Evaluación de Clasificadores

Minería de Datos

José T. Palma

Departamento de Ingeniería de la Información y las Comunicaciones
Universidad de Murcia

DIIC, UMU, 2021

Contenidos de la presentación

Introducción

- Hasta ahora hemos presentado varias técnicas que nos permiten generar modelos a partir de un conjunto de datos.
- ¿Cómo sabemos si el modelo es válido para nuestro propósito?
    - Necesitamos evaluar la calidad de los modelos de forma lo más exacta posible.
- ¿Cómo podemos comparar varios modelos entre sí?

Definición del problema

- Como ya hemos visto el objetivo de las técnicas de aprendizaje automático es calcular una función objetivo f (la función que predice la clase) 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.
- Como hemos visto, a partir de una única evidencia podemos obtener un conjunto bastante grande de hipótesis distintas.
- Necesitamos alguna medida sobre la calidad del modelo.

Exactitud/Error de predicción

- Las medidas más utilizadas para evaluar clasificadores se basan en la exactitud de la hipótesis, o su error, respecto a f.
- **Situación Ideal:** disponer de un conjunto de ejemplos completos, o la de distribución de probabilidad de los mismos.
  - Esto nos permitiría calcular el error verdadero $E_v(h)$

$$E_v(h) = \frac{1}{|U|} \sum_{x \in U} \delta(f(x) \neq h(x)) \; ; \; \delta(verdadero) = 1 \, , \, \delta(falso) = 0$$

  - donde U representa el conjunto de todos los ejemplos posibles.
  - Si no disponemos del conjunto U pero tenemos la distribución de probabilidad D:

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

Exactitud/Error de predicción

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

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

- La única muestra del comportamiento de f sólo se puede obtener a partir de la evidencia S.
- Análogamente, la exactitud de la clasificación se puede medir cómo:

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

Exactitud/Error de predicción

- Desde un punto de vista más práctico, sean
  - n el número total de instancias
  - n_c el número total de instancias clasificadas correctamente.

Exactitud (Accuracy)	Error de clasificación
$\frac{n_c}{n}$	$\frac{n - n_c}{n}$

- Existen otras medidas que están basadas en el análisis de la matriz de confusión.

Matriz de Confusión

- Una matriz de confusión 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 como:

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

Evaluación basada en el coste

- Para obtener una medida de la exactitud del modelo bastaría con definir la matriz de costes como

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

- Además, la matriz de costes c_{ij} se puede adaptar a cualquier problema en el que los distintos tipos de error/aciertos tengan distinta importancia.
- Aunque en la mayoría de casos, encontrar la matriz de costes sea complicado.

Índice Kappa

- La exactitud del modelo, tal y como la hemos definido anteriormente, tiene le problema de que también cuenta como favorables los aciertos debidos a la casualidad.
- Para resolver este problema podemos utilizar el índice **kappa**, que se calcula de las siguiente forma:

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

- donde P_o es el acuerdo observado, es decir, la exactitud del modelo y P_c es el acuerdo debido a la casualidad.

Índice Kappa

- Sea la siguiente matriz de confusión:

		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 $n'_{ij} = n_{ij}/N$
- En este caso:

$$P_c = \sum_{i=1}^{3} \left(\sum_{j=1}^{3} n'_{ij} \cdot \sum_{j=1}^{3} n'_{ji} \right) \qquad P_o = Accuracy = \sum_{i=1}^{3} n'_{ii}$$

Índice Kappa: Ejemplo

- Sea 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

$$P_0 = Acurracy = 0,33 + 0,32 + 0,31 = 0,96$$

$$P_C = 0,33 * 0,33 + 0,33 * 0,35 + 0,34 * 0,32 = 0,33$$

$$kappa = \frac{P_o - P_c}{1 - P_c} = \frac{0,96 - 0,33}{1 - 0,33} = 0,94$$

Matriz de Confusión para dos clases

- Para un problema con dos clases la matriz de confusión tiene la siguiente forma:

		Estimadas	
		+	-
Reales	+	VP	FN
	-	FP	VN

  - VP: Verdaderos positivos.
  - FN: Falsos negativos.
  - FP: Falsos positivos.
  - VN: Verdaderos negativos.
- El número de elementos en el conjunto viene determinado por $N = VP + FN + FP + VN$.

