APRENDIZAJE AUTOMÁTICO: TERCER CURSO DEL GRADO DE ING. INFORMÁTICA EN COMPUTACION

Aprendizaje supervisado: Introducción y evaluación del rendimiento de un clasificador

César Hervás-Martínez
Grupo de Investigación AYRNA

Departamento de Informática y Análisis Numérico
Universidad de Córdoba
Campus de Rabanales. Edificio Einstein.
Email: chervas@uco.es

2019-2020

Clasificación automática y Reconocimiento de patrones

Sistema de decisión automatica donde se reconocen los patrones de unos datos muestrales mediante modelos aprendidos (estimados) utilizando técnicas de computación inteligente.

Mediante esta metodología un algoritmo extrae información de un conjunto de datos etiquetados de forma tal que el modelo aprendido sea capaz de predecir las etiquetas asociadas a un conjunto de nuevos datos no utilizados en el aprendizaje

Estas técnicas han atraído la atención de la sociedad de la información y de la industria.

Aplicaciones en motores de búsqueda, diagnósticos médicos, detección de fraude, clasificación de secuencias de ADN, reconocimiento de caracteres, etc.

UN PROBLEMA DE APRENDIZAJE

Reconocimiento de caracteres (handwritten digit recognition)

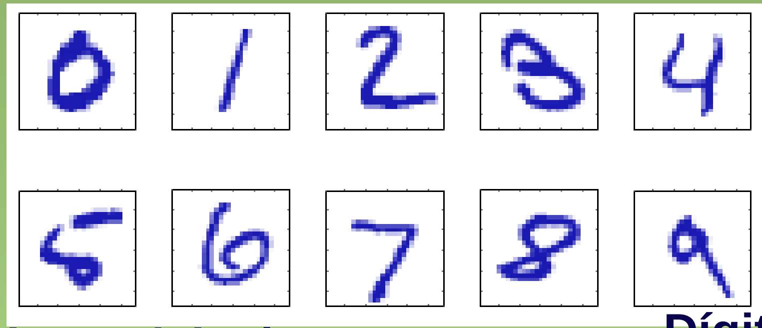
- **¿En qué consiste un problema de aprendizaje?**
- **Conceptos básicos:**

1. **Fase de entrenamiento (conjunto de entrenamiento)**
2. **Fase de test o generalización (conjunto de test)**
3. **Función de error**

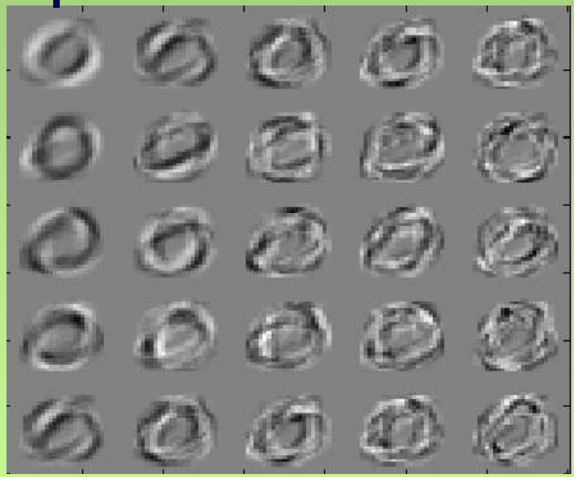
OCR

Reconocimiento de caracteres escritos a mano

UN PROBLEMA DE APRENDIZAJE: Handwritten digits recognition



Muestra de los autovalores del número O generado por el clasificador PCA



Dígitos muestrales utilizados para entrenar al clasificador

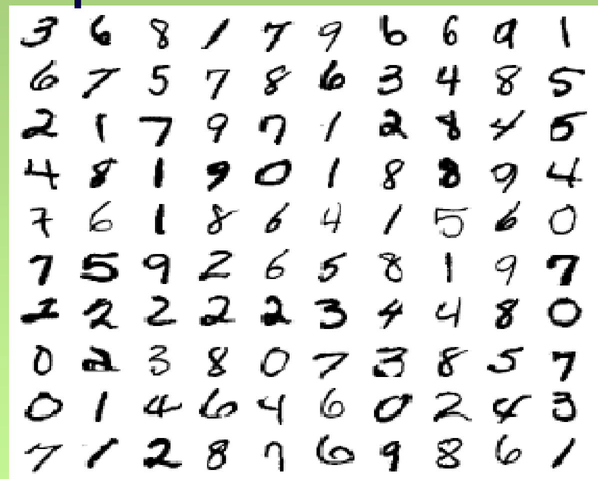


Imagen digital → Pre procesamiento → Extracción de características → Clasificación 1NN → Reconocimiento

❑ Formulación matemática del problema:

- **Imágenes de 28x28 pixeles**

- **Representar cada imagen como un vector** $\mathbf{x} \in \mathbb{R}^{784}$

- **Clasificador** $f(\mathbf{x})$

$$f : \mathbf{x} \rightarrow \{0,1,2,3,4,5,6,7,8,9\}$$

UN PROBLEMA DE APRENDIZAJE

- Extraemos una m.a.s o Conjunto de Entrenamiento:

$$D = \{(\mathbf{x}_i, y_i)\} \quad \text{para i=1,...,n}$$

- Siendo y_i el valor de la clase a la que pertenece el vector de pixeles $\mathbf{x}_i$ correspondiente al patrón i-ésimo

$$x_i = (1,1,0,0,....,1,0,...,0,0,1)$$

$$y_i \rightarrow \{0,1,2,3,4,5,6,7,8,9\}$$

UN PROBLEMA DE APRENDIZAJE

- **Funciones de error en el entrenamiento de un clasificador:**

  Métodos de estimación de la probabilidad de clasificación correcta: "Porcentaje de patrones mal clasificados"

  "Minima sensibilidad"

  "RMSE"

  La curva ROC y el área bajo la curva ROC "AUC"

