

Métodos Estadísticos de Remuestreo

Estadística Computacional I

Departamento de Estadística e Investigación Operativa (Univ. Sevilla)

Introducción

Tests de Permutaciones

Jackknife

Bootstrap

Introducción

3

INTRODUCCIÓN

- Se basan en la generación de nuevas muestras a partir de la muestra disponible.
- Utilizan potencia computacional como alternativa cuando no se dan las condiciones necesarias para aplicar las técnicas tradicionales de la inferencia estadística
- En ciertas ocasiones, este proceso de remuestreo se aplica también a cada una de las muestras generadas a partir de la muestra original
- Son especialmente útiles en situaciones como las siguientes:
 - Tamaño muestral reducido
 - Modelos no paramétricos
 - No cumplimiento de condiciones (independencia de observaciones, homocedasticidad, etc).
- Deben considerarse como métodos complementarios a los ya conocidos

2

4

Tests de Permutaciones

5

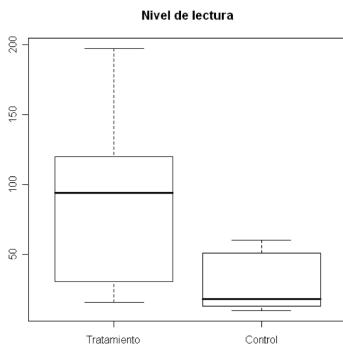
TESTS DE PERMUTACIONES

1. Tests de Permutaciones

- Se basan en la construcción de estadísticos de distribución libre
- Suelen ser exactos: $P[pvalor \leq \alpha] = \alpha$
- **Ejemplo:** ¿Mejora la capacidad de lectura de los alumnos de Primaria si realizan actividades de lectura guiada?
- Se dispone de dos muestras independientes:
 - Tratamiento: 94,38,23,197,99,16,141
 - Control: 52,10,12,14,51,18,60,13,31
- Denotando por X e Y las puntuaciones de un test de aprovechamiento de la lectura para los que siguen el tratamiento y para el grupo de control, se plantea:
$$\begin{cases} H_0: E[X] \leq E[Y] \\ H_1: E[X] > E[Y] \end{cases}$$

3

6



```
> shapiro.test(x)
Shapiro-Wilk normality test
data: x
W = 0.9233, p-value = 0.4955

> shapiro.test(y)
Shapiro-Wilk normality test
data: y
W = 0.8228, p-value = 0.03703
```

¿Inferencia paramétrica?

- Normalidad cuestionable
- Tamaños muestrales reducidos (no aplicable el Teorema Central del Límite)

4

7

Un test de permutaciones:

1. Se elige un estadístico apropiado, por ejemplo la diferencia de medias aritméticas, y se calcula para las muestras disponibles.
2. Se toman aleatoriamente 7 de las 16 puntuaciones, y pasan a formar la nueva muestra del grupo Tratamiento.*
3. Las restantes 9 definen la nueva muestra del grupo Control.*
4. Se calcula la nueva diferencia de medias.
5. Se repiten los pasos 2, 3 y 4 todas las veces posibles.
6. Se calcula el p-valor a partir de la distribución de la nueva diferencia de medias y de la diferencia inicial.

*Un planteamiento equivalente: las etiquetas de pertenencia a las clases Control/Tratamiento se permutan aleatoriamente

5

8

Etapas aconsejadas de un test de permutaciones

1. Análisis del problema y planteamiento del contraste
2. Elegir un estadístico S para el test
3. Calcular el estadístico S_0 para las observaciones originales
4. Calcular el estadístico S^* para cada una de las M posibles permutaciones suponiendo cierta la hipótesis nula
5. Calcular el p -valor: Probabilidad de obtener un valor de S^* tan extremo o más que S_0 bajo H_0

En el ejemplo anterior
$$p = \frac{\#\{S^* \geq S_0\}}{M}$$

6

9

Las permutaciones, en el ejemplo de las dos muestras, se pueden expresar en función de las etiquetas:

Sea Z la muestra conjunta ordenada de tamaño $n+m$, y v el vector de etiquetas identificativas de la muestra de pertenencia

$$Z = (10 \ 16 \ 23 \ 27 \ 30 \ 38 \ 40 \ 46 \ 51 \ 52 \ 94 \ 99 \ 104 \ 141 \ 146 \ 197)$$

$$v = (B \ A \ A \ B \ B \ A \ B \ B \ B \ B \ A \ A \ B \ A \ B \ A)$$

$$S = S(Z, v) = S_0$$

$$S^* = S(Z, v^*)$$

donde v^* representa una permutación aleatoria del vector v

Número de permutaciones distintas de las etiquetas:

$$\binom{n+m}{n}$$

7

10

Lema de permutaciones: Si bajo H_0 , las funciones de distribución de ambas poblaciones coinciden, entonces todas las permutaciones son equiprobables

Por eso en el ejemplo se calcularía el p-valor mediante

$$p = \frac{\# \{S^* \geq S_0\}}{\binom{n+m}{n}}$$

El resultado anterior implica que los datos deben ser **totalmente intercambiables bajo la hipótesis nula**, por tanto en una comparación de medias se supone implícitamente que las varianzas son iguales. Más aún, en realidad la hipótesis nula establece la igualdad de ambas funciones de distribución

8

11

Tests de Permutaciones

Aproximación por el método de Monte Carlo

12

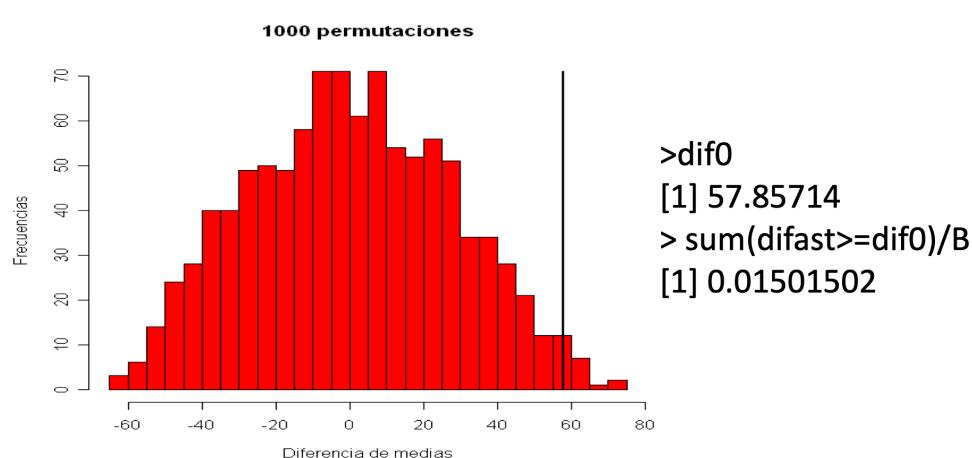
Aproximación por el método de Monte Carlo:

