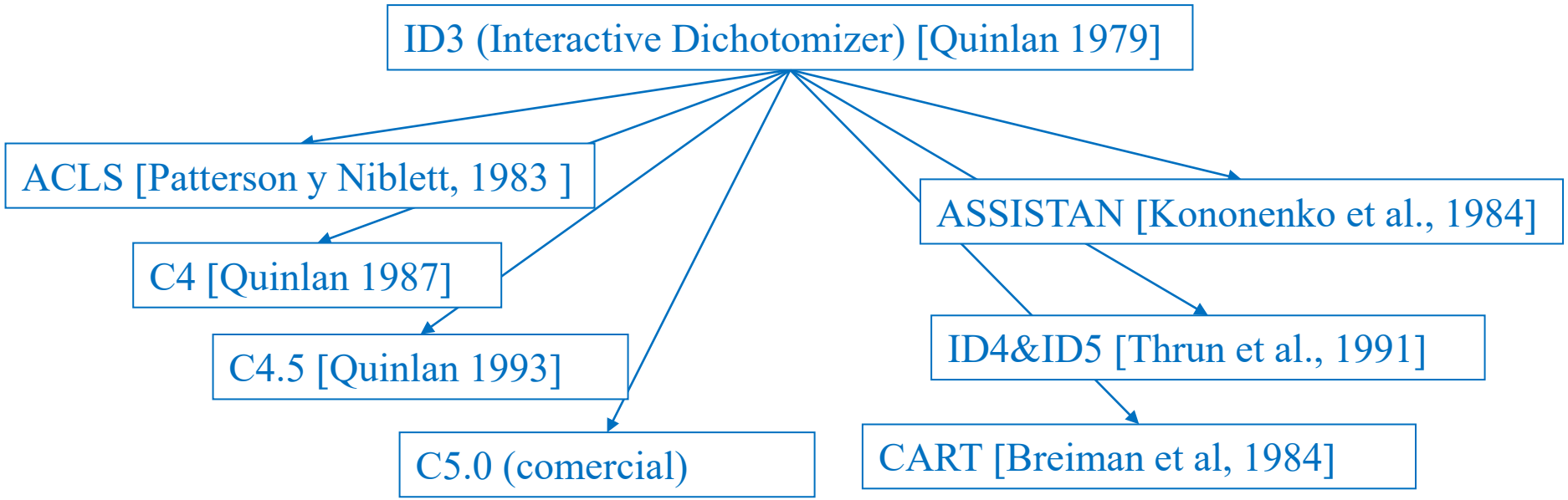
# Origen de los árboles de decisión

**Tienen su origen en  la creación de modelos de simulación cognoscitiva del aprendizaje de conceptos (Años 60)**

CLS (*Concept Learning System*) [Hunt et al. 1966]

(Origen de la familia de algoritmos TDIDT)

⇓

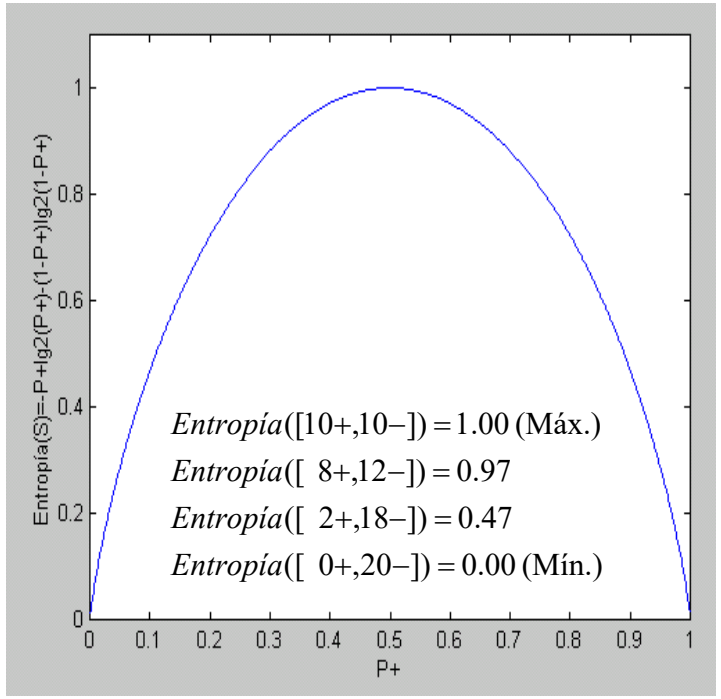


# Cómo elegir el mejor nodo en cada división

**La Entropía de la información (Shanon 1948) caracteriza el "grado de impureza" de un conjunto de ejemplos *S***

***Sólo hay dos clases:***

$$Entropía(S) \equiv -(p_+ \log_2 p_+ + p_- \log_2 p_-)$$



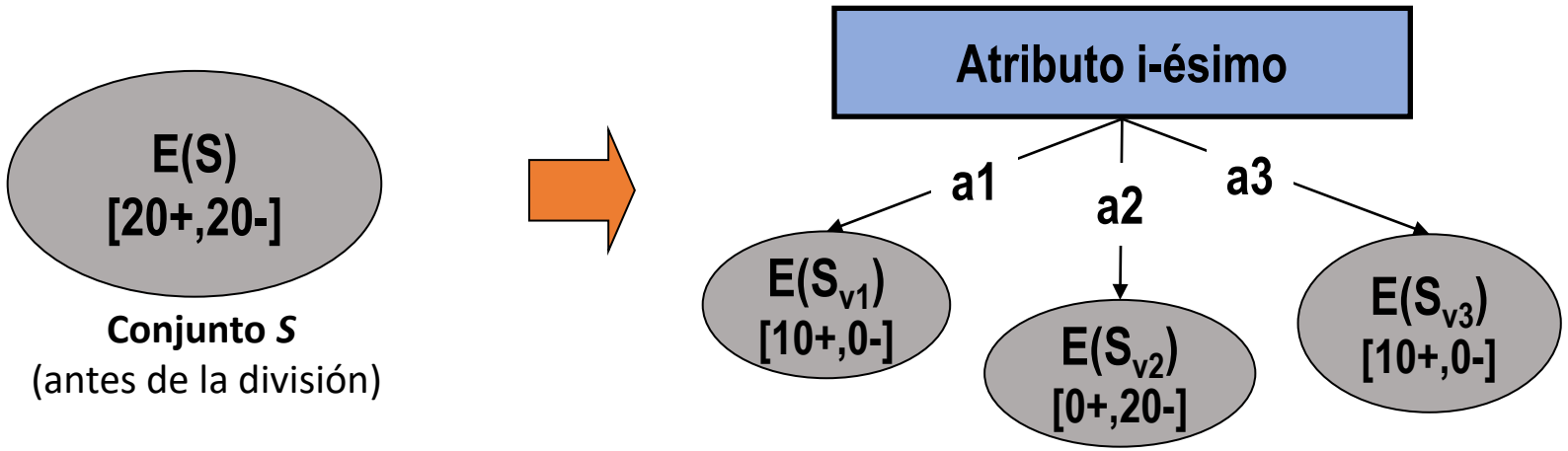

***Hay "c" clases, donde c≥3:***

$$Entropía(S) \equiv -\sum_{i=1}^{c} p_i \log_2 p_i$$

# Cómo elegir el mejor atributo en cada división

***Solución*: Seleccionar el atributo que, tras la partición, consigue una mayor disminución de entropía (mayor ganancia de información)**

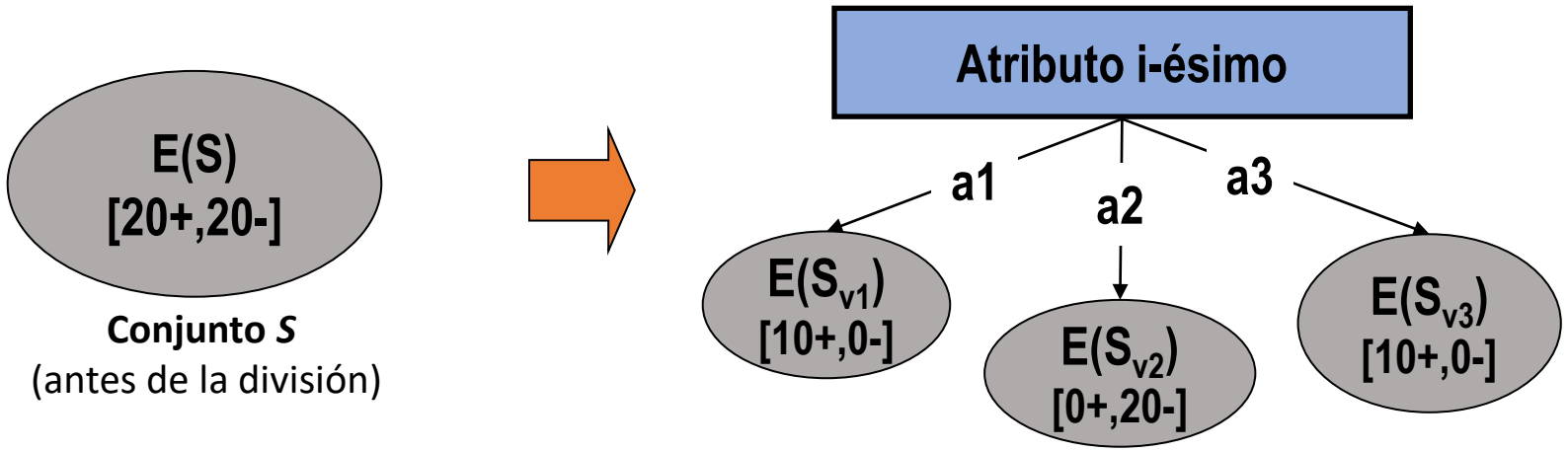
$$\boldsymbol{Ganancía(S,A) \equiv Entropía(S) - \sum_{v \in Valores(A)} Entropía(S_v)}$$



# Cómo elegir el mejor atributo en cada división

***Solución*: Seleccionar el atributo que, tras la partición, consigue una mayor disminución de entropía (mayor ganancia de información)**

$$\boldsymbol{Ganancía(S,A) \equiv Entropía(S) - \sum_{v \in Valores(A)} Entropía(S_v)}$$



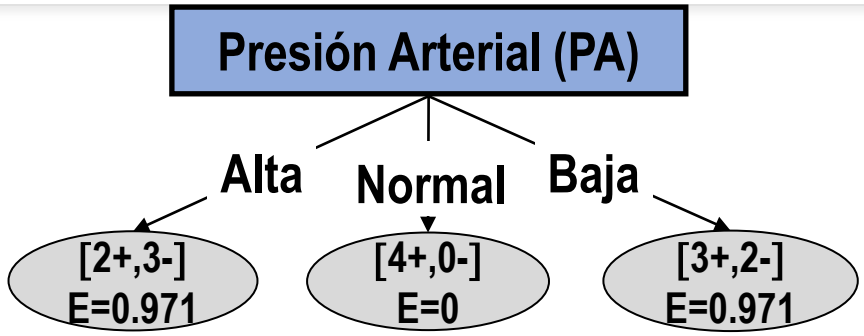

$$\boldsymbol{Ganancía(S,A) \equiv Entropía(S) - \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} Entropía(S_v)}$$

UNED ETS de Ingeniería Informática

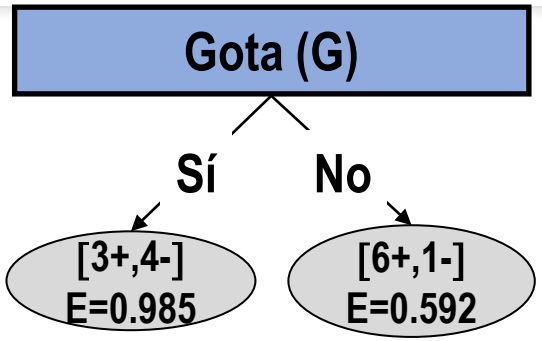
# Ejemplo: ¿Cuándo aplicar un tratamiento?

| Paciente | Presión Aterial | Urea en sangre | Gota | Hipotiroidismo | Administrar Tratamiento |
|---|---|---|---|---|---|
| 1 | Alta | Alta | Sí | No | **No** |
| 2 | Alta | Alta | Sí | Sí | **No** |
| 3 | Normal | Alta | Sí | No | **Sí** |
| 4 | Baja | Normal | Sí | No | **Sí** |
| 5 | Baja | Baja | No | No | **Sí** |
| 6 | Baja | Baja | No | Sí | **No** |
| 7 | Normal | Baja | No | Sí | **Sí** |
| 8 | Alta | Normal | Sí | No | **No** |
| 9 | Alta | Baja | No | No | **Sí** |
| 10 | Baja | Normal | No | No | **Sí** |
| 11 | Alta | Normal | No | Sí | **Sí** |
| 12 | Normal | Normal | Sí | Sí | **Sí** |
| 13 | Normal | Alta | No | No | **Sí** |
| 14 | Baja | Normal | Sí | Sí | **No** |

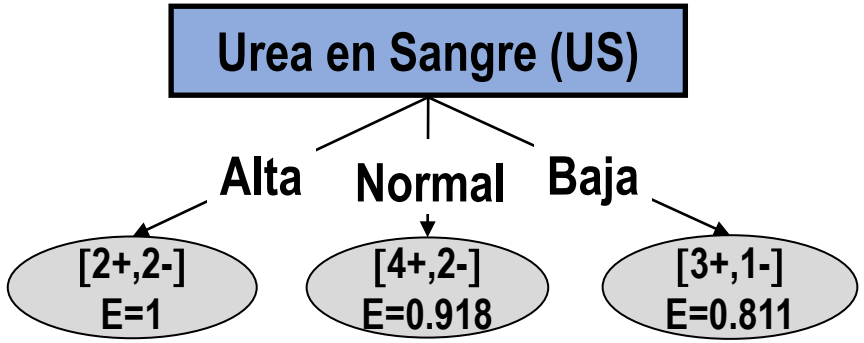
# Ejemplo: ¿Cuándo aplicar un tratamiento?



