Tema 5: Evaluación de Modelos de Clasificación Supervisada

Pedro Larrañaga, Iñaki Inza, Abdelmalik Moujahid

Departamento de Ciencias de la Computación e Inteligencia Artificial

Universidad del País Vasco

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

- Introducción
- Estimación de la probabilidad de clasificación correcta
- Brier score
- La curva ROC

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

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

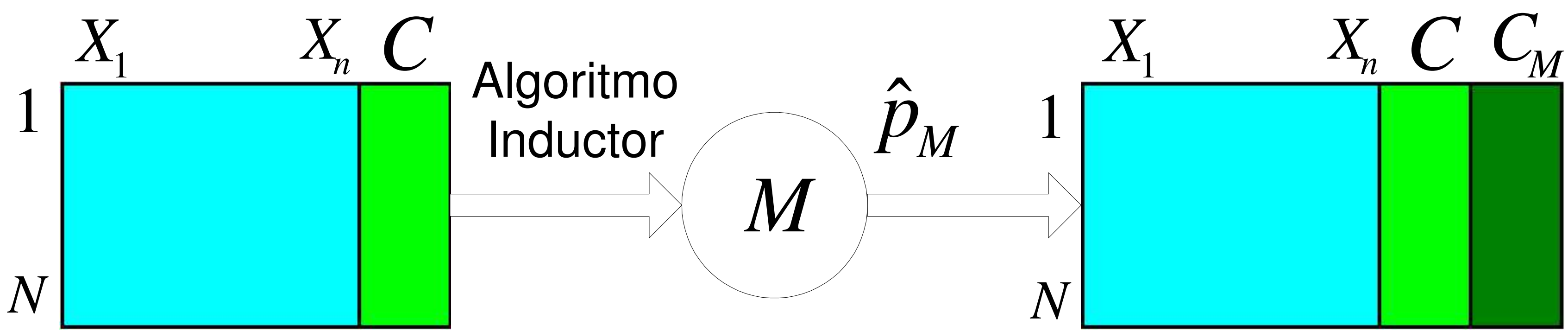
$$\delta(c^{(i)}, c_M^{(i)}) = \begin{cases} 1 & \text{si } c^{(i)} = c_M^{(i)} \\ 0 & \text{si } c^{(i)} \neq c_M^{(i)} \end{cases}$$

Introducción

		C Clase real	
		+	-
C_M Clase predicha	+	a	b
	-	c	d

- Tasa de acierto: $\frac{a+d}{a+b+c+d}$
- Tasa de error: $\frac{c+b}{a+b+c+d}$
- Proporción de verdaderos positivos (sensibilidad): $\frac{a}{a+c}$
- Proporción de verdaderos negativos (especifidad): $\frac{d}{b+d}$
- Proporción de falsos positivos: $\frac{b}{a+c}$
- Proporción de falsos negativos: $\frac{c}{b+d}$