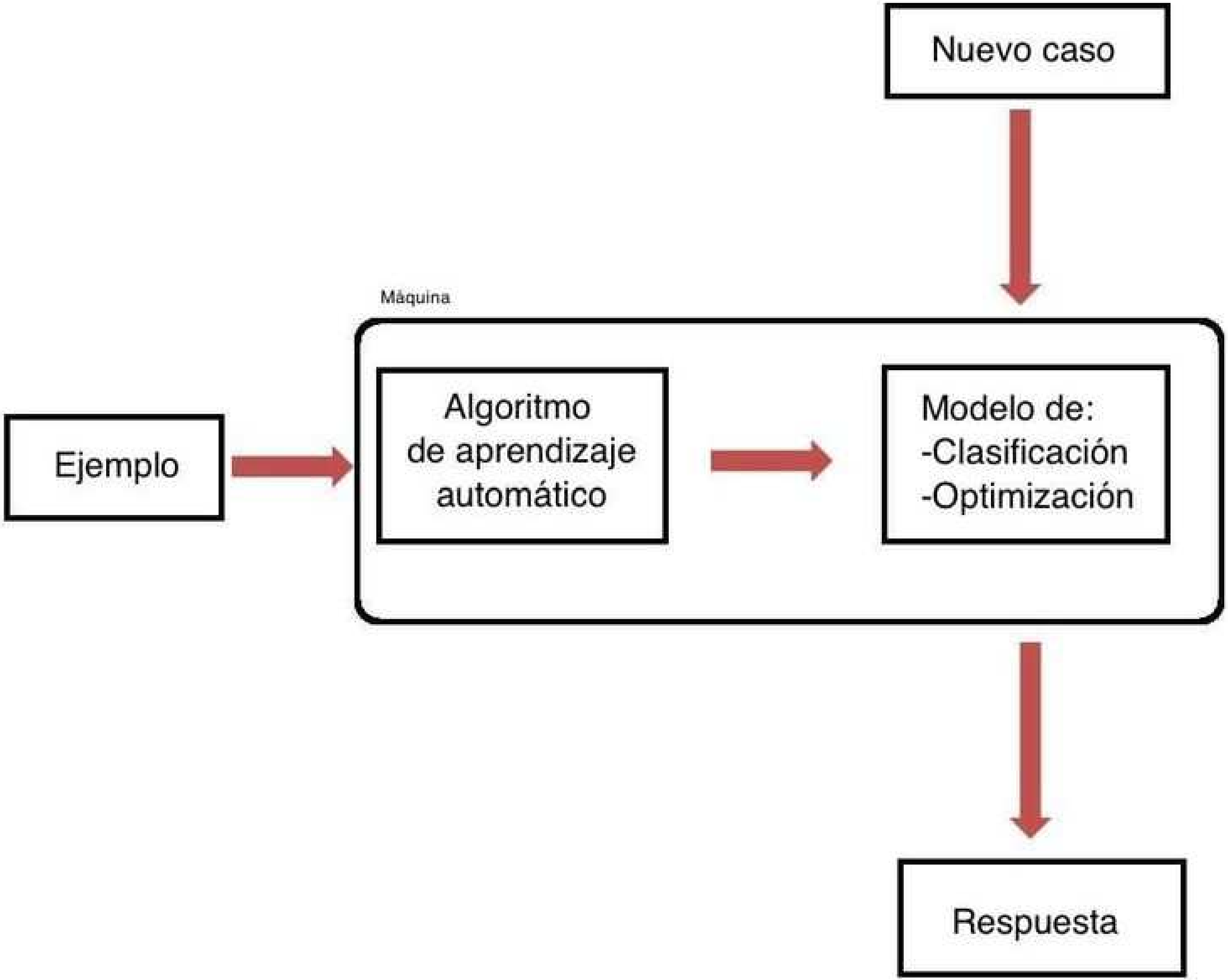
  Puntuación de Brier

- **Evaluación sensible al coste**

Objetivo:

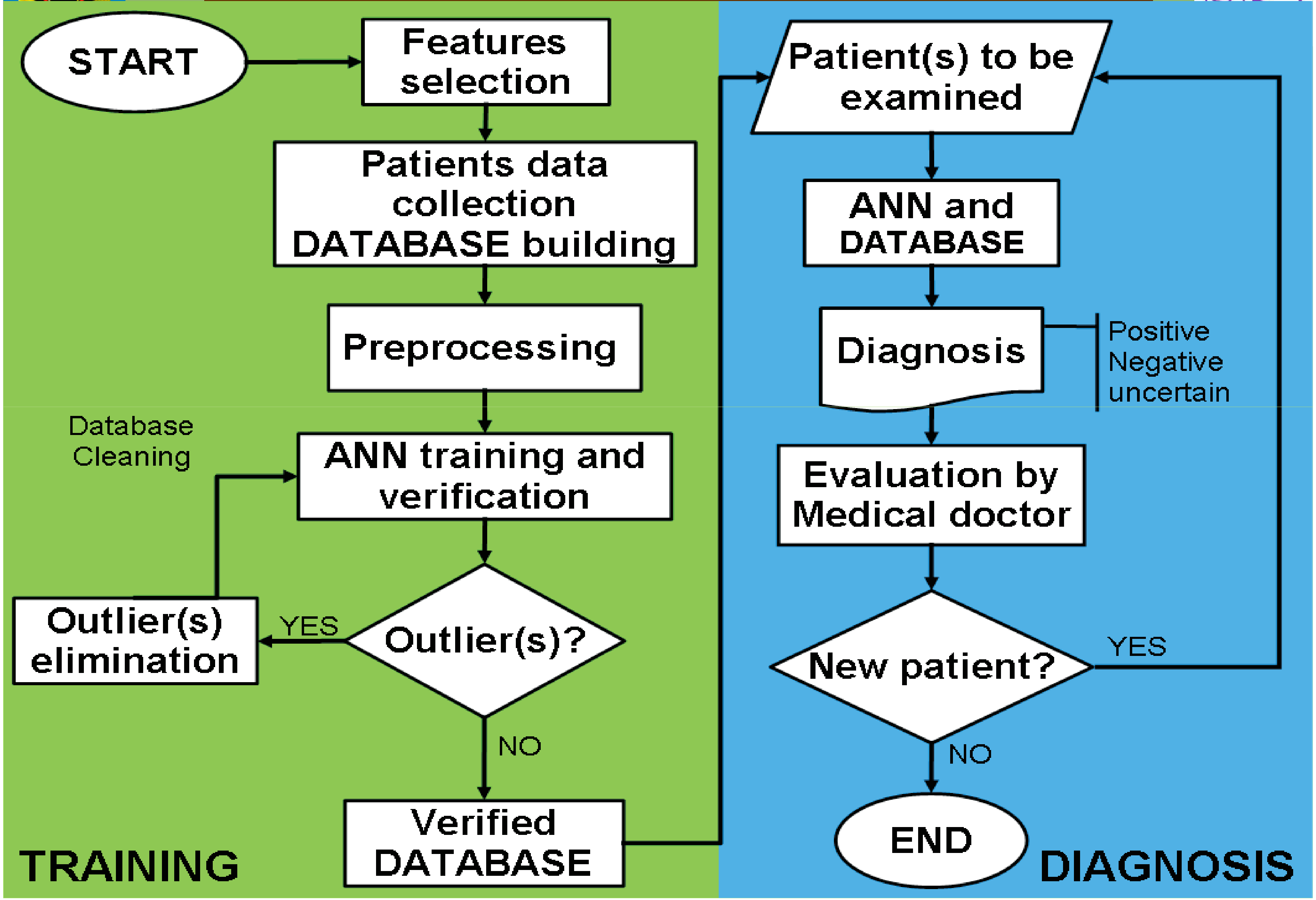
"Minimizar la función de error en el conjunto de test"

¿QUE ES EL APRENDIZAJE DE MÁQUINAS?



¿QUE ES EL APRENDIZAJE DE MÁQUINAS?

¿QUE ES EL APRENDIZAJE DE MÁQUINAS?



