Clasificación Bayesiana: Técnicas no paramétricas

FaMAF

2019

Clasificación Bayesiana: Técnicas no paramétricas

- La mayoría de las densidades paramétricas son unimodales (tienen un único máximo local), mientras que muchos problemas prácticos envuelven densidades multi-modales.
- La estimación paramétrica de la mezcla de gaussianas precisa muchos recursos, tanto en datos como en tiempo de cómputo.
- Los procedimientos no paramétricos pueden ser usados con distribuciones arbitrarias y sin la hipótesis de que las densidades son conocidas.

Clasificación Bayesiana: Técnicas no paramétricas

Enfoques de estimación de densidad no paramétrica en el problema de clasificación:

- Estimar funciones de verosimilitud$p\left(\mathbf{x}|\omega_j\right)$.
- Estimar directamente probabilidades posteriores $p\left(\omega_j|\mathbf{x}\right)$.

Cuando la densidad es desconocida, la estimación paramétrica intenta ajustar una forma global a la muestra de entrenamiento, como vimos con la mezcla de Gaussianas.

En cambio la estimación no paramétrica estima localmente valores para la densidad.

Estimación densidad no paramétrica

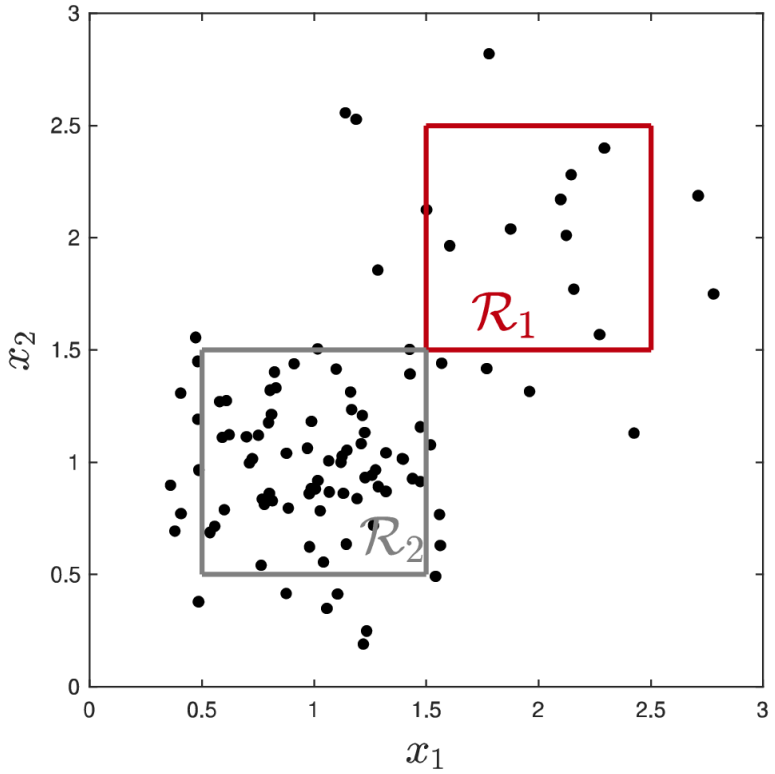
Idea Básica de estimación de densidad $p(\mathbf{x})$:

- Estimamos $p(\mathbf{x})$ con la probabilidad P de que un patrón $\mathbf{x}_i$ caiga en una región $\mathcal{R}$ con volumen 1 que contiene a $\mathbf{x}$.

$$P = \int_{\Re} p\left(\mathbf{x}'\right) d\mathbf{x}' \tag{1}$$

- Intuitivamente, un estimador de P es la fracción de muestras que caen en $\mathcal{R}$.
- P es una version suavizada o promediada de la función de densidad $p(\mathbf{x})$

Estimación densidad no paramétrica



Se tienen $n = 100$ muestras en un espacio bidimensional. En la región $\mathcal{R}_1$ caen $k = 9$ muestras:$\hat{p}(\hat{\mathbf{x}}) = 0.09$. En la región $\mathcal{R}_2$ caen $k = 60$ muestras: $\hat{p}(\hat{\mathbf{x}}) = 0.6$. Ambas regiones tienen área 1.

Estimación densidad no paramétrica

- Si tenemos una muestra de tamaño n, $\mathbf{x}_1, \ldots, \mathbf{x}_n$ tomadas i.i.d. de una distribución $p(\mathbf{x})$.
- Ley binomial: probabilidad de que k de estas n muestras caigan en una región arbitraria $\mathcal{R}$ es:

$$P(N = k) = \left(\begin{array}{c} n \\ k \end{array} \right) P^k (1 - P)^{n-k} \tag{2}$$

- Estimador de máxima verosimilitud para P:

$$\hat{P} = \frac{k}{n} \tag{3}$$

  es un buen estimador de la probabilidad P.

Estimación densidad no paramétrica

- Si p es continua y si la región $\mathcal{R}$ es tan pequeña que p no varía mucho en ella podemos escribir

$$\int_{\mathcal{R}} p\left(\mathbf{x}'\right) d\mathbf{x}' \cong p(\hat{\mathbf{x}}) \int_{\mathcal{R}} d\mathbf{x}' = p(\hat{\mathbf{x}})V \tag{4}$$

  donde $\hat{\mathbf{x}}$ es el punto medio de $\mathcal{R}$ y V es el volumen de $\mathcal{R}$.