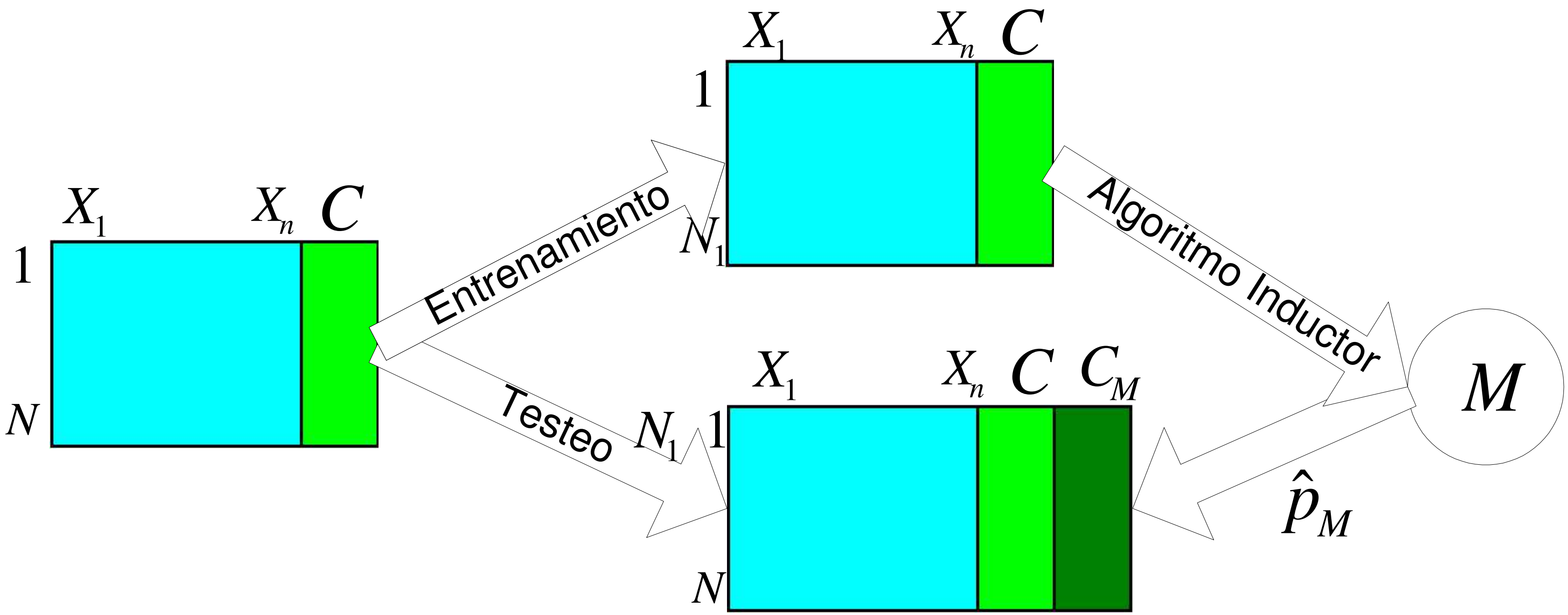
Estimación de la probabilidad de clasificación correcta



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

Método no honesto de estimación

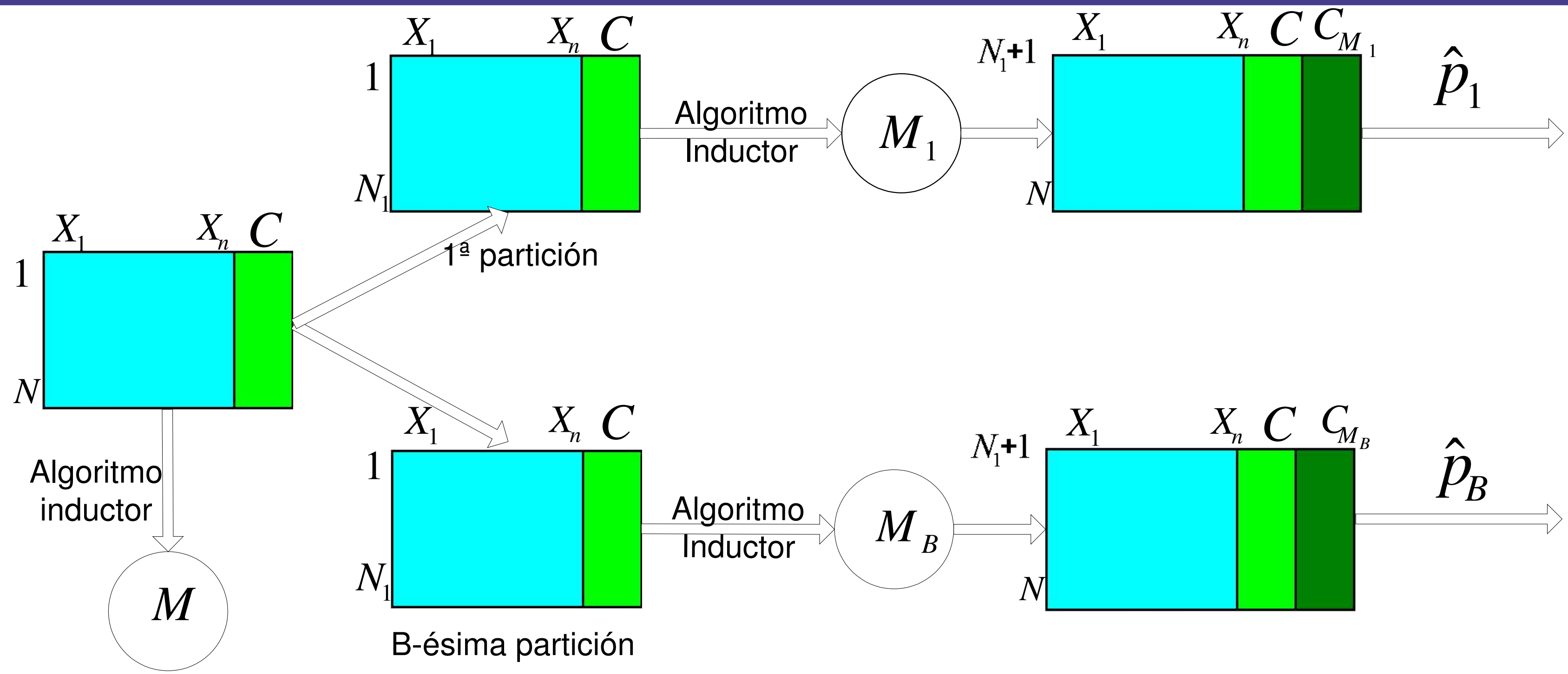
Estimación de la probabilidad de clasificación correcta



