Evaluación de modelos

Minería de datos

Master Universitario en Tecnologías de Análisis de Datos Masivos
Escola Técnica Superior de Enxeñaría (ETSE)
Universidade de Santiago de Compostela

Contenidos de la presentación

- Introducción
- Medidas de calidad para la clasificación
  - Error del modelo
  - Medidas basadas en la matriz de confusión
- Medidas de calidad para la regresión
- Estimación de la eficacia del modelo
  - Hold-out
  - Validación cruzada
  - Validación cruzada dejando uno fuera
  - Bootstrap
  - Estimación del intervalo de confianza
- Ajuste de los parámetros del modelo

Introducción

- Supóngase que hemos generado un modelo a partir de un conjunto de datos:
  - ¿Cómo sabemos si el modelo es válido para nuestro propósito?
- Además:
  - ¿Cómo podemos evaluar la calidad de los modelos de forma lo más exacta posible?
  - ¿Cómo podemos comparar varios modelos entre sí?

Definición del problema

- El objetivo de las técnicas de aprendizaje automático es calcular una función objetivo f considerando un espacio de posibles hipótesis H.
  - Las distintas técnicas emplearán una evidencia o muestra S, formadas por ejemplos de f de acuerdo con una distribución D.
- A partir de una única muestra, lo más probable es que obtengamos un conjunto bastante grande de hipótesis distintas.

Medidas de calidad para la clasificación

- Las medidas más utilizadas para evaluar modelos de clasificación se basan en el error/exactitud de la hipótesis respecto a f.
- **Situación Ideal -** Disponer de un conjunto de ejemplos completos, o de la distribución de probabilidad de estos:

  Esto nos permitiría calcular el error verdadero $E_v(h)$:

$$E_v(h) = \frac{1}{|U|} \sum_{x \in U} \delta(f(x) \neq h(x))$$

  donde $\delta(\top) = 1$, $\delta(\bot) = 0$ y U representa el conjunto de todos los ejemplos posibles .

  Si no disponemos del conjunto U pero tenemos su distribución de probabilidad D:

$$E_v(h) = P_{x \in D}[\delta(f(x) \neq h(x))]$$

Error del modelo

- Sin embargo, normalmente solo disponemos de una muestra S de U, con lo que solo podemos calcular el error de muestra E_S de h.

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

A través de S podemos obtener la única muestra del comportamiento de f.

- Análogamente, la exactitud de la clasificación (*accuracy*) se puede medir como:

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

- Desde el punto de vista práctico, sean n el número total de instancias y n_c el número de instancias clasificadas correctamente:

$$Exactitud = \frac{n_c}{n} \qquad Error = \frac{n - n_c}{n}$$

Matriz de confusión

- Es una forma tabulada de visualizar el rendimiento de un modelo predictivo. Tiene la siguiente forma:

		Estimadas		
		C_1	C_2	C_3
Reales	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} indica el número de ejemplos que perteneciendo a la clase C_i han sido clasificados como la clase C_j.

Evaluación basada en el coste

- A partir de la matriz de confusión se pueden definir varias medidas de calidad del modelo que tienen la siguiente forma:

$$C(\epsilon) = \sum_{i=1}^{n}\sum_{j=1}^{n} n_{ij}c_{ij}$$

  donde c_{ij} es el coste asociado a cada elemento de la matriz de confusión.

- Por ejemplo, para calcular el error del modelo bastaría con definir la matriz de costes de la siguiente forma:

$$c_{ij} = \begin{cases} 1 & i \neq j \\ 0 & i = j \end{cases}$$

Problema: cuenta como favorables los aciertos debidos a la casualidad.

- Por ejemplo, para calcular la exactitud del modelo bastaría con definir la matriz de costes de la siguiente forma:

$$c_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

El coste se podría adaptar en base a la importancia del error en el problema.

Índice Kappa

- Para desvincular el acierto de la casualidad, se puede utilizar el índice kappa, que se calcula de la siguiente forma:

$$Kappa = \frac{P_0 - P_c}{1 - P_c}$$

donde P_0 es el acuerdo observado (*accuracy*), es decir, la exactitud del modelo y P_c es el acuerdo debido a la casualidad.

- La matriz de confusión se transformaría como sigue:

		Estimadas		
		C_1	C_2	C_3
Reales	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:

- $n'_{ij} = n_{ij}/N$
- $P_c = \sum_{i=1}^{n}\left(\sum_{j=1}^{n} n'_{ij} \cdot \sum_{j=1}^{n} n'_{ji}\right)$
- $P_o = \sum_{i=1}^{n} n'_{ii}$

Ejemplo – Índice Kappa

Dada la siguiente matriz de confusión normalizada para $N = 150$:

		Estimadas		
		C_1	C_2	C_3
Reales	C_1	0.33	0	0
	C_2	0	0.32	0.01
	C_3	0	0.03	0.31

Calcular el índice Kappa.

$$P_0 = 0.33 + 0.32 + 0.31 = 0.96$$

$$P_C = 0.33 \cdot 0.33 + 0.33 \cdot 0.35 + 0.34 \cdot 0.32 = 0.33$$

$$Kappa = \frac{Po - Pc}{1 - Pc} = \frac{0.96 - 0.33}{1 - 0.33} = 0.94$$

Matriz de confusión para 2 clases

- Para un problema de 2 clases, la matriz de confusión tiene la siguiente tabla:

		Estimadas	
		$+$	$-$
Reales	$+$	VP	FN
	$-$	FP	VN

El número de elementos viene determinado por $N = VP + FN + FP + VN$

donde:

- VP (verdaderos positivos) se refiere al número de predicciones donde el clasificador predice correctamente la clase positiva como positiva.
- VN (verdaderos negativos) se refiere al número de predicciones donde el clasificador predice correctamente la clase negativa como negativa.
- FP (falsos positivos) se refiere al número de predicciones donde el clasificador predice incorrectamente la clase negativa como positiva.
- FN (falsos negativos) se refiere al número de predicciones donde el clasificador predice incorrectamente la clase positiva como negativa.