- Fijar B (número de permutaciones a generar)
- Generar aleatoriamente y de forma independiente B permutaciones $v_1^*, v_2^*, \dots, v_B^*$
- Calcular $S_b^* = S(Z, v_b^*) \quad b = 1, 2, \dots, B$
- Aproximar p mediante

$$\hat{p}_B = \frac{\#\{S_b^* \geq S_0\}}{B}$$

9

13



Por tanto, se rechaza la hipótesis nula, se ha encontrado evidencia estadística a favor de la siguiente conclusión: *realizar actividades guiadas de lectura tiende a mejorar la capacidad de aprovechamiento de la lectura.*

10

14

Elección de B

$$B\hat{p}_B \sim Bi(B, p)$$

$$CV(\hat{p}_B) = \left[\frac{1-p}{pB} \right]^{1/2}$$

Si deseamos un CV de 0.1, de forma que el error de simulación no afecte a la estimación de p (p-valor a estimar) en más de un 10%:

$p:$	0.5	0.25	0.1	0.05	0.025
$B:$	100	299	900	1901	3894

Se recomienda en torno a 10000 permutaciones.

11

15

El proceso de generación de permutaciones conlleva implícitamente la posibilidad de que se repitan algunas.

En realidad el cálculo correcto del p -valor requiere tener en cuenta que las permutaciones se generan con reemplazamiento.

Para mejorar la estimación del p -valor se puede añadir 1 en el numerador y denominador:

$$\hat{p}_B = \frac{\#\{S_b^* \geq S_0\} + 1}{B + 1}$$

En realidad esta aproximación es conservativa: tiende a ser ligeramente superior al p -valor exacto pero la discrepancia es muy reducida cuando el número total de permutaciones es elevado.

$B=9999$ es el valor recomendado con esta corrección.

12

16

Jackknife

17

JACKKNIFE

2. Jackknife

- 1949: Quenouille propuso esta técnica para estimar el sesgo.
- 1958: Tukey la bautizó como “jackknife” (navaja-suiza) y la usó para estimar el error estándar de un estimador.
- El nombre alude a la versatilidad atribuida a la técnica.
- Se utiliza poco para calcular intervalos de confianza.
- Otro uso: estimar el rendimiento de un modelo de predicción o clasificación, de hecho es la base de los métodos de validación cruzada, muy usados en Minería de Datos.

13

18

Sea una m.a.s. X_1, \dots, X_n i.i.d. $\sim F(x)$. Se considera un estimador $T_n = T(X_1, \dots, X_n)$ de un parámetro θ .

El sesgo de este estimador viene definido por

$$\text{sesgo}(T_n) = E[T_n] - \theta$$

El remuestreo jackknife se basa en la generación de todas las muestras posibles de tamaño $n-1$ que se obtienen al eliminar cada posible observación: $T_{n-1,i} = T(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$

Definición 2.1. El estimador jackknife del sesgo, que denotaremos s_{jack} , se define como

$$s_{jack}(T_n) = (n-1)(\bar{T}_n - T_n)$$

donde

$$\bar{T}_n = \frac{1}{n} \sum_{i=1}^n T_{n-1,i}$$

14

19

El estimador inicial puede ser corregido teniendo en cuenta la estimación del sesgo.

Definición 2.2. Se define el estimador jackknife de θ , basado en T_n , que denotaremos T_{jack} como

$$T_{jack} = T_n - s_{jack} = nT_n - (n-1)\bar{T}_n$$

Se puede demostrar que para estadísticos T_n que sean cuadráticos, $s_{jack}(T_n)$ es un estimador insesgado del sesgo(T_n). (Un estadístico es cuadrático si puede expresarse como suma de funciones de cada una de las X_i , y de funciones de los pares (X_i, X_j)).

En otras situaciones, no está asegurado que T_{jack} tenga menor sesgo que T_n .

El estimador jackknife se puede expresar como la media aritmética de los llamados pseudovalores:

$$T_{jack} = \frac{1}{n} \sum_{i=1}^n \{nT_n - (n-1)T_{n-1,i}\} = \frac{1}{n} \sum_{i=1}^n \tilde{T}_{n-1,i}$$

Cada pseudovalor se puede interpretar como una versión corregida de T_n :

$$\tilde{T}_{n-1,i} = nT_n - (n-1)T_{n-1,i} = T_n + (n-1)(T_n - T_{n-1,i})$$

15

20

Teniendo en cuenta que la cuasivarianza muestral es un estimador insesgado de la varianza, y dado que el estimador jackknife es la media de los pseudovalores, Tukey propuso el siguiente estimador de la varianza.

Definición 2.3. El estimador jackknife de la varianza de T_n , que denotaremos v_{jack} , se define como

$$v_{jack} = \frac{1}{n(n-1)} \sum_{i=1}^n \left(\bar{T}_{n-1,i} - \frac{1}{n} \sum_{j=1}^n \bar{T}_{n-1,j} \right)^2 = \frac{n-1}{n} \sum_{i=1}^n (T_{n-1,i} - \bar{T}_n)^2$$

Se puede demostrar que para estadísticos T_n que sean lineales, $v_{jack}(T_n)$ es un estimador insesgado de $\text{var}(T_n)$. (Un estadístico es lineal si puede expresarse como suma de funciones de cada una de las X_i).

La estimación de la varianza puede utilizarse para calcular Intervalos de Confianza basados en la aproximación normal:

$$(T_n - Z_{1-\alpha/2} \sqrt{v_{jack}}, T_n + Z_{1-\alpha/2} \sqrt{v_{jack}})$$

Sin embargo los métodos bootstrap suelen dar mejor cobertura.

