Análisis Avanzado de Datos.

Nicolás López

Repaso funciónes de densidad y distribución

Repaso funciónes de densidad y distribución

Recordando elementos de teoría estadística de la medida (probabilidad):

- Experimento aleatorio.
- Espacio medible: (Ω, F_Ω) donde $A \in F_\Omega$ es un evento:
  - $\Omega \in F_\Omega$
  - Si $A \in F_\Omega$, $A^c \in F_\Omega$
  - Si $A_1, ... \in F_\Omega$, $\cup_i A_i \in F_\Omega$
- Espacio de probabilidad: (Ω, F_Ω, P_F) donde $P : F_\Omega \longrightarrow [0, 1]$:
  - $P_F(A) \geq 0$ pt $A \in F_\Omega$.
  - $P_F(\Omega) = 1$.
  - Para $A_1, ..,$ donde $A_i \cap A_j = \emptyset$ pt $i \neq j$ se tiene $P(\cup_i A_i) = \sum_i P(A_i)$

Note que P está definida en F, pero, ¿en dónde observamos nuestros datos?

Generalmente me interesan números, no eventos, del experimento. Por lo cual hacemos el siguiente mapeo (para el caso univariado):

$$X : \Omega \longrightarrow R$$

Con R el conjunto de números reales:

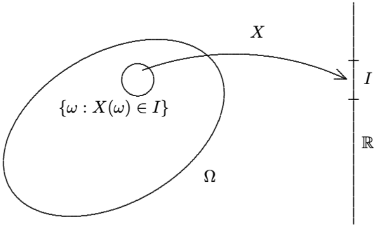


Figure 1: Tomado de:
https://www.cimat.mx/~jortega/MaterialDidactico/probabilidad17/Cap2.pdf

Para todo I, $X^{-1}(I) = \{\omega \in \Omega : X(\omega) \in I\}$ es medible (está en F).

Al **medir** la probabilidad de estos intervalos I (que por cierto, también pertenecen a una estructura de conjuntos, G_R), contamos con medida de probabilidad nueva inducida por X, ahora, transportada a R:

$$P_X : G_R \longrightarrow [0,1] : I \longrightarrow P_X(I) = P(\{\omega \in \Omega : X(\omega) \in I\})$$

Tenemos una medida en los números reales para conjuntos I, llamada la distribución de X.

Cuando tomamos los intervalos I de la forma $(-\infty, x]$, tenemos la función de distribución (acumulada) de X definida sobre R:

$$F_X(x) : R \longrightarrow [0,1] : x \longrightarrow P(\{\omega \in \Omega : X(\omega) \leq x\}) = P_X((-\infty, x])$$

Y la distribución de X (P_X) está determinada de manera única por su correspondiente F_X, por lo cual se suele caracterizar la aleatoriedad de una variable real por su función de distribución.

Si F tiene saltos, X es discreta, y la magnitud del salto en a (arbitrario) indica $P(X = a)$. Y con ello F está completamente caracterizada por la función:

$$P(X = x) = F_X(x) - F_X(x^-)$$

Para todo x en R. Por lo cual se suele caracterizar la aleatoriedad de una variable real **discreta** por su **función másica de probabilidad** $P(X = x)$.

En la estadística paramétrica, hay múltiples modelos en los cuales $P(X = x) = P(x)$ se caracteriza de manera determinística dependiendo de un conjunto de parámetros θ

- $P_\theta(x)$ - Modelo binomial.
- $P_\theta(x)$ - Modelo poisson.
- $P_\theta(x)$ - Modelo bernoulli.
- $P_\theta(x)$ - Modelo binomial negativo.
- $P_\theta(x)$ - Modelo uniforme discreto.
- $P_\theta(x)$ - Modelo hipergeométrico.
- $P_\theta(x)$ - Modelo geométrico.
- ...

La estadística no paramétrica no impone una familia paramétrica sobre $P(x)$.

Por su parte, si F no tiene saltos, X es continua. Y podemos caracterizar F en el caso mediante $f(x) : R \longrightarrow [0, +\infty)$:

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx$$

Para todo x en R. Por lo cual se suele caracterizar la aleatoriedad de una variable real **continua** por su **función de distribución** $f_X(x)$. La existencia de este objeto se basa en elementos de teoría de la medida.