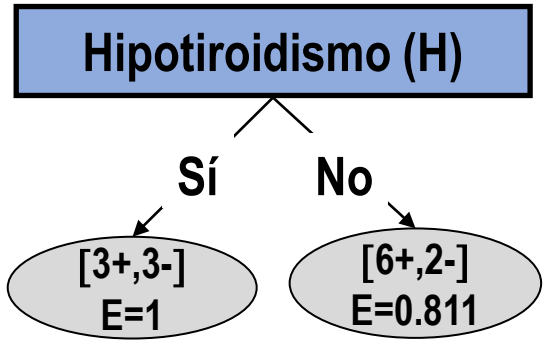

$$G(S,"PA") = 0.940 - (5/14)\cdot 0.971 - (4/14)\cdot 0 - (5/14)\cdot 0.971 = 0.246$$




$$G(S,"G") = 0.940 - (7/14)\cdot 0.985 - (7/14)\cdot 0.592 = 0.151$$

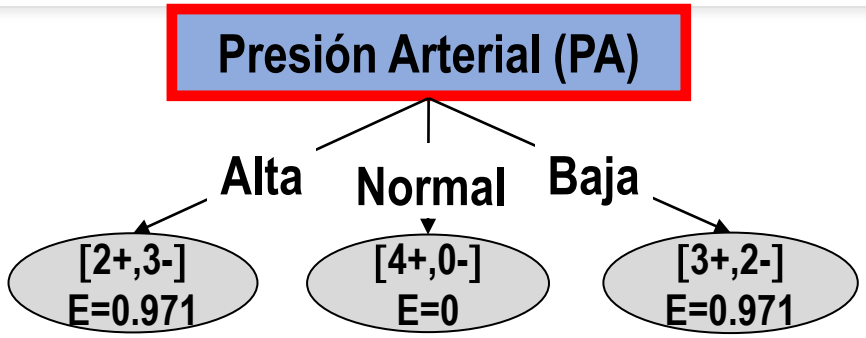



$$G(S,"US") = 0.940 - (4/14)\cdot 1 - (6/14)\cdot 0.918 - (4/14)\cdot 0.811 = 0.029$$

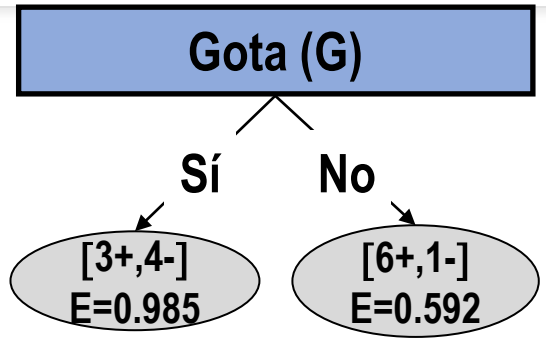



$$G(S,"H") = 0.940 - (6/14)\cdot 1 - (8/14)\cdot 0.811 = 0.048$$

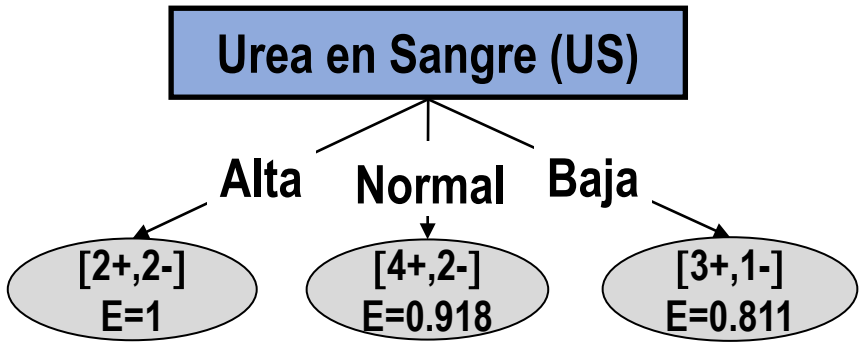
# Ejemplo: ¿Cuándo aplicar un tratamiento?



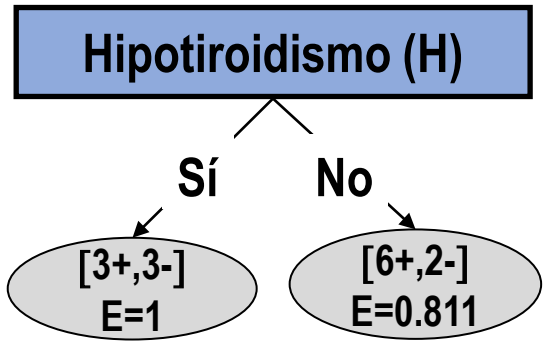

$$G(S,"PA") = 0.940 - (5/14)\cdot 0.971 - (4/14)\cdot 0 - (5/14)\cdot 0.971 = 0.246$$




$$G(S,"G") = 0.940 - (7/14)\cdot 0.985 - (7/14)\cdot 0.592 = 0.151$$

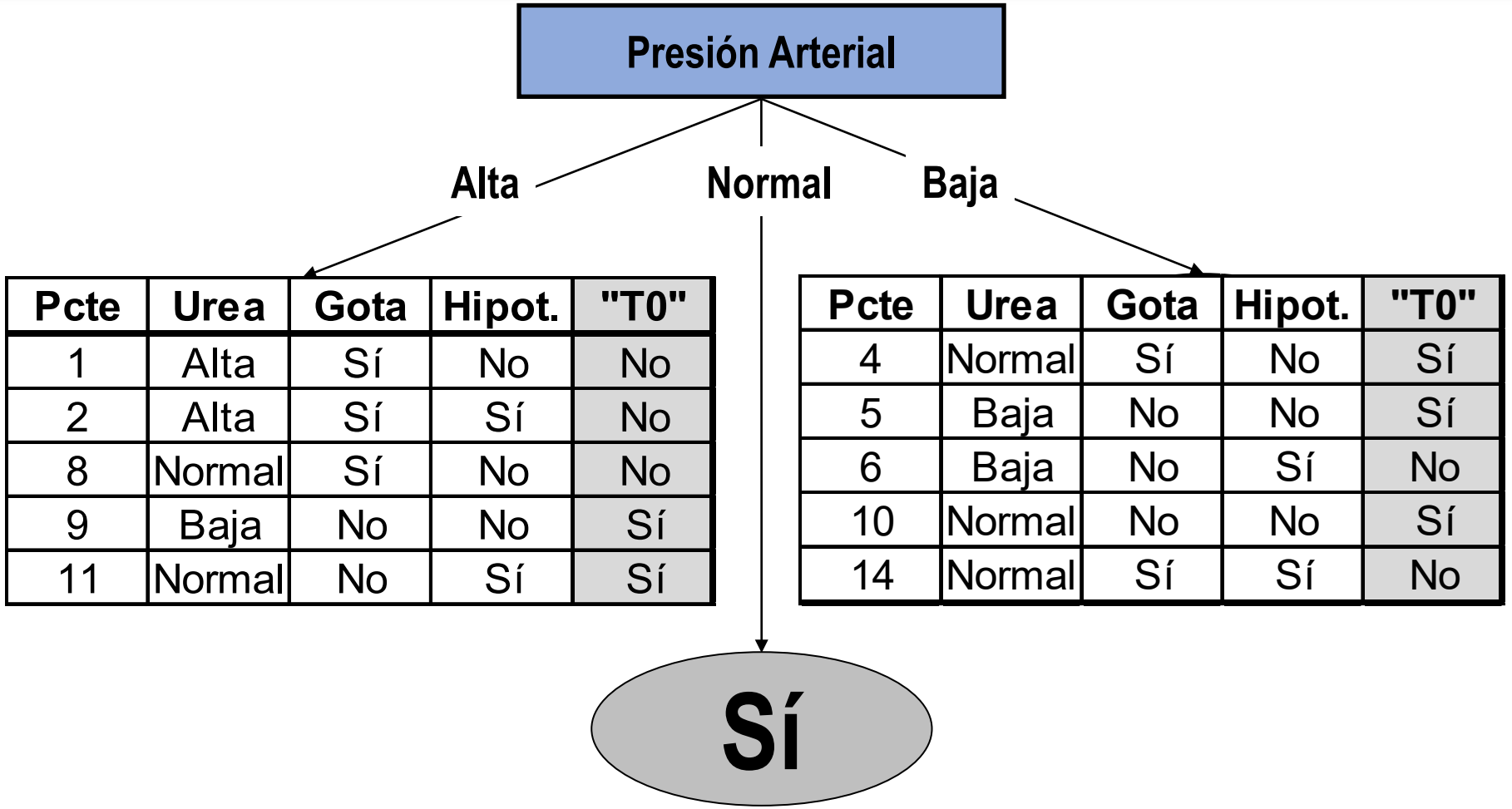



$$G(S,"US") = 0.940 - (4/14)\cdot 1 - (6/14)\cdot 0.918 - (4/14)\cdot 0.811 = 0.029$$




$$G(S,"H") = 0.940 - (6/14)\cdot 1 - (8/14)\cdot 0.811 = 0.048$$

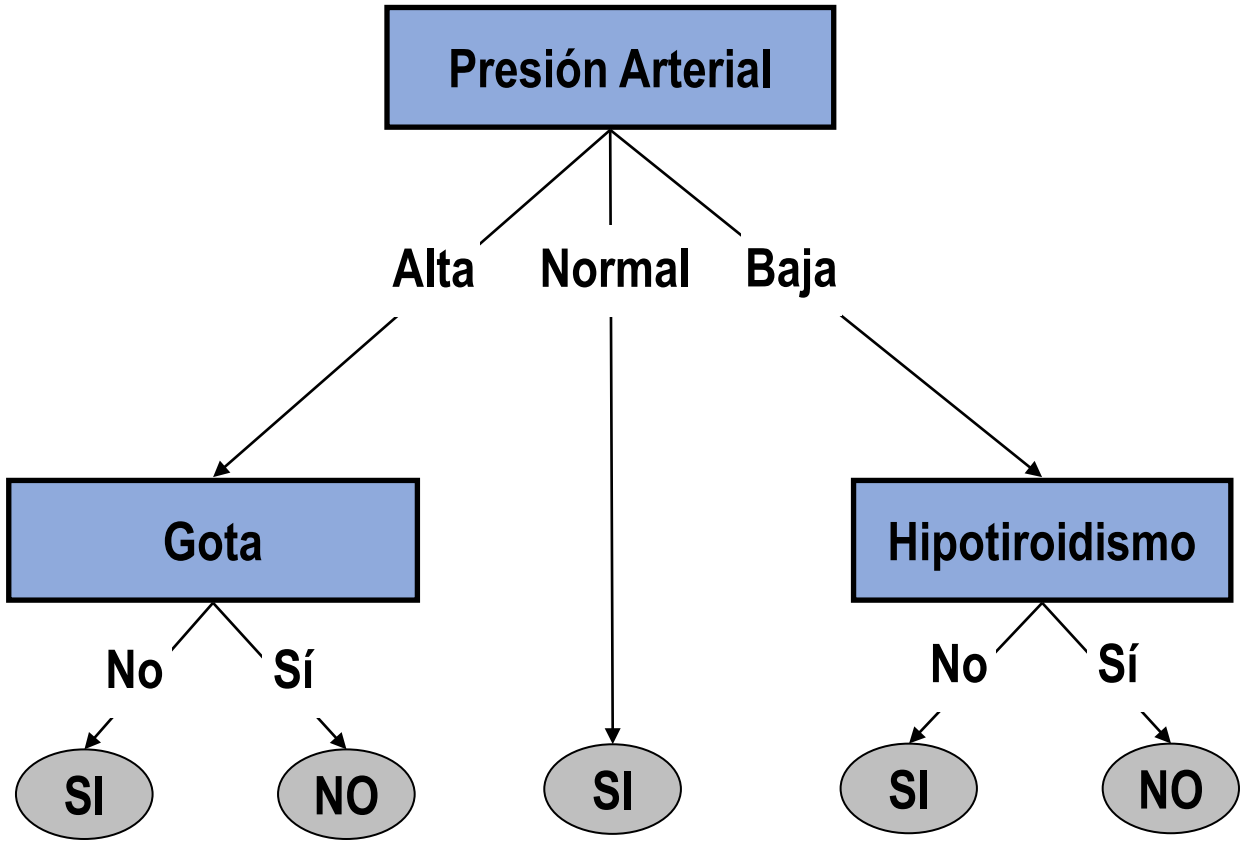
# Ejemplo de AD: ¿Cuándo aplicar un tratamiento?



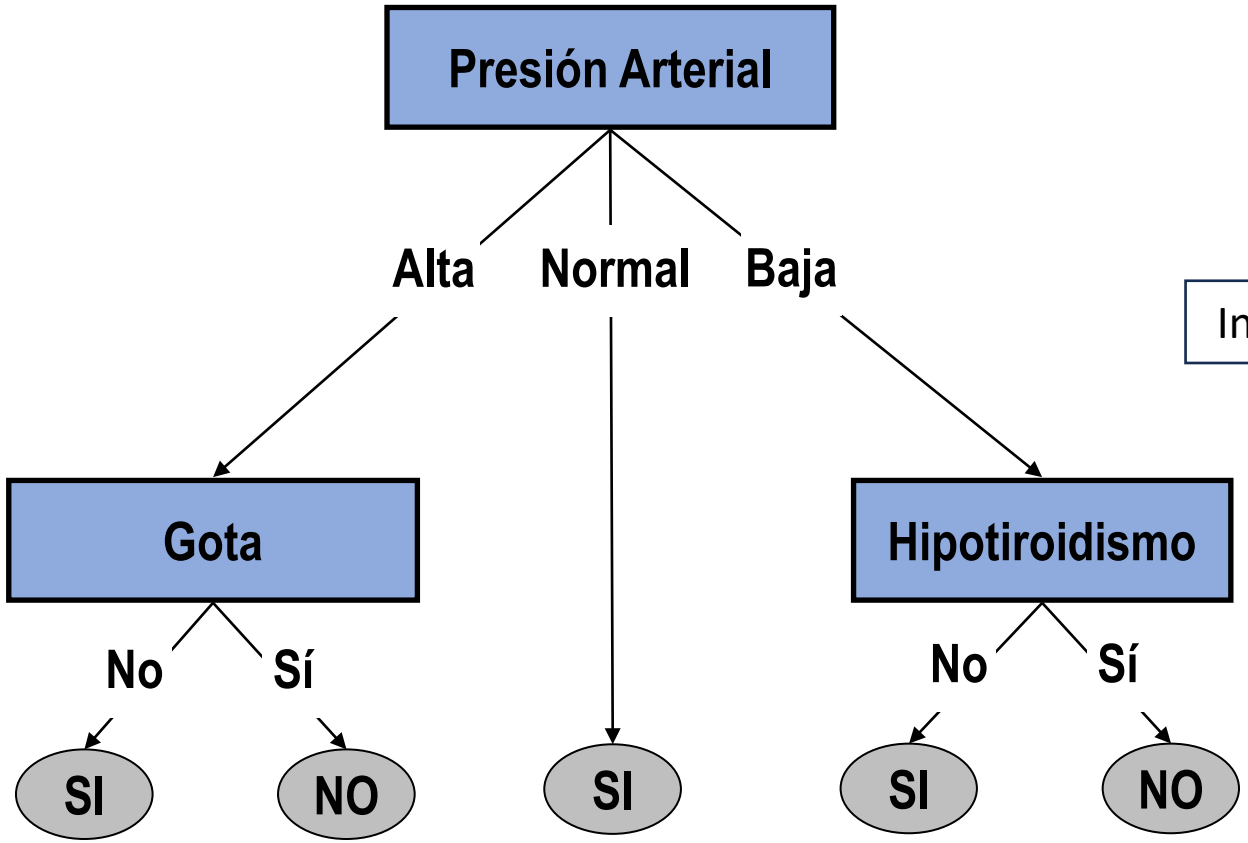

| Pcte | Urea | Gota | Hipot. | "T0" |
|---|---|---|---|---|
| 1 | Alta | Sí | No | No |
| 2 | Alta | Sí | Sí | No |
| 8 | Normal | Sí | No | No |
| 9 | Baja | No | No | Sí |
| 11 | Normal | No | Sí | Sí |

| Pcte | Urea | Gota | Hipot. | "T0" |
|---|---|---|---|---|
| 4 | Normal | Sí | No | Sí |
| 5 | Baja | No | No | Sí |
| 6 | Baja | No | Sí | No |
| 10 | Normal | No | No | Sí |
| 14 | Normal | Sí | Sí | No |

# Ejemplo: ¿Cuándo aplicar un tratamiento?



# Ejemplo: ¿Cuándo aplicar un tratamiento?




Interpretabilidad

R1: (PA=*alta*) $\wedge$ (G=*no*) $\Rightarrow$ (T=*sí*)

