Evaluación de Modelos (II)

Carlos Hurtado L.

Departamento de Ciencias de la Computación, U de Chile

Comparación de Modelos

- Podemos estimar el error de cada modelo y rankear.
- Una alternativa más sólida es usar técnicas de inferencia estadística

Comparación de Modelos

- Supongamos que tenemos dos modelos M1 y M2 con errores sobre datos objetivos π_1, π_2
- No conocemos estos errores, sólo podemos estimarlos.
- Para compararlos, debemos diseñar una prueba para estimar la diferencia:

$$\pi_d = \pi_1 - \pi_2$$

Veremos cuatro Situaciones en que Comparamos Modelos

- Estrategia de Prueba de los modelos:
  - Holdout vs. Validación Cruzada
- Relación entre los datos de prueba de los modelo:
  - Independiente: datos de prueba son independientes.
  - Pareado: datos de prueba se pueden parear (folds iguales en validación cruzada).

Test Estadísticos

- Holdout y Validación Cruzada- Pareado:
  - Test t para diferencia de media: dos muestras pareadas con varianza distinta y desconocida.
- Holdout – Independiente:
  - Test z para diferencia de proporciones: dos muestras no pareadas con varianza distinta y desconocida.

Test Estadísticos (cont.)

- Validación Cruzada – Indep.:
  - Test t para diferencia de medias: dos muestras no pareadas con varianza distinta y desconocida.

Hodout-Independiente

- Los errores de los modelos se estimaron con datos de prueba diferentes e independientes.
  - Este caso sucede frecuentemente cuando queremos comparar modelos propuestos en distintos artículos científicos.
- Para el test tenemos:

$$M_1 : p_1, n_1$$
$$M_2 : p_2, n_2$$

Holdout - Independiente

- Estimamos la diferencia como

$$d = p_1 - p_2$$

- La variable d tiene media $\pi_d = \pi_1 - \pi_2$ y varianza

$$\sigma_d^2 = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$$

- Y tenemos la sgte aproximación

$$\frac{d - \pi_d}{\sigma_d} \sim N(0,1)$$

Holdout - Independiente: Intervalo de Confianza para la Diferencia

- Si las muestra son suficientemente grandes, podemos estimar la varianza como

$$S_d^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

- Y tenemos el sgte intervalo para la diferencia

$$d - z_{0.025}\ S_d \leq \pi_d \leq d + z_{0.025}\ S_d$$

Ejemplo: Holdout – Independiente: Intervalo de Confianza

	p	n
M1	0.9	100
M2	0.8	50
d	0.1	
S_d	0.06	
Confianza C	0.95	
z_(1-C/2)	1.96	
Error Estándar (z * S_d)	0.13	
LimInf (d – ErrorEst)	-0.03	
LimSup (d + ErrorEst)	0.23	

Holdout - Idependiente: Verificación de Hipótesis

- Sea $z_0 = \frac{d}{S_d} \sim N(\pi_d, 1)$
- Definimos las hipótesis

$$H_0 : \pi_d = 0 \text{ vs. } H_1 : \pi_d \neq 0$$

- Si H_0 es cierta tenemos $z_0 \sim N(0, 1)$

Holdout - Idependiente: Verificación de Hipótesis

- Rechazamos H_0 ssi:

$$z_0 > z_{\alpha/2} \text{ or } z_0 < -z_{\alpha/2}$$

- α es el nivel de error (prob. de rechazar H_0 si es cierta).

Holdout - Idependiente: Verificación de Hipótesis

- Para verificar $H_0 : \pi_d = 0 \text{ vs. } H_1 : \pi_d < 0$
  rechazamos H_0 ssi $z_0 < -z_\alpha$
- Para verificar $H_0 : \pi_d = 0 \text{ vs. } H_1 : \pi_d > 0$
  rechazamos H_0 ssi $z_0 > z_\alpha$

Ejemplo: Holdout - Idependiente: Verificación de Hipótesis

	p	n
M1	0.9	100
M2	0.6	50
d	0.3	
S_d	0.08	
Nivel de Error (alfa)	0.05	
z_0	3.97	
z_alfa/2	1.96	
z_alfa	1.64	
¿Se rechaza H0:d=0 vs H1: d=/=0?	si	
¿Se rechaza H0:d=0 vs H1:d<0?	no	
¿Se rechaza H0:d=0 vs H1:d>0?	si	

Holdout - Pareado

- Es posible correr los modelos sobre los mismos datos de prueba.
- Podemos medir el error en cada dato
- En general esto permite intervalos más exactos.
- Medimos los errores como variables binarias:

$$M_1 : X_{11}, X_{12}, \ldots, X_{1n}$$
$$M_2 : X_{21}, X_{22}, \ldots, X_{2n}$$

