Introducción

- **¿Por qué una aproximación estadística en el RP?**
  - 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.
- **El Reconocimiento Estadístico de Patrones (REP) se basa en:**
  - Considerar un patrón 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 en el área de RP (desde finales de los años 30).
  - Su mayor aplicabilidad:
    - Clasificación con valores de las características perdidas
    - Toma de decisiones que minimizan la pérdida esperada

Fernando Pérez Nava

Recordatorio de Probabilidad (1)

- **Cuando estamos en un entorno en el que no existe certeza absoluta es necesario tener alguna forma de modelar la incertidumbre.**
- **Dentro de la IA existen muchas formas de modelar la incertidumbre: probabilidad, lógica difusa, teoría de Dempster-Shaffer.**
- **Puede comprobarse (Cox 1946) que si se pretende trabajar de forma consistente con niveles de certeza, éstos números deben cumplir las reglas de la teoría de la probabilidad.**
- **La Teoría de la Probabilidad (TP) asocia un valor numérico entre 0 y 1 a la certeza en un evento. La certeza absoluta de que un evento ocurrirá toma el valor 1 y la certeza completa de que un evento no ocurrirá toma el valor 0.**

(Cox, 1946) Cox R.T, Probability, Frequency, and Reasonable Expectation, *Am. Jour. Phys.*, 14, 1-13, (1946).

Fernando Pérez Nava

Recordatorio de Probabilidad (2)

- **Las probabilidades se manipulan con dos reglas sencillas:**
  - Regla del Producto
    - Dadas dos variables X e Y que pueden tomar un conjunto finito de valores si llamamos $P(x,y)$ a la *probabilidad conjunta* de que ocurran $X=x$ e $Y=y$ entonces:

      $$P(x,y)=P(y|x)P(x)$$

      donde:$P(y|x)$ es la *probabilidad condicional* de que $Y=y$ *dado* que $X=x$

      $P(x)$ es la probabilidad *marginal* de que $X=x$ independientemente de Y

      De forma similar: $P(x,y)=P(x|y)P(y)$
  - Regla de la suma
    - Dadas de nuevo las variables X e Y se tiene: $P(y)=\sum_x P(x,y)$

      donde la suma se hace sobre todos los valores x de la variable X

      De forma similar: $P(x)=\sum_y P(x,y)$

Fernando Pérez Nava

Recordatorio de Probabilidad (3)

- **A partir de la regla del producto se obtiene la *Regla de Bayes*:**

$$P(x \mid y) = \frac{P(y \mid x)P(x)}{P(y)}$$

con: $P(y) = \sum_x P(x, y) = \sum_x P(y \mid x)P(x)$

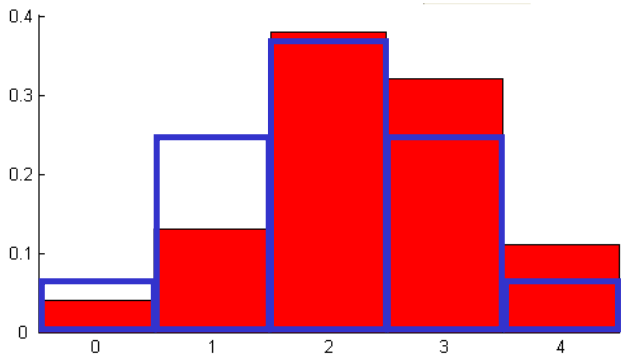
  - Podemos considerar P(x) como la probabilidad a priori (inicial) de que X=x antes de observar la variable Y.

    Entonces P(x|y) nos dice la probabilidad de que X=x después de observar la variable Y.

  - La regla de Bayes proporciona por tanto la forma de adaptar nuestras creencias iniciales a la vista de nueva información

Frecuencias Relativas y Probabilidades

- **La frecuencia relativa de un evento es el cociente entre el número de veces que se presenta un evento y el número total de observaciones**
- **Las frecuencias relativas y las probabilidades tienen propiedades muy parecidas:**
  - Ambas toman valores entre 0 y 1
  - Ambas cumplen la Regla del Producto, la Regla de la Suma y la Regla de Bayes
- **De hecho, la frecuencia relativa de un evento converge* a su probabilidad cuando el número de observaciones tiende a infinito.**



Ejemplo de convergencia de frecuencias relativas a probabilidades
Azul: Probabilidad de obtener n caras al tirar 4 monedas
Rojo: Frecuencia relativa del número de caras tras 100 lanzamientos

*Converge con probabilidad 1

Fernando Pérez Nava

Teoría de Decisión Bayesiana (TDB): Motivación (1)

