



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Introducción a la estimación del error en Reconocimiento de Formas

Alfons Juan
Albert Sanchis
Jorge Civera

DSIC

Departamento de Sistemas
Informáticos y Computación

Objetivos formativos

- Calcular el error teórico de un clasificador
- Calcular el error de Bayes
- Estimar el error de un clasificador por resubstitución
- Estimar el error de un clasificador por *holdout* y añadir un intervalo de confianza al 95 %

Índice

1	Error teórico de un clasificador	3
2	Error del clasificador de Bayes	4
3	Estimación del error por resubstitución	5
4	Estimación del error por <i>holdout</i>	6
5	Conclusiones	8

1. Error teórico de un clasificador

El **error** (esperado) de un clasificador $c(x)$, para todo $x \in E$, es:

$$\varepsilon = E(\varepsilon(c(x))) = \begin{cases} \sum_x P(x) \varepsilon(c(x)) & \text{si } E \text{ es discreto} \\ \int p(x) \varepsilon(c(x)) dx & \text{si } E \text{ es continuo} \end{cases}$$

donde $\varepsilon(c(x))$ es la probabilidad de error de $c(x)$ para x :

$$\varepsilon(c(x)) = 1 - P(c = c(x) \mid x)$$

Ejemplo (problema y clasif.): $E = [0, 1]^2$, $C = 2$, $\eta_c(\mathbf{x}) \triangleq P(c \mid \mathbf{x})$

x_1	x_2	$\eta_1(\mathbf{x})$	$\eta_2(\mathbf{x})$	$P(\mathbf{x})$	$c(\mathbf{x})$	$\varepsilon(c(\mathbf{x}))$
0	0	1	0	1/2	1	0
0	1	3/4	1/4	1/4	1	1/4
1	0	1/4	3/4	1/4	1	3/4
1	1	0	1	0	2	0

$$\Rightarrow \varepsilon = \frac{1}{4} \cdot \frac{1}{4} + \frac{1}{4} \cdot \frac{3}{4} = \frac{1}{4}$$

2. Error del clasificador de Bayes

El **clasificador de Bayes** elige una clase de máxima probabilidad a posteriori:

$$c^*(x) = \arg \max_c P(c \mid x)$$

Su probabilidad de error para un x cualquiera es mínima:

$$\varepsilon(c^*(x)) = 1 - P(c^*(x) \mid x) = 1 - \max_c P(c \mid x)$$

por lo cual también lo es su error, el **error de Bayes**:

$$\varepsilon^* = E(\varepsilon(c^*(x))) = \begin{cases} \sum_x P(x) \varepsilon(c^*(x)) & \text{si } E \text{ es discreto} \\ \int p(x) \varepsilon(c^*(x)) dx & \text{si } E \text{ es continuo} \end{cases}$$

Ejemplo: $\varepsilon^* = \frac{1}{8}$ (para el problema ejemplo)

3. Estimación del error por resubstitución

Sea $c_N(x)$ un clasificador aprendido con un conjunto de N muestras, $S_N = \{(x_1, c_1), (x_2, c_2), \dots, (x_N, c_N)\}$, y sea ε_N su error.

Denominamos estimador por **resubstitución** de ε_N a:

$$\hat{\varepsilon}_N^r = \frac{1}{N} \sum_{n=1}^N [c_N(x_n) \neq c_n] = \frac{\text{número de errores}}{N}$$

Es **optimista**, sobre todo con clasificadores complejos y $N \ll$

Ejemplo: $N = 4$ (para el problema ejemplo)

x_1	x_2	$\eta_1(\mathbf{x})$	$\eta_2(\mathbf{x})$	$P(\mathbf{x})$	$N_1(\mathbf{x})$	$N_2(\mathbf{x})$	$c_N(\mathbf{x})$
0	0	1	0	1/2	2	0	1
0	1	3/4	1/4	1/4	1	0	1
1	0	1/4	3/4	1/4	1	0	1
1	1	0	1	0	0	0	—

$$\Rightarrow \hat{\varepsilon}_N^r = \frac{0}{4}$$

4. Estimación del error por *holdout*

Sea $S_M = \{(x_1, c_1), (x_2, c_2), \dots, (x_M, c_M)\}$ un **conjunto de test** de muestras M independientes de las N de entrenamiento.

Denominamos estimador **holdout** de ε_N a:

$$\hat{\varepsilon}_{N,M} = \frac{1}{M} \sum_{m=1}^M [c_N(x_m) \neq c_m] = \frac{\text{número de errores}}{M}$$

Aproxima bien ε_N cuando M es grande, pero **“desaprovecha” muestras**.

Ejemplo: $S_M = \{((0, 0)^t, 1), ((0, 1)^t, 1), ((1, 0)^t, 2)\} \rightarrow \hat{\varepsilon}_{N,M} = \frac{1}{3}$
(para el problema y clasificador ejemplo)

Intervalo de confianza al 95 %

Si $\text{Var}(\varepsilon_N)$ es despreciable y M es grande, podemos asumir que:

$$\hat{\varepsilon}_{N,M} \sim \mathcal{N} \left(E(\varepsilon_N), \frac{E(\varepsilon_N)(1 - E(\varepsilon_N))}{M} \right)$$

y podemos construir un **intervalo de confianza al 95 %** para ε_N ,

$$P(\varepsilon_N \in I) = 0.95 \quad \text{amb} \quad I = \left[\hat{\varepsilon}_{N,M} \pm 1.96 \sqrt{\frac{\hat{\varepsilon}_{N,M}(1 - \hat{\varepsilon}_{N,M})}{M}} \right]$$

Ejemplo: $M = 2000$, $\hat{\varepsilon}_{N,M} = 0.05$

$$I = \left[0.05 \pm 1.96 \sqrt{\frac{0.05 \cdot 0.95}{2000}} \right] = [0.05 \pm 0.01] = [4 \%, 6 \%]$$

5. Conclusiones

Hemos visto:

- La estimación del error teórico de un clasificador
- La estimación del error del clasificador de Bayes
- La estimación del error por resubstitución y holdout