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

Método H de estimación basado en entrenamiento y testeo

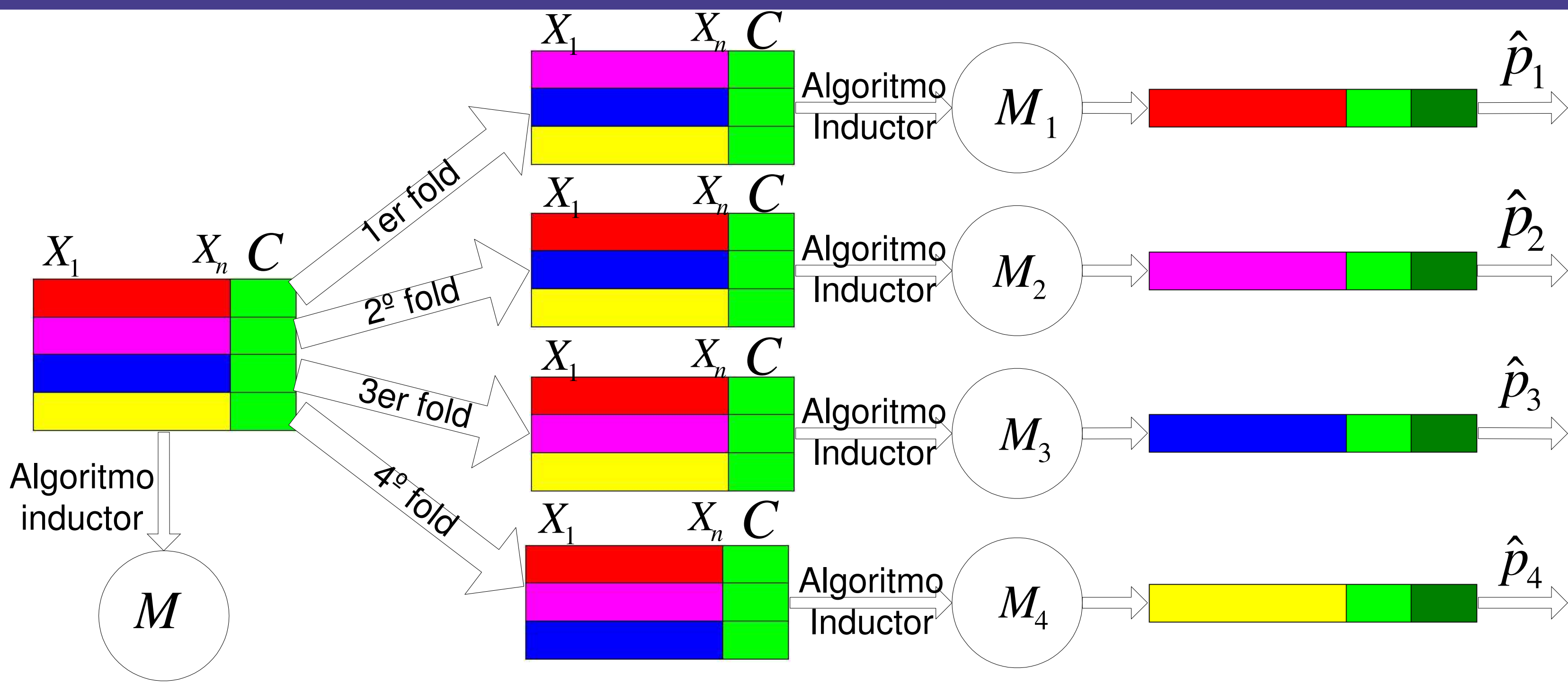
Estimación de la probabilidad de clasificación correcta



