Evaluación de Modelos (II)

Carlos Hurtado L.

Departamento de Ciencias de la Computación, U de Chile

Técnicas para evitar sesgo 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”)

- Forma de Hold-out repetitivo
- Validación cruzada de “**n**-fold”
  - Datos se dividen en un número **n** fijo de subconjuntos
  - Dado un subconjunto s, se usa s como prueba y los datos restantes como entrenamiento.
  - Esto se repite para cada subconjunto

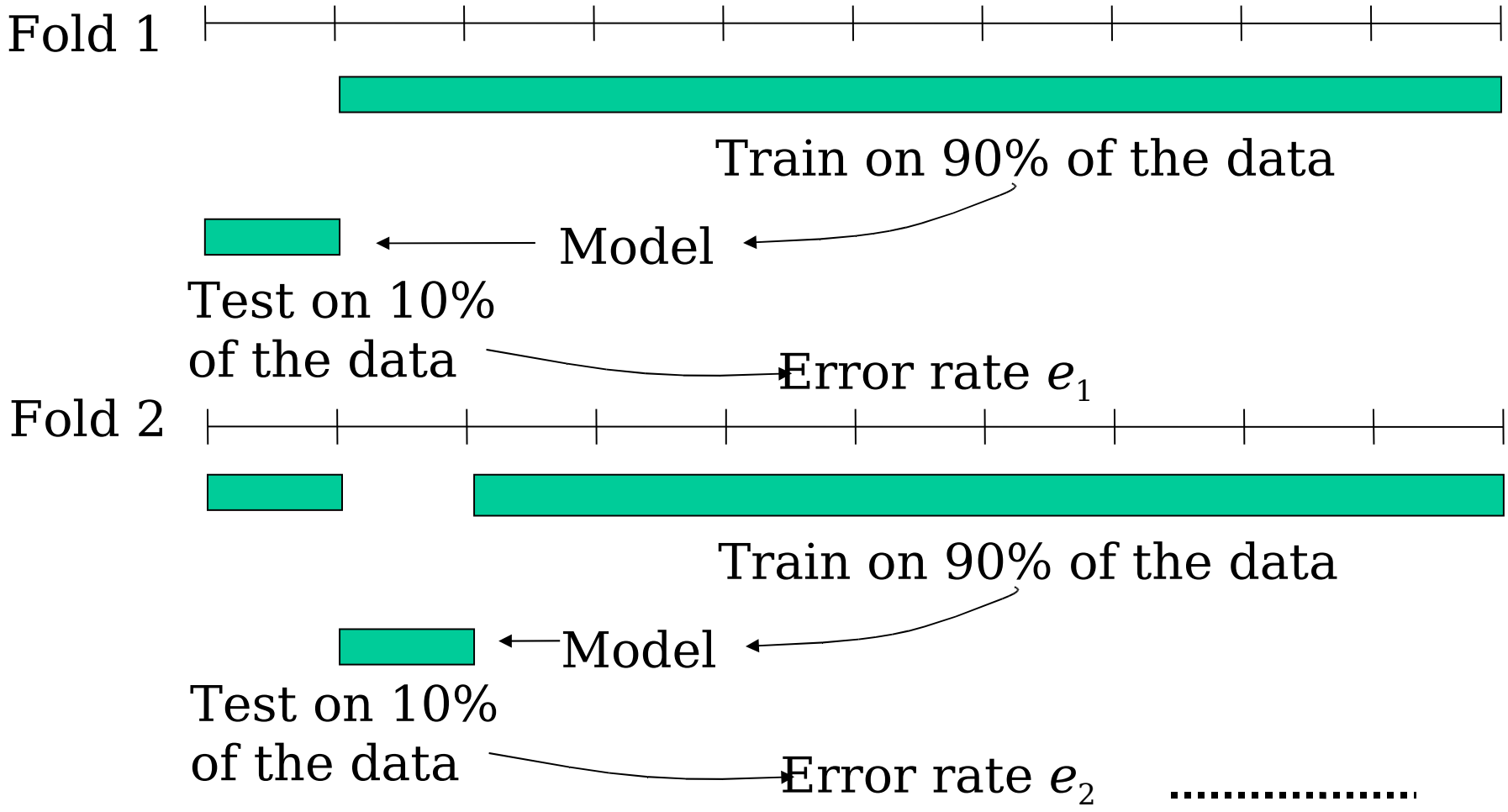
Validación cruzada (“cross validation”)

