METODOS DE VALIDACIÓN

0.1 Introducción

Tal y como se ha comentado en el tema anterior, el objetivo de la clasificación supervisada es la inducción de modelos clasificatorios que tengan una buena capacidad generalizadora. Es decir modelos clasificatorios que ante un nuevo caso del que se conocen los valores de las variables predictoras sean capaces de clasificarlo correctamente con una alta probabilidad. Precisamente el objetivo de este tema es el de estudiar métodos que estimen dicha probabilidad con objeto de que tengamos una idea de la fiabilidad del modelo clasificatorio inducido.

En el apartado siguiente se introducen algunos conceptos básicos, para posteriormente presentar los distintos métodos de validación.

0.2 Conceptos básicos

Vamos a introducir algunos conceptos relacionados con la bondad de un clasificador a partir del supuesto de que la variable clase, C, puede tomar dos posibles valores que denotamos por 0 o $(+)$ y 1 o $(-)$.

La tabla 10.1 muestra una tabla de contingencia recogiendo los resultados obtenidos por un modelo clasificatorio en un problema con dos clases en el que se han clasificado un total de $a+b+c+d$ casos.

Tabla 10.1: Matriz de mala clasificación.

		Clase verdadera C	
		$0(+)$	$1(-)$
Clase predecida por el modelo de clasificación C_M	$0(+)$	a	b
	$1(-)$	c	d

En dicha tabla, la variable C_M recoge los resultados obtenidos con el modelo clasificatorio. Las 4 letras incluidas en la tabla de contingencia se interpretan de la siguiente manera:

- a es el número de casos en los que la clase verdadera es 0 y el modelo clasificatorio predice tambien 0
- d es el número de casos en los que la clase verdadera es 1 y el modelo clasificatorio predice tambien 1
- c es el número de casos en los que la clase verdadera es 0 y el modelo clasificatorio predice 1
- b es el número de casos en los que la clase verdadera es 1 y el modelo clasificatorio predice 0

Por tanto en $a+d$ casos la clasificación es correcta, mientras que en $c+b$ casos se han cometido errores en la clasificación. Definimos

- tasa de acierto $= \frac{a+d}{a+b+c+d}$
- tasa de error $= \frac{c+b}{a+b+c+d}$

Es importante notar que en las definiciones de la tasa de acierto, los dos tipos de acierto se están valorando de la misma forma. La situación es análoga al definir la tasa de error, ya que tampoco en este caso se llega a diferenciar entre los dos tipos de error.

Tratando de matizar lo anterior, podemos definir:

- sensibilidad del clasificador, $\frac{a}{a+c}$, como la proporción de verdaderos positivos
- espeficidad del clasificador, $\frac{d}{b+d}$, como la proporicón de verdaderos negativos.

0.3 Métodos para estimar la probabilidad de clasificación correcta de un clasificador

En este apartado se van a explicar distintos métodos para estimar la probabilidad de que un mo-delo clasificatorio clasifique correctamente un nuevo caso.

Antes de nada conviene aclarar que no parece justo estimar dicha probabilidad a partir del porcentaje de casos que el modelo clasificatorio inducido es capaz de clasificar correctamente en el conjunto de casos a partir del cual se ha inducido el clasificador. Es intuitivo que el proceder de esta manera proporcionaría unas estimaciones demasiado buenas. Tal situación se refleja en la figura 10.1

Figura 10.1: Estimación *no honesta* de la probabilidad de éxito de un clasificador.

En lo que sigue se presentan tres formas básicas para llevar a cabo estimaciones *honestas* de la probabilidad de éxito de un clasificador. Dichas formas son: *método H*, *método basado en el remuestreo* y *método bootstraping*.

0.3.1 Método H

El *método H* (*holdout*) tambien conocido como método de entrenamiento–testeo, se basa en particionar el fichero de entrenamiento –cuyo tamaño es N– en dos subficheros –de tamaños respectivos N_1 y N_2 ($N_1 + N_2 = N$)–. El primero de dichos ficheros se denomina de entrenamiento, ya que a partir del mismo se induce el modelo clasificatorio, M. La bondad de dicho modelo clasificatorio, es decir la estimación de la probabilidad de éxito de dicho modelo frente a casos nuevos, se obtiene por medio del porcentaje de casos bien clasificados obtenido en el segundo subfichero, el cual tiene dimensión N_2. A este segundo fichero se le denomina de testeo, ya que es en el se testa la bondad del modelo clasificatorio inducido.