Matriz de Confusión para dos clases I

- A partir de la matriz de confusión podemos definir los siguiente estadísticos:
- **Ratio de verdaderos positivos, Sensibilidad, Recall**: Mide la capacidad para acertar los casos positivos

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

- **Ratio de falsos positivos**: Mide la tasa de falsas alarmas del modelo

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

Matriz de Confusión para dos clases II

- **Ratio de verdaderos negativos, Especificidad**: Mide la capacidad del modelo para acertar los casos negativos

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

- **Precisión, Valor predictivo postivo**: Mide la tasa de aciertos entre todas las veces que se clasifica una instancia como positiva

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

- **F-score**: la media armónica entre la precisión y el recall:

$$F - score = \frac{2 \cdot Precicion \cdot Recall}{Precision + Recall}$$

Matriz de Confusión para dos clases III

- **Exactitud**: Mide la tasa de aciertos global del modelo:

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

- El problema de la exactitud es que no tiene en cuenta los aciertos debidos a la casualidad.
- Si queremos evitar esto debemos utilizar el índice **Kappa**

Medidas en modelos de 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 de evaluación más utilizadas el **Error Cuadrático Medio**:

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

Medidas en modelos de regresión

- El problema del *ECM* es que no nos ofrece una medida fidedigna de la magnitud del error.
- Para obtener una mejor aproximacion al error se suele utilizar la **Raíz Cuadrada del Error Cuadrático Medio**:

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

- Sin embargo estas medidas tienden a exagerar el efecto de los errores más extremos (outliers). Para evitar esto se suele utilizar el **Error Absoluto medio**:

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

Medidas en modelos de regresión

- En algunos casos lo que nos interesa es el error relativo en cuyo caso se utiliza el **Error cuadrático relativo**:

$$ECR = \frac{1}{n}\sum_{x\in D}\frac{(h(x)-f(x))^2}{(h(x)-\bar{f})^2} \text{ donde } \bar{f} = \frac{1}{n}\sum_{x\in D} f(x)$$

- A esta medida se le puede aplicar todas las variantes que anteriormente hemos visto: raíz cuadrada o utilizar el valor absoluto.

Medidas en modelos de regresión

- La elección de la medida a utilizar depende del problema que estamos tratando:
  - ¿Qué estamos tratando de minimizar?
  - ¿Cuál es el coste de los diferentes tipos de error?
- Las medidas basadas en el cuadrado del error tienden a dar mas importancia a las grandes discrepancias frente a las pequeñas.
- Al utilizar la raíz cuadrada sólo acercamos la magnitud del error a las cantidades que están siendo predichas.
- Las medidas basadas en los errores relativos tienen a compensar la predictibilidad o impredicibilidad de la variable de salida.
- En la práctica un buen modelo de regresión seguirá siendo igualmente bueno independientemente de la medida utilizada.

Estimación de la exactitud/error del modelo

- Como ya hemos mencionado, las anteriores medidas sólo las podemos obtener a partir de la muestra disponible, por lo tanto, estamos trabajando con estimaciones.
- Además, la estimación de la medida obtenida a partir de la evidencia utilizada en el entrenamiento (error de entrenamiento).
- Dicha medida no se puede utilizar para determinar el comportamiento del modelo (demasiado optimista)

Motivación

- ¿Cómo podemos estimar la eficacia del modelo cuando lo utilicemos en fase de producción?
- Puede que una sola ejecución del algoritmo no sea suficiente:
  - La muestra S es pequeña.
  - Existen factores aleatorios que afectan a la construcción del modelo.
  - Nos puede interesar 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.
  - entre diferentes configuraciones del mismo modelo.

Motivación

- Como primera opción podemos utilizar toda la evidencia completa, S, para construir el modelo y para su posterior evaluación.
- El problema el error medido sobre los datos utilizados para construir el modelo no es un buen indicador sobre cómo se comportará el modelo en el futuro.
  - Datos desconocidos no tiene porque ser parecidos a los utilizados en el entrenamiento.