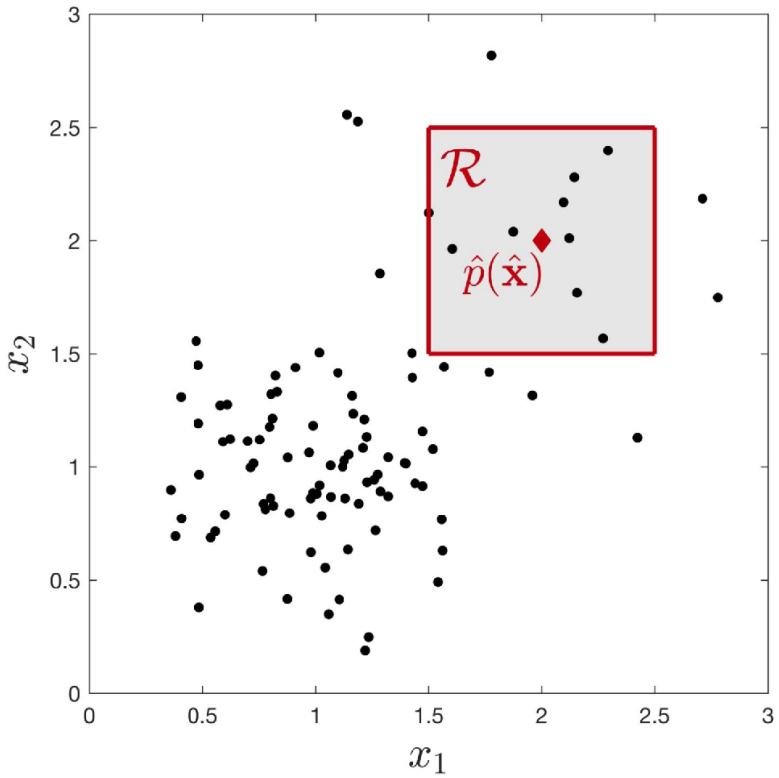
- Por lo cual

$$p(\hat{\mathbf{x}}) \cong \frac{k/n}{V} = \hat{p}(\hat{\mathbf{x}}) \tag{5}$$

  donde $\hat{\mathbf{x}}$ es el punto medio de $\mathcal{R}$ y V el volumen de $\mathcal{R}$.

- El estimador $\hat{p}(\hat{\mathbf{x}})$ depende del volumen V de la region considerada. Si $V = 1$ el estimador es la fracción de muestras que caen en V.

Estimación densidad no paramétrica



Densidad $p(.)$ se asume constante en todo $\mathcal{R}$, $p(\mathbf{x}) \equiv \hat{p}(\hat{\mathbf{x}}) \approx \frac{1}{V}\frac{k}{n}$ donde $\hat{\mathbf{x}}$ es el punto medio y V el volumen de $\mathcal{R}$.

Estimación densidad no paramétrica

- La fracción $k/(nV)$ es un valor promediado de $p(\mathbf{x})$.
- $p(\mathbf{x})$ se obtiene exacta solo si V se acerca a cero.
- Si n es un número fijo, puede no haber muestras en $\mathcal{R}$, por lo cual

$$\lim_{V \to 0, \mathrm{k}=0} \hat{p}(\mathbf{x}) = 0 \tag{6}$$

- Si alguna o mas muestras coinciden con $\mathbf{x}$, el estimador diverge

$$\lim_{V \to 0, k \neq 0} \hat{p}(\mathbf{x}) = \infty \tag{7}$$

Clasificación Bayesiana: Técnicas no paramétricas

- El volumen V necesita ir a cero o no se podria usar esta estimación.
- Sin embargo, en la práctica V no puede ser muy chico pues el número de muestras es siempre limitado.
- Se tiene que aceptar variabilidad en el radio k/n y un promedio en la densidad $p(\mathbf{x})$.

Clasificación Bayesiana: Técnicas no paramétricas

- Teóricamente, con infinitas muestras, para estimar la densidad en $\mathbf{x}$, se forma una sucesión de regiones $\mathcal{R}_1, \mathcal{R}_2, \ldots$ que contienen $\mathbf{x}$: la primera con una muestra, la segunda con dos, etc...
- Si V_n es el volumen de $\mathcal{R}_n$, k_n el número de muestras en $\mathcal{R}_n$ y $\hat{p_n}(\mathbf{x})$ el n-esimo estimador de $p(\mathbf{x})$.

$$\hat{p_n}(\mathbf{x}) = \frac{1}{V_n}\frac{k_n}{n} \tag{8}$$

- Hay tres condiciones necesarias si para asegurar la convergencia $\hat{p_n}(\mathbf{x}) \to p(\mathbf{x})$ cuando $n \to \infty$
  - $\lim_{n\to\infty} V = 0$
  - $\lim_{n\to\infty} k = \infty$
  - $\lim_{n\to\infty} k/n = 0$, n y k suficientemente grandes y V suficientemente pequeño.

Clasificación Bayesiana: Técnicas no paramétricas

- La primera condición
  - $\lim_{n\to\infty} V = 0$

  asegura que el radio P/V converge a $p(\mathbf{x})$.
- La segunda condición
  - $\lim_{n\to\infty} k = \infty$

  asegura que la frecuencia de observación converge en probabilidad a P
- La tercera condición
  - $\lim_{n\to\infty} k/n = 0,$

  asegura que si bien hay infinitas muestras en la ventana, solo son una parte pequeña del total de muestras.

Métodos de estimación

Hay dos métodos principales para obtener estas condiciones :

a) Determinar ventana mediante una función como $V_n = 1/\sqrt{n}$ de tal forma que el k_n resultante se comporte bien y se cumpla

$$\hat{p}_n(\mathbf{x}) \underset{n\to\infty}{\to} p(\mathbf{x}) \tag{9}$$

Este método es llamado “método de estimación de la ventana de Parzen”

b) Especificar k_n como una función de n, como $k_n = \sqrt{n}$; el volumen V_n crece hasta que engloba k_n vecinos de $\mathbf{x}$. Este método se llama “estimacion por los vecinos mas cercanos”.