APRENDIZAJE SUPERVISADO

- **Entrenamiento**

Funciones (modelo)

$$D = \{(\mathbf{x_i}, y_i) \in X \times Y\}$$

$$F = \{ f : X \to Y \}$$

- **Aprendizaje**

$$\hat{f} \in F$$

$$tal\ que\ \ y_i \approx \hat{f}(\mathbf{x_i})$$

- **Predicción**

Nuevos datos

$$\hat{y} = \hat{f}(\mathbf{x})$$

$$\mathbf{x}$$

APRENDIZAJE SUPERVISADO Y NO-SUPERVISADO

Ejemplos canónicos de aprendizaje:

1. **Regresión** **(aprendizaje supervisado), la variable dependiente es continua**

2. **Clasificación** **(aprendizaje supervisado), la variable dependiente es nominal**

3. **Clasificación ordinal** **(aprendizaje supervisado) la variable dependiente es ordinal**

4. **Agrupamiento "Clustering"** **(aprendizaje no supervisado) no se conoce el valor de la variable dependiente**

APRENDIZAJE SUPERVISADO Rendimiento de un clasificador

Consideremos un problema de clasificación con J clases y n patrones de entrenamiento o test

El rendimiento de un clasificador g se puede obtener a partir de su matríz de contingencia o confusión definida en la forma

C/C*	C^*_1	...	C^*_j	...	C^*_J	
C_1	n_{11}	...	...	...	...	$n_{1.}$
...						
C_i			n_{ij}			$n_{i.} = \sum_{i=1}^{J} n_{ij}$
...						
C_J	n_{k1}				n_{kl}	$n_{k.}$
	$n_{.1}$	.....	$n_{.j} = \sum_{j=1}^{J} n_{ij}$	...	n_l	$n = \sum_{i=1}^{J}\sum_{j=1}^{J} n_{ij}$

$$M(g) = \left(n_{ij} \,/ \sum_{i=1}^{J}\sum_{j=1}^{J} n_{ij} = n \right)$$

Donde n_{ij} representa el número de veces que los patrones son predichos por el clasificador g en la clase C^*_j cuando realmente pertenecen a la clase C_i

Coste de una mala clasificación

No podemos comparar numéricamente las clases Ci y Cj pero podemos asignar un coste artificial cuando confundimos, con el clasificador, la clase Ci por la clase Cj

Los costes se pueden definir mediante una matriz de costes Cij, donde este valor indica el coste de clasificar una instancia o patrón de la clase Ci como perteneciente a la clase Cj

Algunas elecciones de matrices de costes y de transformación de variables

$$C_{01} = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

Coste 0,1

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 1 \\ -1 & 0 & 1 & 1 & 1 \\ -1 & -1 & 0 & 1 & 1 \\ -1 & -1 & 1 & 0 & 1 \\ -1 & -1 & -1 & -1 & 0 \end{bmatrix}$$

Matriz de transformación

$$C_{abs} = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 2 & 3 \\ 2 & 1 & 0 & 1 & 2 \\ 3 & 2 & 1 & 0 & 1 \\ 4 & 3 & 2 & 1 & 0 \end{bmatrix}$$

Coste absoluto

Medidas asociadas, CCR, MS

Medida asociada, τ_b

Medidas asociadas, RMSE, MAE, AMAE, MMAE, mMAE r_S

Clasificación binaria:

- Supongamos que tenemos un conjunto de N observaciones (conjunto de entrenamiento)

$$(\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N})$$

$$(y_1, y_2, \ldots, y_N)$$

$$\mathbf{x_i} \in \mathbb{R}^d,\ y_i \in \{-1, 1\}$$

- Problema de clasificación: estimar $f(\mathbf{x})$ tal que

$$f(\mathbf{x_i}) = y_i$$

CLASIFICACIÓN

- **Función de error: Porcentaje de patrones mal clasificados**

$$\frac{1}{N}\sum_{i=1}^{N}\left[y_i \neq f(\mathbf{x_i})\right]$$

- **Otras denominaciones: Función de perdida o Función de riesgo.**
- **Función de evaluación: Porcentaje de patrones correctamente clasificados**

$$C = \frac{1}{N}\sum_{i=1}^{N}\left[y_i = f(\mathbf{x_i})\right]$$

METRICAS EN CLASIFICACIÓN BINARIA

- **Matriz de confusión (caso binario):**

Clase real de pertenencia

Clase asignada o predicha

$$\begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix}$$

Eficacia, o tasa de aciertos,

$$C = \frac{TP+TN}{N}$$

TP= Verdadero Positivo, FP= Falso Positivo,

FN= Falso Negativo , TN= Verdadero Negativo

MEDIDAS DE MERITO EN CLASIFICACIÓN BINARIA

Clase asignada o predicha

Clase real de pertenencia

$$\begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Especificidad} = \frac{TN}{TN+FP}$$

$$\text{Recall} = \text{Sensibilidad} = \frac{TP}{TP+FN}$$

$$F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Kappa} = \frac{TP+TN-E(TP+TN)}{TP+TN+FP+FN-E(TP+TN)}$$

❑ Ejemplo de matriz de confusión (caso binario):

Clase asignada o predicha

Clase real de pertenencia

$$\begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix} = \begin{pmatrix} 20 & 10 \\ 15 & 15 \end{pmatrix}$$

Sensibilidad Especificidad

$$\text{Porcentaje de aciertos}, C = \frac{35}{60}$$

$$\text{Especificidad} = \frac{TN}{TN+FP}$$

$$\text{Recall} = \text{Sensibilidad} = \frac{TP}{TP+FN}$$

$$\text{Especificidad} \equiv TNR = \frac{15}{15+15} = 0.5$$

$$\text{Sensibilidad} \equiv TPR = \frac{20}{20+10} = \frac{2}{3}$$

METRICAS EN CLASIFICACIÓN BINARIA

- **Problemas no-balanceados**
- **Ejemplo: enfermos de cáncer (10) e individuos sanos (991):**

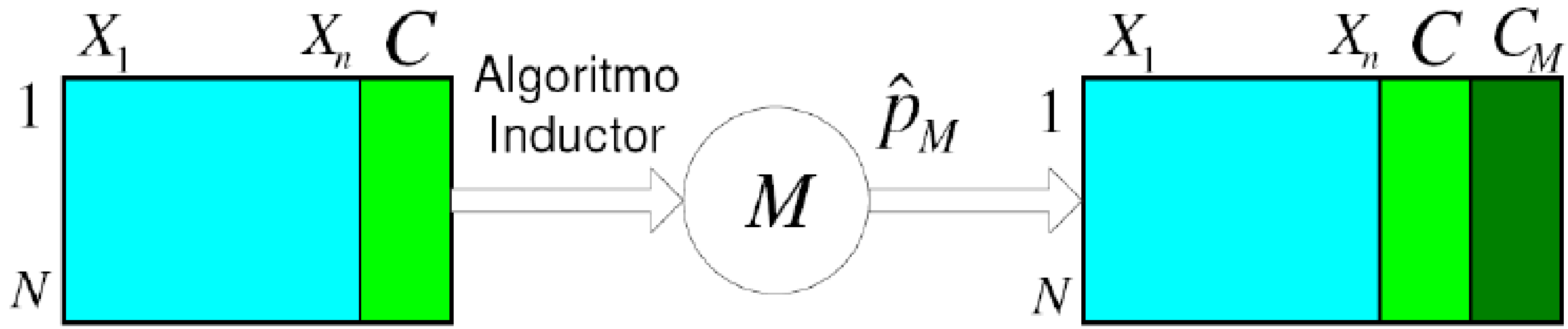
Clase real

$$\begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix} = \begin{pmatrix} 0 & 10 \\ 1 & 990 \end{pmatrix}$$

$$C = \frac{990}{1001} \qquad Sensi = 0,\ Espec = \frac{990}{991}$$

$$\text{Recall=Sensibilidad=}\frac{\text{TP}}{\text{TP+FN}} \qquad \text{Especificidad=}\frac{\text{TN}}{\text{TN+FP}}$$

