Introducción

- **Recordamos:**
  - La forma óptima de realizar el proceso de clasificación consiste en la utilización del clasificador bayesiano:

    Elegir ω_i si $p(\boldsymbol{x}|\omega_i)\,P(\omega_i) > p(\boldsymbol{x}|\omega_j)\,P(\omega_i)\ \forall j \neq i$

  - Para utilizarlo, sin embargo, es necesario conocer la forma de la distribución condicional en cada clase $p(\boldsymbol{x}|\omega_i)$ y la probabilidad a priori $P(\omega_i)$.
- **Problema:**
  - En la práctica las distribuciones de probabilidad no se conocen.
- **Solución (aproximación generativa):**
  - Estimar todas las distribuciones de probabilidad mediante un conjunto de entrenamiento *H*. De esta forma obtenemos el modelo probabilístico mediante el cual se generó el conjunto de entrenamiento

Aproximación Generativa

- **Objetivo:**
  - Estimar $p(\boldsymbol{x}|w_i)$, $P(w_i)$, necesarios para aplicar el modelo de Decisión Bayesiano.
- **Información disponible:**
  - Un conjunto de muestras de entrenamiento H representativas de las distintas clases, correctamente "etiquetadas" con su clase de pertenencia.
  - Esto es, $H = H_1 \cup H_2 \cup ... \cup H_c$, donde cada H_i tiene las muestras de la clase w_i
- **Asumiremos:**
  - Que las muestras de cada clase no proporcionan información acerca de la otra clase.
  - Las muestras en cada clase son independientes
- **Esto permite:**
  - Estimar $p(\boldsymbol{x}|w_i)$, $P(w_i)$ únicamente a partir de las muestras en H_i.
  - Tenemos que resolver el problema de estimación para cada clase
- **Problema:**
  - La estimación de $P(w_i)$ es simple, sin embargo la estimación de $p(\boldsymbol{x}|w_i)$ es un problema complejo

Estrategias de Estimación

- **Estimación Paramétrica**
  - Se basa en suponer que la forma de $p(\boldsymbol{x}|w_i)$ es conocida (gausiana, beta, etc...) y depende de un conjunto de parámetros $\boldsymbol{\theta}_i$.
    - Principal Ventaja: Los métodos de estimación son más simples y precisos
    - Principal Desventaja: Es necesario conocer la forma de la distribución. Los métodos suelen ser sensibles a errores en dicha forma.

    Métodos más importantes:
    - Estimación por Máxima Verosimilitud.
    - Estimación máximo *a posteriori*
    - Estimación Bayesiana.
- **Estimación no Paramétrica.**
  - No se realiza ninguna asunción acerca de la forma de $p(\boldsymbol{x}|w_i)$
    - Principal Ventaja: Métodos robustos
    - Principal Desventaja: Métodos complejos y que requieren un gran número de muestras para una estimación precisa.
  - Métodos más importantes
    - Ventanas de Parzen.
    - Vecinos más próximos.

Estimación Paramérica (1)

- **Métodos paramétricos**
  - Se asume que la forma de las funciones de densidad condicionales son conocidas y dependen de un conjunto de parámetros $\boldsymbol{\theta}_i$. Escribiremos esta dependencia como $p(\boldsymbol{x}|w_i, \boldsymbol{\theta}_i)$.

    Por ejemplo para una normal multidimensional tendremos que $\boldsymbol{\theta}_i=\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$
  - Sin embargo, se desconoce el valor verdadero del conjunto de parámetros que la determinan completamente. Este verdadero valor se estima a partir de un conjunto de entrenamiento mediante un estimador.
- **Es importante recordar que:**
  - El valor del estimador (estimación) depende del conjunto de entrenamiento y distintos conjuntos de entrenamiento proporcionan distintas estimaciones.
  - La estimación no tiene por qué coincidir con el verdadero valor del parámetro.