Estadísticos para 2 clases

- A partir de la matriz de confusión se pueden definir los siguientes estadísticos:
  - Ratio de verdaderos positivos (sensibilidad - *recall*). Capacidad para acertar los casos positivos:

$$RVP = \frac{VP}{VP + FN}$$

  - Ratio de falsos positivos. Tasa de falsas alarmas del modelo:

$$RFP = \frac{FP}{FP + VN}$$

  - Ratio de verdaderos negativos (especificidad - *specificity*). Capacidad del modelo para acertar los casos negativos:

$$RVN = \frac{VN}{FP + VN}$$

  - Valor predictivo positivo (precisión - *precision*). Capacidad del modelo para acertar los casos negativos:

$$VPP = \frac{VP}{FP + VP}$$

Estadísticos para 2 clases

- A partir de la matriz de confusión se pueden definir los siguientes estadísticos:
  - Fß-score. La medida armónica entra la precisión y el *recall*:

$$F\beta-score = (1+\beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

    El más usado es el F1-score:

$$F1-score = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

  - Exactitud (*accuracy*) Tasa de aciertos global del modelo:

$$Exactitud = \frac{VP + VN}{N}$$

  - Kappa. Ya visto anteriormente.

Ejemplo – Matriz de confusión

Dada la siguiente matriz de confusión:

		Estimadas		
		$Manzanas$	$Naranjas$	$Mangos$
Reales	$Manzanas$	7	1	3
	$Naranjas$	8	2	2
	$Mangos$	9	3	1

Calcular la precisión, recall y F1.

Ahora ya no hay solo clases positivas y negativas.

Sin embargo, se puede hacer igualmente para cada clase por separado.

Calculemos los valores para las manzanas:

- $VP = 7$
- $VN = (2 + 3 + 2 + 1) = 8$
- $FP = (8 + 9) = 17$
- $FN = (1 + 3) = 4$

Entonces:

- $precision = 7/(7 + 17) = 0.29$
- $recall = 7/(7 + 4) = 0.64$
- $F1 = 2 \cdot (0.29 \cdot 0.64)/(0.29 + 0.64) = 0.40$

Ejemplo – Matriz de confusión

Si hacemos el cálculo para las demás clases:

¿Qué hacemos ahora con 3 medidas de F1?

	$Precision$	$Recall$	$F1-score$
$Manzanas$	0.29	0.64	0.40
$Naranjas$	0.33	0.17	0.22
$Mango$	0.17	0.08	0.11

Las medidas de $F1\text{-}score$ de cada una de las clases se puede combinar para tener una medida del modelo completo. Existen varias formas de hacerlo:

Micro F1 (*micro-averaged F1-score*). Se calcula considerando los VP, FP y FN totales. No considera las clases individuales.

- $Total\ TP = (7 + 2 + 1) = 10$
- $Total\ FP = (8 + 9) + (1 + 3) + (3 + 2) = 26$
- $Total\ FN = (1 + 3) + (8 + 2) + (9 + 3) = 26$

Ejemplo – Matriz de confusión

A partir de lo anterior podemos calcular la precisión y el recall:

- $Precision = 10/(10+26) = 0.28$
- $Recall = 10/(10+26) = 0.28$

Y aplicando la fórmula de la F1, podemos calcular la Micro F1:

- $Micro\ F1 = 0.28$

Cuando calculamos las métricas anteriores de forma global, todas las medidas se vuelven iguales:

$$Precision = Recall = MiF1 = Accuracy$$

Macro F1 (*macro-averaged F1-score*). Se calculan las métricas para cada clase de forma individual y se hace la media no pesada de las medidas.

- $F1\ Manzanas = 0.40$
- $F1\ Naranjas\ Total\ FP = 0.22$
- $F1\ Mangos = 0.11$

- $Macro\ F1 = \frac{0.40+0.22+0.11}{3} = 0.24$

F1 Pesada (*weighted-averaged F1-score*). Al contrario que la Macro F1, calcula la media pesada de las medidas. El peso de cada clase es igual al total de muestras de cada clase.

- $F1\ Pesada = \frac{(0.40 \cdot 11)+(0.22 \cdot 12)+(0.11 \cdot 13)}{(11+12+13)} = 0.24$

Medidas de calidad para la regresión

- En el análisis de modelos de regresión no tiene sentido evaluar la calidad teniendo en cuenta el número de aciertos/fallos.
- En los modelos de regresión es más interesante calcular la diferencia entre las predicciones del modelo y las de la función objetivo.
- Supongamos que tenemos una función objetivo f modelada mediante una hipótesis h y un conjunto de datos D con n elementos.
- Una de las medidas más utilizadas el error cuadrático medio (*mean squared error*):

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

- El problema del MSE es que no ofrece una medida fidedigna de la magnitud del error.

La diferencia entre la predicción y el valor real se eleva al cuadrado para eliminar el signo y tener siempre un error positivo.

