Estadística No Paramétrica

Clase 12: Bootstrap

Joaquin Cavieres G.

Ingeniería en Estadística

Facultad de Ciencias, Universidad de Valparaíso

Introducción

Introducción

El método bootstrap propuesto por Efron (1979) es un procedimiento de remuestreo basado sobre el supuesto de observaciones independientes. En diversos análisis se ha demostrado que el método Bootstrap tiene bastante éxito, por lo que es posible aceptarlo como una alternativa a los métodos asintóticos. La idea clave es realizar cálculos sobre los datos en sí para estimar la variación de los estadísticos que se calculan a partir de los mismos.

Introducción

Usos generales:

- Para evaluar la incertidumbre asociada a una estimación cuando existen pequeños cambios en los datos, el método de Bootstrap es uno de los métodos más comunmente utilizados (además del método de Jackknife).
- El Bootstrap es una idea extremadamente importante en las estadística no paramétrica; de hecho, Casella y Berger (2002) lo llaman "quizás el desarrollo más importante en metodología estadística en los últimos tiempos"

Fundamentación del método de Bootstrap

- Si estamos interesados en evaluar la varianza de un estimador de la forma $\hat{\theta} = \theta(\boldsymbol{x})$

Fundamentación del método de Bootstrap

- Si estamos interesados en evaluar la varianza de un estimador de la forma $\hat{\theta} = \theta(\boldsymbol{x})$
- La expresión matemática para estimar ese parámetro podría ser:

$$\mathbb{V}(\hat{\theta}) = \int ... \int \{\theta(x_1, ..., x_n - \mathbb{E}(\hat{\theta})\}^2 dF(x_1)...dF(x_n)$$

Sin embargo, aquí nos encontramos con dos problemas: 1) No conocemos la distribución F y el segundo es que 2) la integral es dificil de evaluar análiticamente

Fundamentación del método de Bootstrap

- Para el caso 1) podriamos utilizar el principio plug-in (llamado '*ideal bootstrap estimate*')
- Para el caso 2), asumiendo que $\hat{F}$ es discreto, entonces $\mathbb{V}(\hat{\theta}) = \sum_j 1/n^n \{\theta(\boldsymbol{x}_j) - \hat{\mathbb{E}}(\hat{\theta})\}^2$

donde $\boldsymbol{x}_j$ son los rangos a través de todos los n^n posibles combinaciones de los datos observados $\{x_i\}$.

Introducción

El Bootstrap es un método de simulación el cual considera los datos observados para realizar la inferencia. La idea principal es asumir que si la muestra se aproxima a la función de distribución de una población de la cual fue tomada, entonces podemos hacer simulaciones sobre la misma muestra para obtener un estadístico de interés junto a la incertidumbre asociada.

Introducción

En resumen:

- El método de Bootstrap sirve para estimar la varianza y la distribución de un estadístico.
- También sirve para construir los intervalos de confianza asociados al estadístico.

Plug-in estimador

Plug-in estimador

El *plug-in* nos permite estimar parámetros a partir de los datos observados (muestras), por ejemplo, la estimacion de $\theta = f(E)$ se define como:

$$\hat{\theta} = f(E_n)$$

Esto quiere decir que podemos estimar $\theta = f(E)$ desde una funcion de distribucion E con la misma función pero utilizando la función de distribución empírica $\hat{\theta} = f(E_n)$

Como el método de *plug-in* sólo nos permite realizar la estimación de θ entonces podemos utilizar el Bootstrap para determinar la incertidumbre asociada a este método

Método de Bootstrap

Método de Bootstrap

Una muestra aleatoria $x_1, x_2, ..., x_n$ son $i.i.d$ provenientes desde una distribución poblacional F, por lo que nosotros estamos interesados en encontrar:

$$H_n(x) = \mathbb{P}\{R_n \leq x\},$$

donde $R_n = R_n(T_n, F)$ es el valor funcional real de F y $T_n = T_n(x_1, x_2, ..., x_n)$ el estadístico de interés.

Definición Bootstrap

Se tiene $x_1^*, x_2^*, ..., x_n^*$ muestras aleatorias $i.i.d$ desde la distribución empírica F_n basada en los datos observados $x_1, x_2, ..., x_n$, entonces:

$$T_n^* = T_n(x_1^*, x_2^*, ..., x_n^*)$$

y

$$R_n^* = R_n(T_n^*, F_n)$$

Método de Bootstrap

F_n (distribución empírica) esta construida por reemplazar $1/n$ en cada observación x_i, esto es:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{I}(X_i \leq x)$$

Un estimador de H_n entonces sería:

$$\hat{H}_n(x) = \mathbb{P}_*\{R_n^* \leq x\},$$

para $x_1, x_2, ..., x_n$, donde $\mathbb{P}_*$ es la probabilidad condicional respecto a las muestras generadas por el método de Bootstrap. Como las muestras son generadas desde F_n, este método es llamado el método de Bootstrap no paramétrico.

Bootstrap estimador de la varianza

Consideremos a $\mathbb{V}_F(T_n)$ como la varianza del estadístico T_n. Aquí el subíndice $_F$ enfátiza que la varianza está en función de F. Si nosotros conocieramos F podriamos calcular fácilmente la varianza, por ejemplo si $T_n = n^{-1}\sum_{i=1}^{n} X_t$, entonces:

$$\mathbb{V}_F(T_n) = \frac{\sigma^2}{n} = \frac{\int x^2 dF(x) - (\int x dF(x))^2}{n},$$

la cual claramente está en función de F.

Método de Bootstrap

Con el Bootstrap podemos estimar $\mathbb{V}_F(T_n)$ mediante $\mathbb{V}_{\hat{F}_n}(T_n)$, esto es, usar el estimador *plug-in* de la varianza.

Bootstrap estimador de la varianza

- Considerar $x_1^*, ..., x_n^* \sim \hat{F}_n$
- Calcular $T_n^* = g(x_1^*, ..., x_n^*)$
- Repetir los puntos 1 y 2 B veces hasta $T_{n,1}^*, ..., T_{n,B}^*$
- Hacer:

$$v_{\text{boot}} = \frac{1}{B}\sum_{b=1}^{B}\left(T_{n,B}^* - \frac{1}{B}\sum_{r=1}^{B} T_{n,r}^*\right)^2$$

Método de Bootstrap

El método de Bootstrap también nos permite estimar la CDF de $\hat{\theta}$.

CDF para $\hat{\theta}$

Estimación del sesgo (bias):

- $b_{boot} = \bar{\theta}^* - \hat{\theta}$, donde $\bar{\theta}^* = B^{-1} \sum_b \hat{\theta}^*_b$
- Se puede utilizar para estimar cualquier caracteristica del muestreo sobre $\hat{\theta}$.
- Si G es la CDF de $\hat{\theta}$, entonces para cualquier t

$$\hat{G}(t) = \frac{1}{B} \sum_{b=1}^{B} I(\hat{\theta}^*_b \leq t)$$

- Si $\theta = T(F)$, entonces $\hat{G}(t)$ es un estimador consistente de G.

Bootstrap para los intervalos de confianza

Lo anterior discutido previamente nos indica la forma de usar Bootstrap y la estimación de la incertidumbre asociada a un parámetro de interés, por tanto:

- Los errores estándar a menudo son utilizados para construir intervalos basados en la estimación que tiene una distribución Normal (en la muestra)

  $$\hat{\theta} \pm z_{1-\alpha/2}\text{se}$$

  (o alternativamente el intervalo puede estar construido basado en a distribución t)
- Así los se pueden ser utilizados aquí también.

Método de Bootstrap

Bootstrap para los intervalos de confianza

Como Bootstrap también puede ser utilizado para estimar la CDF (G) de $\hat{\theta}$, nosotros podemos no asumir ningún supuesto respecto a $\hat{\theta}$, así lo podemos estimar dentro del procedimiento para los intervalos de confianza.

Método de Bootstrap

Bootstrap para los intervalos de confianza

Ejemplo

Suponga que los se de un estimador varía con el tamaño de la estimación, así, podemos estimar los se para cada replicación del Bootstrap (*Bootstrap t-interval*)

Bootstrap para los intervalos de confianza

- Por cada muestra del Bootstrap, calculamos:

$$z_b^* = \frac{\hat{\theta}_b^* - \hat{\theta}}{\hat{\text{se}}_b^*},$$

donde $\hat{\text{se}}_b^*$ es el error estándar estimado de $\hat{\theta}^*$ basado en el b-ésimo muestra del Bootstrap.

- Estimar el α percentil de z^* mediante el valor $\hat{t}_\alpha$ tal que:

$$B^{-1} \sum_b I(z_b^* \leq \hat{t}_\alpha) = \alpha$$

- Un intervalo de confianza $1 - \alpha$ para θ entonces es:

$$(\hat{\theta} - \hat{t}_{1-\alpha/2}\hat{\text{se}}, \hat{\theta} - \hat{t}_{\alpha/2}\hat{\text{se}}$$

Método de Bootstrap

Ejemplo en R