Métodos de estimación

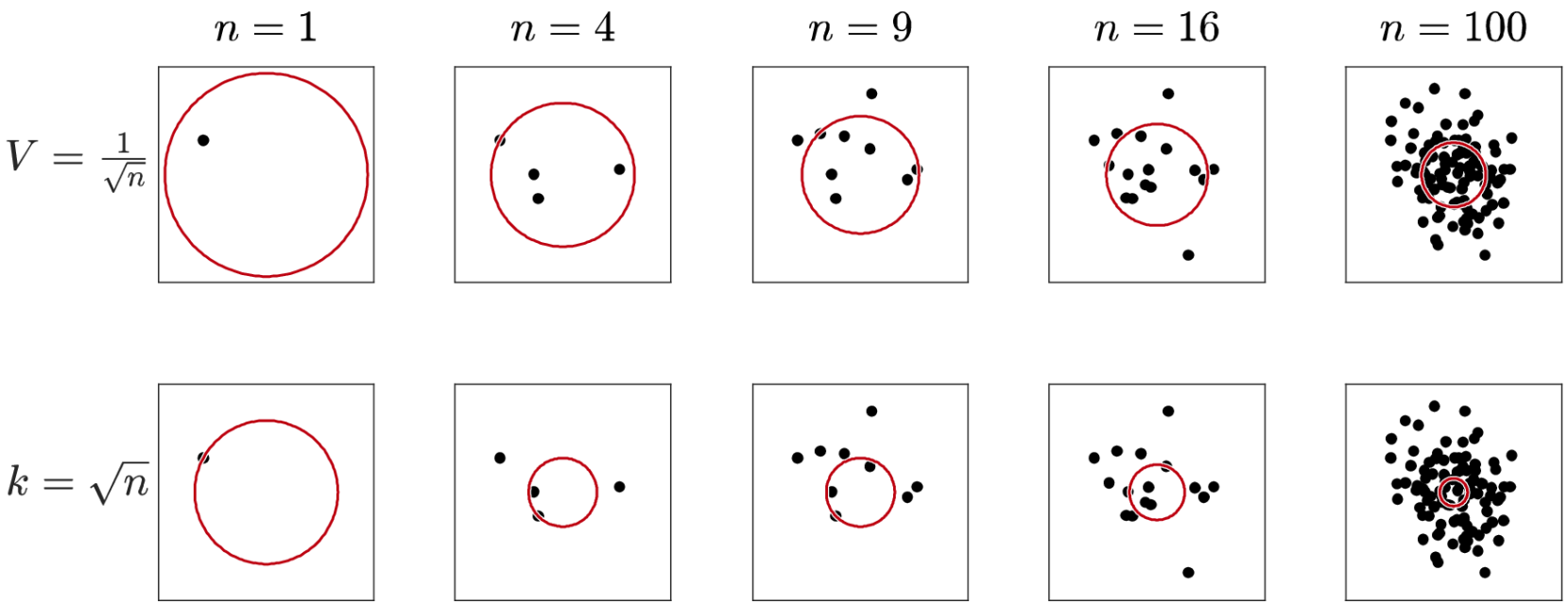


Figure: (a) ventanas de Parzen, (fila superior) se especifica el volumen V en función de n, por ejemplo, $V = 1/\sqrt{n}$, (b) k-vecinos más cercanos, (fila inferior) se especifica k en función de n, por ejemplo, $k = \sqrt{n}$.

Métodos de estimación: Ventana de Parzen

El método de la ventana de Parzen clásico para estimar densidades asume que:

- La region $\mathcal{R} \in \mathbb{R}^d$ es un hipercubo con volumen $V = h^d$, donde h es la longitud de uno de sus bordes.
- Número de muestras que caen en el hipercubo centrado en $\hat{\mathbf{x}}$

$$k = \sum_{i=1}^{n} \phi_\delta \left(\frac{\|\hat{\mathbf{x}} - \mathbf{x}_i\|_\infty}{h} \right) \tag{10}$$

donde

$$\|\mathbf{p} - \mathbf{q}\|_\infty = \max_{j=1,\ldots,d} |p_j - q_j| \tag{11}$$

y $\phi_\delta(.)$ es la función de ventana

$$\phi_\delta(u) = \begin{cases} 1 & \text{si } u \leq 1/2 \\ 0 & \text{otro caso} \end{cases} \tag{12}$$

Ventana de Parzen clásica

Sustituyendo

$$k = \sum_{i=1}^{n} \phi_\delta \left(\frac{\|\hat{\mathbf{x}} - \mathbf{x}_i\|_\infty}{h} \right) \tag{13}$$