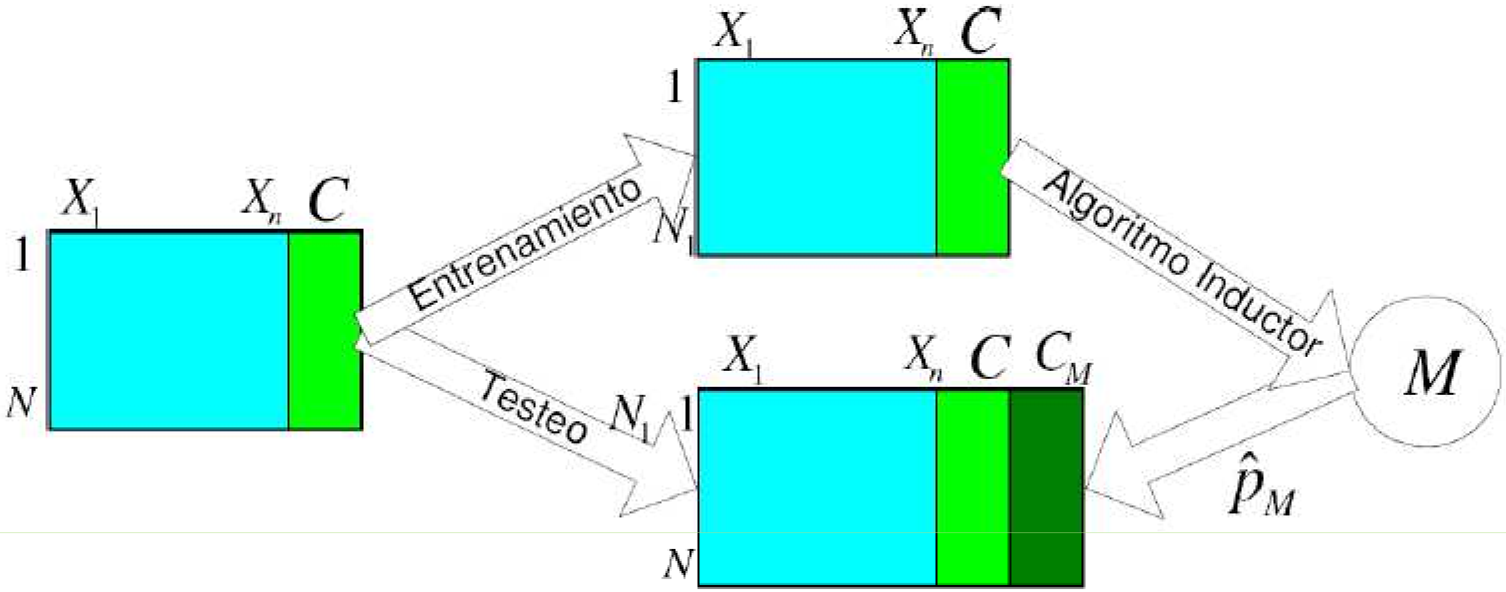
CLASIFICACIÓN DESHONESTA



Estimación de la Precisión $$\hat{p}_M = \frac{1}{N}\sum_{i=1}^{N}\delta(C_i = C_{i,M})$$

C= Clase real, C_M= Clase estimada por el modelo

CLASIFICACIÓN MEDIANTE HOLDOUT 75-25

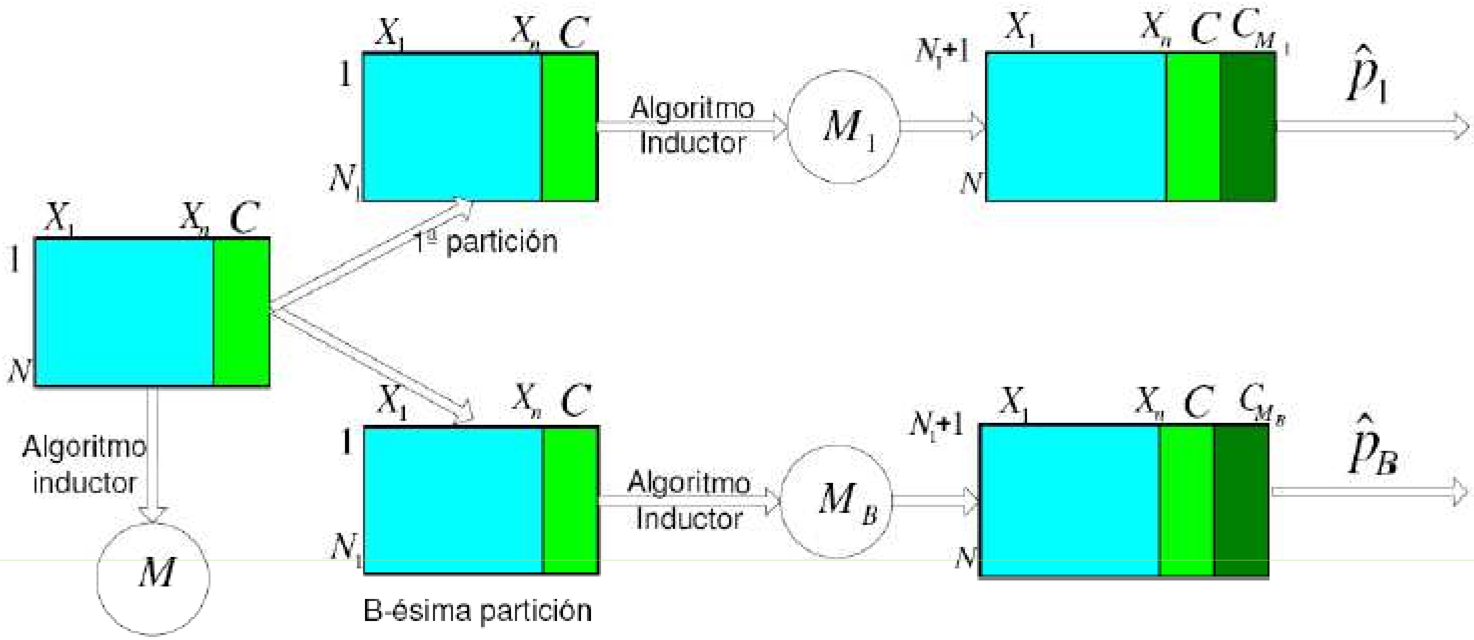


Estimación de la Precisión

$$\hat{p}_M = \frac{1}{N - N_1} \sum_{i=1}^{N-N_1} \delta(C_{N_1+i} = C_{N_1+i,M})$$

$$N_1 \simeq 0,75 \times N$$

CLASIFICACIÓN MEDIANTE UN HOLDOUT REPETIDO, B veces



En realidad es un *ensemble* (modelo de modelos) promedio de los resultados de los modelos $M_1, M_2, \ldots, M_B$