No se recomienda el jackknife cuando T_n es demasiado discontinuo como función de cada observación, o depende solo de algunos valores (por ejemplo la mediana). 16

21

- Procedimiento de validación cruzada: utilizada para estimar el rendimiento de un modelo de predicción o clasificación, por ejemplo la regla discriminante lineal de Fisher o una red de neuronas artificiales. O bien para configurar la complejidad de un modelo.
- La muestra D_n se divide aleatoriamente en K subconjuntos disjuntos $\{V_1, \dots, V_K\}$.
- A diferencia de la estimación directa en la muestra, este procedimiento da un estimador insesgado del *error de generalización* del modelo ajustado.
- $K=n$ corresponde al remuestreo jackknife

D_n	Entrenamiento	Evaluación
V_1	$D_n - V_1$	V_1
V_2	$D_n - V_2$	V_2
V_3	$D_n - V_3$	V_3
V_4	$D_n - V_4$	V_4
V_K	$D_n - V_K$	V_K

$$E_{VC} = \frac{1}{K} \sum_{i=1}^K E_i$$

17

22

Bootstrap

23

Bootstrap

Método Bootstrap

24

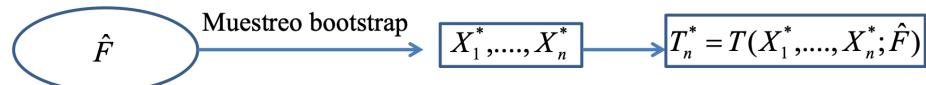
3. Bootstrap

3.1. Métodos Bootstrap

Sea una m.a.s. X_1, \dots, X_n i.i.d. $\sim F(x)$. Se considera un estadístico $T_n = T(X_1, \dots, X_n; F)$.



La idea del bootstrap es reemplazar la función de distribución F por alguna estimación apropiada, y aproximar la distribución del estadístico T_n por la distribución del estadístico evaluado sobre muestras de dicha función de distribución aproximada, a las que se les conoce como muestras bootstrap:



El término bootstrap está inspirado en un episodio de “Las aventuras del Barón de Münchhausen”, donde el protagonista sale de un agujero tirándose de los cordones de sus botas.

Este término también denota el proceso de arranque de los ordenadores, y realizar una actividad emprendedora a partir de muy pocos recursos.

18

25

A partir de la distribución bootstrap se pueden calcular estimaciones del sesgo y varianza de T_n , así como otras características de su distribución. Otras aplicaciones interesantes son la construcción de intervalos de confianza y la realización de contrastes de hipótesis.

Estimación del sesgo, denotando por T_0 el valor observado para T_n en la muestra disponible:

$$\text{sesgo}_{\text{boot}}(T_n) = E[T_n^* / \hat{F}] - T_0$$

Estimación de la varianza: $\text{var}_{\text{boot}}(T_n) = \text{var}[T_n^* / \hat{F}]$

Estimación del percentil de orden p , denotando por G_{boot} la función de distribución del estadístico bootstrap: $\xi_p^* = G_{\text{boot}}^{-1}(p)$

Se distinguen tres variantes principales, según la aproximación a la función de distribución:

Bootstrap no paramétrico: F es totalmente desconocida, se suele aproximar por la función de distribución empírica $F_n(x)$.

Bootstrap paramétrico: F es conocida salvo un vector de parámetros θ . En este caso se considera la función de distribución en la que θ es previamente estimada: $F_{\hat{\theta}} = F(x, \hat{\theta})$

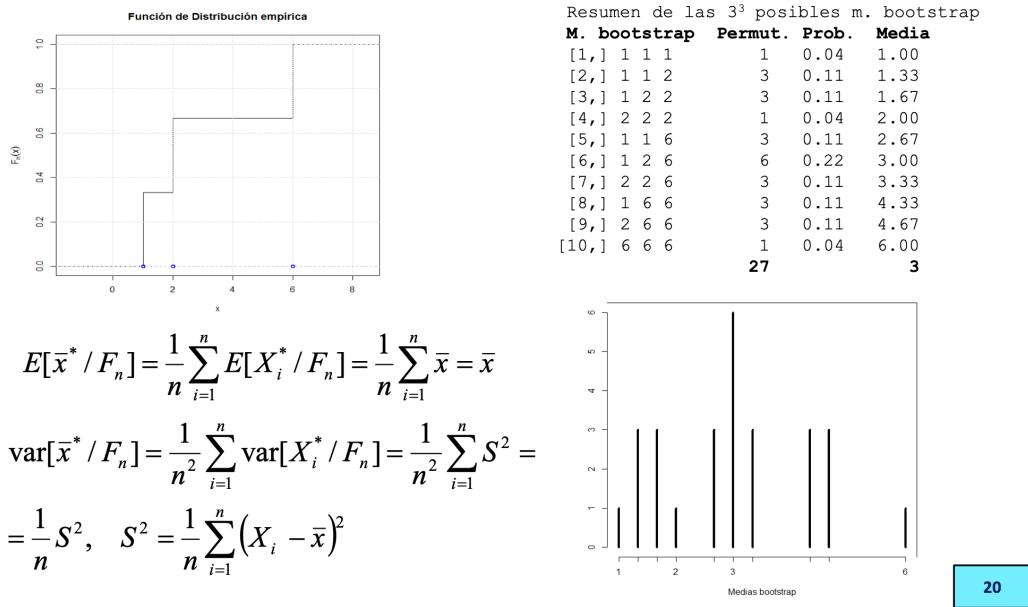
Bootstrap suavizado: si la forma paramétrica de F es desconocida pero se sabe que es continua, se puede aproximar a partir de la estimación no paramétrica de la función de densidad.

19

26

Las aplicaciones prácticas del bootstrap suelen emplear la primera variante, a partir de la función de distribución empírica.

Ejemplo: Se tiene la muestra (1,2,6), y se considera el estadístico media aritmética.



27

Sin embargo, en aplicaciones prácticas o bien no se puede conocer por medios analíticos la distribución del estadístico bootstrap, o bien es inviable generar todas las muestras bootstrap posibles.

En realidad lo usual es generar un número B de muestras bootstrap, y aproximar la distribución del estadístico bootstrap mediante la distribución de los B valores calculados. Este método es conocido como **Método II de Efron**, si bien se suele emplear directamente el término de Método Bootstrap.

Algoritmo:

Fijar un entero $B > 0$.

Para $b=1$ hasta B

Generar una muestra bootstrap y calcular el estadístico sobre dicha muestra

$$\text{Siguiente } b \qquad T_b^* = T(X_1^{*b}, \dots, X_n^{*b}; \hat{F})$$

Las estimaciones bootstrap pueden ser aproximadas como sigue:

$$\text{sesgo}_B^*(T_n) = \frac{1}{B} \sum_{b=1}^B T_b^* - T_0 \qquad \text{var}_B^*(T_n) = \frac{1}{B-1} \sum_{b=1}^B (T_b^* - \bar{T}_B^*)^2$$

$$G_{\text{Boot}} \approx G_B^*(t) = \frac{1}{B} \sum_{b=1}^B I\{T_b^* \leq t\} \qquad \xi_{p,B}^* = G_B^{-1}(p)$$

Se obtienen aproximaciones de los estimadores bootstrap, por lo que además de la variabilidad atribuible a la muestra original se introduce un error o variabilidad debido a este remuestreo. Para el sesgo y varianza suele ser suficiente $B=200$, para otras aplicaciones es recomendable al menos $B=2000$.

21

28

Existen diversos resultados teóricos basados en desarrollos asintóticos sobre la consistencia y la tasa de convergencia de los métodos bootstrap.

Bajo ciertas condiciones, se verifica la consistencia:

$$P\left|P[T(X^*, \hat{F}) \leq q] - P[T(X, F) \leq q]\right| > \varepsilon \xrightarrow{n \rightarrow \infty} 0, \quad \forall \varepsilon, q$$

Este resultado cubre las situaciones inferenciales más comunes.

Sin embargo hay situaciones donde el bootstrap falla, por ejemplo la estimación de extremos (máximos o mínimos).

También puede fallar en el tratamiento de datos dependientes, como las series temporales, donde deben aplicarse mecanismos de muestreo específicos para esa situación.

En comparación con los tests de permutaciones, éstos son exactos si se consideran todas las permutaciones posibles, mientras que las propiedades bootstrap son asintóticas. A cambio, el bootstrap proporciona mejores intervalos de confianza y una mejor estimación del error estándar.

22

29

Bootstrap

Contrastes de Hipótesis

30

3.2. Contrastes de Hipótesis

En primer lugar, un intervalo de confianza bootstrap podría ser empleado para realizar un contraste, pero la precisión y la potencia pueden ser mucho mejores construyendo directamente el contraste bootstrap. Dos consejos importantes, el primero ayuda a incrementar la potencia del test, mientras que el segundo ayuda a obtener una probabilidad de error de tipo I más cercana a α :

I) *Usar estadísticos pivotales,* $\frac{\hat{\theta}^* - \hat{\theta}}{\hat{es}^*(\hat{\theta})}$

II) *El muestreo bootstrap debe reflejar la hipótesis nula.* Si por ejemplo la hipótesis nula establece la igualdad de dos medias, las muestras bootstrap se forman con elementos extraídos de forma indistinta de cualquiera de las submuestras originales.

Si $H_0: \theta = \theta_0$, si el estadístico es la distancia $|\hat{\theta} - \theta_0|$, se podría tener la tentación de utilizar como estadístico bootstrap $|\hat{\theta}^* - \theta_0|$.

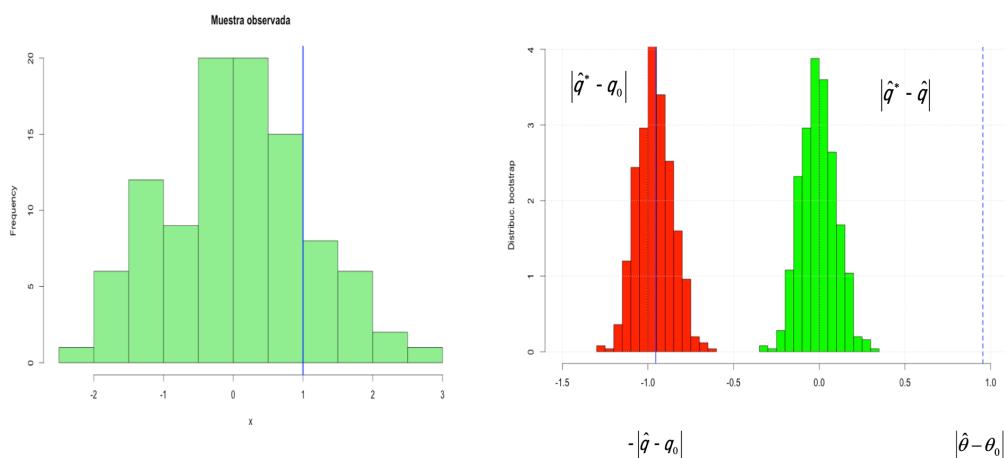
Sin embargo si fuera falsa la hipótesis nula $|\hat{\theta}^* - \theta_0|$ no se vería grande en comparación con los $|\hat{\theta} - \theta_0|$, pero los valores de $|\hat{\theta} - \theta_0|$ sí serían pequeños en comparación con $|\hat{\theta}^* - \theta_0|$.

23

31

Supongamos que se desea contrastar la hipótesis nula de que la media es 1, dada una muestra de la $N(0,1)$.

$$P^* \{ |\hat{q}^* - q_0| > |\hat{q} - q_0| \} = 0.53 \quad P^* \{ \hat{\theta}^* - \hat{\theta} > |\hat{\theta} - \theta_0| \} = 0$$



24

32

Bootstrap

Intervalos de Confianza y Contrastes Bootstrap

33

INTERVALOS DE CONFIANZA Y CONTRASTES BOOTSTRAP

3.3. Intervalos de Confianza y Contrastes Bootstrap

Método basado en la ley Normal.

Los estimadores usuales (e.m.v.) verifican $\frac{\hat{\theta} - \theta}{e\hat{s}(\hat{\theta})} \rightarrow N(0,1)$

De donde se tiene el IC aproximado

$$(\hat{\theta} - Z_{1-\alpha/2} e\hat{s}(\hat{\theta}), \hat{\theta} + Z_{1-\alpha/2} e\hat{s}(\hat{\theta}))$$

Si el estadístico bootstrap sigue una ley normal:

$$\hat{\theta}^* \sim N(\hat{\theta}, e\hat{s}^*(\hat{\theta})^2)$$

se puede entonces construir el siguiente I.C. aproximado:

$$\hat{\theta}_{\alpha/2}^* = \hat{\theta} - Z_{1-\alpha/2} e\hat{s}^*(\hat{\theta}); \hat{\theta}_{1-\alpha/2}^* = \hat{\theta} + Z_{1-\alpha/2} e\hat{s}^*(\hat{\theta})$$

Supone la normalidad del estadístico bootstrap, lo que debe ser comprobado.

25

34

Método Percentil.

Se basa en los percentiles de la distribución bootstrap $\hat{\theta}^* = \hat{\theta}(X^*)$

$$\hat{\theta}_{\alpha/2}^* = \hat{G}^{*-1}(\alpha/2) \quad \hat{\theta}_{1-\alpha/2}^* = \hat{G}^{*-1}(1-\alpha/2)$$

\hat{G}^* es la función de distribución bootstrap, que en general es aproximada a partir de B muestras bootstrap.

- Es proclive a cierto sesgo y a un cubrimiento inadecuado.
- Funciona mejor con parámetros de localización.
- Equivalente, de forma asintótica, al basado en la normal.
- Respeta transformaciones: los extremos del intervalo para una transformación monótona de θ se obtienen aplicando la transformación a los extremos del intervalo para θ .

Justificación:

Supongamos que existe una transformación continua y estrictamente creciente ϕ , y una función de distribución H continua y simétrica, $H(z)=1-H(-z)$, verificando

$$P[h_{\alpha/2} \leq \phi(\hat{\theta}) - \phi(\theta) \leq h_{1-\alpha/2}] = 1 - \alpha \quad (1)$$

Si ϕ es por ejemplo una transformación estabilizadora de la varianza y que conduce a normalidad, H es la $N(0,1)$.

Si F es continua, cualquier v.a. $X \sim F$ se puede transformar a cualquier f.D. G mediante $G^{-1}(F(x))$.

26

35

Aplicando el principio bootstrap a la expresión (1):

$$\begin{aligned} 1 - \alpha &\approx P^*[h_{\alpha/2} \leq \phi(\hat{\theta}^*) - \phi(\hat{\theta}) \leq h_{1-\alpha/2}] = P[h_{\alpha/2} + \phi(\hat{\theta}) \leq \phi(\hat{\theta}^*) \leq h_{1-\alpha/2} + \phi(\hat{\theta})] = \\ &= P[\phi^{-1}(h_{\alpha/2} + \phi(\hat{\theta})) \leq \hat{\theta}^* \leq \phi^{-1}(h_{1-\alpha/2} + \phi(\hat{\theta}))] = P[\xi_{\alpha/2}^* \leq \hat{\theta}^* \leq \xi_{1-\alpha/2}^*] \end{aligned} \quad (2)$$

Desarrollando la expresión (1) y teniendo en cuenta la simetría de H :

$$\begin{aligned} 1 - \alpha &= P[h_{\alpha/2} \leq \phi(\hat{\theta}) - \phi(\theta) \leq h_{1-\alpha/2}] = P[h_{\alpha/2} - \phi(\hat{\theta}) \leq -\phi(\theta) \leq h_{1-\alpha/2} - \phi(\hat{\theta})] = \\ &= P[-h_{1-\alpha/2} + \phi(\hat{\theta}) \leq \phi(\theta) \leq -h_{\alpha/2} + \phi(\hat{\theta})] = P[h_{\alpha/2} + \phi(\hat{\theta}) \leq \phi(\theta) \leq h_{1-\alpha/2} + \phi(\hat{\theta})] = \\ &= P[\phi^{-1}(h_{\alpha/2} + \phi(\hat{\theta})) \leq \theta \leq \phi^{-1}(h_{1-\alpha/2} + \phi(\hat{\theta}))] \end{aligned} \quad (3)$$

Comparando (2) y (3) vemos que el bootstrap proporciona estimadores de los extremos en (3), sin necesidad de conocer la transformación ϕ ni la f.D. H .

Sin embargo, no siempre existe una función ϕ como la descrita, de ahí el sesgo y la ineficiencia del método percentil, más aún con muestras pequeñas.

27

36

Método bootstrap-t

Para mejorar el rendimiento, una buena estrategia es trabajar con estadísticos bootstrap pivotales, es decir, su distribución no debe depender del parámetro θ .

En general es razonable esperar que el siguiente estadístico sea pivotal:

$$Z = \frac{\hat{\theta} - \theta}{es(\hat{\theta})}$$

La aplicación del bootstrap conduce a definir $Z^* = \frac{\hat{\theta}^* - \hat{\theta}}{es^*(\hat{\theta}^*)}$

De este modo, el Intervalo de Confianza que se define a partir de Z , donde intervienen los percentiles de la distribución de $\hat{\theta}$

$$(\hat{\theta} - \xi_{1-\alpha/2} es(\hat{\theta}), \hat{\theta} - \xi_{\alpha/2} es(\hat{\theta}))$$

sería aproximado por

$$(\hat{\theta} - \xi_{1-\alpha/2}^* es(\hat{\theta}), \hat{\theta} - \xi_{\alpha/2}^* es(\hat{\theta}))$$

A su vez la desviación típica del estadístico puede ser estimada por el bootstrap:

$$(\hat{\theta} - \xi_{1-\alpha/2}^* es^*(\hat{\theta}), \hat{\theta} - \xi_{\alpha/2}^* es^*(\hat{\theta}))$$

Como es habitual, la distribución bootstrap será aproximada mediante la generación de B muestras.

28

37

Funciona mejor con parámetros de varianza estabilizada, es decir, cuando el error típico bootstrap es constante. Se presenta seguidamente un procedimiento para estabilizar la varianza.

En primer lugar, se generan B_1 muestras bootstrap

A partir de cada una de estas B_1 muestras bootstrap, se generan B_2 muestras bootstrap, lo que se conoce como doble bootstrap, obteniendo

$$\{\hat{\theta}_{j1}^{**}, \dots, \hat{\theta}_{jB_2}^{**}; j=1,2,\dots, B_1\}$$

Se estima el error estándar a partir de la cuasidesviación típica de los B_2 valores de cada muestra bootstrap:

$$es(\theta / \hat{\theta} = \hat{\theta}_j^*) = es^*(\hat{\theta}_j^*) = \left\{ \frac{1}{B_2 - 1} \sum_{k=1}^{B_2} (\hat{\theta}_{jk}^{**} - \bar{\theta}_j^{**})^2 \right\}^{1/2} \quad j = 1, \dots, B_1$$

La siguiente nube de puntos puede ser ajustada por un algún procedimiento no paramétrico: $(\hat{\theta}_b^*, es^*(\hat{\theta}_b^*))$. Se obtiene así una función s : $es^*(\hat{\theta}_b^*) \cong s(\hat{\theta}_b^*)$

Por otra parte, se sabe que si X es una v.a. con media θ y desviación típica $s(\theta)$, aplicando un desarrollo de Taylor, para una transformación $g(X)$ se tiene

$$\text{var}(g(X)) \approx g'(\theta)^2 s^2(\theta)$$

29

38

Si se desea que la varianza de $g(X)$ sea constante, igualando la anterior expresión a una constante e integrando se llega a la siguiente transformación, que conduce a que $\text{var}(g(X))$ sea constante e igual a 1, siendo a una constante, y $1/s(u)$ es continua en $[a,x]$:

$$g(x) = \int_a^x \frac{1}{s(u)} du$$

Aplicando este método a los pares $(\hat{\theta}_b^*, \hat{s}^*(\hat{\theta}_b^*))$ se tiene por tanto una transformación \hat{g} (la integral puede ser aproximada por métodos numéricos).

A continuación se generan B_3 muestras bootstrap extraídas de la muestra inicial, y se aplica el método bootstrap-t para calcular un intervalo de confianza para $\hat{g}(\theta)$. Como este error estándar es aproximadamente constante e igual a 1, se utiliza

$$Z^* = \hat{g}(\hat{\theta}^*) - \hat{g}(\hat{\theta})$$

Finalmente, los extremos de este intervalo son transformados mediante \hat{g}^{-1} para obtener un intervalo de confianza para θ .

30

39

Método BCa (sesgo corregido y acelerado)

Para un buen funcionamiento del método percentil es necesario que el estimador transformado $\phi(\hat{\theta})$ sea insesgado y que su varianza no dependa de θ . El método BCa incluye en la transformación dos términos que ayudan a satisfacer ambas condiciones, obteniendo así una cantidad pivotal aproximada.

Supongamos que existe una función monótona creciente ϕ y dos constantes a y b de modo que la siguiente transformación U siga una ley $N(0,1)$, donde el denominador se supone positivo:

$$U = \frac{\phi(\hat{\theta}) - \phi(\theta)}{1 + a\phi(\hat{\theta})} + b$$

Cuando $a=b=0$ se tiene el método percentil. Aplicando el bootstrap, U^*

$$U^* = \frac{\phi(\hat{\theta}^*) - \phi(\hat{\theta})}{1 + a\phi(\hat{\theta})} + b \approx N(0,1)$$

Por tanto para cualquier cuantil Z_α de la $N(0,1)$:

$$\alpha \approx P^*[U^* \leq Z_\alpha] = P^*\left[\hat{\theta}^* \leq \phi^{-1}\left(\phi(\hat{\theta}) + (Z_\alpha - b)[1 + a\phi(\hat{\theta})]\right)\right]$$

Como el cuantil de la distribución bootstrap puede ser conocido al menos aproximadamente, $\xi_\alpha^* \approx \phi^{-1}\left(\phi(\hat{\theta}) + (Z_\alpha - b)[1 + a\phi(\hat{\theta})]\right)$

31

40

Por otra parte

$$1 - \alpha = P[U > Z_\alpha] = P[\theta \leq \phi^{-1}(\phi(\hat{\theta}) + u(a, b, \alpha)[1 + a\phi(\hat{\theta})])], \quad u(a, b, \alpha) = \frac{b - Z_\alpha}{1 - a(b - Z_\alpha)}$$

Comparando las dos últimas expresiones se deduce que si se puede encontrar un valor β tal que $u(a, b, \alpha) = Z_\beta - b$, entonces

$$P[\theta < \xi_\beta^*] \approx 1 - \alpha$$

Una forma directa de lograr la anterior igualdad es

$$\beta = \Phi(b + u(a, b, \alpha)) = \Phi\left(b + \frac{b - Z_\alpha}{1 - a(b - Z_\alpha)}\right) = \Phi\left(b + \frac{b + Z_{1-\alpha}}{1 - a(b + Z_{1-\alpha})}\right)$$

Así, conociendo valores apropiados para a y b , se calcularía el cuantil bootstrap de orden β . Para un intervalo de confianza $100(1-\alpha)\%$ se calcularían

$$\beta_1 = \Phi\left(b + \frac{b + Z_{\alpha/2}}{1 - a(b + Z_{\alpha/2})}\right) \quad \beta_2 = \Phi\left(b + \frac{b + Z_{1-\alpha/2}}{1 - a(b + Z_{1-\alpha/2})}\right)$$

Como ocurría en el método percentil, la especificación de la transformación ϕ no es necesaria, además el método BCa también respeta transformaciones. El I.C. bootstrap BCa viene entonces definido por

$$(\xi_{\beta_1}^*, \xi_{\beta_2}^*)$$

32

41

Elecciones habituales de a y b :

$$b = \Phi^{-1}(\hat{F}^*(\hat{\theta})) \approx \Phi^{-1}\left(\frac{\#\{\hat{\theta}_b^* < \hat{\theta}\}}{B}\right) \quad \Phi \text{ es } f.D.N(0,1)$$

Este valor es conocido como corrección del sesgo. Si la distribución bootstrap está centrada a torno a la estimación original, la distribución bootstrap en ese punto será aproximadamente $1/2$ y en tal caso b será aproximadamente 0.

$$a = \frac{\sum_{i=1}^n (\hat{\theta}_{(i)} - \hat{\theta}_{(i)})^3}{6 \left\{ \sum_{i=1}^n (\hat{\theta}_{(i)} - \hat{\theta}_{(i)})^2 \right\}^{3/2}}$$

a es la aceleración, sirve para estabilizar la varianza del estimador.

Si la aceleración y la corrección del sesgo son 0, BCa coincide con el IC percentil BCa: su precisión es de segundo orden, como el bootstrap-t, mientras que el percentil y el normal sólo lo son de primer orden.

$$\text{Segundo orden} \quad P\{\theta < I\} \approx \frac{\alpha}{2} + \frac{c_I}{n}; \quad P\{\theta > S\} \approx \frac{\alpha}{2} + \frac{c_S}{n}$$

$$\text{Primer orden} \quad P\{\theta < I\} \approx \frac{\alpha}{2} + \frac{c_I}{\sqrt{n}}; \quad P\{\theta > S\} \approx \frac{\alpha}{2} + \frac{c_S}{\sqrt{n}}$$

33

42

Bootstrap

Bootstrap en modelos de regresión

43

BOOTSTRAP EN MODELOS DE REGRESIÓN

3.4. Bootstrap en modelos de regresión.

Se supone un modelo de regresión (no necesariamente lineal)

$$y_i = f(x_i, \beta) + \varepsilon_i \quad i=1, \dots, n; \quad x_i \in R^p; y_i \in R; \beta \in R^q$$

Bootstrap de pares:

Cada muestra bootstrap se obtiene muestreando en el conjunto de casos (registros):

$$(x_1, y_1), \dots, (x_n, y_n) \Rightarrow (x_{i_1}^*, y_{i_1}^*), \dots, (x_{i_n}^*, y_{i_n}^*)$$

$$\hat{\beta}^{*(b)} = \arg \min_{\beta} \sum_{i=1}^n (y_i^* - f(x_i^*, \beta))^2$$

Bootstrap de residuos:

Se muestrean los residuos resultantes del ajuste del modelo sobre la muestra original:

$$e_i = y_i - f(x_i, \hat{\beta}) \quad i=1, \dots, n$$

$$(e_1, \dots, e_n) \Rightarrow (e_1^*, \dots, e_n^*)$$

$$y_i^* = \hat{y}_i + e_i^* \quad i=1, \dots, n$$

$$\hat{\beta}^{*(b)} = \arg \min_{\beta} \sum_{i=1}^n (y_i^* - f(x_i, \beta))^2$$

34

44

En el bootstrap de residuos, las variables predictoras quedan fijadas, algo importante en los modelos estadísticos dentro del Diseño de Experimentos.
En otras situaciones pueden explorarse ambas modalidades.

Algunas estimaciones bootstrap:

$$\hat{\beta}^{*(\cdot)} = \frac{1}{B} \sum_{b=1}^B \hat{\beta}^{*(b)} \quad \hat{V}(\hat{\beta}) = \frac{1}{B-1} \sum_{b=1}^B [\hat{\beta}^{*(b)} - \hat{\beta}^{*(\cdot)}] [\hat{\beta}^{*(b)} - \hat{\beta}^{*(\cdot)}]^t$$

$$\hat{s}_B^*(\hat{\beta}) = \hat{\beta}^{*(\cdot)} - \hat{\beta} \quad \hat{V}(\hat{Y}) = \frac{1}{B-1} \sum_{b=1}^B [\hat{Y}^{*(b)} - \hat{Y}^{*(\cdot)}] [\hat{Y}^{*(b)} - \hat{Y}^{*(\cdot)}]^t$$

En el caso de Regresión Lineal Múltiple, y método Bootstrap de Residuos:

$$\hat{V}_{\infty}^*(\hat{\beta}) = \frac{n-p-1}{n} \hat{\sigma}^2 (X^t X)^{-1}$$

35

45

Bootstrap

Métodos Bootstrap alternativos

46

3.4. Métodos Bootstrap alternativos.

El método II de Efron presenta un error o variabilidad inherente debida al muestreo, por lo que se han desarrollado algunas técnicas que pretenden reducir su efecto, mejorando las estimaciones resultantes.

Bootstrap balanceado

La idea es asegurar que cada observación de la muestra original aparezca con la misma frecuencia en el conjunto de las B muestras bootstrap generadas.

La forma más simple de lograrlo es copiar la muestra B veces, permutar aleatoriamente el vector de tamaño nB resultante, y tomar como muestras bootstrap las B secuencias consecutivas de tamaño n de dicha permutación.

Por ejemplo, se sabe

$$E[\bar{x}^* / F_n] = \frac{1}{n} \sum_{i=1}^n E[X_i^* / F_n] = \frac{1}{n} \sum_{i=1}^n \bar{x} = \bar{x}$$

Pero es poco probable obtener un conjunto de B muestras bootstrap verificando

$$\bar{x}_B^* = \frac{1}{B} \sum_{b=1}^B \frac{1}{n} \sum_{i=1}^n X_{b,i}^* = \bar{x}$$

Con esta variante:

$$\bar{x}_{B,bal}^* = \frac{1}{B} \sum_{b=1}^B \frac{1}{n} \sum_{i=1}^n X_{b,i}^* = \frac{1}{B} \frac{1}{n} \sum_{b=1}^B \sum_{i=1}^n X_{b,i}^* = \frac{1}{B} \frac{1}{n} \sum_{i=1}^n BX_i = \bar{x}$$

36

47

Bootstrap contrario

Sea $X_{(1)}, \dots, X_{(n)}$ la muestra ordenada de menor a mayor.

Dada una muestra bootstrap $X^* = (X_1^*, \dots, X_n^*)$ sea $X^{*c} = (X_1^{*c}, \dots, X_n^{*c})$ donde cada ocurrencia de $X_{(i)}$ en X^* es reemplazada por $X_{(n-i+1)}$.

Por ejemplo, si en X^* hubiera demasiada presencia de los valores más elevados, en la muestra contraria X^{*c} predominarían los valores más bajos.

Se calculan dos estadísticos bootstrap, en general correlacionados negativamente, cuya media se utiliza como nuevo estimador:

$$Q_n^* = \frac{T_n^* + T_n^{*c}}{2} \quad T_n^* = T(X^*) \quad T_n^{*c} = T(X^{*c})$$

La varianza del nuevo estimador es menor o igual que la del estimador bootstrap si la covarianza es negativa:

$$\text{var}(Q_n^*) = \frac{1}{4} (\text{var}(T_n^*) + \text{var}(T_n^{*c})) + 2 \text{cov}(T_n^*, T_n^{*c}) \leq \text{var}(T_n^*)$$

Se puede aplicar en datos multivariantes utilizando criterios apropiados de ordenación.

37

48

Bootstrap

Estimación del error de generalización

49

ESTIMACIÓN DEL ERROR DE GENERALIZACIÓN

3.5. Estimación del error de generalización.

- La estimación del rendimiento de un modelo de predicción no puede efectuarse sobre la muestra de entrenamiento
- Conviene disponer de otros mecanismos para evaluar el rendimiento de un modelo. Por ejemplo, disponer de otra muestra, llamada **muestra o conjunto test** o de prueba, donde se evalúe el rendimiento.
- En caso de no disponer de suficientes datos para llevar a cabo esa partición, se pueden utilizar los métodos de validación cruzada o el bootstrap para estimar el rendimiento esperado del modelo.
- Otra cuestión es la de configurar la complejidad de un modelo de predicción (parámetros o tamaño de un modelo de Minería de Datos, número de términos en una regresión polinomial, etc) donde también pueden emplearse técnicas de remuestreo como validación cruzada.

38

50

Medidas de error:

Dado un modelo de predicción $g(X)$ para una variable dependiente Y , se considera una medida del error de predicción $L(Y, g(X))$. Algunas medidas usuales:

Si Y es cuantitativa (Regresión):

Error Cuadrático: $(Y - g(X))^2$

Error Absoluto: $|Y - g(X)|$

Si Y es cualitativa (Clasificación):

Pérdida 0-1: $I(Y \neq g(X))$

Log. Verosimilitud, Entropía Cruzada o Desviación (K es el número de clases):

$$\begin{aligned} -2 \sum_{k=1}^K I(Y = k) \log \hat{P}[Y = k / X] &= -2 \log \hat{P}_Y(X) \\ g(X) &= \arg \max_{1 \leq k \leq K} \hat{P}[Y = k / X] \end{aligned}$$

39

51

DEFINICIÓN. Dado un modelo de predicción $g_\lambda(x)$, siendo λ un parámetro de control de complejidad, construido a partir de un conjunto de entrenamiento $T_n = \{(x_i, y_i) | i=1, 2, \dots, n\}$, se define el **error empírico o error de entrenamiento**

$$v_{T_n}(\lambda) = \frac{1}{n} \sum_{i=1}^n L(y_i, \hat{g}_\lambda(x_i))$$

En el caso de modelos de regresión y clasificación, se suele trabajar con el ECM y la tasa de error, respectivamente:

$$ECM_{T_n}(\lambda) = \frac{1}{n} \sum_{i=1}^n (\hat{g}_\lambda(x_i) - y_i)^2 \quad TE_{T_n}(\lambda) = \frac{1}{n} \sum_{i=1}^n I(\hat{g}_\lambda(x_i) \neq y_i)$$

DEFINICIÓN. Dado un modelo de predicción $g_\lambda(x)$, ajustado sobre un conjunto de entrenamiento T_n , se define el **error de generalización**

$$Err_{T_n}(\lambda) = E_{(X,Y)}[L(Y, \hat{g}_\lambda(X)) / T_n]$$

Si se toma valor esperado sobre los conjuntos de entrenamiento posibles, se tiene el error esperado:

DEFINICIÓN. Dado un modelo de predicción $g_\lambda(x)$, y dado un tamaño muestral n , se define el **error esperado**

$$Err(\lambda) = E_{T_n} \{ Err_{T_n}(\lambda) \} = E_{T_n} \{ E_{(X,Y)} [L(Y, \hat{g}_\lambda(X)) / T_n] \}$$

40

52

Volviendo al problema de la estimación del error esperado, se utiliza el bootstrap cuando el conjunto de entrenamiento es de tamaño reducido y no es aconsejable la subdivisión de los datos disponibles.

Una primera idea consiste en generar B muestras bootstrap a partir de la muestra de entrenamiento, ajustar el modelo sobre cada muestra, calcular el error empírico mediante cada uno de los modelos bootstrap y obtener finalmente la media de los B errores:

$$\hat{Err}_B^* = \frac{1}{n} \frac{1}{B} \sum_{b=1}^B \sum_{i=1}^n L(y_i, g_\lambda^{*b}(x_i))$$

Se supone que previamente la complejidad del modelo ya ha sido determinada, por ejemplo usando validación cruzada.

Inconveniente: Las muestras bootstrap y el conjunto de entrenamiento tienen observaciones en común, lo que produce un importante sesgo en el estimador bootstrap.

41

53

Una alternativa consiste en:

$$\hat{Err}_B^{*(1)} = \frac{1}{n} \sum_{i=1}^n \frac{1}{|C^{(-i)}|} \sum_{b \in C^{(-i)}} L(y_i, g_\lambda^{*b}(x_i))$$

donde $C^{(-i)}$ denota el conjunto de índices de las muestras bootstrap que no incluyen al caso i . Se suele conocer este estimador como **estimador bootstrap VC**, o **también estimador OOB (Out Of the Bag)**.

Este estimador tiene un sesgo optimista, por lo que se propuso el estimador bootstrap **0.632**:

$$\hat{Err}_B^{*(.632)} = 0.368 \nu_{T_n} + 0.632 \hat{Err}_B^{*(1)}$$

Nótese que

$$P[(x_i, y_i) \in m.b.b] = 1 - (1 - 1/n)^n \approx 1 - 1/e = 0.632$$

Según algunos autores, puede fallar en situaciones de sobreajuste.

42

54

Una mejora propuesta por Efron y Tibshirani para tener en cuenta el posible sobreajuste es el estimador bootstrap **0.632+**:

$$\hat{Err}_B^{*(.632+)} = (1-w)\nu_{T_n} + w\hat{Err}_B^{*(1)}$$

$$w = \frac{0.632}{1 - 0.368 \cdot tsr}$$

En esta expresión interviene la tasa de sobreajuste relativo

$$tsr = \frac{\hat{Err}_B^{*(1)} - \nu_{T_n}}{Noinf - \nu_{T_n}}$$

Noinf es una estimación del error en el supuesto de independencia entre x e y .

$$Noinf = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n L(y_i, \hat{g}_\lambda(x_j))$$

En un problema de clasificación con K categorías, con error 0-1, se verifica

$$Noinf = \sum_{l=1}^K \hat{p}_l(1 - \hat{q}_l), \quad \hat{p}_l = \frac{\#\{y_i = l\}}{n}, \quad \hat{q}_l = \frac{\#\{\hat{g}_\lambda(x_i) = l\}}{n}$$

43

55

En general $tsr \in [0,1]$:

$tsr=0$: Nada de sobreajuste,

$$\hat{Err}_B^{*(1)} = \nu_{T_n} \quad \text{en este caso } w=0.632$$

$tsr=1$: Gran sobreajuste,

$$\hat{Err}_B^{*(1)} = Noinf \quad \text{en este caso } w=1$$

Por tanto $\hat{Err}_B^{*(.632+)}$ en general toma valores entre $\hat{Err}_B^{*(1)}$ y ν_{T_n}

Teniendo en cuenta la definición del estimador 0.632, también se puede expresar

$$\hat{Err}_B^{*(.632+)} = \hat{Err}_B^{*(.632)} + (\hat{Err}_B^{*(1)} - \nu_{T_n}) \frac{0.368 \cdot 0.632 \cdot tsr}{1 - 0.368 \cdot tsr}$$

44

56

En ciertas ocasiones tsr puede tomar valores fuera del intervalo $[0,1]$, en concreto:

$$\begin{cases} Noinf < v_{T_n} \Rightarrow tsr < 0 \\ v_{T_n} < Noinf < \hat{Err}_B^{*(1)} \Rightarrow tsr > 1 \end{cases}$$

Para evitar este problema se consideran las siguientes correcciones.

$$\hat{Err}_B^{*(1)'} = \text{Min}(\hat{Err}_B^{*(1)}, Noinf)$$

$$tsr' = \begin{cases} tsr & \text{si } \hat{Err}_B^{*(1)}, Noinf > v_{T_n} \\ 0 & \text{c.c} \end{cases}$$

$$\hat{Err}_B^{*(.632+)} = \hat{Err}_B^{*(.632)} + (\hat{Err}_B^{*(1)'} - v_{T_n}) \frac{0.368 \cdot 0.632 \cdot tsr'}{1 - 0.368 \cdot tsr'}$$

45