La figura 10.2 representa la situación anterior.

Figura 10.2: Método de estimación entrenamiento–testeo.

Conviene tener presente que con este método el modelo clasificatorio obtenido, y que posteriormente se va a aplicar, se ha inducido a partir de N_1 casos. Suele ser habitual el utilizar las proporciones $\frac{2}{3}$ y $\frac{1}{3}$ respectivamente para los conjuntos de entrenamiento y testeo. Es decir $N_1 = \frac{2}{3}N$ y $N_2 = \frac{1}{3}N$. Se utilizará el método H en el caso de que N sea del orden de millares o superior.

0.3.2 Métodos basados en remuestreo

En este punto se explicarán tres métodos que basados en el remuestreo sirven para estimar la probabilidad de éxito de un sistema clasificatorio. La gran diferencia con relación al método H descrito anteriormente radica en que en los métodos basados en el remuestreo el modelo clasificatorio del que se estima la probabilidad de éxito es el inducido con todo el fichero de casos.

Los tres métodos que se exponen en este punto son: *submuestreo aleatorio*, *validaciones cruzadas de k rodajas* y *dejando uno fuera*.

Submuestreo aleatorio (*random–subsampling*)

En el submuestreo aleatorio (*random subsampling*) el método H se aplica B veces, sobre particiones independientes del fichero de casos. La figura 10.3 muestra gráficamente los pasos básicos del método.

Figura 10.3: Método de estimación *random–subsampling*.

Con cada una de las B particiones independientes del fichero de casos, se actúa de manera análoga. A partir del conjunto de entrenamiento se induce el modelo clasificatorio correspondiente a dicha partición. Dicho modelo se testea en el conjunto de testeo correspondiente, contándose el número de aciertos en las predicciones que efectúa el modelo. El porcentaje de casos bien clasificados, $\hat{p}_1, \ldots \hat{p}_B$, obtenido en cada uno de los B conjuntos de testeo se promedia, obteniéndose $\hat{p}_M$, cuyo valor constituye una estimación de la probabilidad de llevar a cabo una clasificación correcta por el modelo M inducido por el algoritmo a partir de todo el fichero de casos.

Validaciones cruzadas de k rodajas (*k–fold croos–validation*)

En el método de validaciones cruzadas de k rodajas *k–fold cross–validation* el conjunto de casos se particiona en k subconjuntos disjuntos de aproximadamente el mismo tamaño. Con cada uno de los k subconjuntos se induce un modelo clasificatorio considerando como conjunto de entrenamiento los $k-1$ subconjuntos restantes. La bondad de dicho modelo clasificatorio se testa con el subconjunto en cuestión, a partir del porcentaje de casos bien clasificados obtenidos con el modelo. Los k porcentajes de casos bien clasificados se promedian para obtener el estimador del modelo inducido con todos los casos. Este último modelo inducido será el que se utilice como clasificador. La figura 10.4 muestra el esquema del *k–fold cross–validation*.

Figura 10.4: Método de estimación *k–fold cross–validation*.

Tal y como se muestra en la figura 10.4, el modelo final, M, es el inducido con todo el conjunto de datos. Los k modelos, $M_1, \ldots, M_k$, cada uno de los cuales han sido inducidos con $\frac{(k-1)N}{k}$ casos, se utilizan para obtener $\hat{p}_1, \ldots, \hat{p}_k$ en sus correspondientes casos de testeo. A su vez $\hat{p}_1, \ldots, \hat{p}_k$ se utilizan para a partir de ellos obtener el valor de $\hat{p}_M$, estimación de la probabilidad de éxito en la clasificación que se obtendrá con el modelo M.