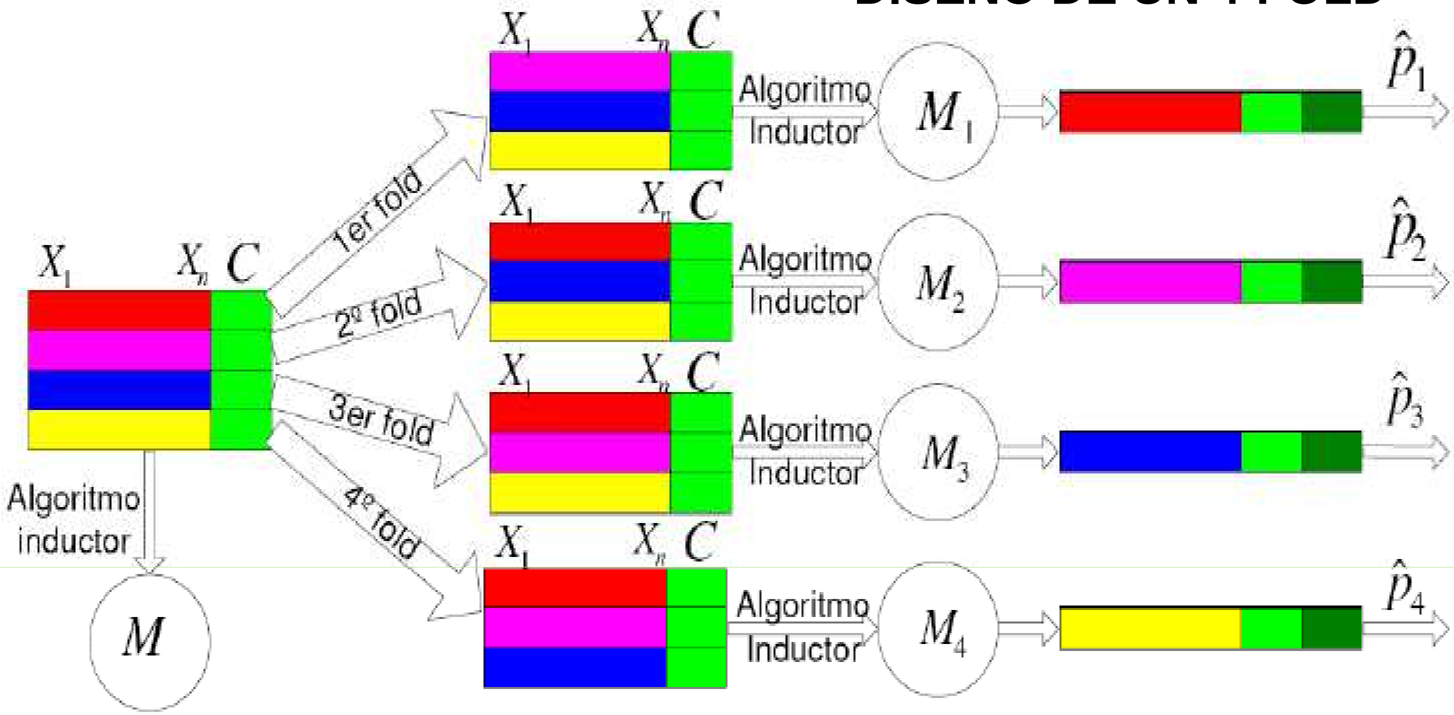
$$\hat{p}_M = \frac{1}{B}\sum_{i=1}^{B}\hat{p}_i$$

$$E(\hat{p}_M) = \frac{1}{B}\sum_{i=1}^{B}E(\hat{p}_i); \quad V(\hat{p}_M) = \frac{1}{B^2}\left(\sum_{i=1}^{B}V(\hat{p}_i) + 2\sum_{i=1<j}^{B}Cov(\hat{p}_i, \hat{p}_j)\right)$$

Luego cuanto más dependientes sean los estimadores $\hat{p}_i$ **más varianza tendrá** $\hat{p}_M$

CLASIFICACIÓN MEDIANTE UN k-FOLD

DISEÑO DE UN 4-FOLD



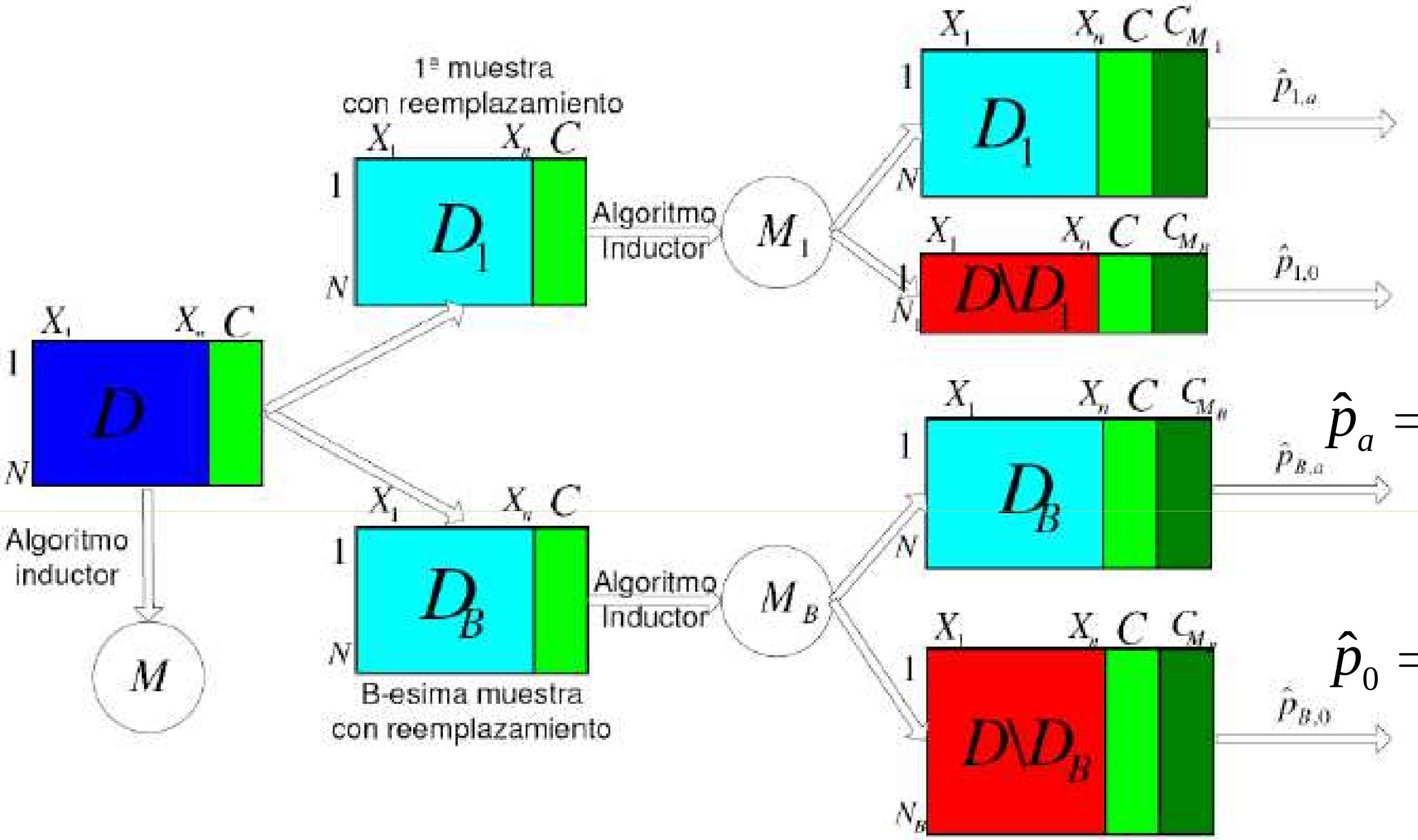
$$\hat{p}_M = \frac{1}{k}\sum_{i=1}^{k}\hat{p}_i$$

En este caso k=4

$$E(\hat{p}_M) = \frac{1}{k}\sum_{i=1}^{k}E(\hat{p}_i); \quad V(\hat{p}_M) = \frac{1}{k^2}\left(\sum_{i=1}^{k}V(\hat{p}_i) + 2\sum_{i=1<j}^{k}Cov(\hat{p}_i, \hat{p}_j)\right)$$

Luego cuanto más dependientes sean los estimadores $\hat{p}_i$
más varianza tendrá $\hat{p}_M$

METODO DE ESTIMACION 0,632 BOOTSTRAPPING



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

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

$$\hat{p}_M = \hat{p}_{0,632B_0} = (0,368\,\hat{p}_a + 0,632\,\hat{p}_0)$$

Es por tanto un *ensemble* de *ensembles*

Consejos de uso de los distintos métodos

Método Holdout: utilizarlo con N grande.