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

Método de estimación H repetidas veces

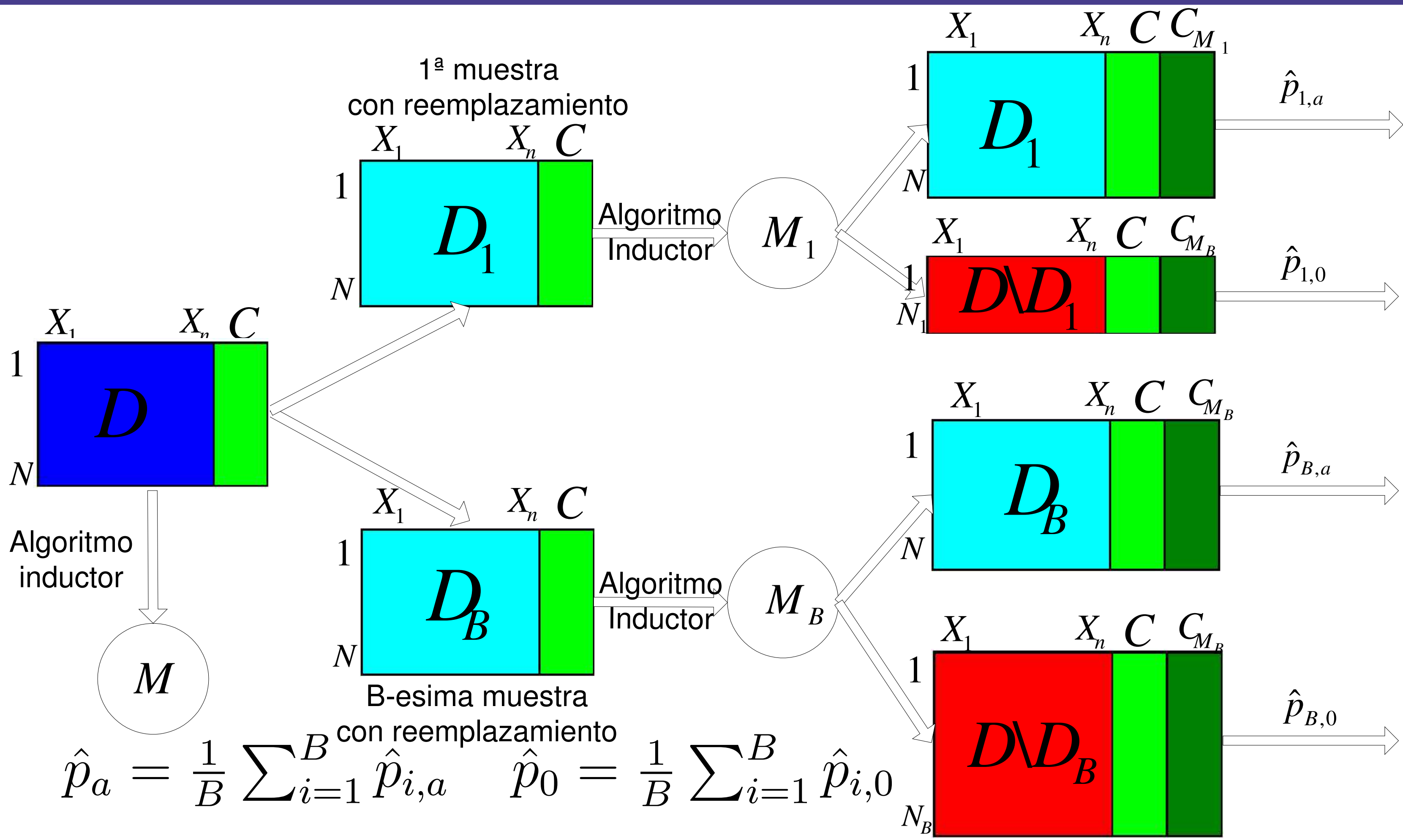
Estimación de la probabilidad de clasificación correcta