en

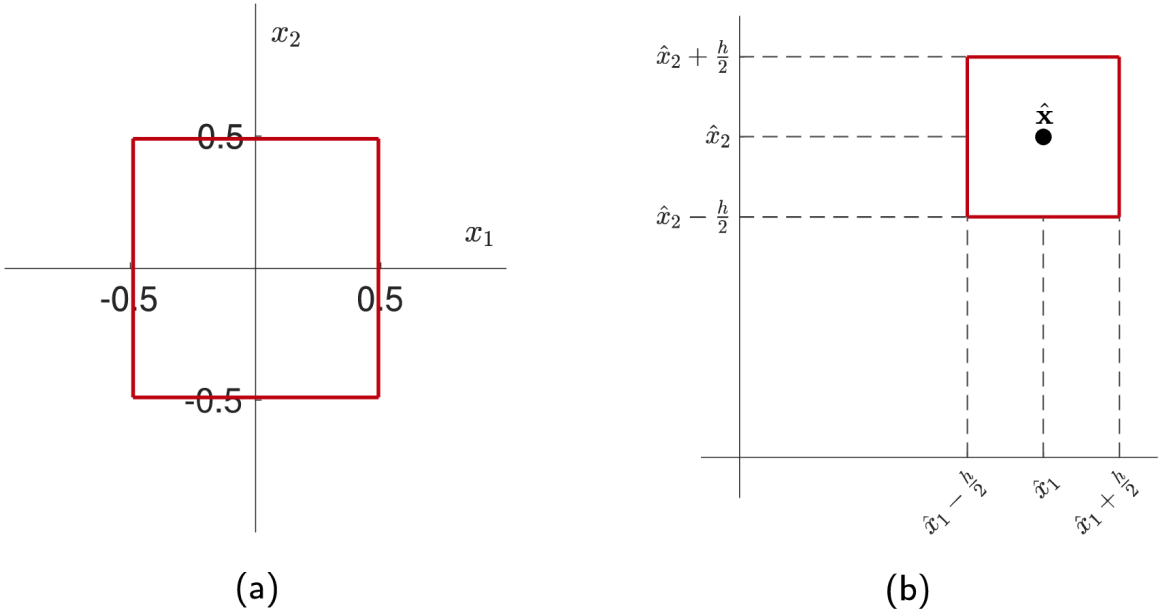
$$p(\mathbf{x}) \equiv \hat{p}(\hat{\mathbf{x}}) \approx \frac{1}{V} \frac{k}{n} \tag{14}$$

se obtiene la probabilidad estimada

$$\begin{aligned} \hat{p}(\hat{\mathbf{x}}) &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{V} \phi_\delta \left(\frac{\|\hat{\mathbf{x}} - \mathbf{x}_i\|_\infty}{h} \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \delta \left(\hat{\mathbf{x}}, \mathbf{x}_i \right) \end{aligned} \tag{15}$$

Ventana de Parzen clásica: ejemplo en un espacio bidimensional:

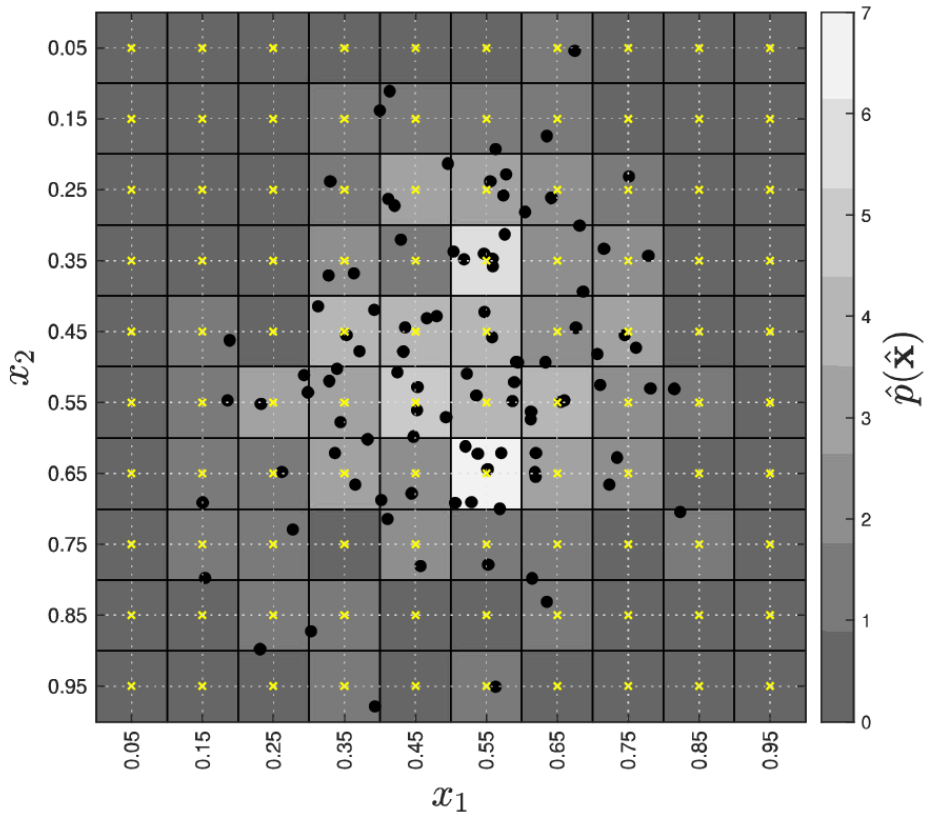
(a) la función $\phi_\delta(\mathbf{x}_i)$ es 1 para cada punto $\mathbf{x}_i$ que cae dentro del cuadro de área unitaria centrado en el origen y será 0 para todo punto fuera de él.



(a) (b)

(b) La función $\phi_\delta(\mathbf{x}_i)$ es 1 si $\mathbf{x}_i$ cae dentro del cuadro de área unitaria centrado en $\hat{\mathbf{x}}$ y 0 en caso contrario. El ancho de banda h reduce o aumenta el lado del cuadrado al valor h.

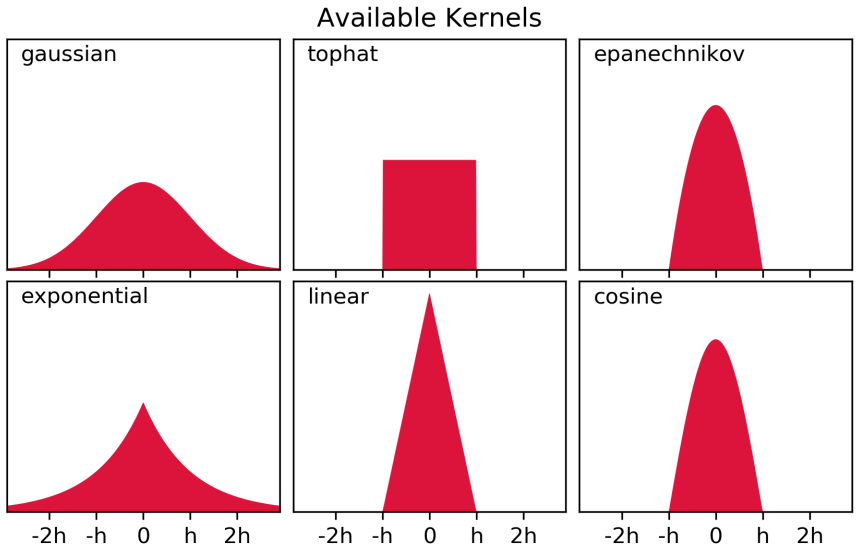
Ventana de Parzen clásica



Estimación de densidad: la cuadrícula divide el espacio de características en regiones de área h^2, cuyos centros $\hat{\mathbf{x}}$ están marcados con puntos amarillos. A mayor densidad de puntos en una región, aumenta el valor de $\hat{p}(\hat{\mathbf{x}})$.

Ventana de Parzen: kernels

- ϕ_δ cuenta puntos dentro de la ventana de ancho h. Produce un estimador muy intermitente.
- El método puede generalizarse reemplazando ϕ_δ por una función simétrica y suave, llamada kernel.
- Opciones generales son: Gaussian Epanechinkov, Exponencial, Linear, Cosine, TopHat.



Ventana de Parzen: kernel Gaussiano

Kernel Gaussiano con media-cero y varianza unitaria:

$$\phi_{\mathcal{N}}(u) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{u^2}{2}\right) \tag{16}$$

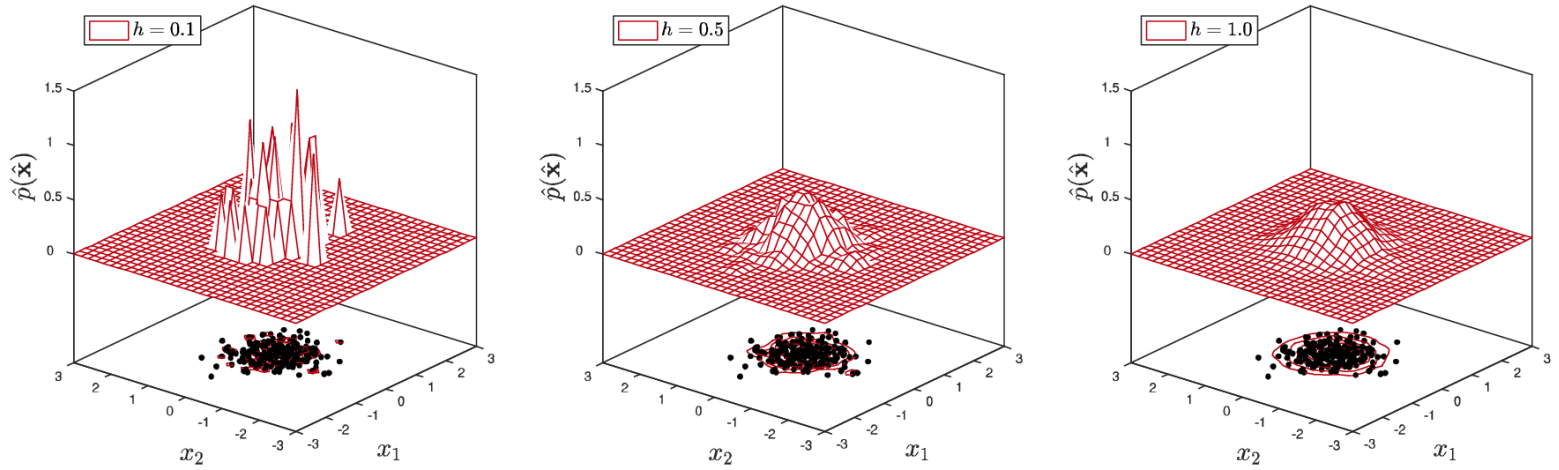
- La estimación de densidad $\hat{p}(\hat{\mathbf{x}})$ resulta

$$\begin{aligned} \hat{p}(\hat{\mathbf{x}}) &= \frac{1}{n}\sum_{i=1}^{n} \frac{1}{V}\phi_{\mathcal{N}}\left(\frac{\|\hat{\mathbf{x}} - \mathbf{x}_i\|_2}{h}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \delta\left(\hat{\mathbf{x}}, \mathbf{x}_i\right) \end{aligned} \tag{17}$$

donde $\|\cdot\|_2$ denota distancia Euclidea y h es el ancho de banda.

Ventana de Parzen en $\mathbb{R}^2$

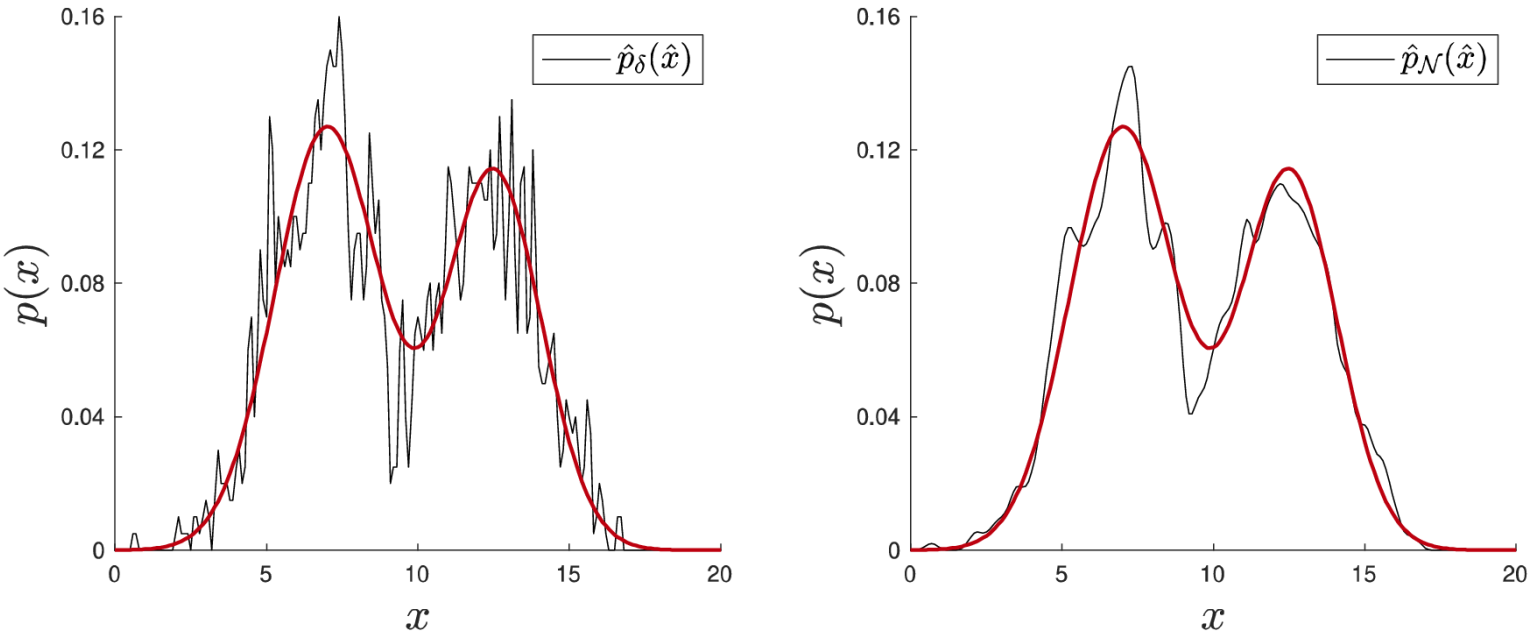
Si h es muy grande, la amplitud de δ es pequeña y el estimador $\hat{p}(\hat{\mathbf{x}})$ está muy suavizado debido a la superposición de varias funciones muy anchas.



Si h es muy pequeña, la amplitud de δ aumenta y el estimador $\hat{p}(\hat{\mathbf{x}})$ es un resultado ruidoso dada la superposición de varias funciones muy angostas.

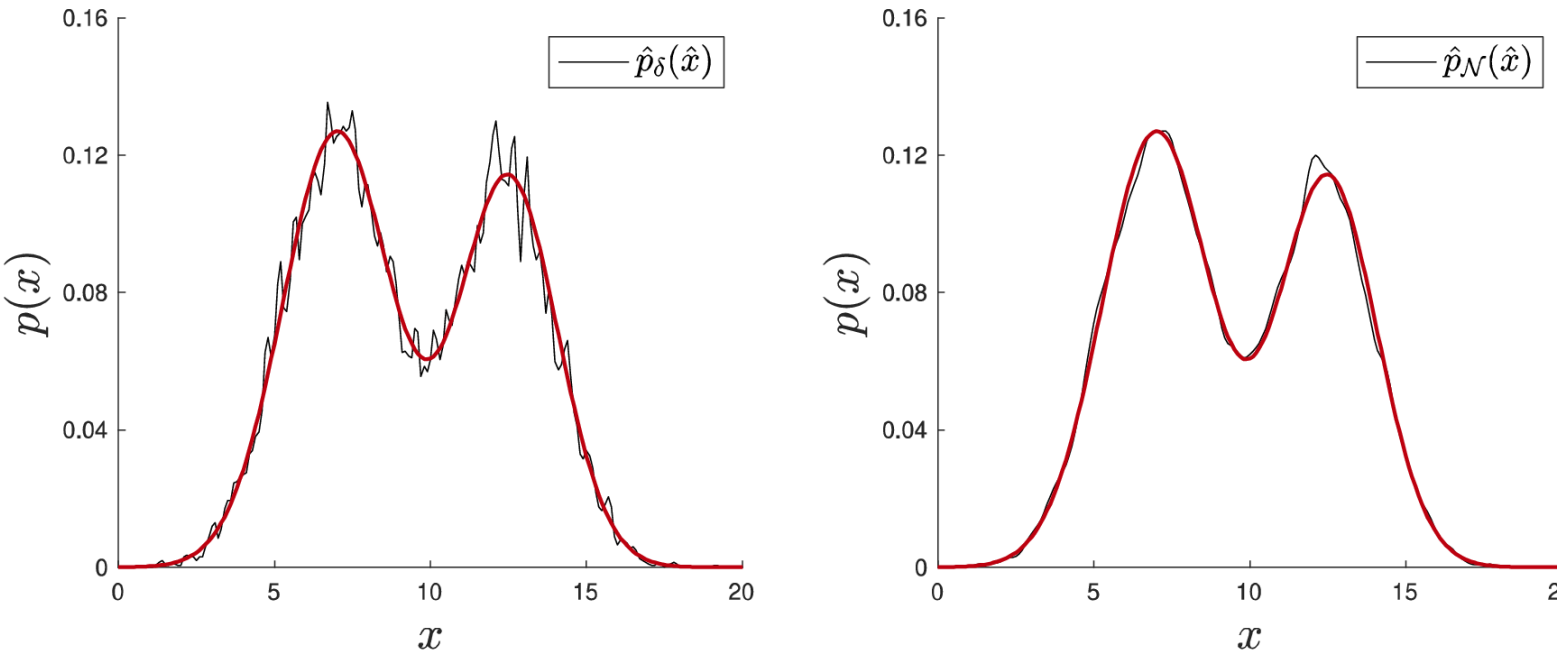
Ventana de Parzen

Efecto del número de muestras y tipo de función de ventana en la estimación de densidad con ventanas de Parzen con $n = 1000$ datos. La curva roja muestra la densidad verdadera.



A la izquierda la ventana es la clásica, a la derecha es el kernel Gaussiano. El ancho de banda es el mismo para ambos kernels, $h = 0.2$.

Ventana de Parzen



En este caso el $h = 0.2$, a la izquierda se estima con el Kernel Cásico y a la derecha con el kernel Gaussiano.

Ventana de Parzen: selección de h

- En el artículo `http://www.ub.edu/stat/personal/minarro/documents/Nonpar.pdf` podemos encontrar en la página 49 un raconto de los principales métodos para selección de ancho de banda.
- En este curso son importantes observar las reglas basadas en distribuciones paramétricas, que son default en sklearn, y la regla plug-in.
- Se deja como ejercicio revisar la regla Plug-in, y verificar que para el caso Gaussiano las transparencias siguientes muestran los pasos necesarios.
- El artículo tiene toda los ”condimentos” para poder entender la mecánica de estimación.

Ventana de Parzen: selección de h regla plug in caso Gaussiano

- Para el kernel Gaussiano en (17) y $\mathbf{x} \in \mathbb{R}$ se calcula:
  - Paso 1 - Estimar ψ_8 dado un estimador de dispersión $\hat{\sigma}$:

$$\hat{\psi}_8^{\hat{\sigma}} = \frac{105}{32\pi^{1/2}\hat{\sigma}(x)^9} \tag{18}$$

  - Paso 2 - Estimar ψ_6 usando el estimador $\hat{\psi}_6\left(g_1\right)$ donde

$$g_1 = \left(\frac{11.9683}{\hat{\psi}_8^{\hat{\sigma}} n}\right)^{1/9} \tag{19}$$

  - Paso 3 - Estimar ψ_4 usando el estimador $\hat{\psi}_4\left(g_2\right)$ donde

$$g_2 = \left(-\frac{2.3937}{\hat{\psi}_6\left(g_1\right) n}\right)^{1/7} \tag{20}$$

Ventana de Parzen: selección de h

- -
  - Paso 4 - El ancho de banda seleccionado es:

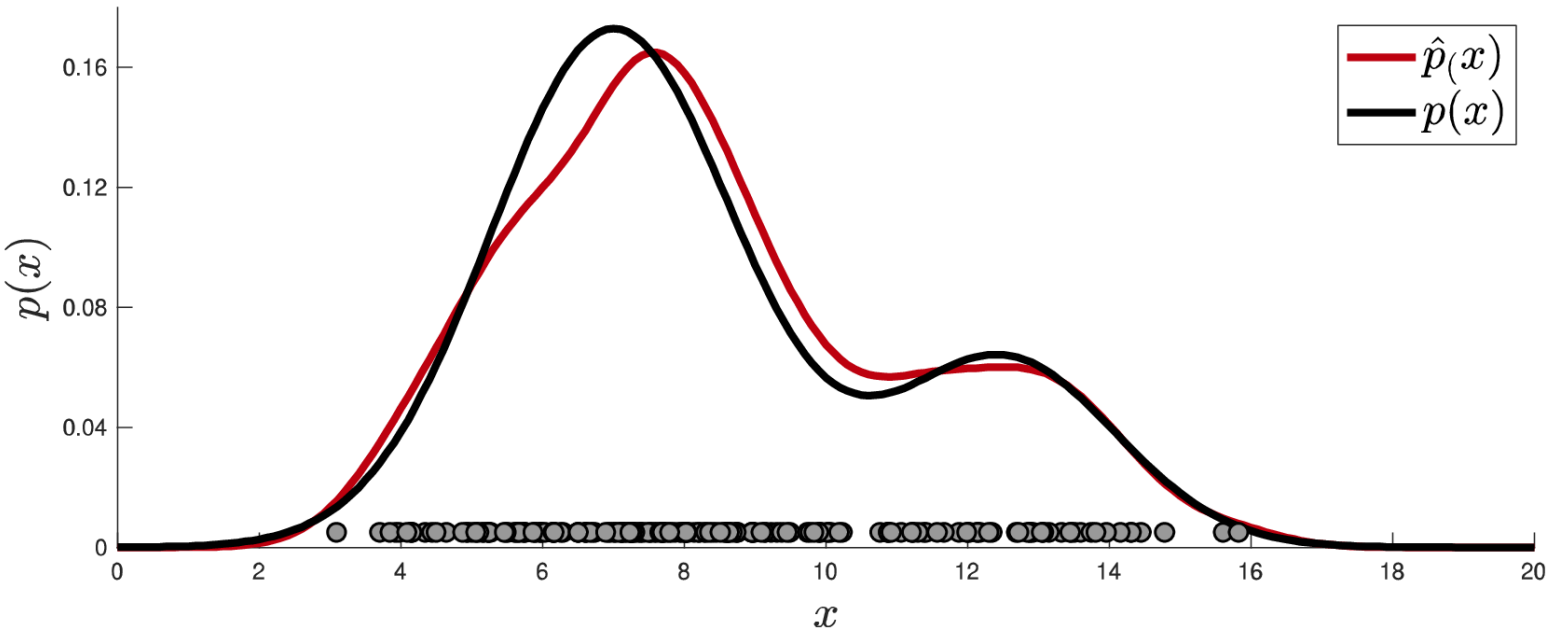
$$h = \left(\frac{0.2821}{\hat{\psi}_4 \left(g_2 \right) n} \right)^{1/5} \tag{21}$$