Método Holdout repetidas veces: no hay control sobre los casos usados como entrenamiento/testeo.

Método de estimación basado en (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.

PUNTUACIÓN DE BRIER PARA CLASIFICACIÓN BINARIA

Salida probabilística del modelo

	X_1	$\dots$	X_n	C	$p(C_M = 0 \mid \boldsymbol{x})$	$p(C_M = 1 \mid \boldsymbol{x})$
$(\boldsymbol{x}^{(1)}, c^{(1)})$	$x_1^{(1)}$	$\dots$	$x_n^{(1)}$	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

Salida real

La puntuación de Brier es

$$B = \frac{1}{N} \sum_{i=1}^{N} \sum_{C=0}^{1} \left[p(C_M = c \mid x_i) - \delta(C_i, C_{m,i}) \right]^2$$

Y en este ejemplo concreto

$$B = \frac{1}{N} \left[(0{,}18 - 0)^2 + (0{,}82 - 1)^2 + (0{,}51 - 0)^2 + (0{,}49 - 1)^2 + \dots \right.$$
$$\left. + (0{,}55 - 0)^2 + (0{,}45 - 1)^2 \right]$$

CLASIFICACIÓN

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 x es 0, se trata de distinguir entre:

$$p(C_M = 0 \mid \mathbf{x}) = 0{,}51 \;\; y \;\; p(C_M = 0 \mid \mathbf{x}) = 0{,}97$$

Interesa clasificadores con bajo valor de Brier (bastante seguros en sus predicciones, dado que las probabilidades a posteriori son altas)

Para problemas con 2 clases: $0 \leq B \leq 2$

CLASIFICADOR CON MENOR COSTE

- Ejemplos:

Matrices de confusión

Real / Pred

c_1	abrir	cerrar
ABRIR	300	500
CERRAR	200	99000

Real

c_2	abrir	cerrar
ABRIR	0	0
CERRAR	500	99500

Real

c_3	abrir	cerrar
ABRIR	400	5400
CERRAR	100	94100

Real / Predicho

	abrir	cerrar
ABRIR	0	100€
CERRAR	2000€	0

Matriz de coste

Matrices resultado

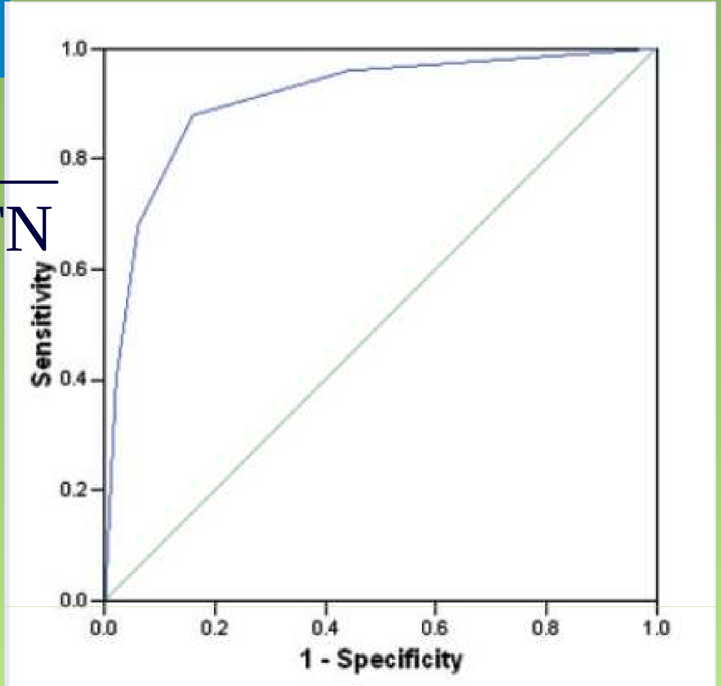
c_1	abrir	cerrar
ABRIR	0€	50.000€
CERRAR	400.000€	0€

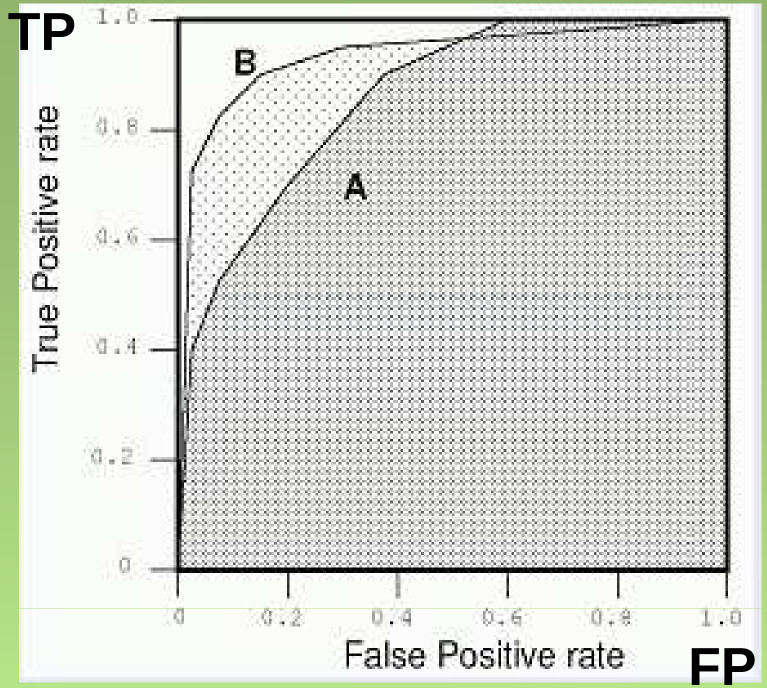
c_2	abrir	cerrar
ABRIR	0€	0€
CERRAR	1.000.000€	0€