R2: (PA=*alta*) $\wedge$ (G=*sí*) $\Rightarrow$ (T=*no*)

R3: (PA=*normal*) $\Rightarrow$ (T=*sí*)

R4: (PA=*baja*) $\wedge$ (H=*no*) $\Rightarrow$ (T=*sí*)

R5: (PA=*baja*) $\wedge$ (H=*sí*) $\Rightarrow$ (T=*no*)

# Y si los atributos no son continuos

1. Se ordenan los valores del atributo junto a la clase a la que pertenecen:

| Edad | 30 | 36 | 44 | 60 | 72 | 78 |
|---|---|---|---|---|---|---|
| Clase | sí | sí | no | no | no | sí |

2. Se mide la ganancia para cada punto de corte posible y se asigna la mayor a este atributo:

G(S, ”edad<33”)
G(S, ”edad<40”) ←
G(S, ”edad<47”)
G(S, ”edad<66”)
G(S, ”edad<75”)

Otra posibilidad → Asignar un nuevo atributo binario a cada rango de valores donde no varía la clase:

**“edad < 40” = {sí, no}**
**“40 ≤ edad < 75” = {sí, no}**
**“75 ≤ edad” = {sí, no}**

# Manejando atributos con costes diferentes

- Se puede sustituir el cálculo de la ganancia por:

$$\frac{\text{Ganancia de información}}{\text{unidad de coste}}$$

- No se garantiza un AD de coste óptimo, pero sesga la búsqueda en favor de los atributos de coste bajo.
- Ejemplos:
  - Medicina (Nuñez 91):

$$\frac{2^{Ganancia(S,A)} - 1}{[Coste(A) + 1]^{w}} \qquad w \in [0,1] \quad \text{importancia relativa } Coste \text{ v.s. } Ganancia$$

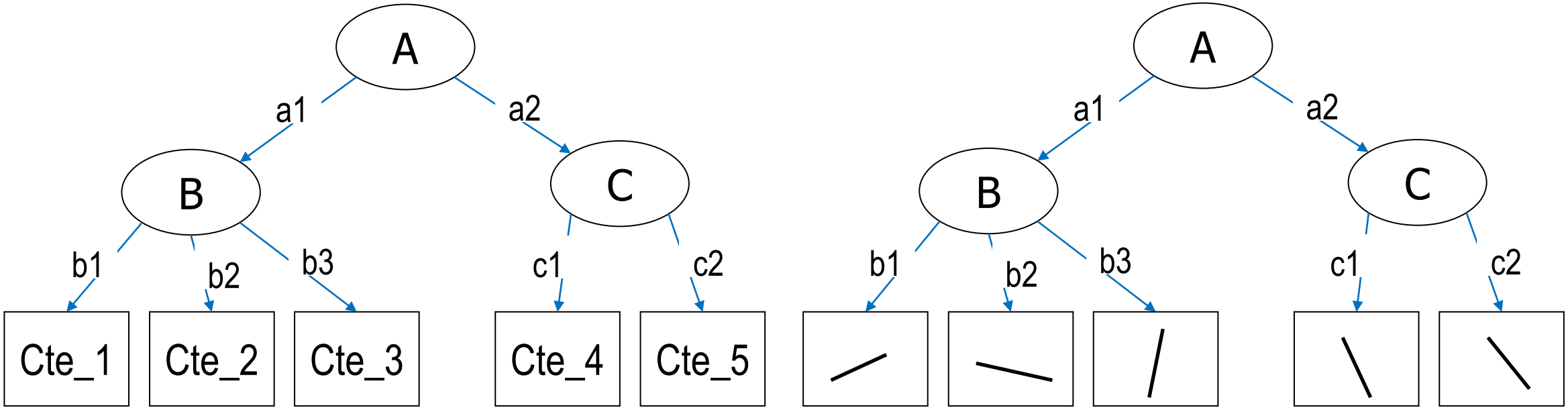
  - Robótica (Tan and Schlimmer 90):

$$\frac{\text{Ganancia}^2(S, A)}{\text{Coste}(A)}$$

# Cómo evitar el *overfitting*

- Pre-poda: Se decide cuando dejar de subdividir
  - Establecer un Nº mínimo de ejemplos en un nodo para seguir dividiendo
  - Dividir sólo si el error de validación tras la división es menor que antes de la división
  - […]
- Post-poda: se construye el AD y después se poda
  - Para cada partición de nodos hoja asociada a un nodo se mide el efecto sobre el error de validación al eliminar dicha partición. Si el error de validación disminuye, se elimina la partición con mayor impacto en dicho error.
  - […]

# Árboles de regresión *vs.* Árboles de modelo

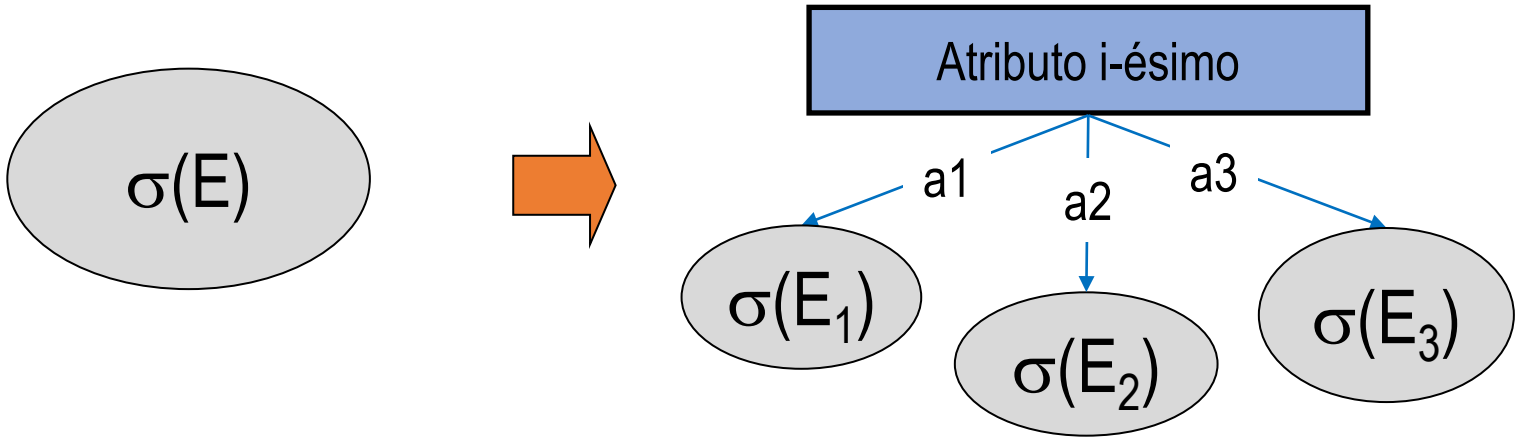



***Árbol de regresión*** ***Árbol de modelo***

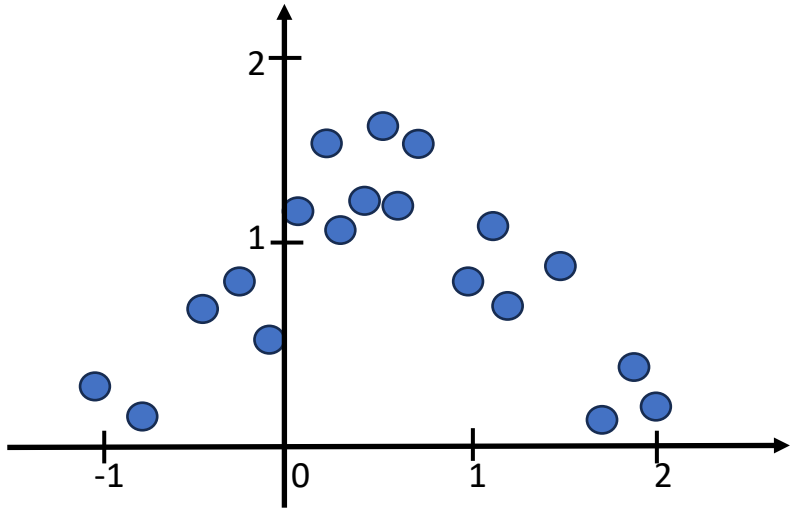
# Cómo elegir el mejor atributo en cada división

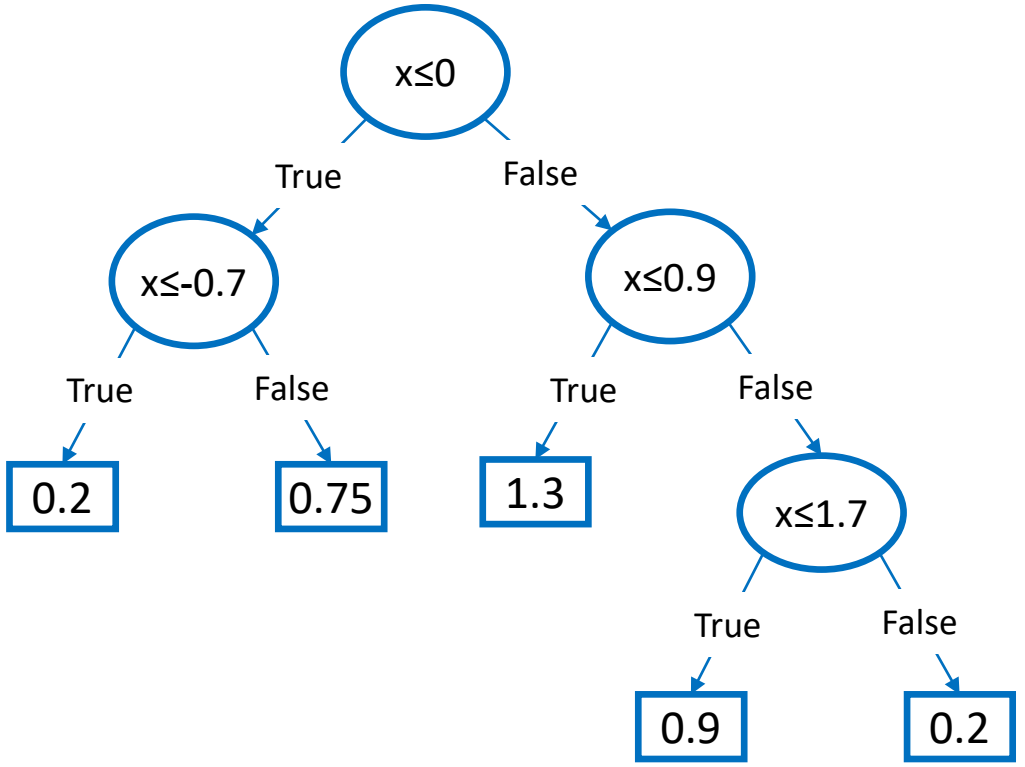
***Solución*: Se elige aquel atributo que maximice la reducción del error esperado tras la partición:**

