Minería de Datos

Tema 3 - Evaluación de Modelos Predictivos

Nombre: Alejandro pérez Belando

---

1. Introducción

El objetivo de las técnicas de aprendizaje automático es calcular la función que predice la clase (f), considerando un espacio de posibles hipótesis (H). Las distintas técnicas emplean una evidencia o muestra (S) formada por ejemplos de la función (f) de acuerdo con una distribución (D).

Una única evidencia puede resultar en muchas hipótesis distintas.

2. Medidas de Calidad

2.1. Exactitud/Error de Predicción

Las medidas más utilizadas para evaluar clasificadores se basan en la exactitud (accuracy) de la hipótesis (h), o su error, respecto a la función (f).

Caso ideal
Disponer de un conjunto de ejemplos completos o de su distribución de probabilidad. En este caso, el error verdadero para:

- Un conjunto de todos los ejemplos posibles (U): $E_v(h) = \frac{1}{|U|} \sum_{x \in U} \delta\left(f(x) \neq h(x)\right)$
- La distribución de probabilidad (D): $E_v(h) = Pr_{x \in D}\left[\delta\left(f(x) \neq h(x)\right)\right]$

Hay que tener en cuenta que si tenemos todos los datos posibles, entonces no tiene sentido entrenar un modelo.

Caso real: solo disponemos de una muestra o evidencia (S) de conjunto de todos los ejemplos posibles (U), de forma que el error de clasificación de la hipótesis (h) en la muestra es:

$$E_S(h) = \frac{1}{|S|} \sum_{x \in S} \delta\left(f(x) \neq h(x)\right)$$

Análogamente, la exactitud (accuracy) de clasificación:

$$A_S(h) = \frac{1}{|S|} \sum_{x \in S} \delta\left(f(x) = h(x)\right)$$

De forma práctica: el error de clasificación y la exactitud (accuracy) se calculan, respectivamente, de la forma:

$$E = \frac{n - n_c}{n} \; ; \qquad A = \frac{n_c}{n}$$

Donde: $n \equiv$ N° total de instancias y $n_c \equiv$ Nº de instancias clasificadas correctamente.

2.2. Matriz de confusión

Una matriz de confusión para tres clases tiene la forma:

Reales	**Estimadas**		
	C_1	C_2	C_3
C_1	n_{11}	n_{12}	n_{13}
C_2	n_{21}	n_{22}	n_{23}
C_3	n_{31}	n_{32}	n_{33}

Donde: $n_{ij} \equiv$ N° de ejemplos que perteneciendo a la clase C_i se han clasificado como la clase C_j.

La diagonal principal son los ejemplos predichos correctamente (es decir, n_{11}, n_{22}, n_{33}).

2.2.1. Evaluación basada en el coste

Medidas de calidad de forma genérica: $C(\epsilon) = \sum_{i=1}^{n} \sum_{j=1}^{n} n_{ij} c_{ij}$, donde $c_{ij} \equiv$ coste asociado a cada elemento de la matriz de confusión.

Entonces, para calcular, o bien el error o bien la exactitud del modelo, bastaría con definir respectivamente las matrices de costes:

$$c_{ij} = \begin{cases} 1 & \text{si } i \neq j \\ 0 & \text{en otro caso} \end{cases} \qquad c_{ij} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{en otro caso} \end{cases}$$

2.2.2. Índice kappa

El problema que tiene la exactitud del modelo es que también cuenta como favorables los aciertos debidos a la casualidad. Para resolver esto, empleamos el índice kappa:

$$\kappa = \frac{P_o - P_c}{1 - P_c}$$

Donde:

- $P_o = \sum_{i=1}^{3} \left(\sum_{j=1}^{3} n'_{ij} \cdot \sum_{j=1}^{3} n'_{ji} \right) \equiv$ acuerdo observado
- $P_c = \ Accuracy \ = \sum_{i=1}^{3} n'_{ii} \equiv$ acuerdo debido a la casualidad

Nota: $n'_{ij} = n_{ij}/N$

Ejemplo de cálculo del índice kappa:
Sea la siguiente matriz de confusión normalizada para $N = 150$:

Reales	**Estimadas**		
	C_1	C_2	C_3
C_1	0,33	0	0
C_2	0	0,32	0,01
C_3	0	0,03	0,31

- $P_o = \text{Accuracy} = 0{,}33 + 0{,}32 + 0{,}31$ (diagonal principal)
- $P_c = 0{,}33 \cdot 0{,}33 + 0{,}33 \cdot 0{,}35 + 0{,}34 \cdot 0{,}32 = 0{,}33$
  Es la suma de los porcentajes reales por los estimados:
  - (suma de la fila 1) · (suma de la columna 1) $= 0{,}33 \cdot 0{,}33$
  - (suma de la fila 2) · (suma de la columna 2) $= (0{,}32 + 0{,}01) \cdot (0{,}32 + 0{,}03)$
  - (suma de la fila 3) · (suma de la columna 3) $= (0{,}03 + 0{,}31) \cdot (0{,}01 + 0{,}31)$
- $\kappa = \dfrac{P_o - P_c}{1 - P_c} = \dfrac{0{,}96 - 0{,}33}{1 - 0{,}33} = 0{,}94$

