Métodos No Lineales

Alvaro J. Riascos Villegas
Universidad de los Andes y Quantil

Agosto de 2017

Contenido

Introducción

- Las funciones de aprendizaje óptimas que hemos descubierto son:
  1. Para el problema de regresión: $f(X) = E[Y \mid X]$ (cuando la función de pérdida es el error cuadrático).
  2. Para el problema de regresión el clasificador de Bayes (cuando la función de pérdida toma solo dos valores).

Introducción

- Las funciones de aprendizaje óptimas que hemos descubierto son:
  1. Para el problema de regresión: $f(X) = E[Y \mid X]$ (cuando la función de pérdida es el error cuadrático).
  2. Para el problema de regresión el clasificador de Bayes (cuando la función de pérdida toma solo dos valores).

Introducción

- Las funciones de aprendizaje óptimas que hemos descubierto son:
  1. Para el problema de regresión: $f(X) = E[Y \mid X]$ (cuando la función de pérdida es el error cuadrático).
  2. Para el problema de regresión el clasificador de Bayes (cuando la función de pérdida toma solo dos valores).

Introducción

- En el caso de la función óptima de regresión la hemos aproximado por una función lineal de las variables explicativas: $f(X) = \beta X$.
- En el caso del clasificador de Bayes lo hemos aproximado por funciones tales que una transformación monótona es lineal (e.g., la función logística).
- En el espacio de características (*features*), estas aproximaciones dan lugar a hiperplanos de separación.
- Por esta razón se llaman métodos lineales.

Introducción

- En el caso de la función óptima de regresión la hemos aproximado por una función lineal de las variables explicativas: $f(X) = \beta X$.
- En el caso del clasificador de Bayes lo hemos aproximado por funciones tales que una transformación monótona es lineal (e.g., la función logística).
- En el espacio de características (*features*), estas aproximaciones dan lugar a hiperplanos de separación.
- Por esta razón se llaman métodos lineales.

Introducción

- En el caso de la función óptima de regresión la hemos aproximado por una función lineal de las variables explicativas: $f(X) = \beta X$.
- En el caso del clasificador de Bayes lo hemos aproximado por funciones tales que una transformación monótona es lineal (e.g., la función logística).
- En el espacio de características (*features*), estas aproximaciones dan lugar a hiperplanos de separación.
- Por esta razón se llaman métodos lineales.

Introducción

- En el caso de la función óptima de regresión la hemos aproximado por una función lineal de las variables explicativas: $f(X) = \beta X$.
- En el caso del clasificador de Bayes lo hemos aproximado por funciones tales que una transformación monótona es lineal (e.g., la función logística).
- En el espacio de características (*features*), estas aproximaciones dan lugar a hiperplanos de separación.
- Por esta razón se llaman métodos lineales.

Contenido

Aproximaciones no lineales al problema de regresión

- Bases de expansión.
- Supongamos que tenemos p variables explicativas y sea $h_m : R^p \rightarrow R$, $m = 1, ..., M$ funciones de transformación de X.
- Vamos a estudiar aproximaciones del tipo:

$$f(x) = \sum_{m=1}^{M} \beta_m h_m(x)$$

- La ventaja de esta aproximación es que una vez se definen las funciones de transformación es un problema lineal en las variables transformadas.
- Ejemplos: la identidad, polinomios, interacciones, funciones indicadoras de regiones.
- El conjunto de funciones de transformación se conoce como un diccionario.

Aproximaciones no lineales al problema de regresión

- Bases de expansión.
- Supongamos que tenemos p variables explicativas y sea $h_m : R^p \rightarrow R$, $m = 1, ..., M$ funciones de transformación de X.
- Vamos a estudiar aproximaciones del tipo:

$$f(x) = \sum_{m=1}^{M} \beta_m h_m(x)$$

- La ventaja de esta aproximación es que una vez se definen las funciones de transformación es un problema lineal en las variables transformadas.
- Ejemplos: la identidad, polinomios, interacciones, funciones indicadoras de regiones.
- El conjunto de funciones de transformación se conoce como un diccionario.

Aproximaciones no lineales al problema de regresión

- Bases de expansión.
- Supongamos que tenemos p variables explicativas y sea $h_m : R^p \to R$, $m = 1, ..., M$ funciones de transformación de X.
- Vamos a estudiar aproximaciones del tipo:

$$f(x) = \sum_{m=1}^{M} \beta_m h_m(x)$$

- La ventaja de esta aproximación es que una vez se definen las funciones de transformación es un problema lineal en las variables transformadas.
- Ejemplos: la identidad, polinomios, interacciones, funciones indicadoras de regiones.
- El conjunto de funciones de transformación se conoce como un diccionario.