$$\Delta error = \sigma(E) - \sum_i \frac{|E_i|}{|E|}\sigma(E_i)$$

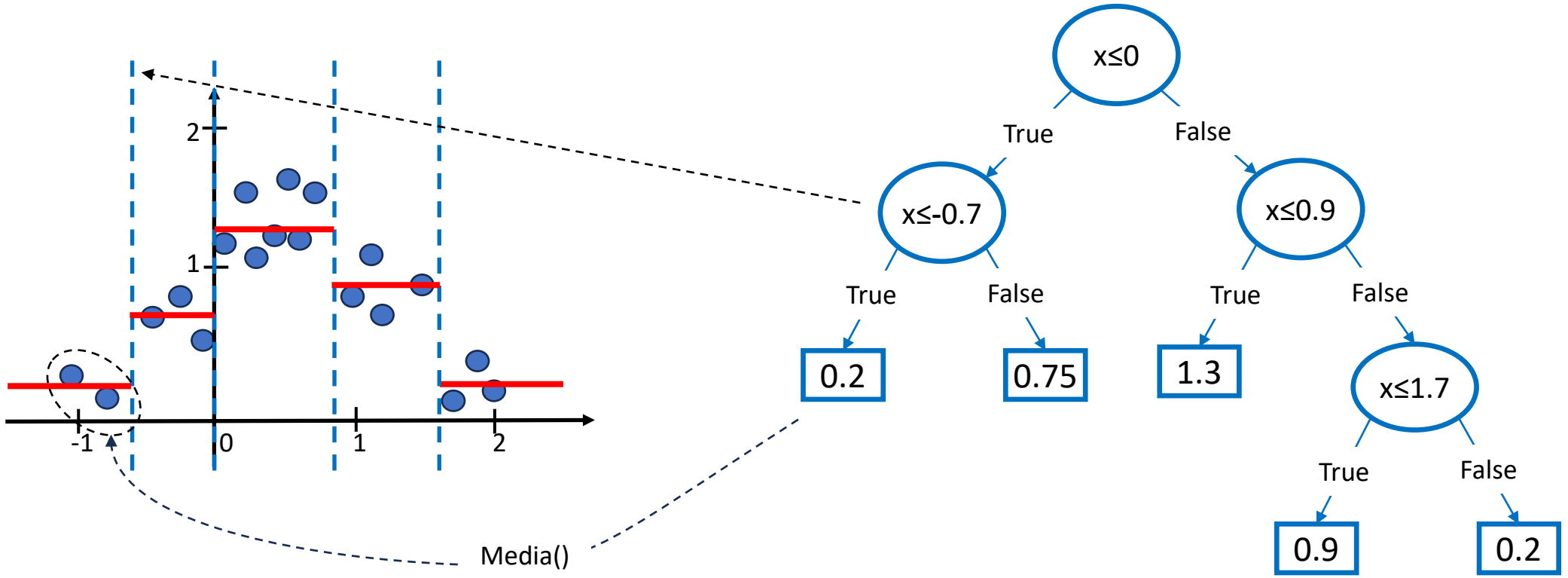


# Ejemplo de árbol de regresión

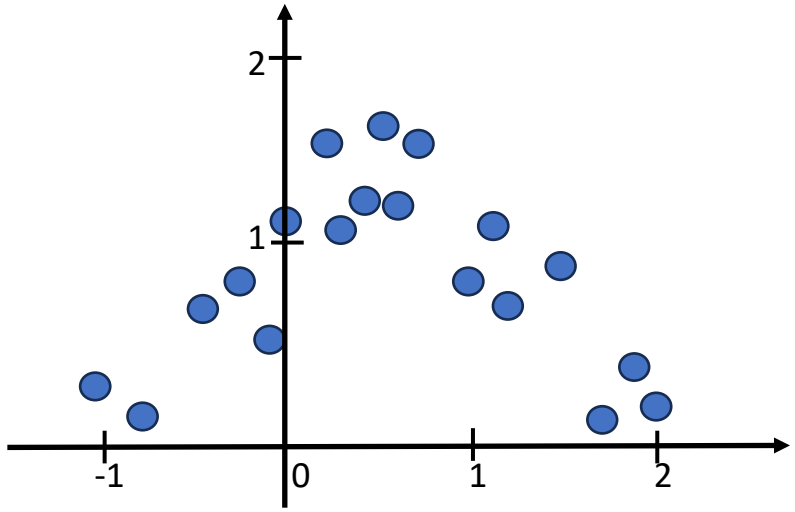


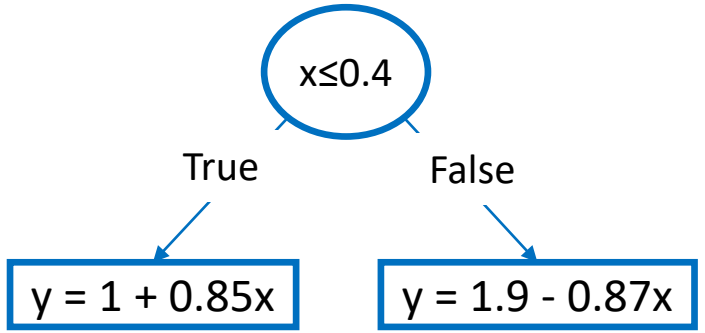



# Ejemplo de árbol de regresión

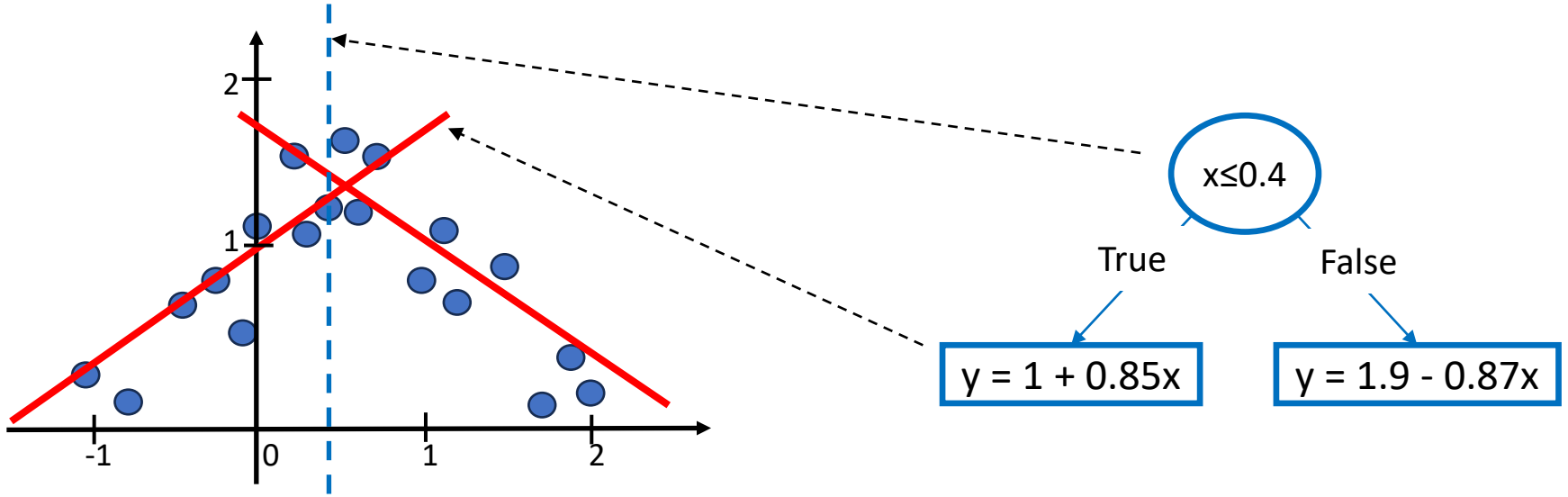


# Ejemplo de árbol de modelo

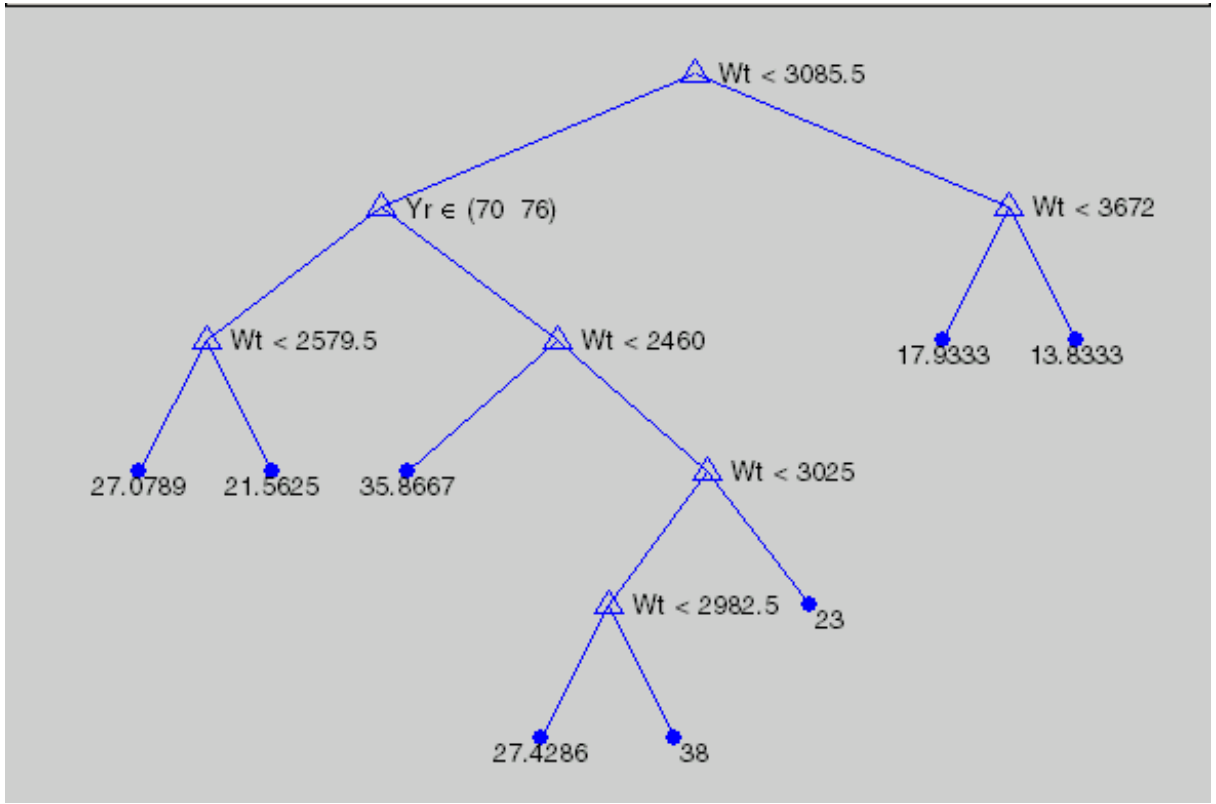


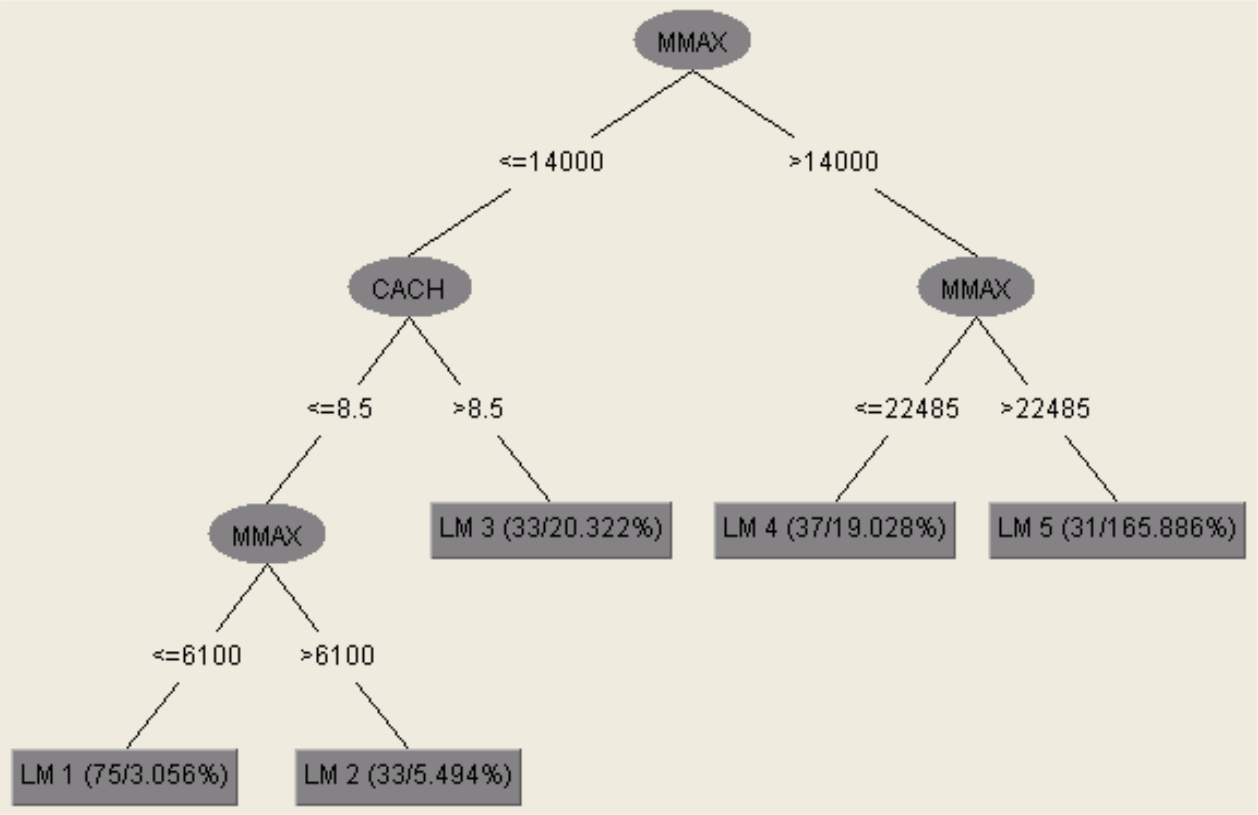
# Ejemplo de árbol de modelo



# Ejemplo de árbol de regresión (más de un atributo)



# Ejemplo de árbol de modelo (más de un atributo)




Ejemplo de modelo de regresión:

LM 4

class =

-21.3079 * vendor=honeywell,ipl,ibm,cdc,ncr

+ 0.7232 * vendor=adviser,sperry,amdahl

+ 20.9924 * vendor=amdahl

- 0.2167 * MYCT

+ 0.0073 * MMIN

+ 0.0122 * MMAX

+ 1.2079 * CACH

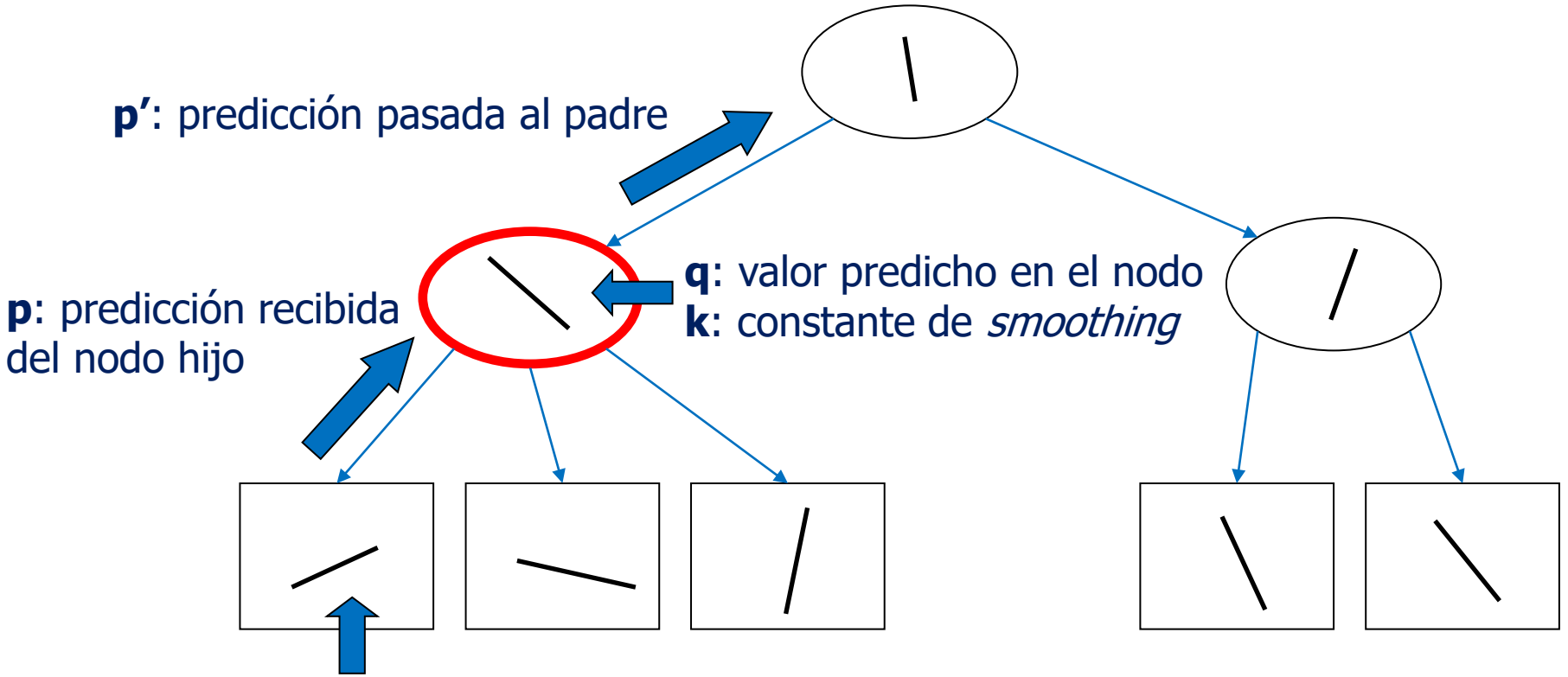
- 0.413 * CHMIN

+ 0.7265 * CHMAX

- 167.7174

# Mejorando la predicción en árboles de modelos

$$p' = \frac{np + kq}{n + k}$$

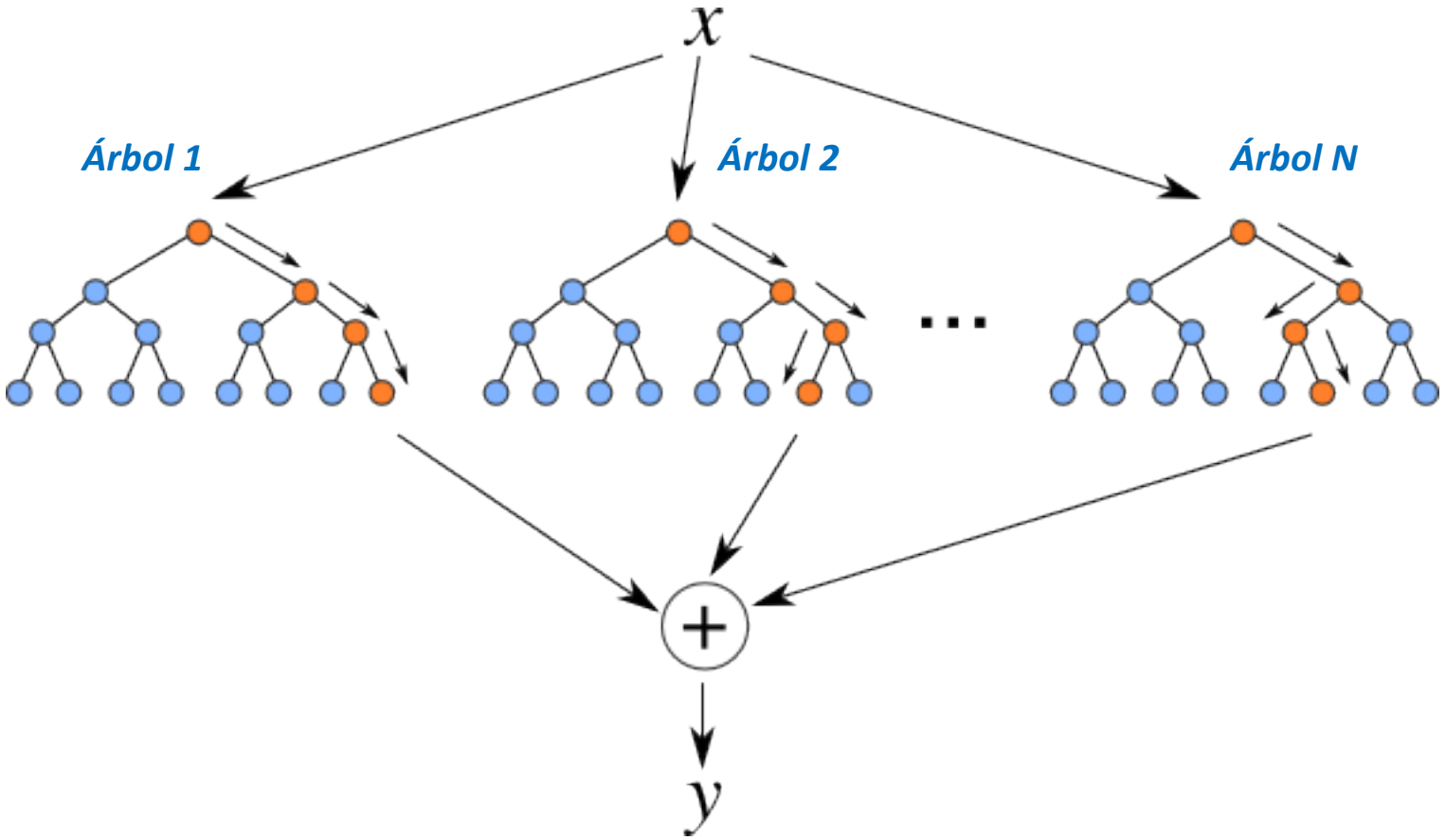


UNED ETS de Ingeniería Informática

# Cómo evitar el overfitting

- Se usan técnicas similares a las usadas con árboles de decisión:
  - Prepoda
  - Postpoda

# *Random Forests*




***Cada Árbol-i:***

- *Se entrena con una muestra del conjunto original de datos, de igual tamaño que éste y obtenida aleatoriamente usando reemplazo (bootstraping).*

- *En cada división del árbol se evalúa un subconjunto aleatorio de atributos, seleccionados aleatoriamente del conjunto de atributos original, y de tamaño prefijado por el usuario.*

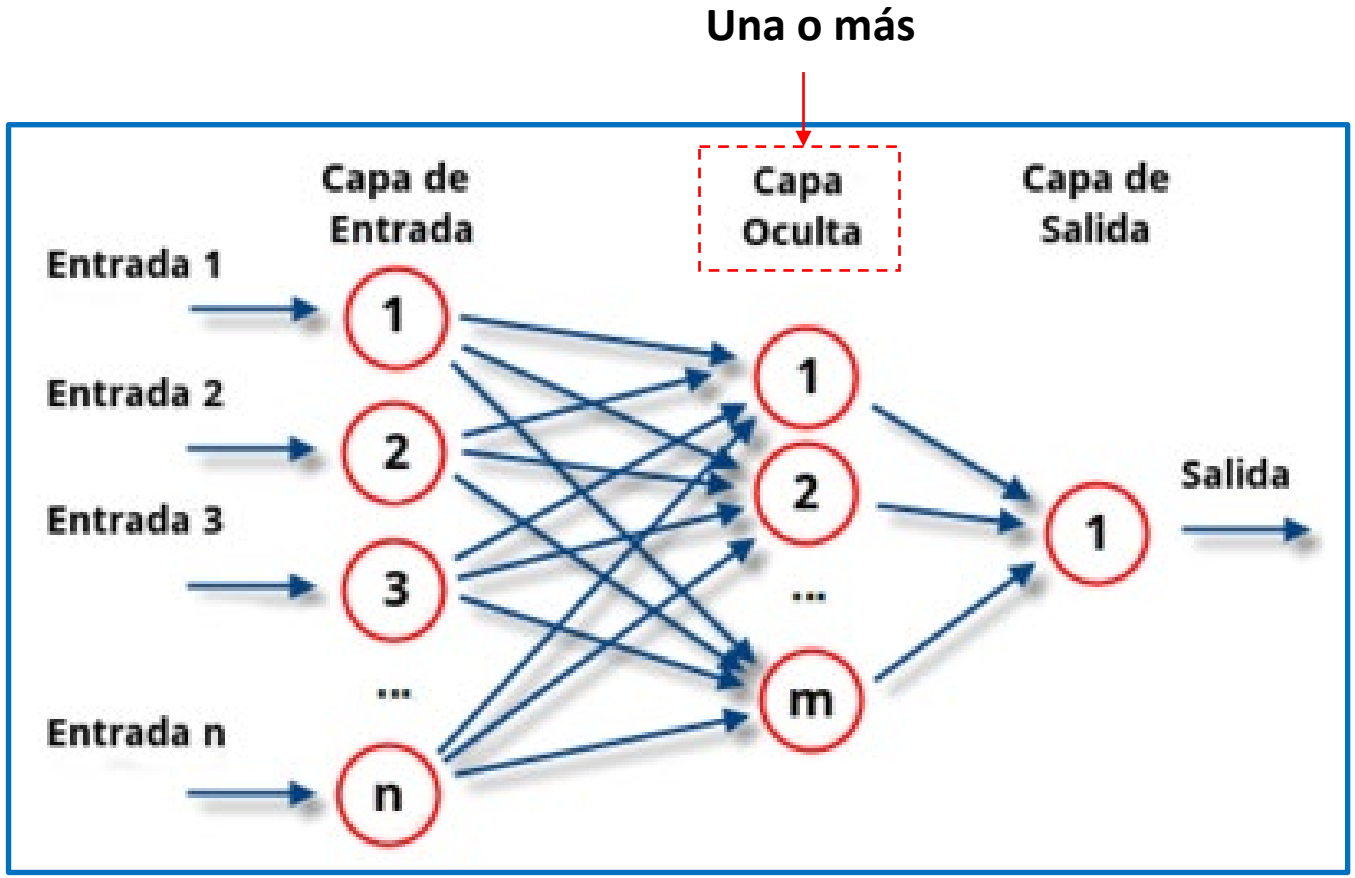
# Contenido

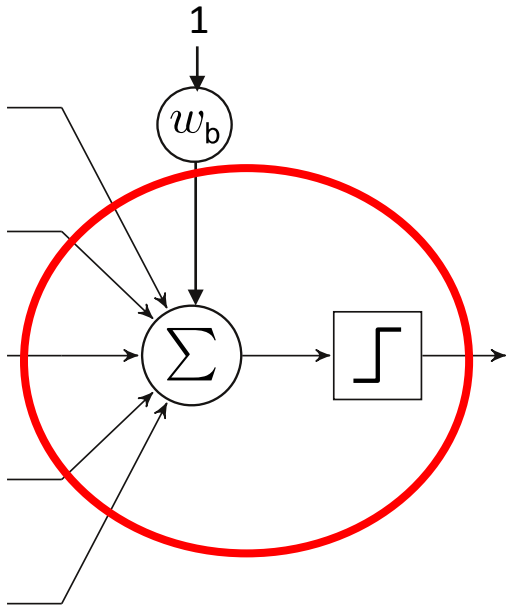
- Algoritmo k-vecinos más cercanos
- Clasificador Naive-Bayes
- Árboles
- Redes neuronales artificiales
- Máquinas de vectores soporte

# Redes Neuronales Artificiales

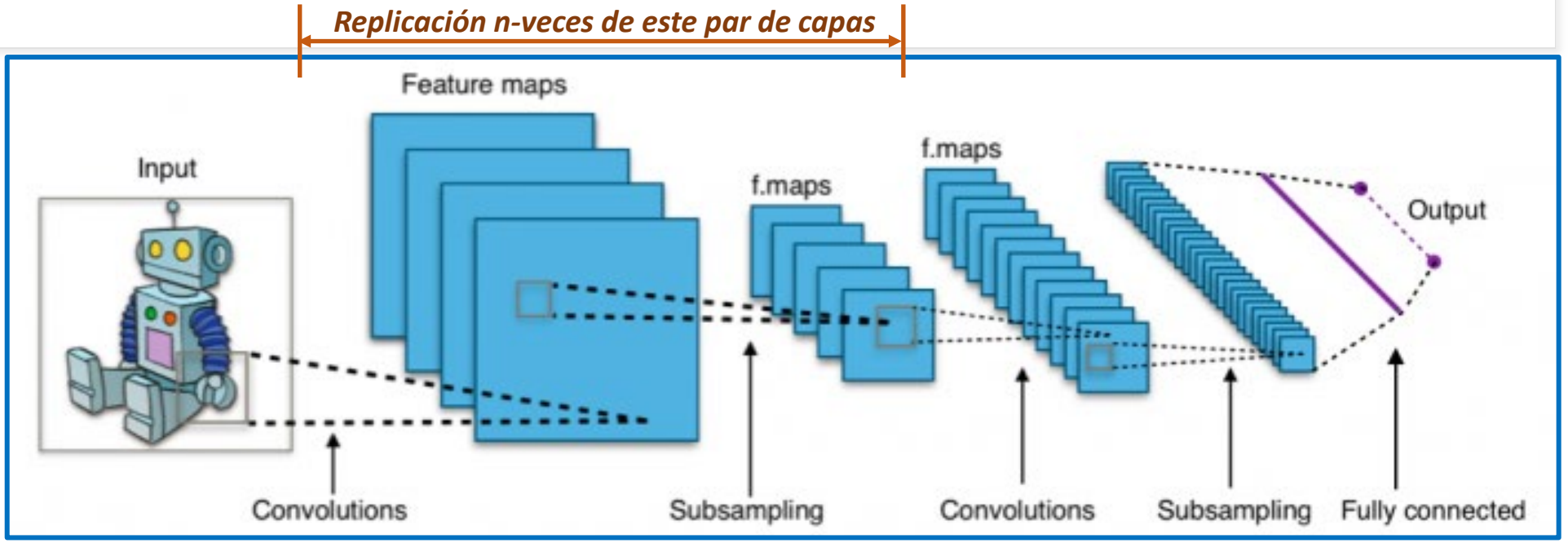
- Clasificación
  - Perceptrón multicapa
  - Redes neuronales convolucionales
- Regresión
  - Perceptrón multicapa
- Método de aprendizaje
  - Algoritmo de retropropagación (basado en el descenso del gradiente)

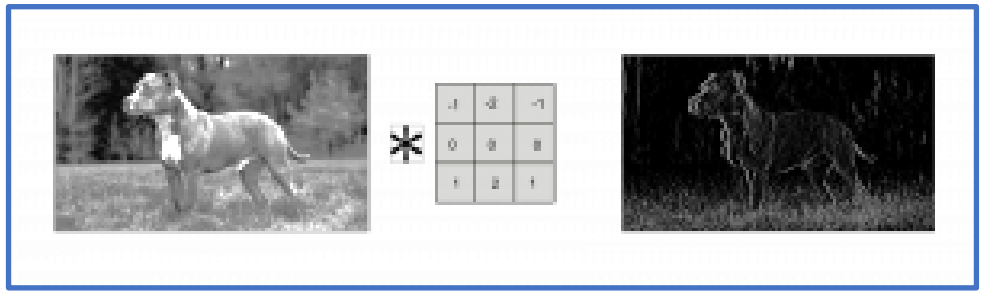
# Perceptrón multicapa

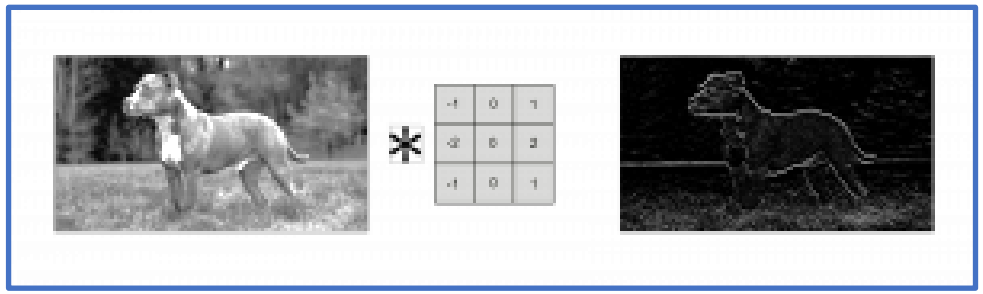


# Redes neuronales convolucionales





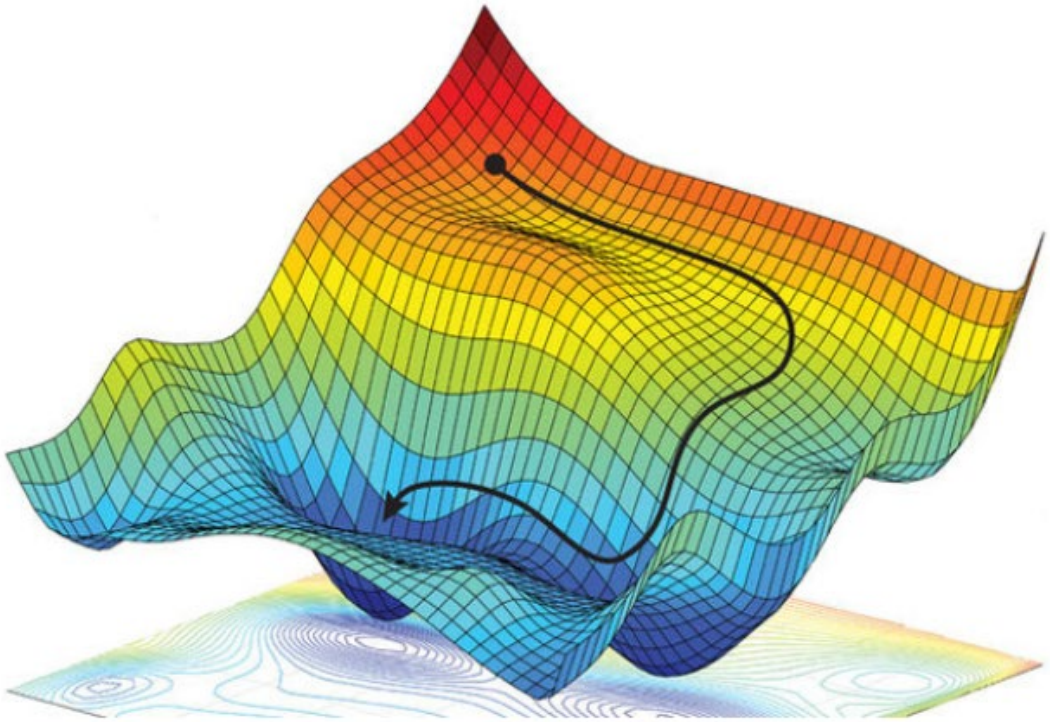
Créditos de las figuras: https://shorturl.at/nrwE9

Aprendizaje Automático en el mundo real

# Cómo se aprenden modelos más complejos

**Descenso del Gradiente**

**Método de *backpropagation* (RNAs)**



*Función de error definida en el espacio de parámetros del modelo*

# Contenido

- Algoritmo k-vecinos más cercanos
- Clasificador Naive-Bayes
- Árboles
- Redes neuronales artificiales
- Máquinas de vectores soporte (SVM)
  - SVM en Clasificación
    - Ejemplos separables linealmente
    - Ejemplos cuasi-separables linealmente
    - Ejemplos no separables linealmente
  - SVM en Regresión (SVR)
    - Regresión lineal o cuasi-lineal
    - Regresión no lineal

UNED ETS de Ingeniería Informática

# Introducción

- Introducidas en los años 90 por Vapnik y colaboradores.
- Permiten abordar tareas de multiclasificación, regresión, agrupamiento, ranking, etc.
- Campos de aplicación:
  - Visión artificial, reconocimiento de caracteres, categorización de texto e hipertexto, clasificación de proteínas, PLN, análisis de series temporales.

