Evaluación de modelos

Luis Carlos Molina Félix
Ramon Sangüesa i Solé

PID_00165734

UOC
Universitat Oberta de Catalunya
www.uoc.edu

Índice

Introducción

El proceso de *data mining* tal como lo hemos descrito no es en absoluto lineal y no consiste en un solo paso ni mucho menos. Corresponde más a un proceso iterativo en el que se obtiene un primer modelo, se evalúa y, dependiendo de si esa evaluación es satisfactoria o no, se da por bueno y se acaba el proceso, o bien se vuelve a un paso anterior, ya sea la selección o la preparación de datos.

En este módulo didáctico nos centramos en el problema de la evaluación del modelo, precisamos la problemática de esta fase e introducimos los métodos y las medidas que se utilizan de forma más habitual para encontrar la calidad de los modelos obtenidos. 

Aunque la calidad de un modelo siempre está en relación con la tarea para la cual ha sido construido (predicción, clasificación, etc.), veremos que hay algunas medidas generales independientes de la tarea que hay que conocer.

Objetivos

Los materiales didácticos asociados a este módulo permitirán que el estudiante alcance los objetivos siguientes:

1. Ser capaces de aclarar los procedimientos y las medidas necesarias para establecer la calidad de los modelos obtenidos a partir de un proceso de *data mining*.

2. Conocer las medidas más adecuadas para cada tipo de tarea.

3. Establecer los procedimientos de comparación entre los resultados que se obtienen con varios tipos de métodos de *data mining*.

1. Evaluación de modelos

Cuando tenemos que evaluar un modelo, partimos de una situación en la que ya hemos utilizado un determinado conjunto de datos, hemos aplicado un método de *data mining*, hemos obtenido un primer modelo y nos encontramos en la necesidad de saber si este modelo es bueno o no con respecto a la tarea que nos hemos propuesto.

Este problema se extiende al hecho de que debemos comparar el funcionamiento de varios métodos de *data mining* con respecto al mismo conjunto de datos.

Ejemplo de problema de comparación de modelos

Hemos construido varios modelos para clasificar a clientes de un banco en 'Morosos' o 'No morosos': un modelo de árbol de decisión, una red neuronal o una lista de reglas de clasificación. Queremos ver qué modelo augura un mejor rendimiento. Evidentemente, nos interesa elegir el que asegure un mayor porcentaje de aciertos en la clasificación.

Pensemos por un momento qué significa exactamente que un modelo sea bueno. En el caso de una tarea de clasificación, el modelo será bueno si asigna correctamente las etiquetas de clase a cada objeto cuando se aplica a objetos nuevos que no se habían utilizado en la construcción del modelo clasificatorio. Para una tarea de agregación de objetos, un modelo es bueno cuando es capaz de crear divisiones del conjunto original que sean bastante diferentes entre sí y que, además, los objetos de cada partición mantengan una alta similitud. Además, es preciso que el modelo se pueda enfrentar con objetos nuevos del mismo dominio y comportar de forma correcta: predecir su etiqueta de clase o integrarlos en la partición más adecuada.

Por tanto, la **calidad de un modelo** es una estimación del comportamiento futuro del modelo con respecto a objetos del dominio.

Calidad de un modelo

La calidad de un modelo es una estimación de su futuro comportamiento con respecto a objetos del dominio.

Una manera típica de enfrentarse al proceso de medir la calidad de un modelo consiste en separar el conjunto de datos disponibles en dos grupos:

- El **conjunto de entrenamiento**. Conjunto de los datos que realmente se utilizan para construir el modelo.

- El **conjunto de prueba**. Conjunto de los datos que se utilizan para ver si el modelo actúa de forma correcta.

Ejemplo de proceso de medida de la calidad de un modelo: los clientes de Hyper-Gym

En nuestro ejemplo del gimnasio podemos separar el conjunto de socios en dos subconjuntos.

Supongamos que utilizamos el primer conjunto para extraer un modelo clasificatorio, por ejemplo un árbol de decisión, que separe los socios en dos clases: los que puede ser que soliciten servicio de entrenador personal y los que no lo solicitarán.

El segundo conjunto de socios, el conjunto de prueba, lo utilizaremos para saber qué comportamiento presenta el árbol de decisión construido. Observad que ya conocemos la clase a la que pertenecen estos clientes del conjunto de pruebas (ya sabemos si solicitan entrenador o no). El objetivo de nuestro modelo es poder predecir si un nuevo cliente lo solicitará. Esperaremos que, si el modelo clasificatorio construido es lo suficientemente bueno, todos o un gran porcentaje de los clientes del conjunto de pruebas sean clasificados correctamente mediante el árbol de decisión en la clase que les corresponde. Los clientes del conjunto de prueba no se han utilizado para construir el árbol de decisión y, en consecuencia, son nuevos para el árbol.