Aproximaciones no lineales al problema de regresión

- Bases de expansión.
- Supongamos que tenemos p variables explicativas y sea $h_m : R^p \to R$, $m = 1, ..., M$ funciones de transformación de X.
- Vamos a estudiar aproximaciones del tipo:

$$f(x) = \sum_{m=1}^{M} \beta_m h_m(x)$$

- La ventaja de esta aproximación es que una vez se definen las funciones de transformación es un problema lineal en las variables transformadas.
- Ejemplos: la identidad, polinomios, interacciones, funciones indicadoras de regiones.
- El conjunto de funciones de transformación se conoce como un diccionario.

Aproximaciones no lineales al problema de regresión

- Bases de expansión.
- Supongamos que tenemos p variables explicativas y sea $h_m : R^p \to R$, $m = 1, ..., M$ funciones de transformación de X.
- Vamos a estudiar aproximaciones del tipo:

$$f(x) = \sum_{m=1}^{M} \beta_m h_m(x)$$

- La ventaja de esta aproximación es que una vez se definen las funciones de transformación es un problema lineal en las variables transformadas.
- Ejemplos: la identidad, polinomios, interacciones, funciones indicadoras de regiones.
- El conjunto de funciones de transformación se conoce como un diccionario.

Aproximaciones no lineales al problema de regresión

- Bases de expansión.
- Supongamos que tenemos p variables explicativas y sea $h_m : R^p \rightarrow R$, $m = 1, ..., M$ funciones de transformación de X.
- Vamos a estudiar aproximaciones del tipo:

$$f(x) = \sum_{m=1}^{M} \beta_m h_m(x)$$

- La ventaja de esta aproximación es que una vez se definen las funciones de transformación es un problema lineal en las variables transformadas.
- Ejemplos: la identidad, polinomios, interacciones, funciones indicadoras de regiones.
- El conjunto de funciones de transformación se conoce como un diccionario.

Splines

- Un Spline de orden M con nodos $\eta_1, ..., \eta_K$ es un polinomio a trozos de orden $M-1$ con derivadas continuas de orden $M-2$.
- Las funciones de transformación son en general:

$$h_j(x) = x^{j-1}, j = 1, ..., M \quad (1)$$

$$h_{M+l} = (x - \eta_l)_+^{M-1}, l = 1, ..., K \quad (2)$$

- El caso de splines cúbicos:

$$f(x) = \sum_{j=0}^{3} \beta_j x^j + \sum_{k=1}^{K} \theta_k (x - \eta_k)_+^3 \quad (3)$$

Splines

- Un Spline de orden M con nodos $\eta_1, ..., \eta_K$ es un polinomio a trozos de orden $M-1$ con derivadas continuas de orden $M-2$.
- Las funciones de transformación son en general:

$$h_j(x) = x^{j-1}, j = 1, ..., M \quad (1)$$

$$h_{M+l} = (x - \eta_l)_+^{M-1}, l = 1, ..., K \quad (2)$$

- El caso de splines cúbicos:

$$f(x) = \sum_{j=0}^{3} \beta_j x^j + \sum_{k=1}^{K} \theta_k (x - \eta_k)_+^3 \quad (3)$$

Splines

- Un Spline de orden M con nodos $\eta_1, ..., \eta_K$ es un polinomio a trozos de orden $M-1$ con derivadas continuas de orden $M-2$.
- Las funciones de transformación son en general:

$$h_j(x) = x^{j-1}, j = 1, ..., M \quad (1)$$

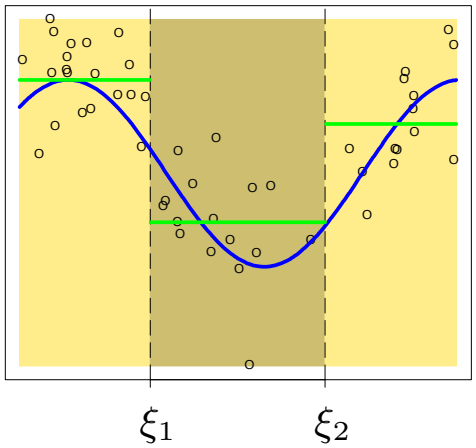
$$h_{M+l} = (x - \eta_l)_+^{M-1}, l = 1, ..., K \quad (2)$$