Motivación

- Cualquier conclusión que obtengamos estará sujeta al conjunto de datos usado para construir el modelo:
  - Los resultados son difícilmente generalizables.
  - No existe el concepto de "mejor modelo".
    - Para cada modelo existirá un conjunto de datos para el que es muy bueno y otro para el que el comportamiento es malo.
  - Cuando afirmamos que un modelo es bueno, estamos diciendo lo bien que se ajusta al sesgo inductivo de los datos utilizados (No Free Lunch Theorem).

Motivación

- La utilización de un único conjunto de datos para estimar la calidad del modelo puede incurrir en:
  - *Sobre-aprendizaje (overfitting):* La hipótesis se ajusta muy bien a la evidencia pero no es preciso con la nueva evidencia.
    - La idea es no centrarse en las particularidades de los datos sino pensar en las generalidades.
  - Para solucionarlo podemos pensar en reducir la evidencia en aras de buscar una mayor generalización, pero podemos caer en *sub-aprendizaje* (underfitting).
- ¿Cómo podemos evitar ese ajuste a los datos utilizados?

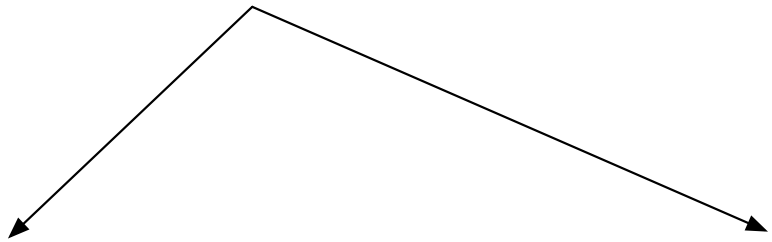
Motivación

- La solución consiste en dividir la evidencia en dos conjuntos : "Entrenamiento (training) y prueba (test)":
  - Entrenamiento: para construir el modelo.
  - Prueba: 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)
  - Bootstrap

Hold-out

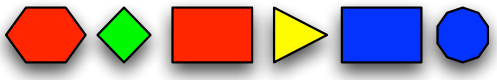
- Es el método más utilizado cuando se tiene un conjunto de datos grande.
- Dividir de forma aleatoria el conjunto de datos en dos conjuntos: *entrenamiento* y *prueba*
  - Normalmente 2/3 para entrenamiento y 1/3 para prueba.
- **Inconvenientes:**
  - Disponemos de menos datos para construir el modelo.
  - El muestreo aleatorio puede introducir sesgos en los conjuntos obtenidos.
- Para resolver este problema **Hold-out stratificado**: se intenta mantener la distribución de clases en cada conjunto