UNED ETS de Ingeniería Informática

# Ejemplos separables linealmente

Dado un conjunto separable de ejemplos:

$$S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}, \text{ donde } \boldsymbol{x}_i \in \mathrm{R}^d \text{ e } y_i \in \{+1, -1\}$$

Existe un **hiperplano**, $D(\boldsymbol{x})=0$, capaz de separar dicho conjunto sin error, donde:

$$D(\boldsymbol{x}) = (w_1 x_1 + \ldots + w_d x_d) + b = < \boldsymbol{w}, \boldsymbol{x} > + b, \quad \boldsymbol{w} \in \mathrm{R}^d, b \in \mathrm{R}$$

El hiperplano de separación cumplirá las siguientes restricciones:

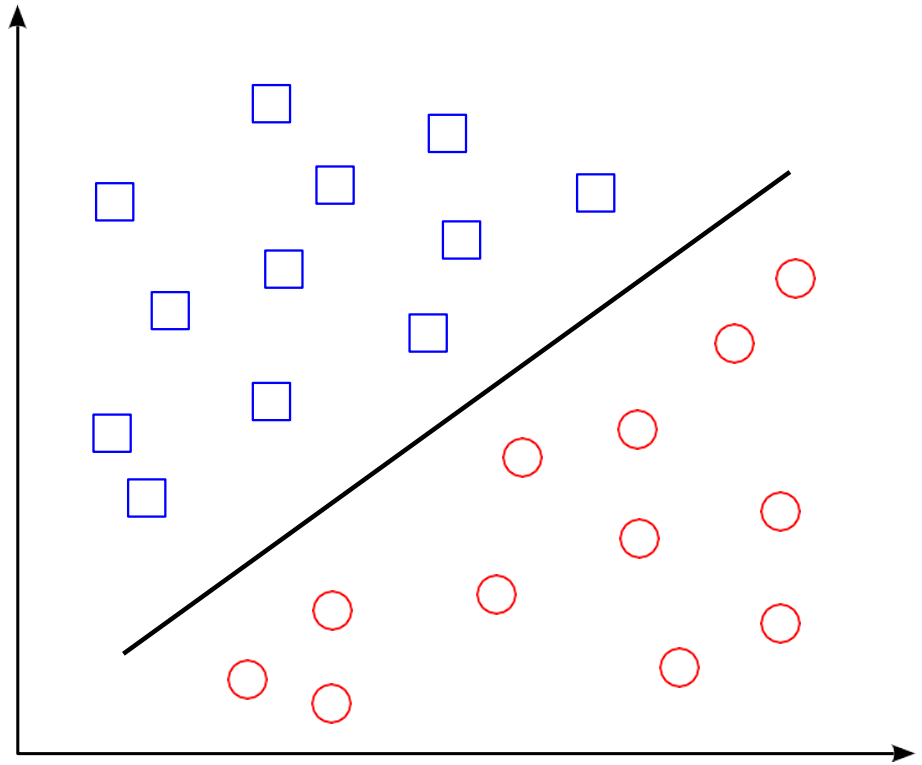
$$y_i (< \boldsymbol{w}, \boldsymbol{x}_i > + b) \geq 0, \ i=1, \ldots, n$$

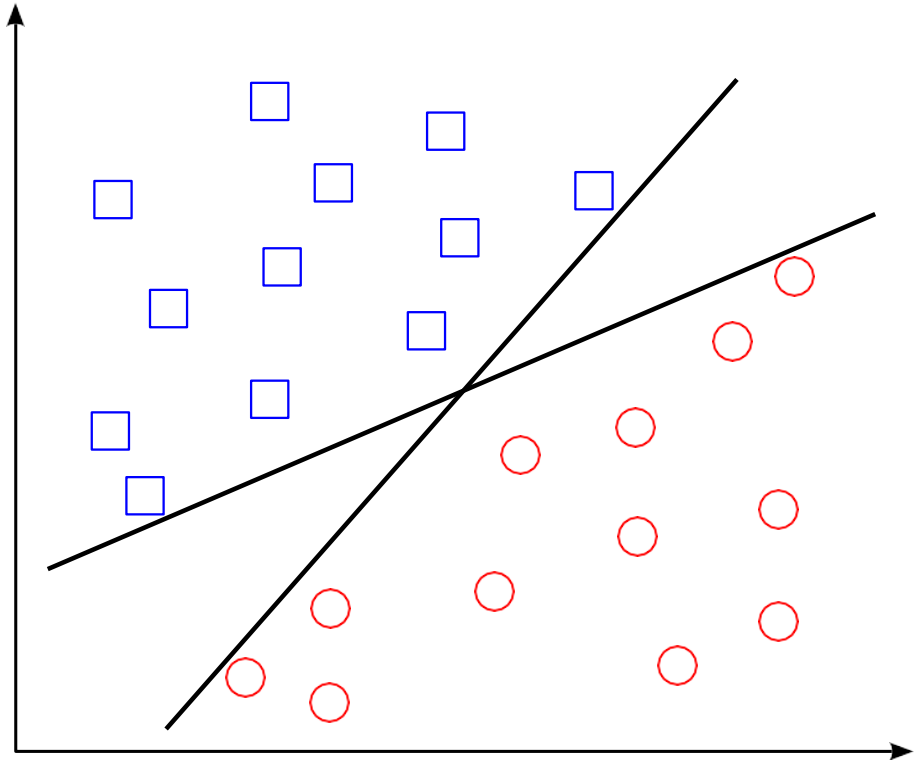
o de forma más compacta:

$$y_i D(\boldsymbol{x}_i) \geq 0, \quad i=1, \ldots, n$$

UNED ETS de Ingeniería Informática

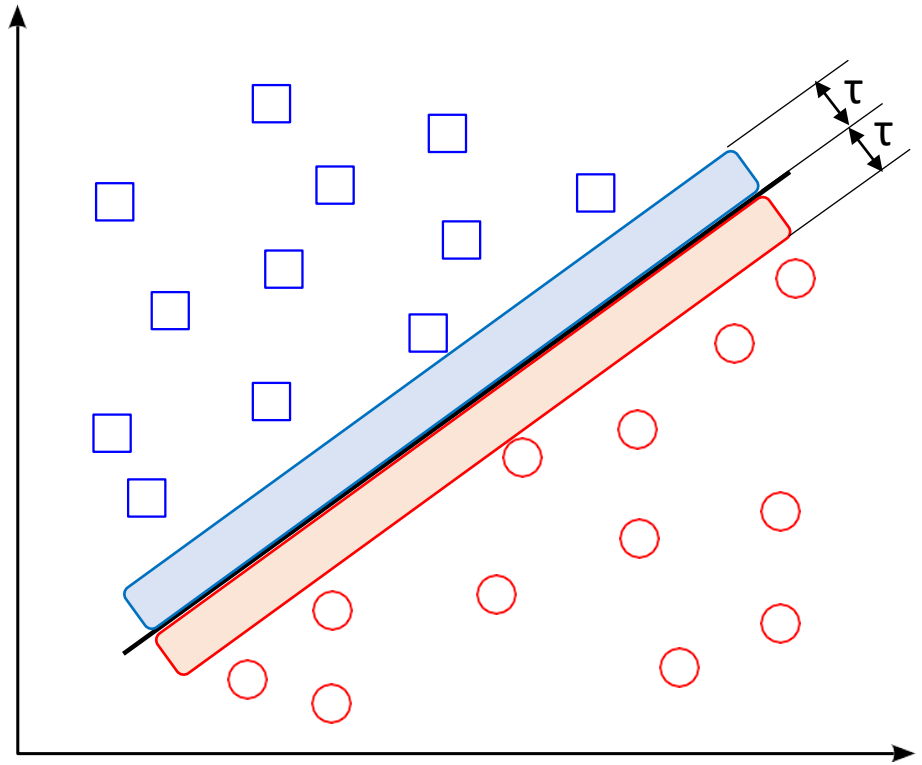
# Infinitos hiperplanos de separación

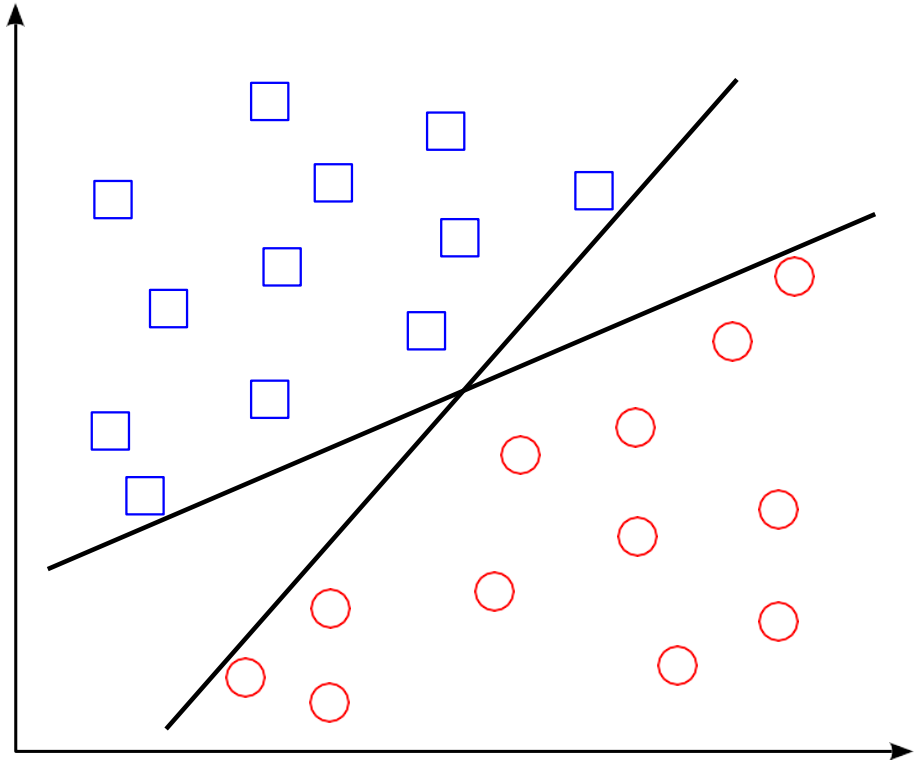




UNED ETS de Ingeniería Informática

# Infinitos hiperplanos de separación



# Hiperplanos paralelos al hiperplano de separación

$$Distancia\left(H_{D(x)}, \boldsymbol{x}'\right) = \frac{|D(\boldsymbol{x}')|}{\|\boldsymbol{w}\|}$$

Por definición de margen máximo:

$$\frac{y_i D(\boldsymbol{x}_i)}{\|\boldsymbol{w}\|} \geq \tau, \qquad i = 1, \ldots, n$$

Los ejemplos situados en las fronteras del margen son, por definición, los vectores soporte ($V_S$):

$$\frac{y_i D(\boldsymbol{x}_i)}{\|\boldsymbol{w}\|} = \tau, \qquad \forall x_i \in V_S$$

La Ec. anterior define dos hiperplanos paralelos al hiperplano original y separados de éste una distancia $\tau$:

$$y_i D(\boldsymbol{x}_i) = \tau \|\boldsymbol{w}\| \quad \Rightarrow \quad \lambda D(x) = 0 \quad \Rightarrow \quad \tau \|\boldsymbol{w}\| = 1$$

# Problema de optimización

$$\tau = 1/\|w\|$$

- La búsqueda del hiperplano óptimo ⇒ **Maximizar $\tau$ ⇒ Minimizar $\|w\|$**

- **Problema de optimización cuadrático (problema primal) ⇒ Problema dual**

$$\begin{aligned} &\text{mín} \quad \tfrac{1}{2}\|\boldsymbol{w}\|^2 \equiv \tfrac{1}{2} < \boldsymbol{w}, \boldsymbol{w} > \\ &\text{s.a}: \quad y_i\left(< \boldsymbol{w}, \boldsymbol{x}_i > + b\right) - 1 \geq 0, \\ &\qquad i = 1, \ldots, n \end{aligned}$$

$$\begin{aligned} &\text{máx} \quad L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \tfrac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j < \boldsymbol{x}_i, \boldsymbol{x}_j > \\ &\text{s.a:} \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \\ &\qquad \alpha_i \geq 0, \; i = 1, \ldots, n \end{aligned}$$

# Cálculo coeficientes del hiperplano separación óptimo

- Solución del problema primal (conocida la solución del problema dual, $\alpha*$):

$$\boldsymbol{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i \boldsymbol{x}_i$$

$$D(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i^* y_i < \boldsymbol{x}, \boldsymbol{x}_i > + b^*$$

$$b^* = \frac{1}{N_{VS}} \sum_{j \in VS} (y_j - < \boldsymbol{w}^*, \boldsymbol{x}_j >)$$

- El hiperplano de separación es función sólo de los $V_S$ ($\alpha_i \neq 0$)

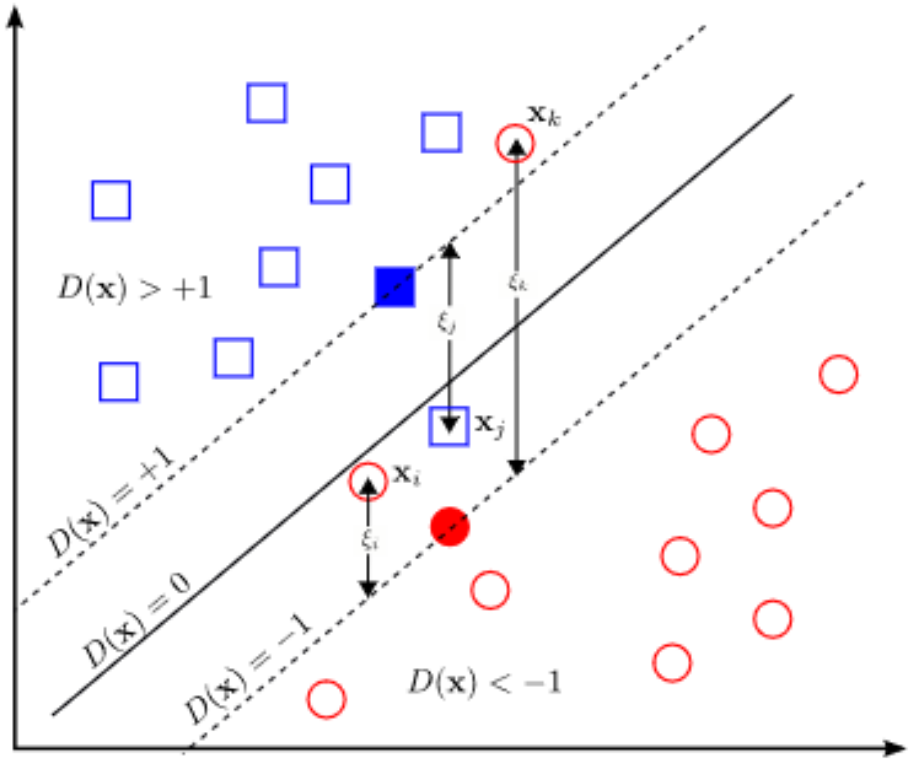
# Ejemplos cuasi-separables linealmente

- La idea es introducir un conjunto de variables de holgura que tolere errores de clasificación:

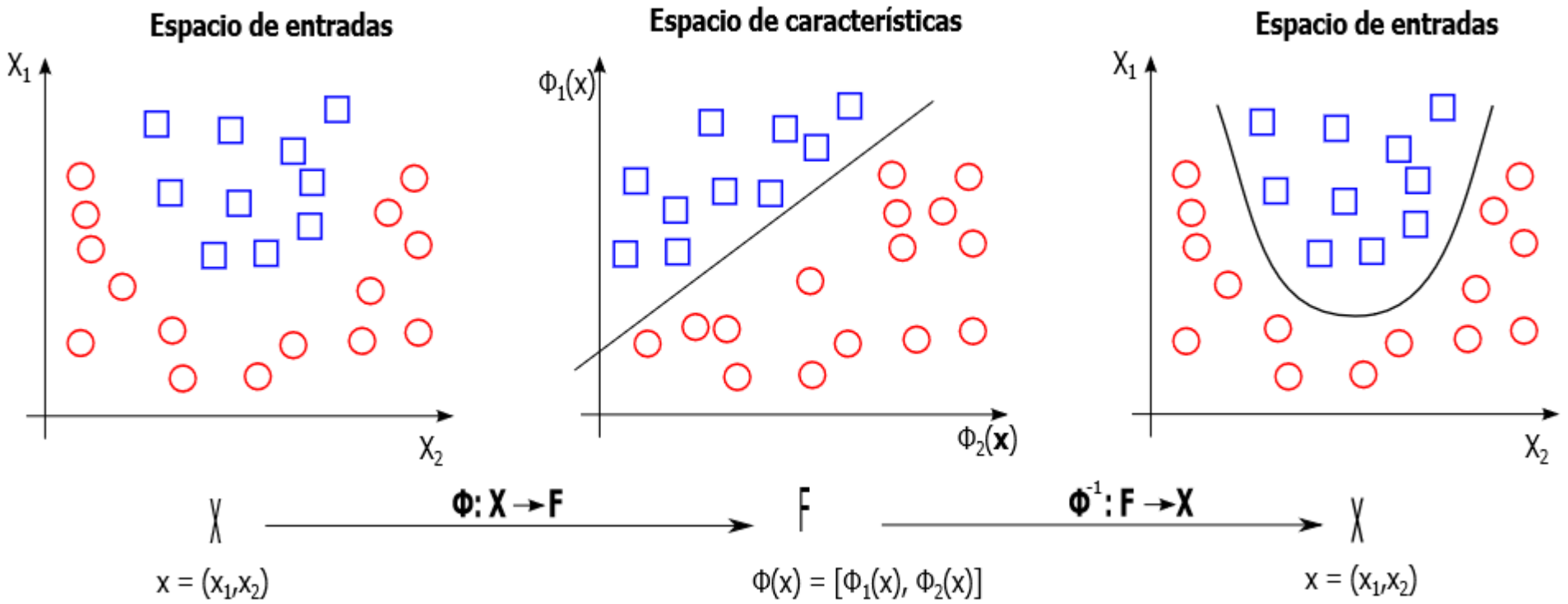
$$\xi_i \geq 0, \ i = 1, \ldots, n,$$

- $(\sum_{i=1}^{n} \xi_i) \Rightarrow$ mide el coste del conjunto de ejemplos no separables.
- Problema optimización primal:

$$\begin{aligned} &\text{mín} \ \ \tfrac{1}{2} < \boldsymbol{w}, \boldsymbol{w} > + C \sum_{i=1}^{n} \xi_i \\ &\text{s.a.} \ \ y_i \left( < \boldsymbol{w}, \boldsymbol{x}_i > + b \right) + \xi_i - 1 \geq 0 \\ &\qquad \xi_i \geq 0, \ i = 1, \ldots, n \end{aligned}$$



# Ejemplos no separables linealmente



# Solución del problema dual en el nuevo espacio

- Sea $\Phi : X \rightarrow F$ una función de transformación, donde:

$$\Phi(\boldsymbol{x}) = [\phi_1(\boldsymbol{x}), \ldots, \phi_m(\boldsymbol{x})] \text{ y } \exists \phi_i(\boldsymbol{x}),\ i = 1, ..., m, \text{ tal que es una función no lineal}$$

- La idea es construir un hiperplano de separación lineal en este nuevo espacio

$$D(\boldsymbol{x}) = (w_1\phi_1(\boldsymbol{x}) + ... + w_m\phi_m(\boldsymbol{x})) =< \boldsymbol{w}, \Phi(\boldsymbol{x}) >$$

- En su forma dual, la función de decisión se obtiene como:

$$D(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i^* y_i < \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}_i) >$$

- Y puede demostrarse que el producto escalar $< \Phi(x), \Phi(x_i) >$ puede calcularse a partir de lo que se denomina función kernel $K(x, x_i)$:

$$D(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i^* y_i K(\boldsymbol{x}, \boldsymbol{x}_i)$$

*Las $\alpha_i^*$ se obtienen como solución del problema dual*

# Funciones base *vs.* Función kernel (Teorema de Aronszajn)

No se necesita conocer el conjunto de funciones base para…

1. construir la función kernel. Basta definir una función que cumpla que la función tiene que ser simétrica y semidefinida positiva.
2. evaluar una función kernel: sólo hay que evaluar esta función
3. resolver el problema dual
4. tener que conocer las coordenadas de los ejemplos transformados en el espacio de características. Por tanto, no importa que el kernel pueda estar asociado a un conjunto infinito de funciones base.

UNED ETS de Ingeniería Informática

# Ejemplos de funciones kernel

- Kernel lineal:

$$K_L(\boldsymbol{x}, \boldsymbol{x}') = < \boldsymbol{x}, \boldsymbol{x}' >$$

- kernel polinómico de grado-$p$:

$$K_P(\boldsymbol{x}, \boldsymbol{x}') = \left[\gamma < \boldsymbol{x}, \boldsymbol{x}' > + \kappa\right]^p, \ \gamma > 0$$

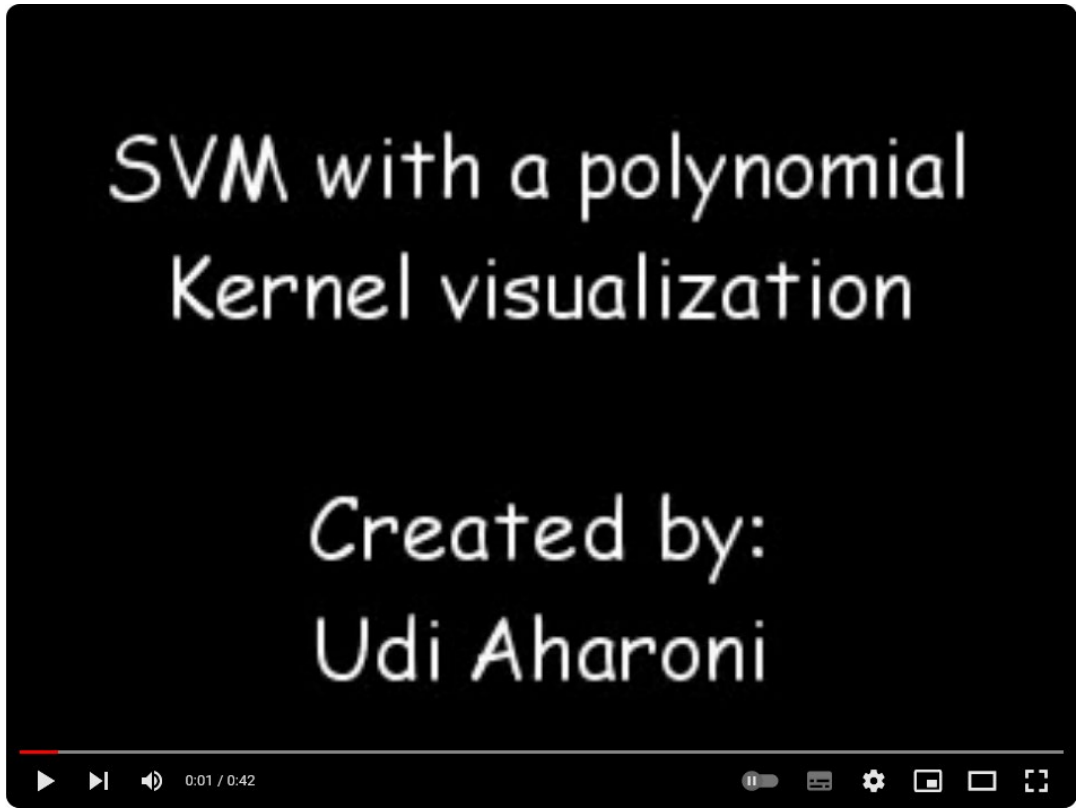
- kernel gaussiano:

$$K_G(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\gamma \left\|\boldsymbol{x} - \boldsymbol{x}'\right\|^2\right) \equiv \exp\left(-\gamma < (\boldsymbol{x} - \boldsymbol{x}'), (\boldsymbol{x} - \boldsymbol{x}') >\right), \ \gamma > 0$$

- kernel sigmoidal:

$$K_S(\boldsymbol{x}, \boldsymbol{x}') = \tanh(\gamma < \boldsymbol{x}, \boldsymbol{x}' > + \kappa)$$

