Técnicas Supervisadas

Aproximación no paramétrica

Reconocimiento de Patrones

2003

- *Notas basadas en el curso Reconocimiento de Formas de F.Cortijo, Univ. de Granada y en el libro Pattern Clasification de Duda, Hart y Storck*
- *Parte del material se extrajo de las notas:Técnicas Supervisadas II: Aproximación no paramétrica de F.Cortijo, Univ. de Granada*

Contenido

1. Introducción
2. Estimación de la función de densidad
    1. Estimadores de Parzen
    2. Estimación mediante los k vecinos más próximos
3. Método de clasificación del vecino más próximo
4. Edición del conjunto de entrenamiento
5. Condensado
6. Métodos de aprendizaje adaptivo
7. Árboles de clasificación

Clasificación

- Considerando que el clasificador de Bayes es el clasificador que proporciona el mínimo error :

$$P(w_i/x)=\frac{p(x/w_i)P(w_i)}{P(x)} \qquad P(x)=\sum_{i}^{J} p(x/w_i)P(w_i)$$

Objetivo: etiquetar x:

6. Estimando directamente la probabilidad a posteriori.
7. Estimando las funciones densidad y determinando la clase a partir de las J funciones discriminantes

Técnicas Supervisadas

- Aprendizaje, estimación basada en un conjunto de entrenamiento, prototipos, a los cuales le conozco la clase a la que pertenecen.

Aprendizaje supervisado

- **Aproximación paramétrica**
  - Conozco la estructura estadística de las clases, funciones de densidad de probabilidad conocida estimo parámetros que las determinan
- **Aproximación no paramétrica**
  - Estimo el valor de la función densidad o directamente la probabilidad a posteriori de que un x pertenezca a la clase w_j a partir de la información proporcionada por el conjunto de prototipos.

Ejemplo sencillo

Sistema de trasmisión digital: 0, 1 sobre un canal ruidoso

Quiero estimar el valor trasmitido a partir del

Valor analógico recibido.

Prototipos de la clase 0:{.1,-.2,.25 ,.2}

Prototipos de la clase 1:{.6,.35,.75 ,1.2}

Cuando recibo **x** quiero saber a cual clase pertenece

Estimación de densidades

- 1 sola clase
- Quiero estimar p(x) a partir de n muestras

$$P = \int_R p(x')\,dx'$$ Estimación suavizada de p(x)

$$P_k = C_n^k P^k (1-P)^k$$

$$E(k) = nP \qquad E(k/n) = P$$

$$P = \int_R p(x')\,dx' \approx p(x)\,V$$

$$p(x) \approx \frac{k/n}{V}$$

Estimación de densidades

$$p_n(x) \approx \frac{k_n / n}{V_n}$$

$p_n(x)$ converge a $p(x)$ si:

$$\lim_{n \to \infty} V_n = 0$$

$$\lim_{n \to \infty} k_n = \infty$$

$$\lim_{n \to \infty} \frac{k_n}{n} = 0$$

Estimación de p(x):

2. Por ventanas o núcleos: fijo V_n en función de n y calculo cuantas muestras caen en V_n. Ej : $V_n = 1/\sqrt{n}$.

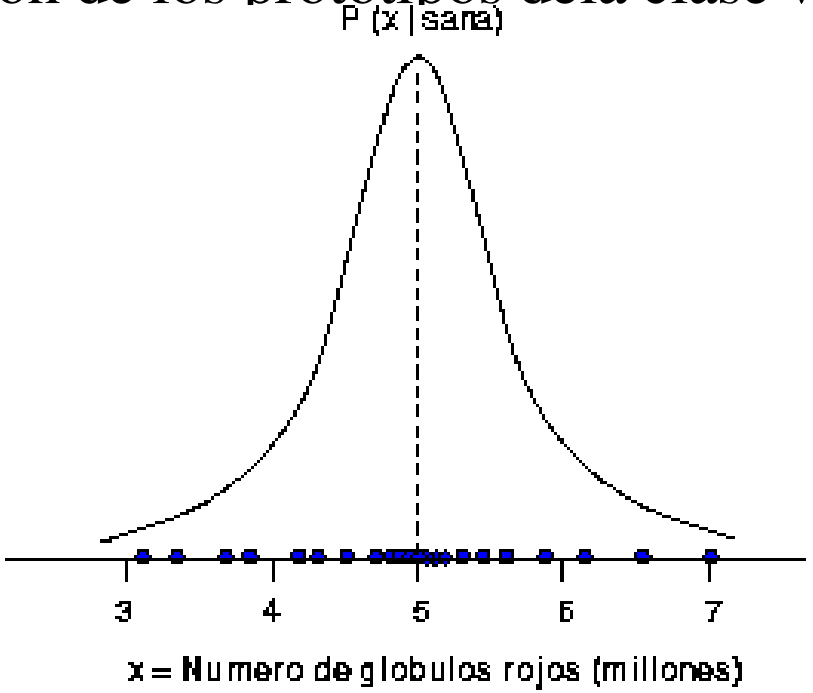
3. k-vecinos: especifico $k_n = \sqrt{n}$ y determino V_n.

Estimación de la función densidad

- ¿De que forma puedo utilizar la información suministrada por el conjunto de entrenamiento para inferir p(x/wi)?

Heurística:

- Muestreo y estimo la densidad de probabilidad de acuerdo a la distribución de los prototipos dela clase wi.



Estimación de densidades

- Sea ki numero de prototipos de la clase wi en un volumen V centrado en x.
- n_i numero total de prototipos de la clase wi

$$p(x/w_i) \approx \frac{k_i/n_i}{V}$$

Probabilidades a priori

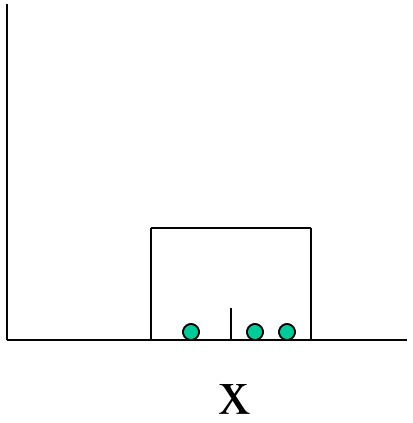
- P(wi) la estimo como $\mathbf{n_i/n}$ o supongo que son equiprobables

Estimación por ventanas

- Considero una ventana centrada en x y veo cuantos prototipos de la clase w_i caen.

$$p(x/w_i) \approx \frac{k_i}{n_i V} = \frac{1}{n_i} \sum_{1}^{m} K(x, Z_i^m)$$

$$K(x, Z_i^m) = \begin{cases} \frac{1}{V} si\ \delta(x, Z_i^m) \leq \rho \\ 0\, si\ \delta(x, Z_i^m) > \rho \end{cases}$$