- El caso de splines cúbicos:

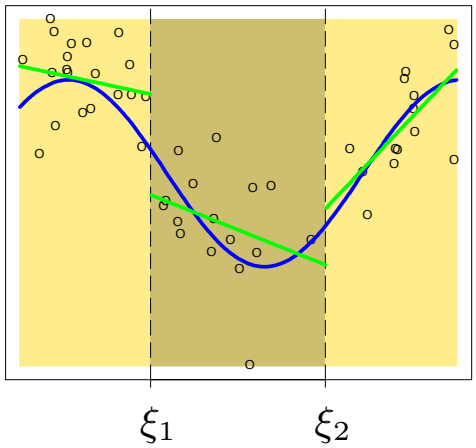
$$f(x) = \sum_{j=0}^{3} \beta_j x^j + \sum_{k=1}^{K} \theta_k (x - \eta_k)_+^3 \quad (3)$$

Splines: Polinomios a trozos

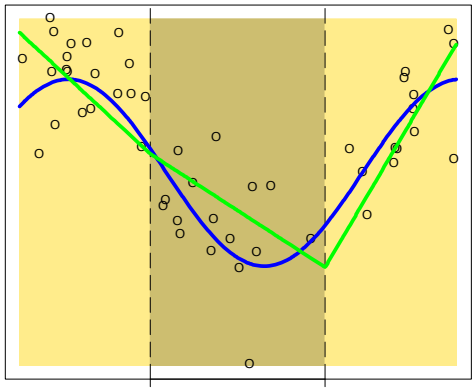
Piecewise Constant



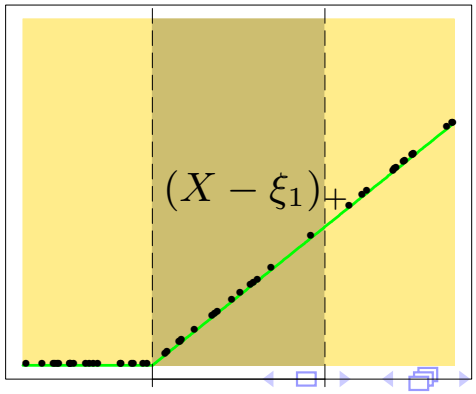
Piecewise Linear



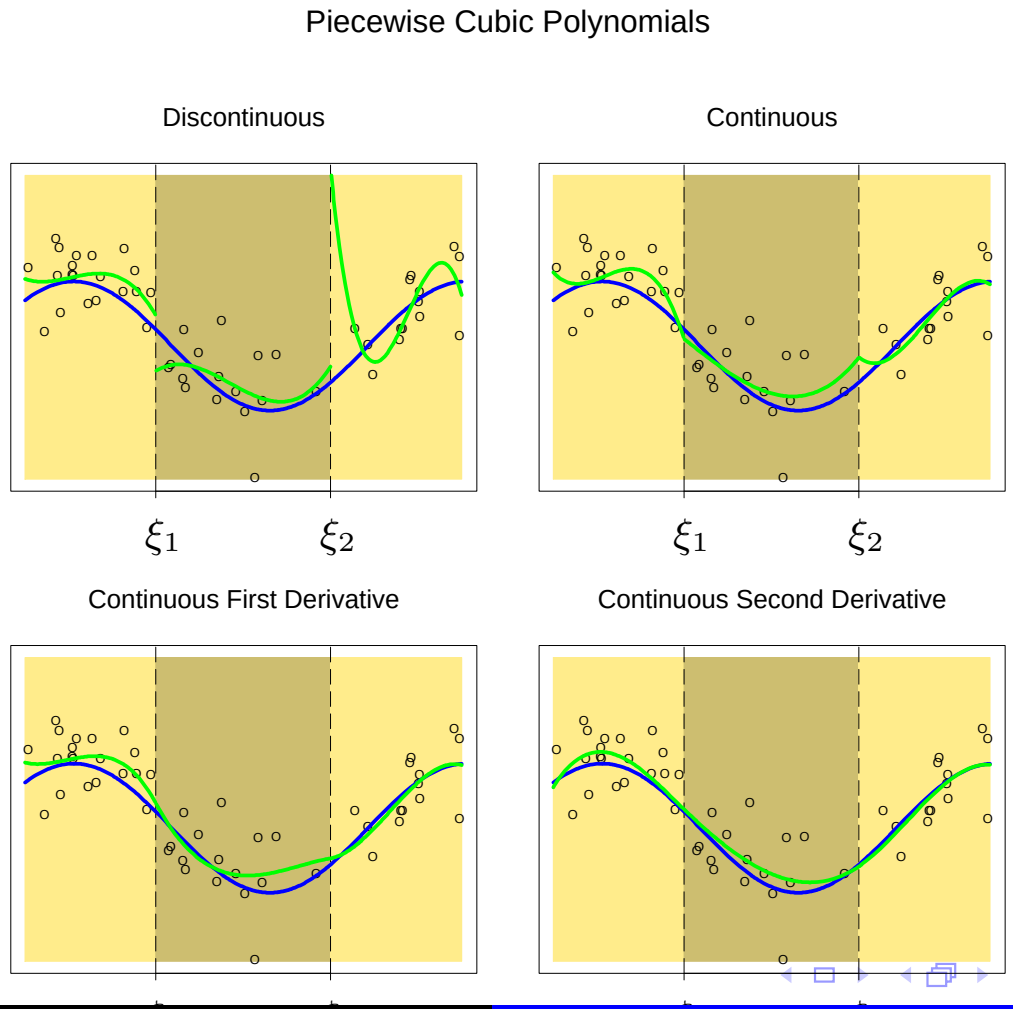
Continuous Piecewise Linear



Piecewise-linear Basis Function



Splines: suavizando



Splines de suavización

- El mayor inconveniente con esta técnica es la necesidad de elegir los nodos exante.
- Los splines de suavización (smoothing splines) buscan resolver este problema.
- Para esto resolvemos el problema de minimización sobre un espacio de funciones dos veces diferenciables:

$$\sum_{i=1}^{N}(y_i - f(x_i))^2 + \lambda \int (f``(t))^2 dt \tag{4}$$

donde λ es el parámetro de suavización.

- La solución a este problema es una función no lineal expresable en términos de una base de expansión.

Splines de suavización

- El mayor inconveniente con esta técnica es la necesidad de elegir los nodos exante.
- Los splines de suavización (smoothing splines) buscan resolver este problema.
- Para esto resolvemos el problema de minimización sobre un espacio de funciones dos veces diferenciables:

$$\sum_{i=1}^{N}(y_i - f(x_i))^2 + \lambda \int (f``(t))^2 dt \tag{4}$$

donde λ es el parámetro de suavización.

- La solución a este problema es una función no lineal expresable en términos de una base de expansión.

Splines de suavización

- El mayor inconveniente con esta técnica es la necesidad de elegir los nodos exante.
- Los splines de suavización (smoothing splines) buscan resolver este problema.
- Para esto resolvemos el problema de minimización sobre un espacio de funciones dos veces diferenciables:

$$\sum_{i=1}^{N}(y_i - f(x_i))^2 + \lambda \int (f``(t))^2 dt \tag{4}$$

donde λ es el parámetro de suavización.

- La solución a este problema es una función no lineal expresable en términos de una base de expansión.

Splines de suavización

- El mayor inconveniente con esta técnica es la necesidad de elegir los nodos exante.
- Los splines de suavización (smoothing splines) buscan resolver este problema.
- Para esto resolvemos el problema de minimización sobre un espacio de funciones dos veces diferenciables:

$$\sum_{i=1}^{N}(y_i - f(x_i))^2 + \lambda \int (f``(t))^2 dt \tag{4}$$

donde λ es el parámetro de suavización.

- La solución a este problema es una función no lineal expresable en términos de una base de expansión.