2.2.3. Matriz de confusión para dos calses

En el caso de la clasificación para dos clases, la matriz de confusión es:

	Estimadas	
Reales	+	−
+	VP	FN
−	FP	VN

- VP: Verdaderos positivos.
- FN: Falsos negativos.
- FP: Falsos positivos.
- VN: Verdaderos negativos.

A partir de esta matriz de confusión, se define los siguientes estadísticos:

- **Ratio de verdaderos positivos / sensibilidad / Recall**: Mide la capacidad del modelo de acertar los casos positivos.
$$RVP = \frac{VP}{VP + FN}$$
- **Ratio de falsos positivos**: mide la tasa de falsas alarmas del modelo.
$$RFP = \frac{FP}{FP + VN}$$
- **Ratio de verdaderos negativos / Especificidad**: mide la capacidad del modelo de acertar los casos negativos.
$$RVN = \frac{VN}{FP + VN}$$
- **Precisión / Valor predictivo positivo**: mide la tasa de aciertos entre todas las veces que se clasifica una instancia como positiva.
$$Precision = \frac{VP}{VP + FP}$$

- **Exactitud (Accuracy)**: mide la tasa de aciertos global del modelo.

$$Acuracy = \frac{VP + VN}{N}\ ; \quad [Recuerdo:\ N = VP + VN + FP + FN \equiv \text{total de elementos}]$$

2.3. Medidas de calidad en modelos de regresión

Al analizar modelos de regresión, no tiene sentido evaluar la calidad teniendo en cuenta el número de aciertos o fallos; es más interesante calcular la diferencia entre las predicciones del modelo y las de la función objetivo.

Tenemos una función objetivo (f) modelada por una hipótesis (h) y un conjunto de datos (D) con n elementos.

Algunas de las medidas que se usan son:

- **Error cuadrático medio**: realmente no nos ofrece una medida fiable de la magnitud del error.

$$ECM = \frac{1}{n}\sum_{x \in D}\left(h(x) - f(x)\right)^2$$

- **Raíz cuadrada del error cuadrático medio**: una mejor aproximación, aunque tanto esta medida como la anterior tiende a exagerar los efectos de los valores atípicos.

$$RECM = \sqrt{\frac{1}{n}\sum_{x \in D}\left(h(x) - f(x)\right)^2}$$

- **Error absoluto medio**:

$$EAM = \frac{1}{n}\sum_{x \in D}\left|h(x) - f(x)\right|$$

- **Error cuadrático relativo**: a esta medida se le puede aplicar todas las variantes anteriores.

$$ECM = \frac{1}{n}\sum_{x \in D}\frac{\left(h(x) - f(x)\right)^2}{\left(h(x) - \bar{f}\right)^2}\ ; \qquad \bar{f} = \frac{1}{n}\sum_{x \in D} f(x)$$

El error medio tiene un problema fundamental, y es que los datos desconocidos no tienen por que ser parecidos a los usados en el entrenamiento.

Para seleccionar qué medida emplear, hay que pensar qué estamos tratando de minimizar y cuál es el coste computacional de las distintas medidas.

Un buen modelo de regresión seguirá siendo bueno independientemente de la medida utilizada.

3. Estimación de la Eficacia del Modelo. Técnicas

En este punto, estamos trabajando con estimaciones. Una buena estimación de la medida de calidad nos permitirá comparar tanto la eficacia de distintos modelos entre sí, como de distintas configuraciones del mismo modelo.

No existe un concepto de 'mejor modelo'; la afirmación de que un modelo es bueno quiere decir lo bien que se ajusta a los datos usados (*No Free Lunch Theorem*).

El uso de un único conjunto de datos puede llevar a:

- **Overfitting**: la hipótesis se ajusta muy bien a la evidencia pero no es preciso con la nueva evidencia (es malo generalizando).
- **Underfitting**: los datos se ajustan muy mal a la evidencia y el modelo no aprende lo suficiente.

Para solventar esto, se divide la evidencia (muestra) en dos conjuntos: **entrenamiento** (para construir el modelo) y **test** (para evaluar la precisión del modelo).

3.1. Hold-out

Se divide de manera aleatoria los datos en conjunto de entrenamiento y prueba, usando 2/3 de los datos originales para el entrenamiento y el 1/3 restante para la prueba.

