

Evaluación de Modelos (I)

Carlos Hurtado L.

Departamento de Ciencias de
la Computación, U de Chile

Evaluación de Desempeño de un Modelo

- Nos centraremos en la capacidad predictiva de un modelo
- Otras dimensiones de desempeño:
 - Tiempo y memoria para clasificar
 - Tiempo y memoria para aprendizaje
 - Complejidad del modelo
 - Interpretabilidad

Evaluación de Desempeño: Decisiones a tomar

- Datos de prueba
- Medidas de efectividad
- Problema típico: predecir desempeño cuando los datos son limitados.
- Estrategia para reducir sesgo
- Considerar vs. no considerar costo de errores
- Evaluación de un modelo vs. comparación de varios modelos

Datos de Entrenamiento, Prueba y Objetivos

- Datos de entrenamiento (DE)
 - Datos que se usan para entrenar el modelo
- Datos de prueba (DP)
 - Datos que se usan para probar el modelo
- Datos objetivos (DO)
 - Datos sobre los cuales se ejecuta el modelo

Error de Predicción

- Medida básica de desempeño
- Tasa de error: fracción de datos objetivos mal clasificados.
- Tasa de éxito (potencia) = $1 - \text{error}$
- El error se estima usando técnicas de inferencia estadística
- El problema fundamental es evitar el efecto varianza en la estimación del error.

Varianza en la Estimación del Error

Ejemplo: weather.nominal

- Supongamos que inducimos la regla
If outlook=rainy then play=yes
Otherwise play=no
- Se evalúa en el siguiente conjunto:

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Rainy	Mild	High	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Sunny	Mild	High	FALSE	No

- Tenemos tasa de error igual a cero

Varianza en la Estimación del Error

Ejemplo: weather.nominal (cont.)

- Ahora inducimos la regla
If outlook=rainy then play=yes
Otherwise play=no
- Esta vez se evalúa en el siguiente conjunto:

Outlook	Temp.	Humidity	Windy	Play
Overcast	Hot	High	FALSE	Yes
Rainy	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Sunny	Cool	Normal	FALSE	Yes
Rainy	Mild	High	TRUE	No

- Ahora tenemos tasa de error igual a 100%

Estrategías de Prueba

- Resustitución
 - Estimamos el error usando los datos de entrenamiento
- Holdout
 - Estimamos el error usando datos de prueba disjuntos a datos de entrenamiento
- Validación Cruzada
 - Estimamos el error usando distintas particiones entrenamiento-prueba

Intervalo de confianza para el Error

- Supongamos que observamos s errores en un conjunto de n datos
- Estimamos el error como la proporción de la muestra

$$p = \frac{s}{n}$$

- Si $np \geq 10$ y $n(1 - p) \geq 10$ se tiene:

$$p \sim N\left(\pi, \frac{\pi(1 - \pi)}{n}\right)$$

Intervalo de Confianza para el Error de Predicción

$$\pi = p \pm z_{.025} \sqrt{\frac{\pi(1 - \pi)}{n}}$$

Despejamos π y obtenemos:

$$\pi = \frac{p + \frac{z^2}{2n} \pm z \sqrt{\frac{p}{n} - \frac{p^2}{n} + \frac{z^2}{4n}}}{1 + \frac{z^2}{n}}$$

Resustitución

- Estimación de error sobre datos de entrenamiento
- En este caso el error estimado se denomina “error de resustitución”
- Problemas:
 - Si usamos el error de resustitución como criterio para elegir un modelo el mejor modelo siempre será el más sobreajustado.
 - En general, la estimación del error sobre datos objetivo es deficiente.

Holdout

- Dividir datos disponibles en dos conjuntos disjuntos: DE y DP
- Se requiere:
 - Los dos conjuntos (construcción y prueba) sean **representativos** de los datos objetivos
- Compromiso tamaño DE vs. tamaño DP:
 - Para producir un buen clasificador necesitamos usar la mayor cantidad de datos para entrenamiento
 - Para obtener una buena estimación del error objetivo necesitamos usar la mayor cantidad de datos para prueba

Datos representativos para Holdout

- En general, no se puede saber si los datos de entrenamiento o prueba son representativos.
- Sí podemos saber si están todas las clases “representadas”
 - Si una clase no está representada en los datos de entrenamiento, el modelo no tendrá un buen desempeño para la clase.
 - Si la clase no está representada en los datos de entrenamiento, no mediremos bien el error asociado a la clase.

Técnicas para Evitar Varianza en Holdout

- Holdout estratificado:
 - Clases ocurren con la misma frecuencia en partición entrenamiento/prueba.
 - Salvaguarda básica para sesgo.
- Holdout repetitivo:
 - Repetir la prueba varias veces pero cambiando la partición entrenamiento/prueba.
 - Error estimado: promedio de errores de cada iteración

Validación Cruzada (“cross validation”)

- Es una forma de Holdout repetitivo
- Validación cruzada de n-subconjuntos (“n-folds”)
 - Datos se dividen en n subconjuntos de igual tamaño
 - Para cada subconjunto s , se usa s como prueba y los datos restantes como entrenamiento.