En el procedimiento de evaluación se plantea una serie de cuestiones como las que mencionamos a continuación:

1) ¿Cuál es la medida de calidad adecuada?

Podéis ver el ejemplo de proceso de medida de la calidad de un modelo más adelante en este mismo subapartado.

Ejemplo de medida de calidad

En el caso de los clientes de Hyper-Gym, aunque no lo mencionamos explícitamente, utilizamos la idea de **error en la predicción**. Si el número de predicciones erróneas es muy elevado, consideramos que el árbol de decisión no es lo suficientemente bueno y no lo utilizamos para clasificar clientes nuevos.

Ahora bien, para otras tareas, por ejemplo para la agregación, ¿qué medidas podríamos utilizar?

2) ¿Cómo fijamos el nivel de calidad?

Podéis consultar el ejemplo de proceso de medida de la calidad de un modelo más adelante en este mismo subapartado.

Ejemplo de nivel de calidad

En el ejemplo de los clientes de Hyper-Gym nos planteamos la cuestión siguiente: ¿estaríamos satisfechos con un árbol de decisión que clasificara correctamente al 89% de los clientes del conjunto de prueba o desearíamos quizá que clasificara el 95%?

La respuesta a esta cuestión no es tan sencilla como parece. Por ejemplo, debemos tener en cuenta el coste que supone clasificar un nuevo objeto de manera incorrecta. ¿Qué es más costoso: asignar un cliente de la clase 1 a la clase 2, o viceversa? Éste es el problema de los falsos positivos o falsos negativos tan conocido en estadística, y especialmente punzante en el diseño de métodos de diagnóstico nuevos en medicina o en herramientas automáticas de diagnosis de sistemas mecánicos complejos.

3) ¿Cómo elegimos los objetos que deben entrar en los conjuntos de entrenamiento y de prueba?

Elección de los conjuntos de entrenamiento y de prueba

En el caso del ejemplo de los clientes de Hyper-Gym, los objetos que elegimos son socios o clientes.

Es evidente que si elegimos un conjunto de entrenamiento con características comunes parecidas, pero muy diferentes de las del conjunto de prueba, el conjunto estará sesgado y será poco representativo. En tal caso, seguro que el modelo resultante confiere valores de calidad bajos.

La estadística permitirá responder a estas preguntas con más precisión. De momento, avanzamos que hay dos situaciones diferentes de cara a hacer frente a la evaluación de modelos:

- Cuando disponemos de **grandes cantidades de datos**, tanto de entrenamiento como de prueba. En principio, en esta situación el método que hemos esbozado podría ser suficiente si tomamos algunas precauciones en la selección de los objetos que se incluyan en cada conjunto. Cuanto mayor

sea la cantidad de datos, más soporte tendremos para valorar como "bueno" o "malo" el modelo resultante.

- Cuando disponemos de **pocos datos**. En este caso, lo que podemos inferir con el modelo tiene menos soporte y se hace más difícil justificar que las calidades del modelo se mantendrán frente a objetos desconocidos, pero es igualmente necesario tener una idea de cómo se comportará el modelo.

La literatura de *data mining* suele ignorar el segundo problema, puesto que las técnicas de *data mining* se aplican justamente a grandes conjuntos de datos. Pero para determinadas aplicaciones el número de datos no es tan elevado. Hay que ser consciente de ello y conocer las técnicas de evaluación de modelos correspondientes.

1.1. Evaluación de modelos clasificatorios

En este subapartado nos planteamos el problema de medir como es debido la calidad de un modelo clasificatorio (árbol de decisión, red neuronal de clasificación, lista de reglas de clasificación, etc.).

La **evaluación de un modelo clasificatorio** es la capacidad para predecir correctamente la clase a que pertenecen objetos no utilizados en su construcción.

Primero veremos las medidas de calidad y después describiremos el procedimiento de evaluación.

Medidas de calidad de los modelos clasificatorios

Como hemos dicho, la base de la medida de la calidad en clasificación es saber si el modelo etiquetará correctamente objetos nuevos.

Simplificaremos el problema suponiendo que hay dos clases, *A* y *B*. En este caso, clasificar se reduce a determinar si un objeto pertenece a la clase *A* o a la *B*.

Establecemos la **medida de la calidad** contando una operación de clasificación como correcta cuando un objeto que pertenece a la clase *A* recibe la etiqueta *A* por parte del clasificador (y de la misma manera, cuando un objeto de la clase *B* recibe la etiqueta *B*), y contando una operación de clasificación como incorrecta cuando un objeto de la clase *A* recibe la etiqueta *B*, y viceversa.