Es el método más usado para conjuntos de datos grandes, aunque el hecho de dividir los datos hace que se disponga de menos datos para construir el modelo, además de que el muestreo aleatorio puede introducir sesgos en los conjuntos obtenidos.

Hold-out estratificado: trata de mantener la distribución de las clases en cada conjunto.

Hold-out con repetición: se repite el proceso de Hold-out un cierto número de veces, de forma que en cada repetición los conjuntos de entrenamiento y prueba son distintos y la estimación final del estadístico se obtiene promediando los resultados de cada repetición.

Sin embargo, los distintos conjuntos de prueba se pueden solapar y podría ocurrir que algún dato nunca apareciera en un conjunto de entrenamiento.

3.2. Validación cruzada

Esta técnica permite evitar el solapamiento de los conjuntos de prueba.

3.2.1. Validación cruzada de k pliegues (k-*fold CV*)

Es un método eficiente cuando no se disponen de muchos datos. El procedimiento consiste en:

1. Dividir el conjunto de datos aleatoriamente en k subconjuntos del mismo tamaño. Normalmente, k se escoge entre 5 y 10 (a medida que k aumenta, el tamaño de los conjuntos de entrenamiento aumenta y del de prueba disminuye).

2. En cada iteración, uno de esos conjuntos (el k-ésimo) se reserva para la evaluación y el resto $(k-1)$ para el entrenamiento.

3. Finalmente, se agregan las diferentes estimaciones del estadístico.

3.2.2. Validación Cruzada dejando uno fuera (LOOCV)

Es un caso especial de la validación cruzada de k pliegues, donde k = número de elementos del conjunto de datos, por lo que en cada iteración el conjunto de prueba está conformado por únicamente un elemento (un dato).

Si bien se emplea solo cuando el conjunto de datos es muy pequeño, esto incrementa la posibilidad de encontrar modelos más precisos en ese tipo de casos.

3.2.3. Validación cruzada dejando un grupo fuera (*leave-group-out*)

Consiste en seleccionar para el conjunto de prueba varios elementos al mismo tiempo, de forma que se reduce el número de veces que hay que calcular el modelo.

3.3. Bootstrap

Esta técnica consiste en crear una cantidad de nuevos conjuntos de datos, del mismo tamaño que el original a partir de un muestreo aleatorio con sustitución. Pueden tener elementos duplicados y el conjunto de instancias no seleccionadas constituyen el conjunto de prueba.

La probabilidad de que un elemento no sea seleccionado nunca es de aproximadamente 36,8 % (es decir, el 63,2 % de los elementos está representado al menos una vez en algún conjunto de entrenamiento). Esto introduce un sesgo importante en los datos.

3.3.1. boot.632

Se redefine el estadístico: $E_s(h) = 0{,}632 \cdot E_{test} + 0{,}368 \cdot E_{training}$ (se da más peso al conjunto de prueba)

Este probablemente sea el mejor método cuando el conjunto de datos es muy pequeño.

3.4. Estimación del intervalo de confianza

Para una muestra con n ejemplos, se pueden establecer unos intervalos de confianza para el error verdadero $E_V(h)$ a partir del error de la muestra $E_S(h)$. Para $n > 30$, se puede usar:

$$E_V(h) = E_S(h) \pm z_c \sqrt{\frac{E_S(h)(1 - E_S(h))}{n}}$$

Donde z_c se obtiene a partir del nivel de confianza ($c\,\%$) según la tabla:

$c\,\%$	50 %	80 %	90 %	95 %	99 %
z_c	0,67	1,28	1,64	1,96	2,58

4. Ajuste de parámetros

Es necesario, para que los modelos funcionen lo mejor posible, determinar la mejor combinación de parámetros (model tunning).

Para ello, el conjunto de entrenamiento se vuelve a dividir en dos, de forma que ahora tenemos tres conjuntos distintos:

- **Entrenamiento**: para obtener los modelos.
- **Validación**: para estimar la precisión/error de los modelos de acuerdo con cada configuración de parámetros.
- **Prueba**: una vez obtenido el modelo con la mejor configuración de parámetros, se estima la precisión (o el error) del modelo para los datos no vistos en el conjunto de prueba.

5. Recomendaciones

- **Si el tamaño del conjunto de datos es pequeño**: repetición de validación cruzada de 10 pliegues (varianza y sesgo buenas y complejidad computacional aceptable).
- **Si queremos comparar modelos**: bootstrap (preferiblemente), ya que introduce menos variabilidad.
- **Para conjuntos de datos grandes**: no hay tanta diferencia a la hora de usar un método u otro, por lo que se usará el que menos complejidad computacional presente.