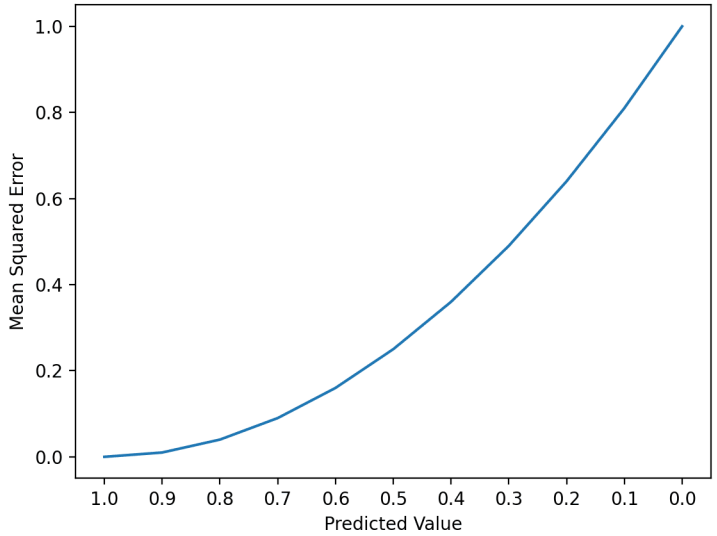
Elevar al cuadrado también tiene el efecto de magnificar el error. A mayor diferencia, mayor el error.

Mean Squared Error (MSE)

Ejemplo: Supongamos que un modelo predictivo que obtiene los siguientes valores respecto a los valores reales:

esperados	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
predichos	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
$(h(x) - f(x))^2$	0.000	0.010	0.040	0.090	0.160	0.250	0.360	0.490	0.640	0.810	1.000

Calculamos el MSE en cada punto



En este ejemplo, podemos apreciar que la curva crece exponencialmente a medida que crece el error

En este caso, el MSE sería igual a 0.35

Root Mean Squared Error (RMSE)

- Para obtener una mejor aproximación al error se suele utilizar la raíz cuadrada del error cuadrático medio (*root mean squared error*):

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

- Al hacer la raíz cuadrada, las unidades del error están en las mismas unidades que los datos originales.
  - Por ejemplo, si nuestros datos están en euros, el error también estará en euros.
- Esta característica hace que normalmente se use el MSE en entrenamiento, mientras que el RMSE en la evaluación para informar al cliente el error.
- En el ejemplo anterior, el $RMSE$ sería igual a 0.5916

Mean Abstolute error (MAE)

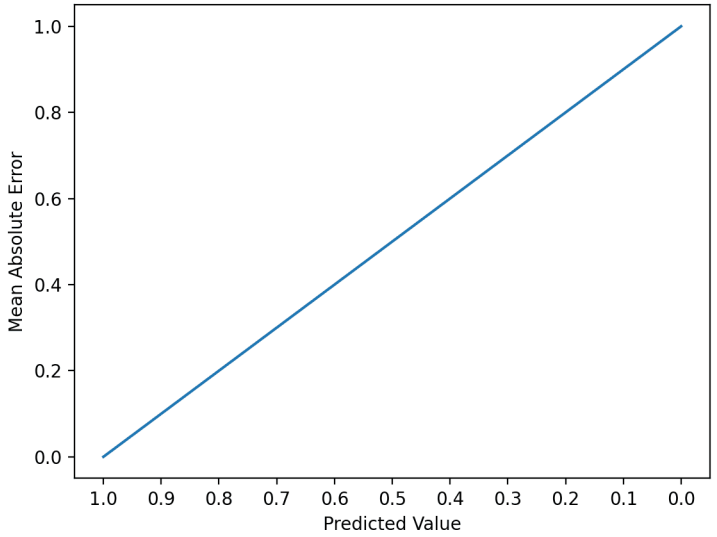
- El MSE y el $RMSE$ suelen exagerar el efecto de los errores más extremos (*outliers*).
- Para evitar esto se suele utilizar el error absoluto medio (***mean absolute error***):

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

- Al igual que el $RMSE$, las unidades del error coinciden con las del problema.
- Al contrario del $RMSE$, los cambios son lineales y más intuitivos.
- Volviendo al ejemplo anterior:

$\|h(x) - f(x)\|$	0.000	0.100	0.200	0.300	0.400	0.500	0.600	0.700	0.800	0.900	1.000

Podemos apreciar que la curva ahora es línea



- En el ejemplo anterior, el MAE sería igual a 0.5

Evaluación de la eficacia de un modelo

- Una sola ejecución de nuestro algoritmo no suele ser suficiente para medir su eficacia:
  - La muestra S es pequeña.
  - Existen efectos aleatorios que afectan la construcción del modelo.
  - Nos interesa obtener distintas estimaciones de la misma medida para determinar una estimación estadísticamente significativa.

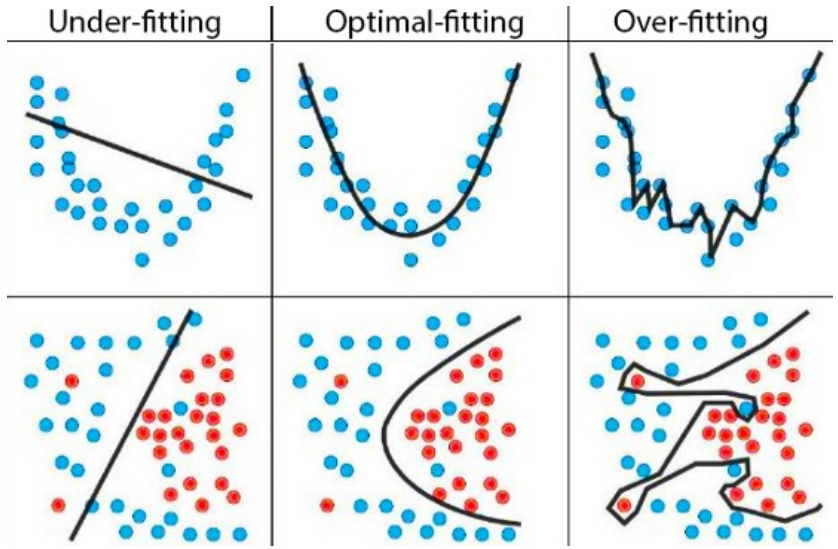
- Una buena estimación de la medida de calidad nos permitiría comparar la eficacia de diferentes modelos entre sí o de diferentes configuraciones del mismo modelo.