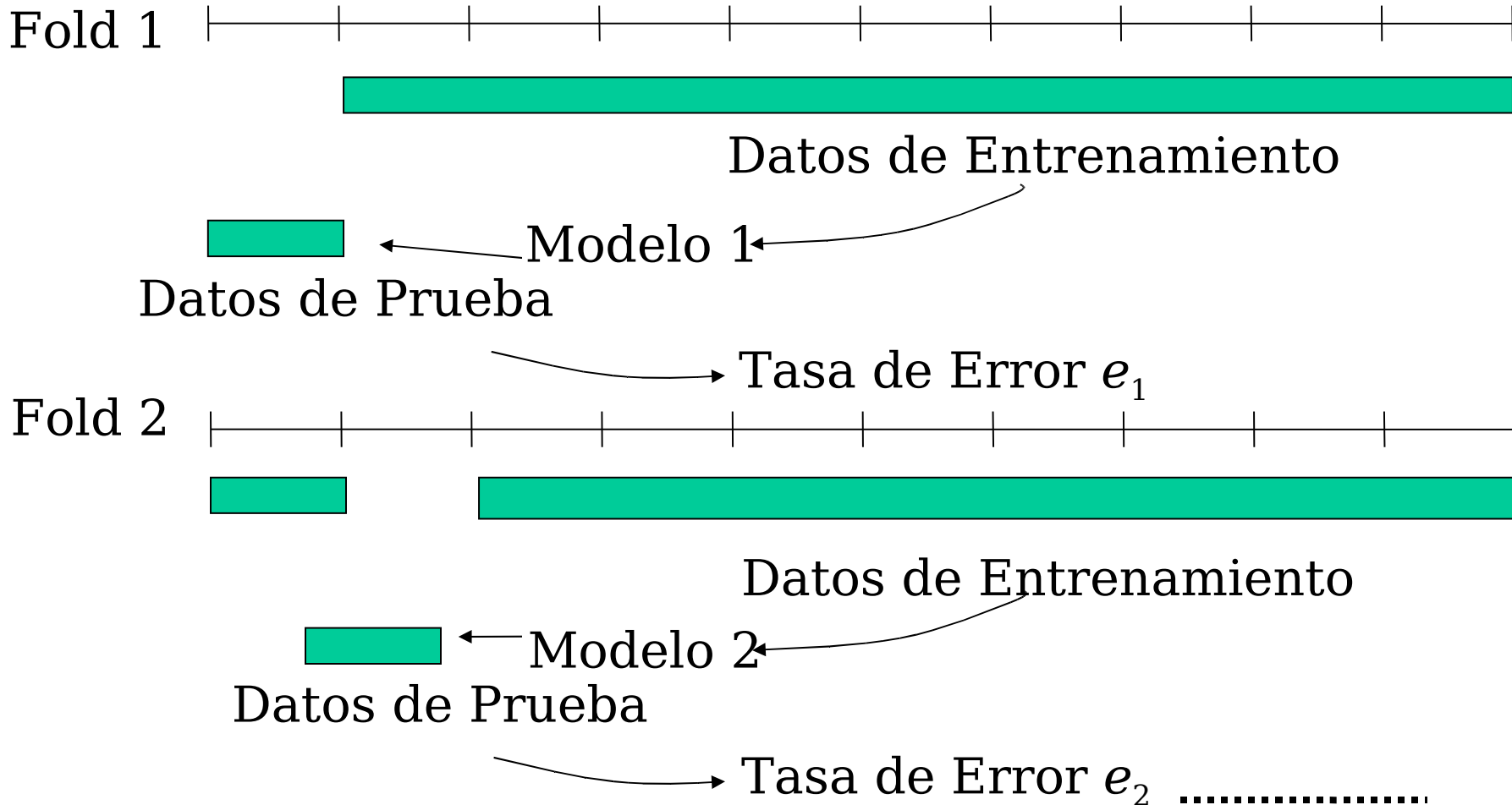
Validación Cruzada

- Error estimado: promedio de los errores en cada iteración
- Se puede usar estratificación
- Desventaja: costo computacional.
 - Se debe inducir el modelo n veces.
 - No es factible para conjuntos de datos grandes.

Validación Cruzada (cont.)

- Uso típico: 10 veces validación cruzada de 10-fold
- “leave-one-out”: caso particular, donde n es número de datos
 - útil cuando se tienen pocos datos.

Validación Cruzada



Estimación del Error en Validación Cruzada con k Folds

- Estimamos el error como:

$$q = \frac{1}{k} \sum_{i=1}^k p_k$$

- La variable q tiene media π
- La varianza de q se puede estimar como:

$$S_q^2 = \frac{1}{(k-1)} \sum_{i=1}^k (p_k - q)^2$$

Estimación del Error en Validación Cruzada con K Folds (cont.)

- Tenemos la siguiente aproximación:

$$\frac{q - \pi}{\frac{S_q}{\sqrt{k}}} \sim t(k - 1)$$

- Obtenemos un intervalo con un nivel de confianza C% para el error:

$$\pi = q \pm t_{1-\frac{C}{2}} \frac{S_q}{\sqrt{k}}$$

Ejemplo

- Luego de una validación cruzada de 6 folds obtenemos los siguientes errores:

0.80, 0.90, 0.95, 0.70, 0.45, 0.80

- Estimamos el error y la varianza como: $q = 0.77, S_q^2 = 0.03$

- Obtenemos la siguiente aproximación: $\frac{p - \pi}{\frac{S_q}{\sqrt{6}}} = \frac{p - \pi}{0.07} \sim t(5)$

Ejemplo (cont.)

- Obtenemos el siguiente intervalo con confianza 95% para el error objetivo:

$$\pi = 0.77 \pm t_{0.025} 0.07$$

$$\pi = 0.77 \pm (2.57 \times 0.07)$$

$$\pi = 0.77 \pm 0.19$$