Splines de suavización

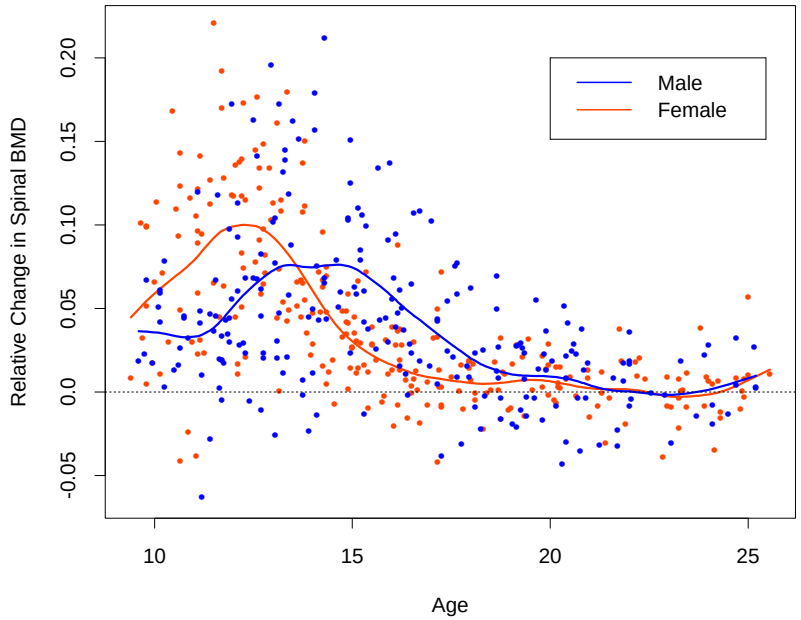


FIGURE 5.6. *The response is the relative change in bone mineral density measured at the spine in adolescents, as a function of age. A separate smoothing spline was fit to the males and females, with* $\lambda \approx 0.00022$. *This choice corresponds to about* 12 *degrees of freedom.*

Contenido

Métodos de Kernels

- En esta sección estudiamos el método de Kernels como funciones de localizacioón en contraste con la literatura en la que representan productos internos en espacios de hilbert.
- La idea de los métodos de kernels es estimar funciones distintas en cada punto observado y de tal forma que la función resultante sea suave.
- Esto se logra usando una función de pesos $K_\lambda(x_0, x_i)$ donde el peso en x_i depende de la distancia de x_i a x_0 y x_0 es el punto donde queremos evaluar la función de aprendizaje.
- El parámetro clave es λ, es el único que hay que entrenar.

Métodos de Kernels

- En esta sección estudiamos el método de Kernels como funciones de localizacioón en contraste con la literatura en la que representan productos internos en espacios de hilbert.
- La idea de los métodos de kernels es estimar funciones distintas en cada punto observado y de tal forma que la función resultante sea suave.
- Esto se logra usando una función de pesos $K_\lambda(x_0, x_i)$ donde el peso en x_i depende de la distancia de x_i a x_0 y x_0 es el punto donde queremos evaluar la función de aprendizaje.
- El parámetro clave es λ, es el único que hay que entrenar.

Métodos de Kernels

- En esta sección estudiamos el método de Kernels como funciones de localizacioón en contraste con la literatura en la que representan productos internos en espacios de hilbert.
- La idea de los métodos de kernels es estimar funciones distintas en cada punto observado y de tal forma que la función resultante sea suave.
- Esto se logra usando una función de pesos $K_\lambda(x_0, x_i)$ donde el peso en x_i depende de la distancia de x_i a x_0 y x_0 es el punto donde queremos evaluar la función de aprendizaje.
- El parámetro clave es λ, es el único que hay que entrenar.

Métodos de Kernels

- El vecino más cercano puede introducir discontinuidades innecesarias.

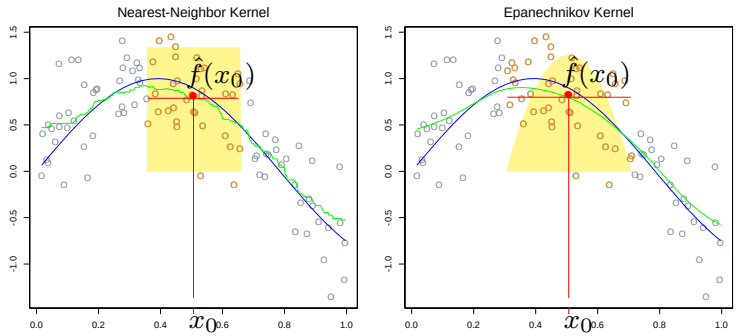


FIGURE 6.1. *In each panel* 100 *pairs* x_i, y_i *are generated at random from the blue curve with Gaussian errors:* $Y = \sin(4X) + \varepsilon$, $X \sim U[0,1]$, $\varepsilon \sim N(0, 1/3)$. *In the left panel the green curve is the result of a* 30*-nearest-neighbor running-mean smoother. The red point is the fitted constant* $\hat{f}(x_0)$, *and the red circles indicate those observations contributing to the fit at* x_0. *The solid yellow region indicates the weights assigned to observations. In the right panel, the green curve is the kernel-weighted average, using an Epanechnikov kernel with (half) window width* $\lambda = 0.2$.

