Clasificación de patrones: Evaluación

Jordi Porta Zamorano

Escuela Politécnica Superior **Dept. de Lingüística Computacional**

Universidad Autónoma de Madrid **Real Academia Española**

jordi.porta@uam.es **porta@rae.es**

abril de 2005

Evaluación: Nada es gratis

No Free Lunch Theorem: no hay razones independientes del contexto y de la aplicaciónn que justifi quen la superioridad de un tipo de clasifi cador sobre otro. Si un algoritmo parece superior a otro en determinadas circunstancias, es consecuencia de su ajuste particular al problema de reconocimiento de patrones, no a su superioridad general como algoritmo. Los aspectos más importantes son: la información a priori, la cantidad de patrones para el entrenamiento, las funciones de coste y recompensa.

Evaluación

- Para clasifi cación, la medida de bondad de un clasifi cador es la *tasa de error* medida como:

$$\text{tasa de error} = \frac{\text{núm. de casos mal clasifi cados}}{\text{núm. de casos totales}}$$

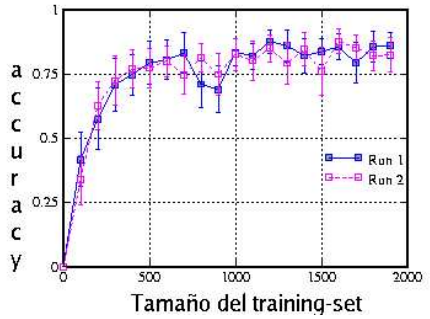
- El error medido sobre el *conjunto de entrenamiento* (*error de resustitución*) es una estimación demasiado optimista.
- El error medido sobre el conjunto de entrenamiento puede servir para depurar errores en los datos.
- El *conjunto de test* es un conjunto independiente de patrones que se usan para medir el error del clasifi cador.
- A veces se usa un *conjunto de validación* (*validation set*) para optimizar los parámetros del clasifi cador.
- No tiene la misma signifi cación un error de un $5\,\%$ estimado sobre 100 ejemplos que sobre 100000.
- La estimación fi nal de los parámetros de un clasifi cador se hace con **TODOS** los patrones disponibles.

Evaluación: La curva de aprendizaje

- podemos defi nir el *acierto* de clasifi cador como:

$$\text{accuracy} = 1 - \text{error} = \frac{\text{núm. de casos correctamente clasifi cados}}{\text{núm. de casos totales}}$$

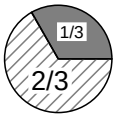
- podemos ver la relación entre el acierto y el número de ejemplos de entrenamiento en la *curva de aprendizaje*:



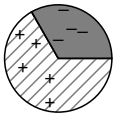
para cada tamaño considerado se ejecutan varios experimentos de los que se muestran la *media* y la *varianza*

Evaluación: Datos insuficientes

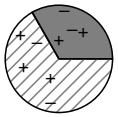
- *Hold-out* por muestreo aleatorio:
  - 1/3 para test y 2/3 para entrenamiento:



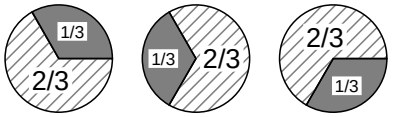
  - Problema: no representatividad de la muestra aleatoria



- *Hold-out* con muestreo estratifi cado:
  asegurar que la distribución de clases en los dos conjuntos es la misma



- *Validación cruzada* (*Cross-validation*):
  - se dividen los datos en un numero m de particiones (*folds*). P. ej. para $m = 3$



  - se entrena m veces y se promedia el error de cada test
  - Problema: no representatividad de las muestras
  - Solución: *validación cruzada estratificada*
- n *stratified* m*-fold cross-validation*:
  - el conjunto de N ejemplos se divide aleatoriamente en m conjuntos distintos de tamaño N/m
  - el clasifi cador se entrena m veces
  - todo el proceso de validación cruzada con m divisiones se repite n veces
    y se calcula el error medio
  - el estándar de facto es $n = m = 10$

- *Leave-one-out*:
  - con n patrones equivale a un n*-fold cross-validation*
  - se entrena n con todos menos con uno y se promedian los n errores
  - computacionalmente costoso cuando hay muchos patrones
  - se usa cuando hay pocos patrones
  - la muestra para test no es estratifi cada (¡sólo hay uno!)