En la estadística paramétrica, existen múltiples modelos en los cuales $f_X(x) = f(x)$ se caracteriza de manera determinística dependiendo de un conjunto de parámetros θ

- $f_\theta(x)$ - Distribución gamma.
- $f_\theta(x)$ - Distribución F.
- $f_\theta(x)$ - Distribución t.
- $f_\theta(x)$ - Distribución Weibull.
- $f_\theta(x)$ - **Distribución normal**.
- ...

La estadística no paramétrica no impone una familia paramétrica sobre f.

Las dos funciones definidas cumplen las siguientes propiedades:

- $\sum_i P(X = x_i) = 1$ y $\int f(x)dx = 1$
- $P(X = x) \geq 0$ y $f(x) \geq 0$ pt x.

Note además que $P(x) \leq 1$ pt x pero no necesariamente $f(x) \leq 1$ pt x. Tome por ejemplo $X \sim U[0, 0.1]$

Conclusiones:

1. El experimento aleatorio da origen a la aleatoriedad del proceso.
2. No siempre me interesan los resultados del experimento, en general me interesan características numéricas de este. Esto se hace a través de una v.a.

Conclusiones:

3. La caracterización de la aleatoriedad de una v.a. se puede realizar de múltiples formas. En general usamos las *fmp* (discreto) y las *fdp* (continuo) y esta puede o no ser parametrizada.
4. Al caracterizar una v.a. X con una *fmp*/*fdp*, decimos que $X \sim f(x)$ / $X \sim P(x)$, y paramétrica o no, podemos calcular la probabilidad de cualquier evento I (continuo *fmp*/discreto *fdp*).

Proceso aleatorio generador de datos

Proceso aleatorio generador de datos

En inferencia estadística:

- Se asume una muestra aleatoria de tamaño n como una colección de dichas variables, igualmente distribuidas e independientes: $X_1, \ldots, X_n \sim f(x) \;/ \sim P(x)$.
- Nosotros, como estadísticos terrenales, observamos sus **realizaciones** $x_1, \ldots x_n$.

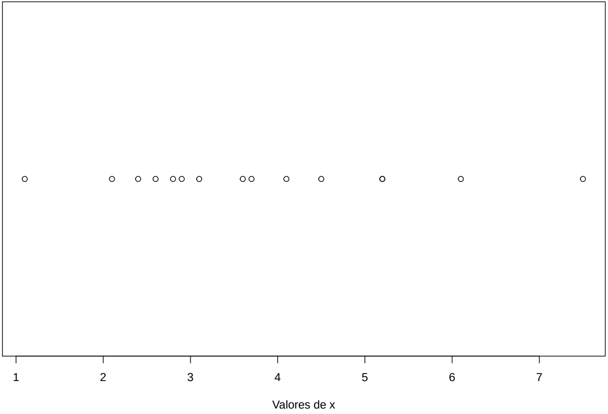
- La perspectiva *paramétrica* asume que la m.a sigue un modelo parametrizado. Es decir $X_1, \ldots, X_n \sim f_\theta(x)$ / $\sim P_\theta(x)$ con θ parámetro desconocido de la distribución: la *forma* de la distribución se asume, mas no sus parámetros.
- La perspectiva *no paramétrica* asume que la m.a sigue un modelo no parametrizado. Es decir $X_1, \ldots, X_n \sim f(x)$ / $\sim P(x)$: no se define una familia para la distribución.

Estimación kernel de la densidad (KDE)

Estimación kernel de la densidad (KDE)

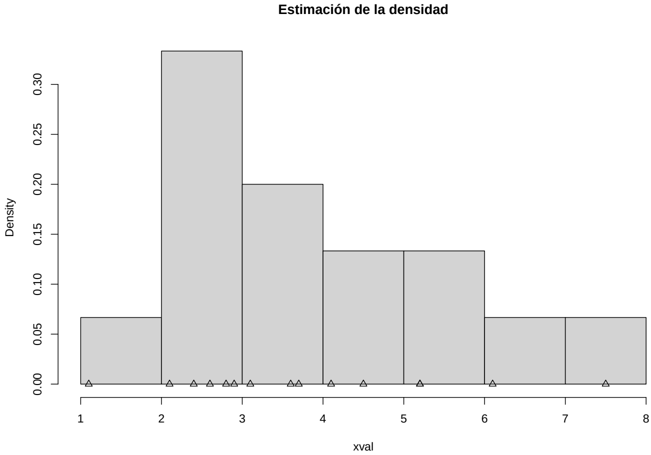
En resumen, con KDE buscamos obtener una estimación de la función de densidad de probabilidad f (ojo, no de $P(X = x)$). Tome estos $n = 15$ datos univariados como ejemplo:

```
xval= c(1.1,2.1,2.4,2.6,2.8,2.9,3.1,3.6,3.7,4.1,4.5,5.2,5.2,6.1,7.5)
plot(xval,y=rep(0,length(xval)),
     yaxt = "n",
     ylab = "",
     xlab = "Valores de x")
```