- Se podría utilizar la evidencia completa S para construir el modelo y para su posterior evaluación.

  Problema 1: El error medido sobre los datos utilizados para construir el modelo no es un buen indicador sobre cómo se comportará en el futuro con otros datos (capacidad de generalización).
  - Los nuevos datos no tienen por qué ser parecidos a los utilizados en el entrenamiento.

Evaluación de la eficacia de un modelo

- Se podría utilizar la evidencia completa S para construir el modelo y para su posterior evaluación.



Problema 2: Cualquier conclusión que obtengamos estará sujeta al conjunto de datos usado para construir el modelo (sesgada a los datos):

- Los resultados son difícilmente generalizables.
- No existe el concepto de "mejor modelo".
  - Un modelo se puede adaptar bien a un conjunto de datos, pero tener un mal comportamiento en otro.

Problema 3: Se puede incurrir en un sobreajuste (*over-fitting*):

La hipótesis se ajusta muy bien a la evidencia, pero no es preciso con la nueva evidencia.
No hay que centrarse en las particularidades de los datos sino pensar en las generalidades.

Problema 4: Para solucionar el sobreajuste podemos pensar en reducir la evidencia en aras de buscar una mayor generalización, pero podemos caer en subajuste (*under-fitting*).

¿Cómo podemos evitar el ajuste a los datos?

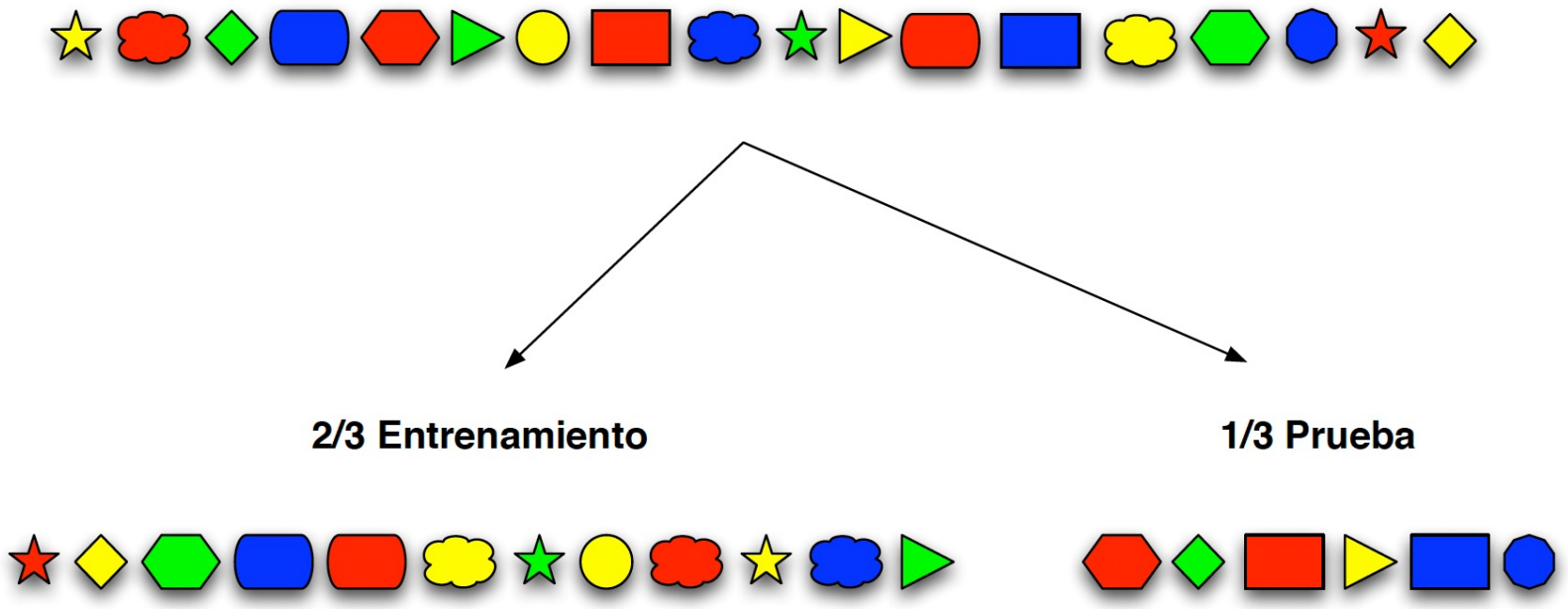
- La solución consiste en dividir la evidencia en dos conjuntos:
  - Entrenamiento (*train*): para construir el modelo.
  - Prueba (*test*): para evaluar la precisión del modelo.

- Asunción: Ambos conjuntos son muestras representativas del problema a modelar.