Una forma de medir la calidad del modelo consiste en contar o predecir el porcentaje de clasificaciones incorrectas que efectuará el modelo (o, en positivo, el porcentaje de operaciones de clasificaciones correctas) a partir de un conjunto de observaciones o casos (clientes o socios, en el ejemplo de HyperGym). Cada una de las operaciones de clasificación incorrectas es un **error de clasificación**.

Por lo tanto, el parámetro de calidad que debemos considerar o estimar es la **tasa de error**:

$$\text{Tasa de error} = \frac{\text{Errores}}{\text{Casos}}$$

Alternativamente, podemos considerar como parámetro de calidad el porcentaje de éxitos (operaciones de clasificación correctas), que también recibe el nombre de **precisión del modelo**:

$$\text{Precisión} = \frac{\text{Exitos}}{\text{Casos}}$$

Ejemplo de medida de la calidad de un modelo

Aquí presentamos un ejemplo sencillo para el caso de la clasificación de clientes del gimnasio entre los que pueden querer entrenador personal y los que no:

Caso	Clase real	Clase predicha
34	Entrenador	Entrenador
24	Entrenador	Entrenador
56	No entrenador	Entrenador
110	No entrenador	Entrenador
21	No entrenador	Entrenador
1	No entrenador	Entrenador
23	Entrenador	Entrenador
77	Entrenador	Entrenador
29	No entrenador	No entrenador
101	No entrenador	No entrenador

La tasa de error es del 40% o, si lo preferís, la precisión es del 60%.

1.1.1. Grandes cantidades de datos

Con la medida de calidad para un modelo de clasificación que hemos presentado en el subapartado anterior ya podemos emprender el proceso de evaluación del modelo. Veamos a continuación cómo.

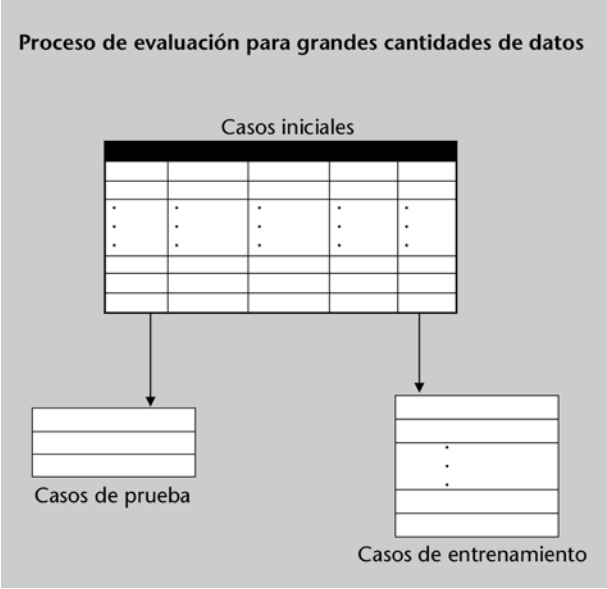
Consideraremos una situación en la que disponemos de grandes cantidades de datos. En tal caso, el procedimiento consiste en disponer de dos conjuntos de datos:

Ejemplo de conjuntos de entrenamiento y de prueba

Se pueden extraer datos de dos oficinas bancarias de un mismo barrio y utilizar los datos de los clientes de la primera oficina para construir el modelo y los datos de los clientes de la segunda, para evaluarlo.

a) El **conjunto de entrenamiento.** Con los datos que contiene este conjunto se construirá el modelo clasificatorio (árbol de decisión, red neuronal de clasificación o lista de reglas de decisión).

b) El **conjunto de prueba**. Se utilizará para comprobar la tasa de error (o precisión del modelo obtenido).



De esta manera, se intenta reducir el problema de la sobreespecialización*, que obtiene modelos demasiado ajustados a datos muy particulares. Si nos descuidamos, podemos inducir modelos demasiado específicos que no darán buenos resultados al aplicarlos a datos reales nuevos.

*** En inglés, *overfitting*.**

Debemos suponer que ambos conjuntos son una buena muestra de la población de la que se obtienen dichos datos. Resulta conveniente utilizar más datos para el conjunto de entrenamiento que para el de prueba. Asimismo, conviene efectuar una extracción aleatoria de casos para constituir cada conjunto.

Este esquema funciona bien con grandes números de datos básicamente porque cuanto mayor sea la muestra de datos, más cerca de los parámetros de la población deberían estar los valores de los parámetros de la muestra:

a) Cuanto mayor es el conjunto de entrenamiento, mejor es el clasificador (aunque hay matices: a partir de cierto número de casos de entrenamiento el rendimiento disminuye).

b) Por otra parte, cuanto mayor es el conjunto de prueba, más cerca de la realidad está el porcentaje de error que obtenemos.