- *Bootstrap*:
  - se usa cuando hay muy pocos patrones (n)
  - el cj. de entrenamiento se usa un muestreo de n elems. con reemplazo
    contendrá aprox. el $63{,}2\,\%$ de los patrones
  - el cj. de test lo forman los patrones que no están en el cj. de entrenamiento
    contendrá aprox. el $36{,}8\,\%$ de los patrones
  - error total $= 0{,}632 \times$ error en el cj. de test $+ 0{,}368 \times$ error en el cj. de entrenamiento
  - el proceso de *bootstrap* se repite varias veces y se promedia el error

Evaluación: matrices de confusión

- también llamadas *tablas de contingencia*
- permite analizar el error para problemas con más de dos clases
- los errores se disponen en la siguiente tabla:

	$\hat{+}$	$\hat{-}$
$+$	verdaderos $+$ (TP)	falsos $-$ (FN)
$-$	falsos $+$ (FP)	verdaderos $-$ (TN)

Evaluación: Medidas

- recuperación de la información:
  - *cobertura* (*recall*): están todos los que son
    $$\text{recall} = \frac{TP}{TP+FN}$$
  - *precision*: son todos los que estan
    $$\text{precision} = \frac{TP}{TP+FP}$$
  - *F-measure*:
    $$\text{F-measure} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$
- teoría de la comunicación:
  - *ruido*: están más de los que son
    $$\text{ruido} = \frac{FP}{TP+FP}$$
  - *silencio*: lo son pero no están
    $$\text{silencio} = \frac{FN}{TN+FN}$$

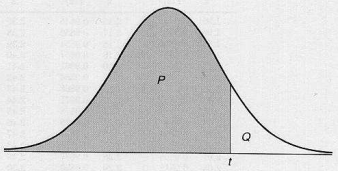
Evaluación: Visualización y otras medidas

- Hay otras curvas que permiten comparar métodos o representar la precisión de un clasifi cador cuyas decisiones dependen de un valor umbral:
  - curvas ROC
  - curvas *lift*
  - curva recall/precision
- en predicción numérica el error se calcula con otras medidas como el *error cuadrático medio*

Evaluación: Comparación de dos métodos

- se trata de saber si el error medio de un método es *significativamente* mayor o menor que el de otro método
- el error medio de cada método se obtiene mediante n*-fold cross-validation*
- es conveniente, pero no obligatorio, usar las mismas particiones para cada método
- se aplica el *test de Student* (t*-test*)
  - sean $x_1 \ldots x_n$ e $y_1 \ldots y_n$ los errores del primer y segundo métodos en las n particiones
  - sea $\bar{d} = \bar{x} - \bar{y}$ la diferencia de los errores medios
  - sea $\sigma_d^2 = \sigma_x^2 + \sigma_y^2$ la variancia de la diferencia d
  - se calcula el valor de $t = \bar{d}/\sqrt{\sigma_d^2/n}$
    - si $t \approx 0$ los métodos no son signifi cativamente diferentes
    - sino, hay que consultar una tabla como la siguiente:

3. A table of the t-distribution



ν \ P	0.75	0.90	0.95	0.975	0.99	0.995	0.999
1	1.00	3.08	6.31	12.71	31.82	63.66	318.3
2	0.82	1.89	2.92	4.30	6.96	9.92	22.33
3	0.76	1.64	2.53	3.18	4.54	5.84	10.22
4	0.74	1.53	2.13	2.77	3.75	4.60	7.17
5	0.73	1.48	2.02	2.57	3.36	4.03	5.89
6	0.72	1.44	1.94	2.45	3.14	3.71	5.21
7	0.71	1.42	1.90	2.36	3.00	3.50	4.78
8	0.71	1.40	1.86	2.31	2.90	3.36	4.50
9	0.70	1.38	1.83	2.26	2.82	3.25	4.30
10	0.70	1.37	1.81	2.23	2.76	3.17	4.14
12	0.70	1.36	1.78	2.18	2.68	3.06	3.93
15	0.69	1.34	1.75	2.13	2.60	2.95	3.73
20	0.69	1.33	1.73	2.09	2.53	2.84	3.55
24	0.68	1.32	1.71	2.06	2.49	2.80	3.47
30	0.68	1.31	1.70	2.04	2.46	2.75	3.38
40	0.68	1.30	1.68	2.02	2.42	2.70	3.31
60	0.68	1.30	1.67	2.00	2.39	2.66	3.23
120	0.68	1.29	1.66	1.98	2.36	2.62	3.37
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.09
ν \ Q	0.25	0.10	0.05	0.025	0.01	0.005	0.001

Source: Adapted from A. Hald, *Statistical Tables and Formulas*, John Wiley and Sons, New York, 1952, Table 4, and D. B. Owen, *Handbook of Statistical Tables*, Addison-Wesley, Reading, Mass., 1962, Table 2.1.

Evaluación: Principios filosóficos

***Occam's Razor*:** La mejor teoría es la más pequeña de las teorías que explican los hechos. Delante de dos teorías, la mejor es la más simple. (Guillermo de Occam, Edad Media)

***Minimun Description Length Principle*:** La mejor teoría para un conjunto de datos es aquella que minimiza el tamaño de la teoría y la cantidad de información necesaria para especifi car las excepciones. (Rissanen, 1985)

***Principio de explicaciones múltiples*:** Si más de una teoría es consistente con los hechos, disfruta de todas. (Epicúreo, Edad Antigua)