- Existen diferentes técnicas basadas en este paradigma:
  - Hold-out
  - Validación cruzada (*cross-validation*)
  - Validación cruzada dejando uno fuera (*leave-one-out cross-validation*)
  - Bootstrap

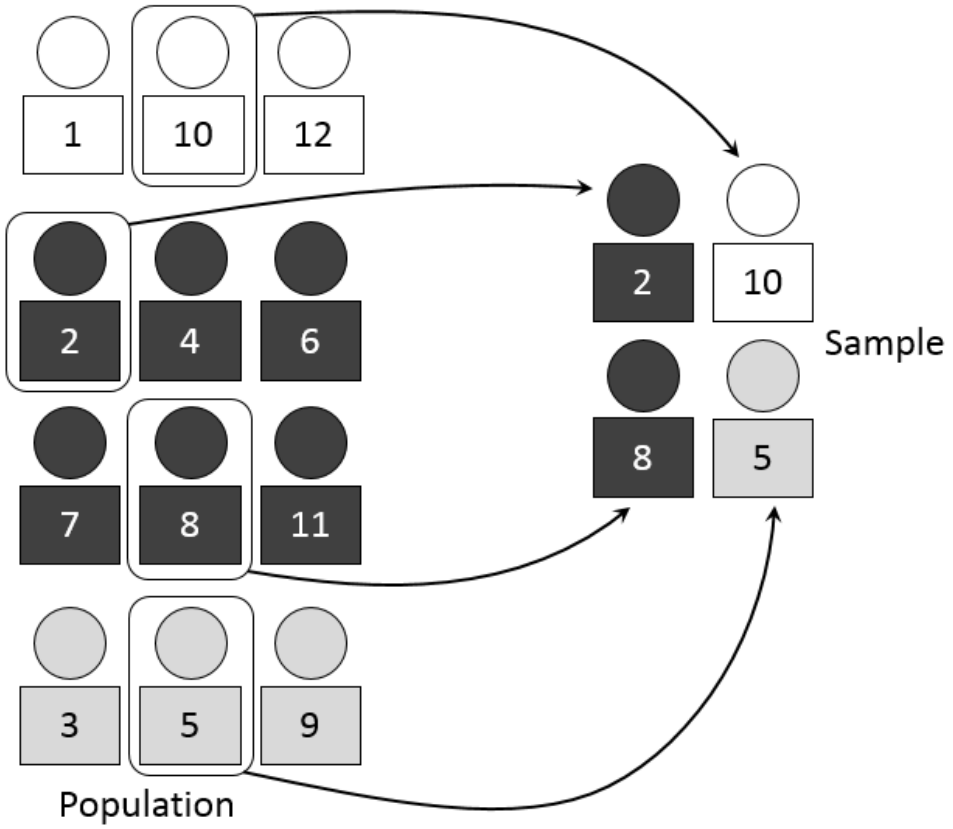
Hold-out

- Es el método más utilizado cuando se tiene un conjunto de datos grande.
- Consiste en dividir de forma aleatoria el conjunto de datos en dos conjuntos: entrenamiento y prueba.
  - Normalmente 2/3 para entrenamiento y 1/3 para prueba.



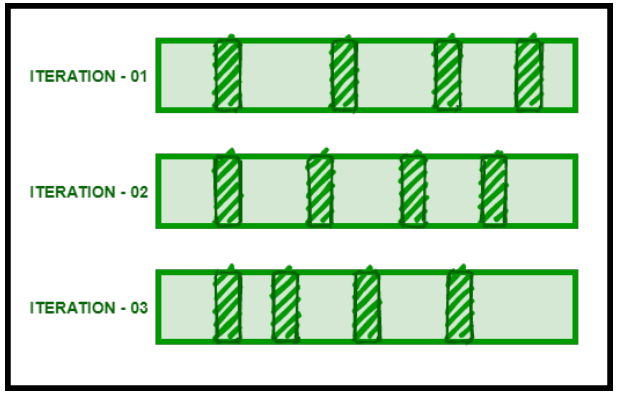
Hold-out

- **Problemas**:
  - Disponemos de **menos datos para construir el modelo**.
  - El muestreo aleatorio puede introducir **sesgos** en los conjuntos obtenidos.
- Para resolver este problema, la variante **hold-out estratificado** intenta mantener la distribución de clases en cada conjunto.

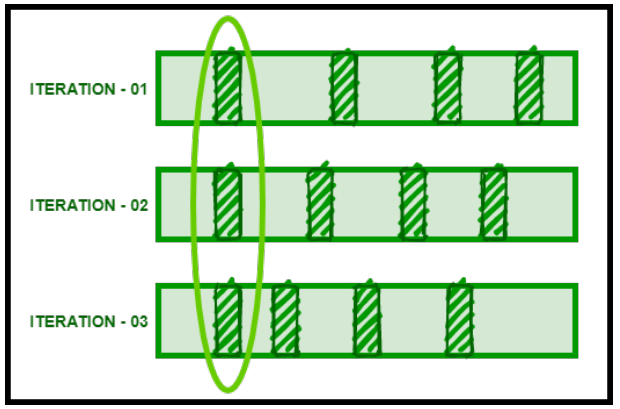


Hold-out con repetición

- Se repite el hold-out un cierto número de veces.
- El muestreo aleatorio hace que en cada repetición los conjuntos de entrenamiento y prueba sean distintos.
- La estimación final del estadístico se obtiene promediando los resultados de cada repetición.



- **Problemas**:
  - Los diferentes conjuntos de prueba se pueden solapar.
  - Puede ocurrir que algunos datos nunca aparezcan en un conjunto de entrenamiento.

