Aprendizaje Estadístico

Modelado y Análisis de Redes de Telecomunicaciones

IIE - Facultad de Ingeniería

Curso 2014

Estimación de Densidades

- Sean $X_1, X_2, \ldots, X_n$ datos *iid* con distribución F
- El teorema de Glivenko-Cantelli prueba que la distribución empírica F_n converge (uniformemente) a la dist. real F:

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

- Es un estimador insesgado y es el de menor varianza
- Sin embargo, F_n siempre es discontinua
- Si además se asume que F es absolutamente continua, es decir existe una densidad f tal que $F(x) = \int_0^x f(t)dt$, interesa estimar f
- Estimarla a partir de F_n no da ninguna información ($1/n$ en cada punto)
- Opción más simple: histogramas

Histogramas

- Dada una partición del soporte de f, $\{B_i\}_{i=1,\ldots,k}$ con $B_i = [t_i, t_{i+1}]$, el histograma se define como:

$$\widehat{f}(x) = \sum_{i=1}^{k} \frac{N_i/n}{t_{i+1} - t_i} \mathbf{1}_{B_i}(x)$$

donde N_i es la cantidad de datos en B_i. Si la longitud de los intervalos es siempre la misma $h_n = t_{i+1} - t_i$ (ancho de la ventana), resulta:

$$\widehat{f}(x) = \frac{1}{nh_n} \sum_{i=1}^{k} N_i \mathbf{1}_{B_i}(x)$$

Histogramas

- Dada una partición $\{B_i\}_{i=1,\dots,k}$ con $B_i = [t_i, t_{i+1}]$ del soporte de f, el histograma se define como:

$$\widehat{f}(x) = \sum_{i=1}^{k} \frac{N_i/n}{t_{i+1} - t_i} \mathbf{1}_{B_i}(x)$$

donde N_i es la cantidad de datos en B_i. Si la longitud de los intervalos es siempre la misma $h_n = t_{i+1} - t_i$ (ancho de la ventana), resulta:

$$\widehat{f}(x) = \frac{1}{nh_n} \sum_{i=1}^{k} N_i \mathbf{1}_{B_i}(x)$$

- Se puede hallar la ventana óptima h_n tal que minimice algún error, por ejemplo el MISE

$$MISE = \int E(\widehat{f}(x) - f(x))^2 dx = \mathrm{Var}(\widehat{f}(x)) + (E(\widehat{f}(x) - f(x)))^2$$

$$= \int \mathrm{Var}(\widehat{f}(x)) + \int \mathrm{Sesgo}(\widehat{\mathrm{f}}(\mathrm{x}))^2$$

Estimador Naif

- Rosenblatt (1956) propone como estimador

$$\widehat{f}(x) = \frac{F_n(x+h) - F_n(x-h)}{2h} = \frac{\sharp\{X_i : X_i \in (x-h, x+h]\}}{2nh}$$

$$= \frac{1}{2nh} \sum_{i=1}^{n} \mathbf{1}_{\{X_i - h \leq x < X_i + h\}}$$

- La idea es que:

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X \leq x + h)$$

- El estimador puede generalizarse de la forma:

$$\widehat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} w\left(\frac{x - X_i}{h}\right)$$

donde $w(y) = \begin{cases} 1/2 & y \in [-1, 1) \\ 0 & \text{otro caso} \end{cases}$

Estimación por Núcleos

- La función w puede generalizarse como:

$$\widehat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}K\left(\frac{x - X_i}{h}\right)$$

  donde K se llama función núcleo o kernel.
- La función K cumple ciertas condiciones de regularidad de manera que f_n converge a f en algún sentido
- Además h_n es una sucesión de constantes positivas (ancho de la ventana, bandwidth): cuando h_n tiende a cero se tienen las deltas.
- La idea es que el kernel le da peso a los puntos más cercanos y suaviza la estimación

Lema de Bochmer

Si la función K cumple las siguientes condiciones:

- K acotado
- $\int K(x)dx = 1$
- simétrico $K(-x) = K(x)$
- $|xK(x)| \to 0$ cuando $|x| \to \infty$

entonces se cumple que

$$\widehat{f}_n(x) \to f(x) \quad \text{cuando } n \to \infty \quad \forall x \text{ pto de continuidad}$$

$$E(\widehat{f}_n(x)) \to f(x) \quad \text{cuando } h_n \to 0$$

$$\widehat{\text{Var}(f_n(x))} \to 0 \quad \text{cuando } h_n \to 0 \text{ y } nh_n \to \infty$$

- Compromiso sesgo/varianza: si h_n converge a cero muy rápido, el sesgo disminuye pero la varianza aumenta
- El h_n óptimo converge a cero pero más lento que $1/n$

