Minería de Datos (MD) estadística

- **¿Por qué una aproximación estadística en la MD?**
  - La utilización de características para representar una entidad provoca una pérdida de información. Esto implica que los valores de las características tienen asociado un determinado nivel de certeza.
- **La MD Estadística se basa en:**
  - Considerar la entidad a estudiar como un conjunto de d características numéricas que se interpretan como un vector *d* dimensional
  - Asumir que la certeza de que el vector represente una determinada entidad viene dada a través de una distribución de probabilidad asociada a las características
- **Es la aproximación más extendida debido a:**
  - La fundamentación de la aproximación en una teoría matemática sólida como la teoría de la probabilidad.
  - Su mayor presencia temporal (trabajos desde finales de los años 30).
  - Su mayor aplicabilidad:
    - Clasificación con valores de las características perdidas
    - Toma de decisiones que minimizan la pérdida esperada

Clasificación:Teoría de la Decisión Bayesiana (TDB)

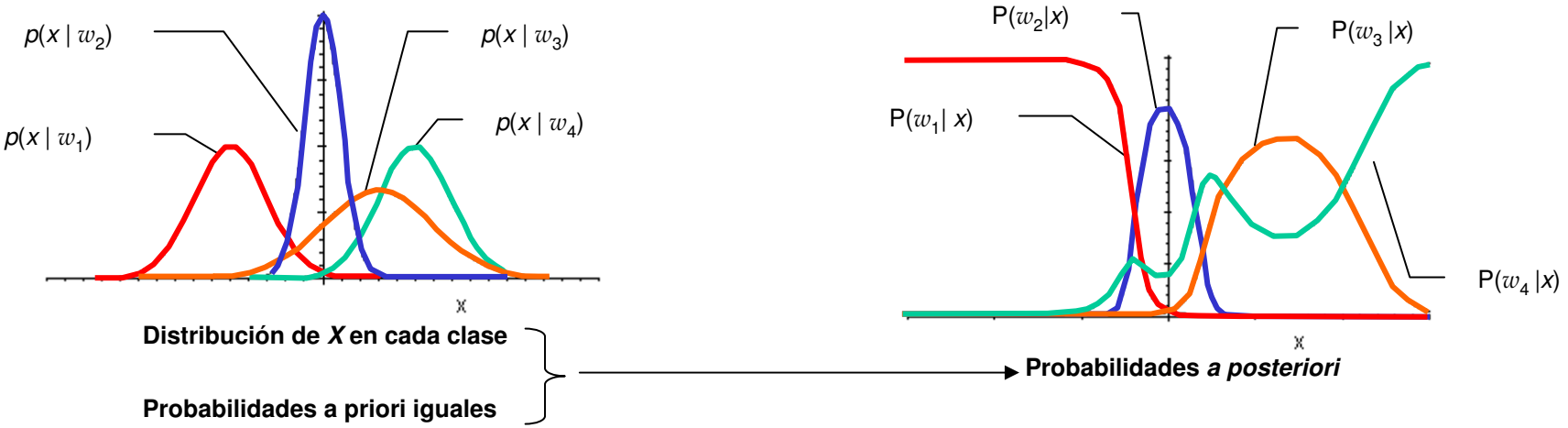
- **La TDB proporciona un marco teórico para tomar decisiones en situaciones de incertidumbre.**
- **En nuestro caso la decisión será la clasificación de una entidad en una determinada clase**
- **La TDB proporciona el clasificador óptimo (clasificador bayesiano) para un conjunto de características dadas**
  - En el marco de la TDB un clasificador es óptimo si produce la mínima probabilidad de error (o el riesgo de la clasificación).
  - La TDB necesita que todas las distribuciones de probabilidad de las características $p(x \mid \omega_j)$ en cada clase sean conocidas.En la práctica esto nunca ocurre, por lo que es necesario inferir (de las muestras) la forma de las distribuciones de probabilidad. También es necesario inferir las probabilidades a priori $P(\omega_j)$

TDB: Enfoque formal (1)

- **Información disponible:**
  - Clases: w_i, i=1...c
  - Características : $\boldsymbol{X}$ variable aleatoria multidimensional.
  - Probabilidades: P(w_i), $p(\boldsymbol{x} \mid w_i)$, i=1...c
  - Mediante la Regla de Bayes:

$$P(w_i \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid w_i)P(w_i)}{p(\boldsymbol{x})}, i = 1...c \quad \text{con} \quad p(\boldsymbol{x}) = \sum_{i=1}^{c} p(\boldsymbol{x} \mid w_i)P(w_i)$$