Estimación Paramétrica (2)

- **Simplificacion:**
  - Las muestras de la clase w_i sólo dan información acerca del parámetro de dicha clase $\boldsymbol{\theta}_i$.
  - Esto permite trabajar con cada clase por separado y obtener *c* problemas de la forma:

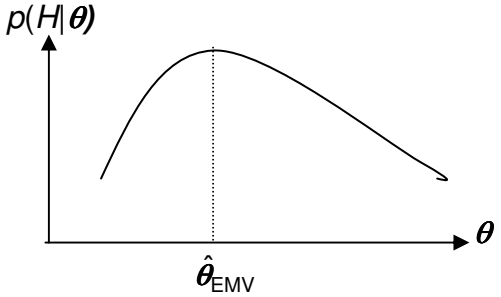
    "Utilizar un conjunto de muestras H_i tomadas de forma independiente de $p(\boldsymbol{x} \mid w_i, \boldsymbol{\theta}_i)$ para estimar $\boldsymbol{\theta}_i$"

- **Notación:**
  - Eliminaremos de la notación la dependencia de la clase para simplificar la escritura y escribiremos $p(\boldsymbol{x} \mid \boldsymbol{\theta})$ en vez de $p(\boldsymbol{x} \mid w_i, \boldsymbol{\theta}_i)$ y H en lugar de H_i.
  - No obstante debemos recordar siempre que estamos utilizando las muestras de una única clase y estimado los parámetros para esa clase.
  - Por tanto para completar el clasificador debemos tener resuelto el problema de estimación para cada clase por separado.

EMV: Método

- **Idea:**
  - Encontrar los valores del conjunto de parámetros que hace máxima la verosimilitud del conjunto de entrenamiento
- **Obtención de la máxima verosimilitud**
  - Si $H=\{\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_n\}$ son muestras generadas de forma independiente de la función de densidad $p(\boldsymbol{x}|\boldsymbol{\theta})$ entonces
    - 1.- Calcular la función de verosimilitud de todas las muestras:

$$L = p(H \mid \boldsymbol{\theta}) = \prod_{k=1}^{n} p(\boldsymbol{x}_k \mid \boldsymbol{\theta})$$



    - 2.- Obtener el valor $\hat{\boldsymbol{\theta}}_{\text{EMV}}$ de $\boldsymbol{\theta}$ que hace máxima la función de verosimilitud L.
      Para ello puede resolverse la ecuación:
      $\nabla_{\boldsymbol{\theta}}\, p(H \mid \boldsymbol{\theta}) = \mathbf{0}$, o de forma equivalente:

$$\nabla_{\boldsymbol{\theta}} \ln(p(H \mid \boldsymbol{\theta})) = \mathbf{0}$$

- **Ejemplo:**
  - Estimar la media $\boldsymbol{\mu}$, y la matriz $\boldsymbol{\Sigma}$ de una distribución normal por EMV, a partir de un conjunto $H=\{\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_n\}$.

$$\hat{\boldsymbol{\mu}}_{\text{EMV}} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{x}_k, \qquad \hat{\boldsymbol{\Sigma}}_{\text{EMV}} = \frac{1}{n}\sum_{k=1}^{n} (\boldsymbol{x}_k - \hat{\boldsymbol{\mu}}_{\text{EMV}})(\boldsymbol{x}_k - \hat{\boldsymbol{\mu}}_{\text{EMV}})^{\mathsf{T}}$$

Fernando Pérez Nava

EMV: Propiedades

- **El EMV tiene propiedades muy deseables:**
  - Es asintóticamente insesgado: $\lim_{n\to\infty} \mathrm{E}(\hat{\boldsymbol{\theta}}_{\mathrm{EMV}}) = \boldsymbol{\theta}$
    - Significa que la media sobre los valores de los posibles conjuntos de entrenamiento da el verdadero valor del parámetro cuando el número de muestras del conjunto de entrenamiento tiende a infinito
  - Es asintóticamente consistente: $\lim_{n\to\infty} \mathrm{P}\left(\left\|\hat{\boldsymbol{\theta}}_{\mathrm{EMV}} - \boldsymbol{\theta}\right\| \le \varepsilon\right) = 1$

    Significa que cuando el número de muestras del conjunto de entrenamiento tiende a infinito el valor del estimador estará arbitrariamente cerca del verdadero valor del parámetro.
  - Es asintóticamente eficiente:
    - Significa que alcanza la menor varianza que cualquier estimador puede tener cuando el número de muestras del conjunto de entrenamiento tiende a infinito

Fernando Pérez Nava

Estimación de las probabilidades a priori

- **La estimación mediante EMV de las probabilidades a priori P(w_i) es simple y se calcula mediante:**
  - $\hat{P}(w_i)=|H_i| / |H|$, $|.|$=número de elementos

    Esto es, el cociente entre el número de elementos de la clase w_i en el conjunto de entrenamiento y el número total de elementos del conjunto de entrenamiento

- **Un inciso...**
  - De la misma forma puede comprobarse que la decisión tomada utilizando el conjunto de entrenamiento para el ejemplo de los ródalos y salmones del tema anterior está basada en la estimación por máxima verosimilitud de las correspondientes funciones de distribución para cada clase.

