Selección de h

jueves, 15 de noviembre de 2018 17:01

Usando distribuciones de referencia

Vimos que mediante el A-MISE que

$$h_0 = \left(\frac{\|K\|_2^2}{\|f''\|_2^2 (\mu_2(K))^2 n} \right)^{1/5}$$

con $\mu_2(K) = \int u^2 K(u)\, du$

En esta ecuación f'' es desconocida pero el resto es conocido.

Podemos para realizar el cálculo suponer que f pertenece a una familia paramétrica y con ello intentar estimar $\|f''\|_2^2$

Veamos un ejemplo, consideremos K el núcleo normal al cual notamos φ

Supongamos que f pertenece a la familia normal de media μ y varianza σ^2 desconocidas. Entonces

$$\|f''\|_2^2 = \sigma^{-5} \int (\varphi''(x))^2\, dx = \sigma^{-5} \frac{3}{-} \cong 0{,}212\, \sigma^{-4}$$

A partir de esto $h_0 = \left(\frac{4}{3}\frac{\hat{\sigma}^5}{n}\right)^{1/5} \cong 1{,}06\,\hat{\sigma}\,n^{-1/5}$

Otra manera es estimar f'' mediante un método no paramétrico y luego ponerlo en la fórmula.

Validación cruzada

Hay 2 maneras
a) V.C mediante máxima verosimilitud
b) VC por mínimos cuadrados

a) Sea $f_n(x)$ un estimador de f y consideremos la razón de verosimilitudes $\frac{f(x)}{f_n(x)}$

Para una buena elección de h esta razón estará cerca de 1. Podemos entonces decir que

$$E_*\left(\log \frac{f(x)}{f_n(x)}\right) = 0$$

Podemos definir la información de Kullback-Leibler mediante

mediante

$$d_{KL}(f, f_n) = \int \log\left(\frac{f}{f_n}\right)(x)\, f(x)\, dx$$

Entonces una buena idea es buscar h la cual minimiza $d_{KL}(f, f_n)$. No podemos calcularla pues no conocemos f

Supongamos que son dadas un conjunto adicional de observaciones X_i independientes de los otros.

La verosimilitud de estas observaciones sería $\prod_i f_n(X_i)$

El valor de este estadístico para diferentes h puede indicar cual h es preferible dado que

$$d_{KL}(f, f_n) = \int \log f(x)\, f(x)\, dx - \int \log f_n(x)\, f(x)\, dx$$

o sea maximizando el log de la verosimilitud minimizamos d_{KL}

Usualmente no agregamos observaciones adicionales

Una forma de lidiar con esto es estimar f_n en el subconjunto $\{X_j\}_{j\neq i}$ y calcular la verosimilitud para X_i

Notamos el estimador "leave-one-out" mediante

$$f_{n,i}(X_i) = \frac{1}{\ldots} \sum_{j\neq i} K\left(\frac{X_i - X_j}{h}\right)$$

$(n-1)h \quad j \neq i$

$$\prod_{i=1}^{n} \hat{f}_{n,i}(X_i) = \frac{1}{(n-1)^n h^n} \prod_{i=1}^{n} \sum_{j \neq i} K\left(\frac{X_i - X_j}{h}\right)$$

Si tomamos el logaritmo y dividimos por n tenemos

$$CV_{KL}(h) = n^{-1} \sum_{i=1}^{n} \log\left(\hat{f}_{n,i}(X_i)\right) =$$

$$= n^{-1} \sum_{i=1}^{n} \log\left[\sum_{j \neq i} K\left(\frac{X_i - X_j}{h}\right)\right] - \log\left[(n-1)h\right]$$

y hallamos $\hat{h}_{KL} = \arg\max_h CV_{KL}(h)$

Si las X_i son idénticamente distribuidas entonces $\log f_{n,i}(X_i)$ son idénticamente distribuidas luego

$$E(CV_{KL}(h)) = E(\log f_{n,i}(X_i))$$

Sacando el efecto del "leave-one-out"

$$E(CV_{KL}(h)) \simeq E\left(\int \log f_n(x)\, f(x)\, dx\right)$$

Pero $E(d_{KL}(f, f_n)) = \int f \log f\, dx - E\left(\int \log f_n(x)\, f(x)\, dx\right)$

$$\cdot \log f_n(x)\, f(x)\, dx)$$

o sea $$E\left(CV_{KL}(h)\right) = -E\left(d_{KL}(f, f_n)\right) + \int \log f(x)\, f(x)\, dx$$

b) Definimos el error cuadrático integrado como

$$d_I(h) = \int (f_n - f)^2 dx = \underbrace{\int f_n^2(x)\, dx}_{\text{Computable a partir de los datos}} - 2\int (f_n f)(x)\, dx + \underbrace{\int f^2(x)\, dx}_{\text{no depende de } h}$$

entonces minimizar $d_I(h)$ es equivalente a minimizar

$$\int f_n^2(x)\, dx - 2\int f_n(x)\, f(x)\, dx$$

Ahora $\int f_n(x)\, f(x)\, dx = E_X\left(f_n(X)\right)$

Para estimar este término $\widehat{E_X\left(f_n(X)\right)} = n^{-1} \sum_{i=1}^{n} f_{n,i}(X_i)$

Definimos $$CV(h) = \int f_n^2(x)\, dx - \frac{2}{n} \sum_{i=1}^{n} f_{n,i}(X_i)$$

$$\hat{h}_{CV} = \underset{h}{\operatorname{argmin}}\ CV(h)$$

Stone demostró en 1982 que si f es acotada $\hat{h}_{CV}$ es asintóticamente óptimo o sea

$$\frac{d_I(h_{cv})}{\inf_{h>0} d_I(h)} \xrightarrow{c.s} 1 \quad (n \to \infty)$$

Algoritmo

Para computar se muestra que

$$\int \hat{f}_h^2(x)\,dx = n^{-2} h^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} K * K\left(\frac{X_j - X_i}{h}\right)$$

donde $K * K(u)$ es la convolución del ~~núc~~ núcleo K

(dem a cargo del lector)

$$CV(h) = n^{-2} h^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{X_j - X_i}{h}\right) - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{n,i}(X_i)$$

$$= \frac{2}{n^2 h}\left[\frac{n}{2} K * K(0) + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} K * K\left(\frac{X_j - X_i}{h}\right) - \frac{2n}{n-1} K\left(\frac{X_j - X_i}{h}\right)\right]$$