Importancia del volumen de datos del conjunto de prueba en la precisión del modelo

Si trabajamos con un conjunto de prueba de quinientos casos y obtenemos un porcentaje de error del 2%, podemos ser muy optimistas y creer que tenemos un clasificador fantástico con una precisión del 98%. Ahora bien, podemos considerar que quinientos casos es una muestra pequeña. Si disponemos de un millón de casos del mismo dominio y resulta que el porcentaje de error es del 40% (y la precisión, del 60%), podemos estar más seguros de que la tasa de error real que mostrará el clasificador estará más cerca del 40% que del 2%. Más adelante veremos las razones de este hecho.

Las suposiciones que se adoptan para llevar a cabo este tipo de método de validación son las siguientes:

- Los conjuntos de datos que se han separado son independientes entre sí.
- Los casos son independientes entre sí.

Hay que procurar que estas dos condiciones se cumplan efectivamente, dado que en caso contrario, los resultados no serán indicativos.

Actividad

1.1. ¿Por qué el tamaño de los diferentes conjuntos aporta más o menos seguridad a las tasas de error y precisión obtenidas?

Confianza en la medida del error

Se espera que el **error de una muestra** se sitúe dentro de dos desviaciones típicas de la media, hecho que representa un acierto en el 95% de los casos. Esto se debe a que las medias para muestras de grandes dimensiones siguen una distribución normal.

Partiendo de que se obtiene una tasa de error determinada mediante un test sobre un conjunto de prueba, ¿qué confianza podemos tener con respecto al verdadero valor del error del clasificador? Supongamos que hemos obtenido un error del 2,5%. ¿Cómo podemos saber y con qué grado de confianza si éste es el verdadero error que mostrará el clasificador? ¿Cómo podemos saber entre qué valores variará el error con una confianza suficiente? Una vez más, la estadística nos ayudará a responder estas cuestiones.

1.1.2. Conjuntos limitados de datos: validación cruzada

Cuando contamos con pocos datos, debemos aprovecharlos y extraer tanto el conjunto de prueba como el de datos de la misma base de datos original. No

hay ninguna proporción fijada con respecto al número relativo de componentes de cada subconjunto, pero la más utilizada acostumbra a ser 2/3 para el conjunto de entrenamiento y 1/3, para el conjunto de prueba.

Hay que hacer una extracción adecuada de observaciones para asegurarnos de no obtener clasificadores sesgados.

No hay recomendaciones generales, pero el sentido común nos obliga a que, después de una extracción aleatoria de casos, efectuemos un análisis de datos mínimo para garantizar que, por ejemplo, no haya una clase sobrerrepresentada en el conjunto de entrenamiento. Un mínimo uso de la estadística descriptiva nos permite averiguar este tipo de cuestiones, aunque no ofrece una garantía completa: puede haber grupos de atributos y valores sobrerrepresentados más allá del atributo clase.

Ejemplo de clase sobrerrepresentada

Si queremos construir un clasificador para distinguir los clientes que solicitarán entrenador de los que no lo harán, tenemos que evitar, por ejemplo, que en el conjunto de entrenamiento haya un 80% de clientes que han solicitado entrenador.

La validación cruzada es el procedimiento más utilizado para asegurar que el error calculado es próximo al real aunque se utilice una cantidad de datos relativamente pequeña.

Puede decirse que el procedimiento de la validación cruzada "aumenta" de alguna manera el tamaño de la base de datos y asegura la independencia entre los conjuntos de pruebas utilizados.

Validación cruzada

La validación cruzada* es un método utilizado para calcular la precisión de un clasificador con respecto a un conjunto de datos.

La validación cruzada divide el conjunto de datos en k subconjuntos mutuamente excluyentes de tamaño aproximadamente igual.