Hold-out



2/3 Entrenamiento

1/3 Prueba



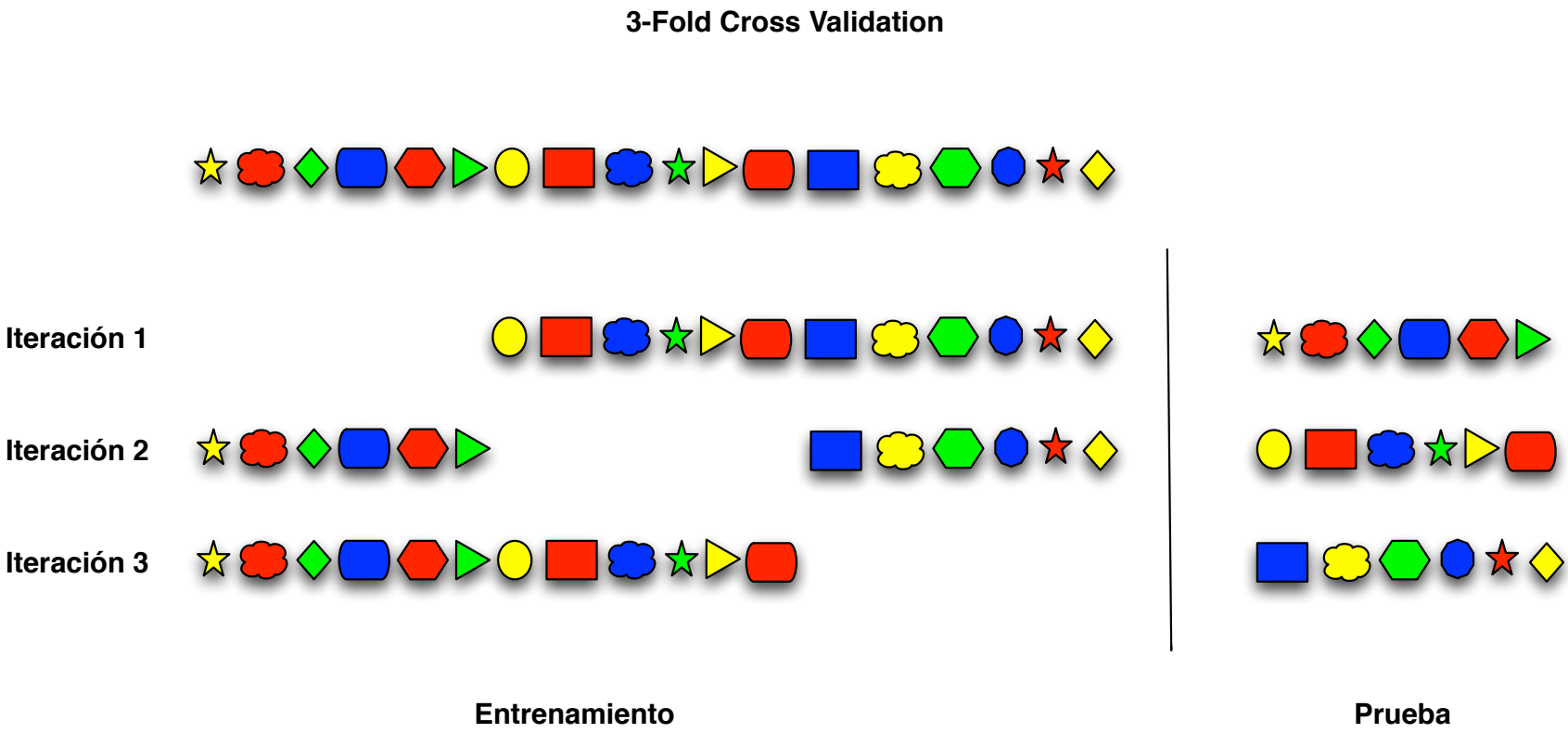
Hold-out

- **Hold-out con repetición:** ser repite el proceso 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 algún dato nunca aparezca en un conjunto de entrenamiento.

validación cruzada

- 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.