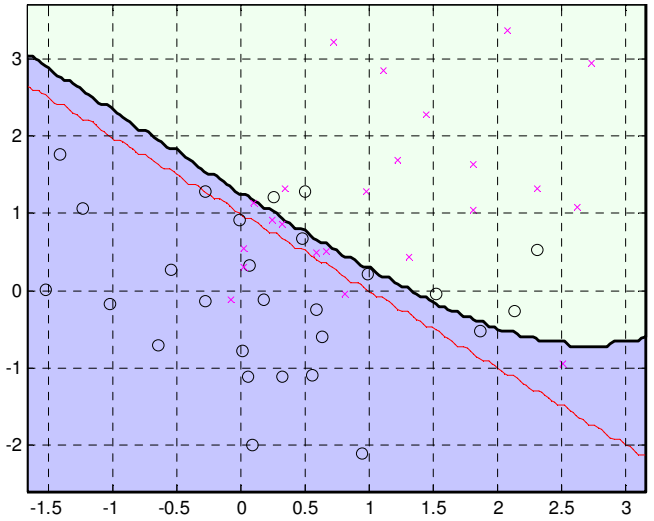
Clasificación tras estimación por EMV: Ejemplo

- **Distribuciones verdaderas:**
  - $p(\boldsymbol{x} \mid w_1, \boldsymbol{\theta}_1) \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$, $p(\boldsymbol{x} \mid w_2, \boldsymbol{\theta}_2) \sim N\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$
  - $P(w_1)=0.5$, $P(w_2)=0.5$
- **Clasificación:**
  - Conjunto de testeo:
    - 50 muestras por clase
  - Conjunto de entrenamiento:
    - 50 muestras por clase
  - Estimación:

$$\hat{p}(\boldsymbol{x} \mid w_1) \sim N\left(\begin{pmatrix}-0.45\\0.32\end{pmatrix}, \begin{pmatrix}0.02 & -0.09\\-0.09 & 0.53\end{pmatrix}\right)$$

$$\hat{p}(\boldsymbol{x} \mid w_2) \sim N\left(\begin{pmatrix}0.52\\0.16\end{pmatrix}, \begin{pmatrix}2.32 & -0.73\\-0.73 & 0.23\end{pmatrix}\right)$$

  - Error de clasificación estimado:
    - 0.24
  - Error bayesiano:
    - 0.23



Ejemplo de clasificación tras estimación mediante EMV
Circulos: muestras de la clase 1
Aspas: muestras de la clase 2
Linea negra: Frontera de decisión a partir de la estimación
Linea roja: Frontera de decisión bayesiana

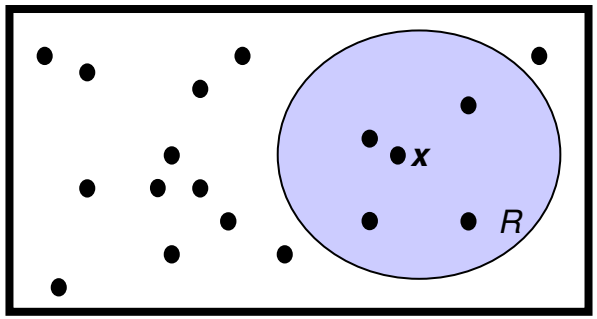
Métodos no Paramétricos (M.n.P.)

- **Métodos no Paramétricos:**
  - Es un conjunto de métodos que no necesita información acerca de la forma de las funciones de densidad condicionales $p(\boldsymbol{x} \mid w_i)$
- **Simplificación:**
  - Se asume que los elementos de H_i solo dan información sobre dicha clase. Esto permite resolver c problemas independientes
- **Notación:**
  - Eliminaremos de la notación la dependencia de la clase para simplificar la escritura y escribiremos $p(\boldsymbol{x})$ en lugar de $p(\boldsymbol{x} \mid w_i)$ y H en lugar de H_i
  - No obstante debemos recordar siempre que estamos utilizando las muestras de una única clase y por tanto para completar el clasificador debemos tener resuelto el problema de estimación para cada clase por separado.
- **Algunos Procedimientos:**
  - Ventanas de Parzen
    - Se estima la función de densidad $p(\boldsymbol{x})$ examinando el conjunto de entrenamiento H en un entorno de $\boldsymbol{x}$ que cuya forma no depende de H
  - k- Vecinos más próximos
    - Se estima la función de densidad $p(\boldsymbol{x})$ examinando el conjunto de entrenamiento H en un entorno de $\boldsymbol{x}$ cuya forma depende de H

