José Antonio Álvarez Ocete
77553417Q
joseantonio.alvarezo@estudiante.uam.es

Entrega de problemas Boostrap
Métodos Avanzados en Estadística

2021-11-15

Ejercicio 6.

Enunciado. Sean $X_1, \dots, X_n$ variables aleatorias *i.i.d.* de una distribución con densidad f. Se considera el estimador del núcleo $\hat{f}$ con núcleo rectangular $K(x) = \mathbb{I}_{[-1/2,1/2]}(x)$ y parámetro de suavizado h.

a) Calcula el sesgo y la varianza de $\hat{f}(x)$, para un valor de x fijo.

b) Demuestra que tanto el sesgo como la varianza tienden a cero si $h \to 0$ y $nh \to \infty$.

Al estudiar el sesgo y la varianza de $\hat{f}(x)$ para un valor de x fijo, estudiamos dichos valores respecto a la muestra tomada de $X_1, \dots, X_n$. En primer lugar, calculemos la esperanza del núcleo rectangular dado por la función indicatriz:

$$\mathbb{E}\left[K\left(\frac{x-t}{h}\right)\right] = \int_{-\infty}^{+\infty} f(t)\ K\left(\frac{x-t}{h}\right)\ dt$$

Utilizamos el cambio de variable $w = \frac{x-t}{h}, dw = -\frac{dt}{h}$ e invirtiendo los límites de integración obtenemos:

$$\begin{aligned} \mathbb{E}\left[K\left(\frac{x-t}{h}\right)\right] &= \int_{-\infty}^{+\infty} f(t)\ K\left(\frac{x-t}{h}\right)\ dt \\ &= \int_{+\infty}^{-\infty} f(x-wh)\ K(w)\ (-h)\ dw \\ &= h\int_{-\infty}^{+\infty} f(x-wh)\ \mathbb{I}_{[-1/2,1/2]}(w)\ dw \\ &= h\int_{-\frac{1}{2}}^{\frac{1}{2}} f(x-wh) dw \end{aligned} \tag{1}$$

Calculemos el sesgo de nuestro estimador del núcleo haciendo uso de la expresión anterior. Para ello calculamos su esperanza:

$$\begin{aligned} \mathbb{E}[\hat{f}(x)] &= \mathbb{E}\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-t}{h}\right)\right] \\ &\overset{\text{iid}}{=} \frac{1}{nh}\ n\mathbb{E}\left[K\left(\frac{x-t}{h}\right)\right] \\ &\overset{(1)}{=} \frac{1}{h}\ h\int_{-\frac{1}{2}}^{\frac{1}{2}} f(x-wh) dw \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} f(x-wh) dw \end{aligned}$$

Este valor únicamente depende del punto fijado x, el parámetro de suavizado h y la función de densidad f. Su valor equivale al área bajo la gráfica de f en el intervalo $[x - h/2, x + h/2]$, como se puede apreciar en la figura 1.

Figure 1: Representación gráfica del valor $\mathbb{E}[\hat{f}(x)]$.

Es decir, estamos aproximando el valor de una función en un punto por su integral en un intervalor centrado en dicho punto. Naturalmente, al tomar $h \to 0$, dicho valor tiende al valor del punto. Es decir, el sesgo tiende a 0. Analíticamente:

$$\begin{aligned}
\text{Sesgo}(f) &= E[\hat{f}(x)] - f(x) \\
&= \int_{-\frac{1}{2}}^{\frac{1}{2}} f(x - wh)dw - f(x) \\
&\overset{h\to\infty}{\longrightarrow} \int_{-\frac{1}{2}}^{\frac{1}{2}} f(x)dw - f(x) \\
&= f(x) \underbrace{\int_{-\frac{1}{2}}^{\frac{1}{2}} dw}_{=1} - f(x) \\
&= f(x) - f(x) = 0
\end{aligned}$$

Por otro lado, podemos darnos cuenta de lo siguiente:

$$K(x)^2 = K(x) \quad \forall x \in \mathbb{R}, \tag{2}$$

pues la función indicatriz tiene por imagen el conjunto $\{0, 1\}$ y la función $x \mapsto x^2$ sobre este conjunto es la identidad. Calculemos la varianza del estimador del núcleo:

$$\begin{aligned}
\text{Var}[\hat{f}(x)] &= \text{Var}\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-t}{h}\right)\right] \\
&\overset{\text{iid}}{=} \frac{1}{n^2h^2}\, n\text{Var}\left[K\left(\frac{x-t}{h}\right)\right] \\
&= \frac{1}{nh^2}\left(\mathbb{E}\left[K\left(\frac{x-t}{h}\right)^2\right] - \mathbb{E}\left[K\left(\frac{x-t}{h}\right)\right]^2\right) \\
&\overset{(2)}{=} \frac{1}{nh^2}\left(\mathbb{E}\left[K\left(\frac{x-t}{h}\right)\right] - \mathbb{E}\left[K\left(\frac{x-t}{h}\right)\right]^2\right) \\
&\overset{(1)}{=} \frac{1}{nh^2}\left(h \int_{-\frac{1}{2}}^{\frac{1}{2}} f(x-wh)dw - h^2 \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} f(x-wh)dw\right)^2\right) \\
&= \frac{1}{nh}\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} f(x-wh)dw - h \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} f(x-wh)dw\right)^2\right)
\end{aligned}$$

Finalmente probamos que la varianza tiende a 0 si $h \to 0$ y $nh \to \infty$:

$$\text{Var}[\hat{f}(x)] = \underbrace{\frac{1}{nh}}_{\hookrightarrow 0}\left(\underbrace{\int_{-\frac{1}{2}}^{\frac{1}{2}} f(x-wh)dw}_{\hookrightarrow f(x)} - \underbrace{h\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} f(x-wh)dw\right)^2}_{\hookrightarrow 0}\right) \longrightarrow 0$$

Aunque ya probamos que el error cuadrático medio para un núcleo cualquiera tiende a 0, en este ejercicio lo hemos probado para el caso del núcleo indicatriz sin el uso de aproximaciones.