Sin embargo, note que esto puede ser realizado mediante el buen conocido histograma

```
h_dens = hist(xval, prob = TRUE, main= "Estimación de la densidad")
points(xval,y=rep(0,length(xval)),pch=2)
```



Del cual destacamos sus $T = 8$ breaks $b_1, \ldots, b_8$.

```
h_dens$breaks
```

```
## [1] 1 2 3 4 5 6 7 8
```

Y la estimación de la densidad estimada en cada uno de sus T intervalos de clase $\hat{f}_h(x_{ic=1}), \ldots, \hat{f}_h(x_{ic=8})$:

```
round(h_dens$density,4)
```

```
## [1] 0.0667 0.3333 0.2000 0.1333 0.1333 0.0667 0.0667
```

Que en efecto cumple las características de una función de densidad

```
# Vector con longitud de c/intervalo de clase (base)
l_base   = diff(h_dens$breaks)
# Vector con densidad est. para c/intervalo de clase (altura)
l_altura = h_dens$density
# Área total de todos los rectángulos del histograma
sum(l_base*l_altura)
```

```
## [1] 1
```

Por lo que afirmamos que hay una v.a real asociada X al proceso que rige los datos.

Esta es la estimación histograma de la densidad $\hat{f}_h(x)$. La cual es **no paramétrica**, ya que no asumimos un modelo para nuestros datos, dada por:

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{I(x_i \in [b_{j-1}, b_j])}{b_j - b_{j-1}}$$

Con $I(x \in [b_{j-1}, b_j])$ y n_j la cantidad de individuos en el intervalo $[b_{j-1}, b_j]$. En caso que los intervalos de clase tengan la misma longitud h:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} I(x_i \in [b_{j-1}, b_j])$$

Aunque es una estimación ampliamente conocida, tiene evidentes problemas:

1. El proceso subyacente f estimado mediante $\hat{f}_h$ probablemente no es escalonado.
2. Asume densidad constante para los puntos en cada intervalo, aún en un intervalo heterogéneo.

La estimación tipo histograma puede centrarse en las observaciones, haciendo que la estimación en cada punto siga el mismo principio de la estimación histograma:

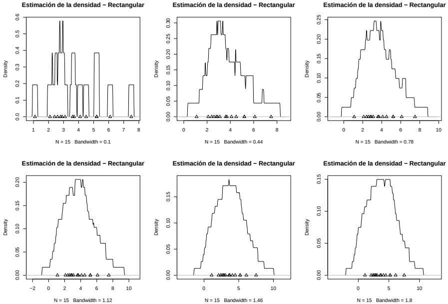
$$\hat{f}_{h_2}(x) = \frac{1}{nh}\sum_{i=1}^{n} I(x_i \in [x - h/2, x + h/2]) = \frac{1}{nh}\sum_{i=1}^{n} I(x - x_i \in [-h/2, +h/2])$$

Equivalente a

$$\hat{f}_{h_2}(x) = \frac{1}{nh}\sum_{i=1}^{n} I\left(\frac{x - x_i}{h} \in [-1/2, +1/2]\right)$$

Sigue siendo la estimación histograma no paramétrica, pero centrada y dependiente del ancho de banda h.

```
par(mfrow=c(2,3))
for(i in seq(0.1,1.8,length.out=6)){
h_dens2 = density(xval,kernel = "rectangular",bw = i)
plot(h_dens2, main= "Estimación de la densidad - Rectangular") ; points(xval,y=rep(0,length(xval)),pch=2)}
```




```
par(mfrow=c(1,1))
```

- Valores bajos de h dan una estimación muy ruidosa de f en x (alta varianza y bajo sesgo).
- Valores altos de h dan una estimación muy suavizada de f en x (baja varianza y alto sesgo).

Para la estimación histograma (centrada) de la densidad, note en la ecuación

$$\hat{f}_{h_2}(x) = \frac{1}{nh}\sum_{i=1}^{n} I\left(\frac{x-x_i}{h} \in [-1/2, +1/2]\right)$$

La podemos escribir como

$$\hat{f}_{h_2}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$$

Con $K(t)$ dada por

$$K(t) = \begin{cases} 1 \text{ si } t \in [-1/2, 1/2], \\ 0 \text{ e.o.c.} \end{cases}$$

A k se le llama el **kernel** de la estimación de la densidad. En este caso es denominado el kernel rectangular/uniforme/caja.