*** En inglés, *cross validation*.**

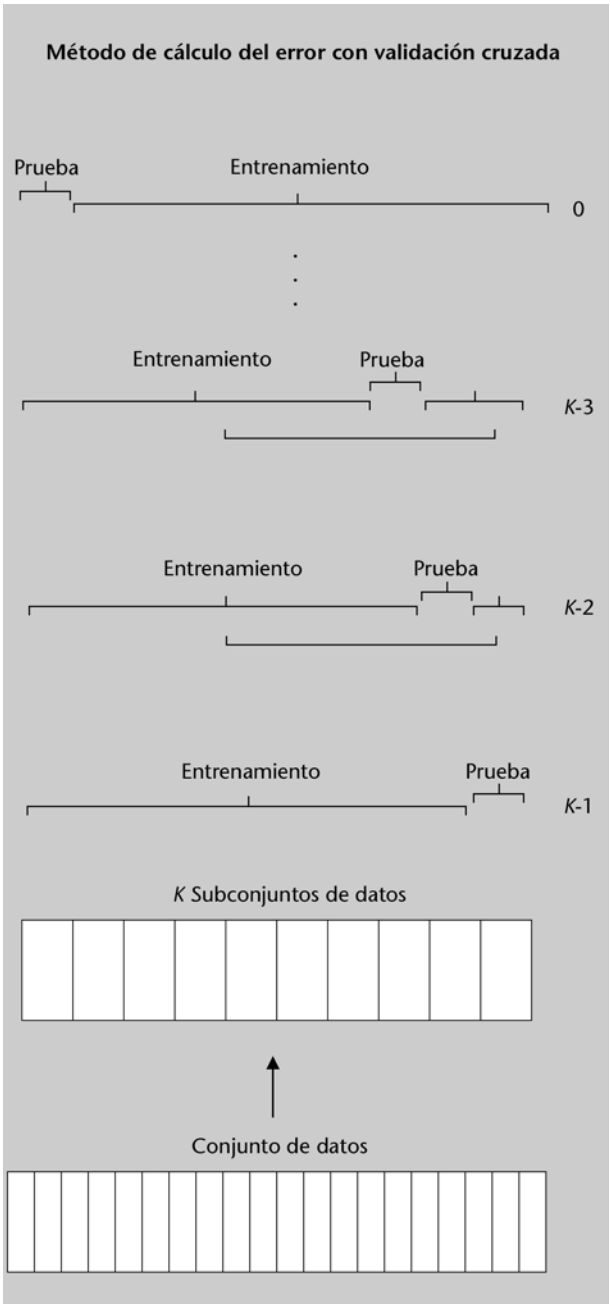
Empíricamente se ha encontrado que el tamaño más adecuado, o el valor adecuado para k es 10.

Una vez formados los subconjuntos, el clasificador tomará $k - 1$ de esos subconjuntos (nueve subconjuntos, en el caso de $k = 10$) como conjunto de entrenamiento y el subconjunto que queda, como conjunto de prueba, y calculará su error de clasificación.

A continuación, se toma el subconjunto siguiente como conjunto de prueba, se toman los demás $k - 1$ como conjuntos de entrenamiento y se vuelve a calcular la tasa de error.

El procedimiento se repite hasta que se agotan todas las combinaciones de subconjuntos posibles.

En la figura siguiente mostramos gráficamente el procedimiento mencionado:



En cada paso se tienen $k - 1$ conjuntos de entrenamiento y un conjunto de prueba. Una vez obtenidas todas las tasas de error de clasificación de los k experimentos que se han realizado en total, se calcula su media. El resultado es el error de clasificación obtenido mediante la validación cruzada.

La ventaja de este método es que utiliza todas las observaciones del conjunto de datos para entrenamiento y para prueba, algo que no ocurre con otros métodos de cálculo del error de clasificación.

Si, además, efectuamos la selección de las particiones de manera que las clases estén representadas del mismo modo que en el total del conjunto de datos,

tanto en el conjunto de prueba como en el de entrenamiento, evitamos los problemas de sesgo que hemos mencionado. Este proceso se denomina **estratificación**, y su combinación con la validación cruzada, **validación cruzada estratificada**.

Supongamos que tenemos mil datos con observaciones que pueden pertenecer a tres clases y la distribución de observaciones por clase es del 30% para la primera clase, del 40% para la segunda, y del 30% para la tercera, y adoptamos una validación cruzada de cinco particiones. Entonces habrá particiones de doscientas observaciones cada una, y cada partición habrá de tener aproximadamente sesenta observaciones de la primera clase, ochenta de la segunda y sesenta de la tercera.

El procedimiento estándar consiste en efectuar una validación cruzada en diez particiones y ejecutarlo diez veces con el fin de acomodar las variaciones aleatorias. Al final del proceso hay que conseguir la media de los errores obtenidos para tener una estimación del error que pueda presentar el modelo de clasificación.

Como podemos ver, la estimación del error de clasificación de un método determinado implica un esfuerzo computacional nada despreciable. No es extraño que en algunas ocasiones, cuando la precisión no es una necesidad extrema y el tiempo de obtención del modelo es una consideración importante, se prefieran esquemas de evaluación más simples.

Complejidad de la validación cruzada

Una validación en diez particiones implica que el método de construcción del modelo clasificatorio se debe ejecutar 10×10 veces.

1.1.3. Comparación de rendimientos

Hemos comentado que una de las utilidades de la evaluación de modelos es la posibilidad de comparar qué método podría dar los mejores resultados.