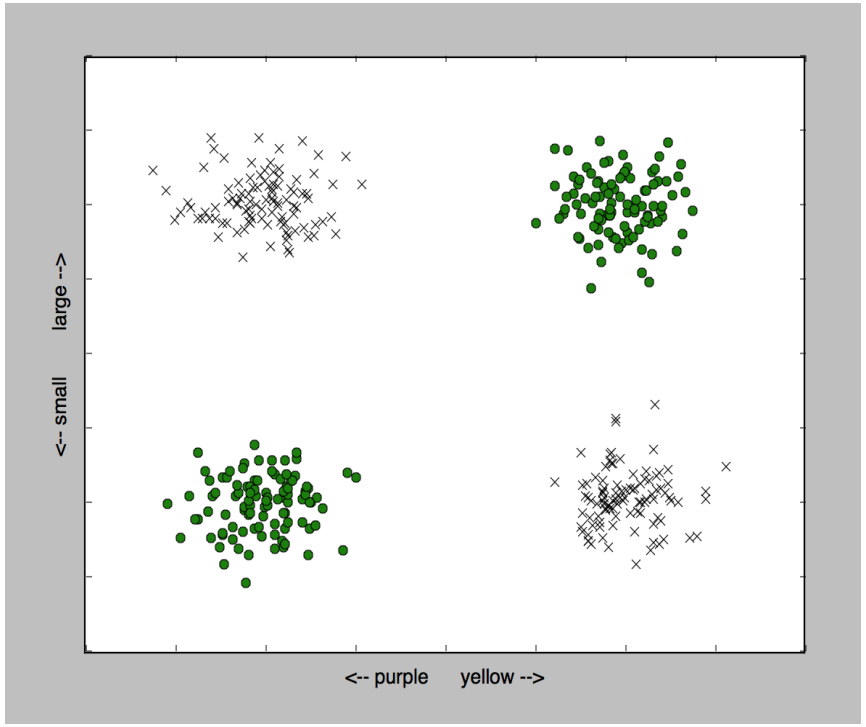
# Ejemplo de SVM con kernels



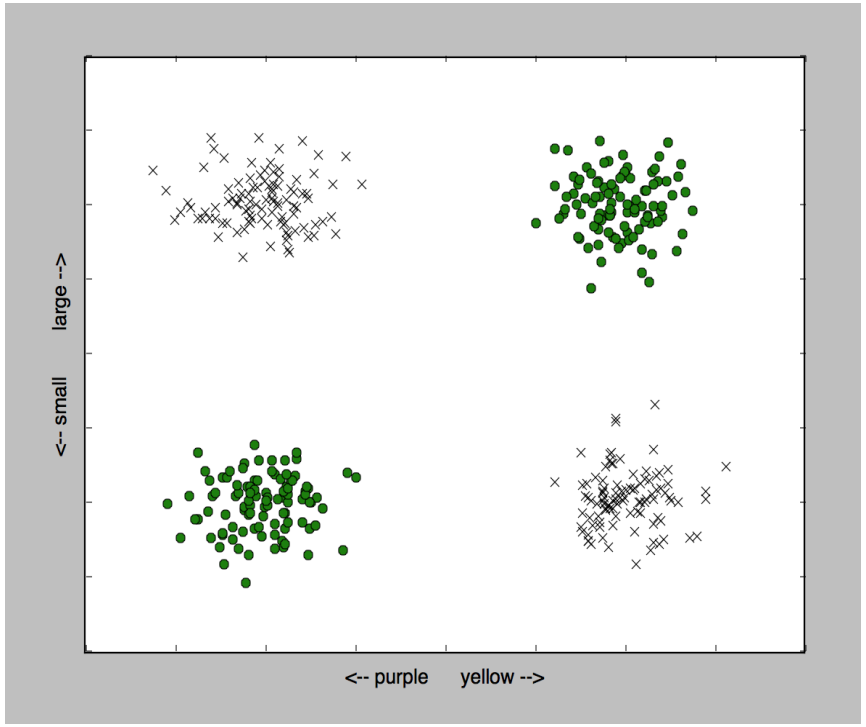

Enlace al vídeo: https://shorturl.at/wFIKT

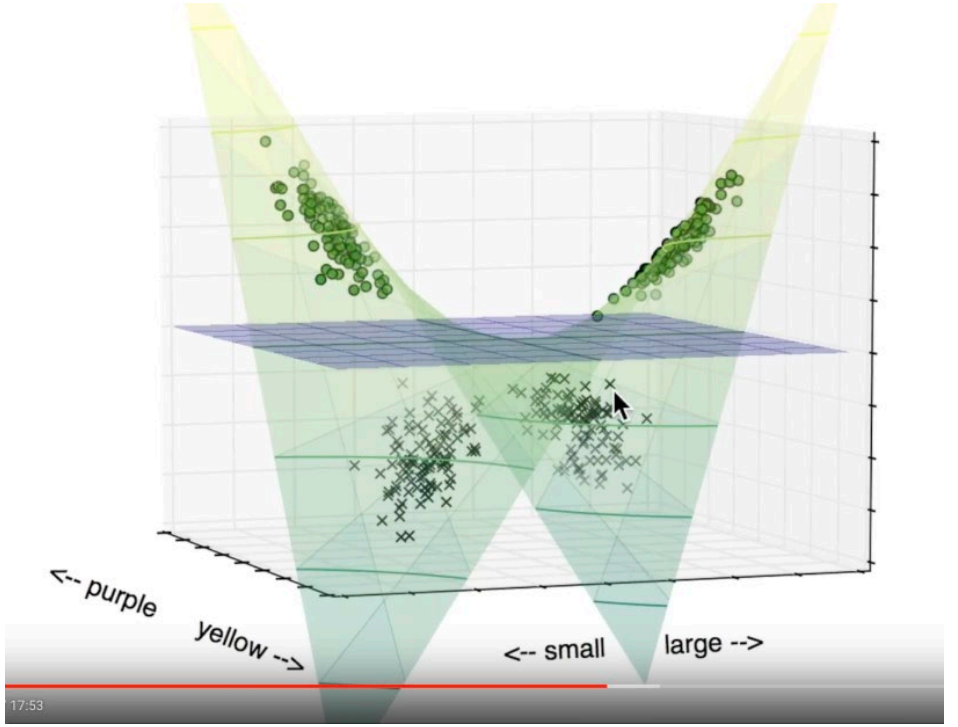
UNED ETS de Ingeniería Informática

# Ejemplo de SVM con kernels



# Ejemplo de SVM con kernels

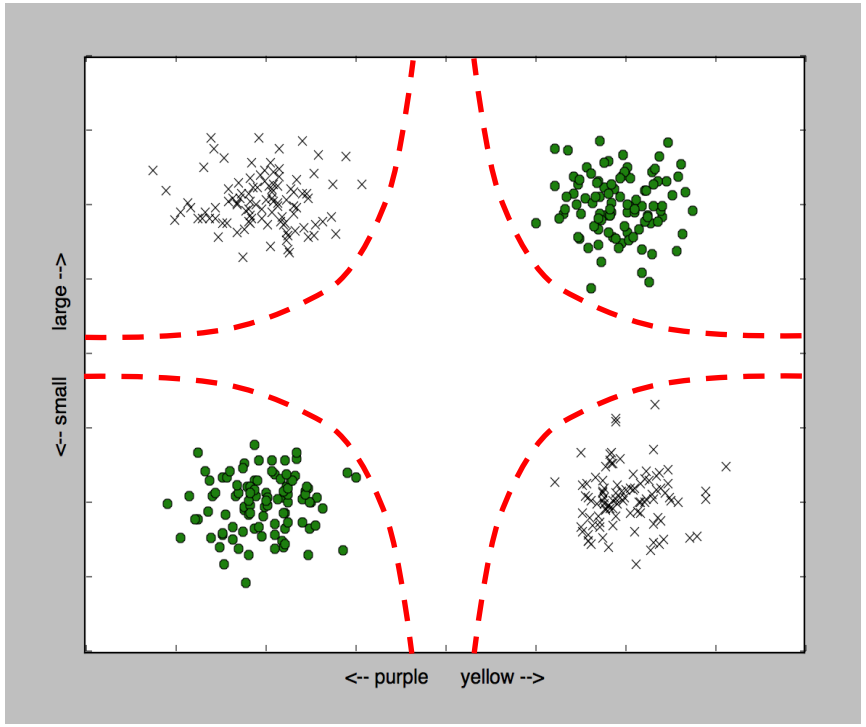


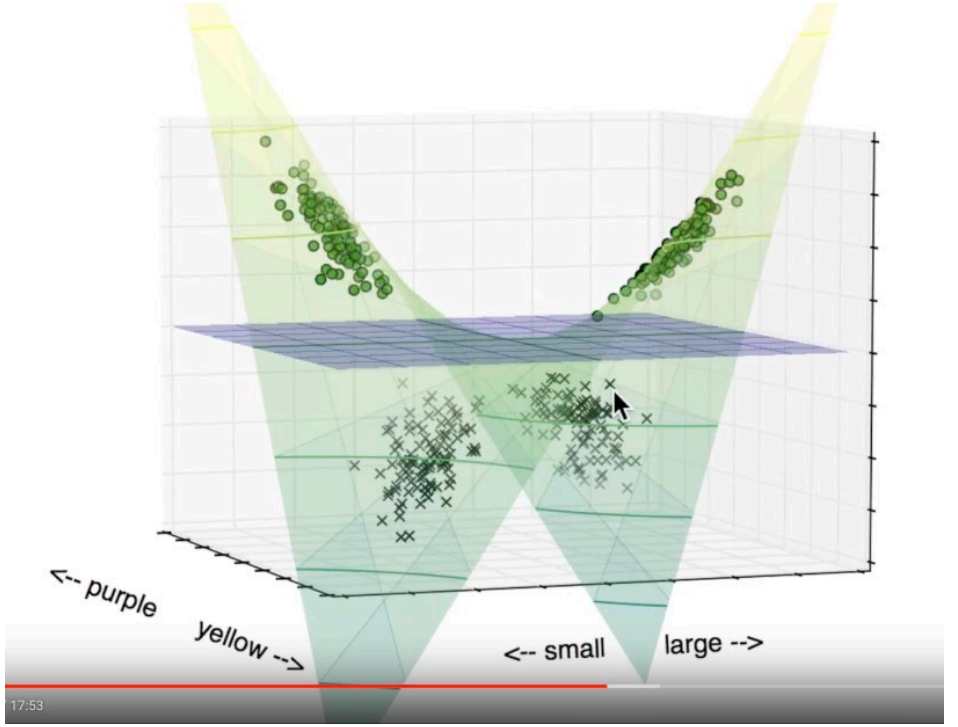




Créditos de figuras: https://shorturl.at/mxDFP

UNED ETS de Ingeniería Informática

# Ejemplo de SVM con kernels



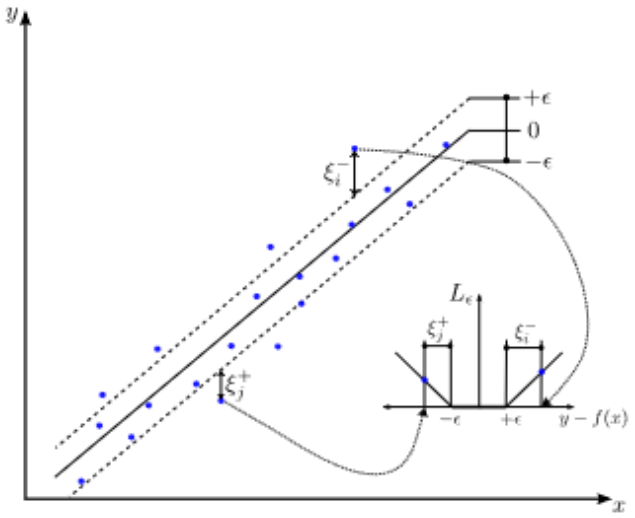
UNED ETS de Ingeniería Informática

# SVM para *k* clases (*k*>2)

- *One-versus-one*
  - Genera y evalúa $k(k-1)/2$ clasificadores SVMs
- *One-versus-all*
  - Genera $k$ clasificadores SVMs
- *DAGSVM* (*Directed Acyclic Graph SVM*)
  - Genera $k(k-1)/2$ clasificadores SVMs, pero no es necesario evaluar todos

# SVM para regresión (SVR)

## *Regresión lineal o cuasi-lineal*




$$\begin{aligned} \text{mín} \quad & \tfrac{1}{2} < \boldsymbol{w}, \boldsymbol{w} > + C \sum_{i=1}^{n} \left(\xi_i^+ + \xi_i^-\right) \\ \text{s.a.} \quad & (< \boldsymbol{w}, \boldsymbol{x}_i > + b) - y_i - \epsilon - \xi_i^+ \leq 0 \\ & y_i - (< \boldsymbol{w}, \boldsymbol{x}_i > + b) - \epsilon - \xi_i^- \leq 0 \\ & \xi_i^+, \xi_i^- \geq 0, \; i = 1, \ldots, n \end{aligned}$$

→ Forma dual y se resuelve

## *Regresión no lineal*

- Los ejemplos se transforman en un nuevo espacio (espacio de características), en el que sea posible el ajuste con un regresor lineal.
- El regresor asociado a la función lineal en el nuevo espacio es:

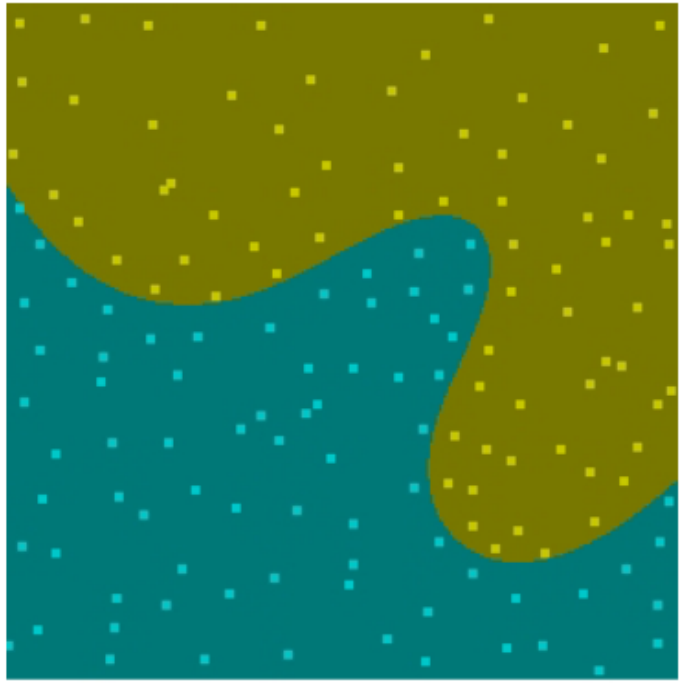
$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \left(\alpha_i^{*-} - \alpha_i^{*+}\right) K(\boldsymbol{x}, \boldsymbol{x}_i)$$

*Las $\alpha_i^{*-}$ y $\alpha_i^{*+}$ se obtienen como solución del problema dual*

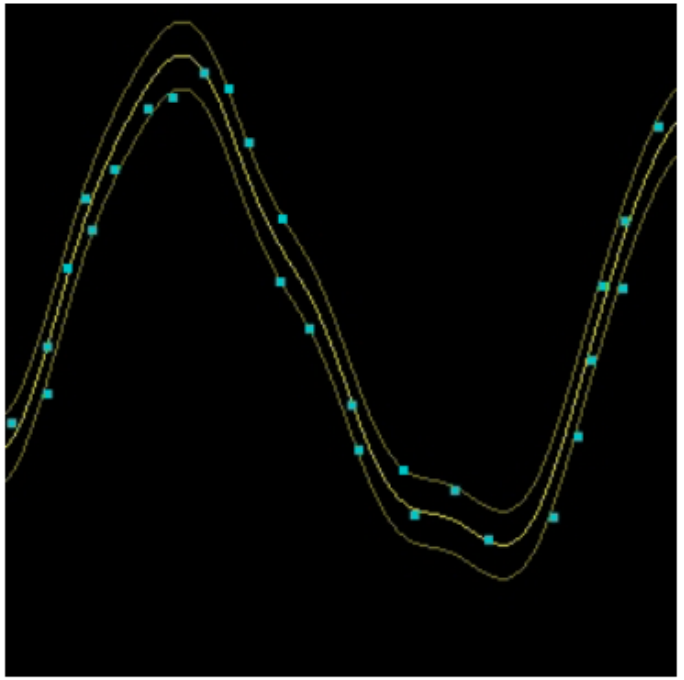
UNED ETS de Ingeniería Informática

# Interfaz gráfico interactivo (SVMs y SVRs)

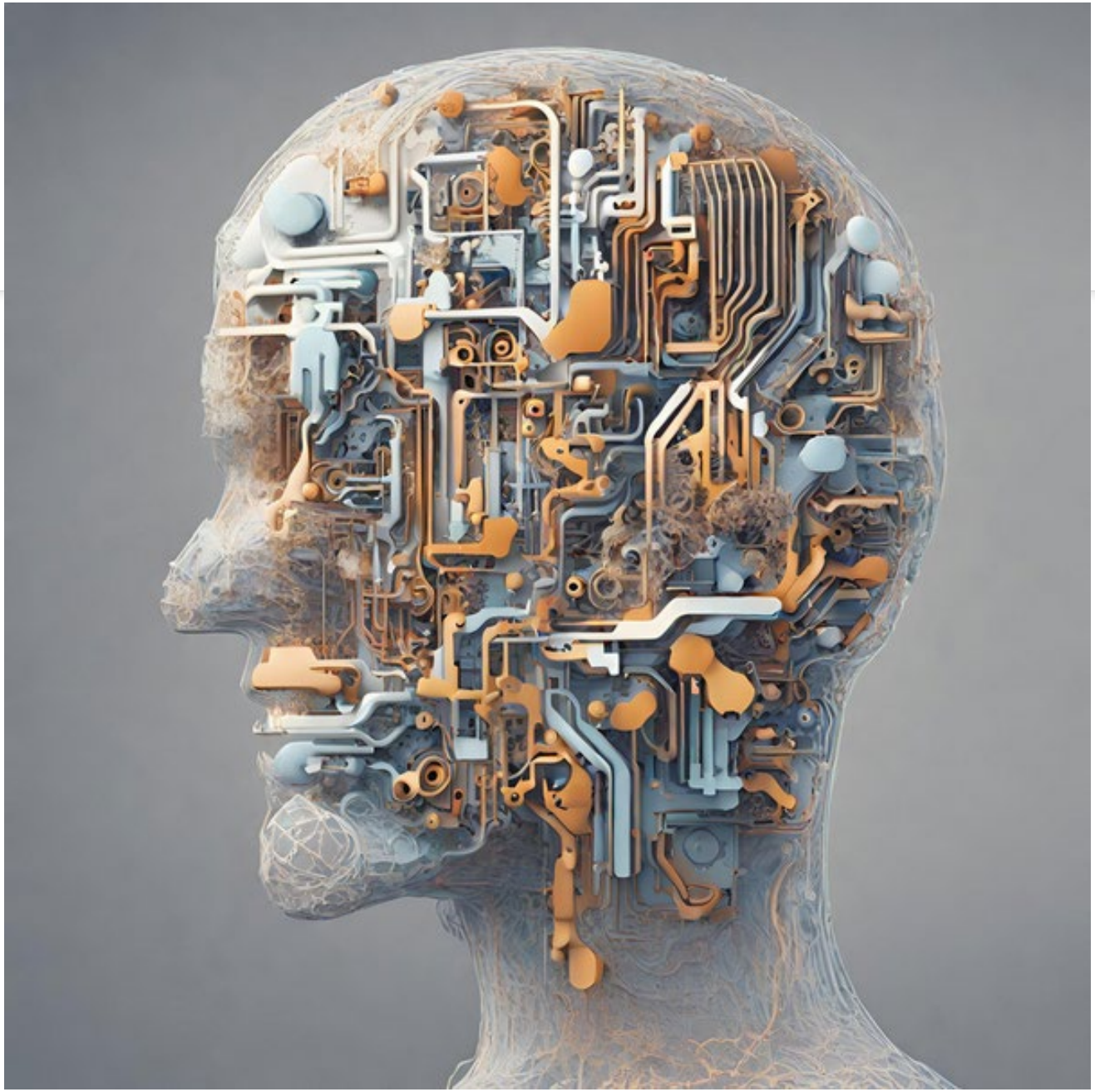
- Librería LBSVM: https://www.csie.ntu.edu.tw/~cjlin/libsvm/



(a) Ejemplo de clasificación



(b) Ejemplo de regresión



UNED ETS de Ingeniería Informática