M.n.P.: Aspectos Generales

- **Objetivo:** Estimar $p(\boldsymbol{x})$ a partir de H
- **Metodología:**
  - Diseñar una región R del espacio de características, que contiene a $\boldsymbol{x}$ y lo suficientemente pequeña para asumir que la función de densidad $p(\boldsymbol{x})$ es aproximadamente constante.
  - A partir de las n muestras independientes presentes en H, generadas de acuerdo a la función de densidad $p(\boldsymbol{x})$, y siendo k el número de muestras que caen en R estimar:

$$\hat{p}(\boldsymbol{x}) = \frac{k/n}{V}, \qquad V = \int_R \mathrm{d}\boldsymbol{x}$$



k=5
n=18
V=área de R

$$\hat{p}(\boldsymbol{x}) = \frac{5/18}{V}$$

Ejemplo de Estimación de $p(\boldsymbol{x})$

Convergencia de la Estimación

- **Convergencia**
  - Una condición deseable es la convergencia de la estimación de $p(\boldsymbol{x})$ a su verdadero valor cuando el tamaño del conjunto de entrenamiento tiende a infinito.
- **Condiciones de Convergencia**
  - Para expresar la dependencia de k y V del tamaño del conjunto de entrenamiento n escribiremos:

$$\hat{p}_n(\boldsymbol{x}) = \frac{k_n / n}{V_n}$$

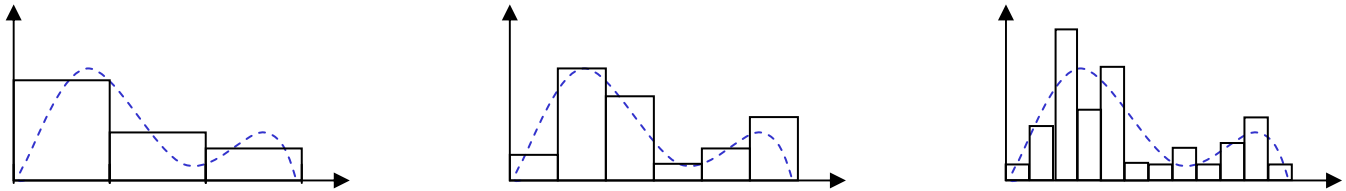
  - Con el siguiente resultado se asegura la convergencia de dicha estimación:

$$\lim_{n\to\infty} V_n = 0,\ \lim_{n\to\infty} k_n = \infty,\ \lim_{n\to\infty} k_n / n = 0 \Rightarrow \lim_{n\to\infty} \hat{p}_n(\boldsymbol{x}) = p(\boldsymbol{x})$$

  - Esto es, cuando el tamaño del conjunto de entrenamiento tiende a infinito tiene que cumplirse:
    1. El volumen de la región V_n debe tender a 0
    2. El número de puntos en la región debe tender a ∞
    3. La frecuencia relativa de los puntos en la región debe tender a 0

Ventanas de Parzen:Preliminares

- **Histogramas**
  - La forma más simple de estimación de funciones de densidad es mediante la creación de un histograma de frecuencias relativas.
  - En un histograma unidimensional se puede elegir el número de celdas *M* y el punto de comienzo de la división en intervalos



Estimación de funciones de densidad mediante histogramas para distintos valores de *M*

- **Desventajas**
  - Cuando se tienen *d* características no es factible construir el histograma debido a que el número de celdas es exponencial (M^d) en el número de características *d.*

Ventanas de Parzen:Introducción

- **Idea inicial:**
  - Fijar un tamaño de región, construirla únicamente alrededor del punto $\boldsymbol{x}$ para el que se desea estimar su probabilidad y aplicar la fórmula de los métodos no paramétricos:

$$\hat{p}(\boldsymbol{x}) = \frac{k/n}{V}$$

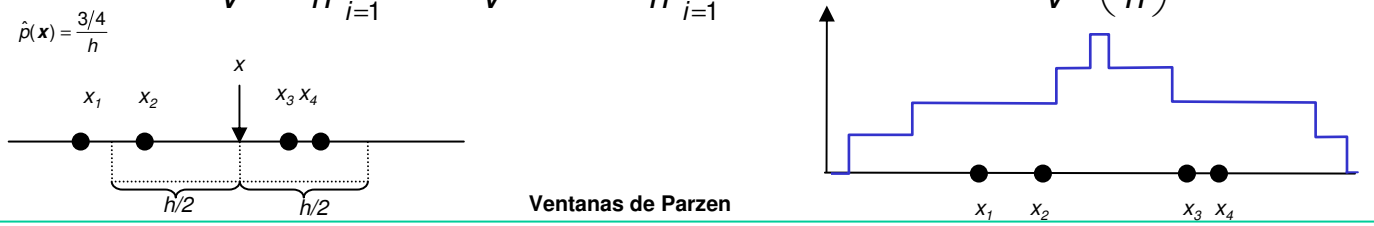
- **Vamos a formalizarlo:**
  - Caso unidimensional
    - La celda es un intervalo centrado en $\boldsymbol{x}$ de longitud h
    - Hallaremos k de una forma un tanto especial:

Primero definimos la función $\phi(t) = \begin{cases} 1 & |t| \le 1/2 \\ 0 & \text{otro caso} \end{cases}$

Entonces $k = \sum_{i=1}^{n} \phi((x - x_i)/h)$

- Finalmente la estimación es:

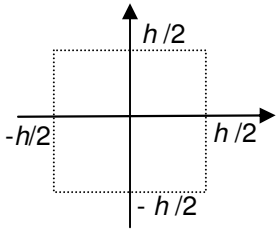
$$\hat{p}(\boldsymbol{x}) = \frac{k/n}{V} = \frac{1}{n}\sum_{i=1}^{n} \frac{\phi((\boldsymbol{x} - \boldsymbol{x}_i)/h)}{V} = \frac{1}{n}\sum_{i=1}^{n} \delta(\boldsymbol{x} - \boldsymbol{x}_i), \quad \delta(\boldsymbol{x}) = \frac{1}{V}\phi\left(\frac{\boldsymbol{x}}{h}\right), \quad V = h$$



Fernando Pérez Nava

Ventanas de Parzen: Caso Multidimensional

- **Caso multidimensional**
  - La celda es un hipercubo centrado en $\boldsymbol{x}$ y la longitud de cada lado es h



Hipercubo en 2-D

  - De nuevo hallamos k de una forma especial:
    Primero definimos la función $\phi(\boldsymbol{x}) = \phi(x_1, x_2, \ldots, x_d) = \begin{cases} 1 & |\, x_i \,| \leq 1/2 \\ 0 & \text{otro caso} \end{cases}$
    igual a 0 salvo dentro del hipercubo centrado en **0** y de longitud de lado 1/2 donde vale 1.

    Entonces $k = \sum_{i=1}^{n} \phi((\boldsymbol{x} - \boldsymbol{x}_i)/h)$

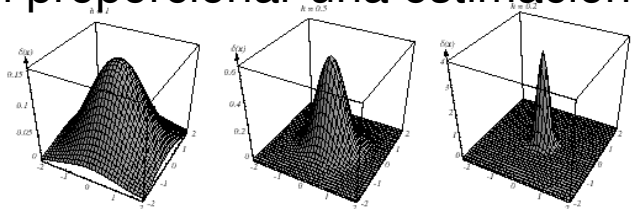
  - Finalmente la estimación es:

$$\hat{p}(\boldsymbol{x}) = \frac{k/n}{V} = \frac{1}{n}\sum_{i=1}^{n} \frac{\phi((\boldsymbol{x} - \boldsymbol{x}_i)/h)}{V} = \frac{1}{n}\sum_{i=1}^{n} \delta(\boldsymbol{x} - \boldsymbol{x}_i), \quad \delta(\boldsymbol{x}) = \frac{1}{V}\phi\left(\frac{\boldsymbol{x}}{h}\right), \quad V = h^d$$

Ventanas de Parzen:Funciones núcleo

- **Problema:**
  - La estimación $\hat{p}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\delta(\boldsymbol{x}-\boldsymbol{x}_i)$ genera funciones de densidad

    discontinuas (pues las funciones δ son discontinuas). Generalmente se suele trabajar con funciones de densidad continuas
- **Solución**
  - Generalizar la noción de histograma variando la función núcleo δ utilizando por ejemplo una gausiana: $\delta_G(\mathbf{x})=1/(2\pi h^2)^{d/2}$ exp((-1/2) $\mathbf{x}'\mathbf{x}$) que suelen proporcionar una estimación más suave.