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

AREA BAJO LA CURVA ROC

$$\frac{TP}{TP+FN}$$





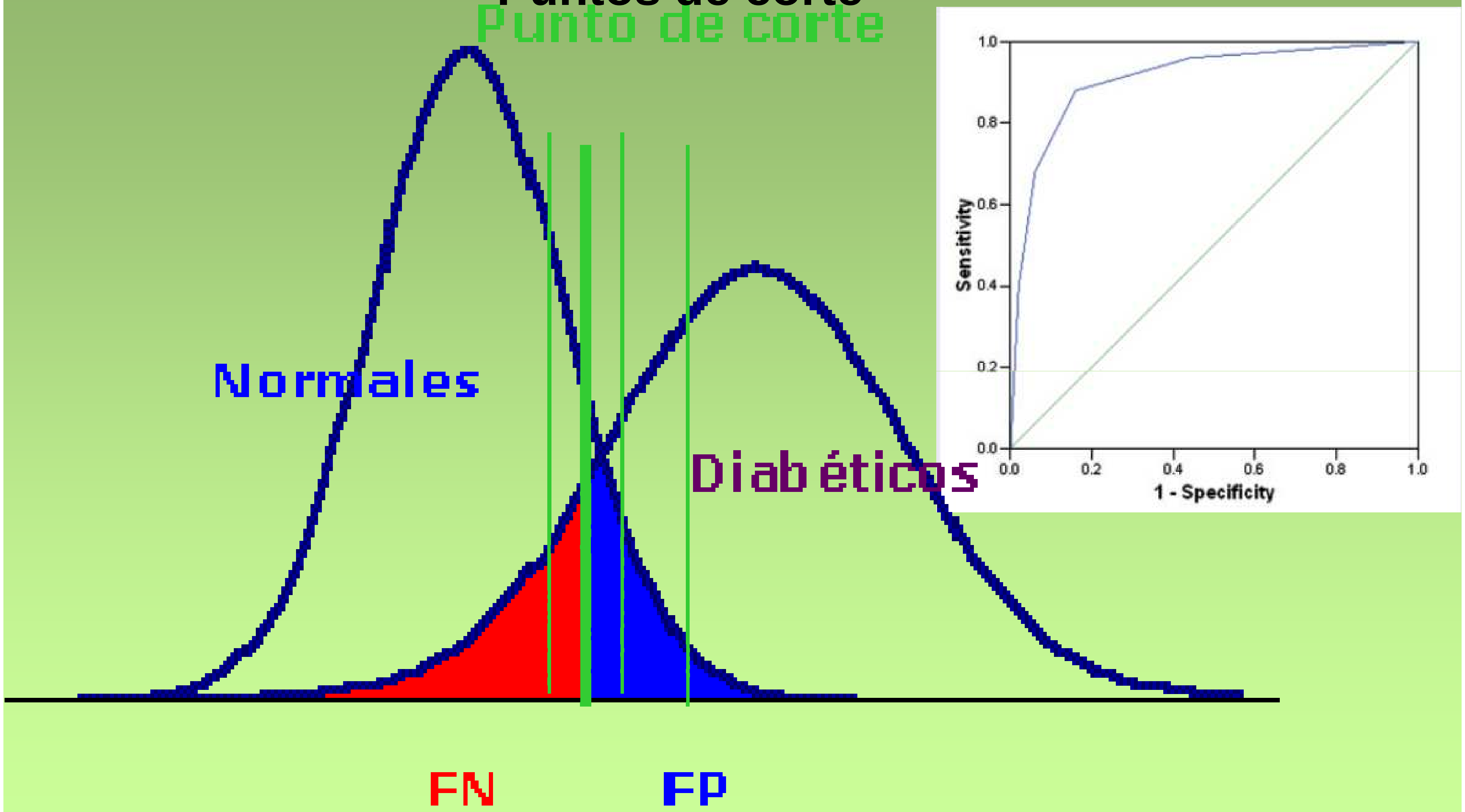
$$1-\frac{TN}{TN+FP}=\frac{FP}{TN+FP}$$

Si cada punto de la curva ROC representa un clasificador: los puntos de la curva cuanto mas a la izquierda y mas arriba mejor será el clasificador

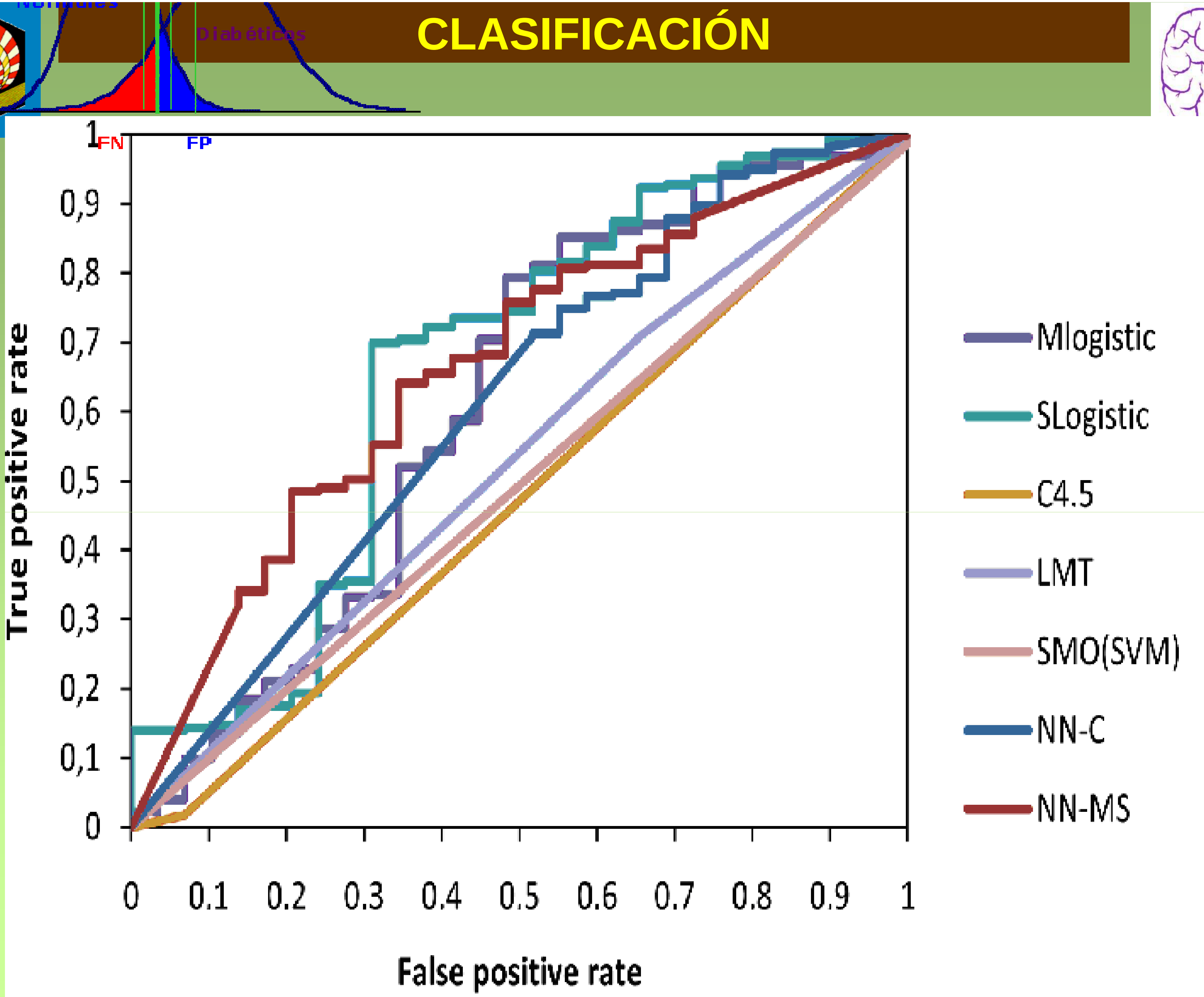
Si cada punto de la curva ROC corresponde a un umbral con el que se toma la decisión: seleccionar el clasificador con mayor área bajo la curva ROC (AUC)

AREA BAJO LA CURVA ROC

Puntos de corte



CLASIFICACIÓN



Clasificación multiclase: Ejemplo de base de datos de clasificación nominal

La base de datos iris es bien conocida en la investigación en estadística y recientemente en aprendizaje puesto que se utiliza para ilustrar la eficacia de algoritmos de clasificación nominal

iris setosa

iris versicolor

iris virginica

La base de datos contiene medidas de las dimensiones de 50 muestras de flores de cada especie.