- 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 (“cross validation”) (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

- Estimación del error

$$\bar{Y} = \frac{1}{k} \sum_{i=1}^{k} \bar{X}_k$$

- La variable $\bar{Y}$ tiene media π

- Desviación estandar de $\bar{Y}$

$$s_{\bar{Y}^2} = \frac{1}{(k-1)} \sum_{i=1}^{k} (\bar{X}_k - \bar{Y})^2$$

Estimación del Error en Validación Cruzada (cont.)

- Tenemos $$\frac{\bar{Y}-\pi}{\sqrt{\frac{s_{\bar{Y}^2}}{k}}} \sim t_{k-1}$$

- Obtenemos el intervalo (95% de confianza):

$$\pi = \bar{Y} \pm t_{.025} \frac{s_{\bar{Y}}}{\sqrt{k}}$$

Ejercicio

- Corra J48 con weather y test de prueba igual a datos de entrenamiento
- Corra J48 con weather y percentage split de 66%
- Corra J48 con weather y 10-fold cross validation
- Compare tasa de error.

Comparación de Modelos

¿Cómo comparamos distintos modelos?

- Método 1:
  - Estimamos el error de cada modelo
  - Rankeamos
  - Seleccionar el mejor
- Método 2:
  - Verificación de Hipótesis

Verificación de Hipótesis

- Definimos una estadística que tiene determinada distribución si la hipótesis es correcta.
- Medimos la estadística en la muestra.
- Si tiene una alta probabilidad de ser obtenida de la distribución, aceptamos la hipótesis.
- En caso contrario la rechazamos.

Verificación de Hipótesis

- Hipótesis nula: $H_0 : \mu = \mu_0$
- Hipótesis alternativa: $H_1 : \mu \neq \mu_0$
- Aceptamos Ho con nivel de significancia 5% si

$$-z_{.025} \leq \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \leq z_{.025}$$

- Este es un “two-side test”.

Error Tipo I

- Es el error que tenemos cuando rechazamos H0 y esta es correcta.
- En el ejemplo anterior tenemos un 5% de prob. de tener error tipo I.
- Denotamos el error tipo I como α, se denomina nivel de significancia:

Error Tipo II

- Es el error que tenemos cuando aceptamos H0 y es falsa, usualmente se denota como β
- Probabilidad de aceptar H0 cuando la media es μ

$$\beta(\mu)=\Pr\left(\mu_0-z_{.025}\frac{\sigma}{\sqrt{n}}\leq\bar{X}\leq\mu_0 z_{.025}\frac{\sigma}{\sqrt{n}}|\mu\right)$$

Valor P

- Muestra el grado de acuerdo entre la muestra y la hipótesis

$$valorp = \Pr\left(Z < \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right) + \Pr\left(Z > \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right)$$

- La decisión equivale a rechazar HO si el valorp <= α. Pero esto se puede flexibilizar.

Verificación de Hipótesis (cont.)