- **Ejemplo:**

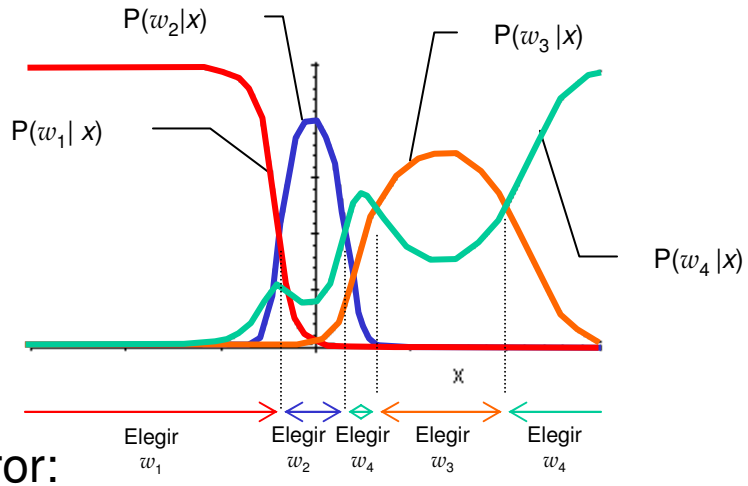


TDB: Enfoque formal (2)

- **Probabilidad de error "Elegir w_i"**

$$P(Error \mid \boldsymbol{x}) = \sum_{k=1,k\neq i}^{c} P(w_k \mid \boldsymbol{x}) = 1 - P(w_i \mid \boldsymbol{x})$$

- **Regla de decisión Bayesiana (óptima):**
  - Elegir w_i si $P(w_i \mid \boldsymbol{x}) \geq P(w_j \mid \boldsymbol{x})\ \forall i \neq j \Leftrightarrow$
    $p(\boldsymbol{x} \mid w_i)P(w_i) \geq p(\boldsymbol{x} \mid w_j)P(w_j)\ \forall i \neq j$

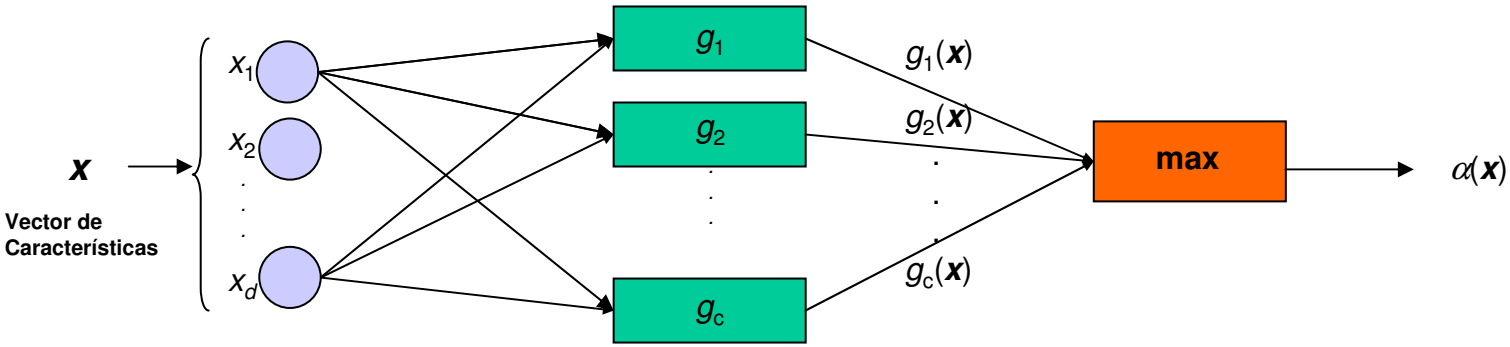


- **Propiedad:**
  - Hace mínima la probabilidad de error:

$$P(Error) = \int P(Error \mid \boldsymbol{x})p(\boldsymbol{x})\,d\boldsymbol{x}$$

Clasificadores y su Representación

- **Definición formal de Clasificador**
  - Mecanismo de elección entre las distintas clases de un problema de clasificadores.
- **Representación**
  - Se suele representar por medio de un conjunto de *funciones discriminantes* $g_i(\boldsymbol{x})$. De esta forma el clasificador asigna el vector de características $\boldsymbol{x}$ a la clase w_i si $g_i(\boldsymbol{x}) \geq g_j(\boldsymbol{x})$ para todo $i \neq j$.