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

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

Estimación de la probabilidad de clasificación correcta



$$\hat{p}_a = \frac{1}{B}\sum_{i=1}^{B}\hat{p}_{i,a} \qquad \hat{p}_0 = \frac{1}{B}\sum_{i=1}^{B}\hat{p}_{i,0}$$

$$\hat{p}_M = \hat{p}_{0{,}632Bo} = (0{,}368\hat{p}_a + 0{,}632\hat{p}_0)$$

Método de estimación $0{,}632$ *booststraping*

Estimación de la probabilidad de clasificación correcta

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

Brier score

	X_1	$\dots$	X_n	C	$p(C_M = 0\|\boldsymbol{x})$	$p(C_M = 1\|\boldsymbol{x}$
$(\boldsymbol{x}^{(1)}, c^{(1)})$	$x_1^{(1)}$	$\dots$	$x_n^{(1)}$	1	0, 18	0, 82
$(\boldsymbol{x}^{(2)}, c^{(2)})$	$x_1^{(2)}$	$\dots$	$x_n^{(2)}$	0	0, 51	0, 49
$\dots$		$\dots$		$\dots$	$\dots$	
$(\boldsymbol{x}^{(N)}, c^{(N)})$	$x_1^{(N)}$	$\dots$	$x_n^{(N)}$	1	0, 55	0,45

$$B = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{2}[p(C_M = c|\boldsymbol{x}^{(i)}) - \delta(c^{(i)}, c_M^{(i)})]^2$$

$$B = \tfrac{1}{N}[(0,18-0)^2 + (0,82-1)^2 + (0{,}51-1)^2 + (0{,}49-0)^2 + \\ \ldots + (0,55-0)^2 + (0{,}45-1)^2]$$

Brier score

- Medida de la *calibración* para un clasificador que asigne, para cada patrón, probabilidades a posteriori a cada valor de la clase
- Suponiendo que la clase real del patrón $\boldsymbol{x}$ es 0, se trata de distinguir:

$$p(C_M = 0|\boldsymbol{x}) = 0{,}51 \quad \text{y} \quad p(C_M = 0|\boldsymbol{x}) = 0{,}97$$

- Interesa clasificadores con bajo valor de Brier (bastante seguros en sus predicciones)
- Para problemas con 2 clases: $0 \leq B \leq 2$

La curva ROC

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	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

La curva ROC

- Ejemplos:

Matrices de confusión

Pred / Real

c_1	abrir	cerrar
ABRIR	300	500
CERRAR	200	99000

c_2	abrir	cerrar
ABRIR	0	0
CERRAR	500	99500

c_3	abrir	cerra
ABRIR	400	5400
CERRAR	100	9410

Matriz de coste

Predicho / Real

	abrir	cerrar
ABRIR	0	100€
CERRAR	2000€	0

Matrices resultado

c_1	abrir	cerrar
ABRIR	0€	50.000€
CERRAR	400.000€	0€

COSTE TOTAL: 450.000€

c_2	abrir	cerrar
ABRIR	0€	0€
CERRAR	1.000.000€	0€

COSTE TOTAL: 1.000.000€

c_3	abrir	cerrar
ABRIR	0€	540.000
CERRAR	200.000€	0€

COSTE TOTAL: 740.000€

La curva ROC

- En muchas situaciones es difícil estimar la matriz de costes
- Análisis *ROC (Receiver Operating Characteristic)*
  - Usado por vez primera para evaluar radares en la segunda guerra mundial
  - Posteriormente se usó para el análisis de respuesta de transistores
  - A partir de 1970 se usa para aplicaciones de diagnóstico médico
  - A finales de los 90 se comienza a usar en minería de datos

La curva ROC

- El espacio ROC
  - Se normaliza la matriz de confusión por columnas: TPR, FNR TNR, FPR.