Métodos de Kernels

- Una forma de suavizar es usando el método de Nadaraya-Watson con función el Kernel cuadrático de Epanechnikov:

$$f(x_0) = \frac{\sum_{i=1}^{N} K_\lambda(x_0, x_i) y_i}{\sum_{i=1}^{N} K_\lambda(x_0, x_i)} \tag{5}$$

$$K_\lambda(x_0, x_i) = D\left(\frac{x_i - x_0}{\lambda}\right) \tag{6}$$

$$D(t) = \frac{3}{4}(1 - t^2) \text{ si } t \leq 1 \tag{7}$$

Métodos de Kernels

- λ controla la varianza de la función de aprendizaje. Entre mayor sea λ menor es la varianza. Kernels comúnes:

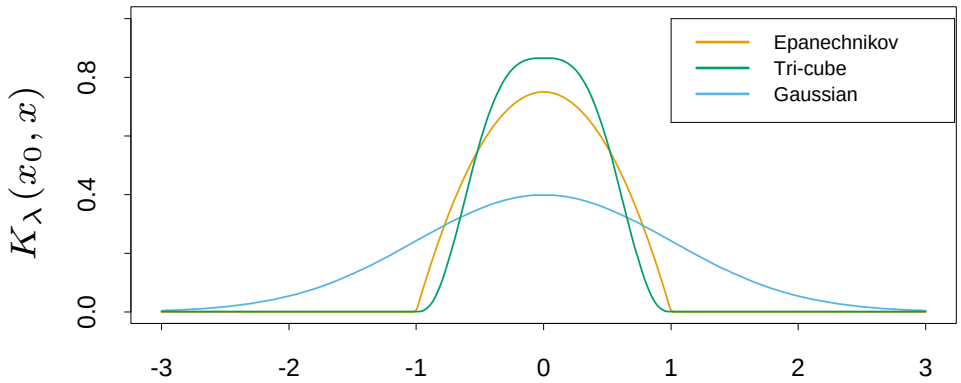


FIGURE 6.2. *A comparison of three popular kernels for local smoothing. Each has been calibrated to integrate to* 1. *The tri-cube kernel is compact and has two continuous derivatives at the boundary of its support, while the Epanechnikov kernel has none. The Gaussian kernel is continuously differentiable, but has infinite support.*

Contenido

Estimación de densidades

- Supongamos que $x_1, ..., x_N$ es una muestra de datos tomada con una distribución con densidad $f_X(x)$.
- Un primer estimador local es:

$$f_X(x_0) = \frac{num\{N(x_0)\}}{N\lambda} \tag{8}$$

donde $N(x_0)$ es una vecindad de tamaño λ.

- Una versión suavizada es:

$$f_X(x_0) = \frac{1}{N\lambda} \sum_{i=1}^{N} K_\lambda(x_0, x_i) \tag{9}$$

donde por ejemplo $K_\lambda(x_0, x_i) = \psi_\lambda(x - x_0)$ y ψ_λ es la densidad Gaussiana $(0, \lambda^2)$.

Estimación de densidades

- Supongamos que $x_1, ..., x_N$ es una muestra de datos tomada con una distribución con densidad $f_X(x)$.
- Un primer estimador local es:

$$f_X(x_0) = \frac{num\{N(x_0)\}}{N\lambda} \tag{8}$$

donde $N(x_0)$ es una vecindad de tamaño λ.

- Una versión suavizada es:

$$f_X(x_0) = \frac{1}{N\lambda} \sum_{i=1}^{N} K_\lambda(x_0, x_i) \tag{9}$$

donde por ejemplo $K_\lambda(x_0, x_i) = \psi_\lambda(x - x_0)$ y ψ_λ es la densidad Gaussiana $(0, \lambda^2)$.

Estimación de densidades

- Supongamos que $x_1, ..., x_N$ es una muestra de datos tomada con una distribución con densidad $f_X(x)$.
- Un primer estimador local es:

$$f_X(x_0) = \frac{num\{N(x_0)\}}{N\lambda} \tag{8}$$

donde $N(x_0)$ es una vecindad de tamaño λ.

- Una versión suavizada es:

$$f_X(x_0) = \frac{1}{N\lambda} \sum_{i=1}^{N} K_\lambda(x_0, x_i) \tag{9}$$

donde por ejemplo $K_\lambda(x_0, x_i) = \psi_\lambda(x - x_0)$ y ψ_λ es la densidad Gaussiana $(0, \lambda^2)$.

