Introducció a l'estimació de l'error en Reconeixement de Formes

Alfons Juan
Albert Sanchis
Jorge Civera

Departament de Sistemes
Informàtics i Computació

Objectius formatius

- Calcular l’error teòric d’un classificador
- Calcular l’error de Bayes
- Estimar l’error d’un classificador per resubstitució
- Estimar l’error d’un classificador per *holdout* i afegir-hi un interval de confiança al 95%

Índex

1 Error teòric d'un classificador

L'***error*** (esperat) d'un classificador $c(x)$, per a tot $x \in E$, és:

$$\varepsilon = \mathrm{E}(\varepsilon(c(x))) = \begin{cases} \sum_x P(x)\,\varepsilon(c(x)) & \text{si } E \text{ és discret} \\ \int p(x)\,\varepsilon(c(x))\,dx & \text{si } E \text{ és continu} \end{cases}$$

on $\varepsilon(c(x))$ és la probabilitat d'error de $c(x)$ per a x:

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

Exemple (problema i classif.): $E = [0,1]^2$, $C = 2$, $\eta_c(\boldsymbol{x}) \triangleq P(c \mid \boldsymbol{x})$

x_1	x_2	$\eta_1(\boldsymbol{x})$	$\eta_2(\boldsymbol{x})$	$P(\boldsymbol{x})$	$c(\boldsymbol{x})$	$\varepsilon(c(\boldsymbol{x}))$
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

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

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

2 Error del classificador de Bayes

El ***classificador de Bayes*** tria una classe de màxima probabilitat a posteriori:

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

La seua probabilitat d'error per a un x qualsevol és mínima:

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

per la qual cosa també ho és el seu error, l'***error de Bayes***:

$$\varepsilon^* = \mathrm{E}(\varepsilon(c^*(x))) = \begin{cases} \sum_x P(x)\,\varepsilon(c^*(x)) & \text{si } E \text{ és discret} \\ \int p(x)\,\varepsilon(c^*(x))\,dx & \text{si } E \text{ és continu} \end{cases}$$

Exemple: $\varepsilon^* = \frac{1}{8}$ (per al problema exemple)

3 Estimació de l'error per resubstitució

Siga $c_N(x)$ un classificador aprés amb un conjunt de N mostres, $S_N = \{(x_1, c_1), (x_2, c_2), \ldots, (x_N, c_N)\}$, i siga ε_N el seu error.

Anomenem estimador per ***resubstitució*** de ε_N a:

$$\hat{\varepsilon}_N^r = \frac{1}{N} \sum_{n=1}^{N} [\, c_N(x_n) \neq c_n \,] = \frac{\text{nombre d'errors}}{N}$$

És ***optimista***, sobretot amb classificadors complexos i N menut.

Exemple: $N = 4$ (per al problema exemple)

x_1	x_2	$\eta_1(\boldsymbol{x})$	$\eta_2(\boldsymbol{x})$	$P(\boldsymbol{x})$	$N_1(\boldsymbol{x})$	$N_2(\boldsymbol{x})$	$c_N(\boldsymbol{x})$
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 Estimació de l'error per *holdout*

Siga $S_M = \{(x_1, c_1), (x_2, c_2), \ldots, (x_M, c_M)\}$ un ***conjunt de test*** de M mostres independents de les N d'entrenament.

Anomenem estimador ***holdout*** de ε_N a:

$$\hat{\varepsilon}_{N,M} = \frac{1}{M} \sum_{m=1}^{M} [\, c_N(x_m) \neq c_m \,] = \frac{\text{nombre d'errors}}{M}$$

Aproxima bé ε_N quan M és gran, però ***"desaprofita" mostres***.

Exemple: $S_M = \{((0,0)^t, 1), ((0,1)^t, 1), ((1,0)^t, 2)\} \rightarrow \hat{\varepsilon}_{N,M} = \dfrac{1}{3}$
(per al problema i classificador exemple)

Interval de confiança al 95%

Si $\mathrm{Var}(\varepsilon_N)$ és menyspreable i M és gran, podem assumir que:

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

i podem construir un ***interval de confiança al 95%*** per a ε_N,

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

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

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