Validación cruzada



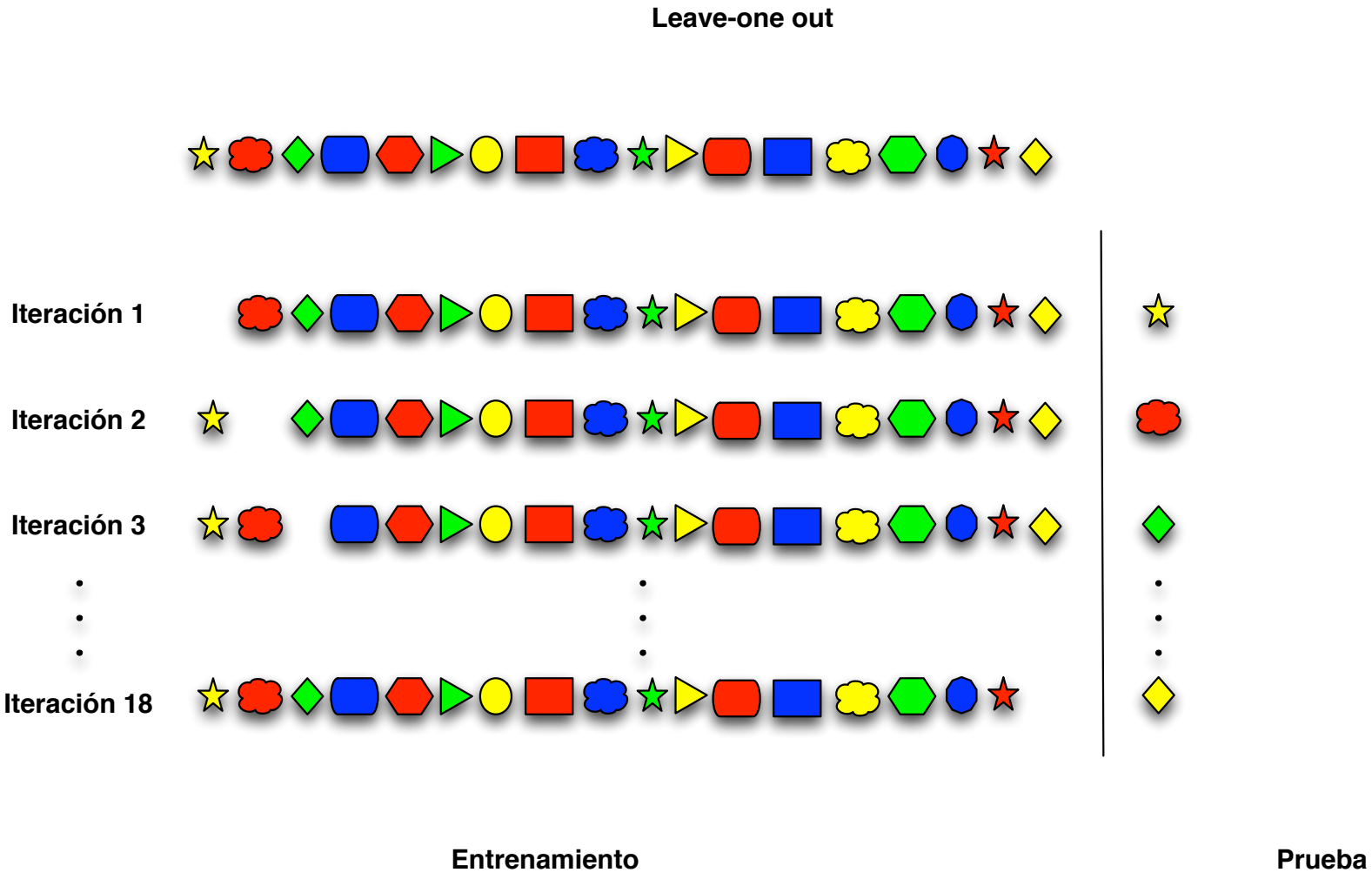
Validación cruzada

- 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 $k = 10$ es una buena opción
- Comparada 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 stratificación).

Validación cruzada dejando uno fuera

- La **validación cruzada dejando uno fuera** (Leave-one-out cross validation) es una caso especial de validación cruzada con $k =$ 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 mas precisos.
- Evita los inconvenientes de un muestreo aleatorio.