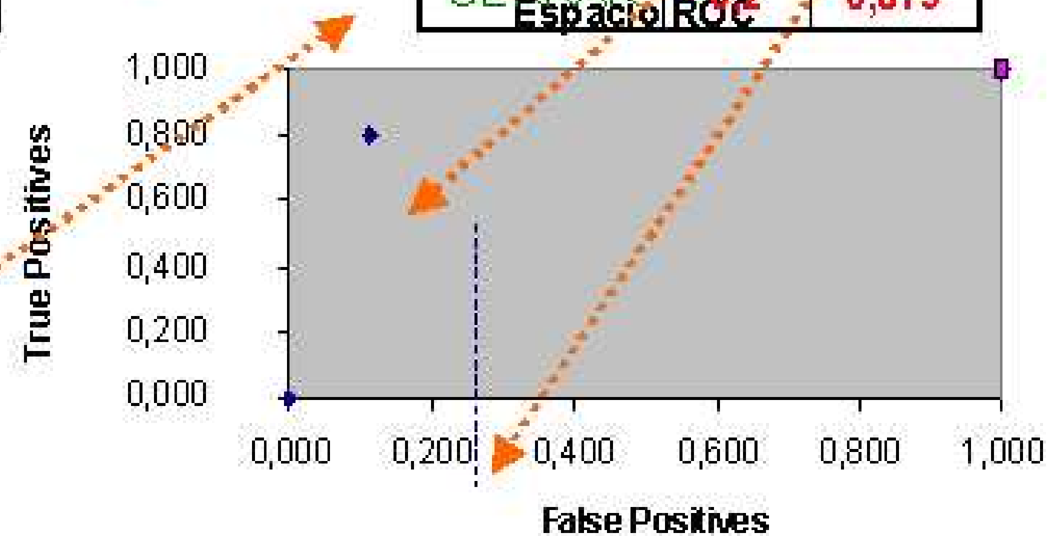
Real

Pred		abrir	cerrar
	ABRIR	400	12000
	CERRAR	100	87500

Real

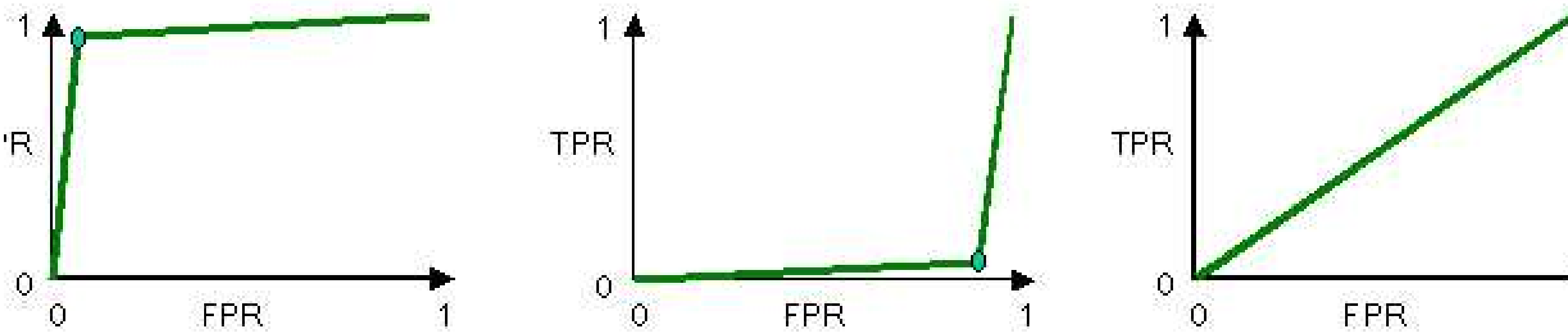
Pred		abrir	cerrar
	ABRIR	0,8	0,121
	CERRAR	0,2	0,879

TPR= 400 / 500 = 80%
FNR= 100 / 500 = 20%
TNR= 87500 / 99500 = 87,9%
FPR= 12000 / 99500 = 12,1%



La curva ROC

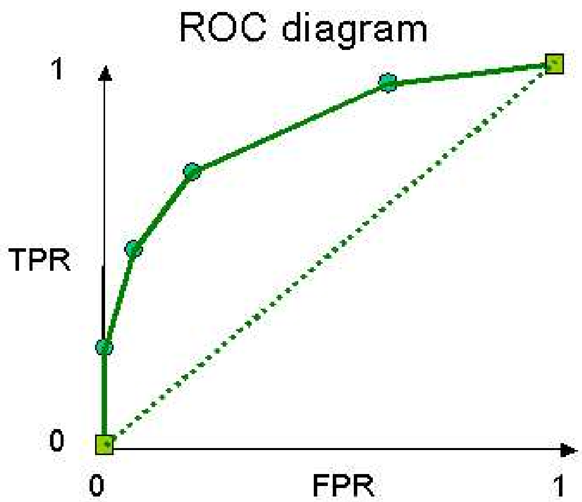
- Espacio ROC: buenos y malos clasificadores.



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

La curva 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)



La curva 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)