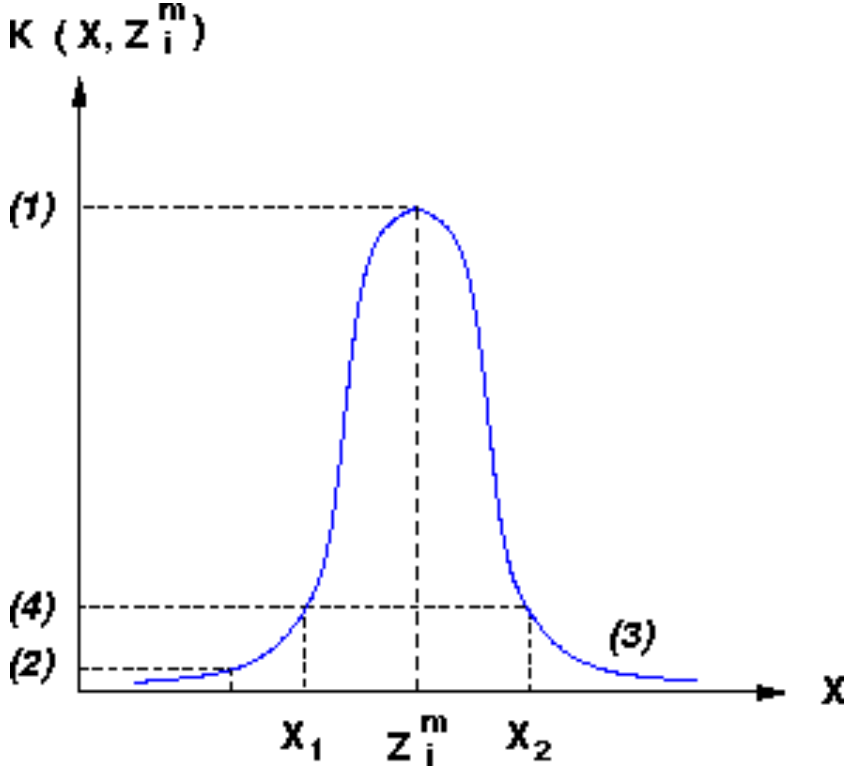


Estimadores de Parzen

- Estimación por núcleos.
- ¿qué información proporciona cada prototipo individualmente de p(x/wi)?
- Z_i^m m-esimo prototipo de la clase wi
  - $P(Z_i^m/wi)>0$
  - Suponiendo continuidad $p(x/wi)\geq 0$ en la vecindad de Z_i^m
  - Cuando me alejo la influencia se hace menor

Estimación mediante núcleos

- $K(x, Z_i^m)$ función centrada en Z_i^m que alcanza un máximo en Z_i^m y que decrece en forma monótona cuando aumenta la distancia



Rango de influencia del núcleo

- La contribución de un prototipo depende del ancho del núcleo y la forma del núcleo.
- El estimador es muy dependiente de los datos disponibles.
- Si las muestras estan muy dispersas ρ tiene que ser grande.
- Si las muestras estan muy agrupadas ρ tiene que ser menor.

Forma general de la función nucleo

$$K(x, Z_i^m) = \frac{1}{\rho^d} h\left[\frac{\delta(x, Z_i^m)}{\rho}\right]$$

$$\lim_{n_i \to \infty} \rho^d(n_i) = 0$$

$n_i \Uparrow \rightarrow \rho \Downarrow$

$\delta(x, Z_i^m)$ métrica determinada por el nucleo

máx(h)=h(0), h monótona decreciente con $\delta(x, Z_i^m)$

Forma general de la función nucleo

Si h>0

$$\int K(x, Z_i^m)\, dx = 1$$

$$\hat{p}(x/w_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} K(x, Z_i^j)$$

$p(x/w_i)$ Función densidad de probabilidad

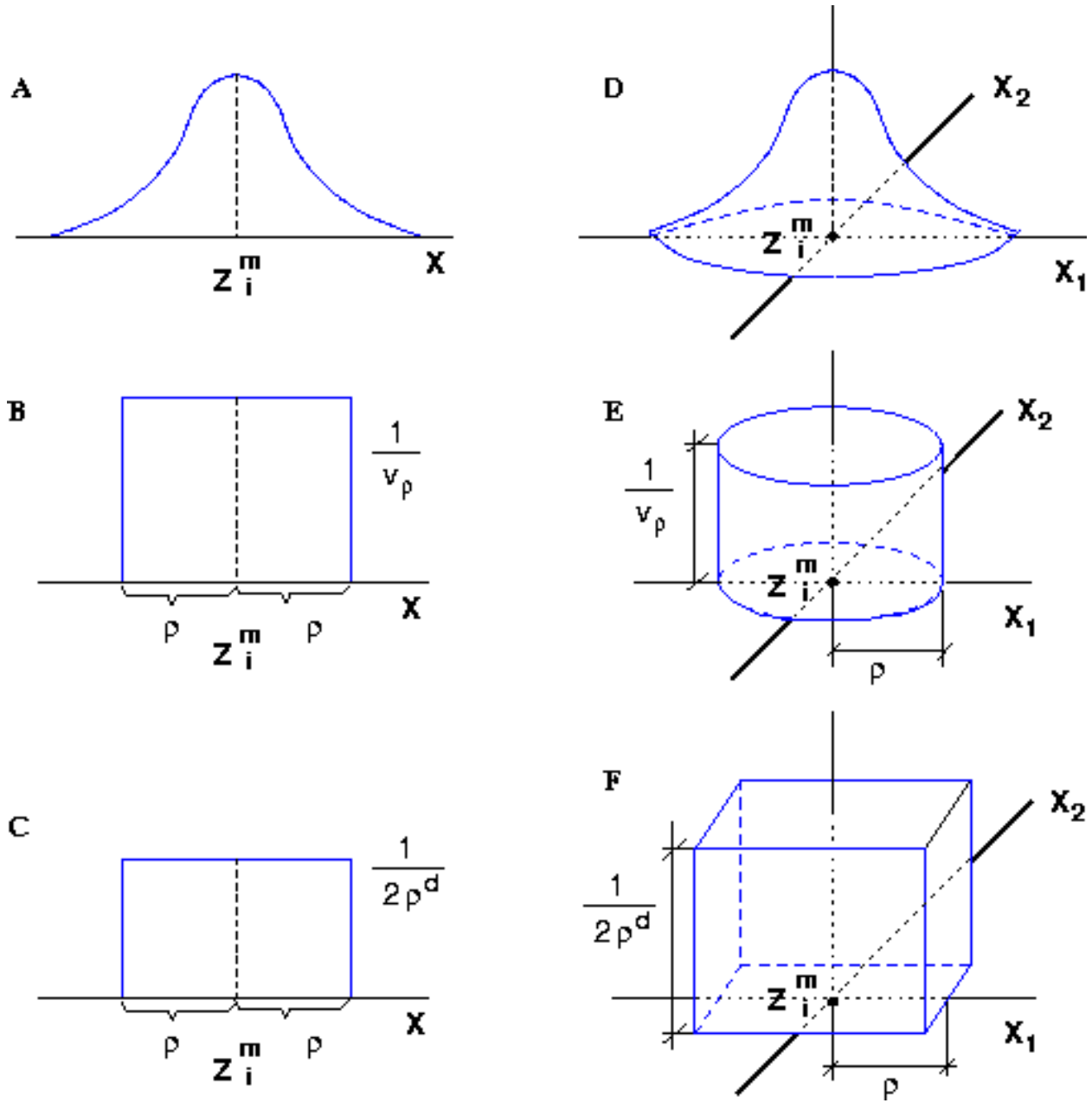
Elección de ρ

$\forall$ ρ=cte

$\forall$ ρ dinámico $\quad K(x, \rho(x), Z_i^m)$

$$\hat{p}(x/w_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} K(x, \rho_j, Z_i^j)$$

$$\rho_j = \sqrt{media\, \delta^2(x_j, x_i)}$$


A
z_i^m
X
B
$\frac{1}{v_\rho}$
X
ρ
ρ
z_i^m
C
$\frac{1}{2\rho^d}$
X
ρ
ρ
z_i^m
D
X_2
z_i^m
X_1
E
X_2
$\frac{1}{v_\rho}$
z_i^m
X_1
ρ
F
X_2
$\frac{1}{2\rho^d}$
z_i^m
X_1
ρ

Núcleo Gaussiano

$$h(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

$$K(x, Z_i^m) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x - Z_i^m)^T \sum^{-1} (x - Z_i^m)}$$

- Distancia de Mahalanobis, ρ distinto en cada dimensión y se contempla correlación entre variables. Si matriz es diagonal, la correlación es 0 y la distancia es euclídea.
- Estimador suave, computacionalmente muy costoso

Núcleo hiperesférico

$$K(x, Z_i^m) = \begin{cases} \frac{1}{V} & si \left\{ d(x, Z_i^m) \leq \rho \right\} \\ 0 & si \left\{ d(x, Z_i^m) > \rho \right\} \end{cases}$$

- Ventaja: eficiencia computacional (cálculo de distancia y suma). Útil cuando tengo muchas muestras.
- Desventaja: estimación constante por tramos

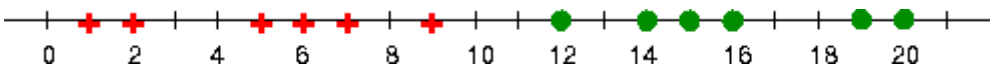
Núcleo hipercúbico

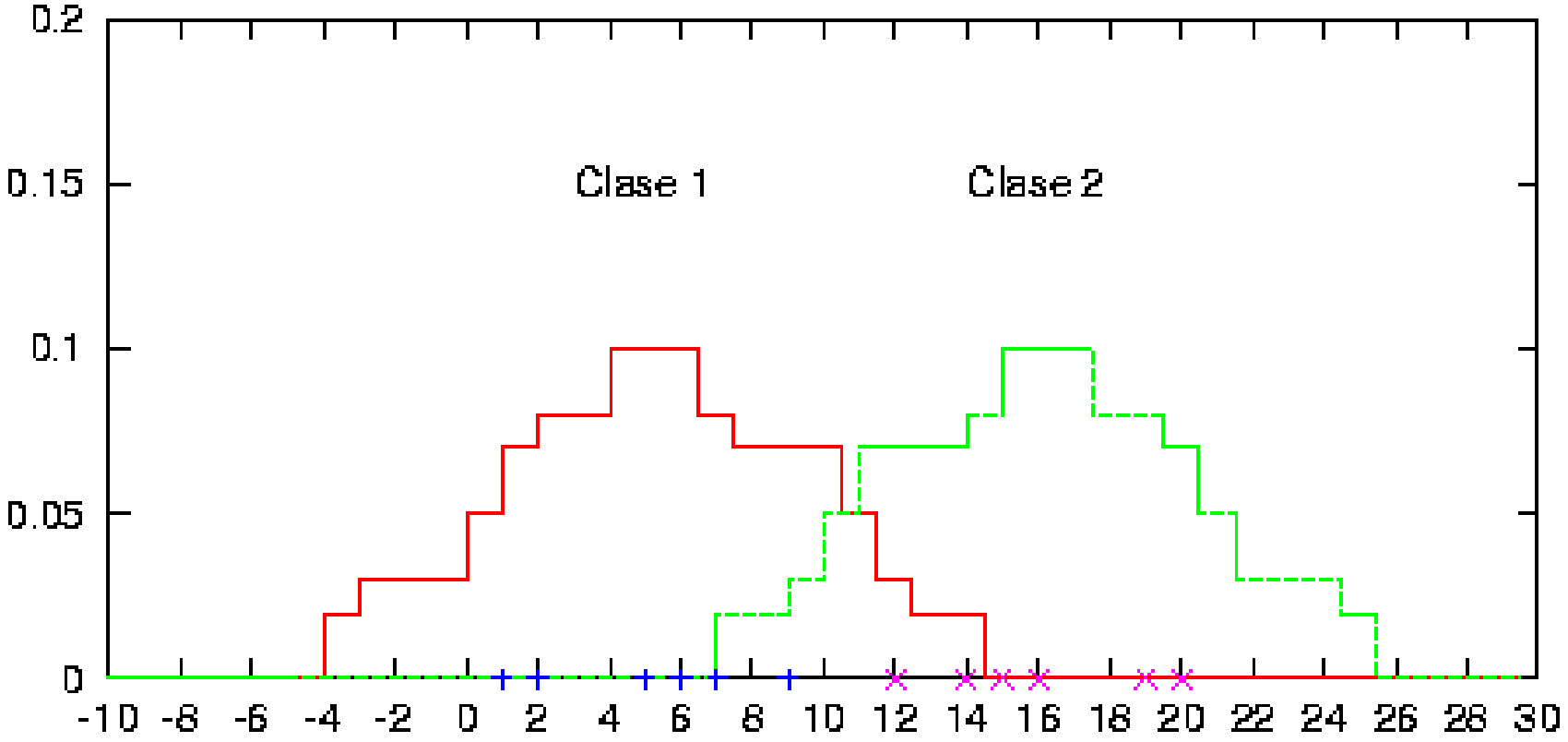
$$K(x, Z_i^m) = \begin{cases} (2\rho)^{-d} & si\left\{\delta_T(x, Z_i^m) \le \rho\right\} \\ 0 & si\left\{\delta_T(x, Z_i^m) > \rho\right\} \end{cases}$$

$$\delta_T(x, Z_i^m) = \underset{j=1..d}{máx}\left\{|x_j - Z_i^m|\right\}$$ Distancia de Chevyshev

- Ventaja: eficiencia computacional .Cálculo de la distancia más eficiente.
- Desventaja: estimación constante por tramos

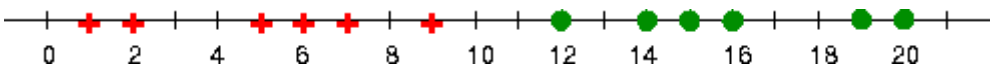
Ejemplo

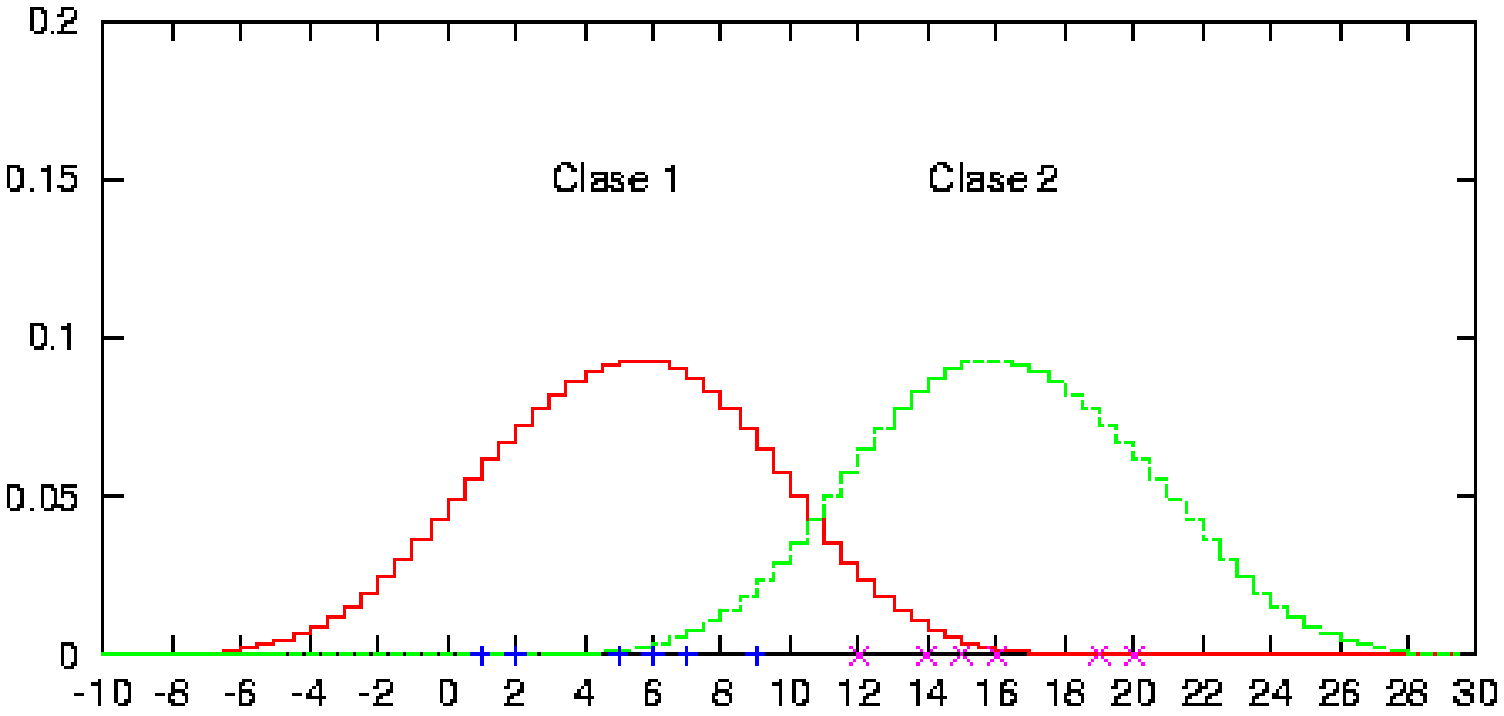




Hipercubo ρ=5

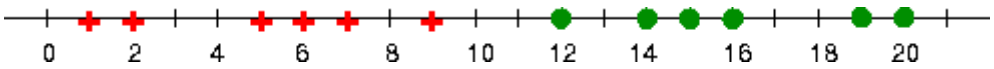
Ejemplo

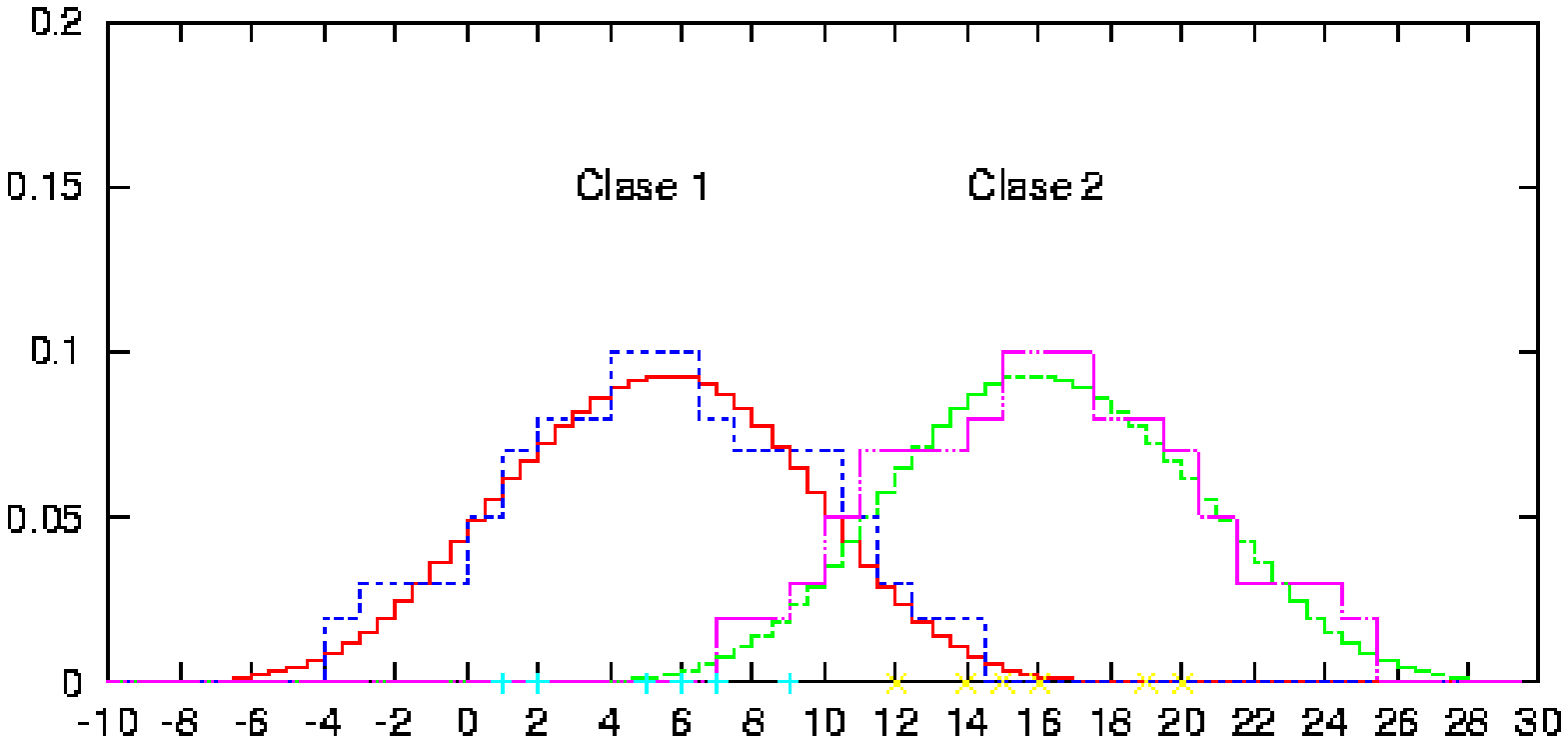




Gaussiano ρ=3

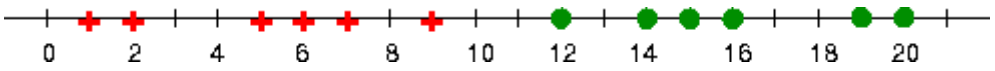
Ejemplo

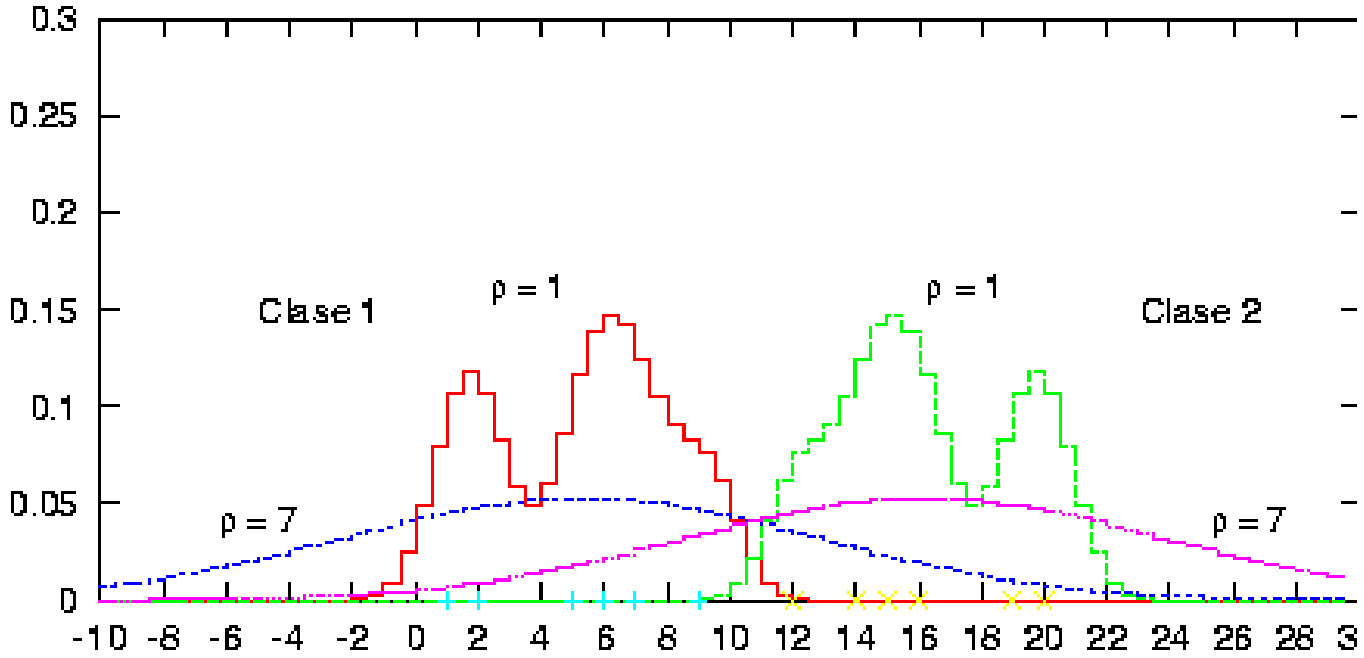




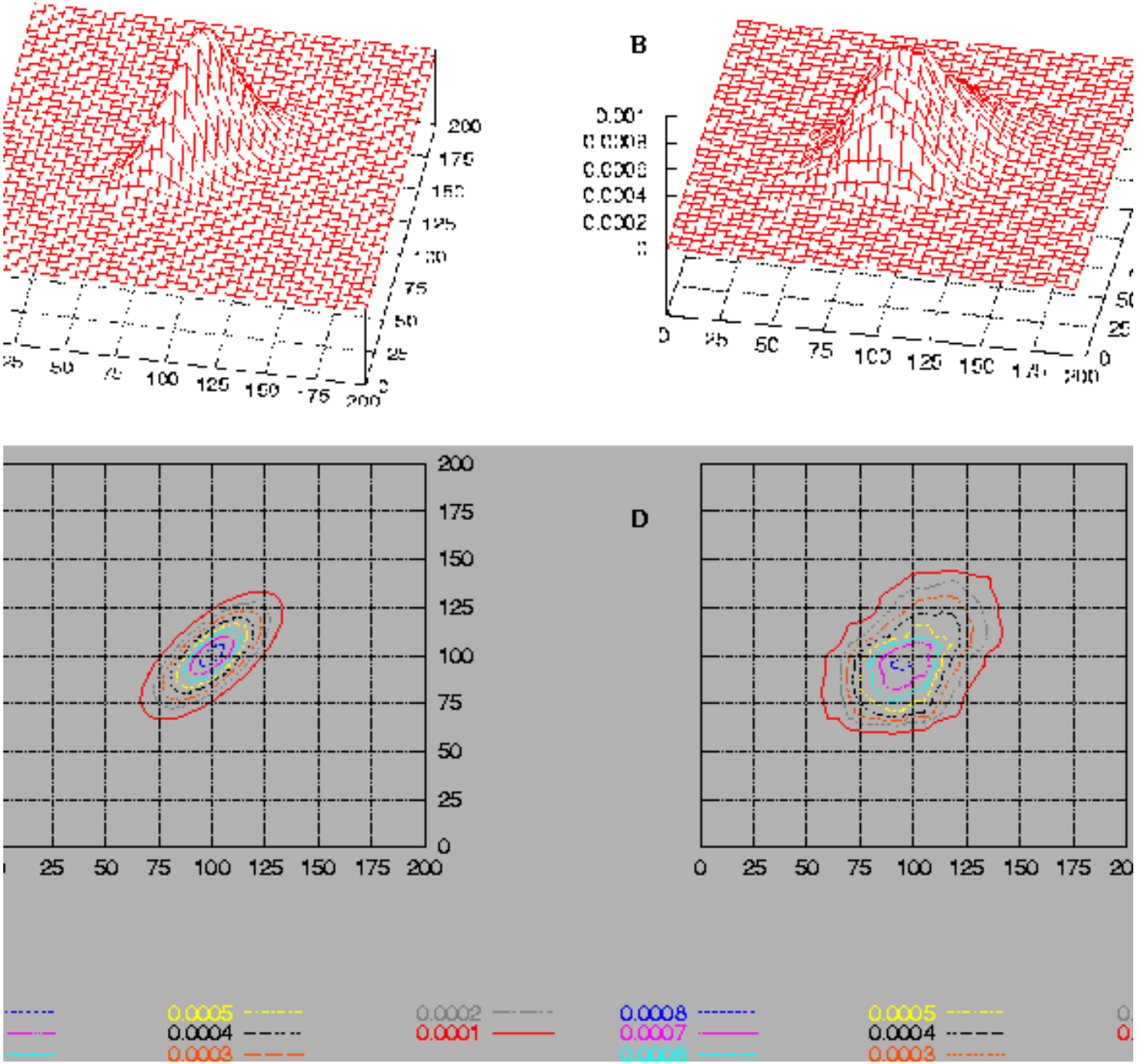
Comparación

Ejemplo





Comparación para un mismo núcleo y distintos anchos

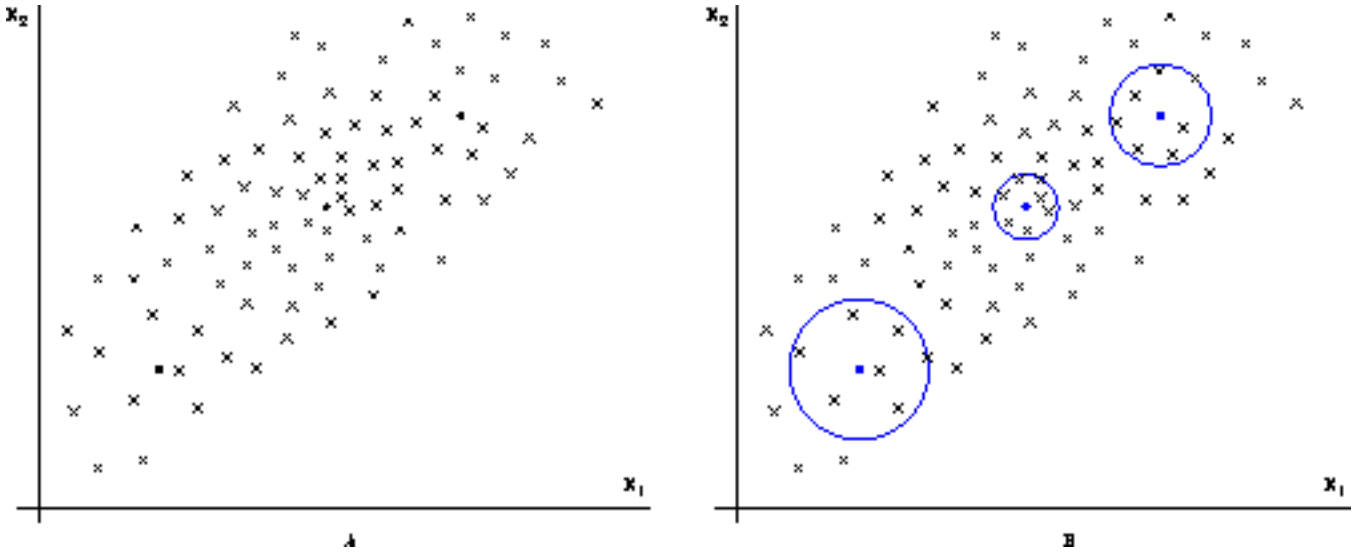


Verdadera

Estimada: hipercubo ρ=20

Estimación mediante los k vecinos más próximos

- Heuristica: El volumen que encierra un número fijo k de prototipos es menor en regiones densamente pobladas que en regiones donde se encuentran más dispersos.
- p(x/wi)=ki/niV(x)



Elección de k

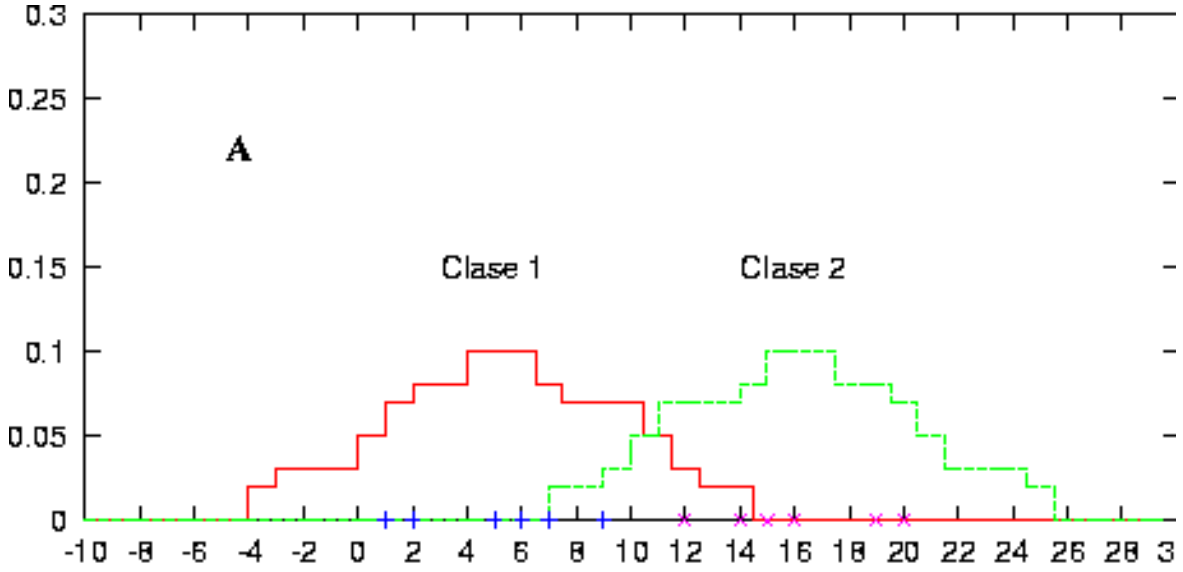
- k depende de n_i
- El estimador es consistente e insesgado si se verifica:

$$\lim_{n_i \to \infty} k_i(n_i) = \infty$$

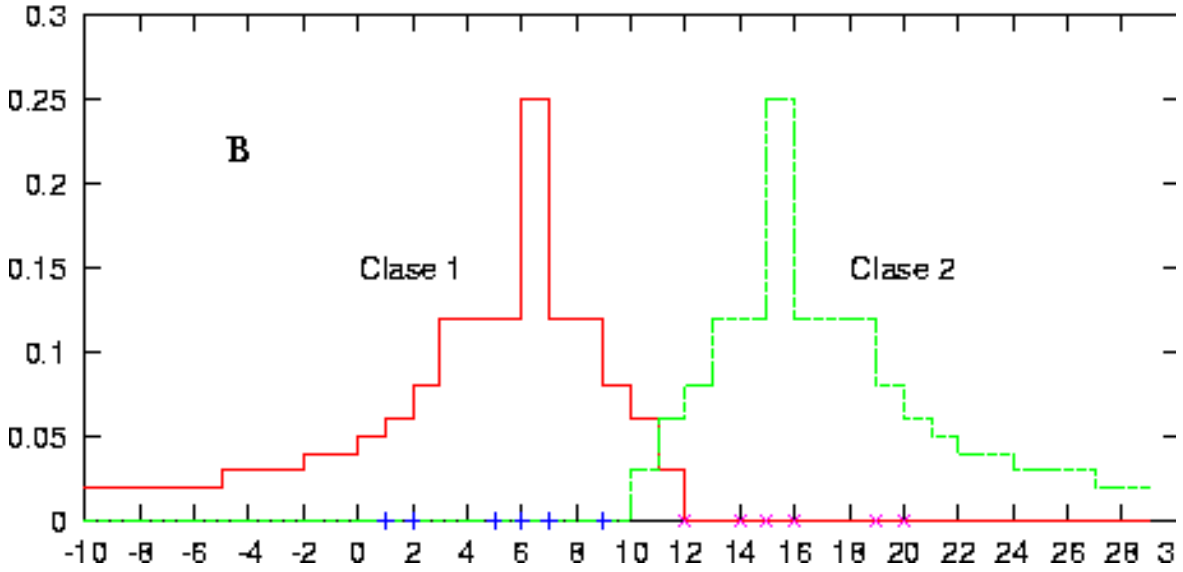
$$\lim_{n_i \to \infty} \frac{k_i(n_i)}{n_i} = 0$$

Se puede elegir: $k_i(n_i) = c\sqrt{n_i}$

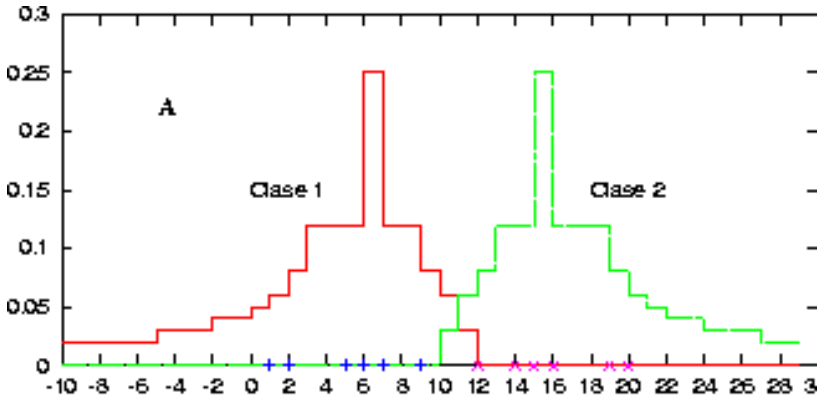
Hipercubo ρ=5



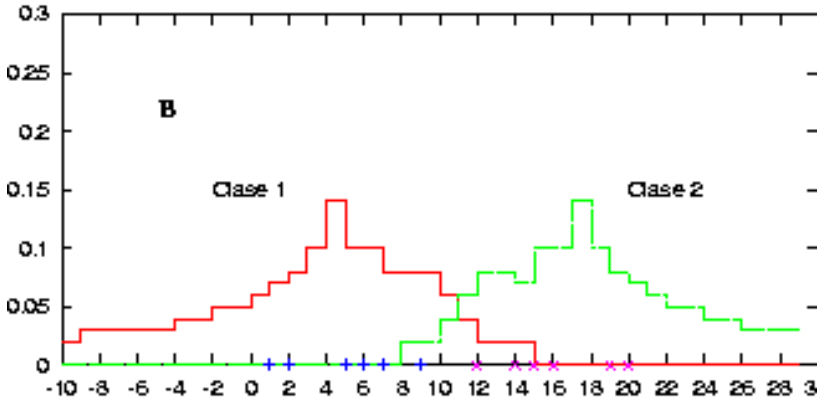
K-vecinos k=3



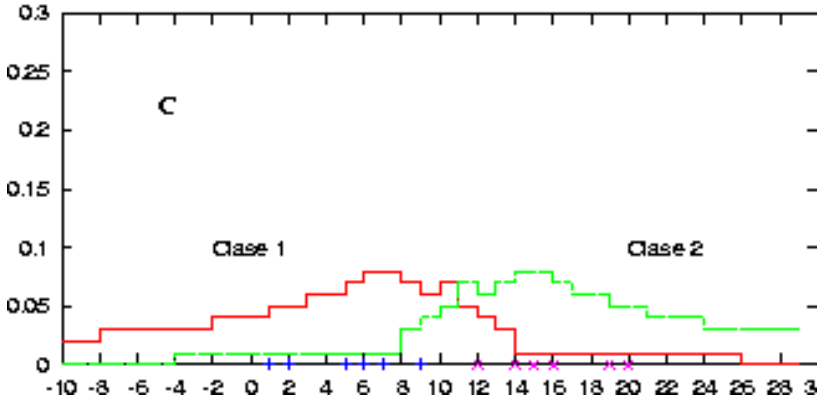
K-vecinos k=3



K-vecinos k=5



K-vecinos k=7



Estimación directa de las probabilidades a posteriori

$$p_n(x, w_i) = p_n(x/w_i)\,P(w_i) = \frac{k_i}{n_i V}\frac{n_i}{n} = \frac{k_i}{nV}$$

$$P_n(w_i/x) = \frac{p_n(x, w_i)}{\sum_{j=1}^{J} p_n(x, w_j)} = \frac{k_i}{k}$$

Regla de clasificación de los k-vecinos más cercanos k-NN

- Elegimos para x la clase más frecuente de la celda
- Selecciono w_c si $$k_c = \underset{i=1..J}{máx} \left\{ k_i(x) \right\}$$
- $k_i(x)$: número de muestras de la clase w_i entre los k vecinos más cercanos a x.