- Hipótesis nula: $H_0: \mu \leq \mu_0$
- Hipótesis alternativa: $H_1: \mu > \mu_0$
- Aceptamos Ho con 5% de confianza si $-\infty \leq \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \leq z_{.95}$
- Este es un “one-side test”. $valorp = \Pr\left(Z > \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right)$

Compáración de Modelos: Test basado en validación cruazada de K-folds

- Estimamos la diferencia:

$$\bar{D} = \frac{1}{k} \sum_{i=1}^{k} \bar{D}_k \qquad \bar{D}_k = \bar{X}_k - \bar{Y}_k$$

- La variable $\bar{D}$ tiene media $\pi_D = \pi_X - \pi_Z$

- Desviación estandar de $\bar{Y}$ estimada

$$s_{\bar{D}^2} = \frac{1}{(k-1)} \sum_{i=1}^{k} (\bar{D}_k - \bar{D})^2$$

Compáración de Modelos: Test basado en validación cruazada de K-folds

- Hipótesis nula: $H_0: \pi_D = 0$
- Hipótesis alternativa: $H_1: \pi_D \neq 0$
- Aceptamos Ho con 5% de confianza si

$$-t_{0.025,k-1} \leq \frac{\bar{D} - \pi_D}{\sqrt{\frac{s_{\bar{D}^2}}{k}}} \leq t_{0.025,k-1}$$

Ejercicio

- Obtener el mejor desempeño para labor midiendo el error con validación cruzada usando “10-folds”
- Pruebe con distintos modelos
- Use el “experimenter” de weka para configurar este experimento

Ejemplos de modelos que parecen buenos pero no lo son

- Aplicación: aprobación de créditos
  - Modelo: todo postulante paga el crédito
  - Baja tasa de error
- Aplicación: diagnóstico de cáncer
  - Modelo: todos los tumores son benignos
  - Baja tasa de error
- Para evitar esto hay que considerar los distintos tipos de error.

Ejemplo: Modelo trivial de Aprobación de Crédito

		Predicted class	
		Yes	No
Actual Class	Yes	90	0
	No	10	0

Este no es un buen modelo ya que los falsos positivos significan un alto costo.

		Predicted class	
		Yes	No
Actual Class	Yes	80	5
	No	5	10

Este puede ser un mejor modelo que el anterior, pese a que tienen la misma tasa de error

Contabilización del Costo

- Los modelos anteriores no son buenos aunque tienen baja tasa de error.
- Pera el tipo de error que nos interesa, se comportan mal.
- En muchos casos, un buen modelo debe detectar excepciones, lo que no mide la tasa de error.

Tipos de Error: Matriz de contingencia

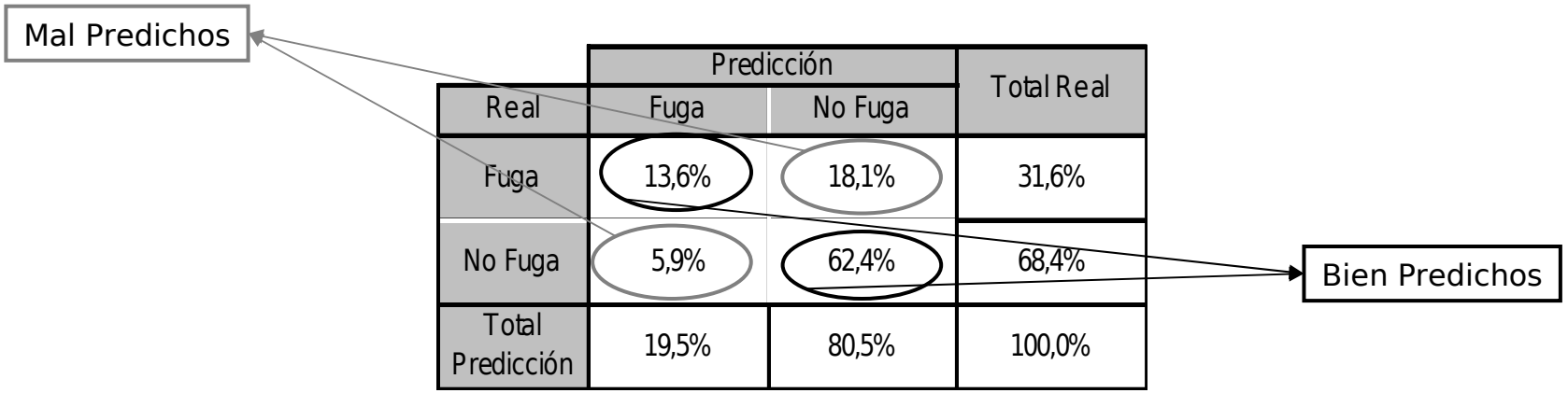
		Predicted class	
		Yes	No
Actual Class	Yes	True positive	False negative
	No	False positiv	True negati