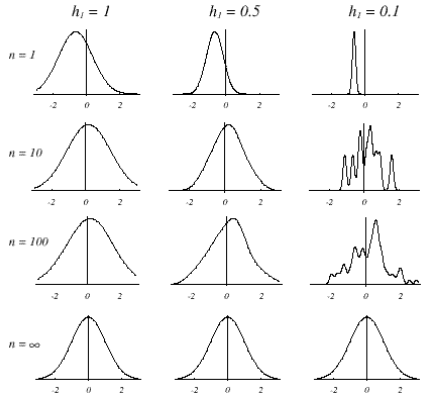
Funciones núcleo gausianas para distintos valores de h

Estimación de Parzen mediante las funciones núcleo gausianas para 5 muestras y distintos valores de h

Gráficos de: Richard O. Duda, Peter E. Hart, and David G. Stork, Pattern Classification. Copyright (c) 2001 por John Wiley & Sons, Inc.

Ventanas de Parzen: La elección de *h*

- **Todavía mas problemas...**
  - La estimación depende de *h.* Si *h* es muy grande la estimación es muy suave. Si por el contrario *h* es muy pequeño la estimación suele tener variaciones bruscas inaceptables (se produce sobreajuste).



Estimación de Parzen de una función de distribución gausiana para distintos valores de h y n

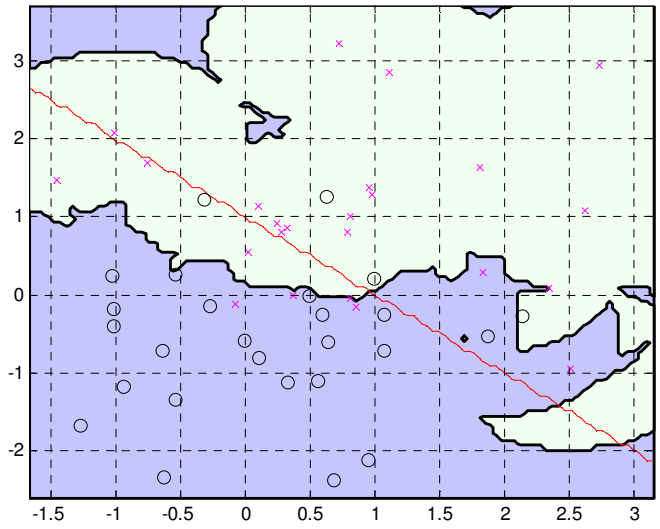
- **Una solución:**
  - Dividir el conjunto de entrenamiento en dos partes: uno para testeo y otro para validación. Utilizar el conjunto de entrenamiento para definir distintas estimaciones en función de *h.* Posteriormente elegir aquel valor de *h* para el que la probabilidad del conjunto de validación sea máxima.

Fernando Pérez Nava

Clasificación por Ventanas de Parzen: Ejemplo

- **Distribuciones verdaderas:**
  - $p(\boldsymbol{x} \mid w_1, \boldsymbol{\theta}_1) \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$, $p(\boldsymbol{x} \mid w_2, \boldsymbol{\theta}_2) \sim N\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$
  - $P(w_1)=0.5$, $P(w_2)=0.5$

- **Clasificación:**
  - Conjunto de testeo:
    - 50 muestras por clase
  - Conjunto de entrenamiento:
    - 50 muestras por clase
  - Valor óptimo calculado para h:
    - 2.154
  - Error de clasificación estimado:
    - 0.32
  - Error bayesiano:
    - 0.23



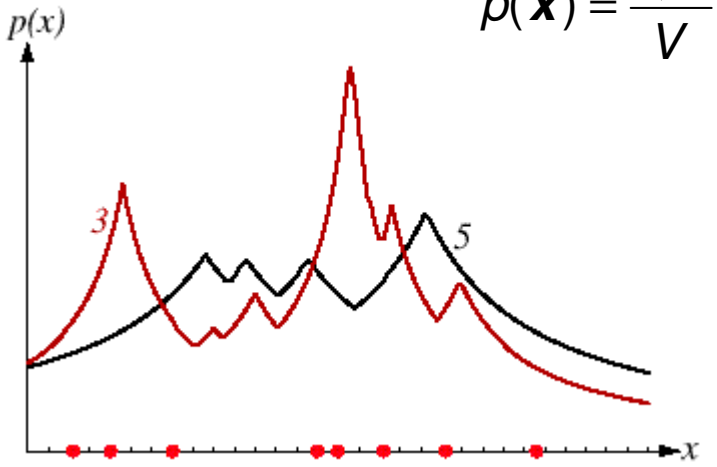
Ejemplo de clasificación tras estimación mediante Parzen
Circulos: muestras de la clase 1
Aspas: muestras de la clase 2
Linea negra: Frontera de decisión a partir de la estimación
Linea roja: Frontera de decisión bayesiana