Dejando uno fuera (*Leaving–one–out*)

La validación dejando uno fuera *leaving–one–out* es un caso particular del método *k–fold cross–validation* para el caso en que k coincida con el tamaño del fichero de casos inicial, N. Es decir en el método de validación dejando uno fuera se inducirán N modelos, $M_1, \ldots, M_N$, cada uno de los cuales parte de un fichero de casos de tamaño $N-1$. Asi por ejemplo el modelo M_i se induce del fichero de casos que contiene a todos los casos excepto al i–ésimo. Dicho modelo M_i se testa en el i–ésimo caso, obteniéndose δ_i, definido como:

$$\delta_i = \begin{cases} 1 & \text{si } c^i_{M_i} = c^i \\ 0 & \text{si } c^i_{M_i} \neq c^i. \end{cases} \tag{1}$$

Es decir δ_i vale 1 si el valor de la clase a la que pertenece el i–ésimo caso coincide con la clase que predice el modelo M_i para dicho i–ésimo caso.

La estimación, $\hat{p}_M$, de la probabilidad de éxito del modelo M inducido con los N casos de los que consta el fichero inicial se obtiene como el porcentaje de éxito logrado por los N modelos anteriores en los casos que han quedado fuera. Es decir:

$$\hat{p}_M = \frac{1}{N} \sum_{i=1}^{N} \delta_i \tag{2}$$

La figura 10.5 refleja gráficamente este proceso.

Figura 10.5: Método de estimación *leaving–one–out*.

0.3.3 Método Bootstraping

En este punto se describe el estimador 0.632*Bootstraping*.

Se comienza escogiendo B muestras independientes de tamaño N extraidas con reemplazamiento del fichero de casos inicial. Con cada una de estas muestras se induce un modelo clasificatorio, el cual es testado en el fichero de casos constituido por aquellos casos que no han sido seleccionados. Denotaremos por $\hat{p}_i$, con $i = 1, \ldots, B$, el porcentaje de éxito obtenido en los casos no seleccionados por el modelo inducio a partir de los casos seleccionados en la i–ésima selección con reemplazamiento de tamaño N. Se denota por $\hat{p}_{global}$ al porcentaje de éxito obtenido por el modelo M inducido del fichero inicial a partir de todos los casos, cuando se considera como testeo dicho fichero inicial.

El estimador 0.632*Bootstraping* para la probabilidad de éxito del modelo M se denota por $\hat{p}_{0.632Bo}$ y se calcula a partir de la siguiente fórmula:

$$\hat{p}_{0.632Bo} = \frac{1}{B} \sum_{i=1}^{B} (0.632\hat{p}_i + 0.368\hat{p}_{global}) \tag{3}$$

La justificación de la utilización del número 0.632 se basa en que:

- si calculamos la probabilidad de que un caso no se seleccione como parte del conjunto de entrenamiento, obtenemos que dicha probabilidad es $(1 - \frac{1}{N})^N$. Dicha cantidad, si N es suficientemente grande, tiende a e^{-1}. Por otra parte se tiene que $e^{-1} \simeq 0.368$
- el número esperado de casos distintos que obtenemos en el conjunto de entrenamiento es: $(1 - 0.368)N = 0.632N$.

La figura 10.6 muestra gráficamente este proceso.

Figura 10.6: Método de estimación *bootstraping*.

0.4 Notas bibliográficas

El método *leaving–one–out* fué introducido por Lachenbruch y Mickey (1968) en relación con el Análisis Discriminante. La *k-fold cross–validation* se presentó por vez primera en Stone (1974), mientras que Efron y Tibshirani (1993) constituye una buena referencia al método *bootstraping*. Referencias generales para este tema son Weiss yKulikowski (1991) y Kohavi (1995).

0.5 Recursos en internet

- http://www.cs.cmu.edu/ rahuls/nips2000
  Contiene información de un congreso reciente sobre los últimos avances en relación con las técnicas de *cross–validation* and *bootstraping*

Referencias

1. B. Efron, R. Tibshirani (1993). *An Introduction to the Bootstrap.* Chapman and Hall.

2. R. Kohavi (1995). A study of cross–validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1137–1143.

3. P. Lachenbruch, R. Mickey (1968). Estimation of error rates in Discriminant Analysis. *Technometrics*, 10: 1–11.

4. M. Stone (1974). Cross–validation choice and assesment of statistical predictions. *Journal of the Royal Statistic Society*, 36: 111–147.

5. S. M. Weiss, C. A. Kulikowski (1991). *Computer Systems that Learn.* Morgan Kaufmann Publishers.