Ejemplo: Modelo trivial de Aprobación de Crédito

		Predicted class	
		Yes	No
Actual Class	Yes	90	0
	No	10	0

Matriz de contingencia: ejemplo

Ejemplo: Modelo de predicción de "fugas" de clientes de tarjetas de crédito



Real	Predicción: Fuga	Predicción: No Fuga	Total Real
Fuga	13,6%	18,1%	31,6%
No Fuga	5,9%	62,4%	68,4%
Total Predicción	19,5%	80,5%	100,0%

Evaluación Sensible al Costo

- Costo del error:

$$C_{fp}\, E_{fp} + C_{fn}\, E_{fn}$$

- Donde:
  - E_{fn}: falsos negativos
  - C_{fn}: costo falso negativo
  - E_{fp}: falsos positivos
  - C_{fn}: costo falso positivo

Entrenamiento Sensible al Costo

- Se puede usar el costo para evaluar un modelo
- Mejor: usar información de costo en el entrenamiento del modelo
- Método simple:
  - “Engrosar” datos que demuestran un comportamiento importante (e.g., excepciones)
  - Se aplica para cualquier algoritmo de aprendizaje
  - Se puede engrosar en forma virtual a medida que se construye el modelo.
  - Por ejemplo, cambiar la distribución de las clases al calcular la entropía.

Modelos Gráficos

- Existen muchos modelos gráficos para representar el error
  - Lift Charts
  - Gráficos de precisión vs. cobertura
  - Curvas ROC

Lift Chart

- Idea: dada una clase, cada éxito representa un beneficio
- Motivación: marketing:
  - Modelo predice si un cliente responde
  - Cuántos folletos debemos enviar para que respondan Y clientes
- Solución
  - Graficar porcentaje de datos vs. número de éxitos (lift chart)
  - Ver qué porcentaje de datos necesito para tener Y éxitos

Construcción de Lift chart

- Fijar una clase
- Ordenar el conjunto de prueba de mayor a menor probabilidad de la clase
- Graficar en el eje X el porcentaje del conjunto de prueba
- Graficar en el eje Y el número de datos correctamente predichos en la clase

Ejemplo: weather.nominal

Table 5.4 Data for a lift chart.

rank	predicted probability	actual class	rank	predicted probability	actual class
1	0.95	yes	11	0.77	no
2	0.93	yes	12	0.76	yes
3	0.93	no	13	0.73	yes
4	0.88	yes	14	0.65	no
5	0.86	yes	15	0.63	yes
6	0.85	yes	16	0.58	no
7	0.82	yes	17	0.56	yes
8	0.80	yes	18	0.49	no
9	0.80	no	19	0.48	yes
10	0.79	yes	...	...	...