Esquema de un clasificador genérico

Funciones Discriminantes y Regiones de Decisión

- **Ejemplos de funciones discriminantes:**
  - Caso Bayesiano: $g_i(\boldsymbol{x})=P(w_i|\boldsymbol{x})$ o alguna expresión equivalente como: $g_i(\boldsymbol{x})=\ln(p(\boldsymbol{x}|w_i)) + \ln(P(w_i))$ para el caso Bayesiano.

- **Regiones de decisión**
  - Todo clasificador divide el espacio de características en regiones de decisión R_i donde se elige la clase *i*. La frontera entre dos regiones de decisión de llama *frontera de decisión*.
  - Utilizando las funciones discriminante las regiones de decisión se escriben para cada clase w_i como $R_i=\{\boldsymbol{x}/g_i(\boldsymbol{x}) \geq g_j(\boldsymbol{x}) \quad i\neq j\}$
  - Si R_i son R_j contiguas entonces la frontera de decisión es la intersección de las dos regiones $R_i \cap R_j=\{\boldsymbol{x}/g_i(\boldsymbol{x})=g_j(\boldsymbol{x})\}$.

Recordatorio de Probabilidad

- **La normal multivariante es la distribución conjunta de varias variables normales.**
- **Función de densidad N(μ,Σ)**

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

, $\boldsymbol{x}, \boldsymbol{\mu} \in \boldsymbol{R}^d$

Σ matriz de dxd elementos, simétrica y definida positiva ($|\Sigma| > 0$)

- **Propiedades**
  - Su valor medio es ahora un vector $E(\boldsymbol{X}) = \boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_d)^{\mathsf{T}}$ con $\mu_i = E(X_i)$
  - La dispersión y relación entre las variables se refleja en la matriz de covarianzas $\Sigma = E((\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^{\mathsf{T}}) = (\sigma_{ij})$ con $\sigma_{ij} = E((X_i - \mu_i)(X_j - \mu_j))$
    - En particular los elementos de la diagonal de la matriz Σ, $\sigma_{ii} = E((X_i - \mu_i)^2)$ son iguales a la varianza de la variable X_i
    - Los elementos fuera de la diagonal σ_{ij} miden la covarianza entre las variables X_i y X_j

      Una covarianza positiva indica que cuando crece X_i crece X_j

      Una covarianza cero indica que X_i es independiente de X_j

      Una covarianza negativa indica que cuando crece X_i decrece X_j

Regiones de Decisión: El caso Normal (1)

- **Estudiaremos las funciones discriminantes y fronteras de decisión que aparecen cuando la distribución de las características en cada clase es normal multidimensional, es decir: $p(\boldsymbol{x}|w_i) \sim N(\mu_i, \Sigma_i)$**
- **Primer caso:**
  - Las matrices de covarianzas de todas las clases son iguales, diagonales y todos los elementos de la diagonal son iguales.

    $\Sigma_i = \sigma^2 \boldsymbol{I}$, donde $\boldsymbol{I}$ es la matriz identidad.
  - Esto significa que dentro de cada clase todas las variables son independientes y tienen la misma varianza σ^2

$$g_i(\boldsymbol{x}) = \boldsymbol{a}_i^{\mathrm{T}} \boldsymbol{x} + a_{i0}$$

$$\boldsymbol{a}_i = \frac{1}{\sigma^2} \boldsymbol{\mu}_i$$

$$a_{i0} = -\frac{1}{2\sigma^2} \boldsymbol{\mu}_i^{\mathrm{T}} \boldsymbol{\mu}_i + \ln(\mathrm{P}(w_i))$$

Función discriminante

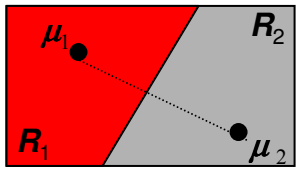
$$\boldsymbol{a}^{\mathrm{T}} (\boldsymbol{x} - \boldsymbol{x}_0) = 0$$

$$\boldsymbol{a} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$$

$$\boldsymbol{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{a}\|^2} \ln\left(\frac{\mathrm{P}(w_i)}{\mathrm{P}(w_j)}\right) \boldsymbol{a}$$

$$\|\boldsymbol{a}\|^2 = \boldsymbol{a}^{\mathrm{T}} \boldsymbol{a}$$

Superficie de decisión



Representación Gráfica

  - La frontera de decisión es lineal y perpendicular a la recta que une las medias de las dos clases