Ejercicio 7.

Enunciado. Considera una variable aleatoria con distribución beta de parámetros $\alpha = 3$, $\beta = 6$.

a) Representa gráficamente la función de densidad y la función de distribución.

Importamos los paquetes necesarios:

```
library(tidyverse)
library(gapminder)
library(comprehenr)
library(ggplot2)
library(dplyr)
library(ggpubr)

defaultW <- getOption("warn")
options(warn = -1)
theme_set(theme_bw())
```

Representamos las funciones especificadas:

```
x = seq(0, 1, length=1000)
pdf = dbeta(x, 3, 6)
cdf = pbeta(x, 3, 6)
df <- data.frame(x, pdf, cdf)

graf1 <- ggplot(df, aes(x=x, y=pdf)) +
  geom_ribbon(aes(ymin=0, ymax=pdf), fill="lightblue", col="blue", alpha=0.5) +
  ylab("Probability Density Function")

graf2 <- ggplot(df, aes(x=x, y=cdf)) +
  geom_ribbon(aes(ymin=0, ymax=cdf), fill="lightblue", col="blue", alpha=0.5) +
  ylab("Cumulative Density Function")

ggarrange(graf1, graf2,
ncol = 2, nrow = 1)
```

1_files/figure-latex/unnamed-chunk-2-1.pdf

b) Simula una muestra de tamañoo 20 de esta distribución. A continuación, representa en los mismos gráficos del apartado (a) las estimaciones de F y f obtenidas respectivamente mediante la función de distribución empírica F_n y un estimador del núcleo $\hat{f}$ obtenidos a partir de la muestra simulada.

```
set.seed(123)

x = seq(0, 1, length=1000)
pdf = dbeta(x, 3, 6)
cdf = pbeta(x, 3, 6)
df <- data.frame(x, pdf, cdf)

muestra <- rbeta(20, 3, 6)
estimador_nucleo <- density(muestra)
df_estimator <- data.frame("x"=estimador_nucleo$x, "y"=estimador_nucleo$y)

graf1 <- ggplot() +
  geom_ribbon(data=df, aes(x=x, y=pdf, ymin=0, ymax=pdf),
fill="lightblue", col="blue", alpha=0.5) +
  geom_line(data=df_estimator, aes(x=x, y=y), col="red") +
  ylab("Probability Density Function") +
  coord_cartesian(xlim = c(0, 1))

graf2 <- ggplot() +
```

```
  geom_ribbon(data=df, aes(x=x, y=cdf, ymin=0, ymax=cdf),
fill="lightblue", col="blue", alpha=0.5) +
  stat_ecdf(data=data.frame(muestra), aes(x=muestra), color="red", geom="step") +
  ylab("Cumulative Density Function") +
  coord_cartesian(xlim = c(0, 1))

ggarrange(graf1, graf2, ncol = 2, nrow = 1)
```

1_files/figure-latex/unnamed-chunk-3-1.pdf

c) Verifica empíricamente el grado de aproximación alcanzado en las estimaciones de F y f. Para ello, genera 200 muestras de tamaño 20 y para cada una de ellas evalúa el error (medido en la norma del supremo, es decir, el máximo de las diferencias entre las funciones) cometido al aproximar F por F_n y f por $\hat{f}$. Por último, calcula el promedio de los 200 errores obtenidos.

```
set.seed(123)

n <- 20
m <- 200
alpha <- 3
beta <- 6

errors_pdf <- NULL
errors_cdf <- NULL
p_values_pdf <- NULL
p_values_cdf <- NULL

for (i in 1:m){
  muestra <- rbeta(n, alpha, beta)

  estimador_nucleo <- density(muestra)
  theoric_pdf_ys <- dbeta(estimador_nucleo$x, alpha, beta)
  ks_pdf <- ks.test(estimador_nucleo$y, theoric_pdf_ys)

  ecdf_estimada <- ecdf(muestra)
  theoric_cdf_ys <- pbeta(muestra, alpha, beta)
  ks_cdf <- ks.test(ecdf_estimada(muestra), "pbeta", alpha, beta)

  errors_pdf <- c(errors_pdf, ks_pdf$statistic)
  p_values_pdf <- c(p_values_pdf, ks_pdf$p.value)
  errors_cdf <- c(errors_cdf, ks_cdf$statistic)
  p_values_cdf <- c(errors_cdf, ks_cdf$p.value)
}

cat("Mean error in cdf: ", mean(errors_cdf), "\n")
```

```
## Mean error in cdf:  0.4115441
```

```
cat("Mean p-value for cdf: ", mean(p_values_cdf), "\n")
```

```
## Mean p-value for cdf:  0.4095036
```

```
cat("Mean error in pdf: ", mean(errors_pdf), "\n")
```

```
## Mean error in pdf:  0.1869434
```

```
cat("Mean p-value for pdf: ", mean(p_values_pdf), "\n")
```

```
## Mean p-value for pdf:  0.0017868
```

```r
Sys.sleep(1)
```