Fíjese que esta función tiene las siguientes características:

1. $K(t)$ es simétrico respecto a cero.
2. $K(t)$ es no negativo pt t.
3. $\int K(t)dt = 1$.
4. Tiende a cero cuando $t \longrightarrow +/-\infty$.

Podemos usar kernels diferentes en la estimación de la densidad.

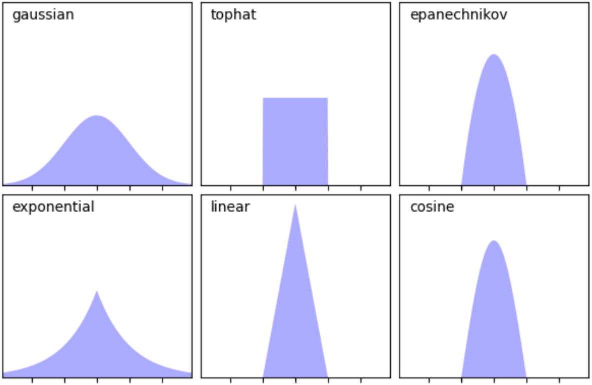


Figure 2: Tomado de: https://scikit-learn.org/stable/modules/density.html

Note que para encontrar $\hat{f}$ bajo KDE se requieren dos “hiperparámetros”: h y K. Con estos se estima la función subyacente $f(x)$ a partir de los datos. Generalmente se toma el kernel gaussiano, pues es más relevante h.

Recordando el ECM de un estimador $\hat{\theta}$ dado por $ECM(\hat{\theta})$ es igual a

$$ECM(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = V(\hat{\theta}) + B^2(\hat{\theta})$$

Para la función de densidad se tiene

$$ECM(\hat{f}(x|h)) = V(\hat{f}(x|h)) + B^2(\hat{f}(x|h))$$

Esto en $X = x$

Una medida resumen de la diferencia entre la función y su estimación para todo el dominio está dada por

$$ECM(\hat{f}|h) = E\left[\int (\hat{f}(x|h) - f(x))^2 dx\right] = \int V(\hat{f}(x|h))dx + \int B^2(\hat{f}(x|h))dx$$

El cual es minimizado* usando K como el kernel gaussiano en

$$h_o \approx 1.06\hat{\sigma} n^{-1/5}$$

Con $\hat{\sigma^2}$ una estimación de la varianza del kernel gaussiano. Una ligera variación de este es implementada en R por defecto (método nrd0 - Silverman's 'rule of thumb').

Métodos de suavizamiento kernel - Regresión local

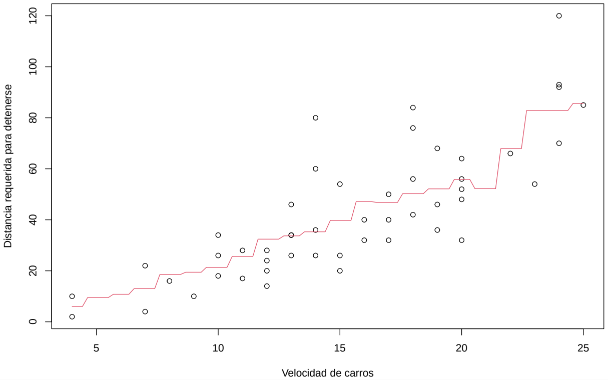
Métodos de suavizamiento kernel - Regresión local

El principio kernel puede ser utilizado para estimar de manera local el modelo de regresión $y = f(x)$ mediante polinomios. A esto se le conoce como **regresión en polinomios locales**. Este ajuste es similar al ajuste de polinomios a trozos, pero en lugar de dividir el dominio, se ajusta un modelo en cada x:

$$y(x) = \text{argmin}_{\alpha} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)(y_i - \alpha)^2$$

Este es llamado el estimador de Nadarya–Watson. Para encontrar $y(x)$ se ajusta un modelo de regresión ponderado con intercepto, esto para cada x, es decir, una función constante en cada x.

```
plot(cars$speed,cars$dist,xlab="Velocidad de carros",ylab="Distancia requerida para detenerse")
sm_reg = ksmooth(cars$speed, cars$dist, "box", bandwidth = 5)
lines(sm_reg, col = 2)
```