Regiones de Decisión: El caso Normal (2)

- **Segundo caso:**
  - Las matrices de covarianzas de todas las clases son iguales, esto es: $\Sigma_i = \Sigma$ con Σ una matriz común.

$$g_i(\boldsymbol{x}) = \boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{x} + a_{i0}$$

$$\boldsymbol{a}_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i$$

$$a_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln(\mathrm{P}(w_i))$$

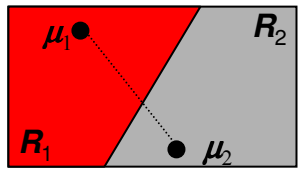
Función discriminante

$$\boldsymbol{a}^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}_0) = 0$$

$$\boldsymbol{a} = \boldsymbol{\Sigma}^{-1}\boldsymbol{d}, \quad \boldsymbol{d} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$$

$$\boldsymbol{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{1}{\left\|\boldsymbol{d}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{d}\right\|^2}\ln\left(\frac{\mathrm{P}(w_i)}{\mathrm{P}(w_j)}\right)\boldsymbol{d}$$

Superficie de decisión



Representación Gráfica

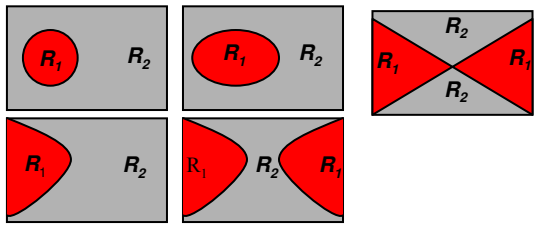
  - La frontera de decisión es lineal pero en general no es perpendicular a la recta que une las medias de las dos clases
- **Tercer caso:**
  - Las matrices de covarianzas son distintas.

$$g_i(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{A}_i\mathbf{x} + \mathbf{a}_i^{\mathrm{T}}\mathbf{x} + a_{i\mathrm{o}}$$

$$\mathbf{A}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}, \quad \mathbf{a}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$$

$$a_{i\mathrm{o}} = -\frac{1}{2}\boldsymbol{\mu}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i + \frac{1}{2}\ln|\boldsymbol{\Sigma}_i^{-1}| + \ln(\mathrm{P}(w_i))$$

Función discriminante



Representación Gráfica

  - Las fronteras de decisión son cuádricas

Aproximación Generativa en Clasificación

- **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 θ_i. Escribiremos esta dependencia como $p(\boldsymbol{x}|w_i, \theta_i)$.

    Por ejemplo para una normal multidimensional tendremos que $\theta_i=\{\mu_i, \Sigma_i\}$
  - 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:

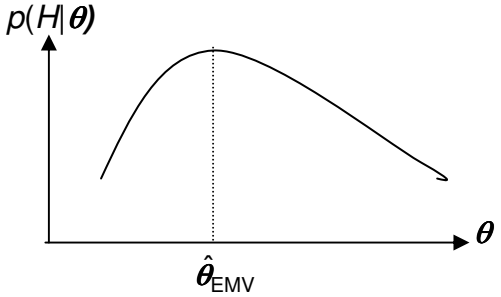
    "Utilizar un conjunto de muestras H_i tomadas de forma independiente de $p(\boldsymbol{x} \mid w_i, \boldsymbol{\theta}_i)$ para estimar $\boldsymbol{\theta}_i$"

- **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 Σ de una distribución normal por EMV, a partir de un conjunto $H=\{\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_n\}$.

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

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

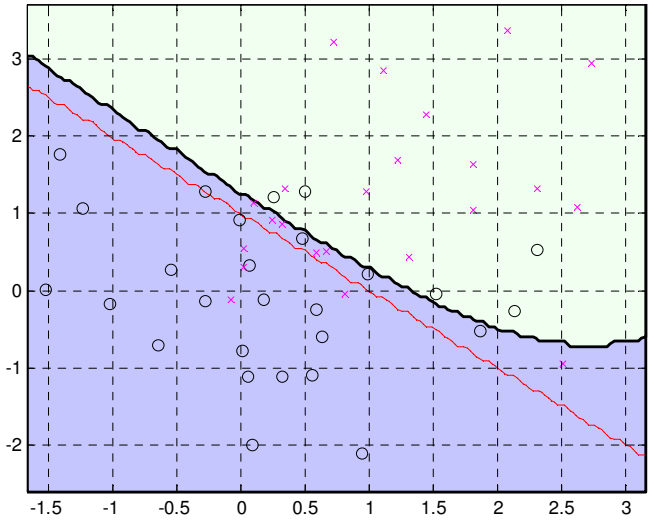
Clasificación tras estimación por EMV: Ejemplo