Fisher, R.A. (1936). "The Use of Multiple Measurements in Taxonomic Problems". *Annals of Eugenics* **7**: 179–188, disponible en: http://digital.library.adelaide.edu.au/coll/special//fisher/138.pdf.

Clasificación MULTICLASE: Ejemplo de base de datos de clasificación nominal

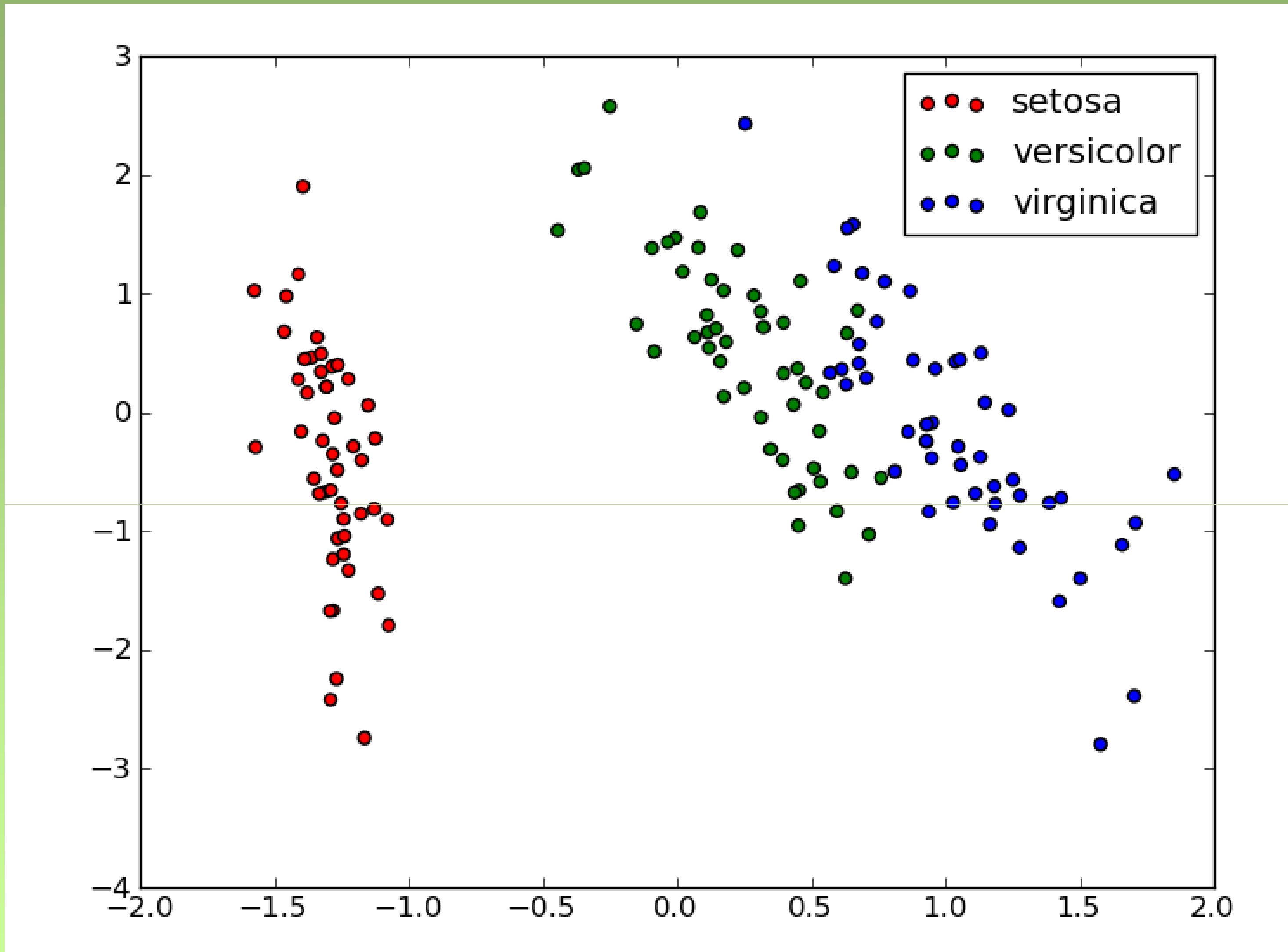
Anderson medió las siguientes dimensiones:

- **Longitud de los sépalos**
- **Ancho de los sépalos**
- **Longitud de los pétalos**
- **Ancho de los pétalos**

- **Las cuatro dimensiones de la flor iris se denominan *atributos o variables de entrada o independientes*.**
- **Las tres especies *(Q=3)* de la flor iris se denominan *clases o variable de salida o dependiente*. Cada ejemplo de flor iris se denomina una *muestra*, un *patrón o una instancia*.**

Las clases se etiquetan en un formato “1 de Q”, de esta forma tendremos:
(1,0,0) para la clase setosa
(0,1,0) para la clase virgínica
(0,0,1) para la clase versicolor

Clasificación MULTICLASE: Ejemplo de base de datos de clasificación nominal



Ejemplo de proyección en dos dimensiones de las tres clases

Métricas de rendimiento

Existen multitud de métricas de rendimiento de un clasificador (multi-clase y binario): CCR, TPR, FPR, Especificidad, AUC, MSE, Entropía, RMSE, etc.

Problema: Un mismo valor de una medida puede representar clasificadores muy distintos, en especial en problemas no balanceados y/o con gran número de clases.

Un buen clasificador debería obtener un alto nivel de precisión global, así como un aceptable nivel de precisión para cada clase.

Medida bidimensional en MOEAS: Precisión-Mínima Sensibilidad.

Métricas de rendimiento

Matriz de confusión de un clasificador y obtención de las medidas MS y C.

$$M = \begin{array}{cc} \begin{pmatrix} n_{11} & n_{12} & \cdots & n_{1n} \\ n_{21} & n_{22} & \cdots & n_{2Q} \\ \vdots & \vdots & \vdots & \vdots \\ n_{Q1} & n_{Q2} & \cdots & n_{QQ} \end{pmatrix} & \begin{matrix} f_1 \\ f_2 \\ \vdots \\ f_Q \end{matrix} \end{array}$$

Sensibilidad de la clase i

$$S_i = n_{ii} / f_i$$

Minima sensibilidad (MS)

$$MS = mín\{S_i ; i = 1, \ldots, Q\}$$

Precisión (C)

$$C = (1/N)\sum_{j=1}^{Q} n_{jj}$$

Métodos de evaluación del rendimiento de un clasificador

GRACIAS POR SU ATENCIÓN

César Hervás-Martínez
Grupo de Investigación AYRNA

Departamento de Informática y Análisis Numérico
Universidad de Córdoba
Campus de Rabanales. Edificio Einstein.
Email: chervas@uco.es

2019-2020