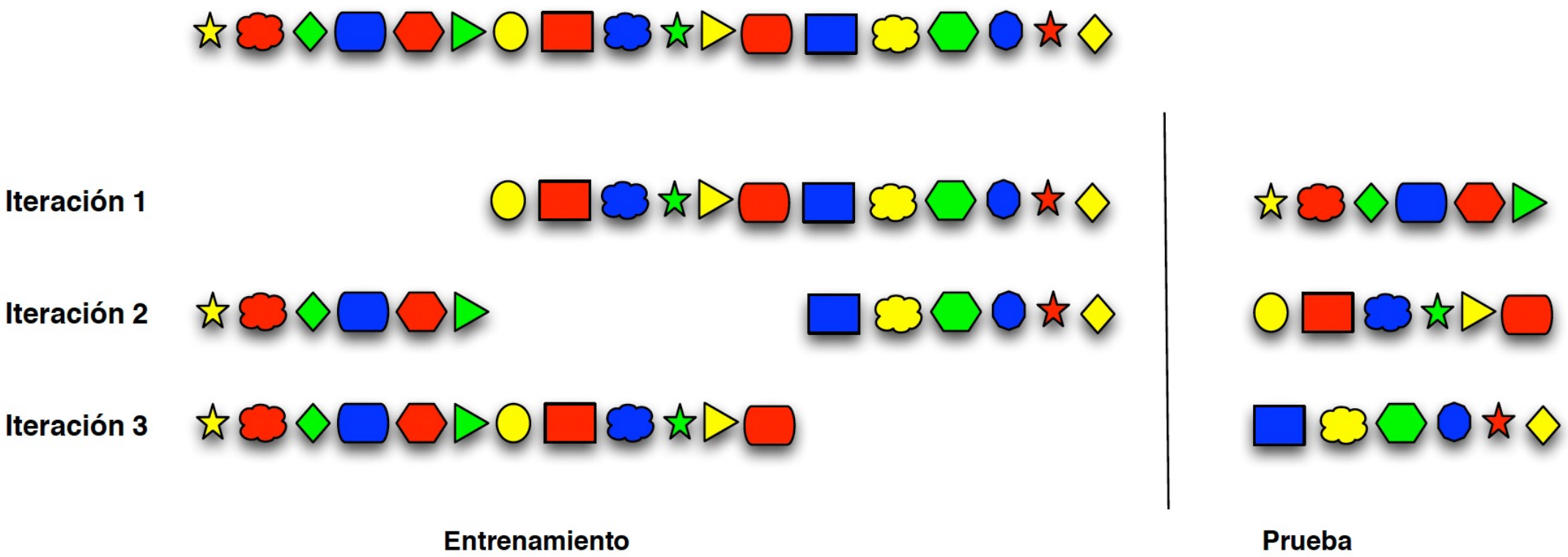


K-fold cross-validation

- La Validación cruzada evita el solapamiento de los conjuntos de prueba.
  1. Se divide el conjunto de datos aleatoriamente en k subconjuntos disjuntos del mismo tamaño.
  2. En cada iteración, uno de esos conjuntos se reserva para la evaluación y el resto se utiliza para el entrenamiento.
  3. Al final, se agregan las diferentes estimaciones del estadístico (media y varianza).

- Validación cruzada con k pliegues (*k-fold cross-validation*)
  - Suele ser eficiente cuando no se disponen de muchos datos.
  - k se suele elegir entre 5 y 10.
  - A medida que k aumenta, el tamaño de los conjuntos de entrenamiento y prueba se hace más pequeño.
    - Esto mejora la estimación del estadístico.
  - Resultados experimentales muestran que $\boldsymbol{k = 10}$ es una buena opción
  - En comparación con otros métodos, presenta una alta variabilidad. Para reducirla:
    - Hacer un muestreo estratificado (*stratified k-fold cross validation*).
    - Repetir el proceso de validación cruzada (*repeated k-fold cross validation*) (con estratificación).