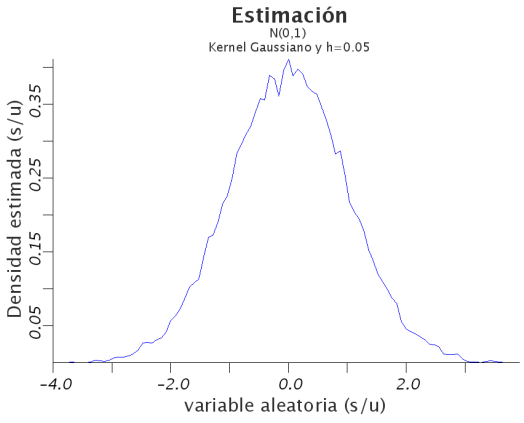
Algunos ejemplos de kernel

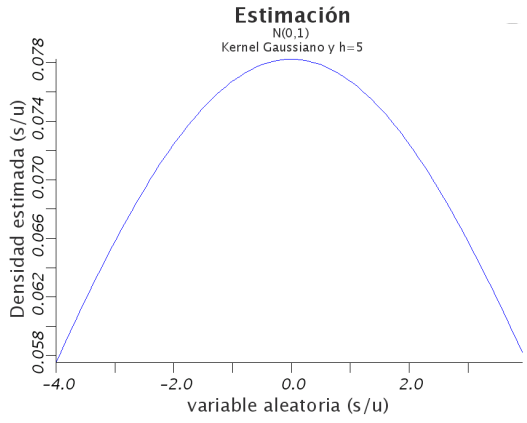
1. Epanechnikov: $\frac{3}{4}(1-t^2)$ si $|t|<1$
2. Gaussiano: $\frac{1}{\sqrt{2\pi}}e^{-t^2/2}$
3. Triangular: $1-|t|$ si $|t|<1$
4. Rectangular: $1/2$ si $|t|<1$
5. Biweight: $\frac{15}{16}(1-t^2)^2$ si $|t|<1$

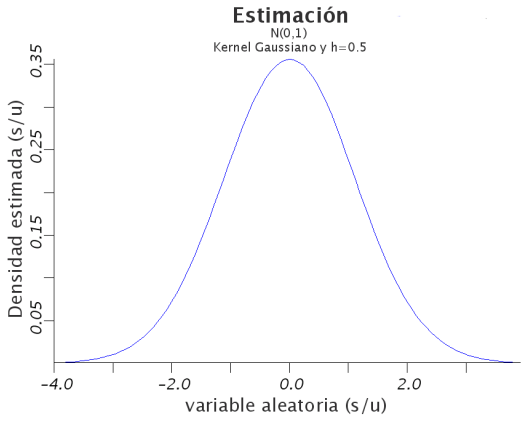
- Todos cumplen con las condiciones del teorema de Bochmer
- Los más usados son los dos primeros
- Epanechnikov tiene buenas propiedades y su cálculo es muy sencillo

Ejemplos

Estimación por kernel gaussiano con $h = 0{,}05$, $h = 5$ y $h = 0{,}5$ de la densidad de una VA normal de media nula y varianza unitaria.







Elección de la ventana óptima

- Hay que elegir h_n de modo que minimice algún tipo de error: e.g. MISE
- Si f tiene derivada segunda continua y acotada, y el núcleo K tiene segundo momento finito, resulta que:

$$\text{Sesgo}(\widehat{f}_n(x)) \approx \frac{1}{2}h_n^2 f''(x)\int u^2 K(u)du = \frac{1}{2}h_n^2 f''(x)K_2$$

$$\text{Var}(\widehat{f}_n(x)) \approx \frac{1}{nh_n}f(x)\int K^2(u)du$$

de donde la ventana óptima h_n^* está dada por:

$$h_n^* = K_2^{-2/5}\left(\int K^2(u)du\right)^{1/5}\left(\int f''(x)^2 dx\right)^{-1/5} n^{-1/5}$$