Ejemplo: Lift Chart

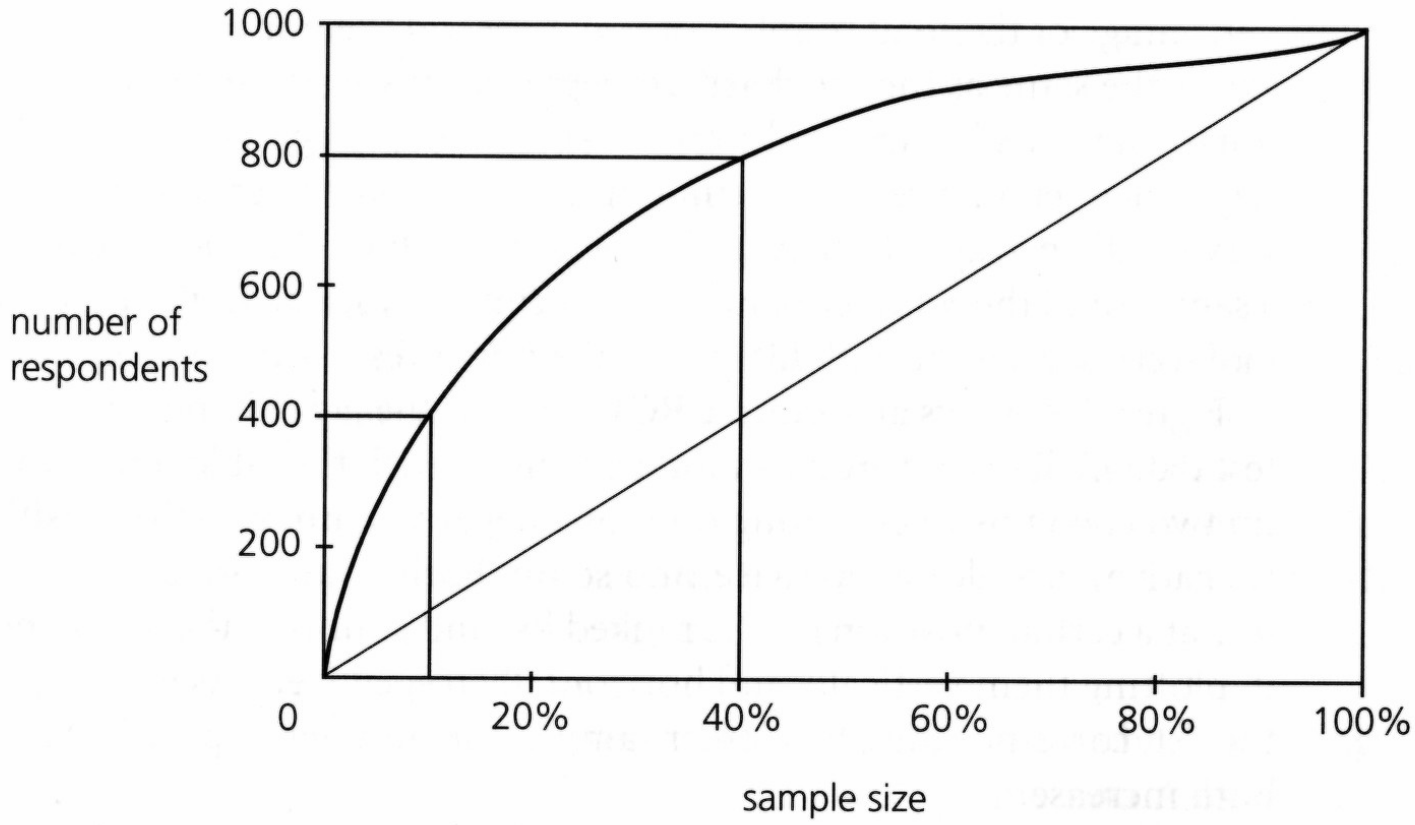


Figure 5.1 A hypothetical lift chart.

BeneficoCampaña(X) = E(X)BM-XMC

Ejercicio

- Graficar el lift chart para los datos anteriores
- Con un 30% de los mejores datos de la muestra, ¿qué número de éxitos se obtiene?

Ejercicio

- Corra J48 con weather y 10-fold cross validation
- Use excell para calcular
  - Mean absolute error
  - (Lift chart)
  - El mean absolute error debe coincidir con lo entregado por Weka
- Suponga que el costo de un falso positivo es 2 veces el costo de un falso negativo, modifique los datos y obtenga un modelo que considere este costo.
- Compare distintas ejecuciones de J48 con respecto a la medida de descripción mínima