Holdout - Pareado

- Estimamos la diferencia:

$$d = \frac{1}{n}\sum_{i=1}^{n} d_i,\ d_i = X_{1i} - X_{2i}$$

- La variable d tiene media $\pi_d = \pi_1 - \pi_2$

- Varianza de d

$$S_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - d)^2$$

Holdout - Pareado

- Tenemos $$\frac{d - \pi_d}{\frac{S_d}{\sqrt{n}}} \sim t(n-1)$$
- Intervalo de 95% de confianza para d

$$d - t_{0.025}\frac{S_d}{\sqrt{n}} \leq \pi_d \leq d + t_{0.025}\frac{S_d}{\sqrt{n}}$$

Holdout - Pareado

- Verificación de Hipótesis es análoga al caso independiente pero con

$$t_0 = \frac{d}{\frac{S_d}{\sqrt{n}}} \sim t(n-1)$$

Validación Cruzada - Pareado

- Realizamos un experimento de validación cruzada con los mismos datos y partición sobre los dos modelos.
- Para cada Modelo tenemos los errores medidos en los folds:

$$M_1 : p_{11}, p_{12}, \ldots, p_{1k}$$
$$M_2 : p_{21}, p_{22}, \ldots, p_{2k}$$

Validación Cruzada - Pareado

- Estimamos la diferencia como:

$$d = \frac{1}{k}\sum_{i=1}^{k} d_i \qquad\qquad d_k = p_{1k} - p_{2k}$$

- La variable d tiene media $\pi_d = \pi_1 - \pi_2$

- Varianza estimada de d

$$S_d^2 = \frac{1}{k-1}\sum_{i=1}^{k}(d_k - d)^2$$

Caso: Validación Cruzada – Pareado

- Tenemos $$\frac{d - \pi_d}{\frac{S_d}{\sqrt{k}}} \sim t(k-1)$$

- Intervalo de 95% de confianza para la diferencia

$$d - t_{0.025}\frac{S_d}{\sqrt{k}} \leq \pi_d \leq d + t_{0.025}\frac{S_d}{\sqrt{k}}$$

Ejemplo: Valiadación Cruzada – Pareado: Intervalo de Confianza

M1	0.8	0.9	0.7	0.6	0.8
M2	0.75	0.7	0.6	0.5	0.6
d_i	0.05	0.2	0.1	0.1	0.2

d	0.13
S_d	0.07
Confianza C	0.95
Grados de libertad (k-1)	4
t_(1-C/2)	2.78
Error Estándar (t * S_d/sqrt(k))	0.08
LimInf (d – ErrorEst)	0.05
LimSup (d + ErrorEst)	0.21

Caso: Validación Cruzada – Pareado

- Hipótesis nula: $H_0 : \pi_d = 0$
- Hipótesis alternativa: $H_1 : \pi_d \neq 0$
- Aceptamos H_0 con α de nivel de error ($1 - \alpha$ de confianza) ssi

$$t_0 > t_{\alpha/2} \text{ or } t_0 < -t_{\alpha/2}$$

$$t_0 = \frac{d}{\frac{S_d}{\sqrt{k}}}$$

Validación Cruzada – Pareado

- Para el test $H_0 : \pi_d = 0 \text{ vs. } H_1 : \pi_d < 0$
  rechazamos H_0 ssi $t_0 < -t_\alpha$
- Para el test $H_0 : \pi_d = 0 \text{ vs. } H_1 : \pi_d > 0$
  rechazamos H_0 ssi $t_0 > t_\alpha$

Ejemplo: Validación Cruzada – Pareado: Verificación de Hipótesis

M1	0.8	0.9	0.7	0.6	0.8
M2	0.75	0.7	0.6	0.5	0.6
d_i	0.05	0.2	0.1	0.1	0.2

d	0.13
S_d	0.07
Grados de libertad (k-1)	4
Nivel de Error (alfa)	0.05
t_0	3.36
t_alfa/2	2.78
t_alfa	2.13

¿Se rechaza H0:d=0 vs H1: d=/=0?	si
¿Se rechaza H0:d=0 vs H1:d<0?	no
¿Se rechaza H0:d=0 vs H1:d>0?	si

Validación Cruzada - Independiente

- Cada modelo se evaluó usando validación cruzada con un conjunto de datos distintos o con los mismos datos pero distintos folds.
- Tenemos los errores (como proporciones) para los folds de cada modelo:

$$M_1 : p_{11}, p_{12}, \ldots, p_{1k_1}$$
$$M_2 : p_{21}, p_{22}, \ldots, p_{2k_2}$$