Utilidad de la evaluación de modelos

La evaluación de modelos nos permite comparar si una red neuronal podría darnos un error de clasificación menor que un conjunto de reglas de clasificación o que un árbol de decisión extraídos de los mismos datos.

Parecería suficiente con establecer un procedimiento como el que hemos comentado antes; es decir, obtener la predicción de error para cada método y seleccionar el método que muestre la tasa de error mínima. Sin embargo, nos volvemos a encontrar en una situación típica de la inferencia estadística: la diferencia entre los resultados de los métodos, ¿se debe a una variabilidad aleatoria o es realmente significativa? Para valorar los resultados y encontrar la manera de plantear la comparación, es preciso que introduzcamos algún otro cambio en el proceso.

Supongamos, para simplificar, que tenemos que comparar dos métodos que denominaremos *método A* y *método B* (una red neuronal de clasificación y un árbol de decisión, por ejemplo).

Ya hemos efectuado la validación cruzada del modelo A y del modelo B obtenidos por los métodos respectivos y ahora tenemos una estimación del error de A, e_A, y otra de B, e_B. Ahora debemos hacer una prueba de hipótesis: debe-

mos decidir, por ejemplo, si e_A es significativamente superior a e_B. Lo haremos con un test típico: la *t* de Student.

Supongamos un conjunto de valores de error de clasificación obtenidos por el primer método, e_{a1}, ..., e_{ak}, donde $k = 10$ si hemos hecho una validación cruzada en diez particiones. Para el caso del método *B*, tenemos e_{b1}, ..., e_{bk}, también con $k = 10$. Indicamos la media de la primera muestra de errores con $\overline{e_a}$ y la media de la muestra de los errores obtenidos con el método *B*, con $\overline{e_b}$. Expresamos la diferencia entre medias con $dif = e_a - e_b$.

La prueba de hipótesis que nos planteamos consta de dos hipótesis alternativas: $dif = 0$ y $dif \neq 0$.

Debemos establecer un nivel de significación y ver si la diferencia apreciada supera el nivel que nos hemos propuesto. Para hacerlo tenemos que derivar la estadística *t* de Student:

$$t = \frac{\overline{dif}}{\sqrt{\sigma^2_{dif} / k}}$$

Explicamos a continuación cada factor de los que contiene esta expresión:

- $\overline{dif}$ es la media de las diferencias entre los errores observados para el método *A* y para el método *B*: $e_{a1} - e_{b1}$, (dif_1), $e_{a2} - e_{b2}$, (dif_2), $e_{ak} - e_{bk}$, (dif_k).

- *k*, en nuestro caso, es 10 (el número de particiones e iteraciones efectuadas en la validación cruzada y, por lo tanto, el número de errores que hemos anotado).

- σ^2_{dif} es la varianza calculada a partir de los valores de las diferencias observadas de la muestra entre los distintos errores: $e_{ai} - e_{bi}$, con *i* que varía de 1 a 10.

La distribución que sigue esta estadística es la distribución *t* de Student con $k - 1$ grados de libertad. En consecuencia, hay que utilizar las tablas de intervalos de confianza correspondientes a esta distribución.

En el caso que consideramos aquí, partimos de la suposición de que cada error $e_{ai} - e_{bi}$ se refiere a la misma partición de los datos. Si no fuera así, deberíamos seguir un método diferente.

1.2. Otras maneras de estimar la calidad de modelos predictivos

Hemos discutido y explicado las necesidades de un tipo de modelo predictivo determinado: los modelos clasificatorios. Hay otros métodos orientados a predecir no ya una etiqueta de clase, sino un valor numérico concreto, por ejemplo los métodos de regresión.

En este caso, los tipos de error que utilizamos no son solamente "aciertos" o "fracasos", sino que también podemos equivocarnos en grado.

La metodología general de evaluación no varía: validación cruzada, cálculo de diferencias entre los dos métodos, establecimiento de la prueba de hipótesis y uso de la distribución *t* de Student para evaluar el nivel de confianza. Ahora bien, hay que dar otra formalización del error.

En estos casos, partimos de un conjunto de atributos conocidos $x_1, \ldots, x_n$ y tenemos que predecir un valor numérico de salida, y. De hecho, el modelo que construimos es una función que establece la relación entre y y el resto de las variables de interés:

$$y = a_0 + w_1 x_1 + \ldots + w_n x_n$$

Indicamos con y el valor real y con y', el valor predicho. En tal caso, hay otras formalizaciones posibles del error:

a) Error cuadrático:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2$$

b) Error estándar:

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{(y_i - y_i')^2}$$

c) Error medio absoluto:

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - y_i'|^2$$

d) Error cuadrático relativo:

$$\frac{\sum_{i=1}^{n}|y_i - y_i'|^2}{\sum_{i=1}^{n}|y_i - \overline{y_i'}|^2}$$

donde $\overline{y_i'}$ es la media de los valores y_i.