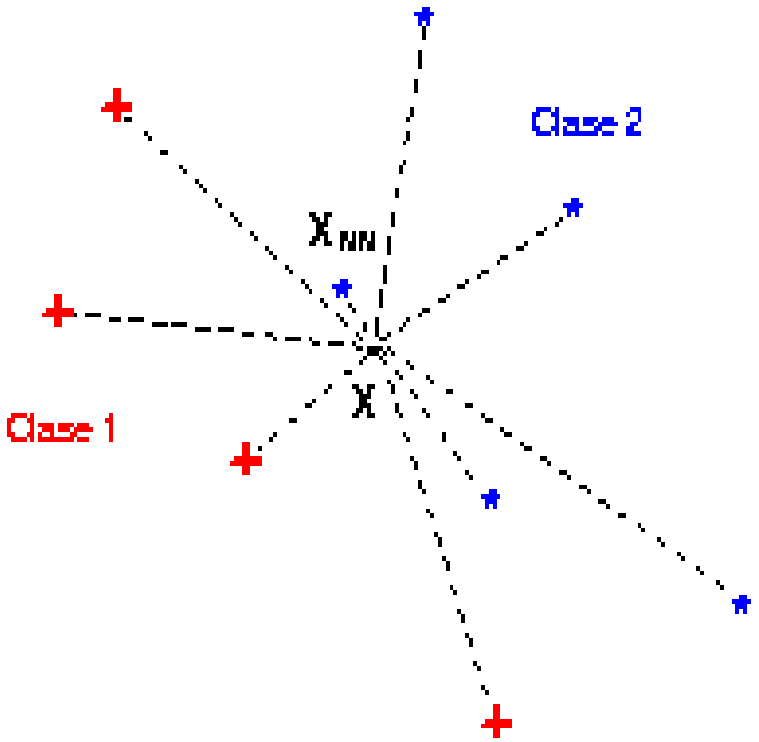
Regla del vecino más cercano 1-NN

- Selecciono w_c si:

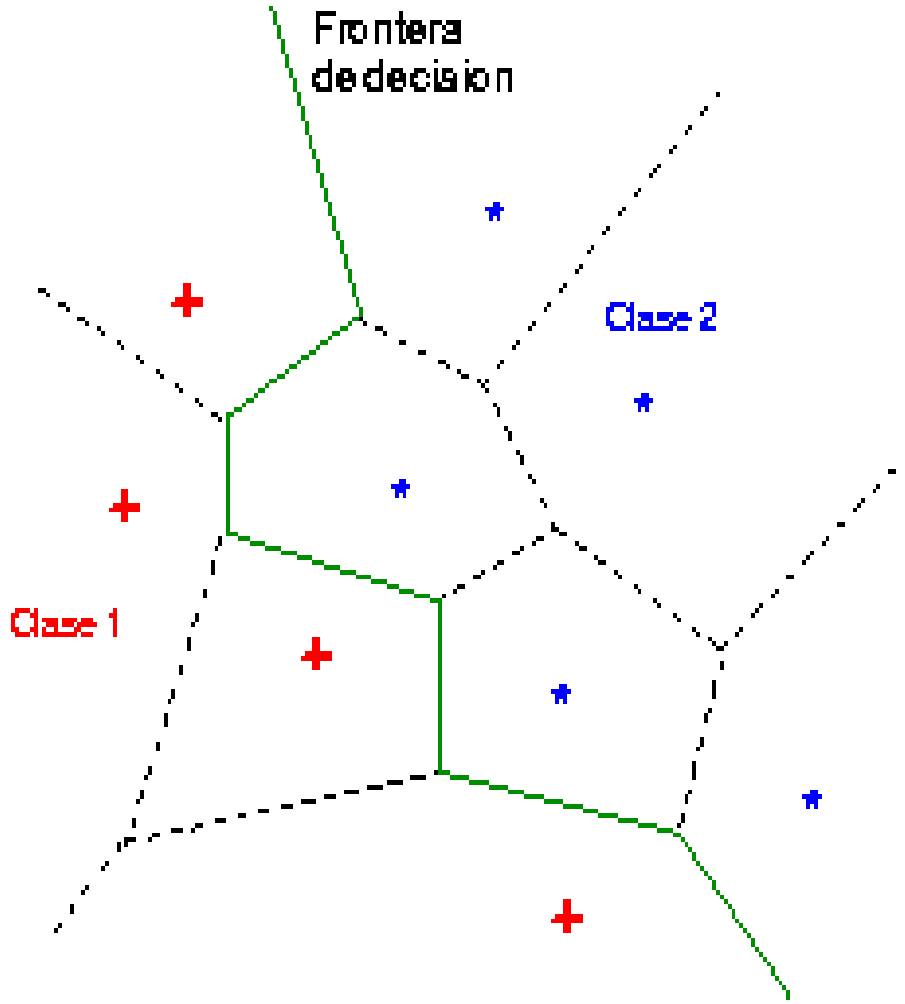
$$d(x, x_{NN}) = \min_{i=1..n} \left\{ \delta(x, x_i) \right\}$$

$$x_{NN} \in w_c$$

- Interpretación: Divide al espacio en n regiones de Voronoi



A



B

Cotas de Error de 1-NN

- Procedimiento subóptimo, tasa de error > que la de Bayes.
- Con un número ilimitado de prototipos el error nunca es mayor que 2 veces Bayes

$$E^{¿} < E_1 \leq E^{¿}\left(2 - \frac{J}{J-1}E^{¿}\right)$$

Error asociado a k-NN

$$E^{¿} \leq \ldots . E_{2k+1} \leq E_{2k} \ldots \leq E_1 \leq 2\,E^{¿}$$

$$Si \quad n \to \infty \quad y \quad k \approx \sqrt{n}$$

$$\lim_{k \to \infty} E_k = E^{¿}$$

Regla (k,t)-NN

- Extensión: clase de rechazo.
- Selecciono la clase w_c si tiene una mayoría cualificada por lo menos t_c son de la clase w_c.
-

$$d(x)=\begin{cases} w_c & k_c = \underset{i=1..J}{máx}\left\{k_i(x)\right\} \geq t_c \\ w_0 & en\ otro\ caso \end{cases}$$

Regla 1-NN(t)

$$d(x)=\begin{cases} w_c & \delta\left(x, x_{NN}\right)=\min\limits_{i=1..n}\left(\delta(x, x_i)\right)\leq t \\ w_0 & en\ otro\ caso \end{cases}$$

.

- t se elige luego de examinar la distribución de las distancias de los patrones

Costo Computacional k-NN

- Requiere explorar todo el conjunto de referencia O(n).
- Calculo de la distancia euclídea lineal con d O(nd)
- Espacio de almacenamiento de todos los prototipos O(nd)
- Inaplicable si tengo un conjunto de referencia grande y alta dimensionalidad

Estrategia:

- Selección de carácterísticas: para bajar **d**
- Disminuir el conjunto de referencia :

  EDICIÓN, CONDENSADO, APRENDIZAJE ADAPTIVO

- Mejorar la eficiencia del cálculo del vecino más cercano: métodos jerárquicos.