Validación cruzada



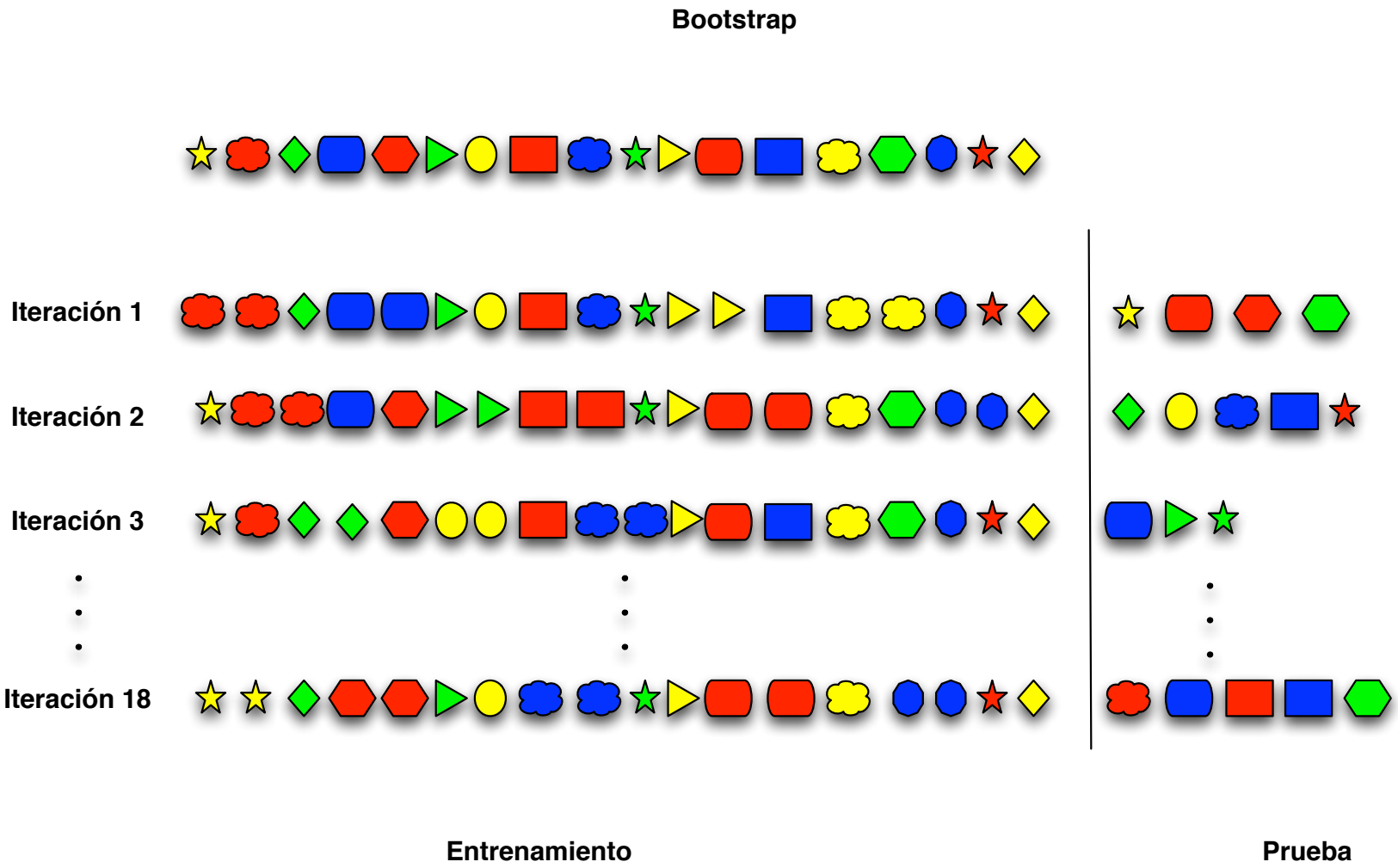
Validación cruzada dejando uno fuera

- Los **Inconvenientes:**
  - Muy costoso computacionalmente. El número de modelos creados es igual al número de elementos en el conjunto de datos.
  - Se utiliza cuando el conjunto de datos es muy pequeño.
  - No es posible una versión con estratificación.
- Muestra resulados similares a una validación cruzada con 10 pliegues.
  - Pero es mucho más ineficiente desde el punto de vista computacional.
- **Validación cruzada dejando un grupo fuera** (leave-group-out). Selecciona para el conjunto de prueba varios elementos al mismo tiempo.
  - Reduce el número de veces que hay que calcular el modelo.

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.
- El **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 sustituación.
  - Este conjunto se utiliza para crear el modelo (pueden existir elementos duplicados)
  - El conjunto de instancias no seleccionadas constituyen el conjunto de prueba.
  - El proceso se repite un número determinado de veces.

Bootstrap


Bootstrap
Iteración 1
Iteración 2
Iteración 3
Iteración 18
Entrenamiento
Prueba

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.

Bootstrap

- **boot.632** mitiga este problema redifiniendo el estadístico como:

$$E_S(h) = 0{,}632 \cdot E_{test} + 0{,}368 \cdot E_{trainig}$$

- 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?
- 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 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 %	99 %
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.
- Pero 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 las diferencias entre métodos se reduce
  - Elegir el que menos complejidad computacional presente.

Ajuste de parámetros

- En muchas ocasiones, los modelos requieren de unos parámetros para poder funcionar y que tienen 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 precisión/error del modelo es necesario determinar la mejor combinación de parámetros (model tunning).
- Normalmente nos quedamos con la configuración que mejor estimador obtiene.

Ajuste de parámetros

- Para ello el conjunto de entrenamiento se vuelve a dividir en dos:
  - Entrenamiento: para obtener los modelos.
  - Evaluación: para estimar la precisión/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 precisión/error del modelo para datos no vistos con 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 una 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 vendrá determinada por el problema.
- Para poder estimar dichos indicadores hemos presentado distintas tecnicas de evaluacion: hold-out, validacion cruzada, validacion cruzada y bootstrap.
- Otro aspecto importante que hemos analizado es el del ajuste de los parametros del modelo, es decir, ¿cómo determinar cuál es la mejor configuración del modelo?

Bibliografía relacionada

- 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.