Estimación de densidades

- En forma reducida:

$$f_X(x) = \frac{1}{N\lambda} \sum_{i=1}^{N} \psi_\lambda(x - x_i) \tag{10}$$

- En p dimensiones:

$$f_X(x) = \frac{1}{N\lambda(2\lambda^2\pi)^{\frac{p}{2}}} \sum_{i=1}^{N} \exp(-\frac{1}{2}(\frac{x - x_i}{\lambda})^2) \tag{11}$$

Estimación de densidades

- En forma reducida:

$$f_X(x) = \frac{1}{N\lambda} \sum_{i=1}^{N} \psi_\lambda(x - x_i) \tag{10}$$

- En p dimensiones:

$$f_X(x) = \frac{1}{N\lambda(2\lambda^2\pi)^{\frac{p}{2}}} \sum_{i=1}^{N} \exp(-\frac{1}{2}(\frac{x - x_i}{\lambda})^2) \tag{11}$$

Estimación de densidades

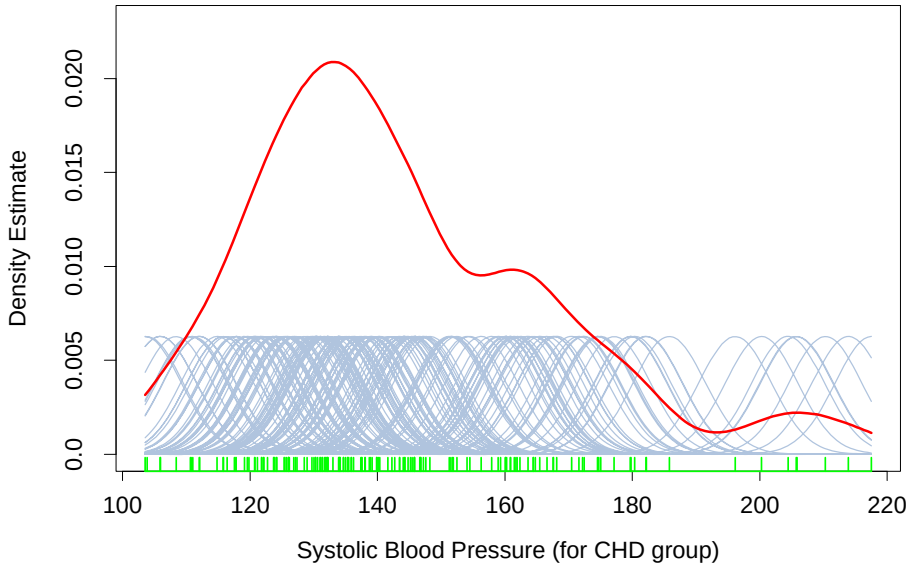


FIGURE 6.13. *A kernel density estimate for systolic blood pressure (for the CHD group). The density estimate at each point is the average contribution from each of the kernels at that point. We have scaled the kernels down by a factor of 10 to make the graph readable.*

Contenido

Clasificación

- Usando las estimaciones de Kernel de las densidades por clase f_j, el clasificador óptimo se puede escribir (usando la regla de Bayes):

$$P(G = j \mid X = x_0) = \frac{\pi_j f_j(x_0)}{\sum_{k=1}^{J} \pi_k f_k(x_0)} \tag{12}$$

donde π_j son las frecuencias relativas de cada clase.

Estimación de densidades

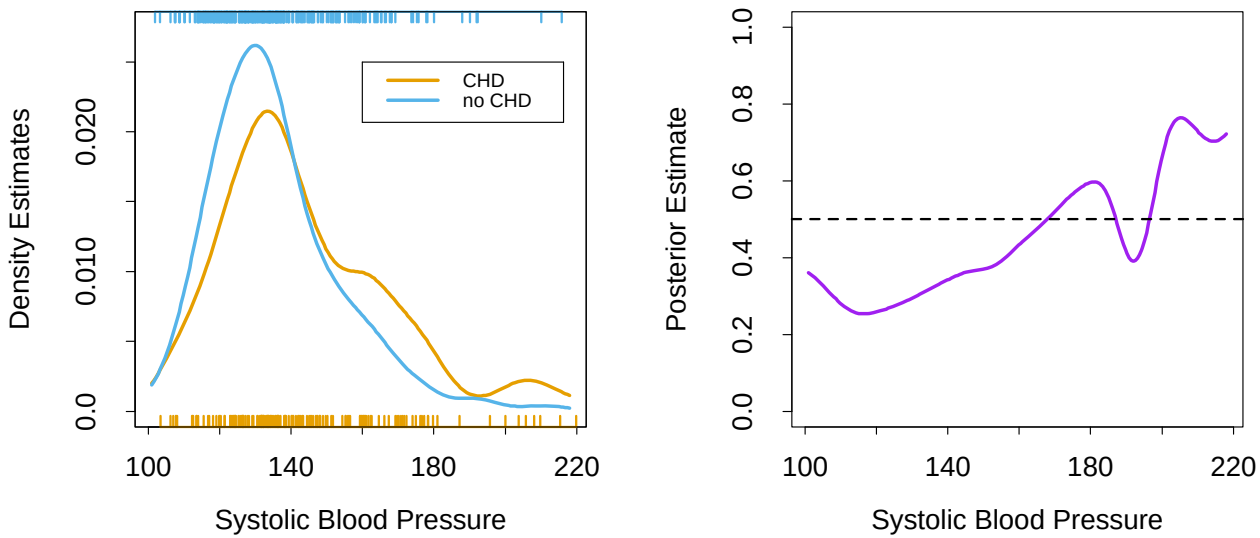


FIGURE 6.14. *The left panel shows the two separate density estimates for systolic blood pressure in the CHD versus no-CHD groups, using a Gaussian kernel density estimate in each. The right panel shows the estimated posterior probabilities for CHD, using (6.25).*

Clasificación: Bayes Naive

- Cuando el espacio de características es muy grande (p grande) la estimación por kernels tiene mucha varianza.
- El clasificador de Bayer asume independencia de las variables y en ese caso el clasificador de Bayes se reduce a:

$$P(G = j \mid X = x_0) = \frac{\pi_j f_j(x_0)}{\sum_{k=1}^{J} \pi_k f_k(x_0)} \tag{13}$$

donde:

$$f_j(x) = \Pi_{l=1}^{p} f_{j,l}(x_l) \tag{14}$$

- Cada $f_{j,l}$ es un kernel unidimensional. La complejidad del problema se reduce enormemente.
- Si alguna variable es dicreta, permite combinar fácilmente variables continuas y categóricas.

Clasificación: Bayes Naive

- Cuando el espacio de características es muy grande (p grande) la estimación por kernels tiene mucha varianza.
- El clasificador de Bayer asume independencia de las variables y en ese caso el clasificador de Bayes se reduce a:

$$P(G = j \mid X = x_0) = \frac{\pi_j f_j(x_0)}{\sum_{k=1}^{J} \pi_k f_k(x_0)} \tag{13}$$

donde:

$$f_j(x) = \Pi_{l=1}^{p} f_{j,l}(x_l) \tag{14}$$

- Cada $f_{j,l}$ es un kernel unidimensional. La complejidad del problema se reduce enormemente.
- Si alguna variable es dicreta, permite combinar fácilmente variables continuas y categóricas.

Clasificación: Bayes Naive

- Cuando el espacio de características es muy grande (p grande) la estimación por kernels tiene mucha varianza.
- El clasificador de Bayer asume independencia de las variables y en ese caso el clasificador de Bayes se reduce a:

$$P(G = j \mid X = x_0) = \frac{\pi_j f_j(x_0)}{\sum_{k=1}^{J} \pi_k f_k(x_0)} \tag{13}$$

donde:

$$f_j(x) = \Pi_{l=1}^{p} f_{j,l}(x_l) \tag{14}$$

- Cada $f_{j,l}$ es un kernel unidimensional. La complejidad del problema se reduce enormemente.
- Si alguna variable es dicreta, permite combinar fácilmente variables continuas y categóricas.

Clasificación: Bayes Naive

- Cuando el espacio de características es muy grande (p grande) la estimación por kernels tiene mucha varianza.
- El clasificador de Bayer asume independencia de las variables y en ese caso el clasificador de Bayes se reduce a:

$$P(G = j \mid X = x_0) = \frac{\pi_j f_j(x_0)}{\sum_{k=1}^{J} \pi_k f_k(x_0)} \tag{13}$$

donde:

$$f_j(x) = \Pi_{l=1}^{p} f_{j,l}(x_l) \tag{14}$$

- Cada $f_{j,l}$ es un kernel unidimensional. La complejidad del problema se reduce enormemente.
- Si alguna variable es dicreta, permite combinar fácilmente variables continuas y categóricas.

Contenido

Aplicaciones

- Predicción del crimen en Bogotá.
- El problema de causalidad y predicción.