# Tema 6: Evaluación de Modelos de Clasificación Supervisada

Abdelmalik Moujahid, Iñaki Inza y Pedro Larrañaga

Departamento de Ciencias de la Computación e Inteligencia Artificial

Universidad del País Vasco

http://www.sc.ehu.es/isg/

# Evaluación de Modelos de Clasificación Supervisada

- Introducción
- Estimación basada en precisión
- Estimación basada en coste
- Análisis ROC (Receiver Operating Characteristic)

# Introducción

*Clasificación Supervisada*

| | $X_1$ | $\ldots$ | $X_n$ | $C$ |
|---|---|---|---|---|
| $(\boldsymbol{x}^{(1)}, c^{(1)})$ | $x_1^{(1)}$ | $\ldots$ | $x_n^{(1)}$ | $c^{(1)}$ |
| $(\boldsymbol{x}^{(2)}, c^{(2)})$ | $x_1^{(2)}$ | $\ldots$ | $x_n^{(2)}$ | $c^{(2)}$ |
| $\ldots$ | | $\ldots$ | | $\ldots$ |
| $(\boldsymbol{x}^{(N)}, c^{(N)})$ | $x_1^{(N)}$ | $\ldots$ | $x_n^{(N)}$ | $c^{(N)}$ |
| $\boldsymbol{x}^{(N+1)}$ | $x_1^{(N+1)}$ | $\ldots$ | $x_n^{(N+1)}$ | ??? |

# Introducción

*Clasificación Supervisada*

| | $X_1$ | $\ldots$ | $X_n$ | $C$ | $C_M$ |
|---|---|---|---|---|---|
| $(\boldsymbol{x}^{(1)}, c^{(1)})$ | $x_1^{(1)}$ | $\ldots$ | $x_n^{(1)}$ | $c^{(1)}$ | $c_M^{(1)}$ |
| $(\boldsymbol{x}^{(2)}, c^{(2)})$ | $x_1^{(2)}$ | $\ldots$ | $x_n^{(2)}$ | $c^{(2)}$ | $c_M^{(2)}$ |
| $\ldots$ | | $\ldots$ | | $\ldots$ | $\ldots$ |
| $(\boldsymbol{x}^{(N)}, c^{(N)})$ | $x_1^{(N)}$ | $\ldots$ | $x_n^{(N)}$ | $c^{(N)}$ | $c_M^{(N)}$ |

Número de aciertos: $\sum_{i=1}^{N} \delta(c^{(i)}, c_M^{(i)})$

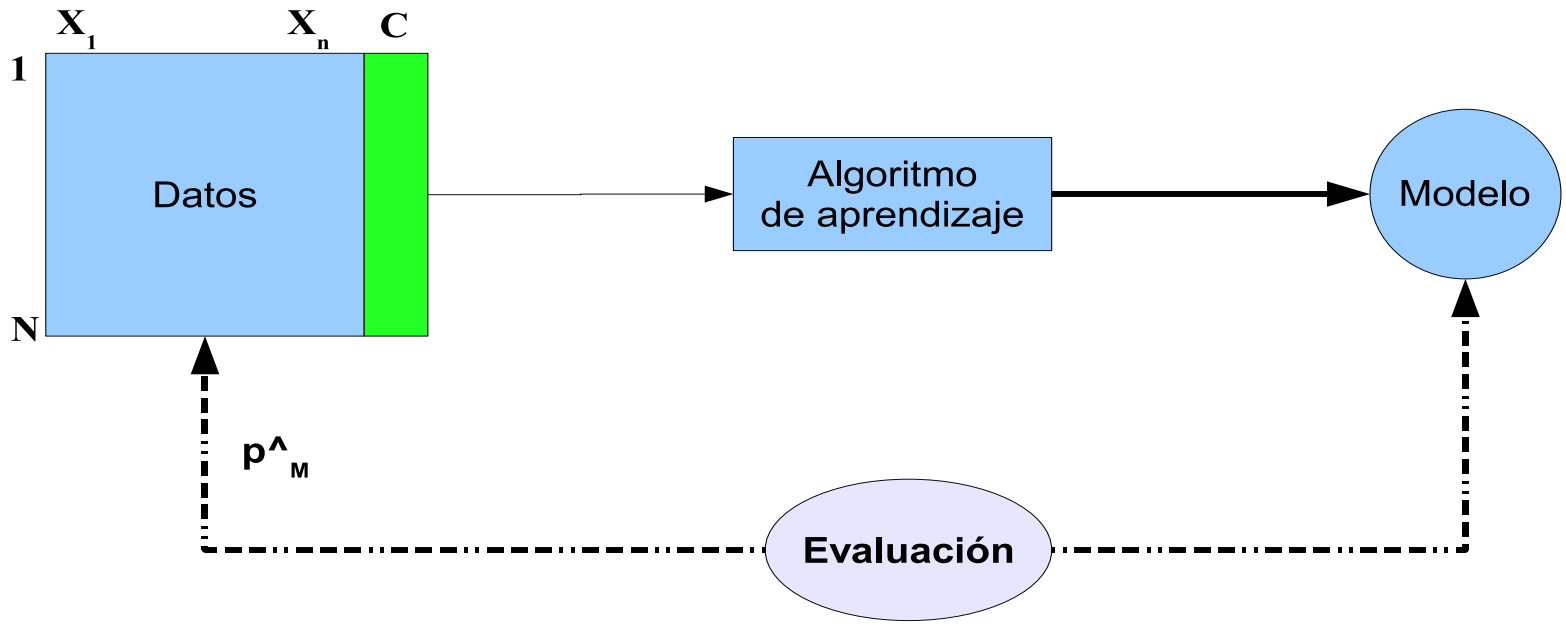
$$\delta(c^{(i)}, c_M^{(i)}) = \begin{cases} 1 & \text{si } c^{(i)} = c_M^{(i)} \\ 0 & \text{si } c^{(i)} \neq c_M^{(i)} \end{cases}$$