- **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

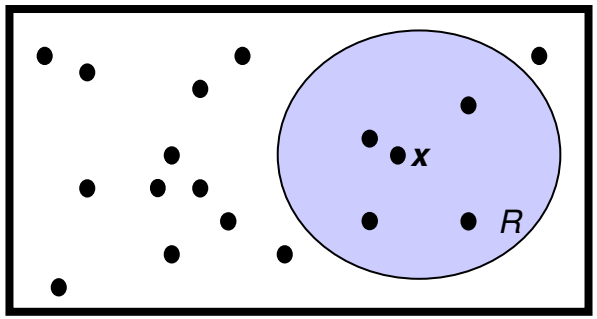
Métodos no Paramétricos (M.n.P.)

- **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

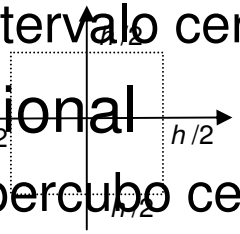
$$\hat{p}(\boldsymbol{x}) = \frac{5/18}{V}$$

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

Ventanas de Parzen:Introducción

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

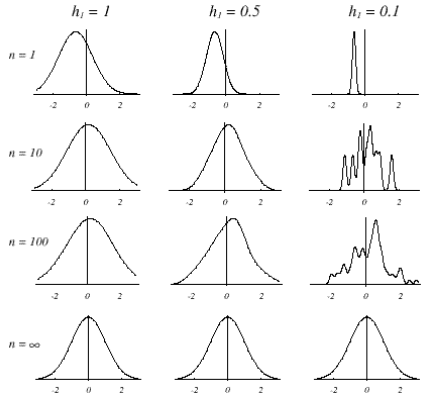
- **Vamos a formalizarlo:**
  - Caso unidimensional
    - La región es un intervalo centrado en ***x*** de longitud *h*
  - Caso multidimensional
    - La celda es un hipercubo centrado en ***x*** y la longitud de cada lado es *h*



Hipercubo en 2-D

Ventanas de Parzen: La elección de *h*

- **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).



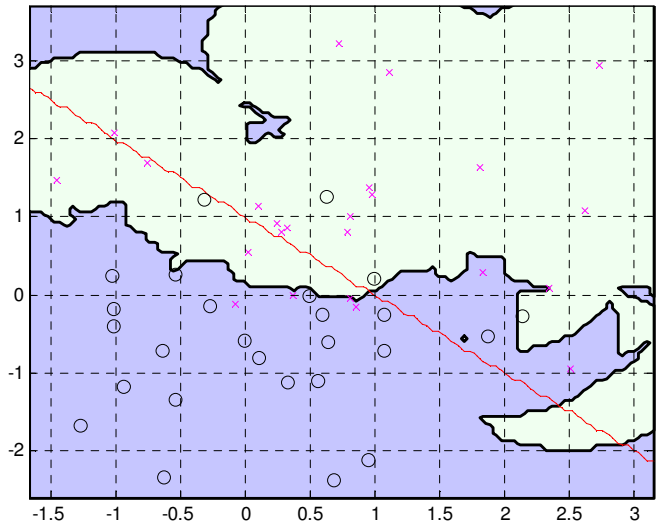
Estimación de Parzen de una función de distribución gausiana para distintos valores de h y n

- **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.