Fernando Perez Nava

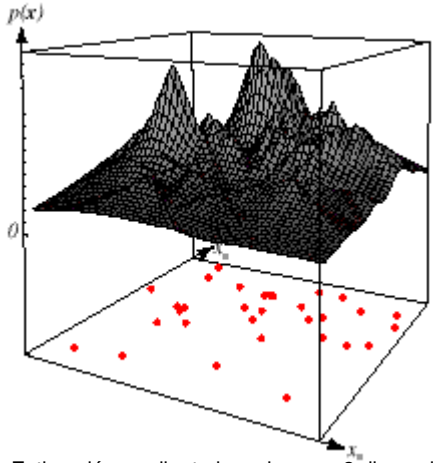
Estimación por *k*- vecinos más próximos

- **Idea:**
  - Parece que en zonas con pocas muestras deberíamos hacer la región grande mientras que en zonas con pocas muestras la podemos hacer pequeña. Una idea sería entonces fijar el número de muestras que queremos en la región alrededor del punto ***x*** para el que se desea estimar su probabilidad y aplicar la fórmula de los métodos no paramétricos:

$$\hat{p}(\boldsymbol{x}) = \frac{k/n}{V}$$



Estimación mediante k-vecinos (k=3,5) y 8 muestras



Estimación mediante k-vecinos en 2 dimensiones

Fernando Pérez Nava

Estimación directa de $p(w_i|\boldsymbol{x})$

- **Recordemos:**
  - El clasificador óptimo bayesiano se puede construir hallando la clase para la que es máxima la probabilidad a posteriori: $p(w_i|\boldsymbol{x})$
- **Entonces:**
  - Supongamos que el conjunto de datos H contiene n_i muestras en la clase w_i y n muestras en total.
  - Supongamos que fijamos una región R de volumen V para todas las clases
  - Como sabemos, debemos resolver un problema de estimación por clase. Para la clase w_i la estimación será:

$$\hat{p}(\boldsymbol{x} \mid w_i) = \frac{k_i / n_i}{V}$$

  - Entonces si utilizamos $\hat{p}(w_i) = \frac{n_i}{n}$ tendremos $\hat{p}(w_i \mid \boldsymbol{x}) = \frac{k_i}{k}$
  - La regla es simple: "Seleccionar la clase con mayor número de elementos en la región R".

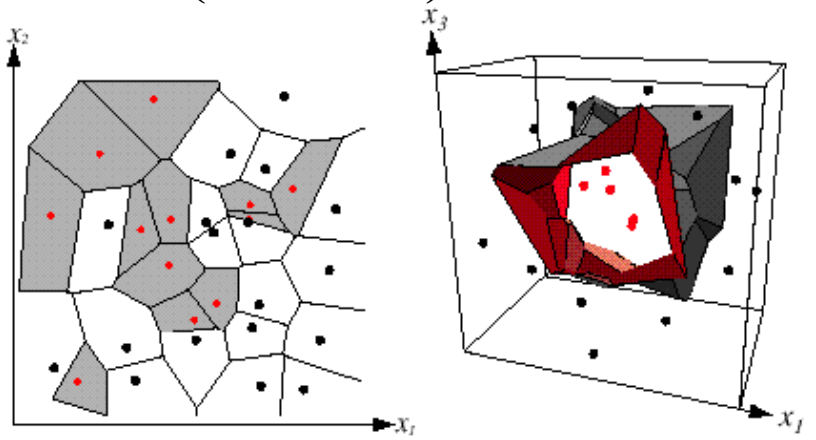
    La región R puede definirse mediante el esquema de las ventanas de Parzen o los k-vecinos. Este último esquema lleva a la clasificación por vecinos más cercanos.

Fernando Pérez Nava

Clasificación por el vecino más próximo

- **Clasificación (1-vecino más próximo)**
  - Dado el conjunto *H* de muestras se clasifica ***x*** como perteneciente a la clase de su vecino más próximo en *H*.
- **Probabilidad de Error**
  - Si P^* es la probabilidad de error bayesiano (mínima), P la de la regla 1-NN, c el número de clases y n el número de muestras en *H:*

$$P^* \leq P \leq P^*\left(2 - \frac{c}{c-1}P^*\right) < 2P^*, \quad \text{para n} \to \infty$$