La mayoría de estas medidas dan una aproximación adecuada del error de predicción en que se incurre. De todos modos, las dos primeras asignan un peso demasiado fuerte a los valores extraños*.

* En inglés, *outliers*.

1.3. Coste

Hasta ahora hemos estudiado la calidad de los métodos de clasificación –o de predicción numérica– teniendo en cuenta tan sólo el error de predicción que pueden mostrar. Naturalmente, tendemos a preferir métodos que posean un porcentaje bajo de error en predicción. Ahora bien, al comparar dos métodos diferentes, nos interesa tener en cuenta también el coste del error. Es decir, queremos poder decidirnos por un método u otro según las consecuencias de su comportamiento erróneo.

Podemos expresar el problema del coste de la manera más sencilla cuando tenemos únicamente dos clases: + y – (o dos valores numéricos, por ejemplo 0 y 1). La tabla siguiente lo esquematiza (matriz de confusión):

		Clase predicha	
		+	–
Clase real	+	Positivo verdadero	Falso negativo
	–	Falso positivo	Negativo verdadero

El coste de predecir un falso positivo o un falso negativo depende de la aplicación a la que se quiera destinar el modelo. Por ejemplo, si queremos crear un clasificador que prediga correctamente si un paciente puede sufrir una enfermedad determinada, un falso positivo pondrá en tratamiento a una persona sana y un falso negativo la dejará sin él. Cada uno de estos dos casos tiene un claro coste económico (y humano).

Lectura recomendada

Para una discusión más profunda de las medidas de utilidad, podéis consultar la obra siguiente:

D. Heckerman (1996). "Bayesian Networks for Knowledge Discovery". En: U. Fayyad; G. Piatetsky-Shapiro; P. Smyth; U. Uthurusamy (ed.). *Advances in Knowledge Discovery and Data Mining* (pág. 273-306). Menlo Park: AAAI Press.

Al utilizar las medidas anteriores, mezclábamos el número de positivos correctos y negativos correctos y lo dividíamos por el total de casos de prueba. Para considerar el coste, debemos tener una idea (no una medida) de la utilidad del modelo de clasificación que hemos construido.

1.3.1. Aproximación de costes

Cuando se desconoce el coste de las distintas alternativas, se puede intentar trabajar con escenarios que permitan relacionar alternativas y resultados posibles. Una manera de hacerlo es tener en cuenta el concepto de *lift* de un modelo predictivo.

El ***lift*** consiste en evaluar el cambio que hay en la probabilidad de una clase cuando el modelo predictivo se aplica a una población que muestra unas características especiales. Se define con la expresión siguiente:

$$Lift = \frac{P(Clase|muestra)}{P(Clase|población)}$$

Este concepto procede del *marketing*. En entornos de *marketing* es muy típico hacer una estimación del retorno que puede generar una campaña de *mailing*.

Imaginemos que tenemos un modelo que nos permite predecir si un cliente responderá al *mailing* recibido. Supongamos también que elegimos un conjunto de clientes para evaluar el modelo con respecto a datos que no se han utilizado para construirlo (el conjunto de evaluación).

El modelo asigna la etiqueta + a aquellos clientes que predice que contestarán al *mailing* y –, a los que no lo harán. Si el modelo es bueno, habrá que esperar que entre los clientes marcados como + haya una proporción mayor de clientes que respondan efectivamente que la que haya en el resto de la población. Es decir:

$$Lift_{mailing} = \frac{P(Clase_{+}|muestra)}{P(Clase_{+}|población)}$$

debería ser mayor que 1.

Por ejemplo, si el conjunto de evaluación contiene el 3% de clientes que responden a *mailings* y el clasificador asigna correctamente la etiqueta + al 60%, entonces el factor de *lift* es:

$$Lift_{mailing} = \frac{0{,}60}{0{,}03} = 20$$

Si tenemos otro método de predicción que posee un factor de *lift* de 25, no podemos afirmar que sea un modelo mejor, porque ese factor se encuentra en relación con el tamaño de la muestra y de la población, así como con los otros costes asociados.

