Tema 4:

CLASIFICADORES NO PARAMÉTRICOS

Febrero-Mayo 2006

INDICE

4.1 ESTIMACIÓN NO PARAMÉTRICA DE LA FDP

Si no podemos asumir un modelo preciso para $f_{\mathbf{x}}(\mathbf{x}\,|\omega_i)$ hay que recurrir a estimadores no paramétricos de la f.d.p., como el histograma:

- La probabilidad de que $\mathbf{x}$ se encuentre en la región R es:

$$P = \int_R f_{\mathbf{x}}(\mathbf{x}')d\mathbf{x}'$$

- Si disponemos de una serie de observaciones independientes $D = \{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$, la probabilidad de que k de entre los n datos se encuentren en la región viene dada por:

$$P_k = \binom{n}{k} P^k (1-P)^{n-k} \qquad \Rightarrow \qquad E\{k\} = nP$$

Una estimación razonable puede ser: $\hat{P} = k/n$

Si $f_{\mathbf{x}}(\mathbf{x})$ es continua y R es tan pequeña que $f_{\mathbf{x}}(\mathbf{x})$ no varía en su interior:

$$\int_R f_{\mathbf{x}}(\mathbf{x}')d\mathbf{x}' \cong f_{\mathbf{x}}(\mathbf{x})\, V_R$$

combinando ambas expresiones obtenemos el histograma:

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{\int_R f_{\mathbf{x}}(\mathbf{x}')d\mathbf{x}'}{\int_R d\mathbf{x}'} \cong \frac{k/n}{V_R}$$

CONDICIONES DE CONVERGENCIA

El histograma está promediando valores en una región, y por tanto está generando una versión distorsionada de $f_{\mathbf{x}}(\mathbf{x})$.

Para reducir el efecto nos interesa hacer $V_R \to 0$, lo cual hace tender a cero k si el numero de muestras es finito.

¿Cómo podemos garantizar la convergencia de

$$f_n(\mathbf{x}) = \frac{k_n / n}{V_{R,n}}$$

cuando $n \to \infty$?

¿De qué forma diseñaremos $V_{R,n}$?

Para que $\lim_{n\to\infty} f_n(\mathbf{x}) = f(\mathbf{x})$ debe cumplirse que:

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

Podemos garantizar las tres condiciones de dos formas:

1. Parzen: haciendo que $V_{R,n}$ sea una funcion de n.
2. k-nearest neighbors: adaptando el volumen en cada región del dominio $\mathbf{x}$ de forma que k crezca a una velocidad inferior a n.

4.2 VENTANAS DE PARZEN

Supongamos que la región *R* queda definida por una función $\varphi(\mathbf{x})$ que encierra un volumen $V_{R,n}$. El número de datos que encierra esa región alrededor de $\mathbf{x}$ es:

$$k_n = \sum_{i=1}^{n} \varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)$$

si $\varphi(\mathbf{x})$ es un hipervolumen

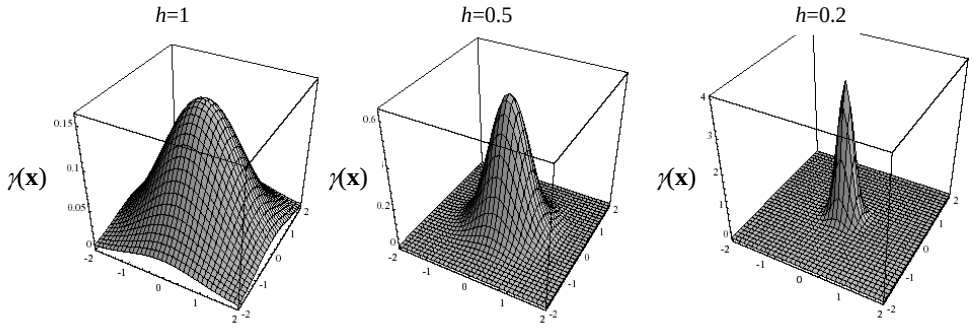
La estimación de la fdp viene dada por:

$$f_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_{R,n}}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right) = \frac{1}{n}\sum_{i=1}^{n}\gamma_n\left(\mathbf{x}-\mathbf{x}_i\right)$$

Para que $f_n(\mathbf{x})$ sea una fdp, $\varphi(\mathbf{x})$ ha de cumplir las condiciones siguientes:

$$\varphi(\mathbf{x}) \geq 0$$
$$\int \varphi(\mathbf{x})d\mathbf{x} = 1$$
$$V_n \propto h_n^d$$