- Observar que h_n^* depende de f

Elección de la ventana óptima

- "Rule of thumb": asumir f distribución conocida (e.g. gaussiana y kernel gaussiano)

$$h_n^* = (4/3)^{1/5} \sigma n^{-1/5}$$

- "Plug-in": obtener una estimación de f con alguna ventana y sustituir en la ecuación anterior. Iterar hasta que el error no decrezca de manera significativa
- Validación cruzada:
  - se busca minimizar el MISE para lo que basta minimizar $R(f) = E(\int \widehat{f}_n^2) - 2E(\int \widehat{f}_n f)$ y un estimador insesgado es

$$CV(h) = \int \widehat{f}_n^2 - \frac{2}{n} \sum_{i=1}^{n} \widehat{f}_n^{-i}(X_i)$$

    donde $\widehat{f}_n^{-i}(x)$ es la estimación en x usando todos los datos menos el i-ésimo
  - por máxima verosimilitud:

$$CV(h) = \frac{1}{n} \sum_{i=1}^{n} \log \widehat{f}_n^{-i}(X_i)$$

El problema de Regresión

- Se tiene el modelo:

$$Y = r(X) + \epsilon$$

  donde ϵ es una v.a. centrada e independiente de X.
- Se busca construir una estimación de Y a partir de un conjunto $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$
- Se asume que $Y \in \mathbb{R}$, sin embargo X puede ser real, vectorial o incluso funcional
- El modelo de regresión puede ser parámetrico o no parámetrico
  - caso parámetrico: se asume un modelo particular (e.g. lineal) y se busca estimar sólo los parámetros, por ejemplo

$$Y = aX + b + \epsilon$$

    y se busca estimar a y b
  - caso no parámetrico: la función r es continua

Regresión Multidimensional y Funcional

- Caso multidimensional: en este caso $X \in \mathbb{R}^d$ y $r : \mathbb{R}^d \to \mathbb{R}$
  - como antes se pueden asumir modelos parámetricos o no:
  - ejemplo de modelo parámetrico: $Y = A^t X + b + \epsilon$ donde $A \in \mathbb{R}^d$
- Caso funcional: en este caso los datos $X \in E$ siendo E un espacio de dimensión infinita (e.g. curvas aleatorias)
  - un ejemplo de modelo funcional parámetrico es:

$$r \in \mathcal{C} = \{\text{operadores lineales continuos}\}$$

Estimador de Nadaraya-Watson

- Dado un conjunto de entrenamiento $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, se busca estimar $r(x)$ para un x cualquiera.
- Como vimos antes un estimador razonable es la esperanza condicional:

$$\widehat{r}(x) = E(Y|X = x)$$

- Sea $f(x, y)$ a densidad conjunta del vector (X, Y) y $g(x)$ la densidad marginal de X, entonces:

$$E(Y|X = x) = \int yf(y|x)dy = \int y\frac{f(x, y)}{g(x)}dy$$
$$= \frac{1}{g(x)}\int yf(x, y)dy = \frac{\Phi(x)}{g(x)}$$

- El estimador de NW consiste en sustituir $g(x)$ y $\Phi(x)$ por las respectivas estimaciones por núcleos.

Estimador de Nadaraya-Watson

- El estimador de NW resulta entonces:

$$\widehat{r}_n(x) = \frac{\widehat{\Phi}_n(x)}{\widehat{g}_n(x)} = \frac{\frac{1}{nh}\sum_{i=1}^{n} Y_i K(\frac{x-X_i}{h})}{\frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-X_i}{h})}$$

- Valen los mismos comentarios que antes sobre la elección de la ventana óptima (compromiso sesgo/varianza)
- En el caso multidimensional se tiene que:

$$\mathrm{Var}(\widehat{f}_n) \approx n^{-\frac{4}{d+4}}$$

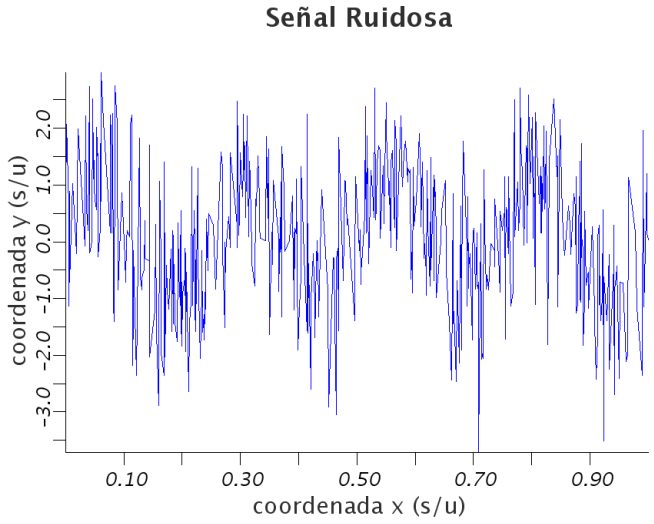
por tanto es necesario un n de orden exponencial en d para mantener una varianza dada.