# Introducción

| | | $C$ Clase real + | $C$ Clase real - |
|---|---|---|---|
| $C_M$ Clase predicha | + | TP | FP |
| | - | FN | TN |

- Tasa de acierto: $\frac{TP+TN}{TP+FP+FN+TN}$
- Tasa de error: $\frac{FN+FP}{TP+FP+FN+TN}$
- Sensibilidad: $TPR = \frac{TP}{TP+FN}$
- Especifidad: $TNR = \frac{TN}{FP+TN}$
- Proporción de falsos positivos: $FPR = \frac{FP}{FP+TN}$
- Proporción de falsos negativos: $FNR = \frac{FN}{TP+FN}$

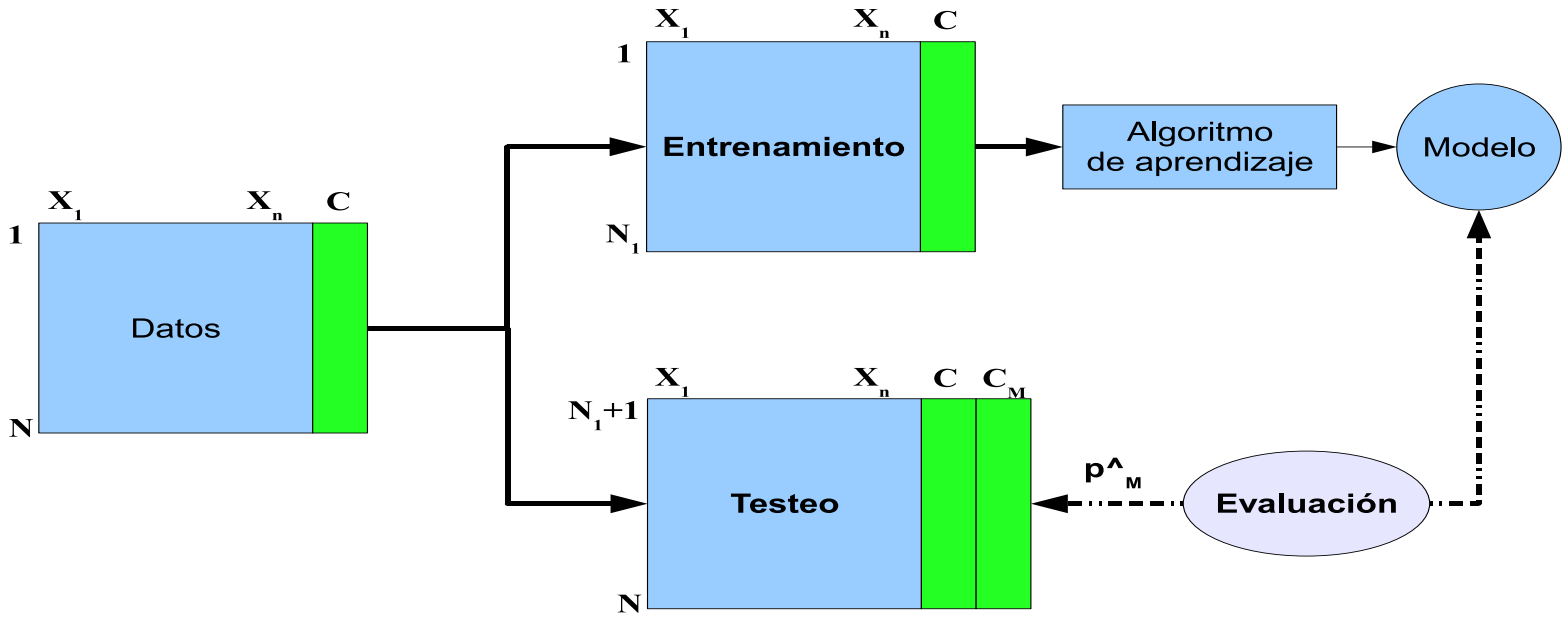
# Estimación basada en precisión




$$\hat{p}_M = \frac{1}{N} \sum_{i=1}^{N} \delta(c^{(i)} = c_M^{(i)})$$

Método no honesto de estimación

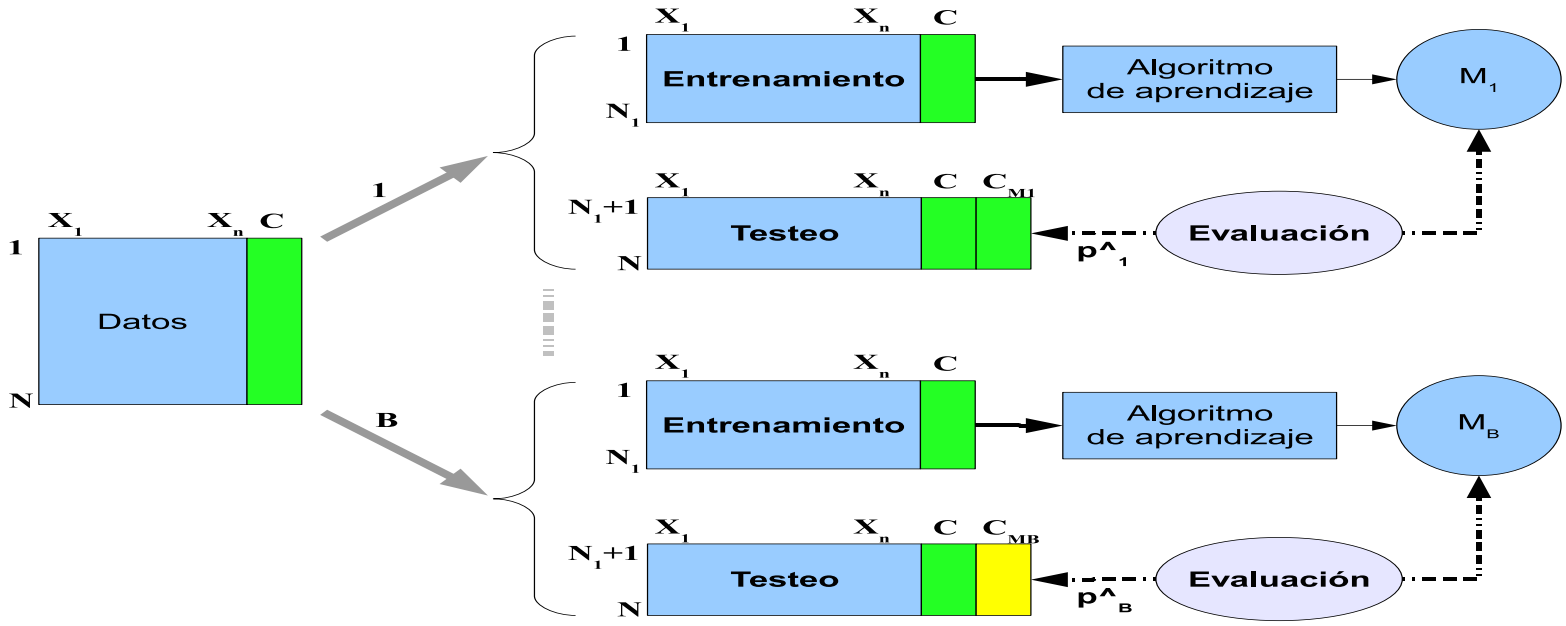
# Estimación basada en precisión




$$\hat{p}_M = \frac{1}{N - N_1} \sum_{i=1}^{N-N_1} \delta\left(c^{(N_1+i)} = c_M^{(N_1+i)}\right)$$