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

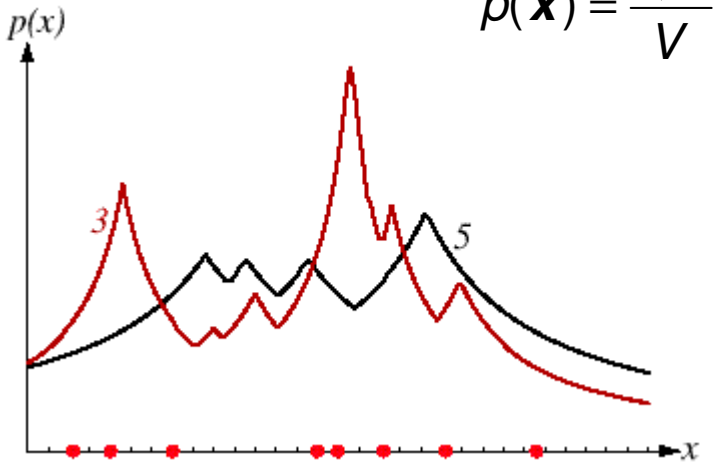


Ejemplo de clasificación tras estimación mediante Parzen
Circulos: muestras de la clase 1
Aspas: muestras de la clase 2
Linea negra: Frontera de decisión a partir de la estimación
Linea roja: Frontera de decisión bayesiana

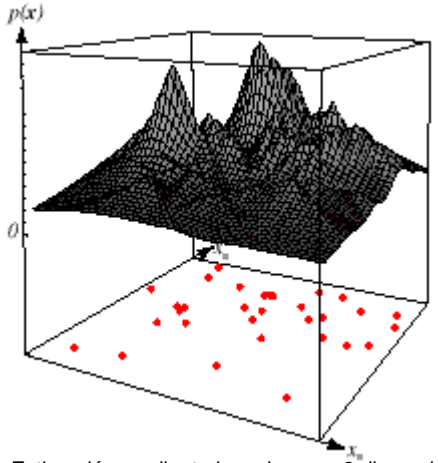
Estimación por *k*- vecinos más próximos

- **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

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:

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

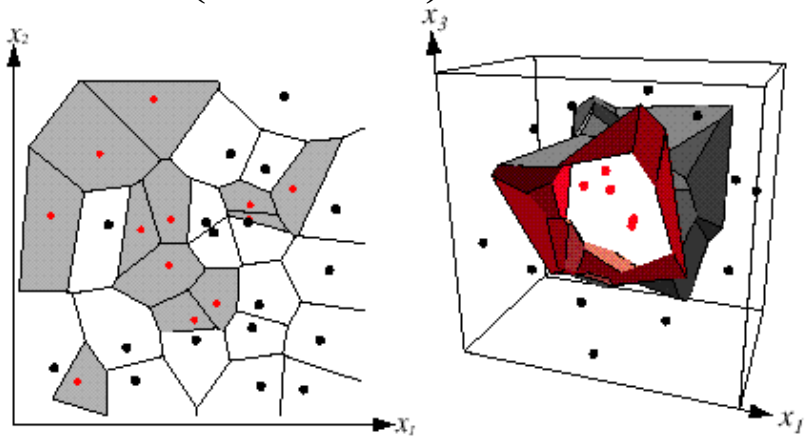
  - Entonces si utilizamos $\hat{p}(w_i) = \frac{n_i}{n}$ tendremos $\hat{p}(w_i \mid \boldsymbol{x}) = \frac{k_i}{k}$
  - La regla es simple: "Seleccionar la clase con mayor número de elementos en la región R".

    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.

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:*

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



Clasificación mediante el vecino más próximo en 1 y 2 dimensiones

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:

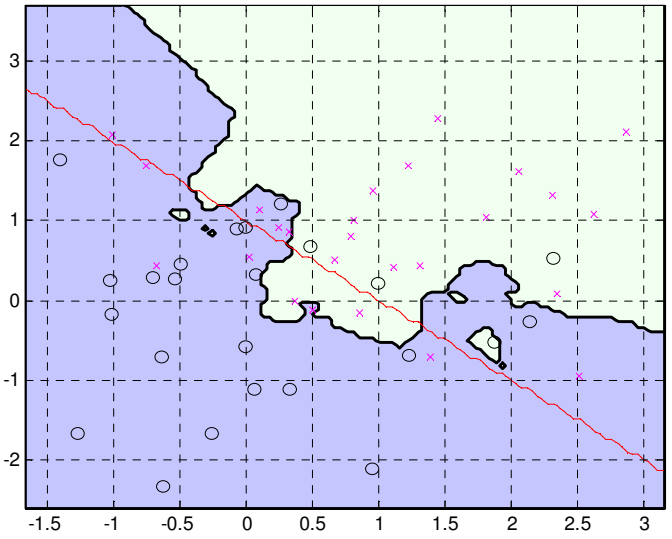
$$P^* \leq P_{\text{kNN}} \leq P^* + \frac{1}{\sqrt{k\,e}}$$

- **¿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