Ejemplo de 3-fold cross-validation

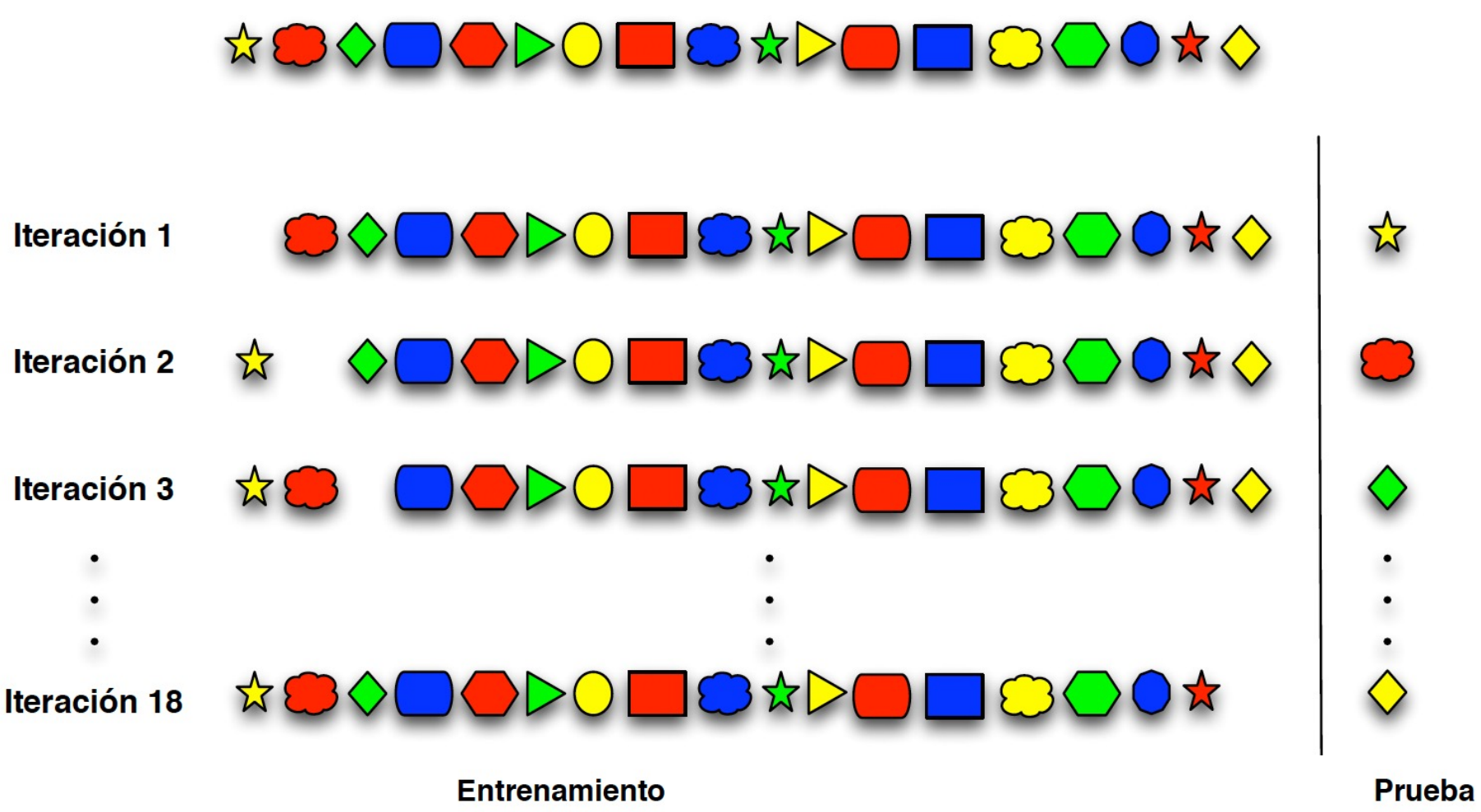


Leave-one-out cross-validation

- Es un caso especial de validación cruzada con k igual al número de elementos en el conjunto de datos.
- En cada iteración se reserva un elemento para evaluar el modelo.
- Se utiliza cuando el conjunto de datos es muy pequeño.
- Hace un mejor uso del conjunto de datos.
- Incrementa la posibilidad de encontrar modelos más precisos.
- Evita los inconvenientes de un muestreo aleatorio.
- Los Inconvenientes:
  - Muy costoso computacionalmente (número de modelos igual al número de elementos en el conjunto de datos).
  - Muestra resultados similares a una validación cruzada con 10 pliegues, pero es computacionalmente más ineficiente.

La variante leave-one-group-out deja un grupo fuera en cada iteración:
- Reduce el número de modelos

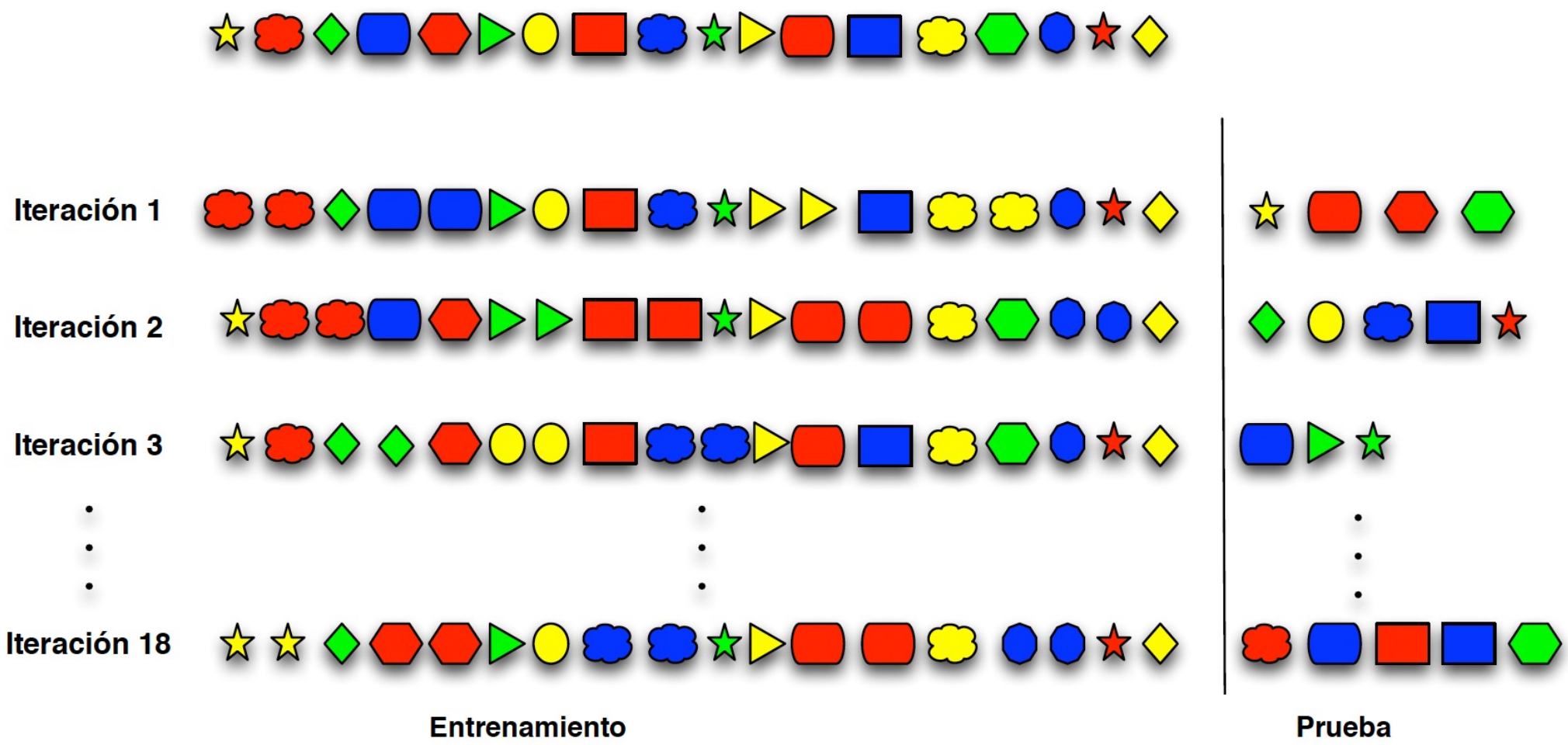
Ejemplo de leave-one-out cross-validation



Bootstrap

- Hasta ahora los métodos analizados aplicaban un muestreo sin sustitución.
  - Una vez seleccionado un elemento, este ya no puede volver a ser seleccionado.
  - No existen duplicados.
- Bootstrap se basa en un muestreo con sustitución.
- Se crea un nuevo conjunto de datos, del mismo tamaño que el original, mediante un muestreo aleatorio con sustitución.
- Este conjunto se utiliza para crear el modelo (pueden existir elementos duplicados).
- El conjunto de instancias no seleccionadas constituye el conjunto de prueba.
- El proceso se repite un número determinado de veces.