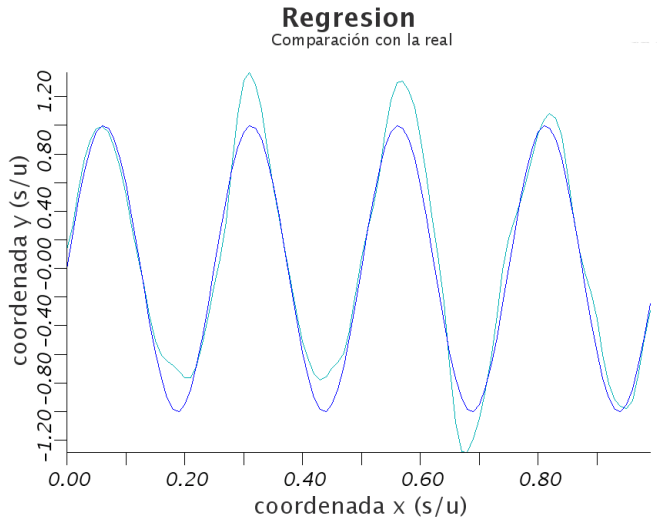
Estimador de Nadaraya-Watson

Se considera el siguiente ejemplo unidimensional:

$$Y_i = r(X_i) + \varepsilon = \sin(2\pi 4 X_i) + \varepsilon$$

donde ε es una va con distribución normal de media nula y varianza unitaria. El kernel utilizado es el de Epanechnikov.





Estimador de Nadaraya-Watson Multidimensional

- Dado $X \in \mathbb{R}^d$, "cerca" implica la definición de alguna distancia
- El estimador de NW resulta entonces:

$$\widehat{r}_n(x) = \frac{\widehat{\Phi}_n(x)}{\widehat{g}_n(x)} = \frac{\frac{1}{nh}\sum_{i=1}^{n} Y_i K(\frac{||x-X_i||}{h})}{\frac{1}{nh}\sum_{i=1}^{n} K(\frac{||x-X_i||}{h})}$$

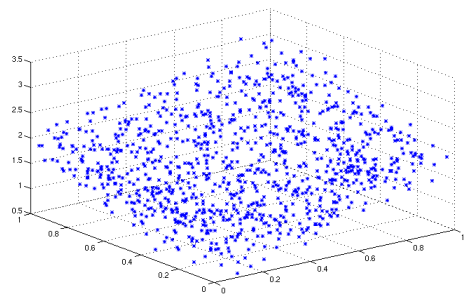
- La norma puede ser cualquiera, y en todos los casos se tiene convergencia del estimador.

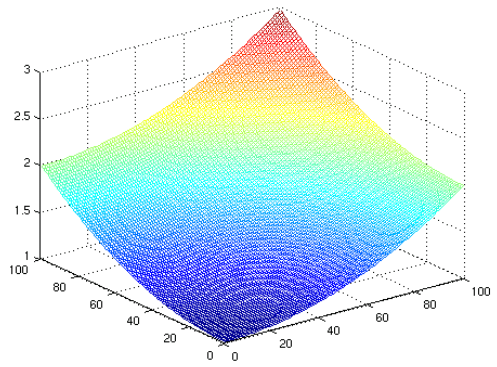
Estimador de Nadaraya-Watson

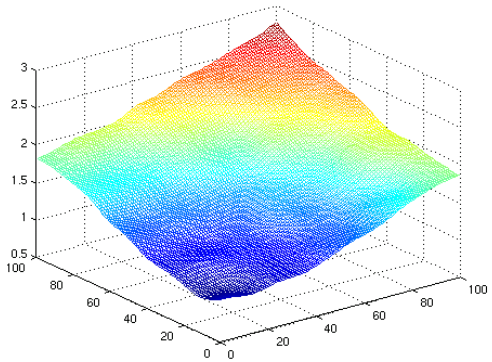
Se considera el siguiente ejemplo de regresión bidimensional:

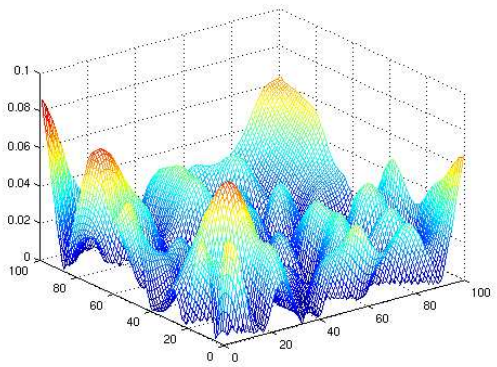
$$Y_i = r(X_i) + \varepsilon = x^2 + y^2 + 1 + \varepsilon$$

donde ε es una va con distribución normal de media nula y varianza 0.25.