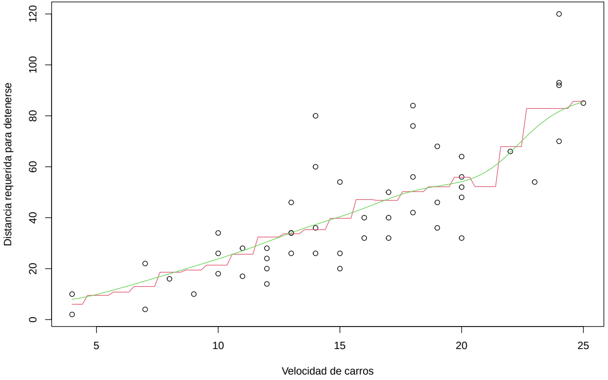



```
fit_val = NULL
for(t in 1:length(sm_reg$x)){
fit_val[t] = mean(cars$dist[abs((sm_reg$x[t] - cars$speed))/5 < 0.5])}
sum(fit_val != sm_reg$y)
```

```
## [1] 0
```

Pueden ser usados múltiples kernels para este estimador:

```
plot(cars$speed,cars$dist,xlab="Velocidad de carros",ylab="Distancia requerida para detenerse")
sm_reg_2 = ksmooth(cars$speed, cars$dist, "normal", bandwidth = 5)
lines(sm_reg  , col = 2)
lines(sm_reg_2, col = 3)
```



Podríamos ajustar polinomios de mayor grado para cada punto y tomar el intercepto como el valor esperado de Y dado el ajuste polinomial. Iniciando con un polinimio lineal, se tendría lo siguiente:

$$y(x) = \text{argmin}_{\alpha} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)\left(y_i - \alpha + \beta_1 (x - x_i)\right)^2$$

A esto también se le llama **regresión lineal local**.

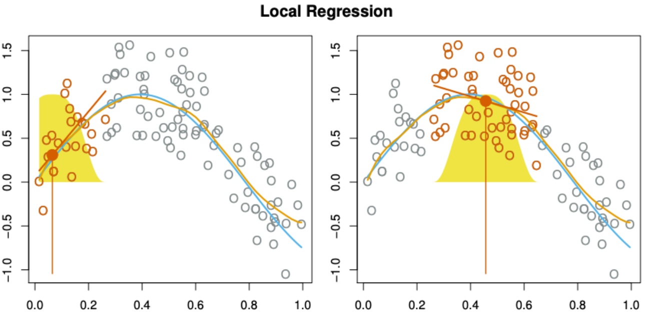


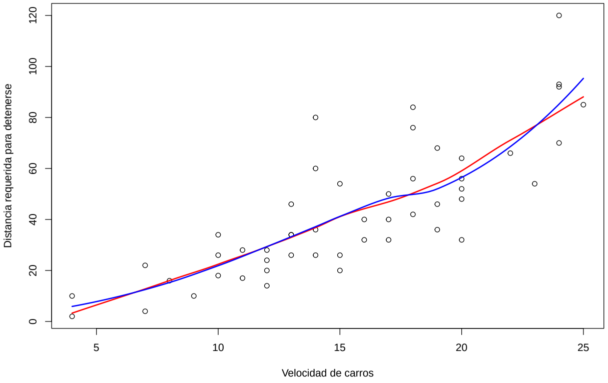
Figure 3: Tomado de: Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning (En naranja $\hat{f}$ y en azul f)

Podríamos ajustar polinomios de mayor grado para cada punto y tomar el intercepto como el valor esperado de Y dado el ajuste polinomial. Podemos llegar a polinomios de grado p:

$$y(x) = \text{argmin}_{\alpha} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)(y_i - \alpha + \sum_{j=1}^{p} \beta_j (x - x_i)^j)^2$$

A esto también se le llama **regresión polinomial local**.

```r
fit_1 = loess(dist ~ speed , degree = 1, data = cars)
fit_2 = loess(dist ~ speed , degree = 2, data = cars)
plot(cars$speed,cars$dist,xlab="Velocidad de carros",ylab="Distancia requerida para detenerse")
lines(sm_reg$x, predict(fit_1, data.frame(speed = sm_reg$x)),col = "red", lwd = 2)
lines(sm_reg$x, predict(fit_2, data.frame(speed = sm_reg$x)),col = "blue", lwd = 2)
```



Referencias

Referencias

1. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning. Springer.
2. Garet, Witten, Hastie, Tibshirani. Introduction to Statistical Learning with R.