Clasificación mediante el vecino más próximo en 1 y 2 dimensiones

Fernando Pérez Nava

Clasificación por *k*-vecinos más próximos

- **Clasificación (k-vecinos más próximos)**
  - Dado el conjunto *H* de muestras se clasifica ***x*** como perteneciente a la clase mayoritaria entre sus *k* vecinos más próximos de *H*.
- **Probabilidad de Error**
  - Se aproxima a la Probabilidad de Error Bayesiano, cuando tanto *k*, como el número de muestras *n*, tienden a infinito.
  - La probabilidad de error se puede acotar:

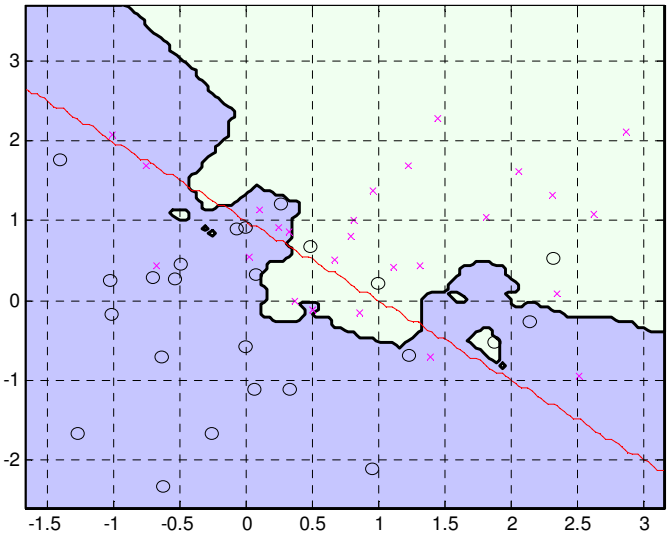
$$P^* \leq P_{\text{kNN}} \leq P^* + \frac{1}{\sqrt{k\,e}}$$

- **¿Qué valor elegir para *k*?**
  - Se suele dividir el conjunto de entrenamiento en dos partes: uno para testeo y otro para validación. Utilizar el conjunto de entrenamiento para construir el clasificador para distintos valores de *k*. Posteriormente elegir aquel valor de *k* para el que la probabilidad de error sea mínima sobre el conjunto de validación

Clasificación por *k*-vecinos: Ejemplo

- **Distribuciones verdaderas:**
  - $p(\boldsymbol{x} \mid w_1, \boldsymbol{\theta}_1) \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$, $p(\boldsymbol{x} \mid w_2, \boldsymbol{\theta}_2) \sim N\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$
  - $P(w_1)=0.5$, $P(w_2)=0.5$

- **Clasificación:**
  - Conjunto de testeo:
    - 50 muestras por clase
  - Conjunto de entrenamiento:
    - 50 muestras por clase
  - Valor óptimo calculado para *k*:
    - 8
  - Error de clasificación estimado:
    - 0.28
  - Error bayesiano:
    - 0.23



Ejemplo de clasificación por k-vecinos
Circulos: muestras de la clase 1
Aspas: muestras de la clase 2
Linea negra: Frontera de decisión a partir de la estimación
Linea roja: Frontera de decisión bayesiana

Resumiendo...

- **En este tema hemos visto métodos para estimar la estructura de probabilidad necesaria para aplicar la regla de clasificación bayesiana.**
- **Las buenas noticias...**
  - Cuando la forma de la función de densidad condicional $p(\boldsymbol{x} \mid w_i)$ es conocida y depende únicamente de un vector de parámetros $\boldsymbol{\theta}$ (caso paramétrico) se dispone de estimadores con buenas propiedades. Además el coste computacional del clasificador depende del número de características.
  - Aun en el caso de que la forma de la función de densidad condicional $p(\boldsymbol{x} \mid w_i)$ sea desconocida (caso no paramétrico) se dispone también de métodos de estimación. Algunos muy simples como los k-vecinos.
  - Además obtenemos un modelo probabilístico de la forma de generación de los datos
- **Las malas noticias...**
  - La forma de $p(\boldsymbol{x} \mid w_i)$ raramente es conocida en problemas complejos.

    Cuando la forma de $p(\boldsymbol{x} \mid w_i)$ es errónea el clasificador suele ofrecer malos resultados (no es robusto frente a errores en forma de la distribución).
  - Los métodos no paramétricos necesitan un gran número de muestras para ofrecer resultados precisos. Además el coste computacional del clasificador depende del número de muestras.
  - Los métodos no paramétricos son tan flexibles que pueden sufrir de sobreajuste. Es necesario fijar determinados parámetros para que esto no ocurra.