Estimador de Nadaraya-Watson Funcional

- Dado $X \in E$, y sea $||\cdot||$ una pseudo-norma (no tiene porque cumplir la desigualdad triangular)
- El estimador de NW resulta entonces:

$$\widehat{r}_n(x) = \frac{\widehat{\Phi}_n(x)}{\widehat{g}_n(x)} = \frac{\frac{1}{nh}\sum_{i=1}^{n} Y_i K(\frac{||x-X_i||}{h})}{\frac{1}{nh}\sum_{i=1}^{n} K(\frac{||x-X_i||}{h})}$$

- Hay hipótesis fuertes sobre la norma (tema de la dimensión y la cantidad de datos)

Elección de la Ventana Óptima NW Multidimensional

- Se considera el error cuadrático medio integrado ponderado (ECMIP):

$$U_n = E \int_{\mathbb{R}^d} (\widehat{r}_n(x) - r(x))^2 \widehat{g}_n^2(x) dx$$

que le da más peso a los valores de x más probables

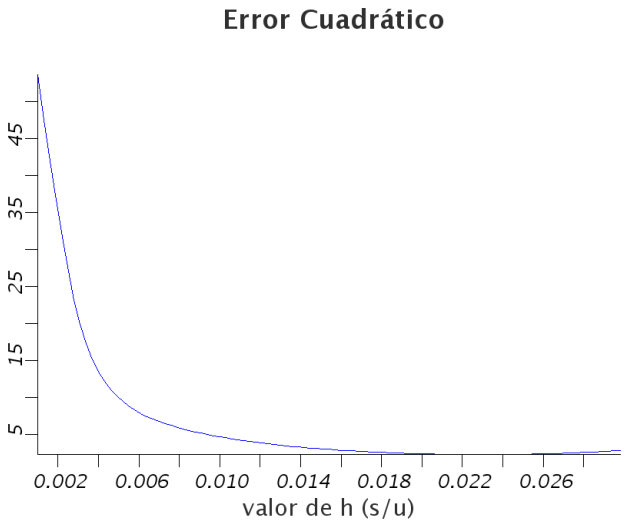
- El valor de h que minimiza U_n está dado por:

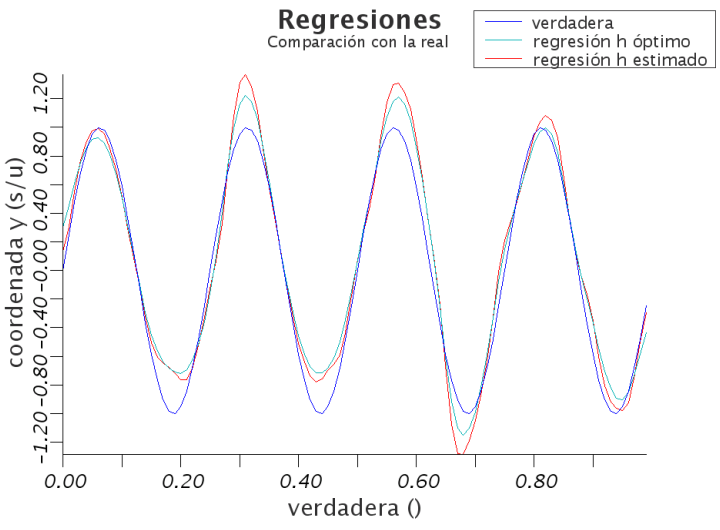
$$h = n^{\frac{-1}{4+p}} \left(\frac{\left(\int_{\mathbb{R}^p} v^2 K(v) dv\right)^2 \int_{\mathbb{R}^p} \left(\sum_{i=1}^p \left(\frac{\partial^2 \Phi(x)}{\partial^2 v_i} - r(x) \frac{\partial^2 g(x)}{\partial^2 v_i} \right) \right)^2 dx}{p \int\limits_{\mathbb{R}^p} K^2(v) dv \int\limits_{\mathbb{R}^p} \left(\frac{V(x)}{g(x)} - r^2(x) \right) g(x) dx} \right)^{\frac{-1}{4+p}}$$

- Cómo antes depende de f, por lo tanto se usa método plug-in explicado antes

Elección de la Ventana Óptima

Comparación de la elección dada por el algoritmo iterativo y la óptima obtenida con una grilla





Referencias

- "Estimación no parámetrica de la función de densidad". Antonio Miñarro. Barcelona, 1988.
- "Regresión no parámetrica. Introducción al estimador de Nadaraya Watson". Federico Larroca, 2006.
- "Cálculo de la ventana óptima para el estimador de Nadaraya Watson". Paola Bermolen, Federico Larroca, 2005.