En este ejemplo siempre debemos preferir un modelo que dé una cantidad significativa de clientes que contesten al *mailing*. En el ejemplo, si en la muestra tenemos una proporción del 3% de clientes que contestan, el máximo factor de *lift* que se puede conseguir es:

$$Lift_{mailing} = \frac{1{,}00}{0{,}03} = 33{,}\overline{3}$$

Ahora bien, no es lo mismo un método que tiene un *lift* de 33,3 con respecto a una población de cien mil clientes que uno de 25 con respecto a una población de un millón de clientes. En principio, nos interesa más un método que lo acierta el 25% de las veces para una población de un millón de personas (250.000 respuestas positivas a la campaña de *marketing*) que uno que lo acierta la tercera parte de las veces para una población de cien mil personas (33.333 clientes que responderán positivamente). Todavía tenemos que ma-

tizar lo expuesto mucho más. En efecto, hay otros factores de coste que debemos tener en cuenta:

- ¿Cuál de los dos métodos es más caro de implementar?

- ¿Cuál es el coste asociado de realizar un envío de *mailing* a 250.000 clientes frente al coste de un envío a 33.333 clientes?

Si no dispusiéramos de ningún modelo predictivo y enviáramos un *mailing* a toda la población objetivo (pongamos por caso un millón de personas), incluyendo todos los clientes que pueden responder, nos encontraríamos con lo siguiente: si sólo enviamos publicidad al 10% de la población, sólo podremos llegar al 10% de los clientes que responden; si enviamos al 25%, al 25% de quienes responden, etc.

En un gráfico en el que pongamos en el eje *X* el porcentaje de población a la que se envía *mailing* y en el eje *Y*, la población de clientes que responden, esta estrategia "desinformada" dará una línea recta inclinada un ángulo de cuarenta y cinco grados.

En cambio, si utilizamos un buen método predictivo, debemos esperar que la línea se halle por encima de la de cuarenta y cinco grados. De hecho, sólo nos interesan las acciones que nos sitúen en la zona superior.

Resumen

La evaluación de modelos es importante por dos motivos:

a) Por una parte, nos da una idea de la calidad del modelo obtenido.

b) Por otra, nos permite comparar el rendimiento de dos o más métodos diseñados para resolver la misma tarea. Por ejemplo, la clasificación obtenida por un método basado en redes neuronales frente a la que se obtiene con un método que utiliza árboles de decisión o clasificación bayesiana "ingenua".

En el caso de los clasificadores el método más común para tener una aproximación a la calidad es el cálculo de la tasa de error en clasificación. Es decir, la proporción de observaciones mal clasificadas con respecto al total de observaciones.

Distinguimos las dos situaciones siguientes:

1) Grandes volúmenes de datos. Cuando hay muchos datos, se suele separar el conjunto original de datos en dos, un conjunto de entrenamiento y otro de prueba.

a) El **conjunto de entrenamiento** recoge los datos que permiten construir el clasificador.

b) El **conjunto de prueba** sirve para calcular la tasa de error.

En ocasiones, se aconseja utilizar un tercer conjunto, el **conjunto de validación**, procedente de un conjunto diferente de datos, aunque se refiere al mismo dominio. Una vez validado el modelo, se puede volver a probar sobre el segmento de datos inicialmente separado como conjunto de prueba.

2) Escaso volumen de datos. Cuando hay pocos datos, es recomendable utilizar el método de validación cruzada y estratificada.

Para comparar el rendimiento de varios métodos, lo más corriente es ver las diferencias en las respectivas tasas de error. Lo más adecuado, sin embargo, es establecer una prueba de comparación refiriéndose a las tablas de la *t* de Student. Finalmente, puede ser interesante tener en cuenta otros factores en la comparación de método, como el coste de cada método o el retorno esperado. El *lift* es un parámetro que trata de reflejar estas cuestiones.

La evaluación de los métodos de agregación se efectúa sobre bases parecidas, pero aquí el parámetro para comparar no es la tasa de error, como en clasificación. Por norma general, en el proceso de validación cruzada se comparan las mismas medidas de calidad utilizadas para construir el modelo.

Ejercicios de autoevaluación

1. Suponiendo que después de un proceso de validación cruzada aplicada sobre el mismo conjunto de datos a dos métodos diferentes, habéis obtenido los resultados que presentamos en la tabla siguiente para el error de clasificación (expresado en fracciones de la unidad), decid cuál de los dos métodos es mejor.

Método 1	0,25	0,05	0,34	0,08	0,02	0,10	0,04	0,08	0,09	0,23
Método 2	0,38	0,15	0,04	0,16	0,05	0,08	0,01	0,09	0,12	0,23

2. Suponiendo que tenéis dos métodos que os han devuelto las respectivas matrices de confusión que os presentamos a continuación, decid cuál de los dos consideráis que es mejor.

		Clase predicha	
		Entrenador personal	**No entrenador personal**
Clase real	***E***	0,45	0,60
	$\neg E$	0,55	0,40

		Clase predicha	
		Entrenador personal	**No entrenador personal**
Clase real	***E***	0,65	0,40
	$\neg E$	0,35	0,60