- **Retomemos el experimento de la clasificación con 2 Clases, salmones y ródalos. (w_1 y w_2)**
- **Supongamos que la característica elegida es la longitud (*X*) y supongamos por simplificar que ésta toma 3 valores:**
  - x_1=corta (0-40 cm), x_2=*media*(40-100 cm) y x_3=*larga* (>100 cm)
- **Supongamos que tenemos el siguiente conjunto de entrenamiento:**

  H={(x_1, w_2), (x_2, w_2), (x_2, w_2), (x_2, w_2), (x_2, w_2), (x_2, w_2), (x_2, w_2), (x_3, w_2), (x_3, w_2), (x_1, w_1), (x_1, w_1), (x_1, w_1), (x_1, w_1), (x_2, w_1), (x_2, w_1), (x_2, w_1), (x_2, w_1), (x_2, w_1), (x_3, w_1), (x_3, w_1)}
- **¿Como diseñarías el clasificador?**
  - Cuál sería tu elección (w_1 o w_2) si:
    - Se observa $X= x_1$ (Corta)
    - Se observa $X= x_2$ (Media)
    - Se observa $X= x_3$ (Larga)

Fernando Pérez Nava

TDB: Motivación (2)

- **Un criterio sencillo: buscar la regla que produzca menos errores o lo que es lo mismo elegir la clase de mayor frecuencia absoluta (o relativa)**

	x_1	x_2	x_3
w_1	4	5	2
w_2	1	6	2

Frecuencias absolutas

	x_1	x_2	x_3
Elijo w_1	1	6	2
Elijo w_2	4	5	2

Errores absolutos sobre el conjunto de entrenamiento. Amarillo: Valores mínimos

	x_1	x_2	x_3
w_1	4/20	5/20	2/20
w_2	1/20	6/20	2/20

Frecuencias relativas

	x_1	x_2	x_3
Elijo w_1	1/20	6/20	2/20
Elijo w_2	4/20	5/20	2/20

Errores relativos sobre el conjunto de entrenamiento. Amarillo: Valores mínimos

Decisión. Naranja:Salmón, Violeta:Ródalo.

- **La frecuencia relativa del error de esta regla es 8/20 y no hay ninguna regla con menor error sobre este conjunto de entrenamiento*.**

*Hay otra regla con el mismo error

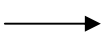
Fernando Pérez Nava

TDB: Motivación (3)

- **¿A que se aproxima la tabla de errores relativos cuando el número de muestras tiende a infinito?**

	x_1	x_2	x_3
Elijo w_1	1/20	6/20	2/20
Elijo w_2	4/20	5/20	2/20

Errores relativos sobre el conjunto de entrenamiento. Amarillo: Valores mínimos



	x_1	x_2	x_3
Elijo w_1	$P(x_1,w_2)$	$P(x_2,w_2)$	$P(x_3,w_2)$
Elijo w_2	$P(x_1,w_1)$	$P(x_2,w_1)$	$P(x_3,w_1)$

Probabilidad de error.

- **Converge a la probabilidad de error. Por tanto en el caso ideal de un número infinito de muestras la relación entre frecuencias relativas y probabilidades sugiere utilizar :**
  - Elegir w_1 si $P(x, w_1) > P(x, w_2)$
  - Elegir w_2 si $P(x, w_2) > P(x, w_1)$

- **La intuición es buena. La regla anterior es óptima.**

TDB: Motivación (4)

- **La regla:**
  - Elegir w_1 si P(x, w_1) > P(x, w_2)
  - Elegir w_2 si P(x, w_2) > P(x, w_1)

se puede escribir como (utilizando la regla del producto):

  - Elegir w_1 si P($x \mid w_1$) P(w_1) > P($x \mid w_2$) P(w_2)
  - Elegir w_2 si P($x \mid w_2$) P(w_2) > P($x \mid w_1$) P(w_1)

    P($x \mid w_i$) se llama *distribución de la característica en la clase* e indica la probabilidad de los valores de X dentro de la clase w_i

    P(w_i) se llama *probabilidad a priori de la clase* e indica la probabilidad de que aparezca un objeto de la clase w_i

o dividiendo en ambos miembros por *p*(*x*) se obtiene:

  - Elegir w_1 si P($w_1 \mid x$) > P($w_2 \mid x$)
  - Elegir w_2 si P($w_2 \mid x$) > P($w_1 \mid x$)

    P($w_i \mid x$) se llama *probabilidad a posteriori de la clase* e indica la probabilidad de la clase tras haber observado la variable X

entonces, la regla óptima consiste en elegir la clase más probable tras haber observado el valor *x*.

TDB: Motivación (5)

- **Volviendo al problema del pescado ¿cómo interpretamos las probabilidades $P(w_i)$, $P(x|w_i)$, $P(w_i|x)$**

	x_1	x_2	x_3
w_1	4	5	2
w_2	1	6	2

Frecuencias absolutas

w_1	11/20
w_2	9/20

Frecuencias relativa de cada clase

	x_1	x_2	x_3
w_1	4/11	5/11	2/11

Frecuencias relativa de X en w_1

	x_1	x_2	x_3
w_2	1/9	6/9	2/9

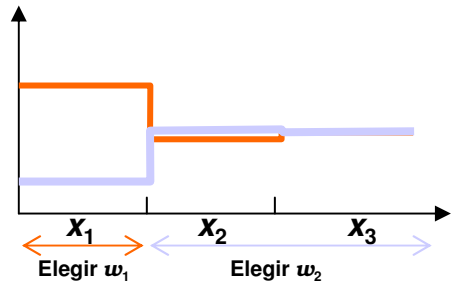
Frecuencias relativa de X en w_2

	x_1	x_2	x_3
w_1	4/5	5/11	2/4

Frecuencias relativa de w_1 dado X

	x_1	x_2	x_3
w_2	1/5	6/11	2/4

Frecuencias relativa de w_2 dado X



Regiones de decisión: Representación gráfica

Recordatorio de Probabilidad (4)

- **Variables Aleatorias Continuas**
  - Cuando una variable X toma valores reales la probabilidad de tomar un valor específico es siempre nula. Por ello se habla de la probabilidad de que tome valores en un intervalo (a,b) mediante una *función de densidad* $p(x)$:

$$\mathrm{P}(x \in (a,b)) = \int_a^b p(x)dx$$

  - En general, todas las definiciones dadas para variables discretas se pasan a continuas cambiando sumas por integrales. Así si X *e* Y son continuas las reglas del producto, suma y de Bayes quedan:

$$p(x,y) = p(y \mid x)p(x) \qquad p(y) = \int_{-\infty}^{\infty} p(x,y)dx \qquad p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}$$

  - Cuando se tiene un vector de variables aleatorias $\boldsymbol{X}=(X_1, X_2,\ldots X_n)^T$ se tiene una *función de densidad multidimensional* $p(\boldsymbol{x})$

$$\mathrm{P}(\boldsymbol{x} \in R) = \int_R p(\boldsymbol{x})d\boldsymbol{x}$$

Fernando Pérez Nava

Teoría de la Decisión Bayesiana (TDB)

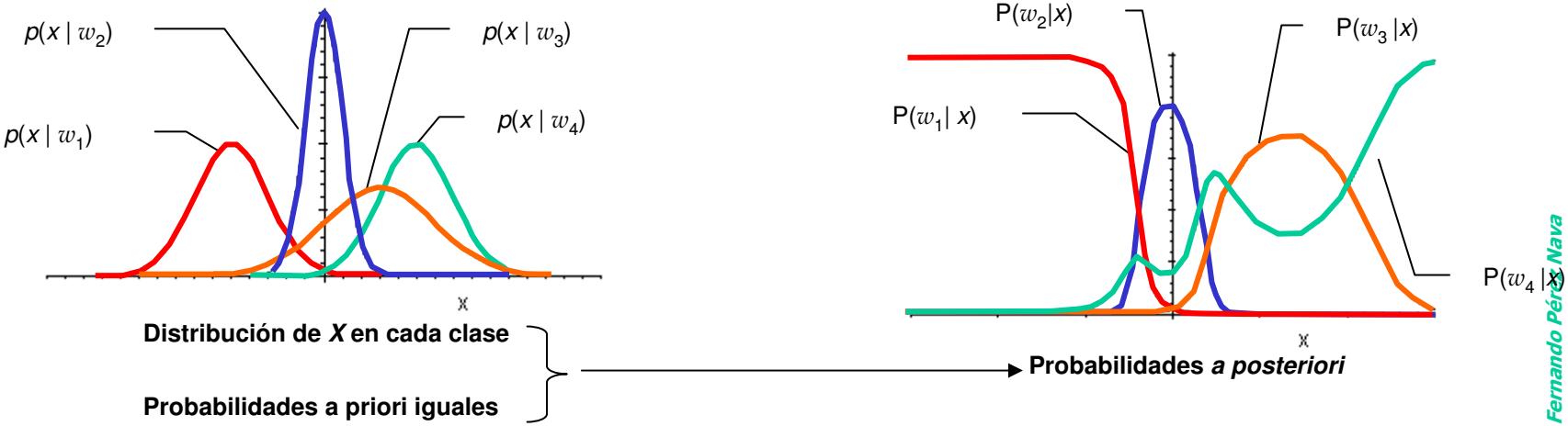
- **La TDB proporciona un marco teórico para tomar decisiones en situaciones de incertidumbre.**
- **En nuestro caso la decisión será la clasificación de un patrón en una determinada clase**
- **La TDB proporciona el clasificador óptimo (clasificador bayesiano) para un conjunto de características dadas**
  - 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 $p(x|\omega_i)$ 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 $P(\omega_i)$

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:**

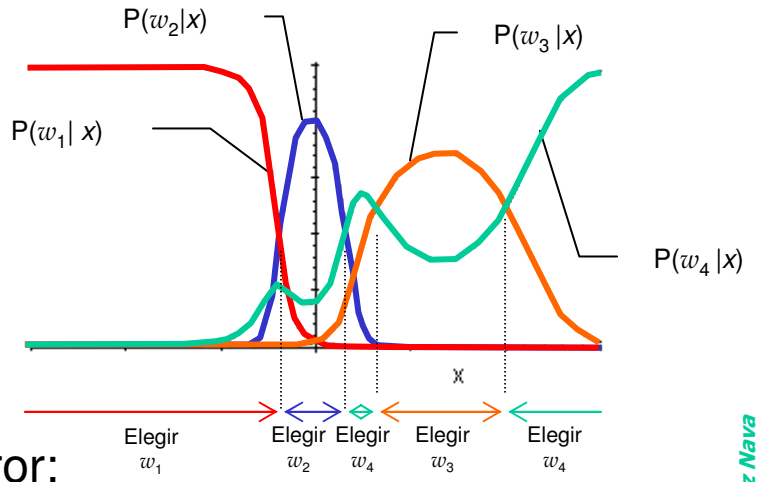


TDB: Enfoque formal (2)

- **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:

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

Fernando Pérez Nava

Decisión Bayesiana con Riesgo (DBR): Motivación (1)

- **Retomemos el experimento anterior con 2 Clases: salmones y ródalos. (w_1 y w_2); una característica: longitud con tres valores x_1=corta, x_2=*media* y x_3=*larga* y el conjunto de entrenamiento:**

  H={(x_1, w_2), (x_2, w_2), (x_2, w_2), (x_2, w_2), (x_2, w_2), (x_2, w_2), (x_2, w_2), (x_3, w_2), (x_3, w_2), (x_1, w_1), (x_1, w_1), (x_1, w_1), (x_1, w_1), (x_2, w_1), (x_2, w_1), (x_2, w_1), (x_2, w_1), (x_3, w_1), (x_3, w_1)}

- **Los errores que aparecen al realizar la clasificación son:**
  - Elegir w_1 (salmón) cuando la clase verdadera es w_2 (ródalo)
  - Elegir w_2 (ródalo) cuando la clase verdadera es w_1 (salmón)
  - El salmón es un pescado más caro que el ródalo. Supongamos que:
    - Si eliges w_1 cuando la clase verdadera es w_1 has detectado un salmón. El costo de procesamiento del sistema es de λ_{11}= 1 unidad monetaria
    - Si eliges w_1 cuando la clase verdadera es w_2 proporcionas un producto de peor calidad de la especificada y eso cuesta en sanciones λ_{12}= 11 unidades monetarias.
    - Si eliges w_2 cuando la clase verdadera es w_1 proporcionas un producto de mayor calidad de la necesaria y eso cuesta λ_{21}= 10 unidades monetarias.
    - Si eliges w_2 cuando la clase verdadera es w_2 has detectado un ródalo. El costo de procesamiento del sistema es de λ_{22}=1 unidad monetaria
- **¿Qué elegirías ahora w_1 o w_2 para $X=x_1$, $X=x_2$ y $X=x_3$?**

Fernando Pérez Nava

DBR:Motivación (2)

- Con la notación utilizada λ_{ij} es el costo de elegir la clase w_i cuando la verdadera es w_j:

		Clase Verdadera w_1	Clase Verdadera w_2
Elijo	w_1	$\lambda_{11}=1$	$\lambda_{12}=11$
	w_2	$\lambda_{21}=10$	$\lambda_{22}=1$

- Una regla que parece lógica es elegir la clase que produzca el menor costo

	x_1	x_2	x_3
w_1	4	5	2
w_2	1	6	2

Frecuencias absolutas

	x_1	x_2	x_3
Elijo w_1	$1\times4+11\times1=15$	$1\times5+11\times6=71$	$1\times2+11\times2=24$
Elijo w_2	$1\times1+10\times4=41$	$1\times6+10\times5=56$	$1\times2+10\times2=22$

Costos absolutos. Amarillo: costos mínimos

	x_1	x_2	x_3
w_1	4/20	5/20	2/20
w_2	1/20	6/20	2/20

Frecuencias relativas

Decisión. Naranja:Salmón, Violeta:Ródalo

	x_1	x_2	x_3
Elijo w_1	$1\times4/20+11\times1/20=15/20$	$1\times5/20+11\times6/20=71/20$	$1\times2/20+11\times2/20=24/20$
Elijo w_2	$1\times1/20+10\times4/20=41/20$	$1\times6/20+10\times5/20=56/20$	$1\times2/20+10\times2/20=22/20$

Costo relativos: Amarillo: costos mínimos

- El costo relativo de esta regla es 93/20 (mínimo sobre H)

Fernando

DBR:Motivación (3)

- **¿A que se aproxima la tabla de costos relativos cuando el número de muestras tiende a infinito?**

	x_1	x_2	x_3
Elijo w_1	1×4/20+11×1/20=15/20	1×5/20+11×6/20=71/20	1×2/20+11×2/20=24/20
Elijo w_2	1×1/20+10×4/20=41/20	1×6/20+10×5/20=56/20	1×2/20+10×2/20=22/20

Costo relativos: Amarillo: costos mínimos

	x_1	x_2	x_3
Elijo w_1	$\lambda_{11}\times P(x_1,w_1)+\lambda_{12}\times P(x_1,w_2)$	$\lambda_{11}\times P(x_2,w_1)+\lambda_{12}\times P(x_2,w_2)$	$\lambda_{11}\times P(x_1,w_1)+\lambda_{12}\times P(x_1,w_2)$
Elijo w_2	$\lambda_{21}\times P(x_1,w_1)+\lambda_{22}\times P(x_1,w_2)$	$\lambda_{21}\times P(x_1,w_1)+\lambda_{22}\times P(x_1,w_2)$	$\lambda_{21}\times P(x_1,w_1)+\lambda_{22}\times P(x_1,w_2)$

Costo medio

- **Por tanto en el caso ideal de un número infinito de muestras la relación entre frecuencias relativas y probabilidades sugiere utilizar:**
  - Elegir w_1 si $\lambda_{11}\, P(x, w_1) + \lambda_{12}\, P(x, w_2) < \lambda_{21}\, P(x, w_1) + \lambda_{22}\, P(x, w_2)$
  - Elegir w_2 si $\lambda_{21}\, P(x, w_1) + \lambda_{22}\, P(x, w_2) < \lambda_{11}\, P(x, w_1) + \lambda_{12}\, P(x, w_2)$

Fernando Pérez Nava

DBR: Motivación (4)

- **La intuición es correcta. La regla:**
  - Elegir w_1 si $\lambda_{11}\, P(x, w_1) + \lambda_{12}\, P(x, w_2) < \lambda_{21}\, P(x, w_1) + \lambda_{22}\, P(x, w_2)$
  - Elegir w_2 si $\lambda_{21}\, P(x, w_1) + \lambda_{22}\, P(x, w_2) < \lambda_{11}\, P(x, w_1) + \lambda_{12}\, P(x, w_2)$

  es óptima
- **La regla se puede escribir dividiendo por P(*x*) como:**
  - Elegir w_1 si $\lambda_{11}\, P(w_1|x) + \lambda_{12}\, P(w_2|x) < \lambda_{21}\, P(w_1|x) + \lambda_{22}\, P(w_2|x)$
  - Elegir w_2 si $\lambda_{21}\, P(w_1|x) + \lambda_{22}\, P(w_2|x) < \lambda_{11}\, P(w_1|x) + \lambda_{12}\, P(w_2|x)$

    Se suele escribir:

    $R(w_1|x) = \lambda_{11}\, P(w_1|x) + \lambda_{12}\, P(w_2|x)$

    $R(w_2|x) = \lambda_{21}\, P(w_1|x) + \lambda_{22}\, P(w_2|x)$

    a $R(w_i|x)$ se le llama *riesgo de elegir* w_i *dado x* e indica el costo de elegir w_i tras haber observado el valor *x*

  entonces, la regla óptima consiste en elegir la clase con menor costo tras haber observado el valor *x*

Fernando Pérez Nava

DBR: 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:

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

  - Acciones: α_i, i=1...c; α_i = "Elegir w_i "
  - Riesgos: $\lambda_{i,j} = \lambda(\alpha_i \mid w_j)$ i=1...c, j=*1*...*c*. Indica el riesgo de elegir w_i cuando la verdadera clase es w_j

  - Función de riesgo dado un valor de $\boldsymbol{x}$:

$$R(\alpha_i \mid \boldsymbol{x}) = \sum_{j=1}^{c} \lambda(\alpha_i \mid w_j)P(w_j \mid \boldsymbol{x}) \quad i = 1,...c$$

Fernando Pérez Nava

DBR: Enfoque formal (2)

- **Regla de decisión bayesiana (óptima):**
  - Elegir α_i si $R(\alpha_i|\ \boldsymbol{x}) \leq R(\alpha_j|\ \boldsymbol{x})\ \forall i \neq j$
  - Esto es, elegir la clase con menor riesgo dado el valor de $\boldsymbol{x}$

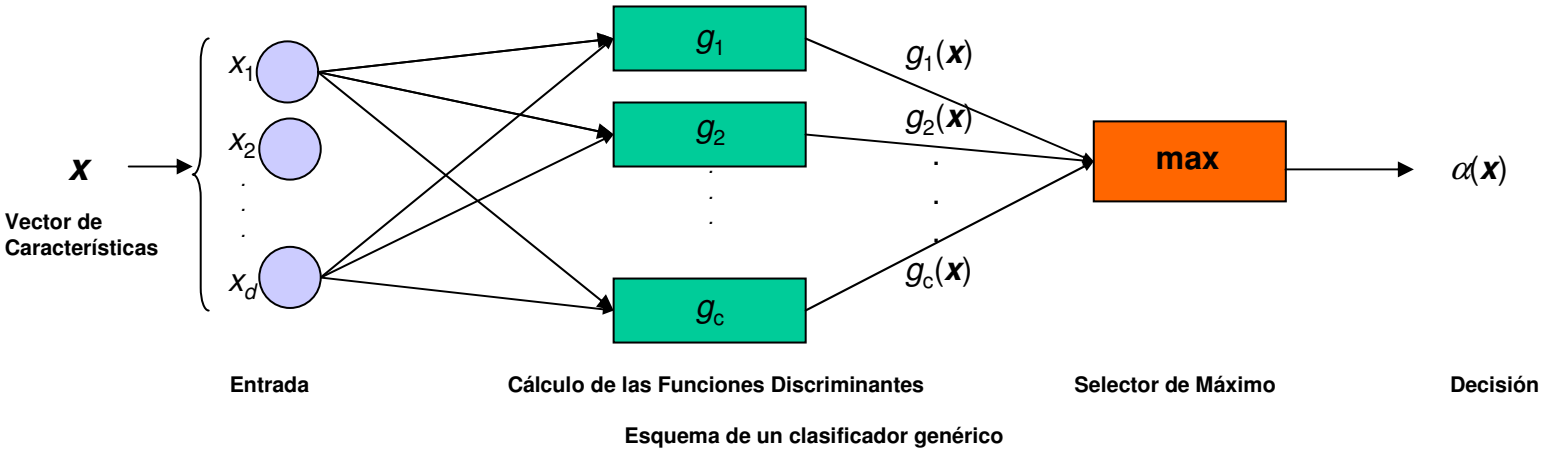
- **Propiedad:**
  - Hace mínimo el riesgo total:

$$R = \int R(\alpha(\boldsymbol{x}) \mid \boldsymbol{x}) p(\boldsymbol{x})\, d\boldsymbol{x}$$

Fernando Pérez Nava

Clasificadores y su Representación

- **Definición formal de Clasificador**
  - Mecanismo de elección entre las distintas clases de un problema de R.P.
- **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: $g_i(\boldsymbol{x})=P(w_i|\boldsymbol{x})$
  - Caso Bayesiano con riesgo: $g_i(\boldsymbol{x})=-R(\alpha_i|\boldsymbol{x})$

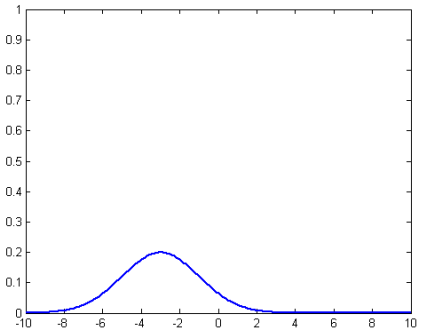
    o alguna expresión equivalente como:

    $g_i(\boldsymbol{x})=\ln(p(\boldsymbol{x}|w_i)) + \ln(P(w_i))$ para el caso Bayesiano.

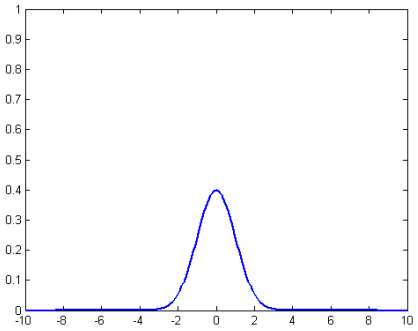
- **Regiones de decisión**
  - Todo clasificador divide el espacio de características en regiones de decisión R_i donde se elige la clase *i*. La frontera entre dos regiones de decisión de llama *frontera de decisión*.
  - Utilizando las funciones discriminante las regiones de decisión se escriben para 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 (5)
Variable Aleatoria Normal

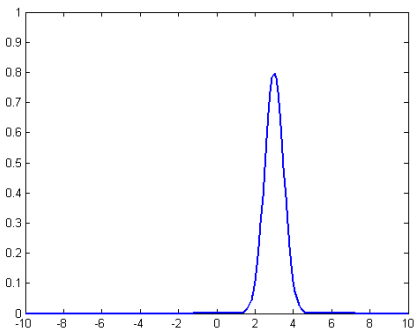
- **La normal es la variable aleatoria continua más importante.**
- **Cuando hay una única variable se llama *normal unidimensional*, cuando hay varias variables que se distribuyen de forma normal a la distribución conjunta se la llama *normal multidimensional***
- **La normal unidimensional N(μ,σ^2)**
  - Función de densidad: $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \quad \sigma > 0$
  - Algunas propiedades
    - Su valor medio E(X) es igual a μ
    - Su varianza es igual a V(X)=σ^2



N(-3,2)



N(0,1)



N(3,0.5)

Normal unidimensional. Representación gráfica

Fernando Perez Nava

Recordatorio de Probabilidad (6)

- **Independencia**
  - Dos variables X e Y son independientes si conocer una no proporciona información sobre la otra, es decir:

$$p(x \mid y) = p(x) \Leftrightarrow p(x,y) = p(x)p(y)$$

- **Esperanza de una variable aleatoria.**
  - Nos informa del valor medio de la variable: $\mathrm{E}(X) = \int_{-\infty}^{\infty} x\, p(x)dx$
  - En el caso multidimensional es un vector: $\mathrm{E}(\boldsymbol{X}) = \int \boldsymbol{x}\, p(\boldsymbol{x})d\boldsymbol{x}$
- **Varianza y covarianza de variables aleatorias.**
  - La varianza es una medida de dispersión: $\mathrm{V}(X) = \int_{-\infty}^{\infty} (x - \mathrm{E}(X))^2 p(x)dx$
  - La covarianza es una medida de relación:

$$\mathrm{Cov}(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \mathrm{E}(X))(y - \mathrm{E}(Y))p(x,y)dxdy$$

  - En el caso multidimensional se tiene la matriz de covarianzas:

$$\mathrm{Cov}(\boldsymbol{X}) = \int (\boldsymbol{x} - \mathrm{E}(\boldsymbol{x}))(\boldsymbol{x} - \mathrm{E}(\boldsymbol{x}))^{'}\, p(\boldsymbol{x})d\boldsymbol{x}$$

Fernando Pérez Nava

Recordatorio de Probabilidad (7)

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

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

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

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

- **Propiedades**
  - Su valor medio es ahora un vector $E(\boldsymbol{X}) = \boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_d)^T$ con $\mu_i = E(X_i)$
  - La dispersión y relación entre las variables se refleja en la matriz de covarianzas $\Sigma = E((\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T) = (\sigma_{ij})$ 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

Fernando Pérez Nava

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

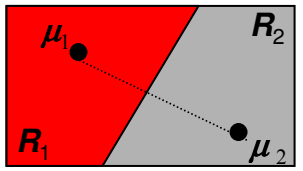
$$\boldsymbol{a}^{\mathrm{T}} (\boldsymbol{x} - \boldsymbol{x}_0) = 0$$

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

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

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

Superficie de decisión



Representación Gráfica

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

Fernando Pérez Nava

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.

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

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

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

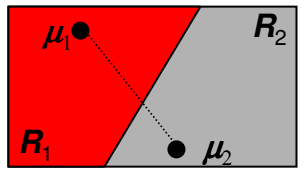
Función discriminante

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

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

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

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

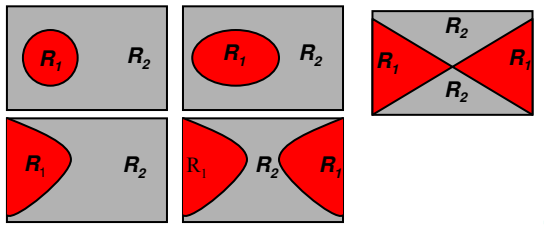
- **Tercer caso:**
  - Las matrices de covarianzas son distintas.

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

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

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

Función discriminante



Representación Gráfica

  - Las fronteras de decisión son cuádricas

Fernando Pérez Nava

Resumiendo...

- **Las buenas noticias;**

  Cuando se conoce la estructura de probabilidad del problema:

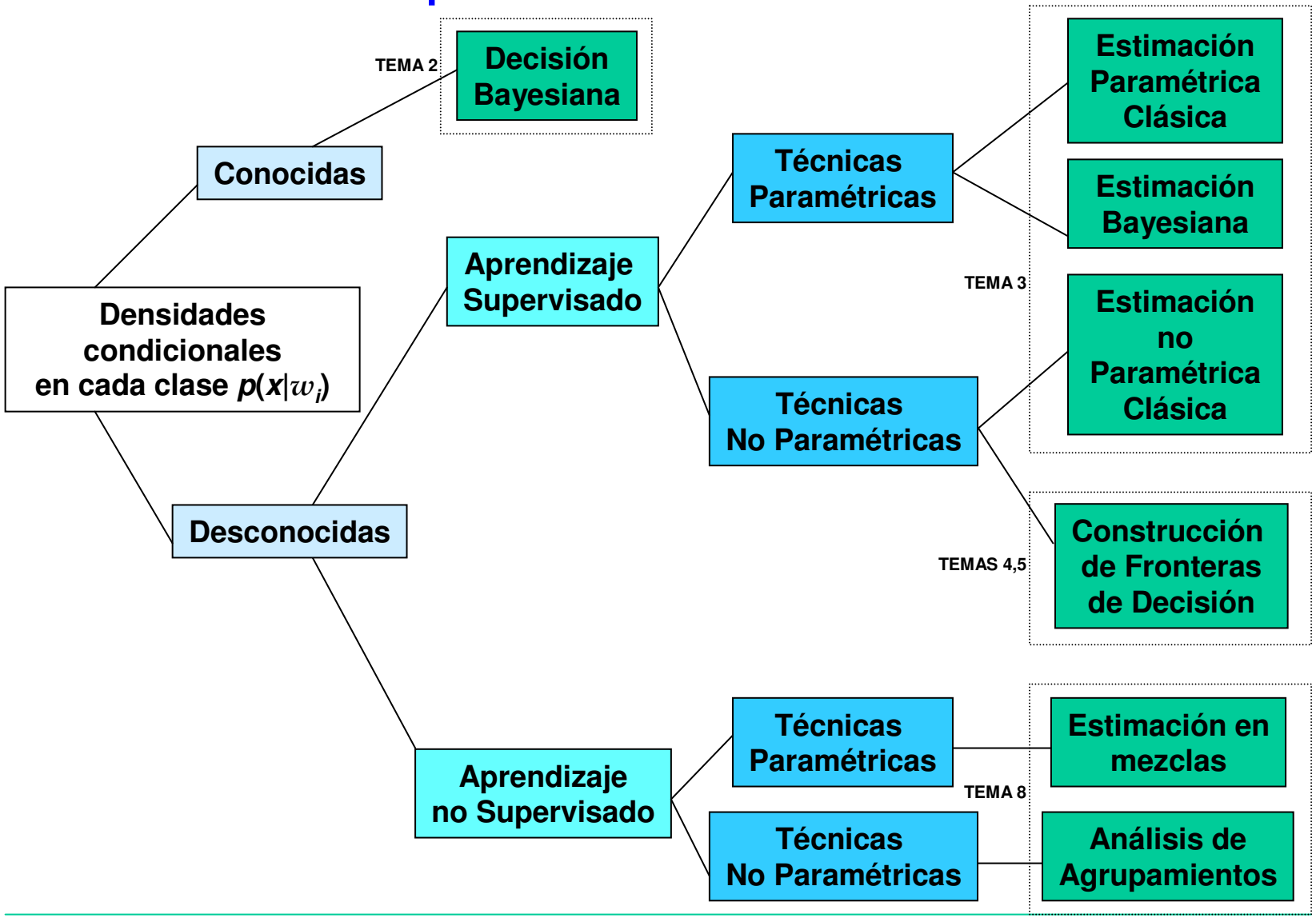
  $$P(w_i),\ p(\boldsymbol{x}|w_i)$$

  siempre se puede encontrar el clasificador óptimo (clasificador bayesiano): Elegir w_i si $P(w_i|\ \boldsymbol{x}) \geq P(w_j|\ \boldsymbol{x})\ \forall i \neq j \Leftrightarrow$

  $$p(\boldsymbol{x}\mid w_i)P(w_i) \geq p(\boldsymbol{x}\mid w_j)P(w_j)\ \forall i \neq j$$

- **Las malas noticias:**
  - En prácticamente ningún problema práctico se conoce la estructura de probabilidad del problema. ¿Qué hacer entonces?
  - Dos ideas:
    - Intentar estimar las probabilidades $P(w_i)$, $p(\boldsymbol{x}|w_i)$ a partir de un conjunto de entrenamiento. Estimar $P(w_i)$ con precisión es fácil. Estimar $p(\boldsymbol{x}|w_i)$ es un problema difícil.
    - Olvidarnos del clasificador bayesiano e introducir otros criterios (por ejemplo geométricos) con la esperanza de obtener un buen clasificador aunque no sea óptimo.

El mapa del RP Estadístico



Fernando Pérez Nava