Método $H$ de estimación basado en entrenamiento y testeo

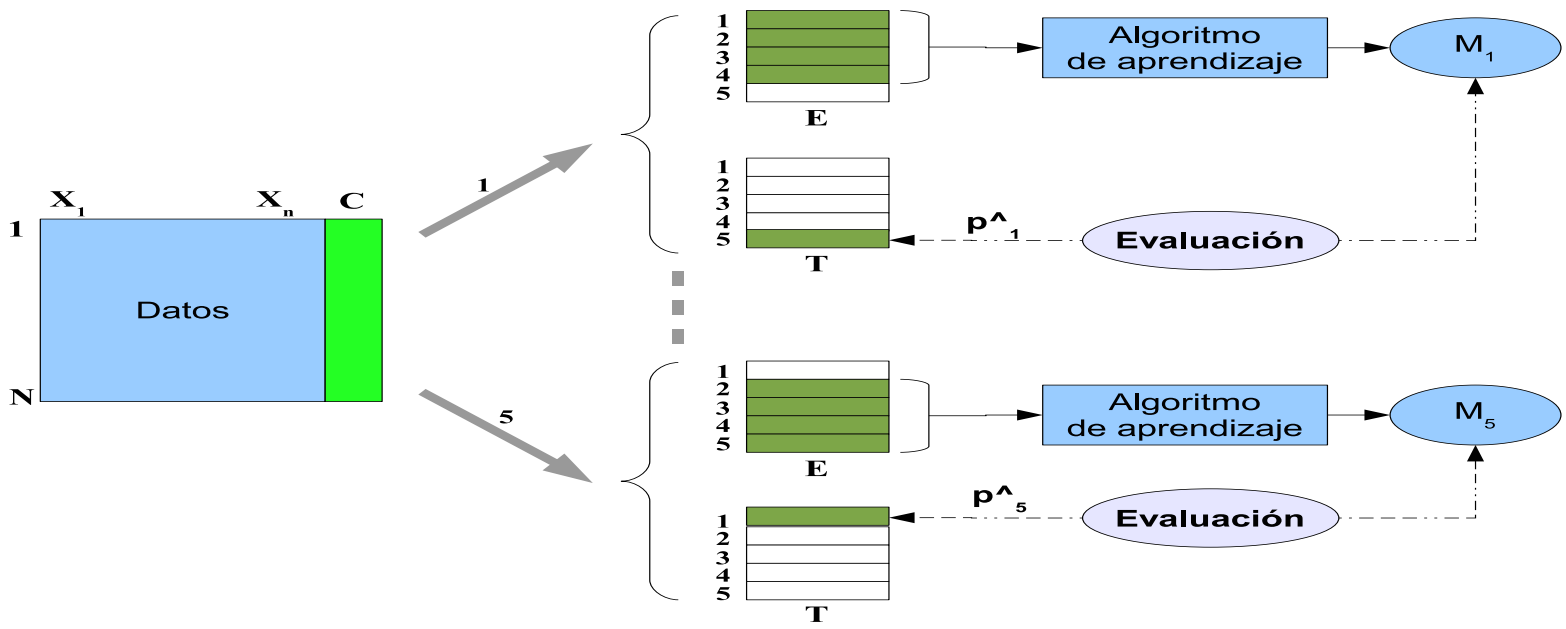
# Estimación basada en precisión




$$\hat{p}_M = \frac{1}{B}\sum_{i=1}^{B}\hat{p}_i$$

Método de estimación $H$ repetidas veces

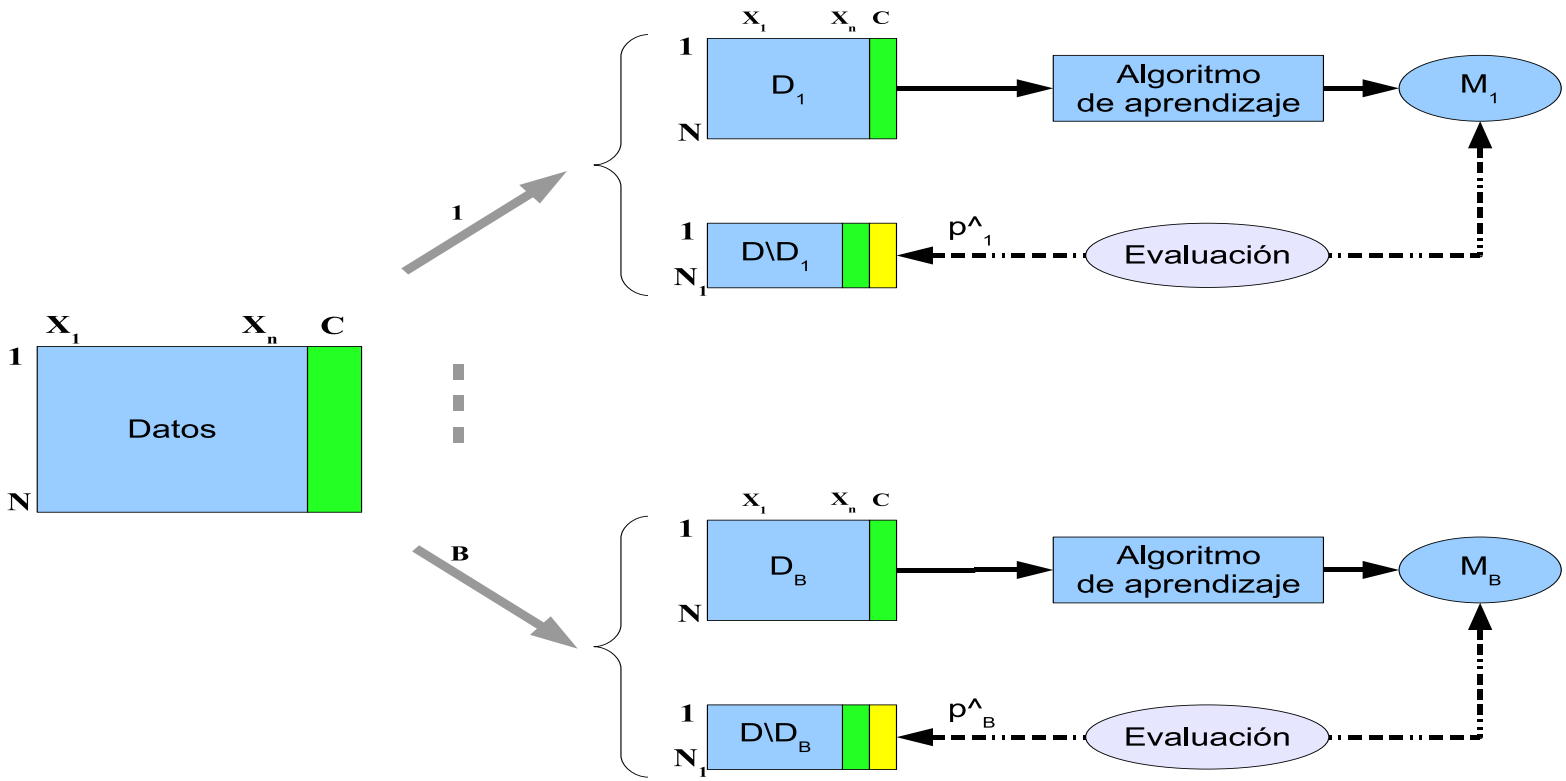
# Estimación basada en precisión




$$\hat{p}_M = \frac{1}{k} \sum_{i=1}^{k} \hat{p}_i$$

Método de estimación basado en $k$ rodajas ($k$–fold cross validation). Si $k = N$ leave one out

# Estimación basada en precisión




$$\hat{p}_{test} = \frac{1}{B}\sum_{i=1}^{B}\hat{p}_i \qquad \hat{p}_M = \hat{p}_{0,632Bo} = (0{,}368\hat{p}_{entrenamiento} + 0{,}632\hat{p}_{test})$$

Método de estimación $0{,}632$ booststraping

# Estimación basada en precisión

Sobre los distintos métodos:

- *Método H*: utilizarlo con $N$ grande
- *Método H repetidas veces*: no hay control sobre los casos usados como entrenamiento (testeo)
- *Método de estimación basado en $k$ rodajas ($k$–fold cross validation)*: estimación insesgada de la probabilidad de acierto, pero con alta varianza
- *Método de estimación $0{,}632$ booststraping*: insesgada en el límite y con baja varianza

# Estimación basada en coste

Evaluación sensible al coste

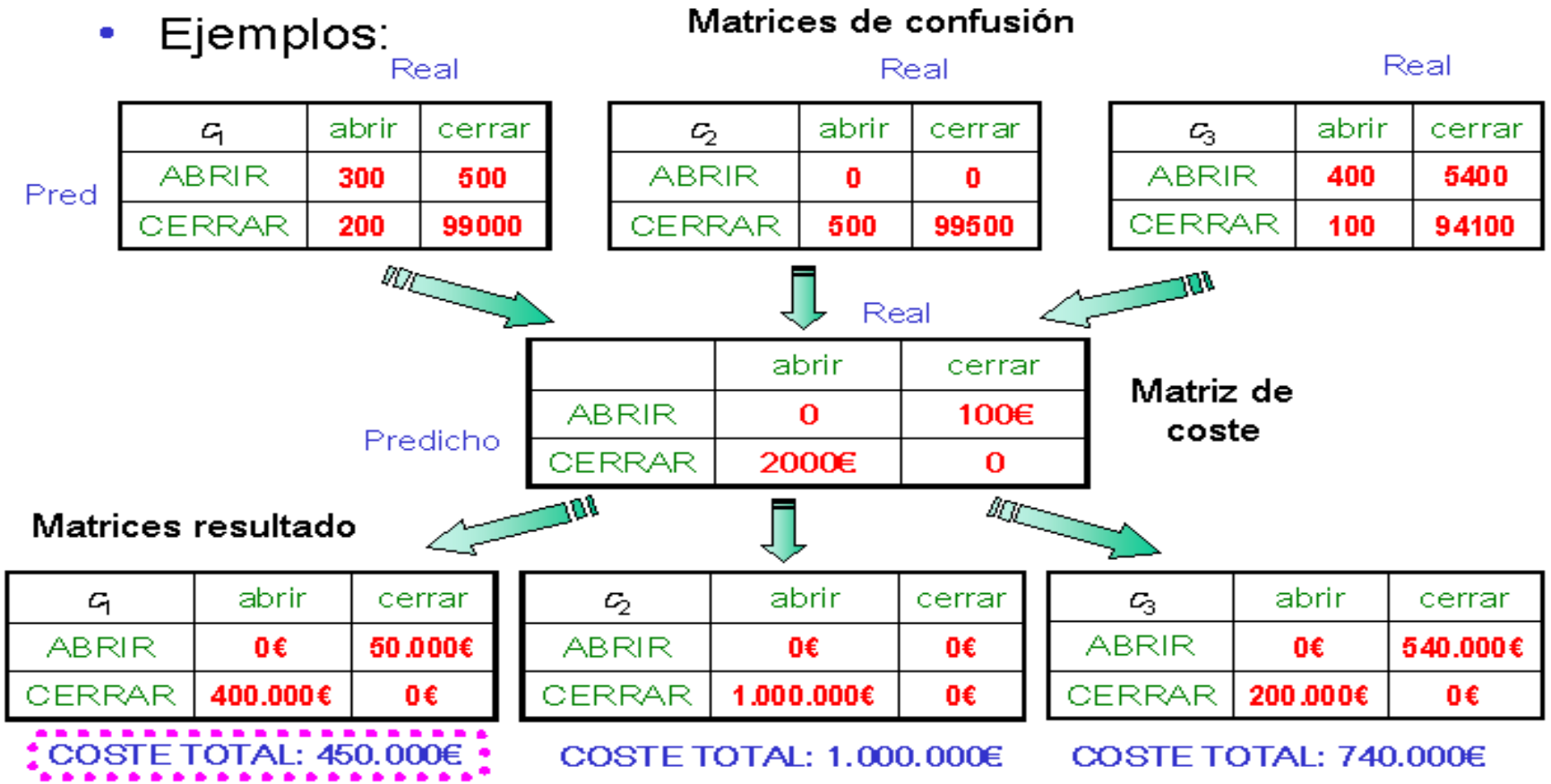
- En muchas situaciones los dos tipos de error que puede cometer un clasificador no tienen las mismas consecuencias
  - Dejar cerrada una válvula en una central nuclear, cuando es necesario abrirla, puede provocar una explosión, mientras que abrir una válvula cuando puede mantenerse cerrada, puede provocar una parada de la central
- Matriz de costes

| | | $C$ Clase real: abrir | $C$ Clase real: cerrar |
|---|---|---|---|
| $C_M$ Clase predicha | ABRIR | 0 | 100 € |
| | CERRAR | 2000 € | 0 |

- Lo importante no es obtener un clasificador que falle lo menos posible, sino uno que tenga coste menor

# Estimación basada en coste

- Ejemplos:



# Análisis ROC

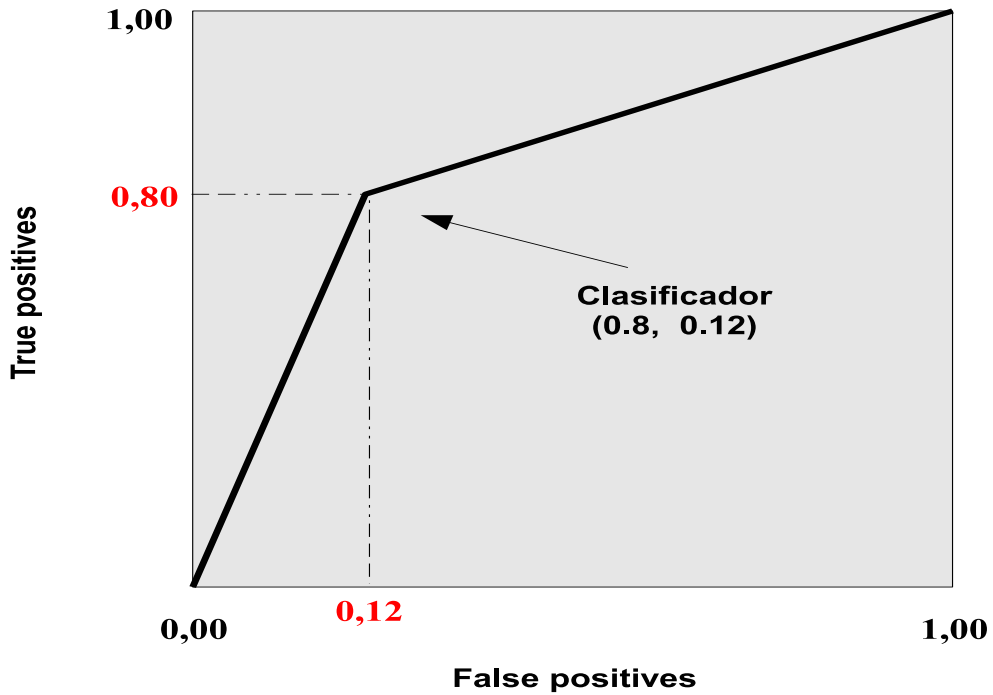
- En muchas situaciones es difícil estimar la matriz de costes
- Aprender un conjunto de clasificadores y seleccionar el que mejor se comporte para unas cicunstancias o contextos de coste determinados a posteriori.
- El análisis ROC provee herramientas que permiten seleccionar el subconjunto de clasificadores que tienen un comportamiento óptimo en general.
- El análisis ROC se utiliza normalmente para problemas de dos clases.