- En los pasos 2 y 3 el estimador $\hat{\psi}_r(g)$ es

$$\hat{\psi}_r(g) = (n(n-1))^{-1} g^{(-r-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \phi_{\mathcal{N}}^{(r)} \left(\frac{x_i - x_j}{g} \right) \tag{22}$$

donde $\phi_{\mathcal{N}}^{(r)}$ es la r-ésima derivada de $\phi_{\mathcal{N}}$

Ventana de Parzen



Estimación de densidad con kernel Gaussiano para un conjunto de muestras tomadas de una distribución bimodal. La regla de plug-in directa estimó $h = 0.71$.

Ventana de Parzen: clasificación

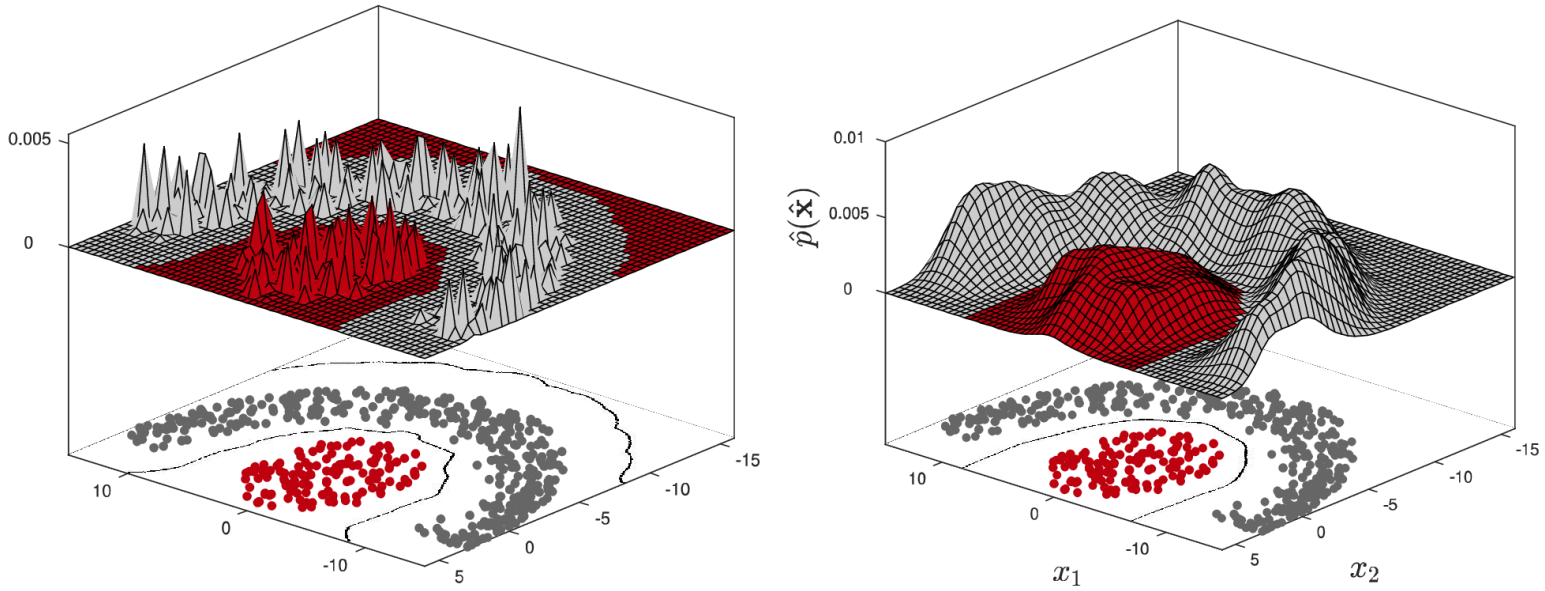
- Un clasificador basado en ventanas de Parzen no requiere una fase explícita de entrenamiento.
- Dado un patrón de prueba $\mathbf{x}_t$, se estima la probabilidad a posteriori de pertenecer a la clase$\omega_i, i = 1, \ldots, c$, como:

$$\begin{aligned} p\left(\omega_i|\mathbf{x}_t\right) &= p\left(\mathbf{x}_t|\omega_i\right) p\left(\omega_i\right) \\ &= \frac{p\left(\omega_i\right)}{n_i} \sum_{j=1}^{n_i} \frac{1}{h^d} \phi_{\mathcal{N}} \left(\frac{\left\|\mathbf{x}_j - \mathbf{x}_t\right\|_2}{h} \right) \end{aligned} \tag{23}$$

- Se aplica la regla de decisión Bayesiana:

$$\text{Decidir } \omega_i \text{ si } p\left(\omega_i|\mathbf{x}_t\right) > p\left(\omega_j|\mathbf{x}_t\right) \text{ para todo } i \neq j \tag{24}$$

Ventana de Parzen



Clasificación con ventanas de Parzen: (a) $h = 0.1$ y (b) $h = 1$. Para valores de h grandes, las fronteras de decisión que se generan son más suaves que aquellas generadas con valores pequeños de h.

Ventana de Parzen: python

Estos son algunos de los recursos Python para hacer los ejercicios.

- `https://sebastianraschka.com/Articles/2014_kernel_density_est.html`
- `https://jakevdp.github.io/PythonDataScienceHandbook/05.13-kernel-density-estimation.html`
- `https://kdepy.readthedocs.io/en/latest/bandwidth.html`