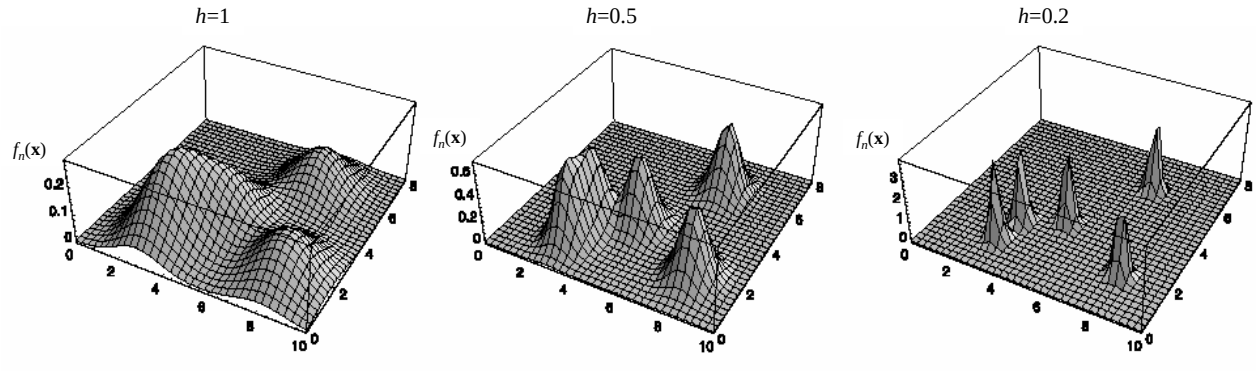
Ejemplo 1:



Dará lugar a una estimación muy promediada

Dará lugar a una estimación muy ruidosa

Estimación de $f_n(\mathbf{x})$ hecha con 5 datos, para los tres distintos anchos de ventana.



MEDIA Y VARIANZA DE LA ESTIMACIÓN

Media de la estimación

$$\overline{f}_n(\mathbf{x}) = E\{f_n(\mathbf{x})\} = \frac{1}{n}\sum_{i=1}^{n} E\left\{\frac{1}{V_{R,n}}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\right\} =$$

$$= \int \frac{1}{V_{R,n}}\varphi\left(\frac{\mathbf{x}-\mathbf{v}_i}{h_n}\right) f(\mathbf{v})d\mathbf{v} = \int \gamma_n(\mathbf{x}-\mathbf{v}_i) f(\mathbf{v})d\mathbf{v}$$

Es una convolución de la ventana con la verdadera fdp

Interpretación

Si $V_{R,n} \to 0$, entonces $\varphi(\mathbf{x}) \to \delta(\mathbf{x})$ y el estimador no es sesgado, pero el número de muestras en cada región va a tender a cero y la estimación no será buena $\Rightarrow$ hay que evaluar la varianza.

Varianza de la estimación

$$\sigma_n^2(\mathbf{x}) = \sum_{i=1}^{n} E\left\{\left(\frac{1}{nV_{R,n}}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right) - \frac{1}{n}\overline{f}_n(\mathbf{x})\right)^2\right\} =$$

$$= nE\left\{\frac{1}{n^2V_{R,n}^2}\varphi^2\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\right\} - \frac{1}{n}\overline{f}_n^2(\mathbf{x}) =$$

$$= \frac{1}{nV_{R,n}}\int \frac{1}{V_{R,n}}\varphi^2\left(\frac{\mathbf{x}-\mathbf{v}}{h_n}\right) f(\mathbf{v})d\mathbf{v} - \frac{1}{n}\overline{f}_n^2(\mathbf{x})$$

Acotando superiormente: $\sigma_n^2(\mathbf{x}) \le \dfrac{\sup(\varphi(.))\,\overline{f}(\mathbf{x})}{nV_{R,n}}$

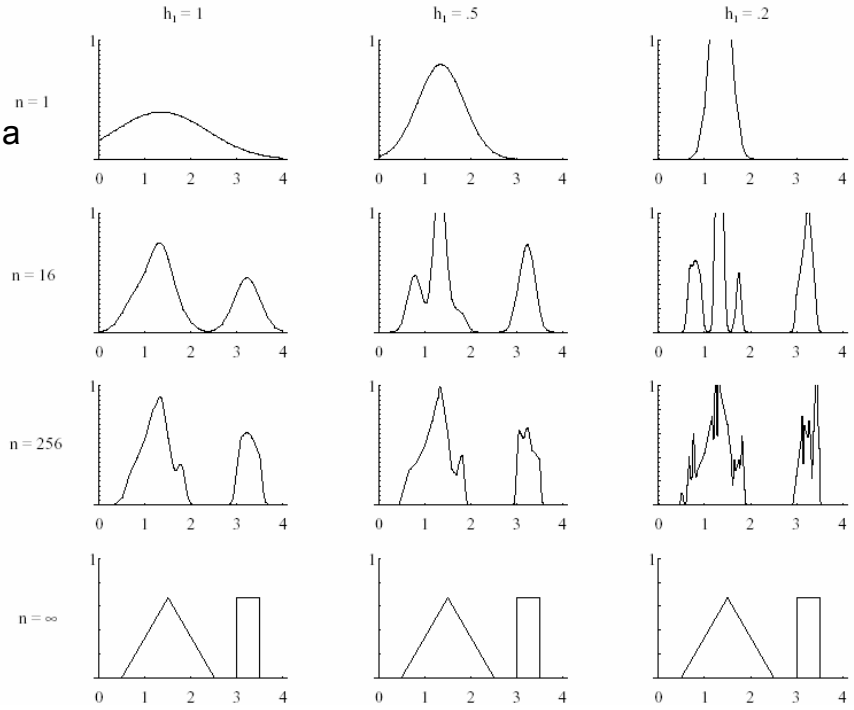
Interpretación

Dado n, para que la varianza sea pequeña $V_{R,n}$ ha de ser grande $\Rightarrow$ mucho sesgo.

Conseguimos que la varianza sea pequeña si $nV_{R,n}$ es grande, cuando $n \to \infty$.

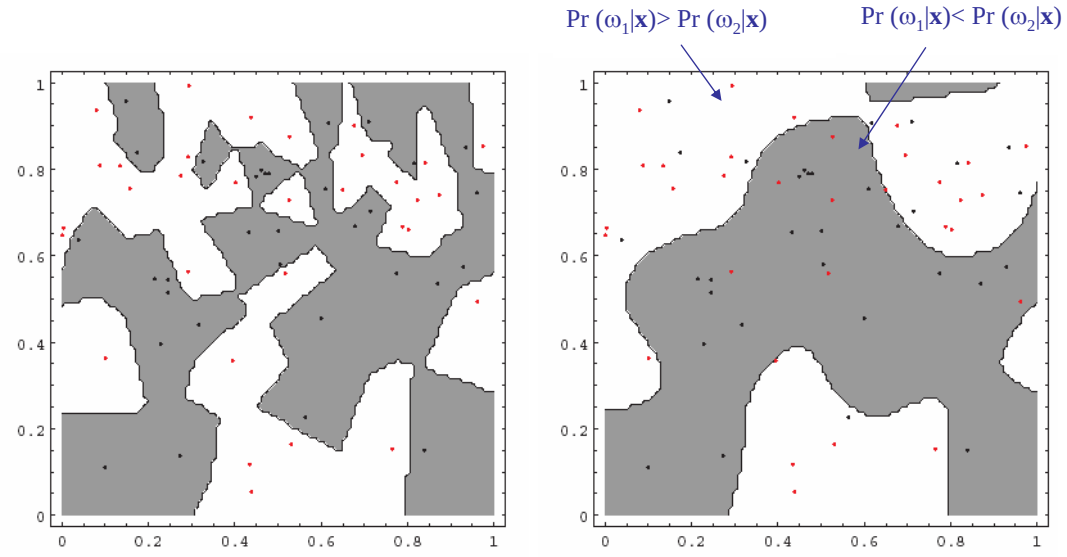
Ejemplo 2:

Ventana Gaussiana



>

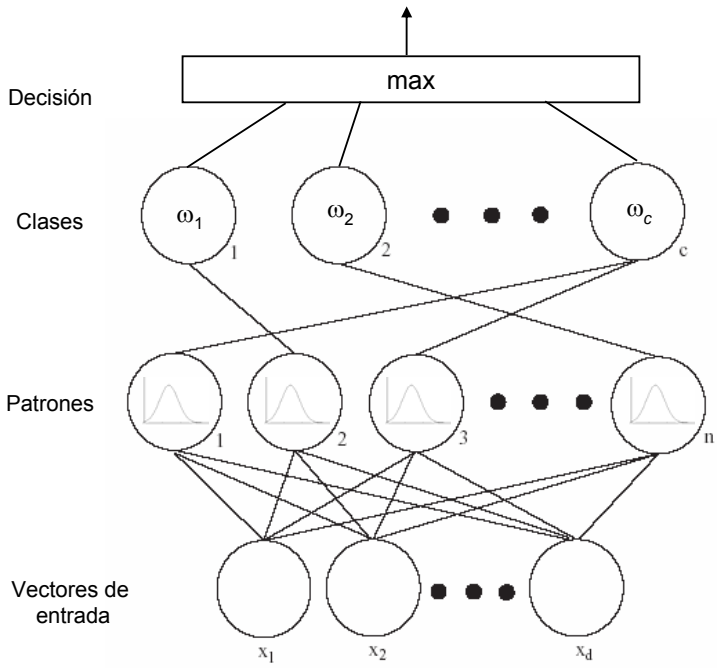
Ejemplo 3: Regiones de decisión para un dicotomizador (2 clases) usando ventanas de Parzen Gaussianas para un valor pequeño de *h* (izquierda) y grande (derecha).



Observación: El tamaño óptimo de la ventana es posiblemente distinto dependiendo de la región que se analice.

4.3 CLASIFICADOR NEURONAL PROBABILÍSTICO

Un clasificador basado en la estimación de la pdf con ventanas de Parzen puede construirse a partir de una red neuronal, que calcule las estimaciones y determine la clase mediante una regla Bayesiana.



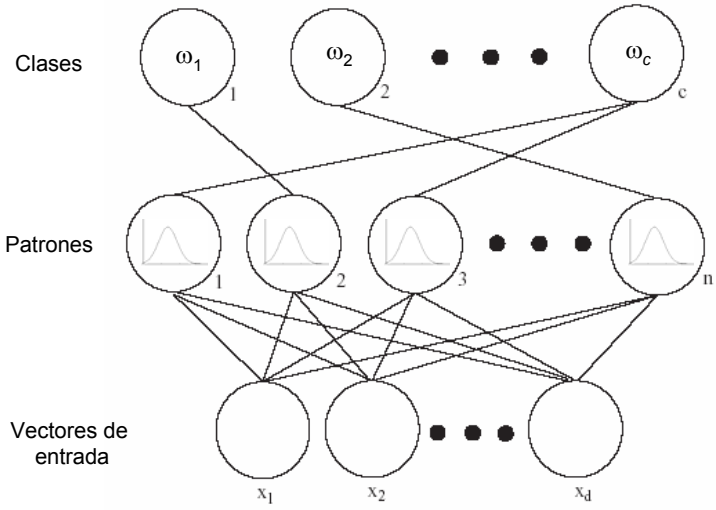
Disponemos de *n* vectores de entrenamiento de dimensión *d*: $\mathbf{x}_i$

Fase de entrenamiento

3. Se asigna un enlace entre el bloque patrón i y la clase a la que pertenece el vector $\mathbf{x}_i$

2. El bloque patrón i ponderará sus entradas con el vector de pesos $\mathbf{w}_i=\mathbf{x}_i$

1. Los vectores de entrenamiento son normalizados.



Fase de clasificación

4. La decisión sobre la clase se toma a partir del máximo $g_j(\mathbf{x})$

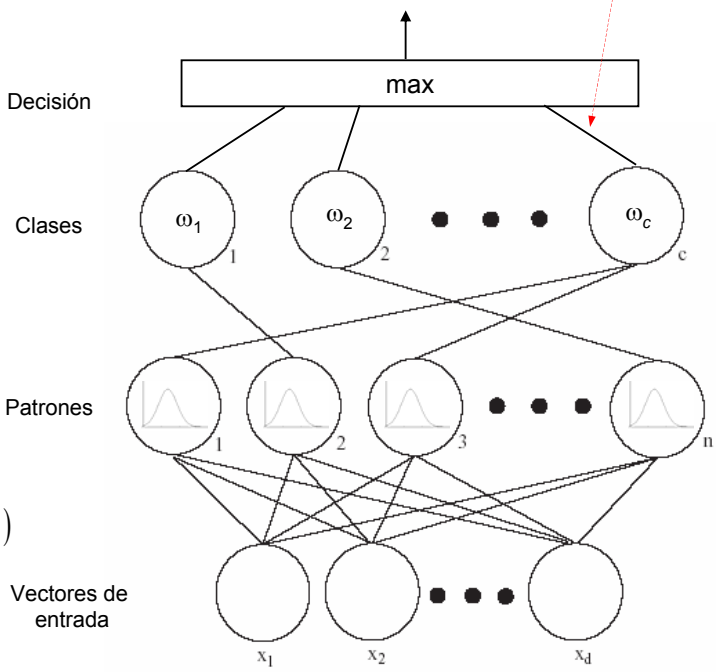
3. Cada bloque-clase suma sus entradas y genera el discriminante $g_j(\mathbf{x})$

2. Los bloques-patron calculan un factor de actividad:

$$\varphi\left(\frac{\mathbf{x}-\mathbf{w}_i}{h_n}\right) \propto \exp\left(-(\mathbf{x}-\mathbf{w}_i)^T(\mathbf{x}-\mathbf{w}_i)/2\sigma^2\right) =$$
$$= \exp\left(-\left(\mathbf{x}^T\mathbf{x}+\mathbf{w}_i^T\mathbf{w}_i-2\mathbf{x}^T\mathbf{w}_i\right)/2\sigma^2\right) =$$
$$= \left\{\mathbf{x}^T\mathbf{x}=\mathbf{w}_i^T\mathbf{w}_i=1\right\} = \exp\left(\left(\mathbf{x}^T\mathbf{w}_i-1\right)/2\sigma^2\right)$$

1. El vector a clasificar es $\mathbf{x}$. Se normaliza.

Podemos ponderar con la probabilidad a priori de cada clase si es conocida



4.4 ESTIMACIÓN K-NEAREST NEIGHBORS

En lugar de buscar la mejor ventana (en forma y tamaño) vamos a hacer que el volumen de la célula en la que estimamos las probabilidades crezca o decrezca en función de los datos de entrenamiento:

Para estimar $f(\mathbf{x})$ haremos crecer una celda de volumen $V_{R,x}$ alrededor de $\mathbf{x}$ hasta que capture k_n datos: los k_n – nearest neighbors.

$$f_n(\mathbf{x}) = \frac{k_n / n}{V_{R,\mathbf{x}}}$$

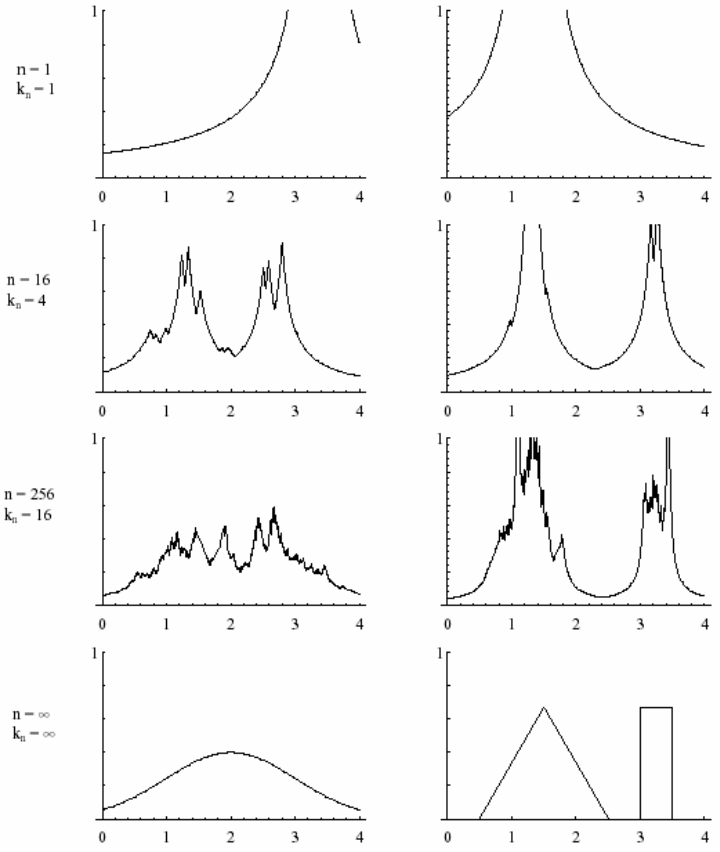
Para que converja hay que asegurar que: $\lim_{n\to\infty} \frac{k_n}{n} = 0$

Ejemplo 4:

$$k_n \le \sqrt{n} \qquad\qquad k_n \le \ln(n)$$

Ejemplo 5:

Estimaciones k-nearest-neighbors. Comparar las obtenidas en la diapositiva 12



Las probabilidades a posteriori $\Pr(\omega_i|\mathbf{x})$ pueden calcularse como:

$$\Pr(\omega_i \mid \mathbf{x}) = \frac{f_{\mathbf{x}}(\mathbf{x} \mid \omega_i)\Pr(\omega_i)}{\sum_{i=1}^{c} f_{\mathbf{x}}(\mathbf{x} \mid \omega_i)\Pr(\omega_i)} = \frac{f_{\mathbf{x}}(\mathbf{x}, \omega_i)}{\sum_{i=1}^{c} f_{\mathbf{x}}(\mathbf{x}, \omega_i)} = \frac{k_i / nV_{R,n}}{k / nV_{R,n}} = \frac{k_i}{k}$$

$$f_{\mathbf{x}}(\mathbf{x}, \omega_i) = f_{\mathbf{x}}(\mathbf{x} \mid \omega_i)\Pr(\omega_i) \simeq \frac{k_i / n_i}{V_{R,n}} \frac{n_i}{n} = \frac{k_i}{nV_{R,n}}$$

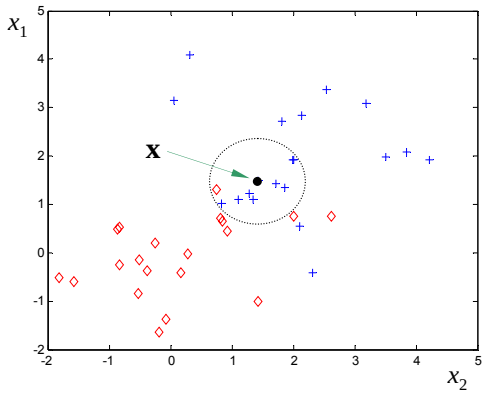
Fracción de datos que pertenecen a la clase ω_i de entre los k datos vecinos de $\mathbf{x}$

4.5 REGLA K-NEAREST NEIGHBORS

Regla de clasificación. A partir de la ecuación anterior podemos clasificar un vector de datos $\mathbf{x}$ con la siguiente regla (que es óptima si n es muy grande):

Seleccionamos la categoría que tiene más datos de entrenamiento representados en la ventana alrededor de $\mathbf{x}$.

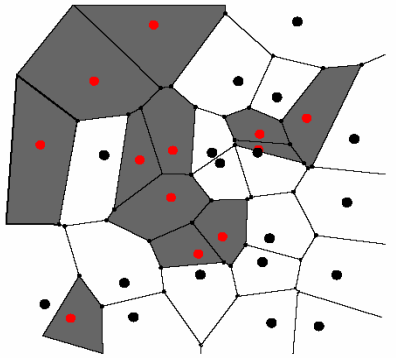
Ejemplo 6:



K=8
La clase seleccionada para el vector $\mathbf{x}$ es '+'

• Parzen: Seleccionamos la clase con más datos (ponderados) en la ventana centrada en $\mathbf{x}_i$ $\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)$

• K-nearest-neighbors: Buscar la ventana que nos proporcione *k* vecinos alrededor de $\mathbf{x}_i$. La clase con más representantes en esa ventana es la escogida.



Regiones de decisión para 1-nearest

Se pueden obtener prestaciones razonables si se decide la clase únicamente a partir de la clase a la que esté asociado el vector de entrenamiento $\mathbf{x}$ más próximo ("1-nearest").

4.6 DISTANCIA

Propiedades:

No-negatividad	$D(\mathbf{x},\mathbf{y}) \geq 0$
Reflexividad	$D(\mathbf{x},\mathbf{y}) = 0$ si y solo si $\mathbf{x} = \mathbf{y}$
Simetria	$D(\mathbf{x},\mathbf{y}) = D(\mathbf{y},\mathbf{x})$
Desigualdad triangular	$D(\mathbf{x},\mathbf{y}) + D(\mathbf{y},\mathbf{t}) \geq D(\mathbf{x},\mathbf{t})$

• La distancia a escoger depende de cada problema concreto.

Ejemplo 7:

$$D(\mathbf{x},\mathbf{y}) = \left(\sum_{i=1}^{d} |x_i - y_i|^p\right)^{1/p}$$

$p = 1$

$p = 2$

$p = \infty$

4.5 CONCLUSIONES

Dos formas no paramétricas de clasificar:

1. Parzen: Cálculo de una función de densidad y posterior clasificacion
2. k-nearest neighbors: Determinación directa la clase a la que pertenece.

El uso de 1-nearest neighbor da una $P(\varepsilon)$ (cuando se tienen muchas muestras de entrenamiento) igual a dos veces la de Bayes, con una complejidad de cálculo extremadamente baja.