# Análisis ROC

| | | Real | |
|---|---|---|---|
| | | Abrir | Cerrar |
| **Pred.** | Abrir | **400** | **12000** |
| | Cerrar | **100** | **87500** |

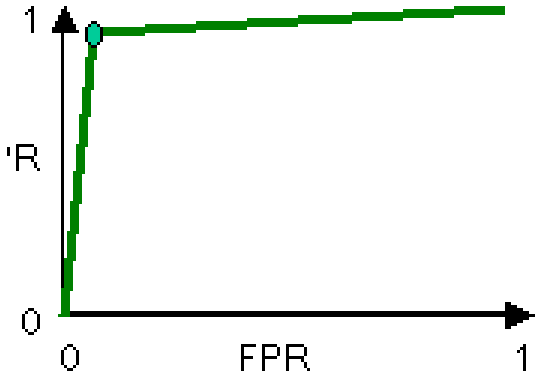
Normalización

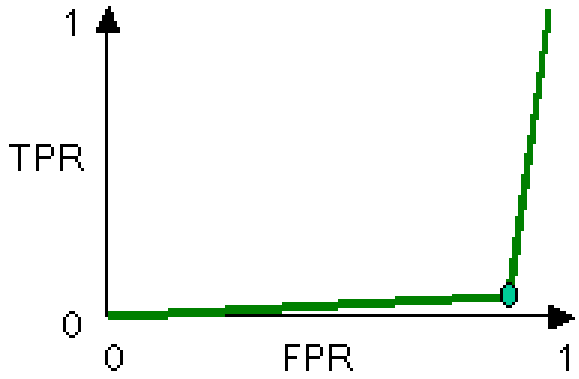
| | | Real | |
|---|---|---|---|
| | | Abrir | Cerrar |
| **Pred.** | Abrir | **0.8** | **0.12** |
| | Cerrar | **0.2** | **0.879** |

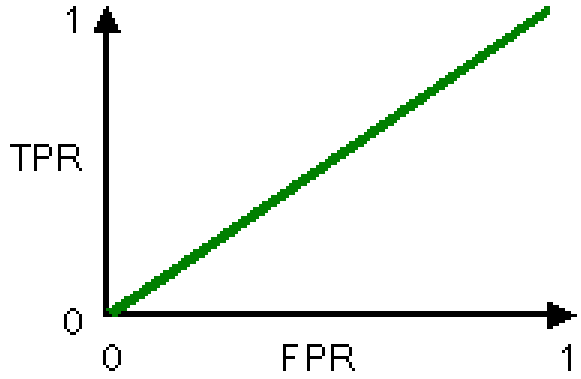


# La curva ROC

- Espacio ROC: buenos y malos clasificadores.



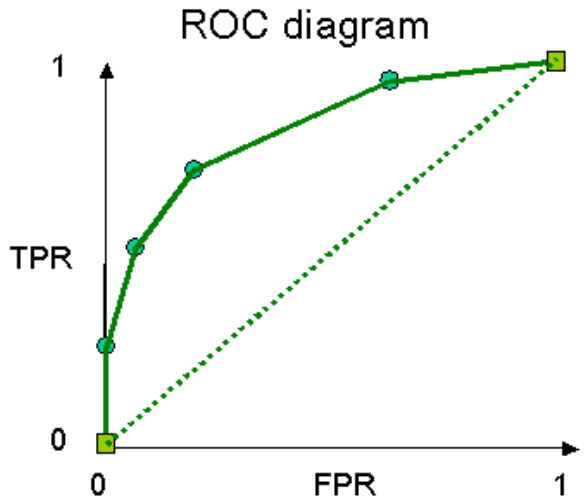




- Buen clasificador.
  - Alto TPR.
  - Bajo FPR.
- Mal clasificador.
  - Bajo TPR.
  - Alto FPR.
- Mal clasificador (en realidad).

# Análisis ROC

- *Convex hull* (casco convexo) a partir de la poligonal uniendo varios puntos (FPR, TPR)
- Dichos puntos pueden provenir de varios clasificadores o de un mismo clasificador (variando el umbral)



# Análisis ROC

Seleccionando el mejor clasificador

- Si cada punto de la curva ROC representa un clasificador: escoger el que tenga mayor valor de: $\frac{FPcost}{FNcost} \cdot \frac{Neg}{Pos}$
- Si cada punto de la curva ROC corresponde a un umbral con el que se toma la decisión: seleccionar el clasificador con mayor área bajo la curva (AUC)