Validación Cruzada – Independiente

- Estimamos la diferencia como:

$$p_1 = \frac{1}{k_1}\sum_{i=1}^{k_1} p_{1i} \quad p_2 = \frac{1}{k_2}\sum_{i=1}^{k_2} p_{2i} \qquad d = p_1 - p_2$$

- La variable d tiene media $\pi_d = \pi_1 - \pi_2$

- Varianza estimada de d

$$S_d^2 = \frac{S_1^2}{k_1} + \frac{S_2^2}{k_2}$$

Validación Cruzada – Independiente

- Tenemos la siguiente aproximación

$$\frac{d - \pi_d}{S_d} \sim t(v)$$

- Grados de libertad:

$$v = \frac{(n_1 - 1)(n_2 - 1)}{(n_2 - 1)c^2 + (n_1 - 1)(1 - c^2)} \qquad c = \frac{\frac{S_1^2}{k_1}}{\frac{S_1^2}{k_1} + \frac{S_2^2}{k_2}}$$

- Intervalo de 95% de confianza para la diferencia:

$$d - t_{0.025} S_d \leq \pi_d \leq d + t_{0.025} S_d$$

Validación Cruzada – Independiente

- Verificación de Hipótesis es análoga al caso anterior pero con

$$t_0 = \frac{d}{S_d} \sim t(v)$$

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

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

- Para evitar lo anterior debemos considerar los distintos tipos de error y su costo:
- Tabla de contingencia:

		Clase Predicha	
		si	no
Clase Real	si	VP	FN
	no	FP	VN

Tipos de Error: ejemplo: Modelos de Aprobación de Crédito

		Clase Predicha	
		si	no
Clase Real	si	90	0
	no	10	0

		Clase Predicha	
		si	no
Clase Real	si	80	10
	no	0	10

Tipos de Error: ejemplo: Modelos de Aprobación de Crédito

		Clase Predicha	
		si	no
Clase Real	si	90	0
	no	10	0

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

		Clase Predicha	
		si	no
Clase Real	si	80	10
	no	0	10

Este modelo es mejor que el anterior,
pese a que tienen la misma tasa de error

Matriz de Costos: Aprobación de Crédito

		Clase Predicha	
		si	no
Clase Real	si	-100	0
	no	1000	0

Evaluación Sensible al Costo

- El costo de un modelo sobre un conjunto de datos se puede expresar como:

$$\text{Costo}(M,D) = C_{vp}\,E_{vp} + C_{fn}\,E_{fn} + C_{fp}\,E_{fp} + Cv_n\,Ev_n$$

- Donde:
  - $E_{vp}, E_{fn}, E_{fp}, E_{vn}$: las frecuencias de cada tipo de error.
  - $C_{vp}, C_{fn}, C_{fp}, C_{vn}$: los costos de cada tipo de error.

Evaluación Sensible al Costo

- Es fácil demostrar que:

$$\text{Costo}(M,D) = (C_{fn} - C_{vp})\, E_{fn} + (C_{vn} - C_{fp})\, E_{fp} + \text{Cte.}$$

- Es decir, para efectos de evaluar modelos, podemos usar una matriz de la forma:

		Clase Predicha	
		si	no
Clase Real	si	0	CFN-CVP
	no	CFP – CVN	0

Entrenamiento Sensible al Costo

- Dada una matriz de costos:

		Clase Predicha	
		si	no
Clase Real	si	0	CVP -CFN
	no	CFP – CVN	0

- El costo de un error de un dato que está en la clase “si” es CVP-CFN.
- El costo de un error de un dato que está en la clase “no” es CFP – CVN.

Entrenamiento Sensible al Costo

- Idea: usar información de costo en el entrenamiento del modelo
- El método más básico es estratificación:
  - Duplicar (engrosar) o eliminar datos de alguna clase de forma que la proporción de datos en cada clase sea igual que la proporción del costo del error de cada clase.

Entrenamiento Sensible al Costo

- Se puede estratificar en forma virtual a medida que se construye el modelo. Por ejemplo, cambiar la distribución de las clases al calcular la entropía.
- Este método es general: se aplica a cualquier algoritmo de aprendizaje

Ejemplo: Aprobación de Créditos

- Tenemos la matriz de costos:

		Clase Predicha	
		si	no
Clase Real	si	-100	0
	no	1000	0

- La transformamos en la sgte matriz de costos:

		Clase Predicha	
		si	no
Clase Real	si	0	100
	no	1000	0

Ejemplo: Aprobación de Crédito

- El costo de cometer error para un dato de la clase “si” es 100
- El costo de cometer un error para un dato de la clase “no” es 1000.
- Luego necesitamos obtener una muestra de entrenamiento donde la relación de la clase “si” sobre la “no” sea 1/10.