Ejemplo de bootstap



Bootstrap

- La estimación del estadístico tiende a ser pesimista:
  - La probabilidad de que un elemento no sea seleccionado nunca es:

$$\lim_{n\to\infty}\left(1-\frac{1}{n}\right)^{n} = e^{-1} \approx 0.368$$

  - Por lo tanto, el 63.2% de los elementos está representado al menos una vez en algún conjunto de entrenamiento.
- Este hecho introduce un sesgo importante en los datos.
  - Tiende a ser importante cuando hay pocos datos, reduciéndose a medida que el conjunto de datos se hace más grande.

- boot.632 mitiga este problema redefiniendo el estadístico como:

$$E_S\,(h) \;=\; 0.632 \;\cdot E_{test} \;+\; 0.368 \cdot E_{training}$$

- En cada iteración se le da más peso al error en el conjunto de prueba.
- Probablemente sea el mejor método cuando el conjunto de datos es muy pequeño.

Intervalo de confianza

- ¿Es fiable el valor del estadístico obtenido por las técnicas de remuestreo anteriores?
- Se pueden establecer intervalos de confianza para el error verdadero $E_v(h)$, a partir del error de la muestra $E_S(h)$, en una muestra con n ejemplos.
- Para ello, el intervalo de error con un nivel de confianza c % es (se suele utilizar la distribución binomial, pero para $n > 30$ se puede utilizar la distribución normal):

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

- Donde z_c se establece a partir del nivel de confianza según la normal:

c %	50 %	80 %	90 %	95 %	100 %
z_c	0.67	1.28	1.64	1.96	2.58

Recomendaciones

- No se puede afirmar que un método de muestreo sea mejor que otro.
- Si el tamaño del conjunto de datos es pequeño se recomienda la repetición de validación cruzada con 10 pliegues (*repeated 10-fold cross validation*):
  - Las propiedades de varianza y sesgo son buenas.
  - La complejidad computacional es adecuada para el tamaño del conjunto de datos.
- Debido a la variabilidad, si lo que queremos es comparar modelos es preferible algún método de bootstrap.
  - Introduce menos variabilidad.
- Para conjunto de datos grandes:
  - Elegir el de menor complejidad computacional.

Ajuste de parámetros

- En muchas ocasiones, los modelos requieren de unos parámetros para poder funcionar
- Los parámetros pueden tener una gran influencia sobre el resultado final:
  - En el MLP: la tasa de aprendizaje y el número de neuronas en la capa intermedia.
  - En una SVM: el coste.
- A parte de estimar la exactitud/error del modelo es necesario determinar la mejor combinación de parámetros (model tunning).
- Normalmente nos quedaremos con la configuración que mejor estimador obtiene.
- Para ello el conjunto de entrenamiento se vuelve a dividir en dos:
  - Entrenamiento: para obtener los modelos.
  - Evaluación: para estimar la exactitud/error del modelo de acuerdo con una configuración de los parámetros.
- Una vez obtenido el modelo con la mejor configuración de los parámetros, se puede estimar la exactitud/error del modelo para el conjunto de prueba.

Ajuste de parámetros

Algoritmo Ajuste de parámetros

1: Definir conjuntos de diferentes valores de los parámetros a ajustar.
2: **para** Cada conjunto de valores **hacer**
3: {Aplicar una técnica de remuestreo sobre el conjunto de entrenamiento.}
4: **para** Cada iteración de remuestreo **hacer**
5: Crear un conjunto de evaluación;
6: Entrenar el modelo;
7: Estimar uno o varios estadísticos sobre el conjunto de evaluación;
8: **fin para**
9: Calcular la estimación del estadístico como promedio de todas las iteraciones;
10: **fin para**
11: Determinar la mejor configuración de parámetros;
12: Estimar el estadístico sobre el conjunto de prueba;

Conclusiones

- En este capítulo hemos revisado un gran número de indicadores para medir la eficacia de los clasificadores y modelos de regresión.
  - En la mayoría de los casos, la medida a utilizar vendría determinada por el problema.
- Para poder estimar dichos indicadores hemos presentado distintas técnicas de evaluación: hold-out, validación cruzada y bootstrap.
- Otro aspecto importante que hemos analizado es el del ajuste de los parámetros del modelo, es decir, ¿cómo determinar cuál es la mejor configuración del modelo?

Bibliografía:

- Ethem Alpaydin. Introduction to Machine Learning. MIT Press 2004.
- Max Khun and Kjell Johnson. Applied Predictive Modeling. Springer.
- Tom Fawcett. An introduction to ROC analysis. Pattern Recognition Letters 27 (2006) 861-874.
- José Hernández Orallo, Mª José Ramírez Quintana and César Ferri Ramirez. Introducción a la Minería de Datos. Pearson-Prentice-Hall. 2004
- Ian H. Witten, Eibe Frank, and Mark A. Hall. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers.