



INTRODUCCIÓN A LOS MODELOS COMPUTACIONALES:

CUARTO CURSO DEL GRADO DE ING. INFORMÁTICA EN COMPUTACION

Análisis Discriminante Lineal

César Hervás-Martínez
Grupo de Investigación AYRNA

**Departamento de Informática y Análisis
Numérico**
Universidad de Córdoba
Campus de Rabanales. Edificio Einstein.
Email: chervas@uco.es

2018-2019



PROBABILIDADES “*A priori y a posteriori*”



La **Teoría de Decisión Bayesiana** se basa en dos suposiciones:
El problema de decisión se puede describir en términos probabilísticos:

Dado un suceso o evento, D ¿Cual es la mejor hipótesis h del conjunto de hipótesis H ?

La mejor hipótesis es la hipótesis mas probable

Todos los valores de las probabilidades *a priori* del problema son conocidas

Las decisiones se toman en función de las observaciones muestrales



PROBABILIDADES “*A priori y a posteriori*”



Espacio de hipótesis: {Cara, C; Cruz, F}, Hipótesis nula

Hipótesis alternativa

Espacio de observaciones: {Brillo, B; Mate, M}

Lanzo una moneda, recibo la observación, D, y genero una hipótesis, h.

P(h): Probabilidad de que la hipótesis h sea cierta.

Refleja el conocimiento que tenemos sobre las oportunidades de que la hipótesis h sea cierta antes de recibir ninguna observación.

Si no tenemos ningún conocimiento *a priori*, se le podrá asignar la misma probabilidad a todas las hipótesis.

P(D): Probabilidad de que obtengamos el suceso D.

Refleja la probabilidad de observar el suceso D, cuando no tenemos ninguna idea sobre cual es la hipótesis real.



PROBABILIDADES “*A priori y a posteriori*”



$P(D|h)$: Probabilidad de observar el suceso D, cuando se cumple la hipótesis h o Probabilidad condicional del suceso D.

$P(h|D)$: Probabilidad de que se cumpla la hipótesis h, dado que se ha obtenido el suceso D, o Probabilidad a posteriori de la hipótesis h

Preguntas:

¿Cual es la mejor hipótesis?

¿Cual es la hipótesis mas probable?

¿Cual es la probabilidad de obtener cara?

¿Cual es la probabilidad de obtener cruz?

¿Cual es la probabilidad de obtener cara, habiendo observado que la moneda tiene brillo?

¿Cual es la probabilidad de obtener cruz, habiendo observado que la moneda es mate?



TEOREMA DE BAYES



En **aprendizaje inductivo**, estamos interesados en calcular las probabilidades de las **hipótesis a posteriori**, ya que son las que se obtienen tras obtener observaciones muestrales o ejemplos de entrenamiento.

Teorema de Bayes:

$$P(h_i / D) = \frac{P(h_i \cap D)}{P(D)} = \frac{P(h_i)P(D / h_i)}{P(D)} = \frac{P(h_i)P(D / h_i)}{\sum_{j=1}^J P(h_j)P(D / h_j)}$$

para $P(D) \neq 0$

Calcula la probabilidad a posteriori de la hipótesis dados unos datos muestrales, en función de las probabilidades a priori y condicionadas.



Maximizar la Probabilidad a Posteriori, MAP

Decisor máxima probabilidad “a posteriori”: MAP, luego la hipótesis a considerar es

$$\begin{aligned} h_{MAP} &= \arg \max_{h \in H} P(h / D) = \arg \max_{h \in H} \frac{P(h)P(D / h)}{P(D)} = \\ &= \arg \max_{h \in H} P(h)P(D / h) \quad \text{para } P(D) \neq 0 \end{aligned}$$

El decisor de máxima verosimilitud, ML (maximum likelihood), asume que todas las hipótesis son equiprobables “a priori”:
Ahora la hipótesis a considerar es

$$h_{ML} = \arg \max_{h \in H} P(D / h)$$



EJEMPLO



Sean los sucesos X = Sacar cara; F =Sacar cruz
y sus probabilidades a priori:

$$P(X) = 0,2, P(F) = 0,8$$

Sean los sucesos B =obtener brillo; M =obtener mate.

Conocemos que las probabilidades a posteriori son:

$P(B|X) = 0,9$; $P(M|X) = 0,1$, (el 90% de las monedas tienen la parte de la cara con brillo)

$P(B|F) = 0,6$; $P(M|F) = 0,4$, (el 60% de las monedas tiene la parte de la cruz con brillo).

Tiro la moneda y obtengo brillo, esto es obtengo el suceso B :

$$\begin{aligned} h_{MAP} &= \arg \max_{X,F} P(B | h)P(h) = \\ &= \arg \max_{X,F} P(P(B | X)P(X), P(B | F)P(F)) = \\ &= \arg \max_{X,F} (0,9 \times 0,2, 0,6 \times 0,8) = \\ &= \arg \max_{X,F} (0,18, 0,48) = F \end{aligned}$$



EJEMPLO



Sin embargo con el método de máxima verosimilitud, la decisión es la contraria

$$\begin{aligned}h_{ML} &= \arg \max_{X,F} P(B | h) = \\&= \arg \max_{X,F} P(P(B | X), P(B | F)) = \\&= \arg \max_{X,F} (0,9, 0,6) = X\end{aligned}$$

frente a

$$h_{MAP} = \arg \max_{X,F} (0,18, 0,48) = F$$

Esto se debe a las diferentes probabilidades *a priori* de ambas hipótesis. $P(X)= 0,2$, $P(F)= 0,8$ en un caso y equiprobables en otro.



CLASIFICADOR MAP



Caso general:

1 Para cada clase (hipótesis) h , calcular la probabilidad *a posteriori*:

$$P(h / D) = \frac{P(h)P(D / h)}{P(D)}$$

para $P(D) \neq 0$

2 Dar como salida a un patrón la clase (hipótesis) con la mayor probabilidad *a posteriori*:

$$h_{MAP} = \arg \max_{h \in H} P(h | D)$$



CLASIFICADOR MAP



Caso general para dos clases:

Tenemos un problema de clasificación binaria (dos hipótesis simples, clase 1 frente a clase 2) con una **función discriminante** del tipo:

$$g(D) = \frac{1}{P(D)} (P(h_1)P(D / h_1) - P(h_2)P(D / h_2))$$

Donde $H_0 = h_1$, clase C_1 frente a la hipótesis alternativa $H_1 = h_2$, clase C_2

$$h_{MAP} = \begin{cases} h_1 \text{ (Clase 1)} & \text{Si } g(\mathbf{x}) \geq 0, \text{ esto es,} \\ h_1 \text{ (Clase 1)} & \text{Si } \frac{P(h_1)P(D / h_1)}{P(D)} \geq \frac{P(h_2)P(D / h_2)}{P(D)} \\ h_2 \text{ (Clase 2)} & \text{Si } g(\mathbf{x}) < 0, \end{cases}$$



PROBABILIDADES DE ERROR DE CLASIFICACIÓN

Un clasificador binario, divide el espacio en dos regiones, R_1 para la clase C_1 , y R_2 para la clase C_2 .

Errores de clasificación de una instancia x :

Error de tipo I:

x pertenece a la clase C_1 pero cae en la región R_2

Error de tipo II:

x pertenece a clase C_2 pero cae en la región R_1

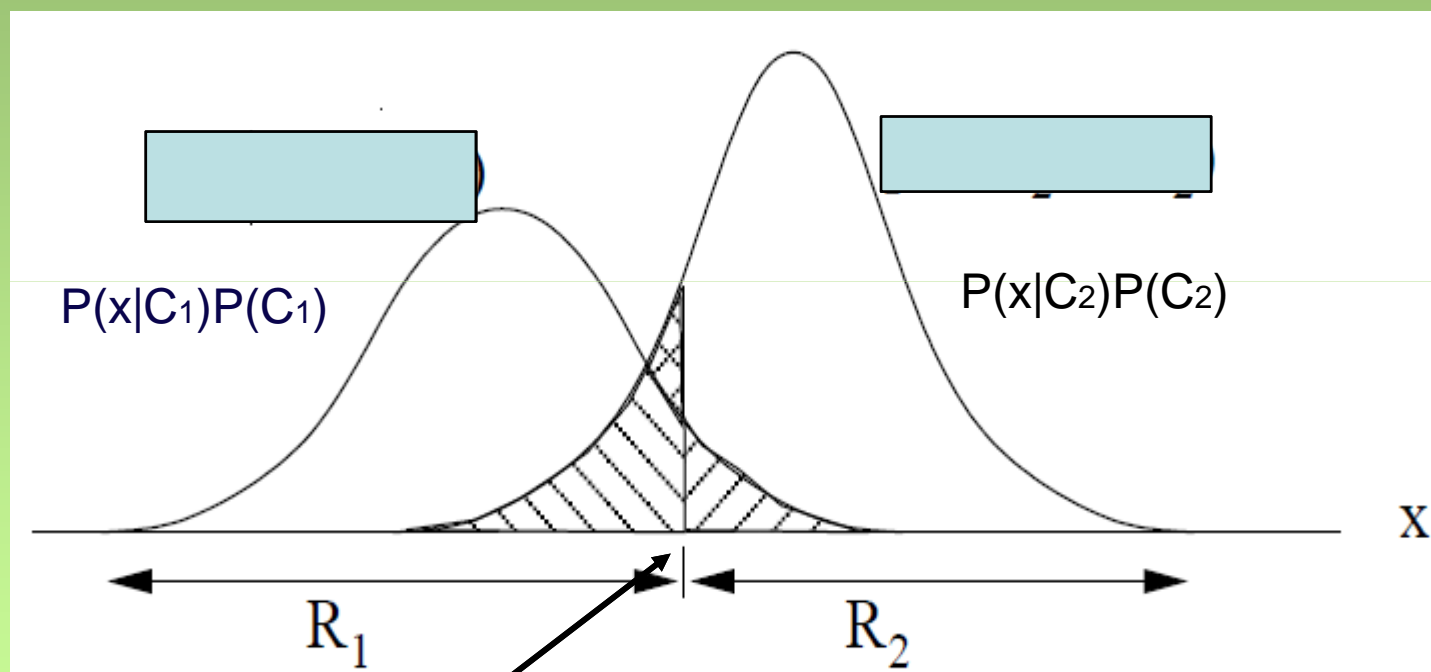
Probabilidad de Error de un clasificador MAP:

$$\begin{aligned} P(E) &= P(\mathbf{x} \in R_2, C_1) + P(\mathbf{x} \in R_1, C_2) = \\ &= P(\mathbf{x} \in R_2 \mid C_1)P(C_1) + P(\mathbf{x} \in R_1 \mid C_2)P(C_2) = \\ &= \int_{R_2} P(\mathbf{x} \mid C_1)P(C_1)d\mathbf{x} + \int_{R_1} P(\mathbf{x} \mid C_2)P(C_2)d\mathbf{x} \end{aligned}$$



PROBABILIDADES DE ERROR DE CLASIFICACIÓN

Si las distribuciones de probabilidad $P(x/C_1)$ y $P(x/C_2)$ son Normales



¿Cuál es el punto que minimiza el error de clasificación?



DISTRIBUCIÓN NORMAL



La función de densidad de una v.a $N(\mu, \sigma)$ unidimensional es

$$f(x; \mu, \sigma^2) = \frac{1}{(2\pi)^{1/2} \sigma} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}, \text{ para } -\infty < x < \infty$$

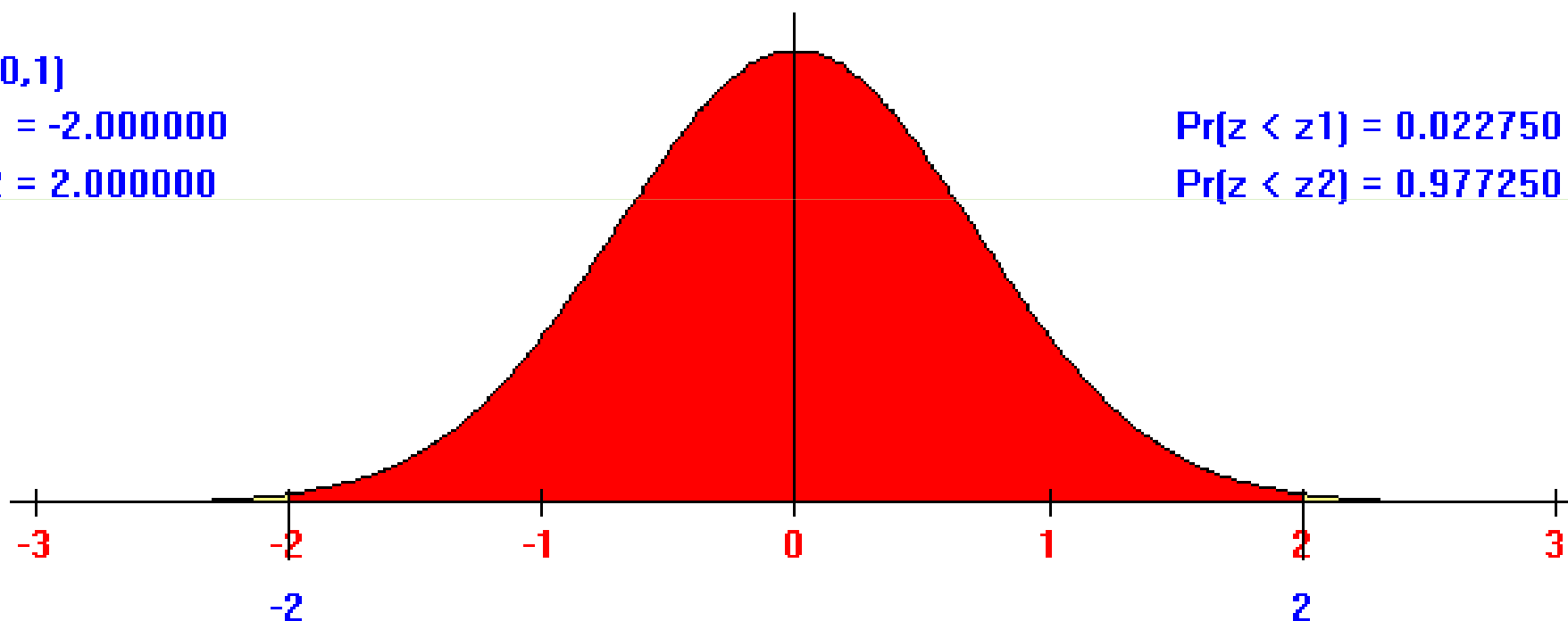
$N(0,1)$

$z1 = -2.000000$

$z2 = 2.000000$

$\Pr\{z < z1\} = 0.022750$

$\Pr\{z < z2\} = 0.977250$





Ejemplo de clasificador lineal: Análisis Discriminante



Análisis discriminante lineal biclase:

Se realiza la clasificación mediante una transformación lineal donde los coeficientes del modelo son determinados mediante uno o varios procedimientos de optimización.

Se trabaja con un conjunto de datos linealmente separables, es decir, conjuntos de datos cuyas clases pueden ser separadas por superficies de decisión lineales.



Ejemplo de clasificador lineal: Análisis Discriminante



Análisis discriminante no lineal: En este caso trataremos con situaciones donde los datos no son linealmente separables como no sea mediante una transformación del espacio a otro de mayor dimensión, a un espacio de Hilbert, es la técnica que se utiliza en las máquinas de vectores soporte, SVM.



Análisis Discriminante



La Figura 1 muestra un ejemplo bidimensional y biclase de dos distribuciones normales donde los puntos definen las muestras tomadas y las líneas continuas los contornos de las funciones de densidad de probabilidad.

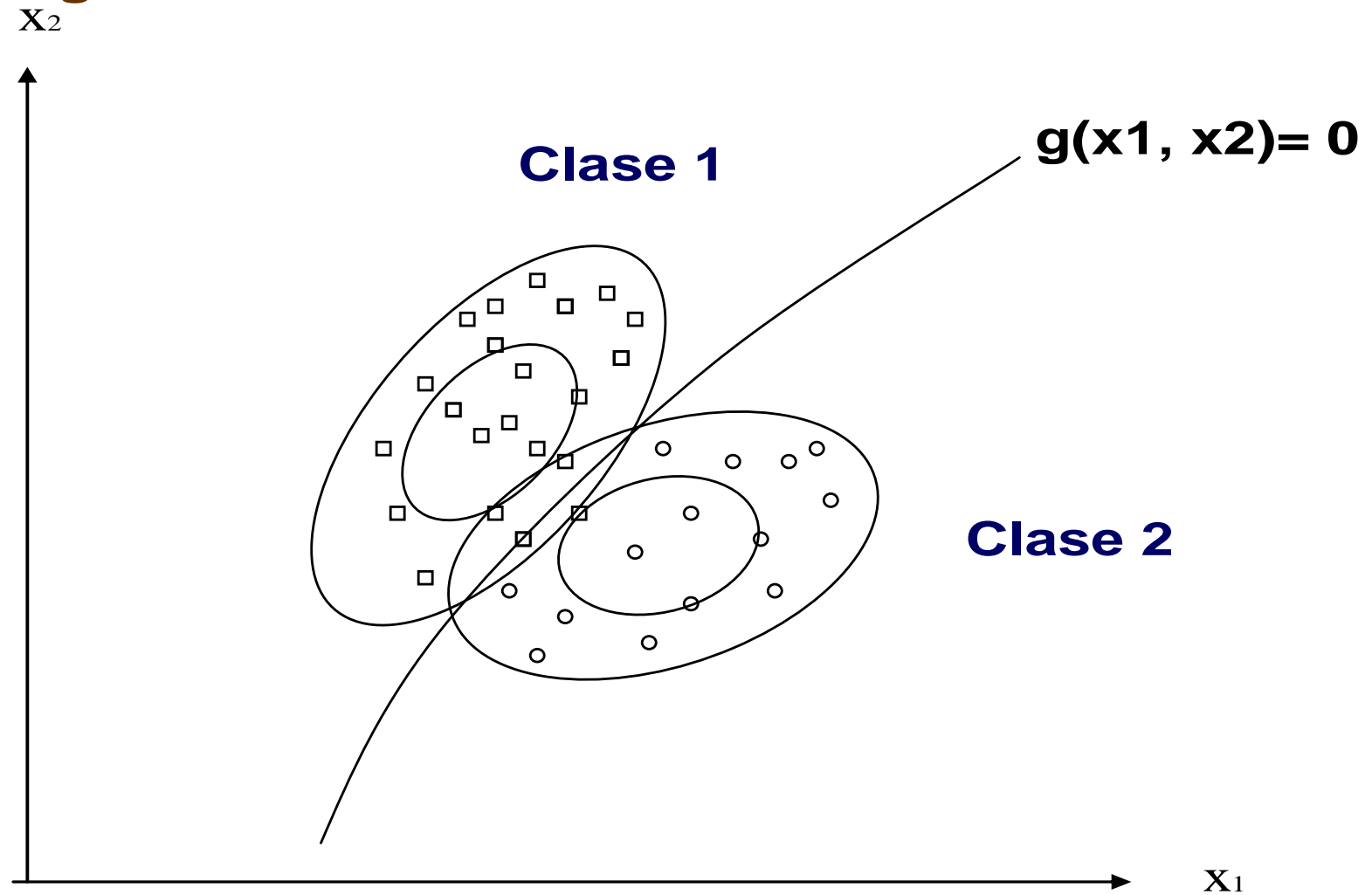
Si sabemos los parámetros de las dos distribuciones de probabilidad de \mathbf{x} , de experiencias pasadas, podemos definir una curva que sirva de límite entre ambas distribuciones, $g(\mathbf{x}_1, \mathbf{x}_2) = 0$. La cual divide el espacio bidimensional en dos regiones.



Análisis Discriminante



Figura 1 Clasificación con dos distribuciones





Análisis Discriminante



Una vez que la línea límite se selecciona, se puede clasificar una muestra o patrón sin una etiqueta de pertenencia a una clase (patrón del conjunto de generalización) en Clase 1 y Clase 2, dependiendo de si $g(x_1, x_2) > 0$ o $g(x_1, x_2) < 0$.

A esta función $g(x_1, x_2)$ se le denomina **función discriminante**, puesto que discrimina o reconoce la pertenencia, o no, de un patrón a una clase, a través del signo de g , a la cual también se le llama **clasificador, categorizador o reconocedor de patrones**.

$$h_{MAP} = \begin{cases} h_1 \text{ (Clase 1)} & \text{Si } g(\mathbf{x}) \geq 0, \text{ esto es,} \\ h_1 \text{ (Clase 1)} & \text{Si } \frac{P(h_1)P(D / h_1)}{P(D)} \geq \frac{P(h_2)P(D / h_2)}{P(D)} \\ h_2 \text{ (Clase 2)} & \text{Si } g(\mathbf{x}) < 0, \end{cases}$$



Análisis Discriminante



El algoritmo LDA (Linear Discriminant Analysis) es un método

usado en estadística, aprendizaje automático y reconocimiento de patrones para reducir la dimensión del espacio de datos y encontrar una combinación lineal de características que separen en dos o mas clases los objetos.

Este método toma en consideración todos los datos además de la distribución de las clases de estos datos.

Su objetivo es proyectar los datos, encontrando así la proyección optima, minimizando la distancia entre los objetos dentro de una misma clase y maximizando la distancia entre clases para conseguir la máxima discriminación a la hora de clasificar.



Análisis Discriminante



La Figura 2 muestra un diagrama de un clasificador en un espacio de características k -dimensional.

Por tanto para diseñar un clasificador se deberán de estudiar las características de la distribución de x para cada categoría y encontrar una función discriminante apropiada.

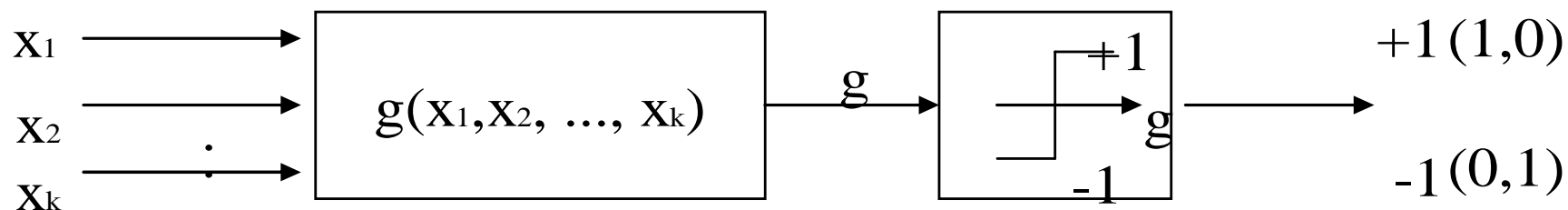
A este procedimiento se le llama **aprendizaje o entrenamiento**, y a las muestras utilizadas para diseñar el clasificador se les llama **conjunto de aprendizaje o de entrenamiento**.



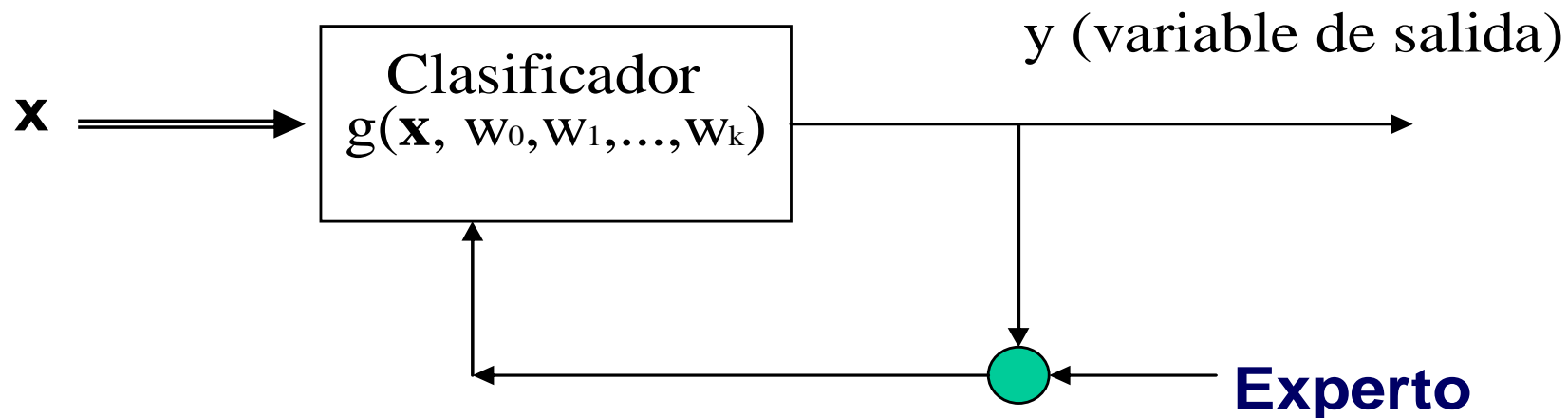
Análisis Discriminante



Figura 2 Diagrama de un clasificador binario



Clasificador



Algoritmo de aprendizaje



REGLA DE DECISIÓN BAYES PARA MINIMIZAR EL ERROR.



CONTRASTE BAYES. FRONTERAS DE DECISIÓN

Dadas dos clases, C_1 y C_2 , tenemos sus funciones discriminantes en las formas:

$$g_i(\mathbf{x}) = P(C_i)P(\mathbf{x} / C_i), \text{ para } i=1,2$$

La frontera de decisión es:

$$g_1(\mathbf{x}) = g_2(\mathbf{x})$$

Una regla de decisión basada solamente en probabilidades se describe en la forma:

$$\text{Si } g_1(\mathbf{x}) - g_2(\mathbf{x}) > 0 \text{ entonces } \mathbf{x} \in C_1$$

$$\text{Si } g_1(\mathbf{x}) - g_2(\mathbf{x}) < 0 \text{ entonces } \mathbf{x} \in C_2$$

$$\text{Si } P(C_1)P(\mathbf{x} / C_1) - P(C_2)P(\mathbf{x} / C_2) > 0 \text{ entonces } \mathbf{x} \in C_1$$

$$\text{Si } P(C_1)P(\mathbf{x} / C_1) - P(C_2)P(\mathbf{x} / C_2) < 0 \text{ entonces } \mathbf{x} \in C_2$$



REGLA DE DECISIÓN BAYES



Si tenemos en cuenta los valores de las funciones discriminantes

Entonces la regla de decisión del clasificador es

$$\text{Si } \frac{P(\mathbf{x} / C_1)}{P(\mathbf{x} / C_2)} - \frac{P(C_2)}{P(C_1)} > 0 \quad \text{ó} \quad \frac{P(\mathbf{x} / C_1)}{P(\mathbf{x} / C_2)} > \frac{P(C_2)}{P(C_1)} \quad \text{entonces } \mathbf{x} \in C_1$$

$$\text{Si } \frac{P(\mathbf{x} / C_1)}{P(\mathbf{x} / C_2)} - \frac{P(C_2)}{P(C_1)} < 0 \quad \text{ó} \quad \frac{P(\mathbf{x} / C_1)}{P(\mathbf{x} / C_2)} < \frac{P(C_2)}{P(C_1)} \quad \text{entonces } \mathbf{x} \in C_2$$

Al cociente $L(\mathbf{x}) = \frac{P(\mathbf{x} / C_1)}{P(\mathbf{x} / C_2)}$ se le llama **razón de verosimilitudes**, y la regla de decisión consiste en que la razón de verosimilitudes supere el umbral dado por $\frac{P(C_2)}{P(C_1)}$ para que \mathbf{x} pertenezca a C_1 o lo que es lo mismo se verifique la hipótesis h_1 .



LA REGLA DE DECISIÓN BAYES ES ENTONCES

$$\text{Si } \frac{P(\mathbf{x} / C_1)}{P(\mathbf{x} / C_2)} > \frac{P(C_2)}{P(C_1)} \text{ entonces } \mathbf{x} \in C_1$$

$$\text{Si } \frac{P(\mathbf{x} / C_1)}{P(\mathbf{x} / C_2)} < \frac{P(C_2)}{P(C_1)} \text{ entonces } \mathbf{x} \in C_2$$

Si tomamos el $-\ln$ de la razón de verosimilitudes tenemos

$$\text{Si } -\ln \frac{P(\mathbf{x} / C_1)}{P(\mathbf{x} / C_2)} < -\ln \frac{P(C_2)}{P(C_1)} \quad ; \text{ o lo que es igual}$$

$$\text{Si } -\ln P(\mathbf{x} / C_1) + \ln P(\mathbf{x} / C_2) < -\ln \frac{P(C_2)}{P(C_1)}$$

Entonces $\mathbf{x} \in C_1$



REGLA DE DECISIÓN BAYES



A veces, desde un punto de vista analítico, es mas conveniente trabajar con la función, $-\logaritmo$ de la razón de verosimilitudes, $H(X)$ o función de entropía, y de esta forma la regla de decisión ahora es que

$$H(\mathbf{x}) = -\ln(\mathbf{x}) = -\ln P(\mathbf{x} / C_1) + \ln P(\mathbf{x} / C_2) < -\ln \frac{P(C_2)}{P(C_1)}$$

Al término $H(x)$ se le llama la función discriminante. Si las probabilidades a priori de pertenencia a las dos clases son iguales, esto es

$$P(C_1) = P(C_2); \text{ entonces } \ln \frac{P(C_2)}{P(C_1)} = 0$$

Y las reglas de decisión son ahora

Si $-\ln P(\mathbf{x} / C_1) + \ln P(\mathbf{x} / C_2) < 0$ Entonces $\mathbf{x} \in C_1$

Si $-\ln P(\mathbf{x} / C_1) + \ln P(\mathbf{x} / C_2) > 0$ Entonces $\mathbf{x} \in C_2$

A estas reglas de decisión se les llama el contraste Bayes de mínimo error



REGLA DE DECISIÓN BAYES: Ejemplo para distribuciones normales



Si consideramos que la distribución de $P(\mathbf{x} / C_i)$, para $i=1,2$ es Normal con vector de medias \mathbf{m}_i y matriz de varianzas-covarianzas Σ_i , la regla de decisión del contraste de hipótesis se convierte en

$$H(\mathbf{x}) = -\ln \left[(2\pi)^{-n/2} |\Sigma_1|^{-1/2} \exp \left\{ -1/2 (\mathbf{x} - \mathbf{m}_1)^T \Sigma_1^{-1} (\mathbf{x} - \mathbf{m}_1) \right\} \right] + \\ \ln \left[(2\pi)^{-n/2} |\Sigma_2|^{-1/2} \exp \left\{ -1/2 (\mathbf{x} - \mathbf{m}_2)^T \Sigma_2^{-1} (\mathbf{x} - \mathbf{m}_2) \right\} \right]$$

ecuación que simplificando queda en la forma

$$H(\mathbf{x}) = (1/2)(\mathbf{x} - \mathbf{m}_1)^T \Sigma_1^{-1} (\mathbf{x} - \mathbf{m}_1) - (1/2)(\mathbf{x} - \mathbf{m}_2)^T \Sigma_2^{-1} (\mathbf{x} - \mathbf{m}_2) + \frac{1}{2} \ln \frac{|\Sigma_1|}{|\Sigma_2|}$$

Y la regla de decisión

$$\text{Si } H(\mathbf{x}) < \ln \frac{P(C_1)}{P(C_2)}, \text{ Entonces } \mathbf{x} \in C_1$$

$$\text{Si } H(\mathbf{x}) > \ln \frac{P(C_1)}{P(C_2)}, \text{ Entonces } \mathbf{x} \in C_2$$



REGLA DE DECISIÓN BAYES: Ejemplo para distribuciones normales



La ecuación se puede expresar en función de la distancia de Mahalanobis, al cuadrado, del vector \mathbf{x} a cada uno de los centroides o medias de clase, \mathbf{m}_1 y \mathbf{m}_2

$$H(\mathbf{x}) = (1/2)(\mathbf{x} - \mathbf{m}_1)^T \Sigma_1^{-1} (\mathbf{x} - \mathbf{m}_1) - (1/2)(\mathbf{x} - \mathbf{m}_2)^T \Sigma_2^{-1} (\mathbf{x} - \mathbf{m}_2) + \frac{1}{2} \ln \frac{|\Sigma_1|}{|\Sigma_2|}$$

$$H(\mathbf{x}) = (1/2) D_{\mathbf{x}, \mathbf{m}_1}^2 - (1/2) D_{\mathbf{x}, \mathbf{m}_2}^2 + \frac{1}{2} \ln \frac{|\Sigma_1|}{|\Sigma_2|}$$

Si las matrices de varianzas-covarianzas son iguales, la regla de decisión en función de las distancias de Mahalanobis es

$$\text{Si } D_{\mathbf{x}, \mathbf{m}_1}^2 < D_{\mathbf{x}, \mathbf{m}_2}^2 \text{ entonces } \mathbf{x} \in C_1$$



REGLA DE DECISIÓN BAYES: Ejemplo para distribuciones normales



Las inecuaciones anteriores muestran que la función discriminante viene dada por:

Una ecuación cuadrática si $\Sigma_1 \neq \Sigma_2$,

Una ecuación lineal si $\Sigma_1 = \Sigma_2 = \Sigma$

En este **segundo caso**, la ecuación del hiperplano es

$$(\mathbf{m}_2 - \mathbf{m}_1)^T \Sigma^{-1} \mathbf{x} + \frac{1}{2} (\mathbf{m}_1^T \Sigma^{-1} \mathbf{m}_1 - \mathbf{m}_2^T \Sigma^{-1} \mathbf{m}_2) - \ln \frac{P(C_1)}{P(C_2)}$$

Y la regla de decisión es para un patrón de vector de características \mathbf{x}

$$\text{Si } (\mathbf{m}_2 - \mathbf{m}_1)^T \Sigma^{-1} \mathbf{x} + \frac{1}{2} (\mathbf{m}_1^T \Sigma^{-1} \mathbf{m}_1 - \mathbf{m}_2^T \Sigma^{-1} \mathbf{m}_2) > \ln \frac{P(C_1)}{P(C_2)}$$

Entonces $\mathbf{x} \in C_1$



REGLA DE DECISIÓN BAYES: Ejemplo para distribuciones normales



Para dos distribuciones normales, la regla de decisión Bayes se puede expresar como una función cuadrática del vector \mathbf{x} de observaciones, pero si las matrices de covarianzas son iguales entonces la función es lineal y si además las matrices de covarianzas son la identidad \mathbf{I} , esto es, las variables son independientes, al estar incorreladas y ser normales, la ecuación es

$$(\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{x} + \frac{1}{2}(\mathbf{m}_1^T \mathbf{m}_1 - \mathbf{m}_2^T \mathbf{m}_2) - \ln \frac{P(C_1)}{P(C_2)} \begin{cases} > 0 & \mathbf{x} \in C_1 \\ < 0 & \mathbf{x} \in C_2 \end{cases}$$

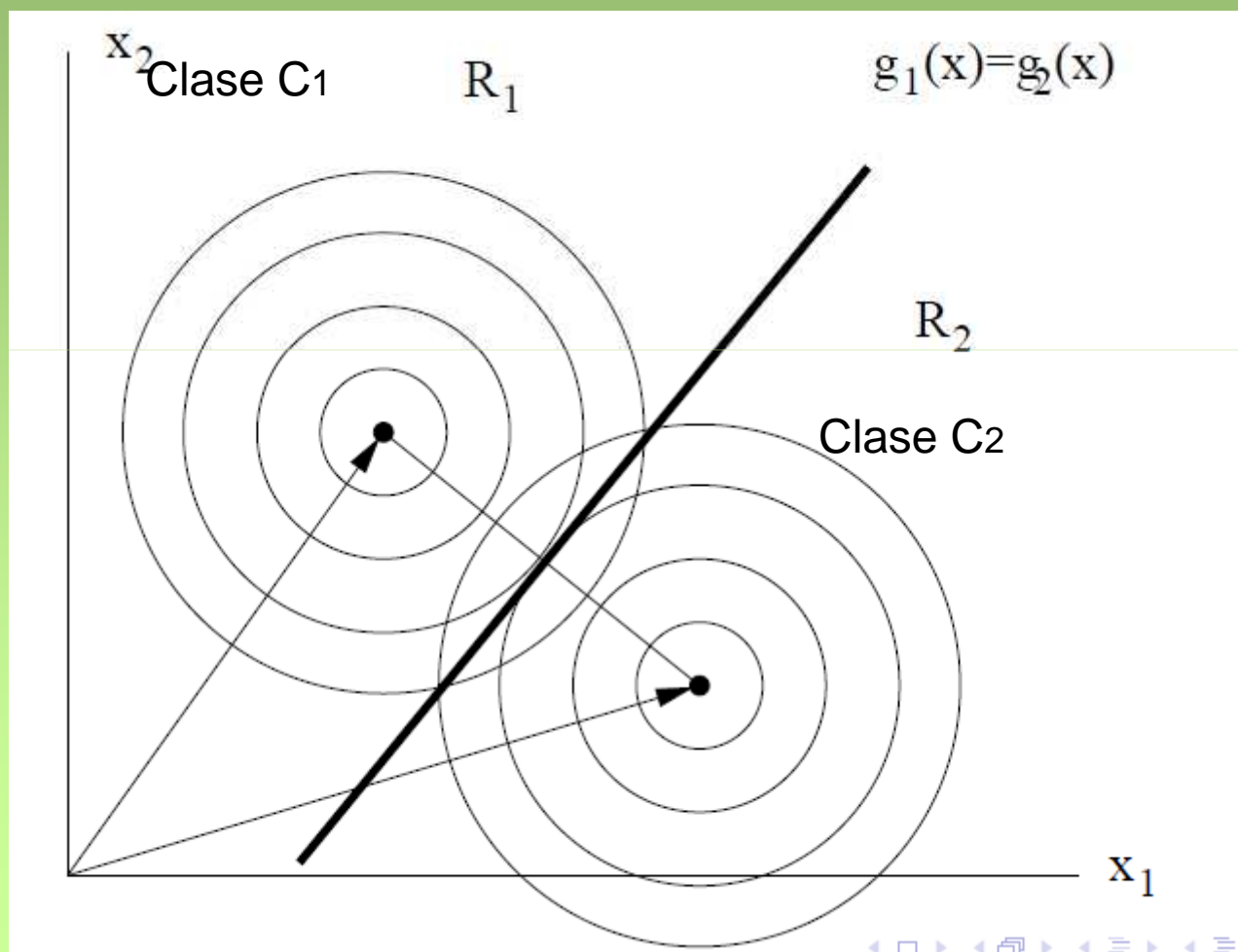


Clasificación lineal con dos distribuciones normales



Una ecuación lineal si $\Sigma_1 = \Sigma_2 = I$

Si $(\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{x} + \frac{1}{2}(\mathbf{m}_1^T \mathbf{m}_1 - \mathbf{m}_2^T \mathbf{m}_2) > 0$ entonces $\mathbf{x} \in C_1$





Bibliografía



Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth edition. Springer.

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press.



REGLA DE DECISIÓN BAYES



Ejemplo de Clasificador lineal Bayesiano para Pima

Como primer ejemplo de análisis discriminante vamos a considerar la clasificación de indias Pima, puesto que es uno de los ejemplos de prueba (benchmark). Es un conjunto de 768 muestras de un problema de diagnóstico en el que se utilizan ocho variables de clasificación propuestas por la OMS (Ver Tabla 1), de forma tal que mediante los ocho valores de las variables se clasifica a un individuo, mujer india de la tribu Pima de EEUU, como que padece diabetes, clase (1, 0), o no padece diabetes, clase (0, 1). La composición de la muestra es de 268 indias pertenecientes a la clase (1, 0) y 500 a la clase (0, 1)



REGLA DE DECISIÓN BAYES



Ejemplo de Clasificador lineal Bayesiano para Pima

Tabla 1

Variable	Rango
Número de embarazos (NE)	[0, 17]
Concentración de glucosa en plasma (CG)	[0, 199]
Presión diastólica de la sangre (PD)	[0, 122]
Grosor de los dobleces de la piel (GD)	[0, 99]
Nivel de insulina (NI)	[0, 846]
Índice de masa corporal (IC)	[0, 67.1]
Estimación de la influencia genética (IG)	[0.08, 2.42]
Edad (E)	[21, 81]



REGLA DE DECISIÓN BAYES



Variables introducidas/eliminadas^{a,b,c,d}

Paso	Introducidas	Lambda de Wilks							
		Estadístico	gl1	gl2	gl3	F exacta			
						Estadístico	gl1	gl2	Sig.
1	CGLUCOS	.780	1	1	574.000	162.069	1	574.000	.000
2	NEMBARA	.744	2	1	574.000	98.664	2	573.000	.000
3	IMASACOR	.712	3	1	574.000	76.994	3	572.000	.000
4	EINFLUGE	.703	4	1	574.000	60.199	4	571.000	.000

En cada paso se introduce la variable que minimiza la lambda de Wilks global.

- a. El número máximo de pasos es 16.
- b. La F parcial mínima para entrar es 3.84.
- c. La F parcial máxima para eliminar es 2.71
- d. El nivel de F, la tolerancia o el VIN son insuficientes para continuar los cálculos.

Variables en el análisis

Paso		Tolerancia	F para eliminar	Lambda de Wilks
1	CGLUCOSA	1.000	162.069	
2	CGLUCOSA	.999	147.009	.935
	NEMBARA	.999	27.716	.780
3	CGLUCOSA	.985	118.813	.860
	NEMBARA	.997	28.683	.748
	IMASACOR	.985	25.288	.744
4	CGLUCOSA	.982	112.052	.841
	NEMBARA	.994	29.951	.740
	IMASACOR	.979	22.655	.731
	EINFLUGE	.986	7.280	.712



REGLA DE DECISIÓN BAYES



Coefficientes estandarizados de las funciones discriminantes canónicas

	Función
	1
NEMBARA	.411
CGLUCOSA	.750
IMASACOR	.363
EINFLUGE	.207

Probabilidades previas para los grupos

CLASE	Previas	Casos utilizados en el análisis	
		No ponderados	Ponderados
nodiabetes	.634	365	365.000
diabetes	.366	211	211.000
Total	1.000	576	576.000

Resultados de la clasificación

a,b

				Grupo de pertenencia pronosticado		Total
CLASE				nodiabetes	diabetes	
Casos seleccionados	Original	Recuento	nodiabetes	316	49	365
			diabetes	89	122	211
	%		nodiabetes	86.6	13.4	100.0
			diabetes	42.2	57.8	100.0
Casos no seleccionados	Original	Recuento	nodiabetes	120	15	135
			diabetes	27	30	57
	%		nodiabetes	88.9	11.1	100.0
			diabetes	47.4	52.6	100.0

a. Clasificados correctamente el 76.0% de los casos agrupados originales seleccionados.

b. Clasificados correctamente el 78.1% de casos agrupados originales no seleccionados.



Ejemplo: Heart Disease Data Set

Esta base de datos contiene 76 atributos, pero todos los experimentos publicados se refieren al uso de un subconjunto de 14 de ellos. En particular, la base de datos Cleveland es la única que ha sido utilizada por investigadores de ML hasta la fecha. El campo "objetivo" se refiere a la presencia de la enfermedad cardíaca en el paciente. Esta codificada con un 0 (ninguna presencia) hasta 4.

Los experimentos con la base de datos de Cleveland se han concentrado en simplemente intentar distinguir presencia (valores 1,2,3,4) de ausencia (valor 0).



Ejemplo: Heart Disease Data Set



Attribute Information:

1. #3 (age) in years V2
2. #4 (sex) (1 = male; 0 = female) V3
3. #9 (cp) chest pain type
 - Value 1: typical angina
 - Value 2: atypical angina
 - Value 3: non-anginal pain
 - Value 4: asymptomatic V4
4. #10 (trestbps) resting blood pressure
(in mm Hg on admission to the hospital) V5
5. #12 (chol) serum cholestoral in mg/dl V6
6. #16 (fbs) (fasting blood sugar > 120 mg/dl)
(1 = true; 0 = false) V7
7. #19 (restecg) resting electrocardiographic results
 - Value 0: normal
 - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
 - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria V8



Ejemplo: Heart Disease Data Set



Attribute Information:

Ejemplo: Heart Disease Data Set

- 8. #32 (thalach) maximum heart rate achieved V9
- 9. #38 (exang) exercise induced angina (1 = yes; 0 = no) V10
- 10. #40 (oldpeak) ST depression induced by exercise relative to rest V11
- 11. #41 (slope) the slope of the peak exercise ST segment
 - Value 1: upsloping
 - Value 2: flat
 - Value 3: downsloping V12
- 12. #44 (ca) number of major vessels (0-3) colored by flourosopy V13
- 13. #51 (thal) 3 = normal; 6 = fixed defect; 7 = reversable defect V14
- 14. #58 (num) (the predicted attribute)
 - num: diagnosis of heart disease (angiographic disease status)
 - Value 0: < 50% diameter narrowing
 - Value 1: > 50% diameter narrowing V15



Ejemplo: Heart Disease Data Set



Visible: 19 de 19 variab

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	Dis_1	Dis1_1	Dis2_1
1	1	46	0	3	142	177	0	2	160	1	1,4	3	0	3	1	0	1	,16619	,83381
2	1	67	0	3	152	277	0	0	172	0	,0	1	1	3	1	0	1	,02943	,97057
3	1	56	1	4	125	249	1	2	144	1	1,2	2	1	3	0	1	0	,70821	,29179
4	1	34	0	2	118	210	0	0	192	0	,7	1	0	3	1	0	1	,00295	,99705
5	1	57	1	4	132	207	0	0	168	1	,0	1	0	7	1	0	0	,55473	,44527
6	1	64	1	4	145	212	0	2	132	0	2,0	2	2	6	0	1	0	,98599	,01401
7	1	59	1	4	138	271	0	2	182	0	,0	1	0	3	1	0	1	,09415	,90585
8	1	50	1	3	140	233	0	0	163	0	,6	2	1	7	0	1	0	,69513	,30487
9	1	51	1	1	125	213	0	2	125	1	1,4	1	1	3	1	0	1	,31596	,68404
10	1	54	1	2	192	283	0	2	195	0	,0	1	1	7	0	1	0	,65032	,34968
11	1	53	1	4	123	282	0	0	95	1	2,0	2	2	7	0	1	0	,99691	,00309
12	1	52	1	4	112	230	0	0	160	0	,0	1	1	3	0	1	1	,18097	,81903



Analizar

Marketing directo

Gráficos

Utilidades

Ventana

Ayuda

Informes

Estadísticos descriptivos

Tablas

Comparar medias

Modelo lineal general

Modelos lineales generalizados

Modelos mixtos

Correlaciones

Regresión

Loglineal

Redes neuronales

Clasificar

Reducción de dimensiones


Escala

Pruebas no paramétricas

Previsiones

Supervivencia

Respuesta múltiple

 Análisis de valores perdidos...

Imputación múltiple



V13

2

1

1

0

2

1

1

0

1

0

2

0

1

2

2

2



Clúster bietápico...



Clúster de K-medias...



Clúster jerárquico...



Árbol...



Discriminante...



Vecino más cercano...

2

0

2

0



Análisis discriminante

Variable de agrupación:
V15(0 1)
[Definir rango...](#)

Independientes:
V2
V3
V4
☒ Introducir independientes juntas
☐ Usar método de inclusión por pasos

Variable de selección:
V1=1 [Valor...](#)

[Estadísticos...](#)
[Método...](#)
[Clasificar...](#)
[Guardar...](#)
[Bootstrap...](#)

[Aceptar](#) [Pegar](#) [Restablecer](#) [Cancelar](#) [Ayuda](#)



Análisis discriminante: Clasificación



Probabilidades previas

- ☐ Todos los grupos iguales
- ☒ Calcular según tamaños de grupos

Usar matriz de covarianzas

- ☒ Intra-grupos
- ☐ Grupos separados

Visualización

- ☐ Resultados para cada caso
 - ☐ Limitar los casos a los primeros:
- ☒ Tabla de resumen
- ☐ Clasificación dejando uno fuera

Gráficos

- ☐ Grupos combinados
- ☐ Grupos separados
- ☒ Mapa territorial

- ☐ Reemplazar los valores perdidos con la media

Continuar

Cancelar

Ayuda



Resumen de funciones discriminantes canónicas



Autovalores

Función	Autovalor	% de varianza	% acumulado	Correlación canónica
1	1,217 ^a	100,0	100,0	,741

a. Se utilizaron las primeras 1 funciones discriminantes canónicas en el análisis.

Lambda de Wilks

Prueba de funciones	Lambda de Wilks	Chi-cuadrado	gl	Sig.
1	,451	153,231	13	,000

Coeficientes de función discriminante canónica estandarizadas

	Función 1
V2	-,130
V3	,280
V4	,355
V5	,194
V6	,029
V7	-,096
V8	,159
V9	-,231
V10	,193
V11	,192
V12	,063
V13	,541
V14	,385

Matriz de estructuras

	Función 1
V14	,559
V13	,474
V11	,401
V4	,395
V9	-,379
V10	,374
V3	,322
V12	,296
V5	,180
V8	,164
V2	,155
V6	,092
V7	,003

Correlaciones dentro de grupos combinados entre las variables discriminantes y las funciones discriminantes canónicas estandarizadas

Variables ordenadas por el tamaño absoluto de la correlación dentro de la función.

Resumen de funciones discriminantes canónicas



Funciones en centroides de grupo

	Función 1
V15	
0	1,219
1	-,988

Las funciones discriminantes canónicas sin estandarizar se han evaluado en medias de grupos



Estadísticas de clasificación



Probabilidades previas para grupos

V15	Previa	Casos utilizados en análisis	
		No ponderados	Ponderados
0	,448	90	90,000
1	,552	111	111,000
Total	1,000	201	201,000

Resultados de clasificación^{a,b}

				Pertenencia a grupos pronosticada		Total
				0	1	
V15						
Casos seleccionados	Original	Recuento	0	73	17	90
			1	10	101	111
		%	0	81,1	18,9	100,0
			1	9,0	91,0	100,0
Casos no seleccionados	Original	Recuento	0	24	6	30
			1	3	35	38
		%	0	80,0	20,0	100,0
			1	7,9	92,1	100,0

a. 86,6% de casos agrupados originales seleccionados clasificados correctamente.

b. 86,8% de casos agrupados originales sin seleccionar clasificados correctamente.



REGLA DE DECISIÓN BAYES PARA DISTRIBUCIONES DISCRETAS



Para toda clase C_i , $P(C_i | D) = \frac{|C_i|}{|D|}$ siendo el D el conjunto de datos de entrenamiento

Para toda posible instancia x Sea M_i el conjunto de todas las ocurrencias de x en C_i

El problema de la dimensionalidad:

Cada ejemplo x debe aparecer en C_i un numero suficientemente grande de veces como para obtener estadísticas significativas.

Si la dimensión de x crece, el numero de posibles valores de x crece exponencialmente, haciendo el problema intratable

¿Que ocurre si el nuevo ejemplo a clasificar, x, no se había dado en D?



Ejemplo para caso discreto: Jugar a tenis

Se contacta con el Departamento de Investigaciones Meteorológicas de Springfield para obtener un registro del tiempo durante las actuaciones previamente programadas, y el centro de entretenimiento nos informa de como de apropiado es el tiempo para poder jugar o no hoy a tenis.

De esta manera tenemos cuatro atributos de entrada, describiendo las condiciones meteorológicas y un atributo de salida asociado a la clase, señalando las decisiones tomadas por el director acerca de si jugar o no.



REGLA DE DECISIÓN BAYES



Ejemplo para caso discreto

Tenemos una muestra de tamaño 14 acerca de las acciones realizadas en el pasado por el director, de jugar o no a tenis, en función de diferentes condiciones meteorológicas.

No tenemos mucho tiempo para entender todo esto, así que lo que queremos es simplemente una forma sencilla de predecir si el director se presentará esta noche a jugar.



REGLA DE DECISIÓN BAYES

Nuevo ejemplo



Outlook	Temp.	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Overcast: Cubierto, encapotado; Sunny: Soleado; Rainy: Lluvioso

Hot: Caliente, Mild: Leve o Moderada; Cool: Fría



REGLA DE DECISIÓN BAYES



Consulta: $\mathbf{x}=(\text{outlook}=\text{sunny}, \text{Temperature}=\text{cool}, \text{Humidity}=\text{high}, \text{Wind}=\text{strong})$

Clasificador Bayesiano:

$$\begin{aligned} h_{MAP} &= \arg \max_{C_1, C_2} P(\mathbf{x} | h) P(h) = \\ &= \arg \max_{C_1, C_2} (P(x | C_1) P(C_1), P(x | C_2) P(C_2)) = \\ &= \arg \max_{C_1, C_2} (?? \times \frac{9}{14}; ?? \times \frac{5}{14}) \end{aligned}$$

Probabilidades a priori:

$$P(C1 \text{ o yes}) = 9/14, P(C2 \text{ o no}) = 5/14$$

Probabilidades a posteriori:

$$P(< \text{sunny}; \text{cool}; \text{high}; \text{strong} > | C1) = ??$$

$$P(< \text{sunny}; \text{cool}; \text{high}; \text{strong} > | C2) = ??$$

No podemos tomar decisiones



INTRODUCCIÓN A LOS MODELOS COMPUTACIONALES:

CUARTO CURSO DEL GRADO DE ING. INFORMÁTICA EN COMPUTACION

Análisis Discriminante Lineal

GRACIAS POR SU ATENCIÓN

César Hervás-Martínez
Grupo de Investigación AYRNA

**Departamento de Informática y Análisis
Numérico**
Universidad de Córdoba
Campus de Rabanales. Edificio Einstein.
Email: chervas@uco.es

2018-2019



Ejercicios



Ejercicio 1.- Se desea atribuir un cuadro a dos pintores. Para ello se tienen dos características o variables del cuadro: La profundidad del trazo en diferentes cuadros del mismo pintor y la proporción que ocupan los cuadros de los pintores sobre la superficie del lienzo. Los valores medios de estas variables son, para el primer pintor P1, (2, 0,8) y para el segundo P2, (2,3 y 0,7). Las desviaciones típicas de estas dos variables son 0,5 y 0,1 y la correlación entre estas medidas es de 0,5. La obra a atribuir tiene como valores de estas variables (2,1 y 0,75). Calcular la regla de decisión y las probabilidades de error.

Solución.- Calculamos las distancias de Mahalanobis de x a m_1 , para P1

$$D_{x,m_1}^2 = (x - m_1)^T \Sigma^{-1} (x - m_1) \quad \text{y para P2} \quad D_{x,m_2}^2 = (x - m_2)^T \Sigma^{-1} (x - m_2)$$

$$D_{P1}^2 = (2.10 - 2.00, 0.75 - 0.80) \begin{pmatrix} 0.25 & 0.025 \\ 0.025 & 0.01 \end{pmatrix}^{-1} \begin{pmatrix} 2.10 - 2.00 \\ 0.75 - 0.80 \end{pmatrix} = 0.52$$

$$D_{P2}^2 = (2.10 - 2.30, 0.75 - 0.70) \begin{pmatrix} 0.25 & 0.025 \\ 0.025 & 0.01 \end{pmatrix}^{-1} \begin{pmatrix} 2.10 - 2.30 \\ 0.75 - 0.70 \end{pmatrix} = 0.81$$

Como $D_{P1}^2 < D_{P2}^2$ esto es, $0.52 < 0.81$, asignamos el cuadro al primer pintor, P1



Ejercicios



El error esperado de clasificación con esta regla de decisión depende de la diferencia de distancias de Mahalanobis entre medias, que es la varianza de la variable z .

$$z = \mathbf{w}^T \mathbf{x}, \text{ siendo } \mathbf{w} = \Sigma^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

$$D^2 = (\mathbf{m}_1 - \mathbf{m}_2)^T \Sigma^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

$$D^2 = (2.00 - 2.30, 0.80 - 0.70) \begin{pmatrix} 0.25 & 0.025 \\ 0.025 & 0.01 \end{pmatrix}^{-1} \begin{pmatrix} 2.00 - 2.30 \\ 0.80 - 0.70 \end{pmatrix} = 2.61$$

Luego $D = 1.62$. La probabilidad de equivocarse es

$$P(P_1 / P_2) = 1 - \phi\left(\frac{1.62}{2}\right) = 1 - \phi(0.81) = 1 - 0.79 = 0.21$$

Siendo ϕ , la función de distribución de una $N(0,1)$. De manera que la clasificación mediante el estudio de estas variables no es muy precisa dado que tenemos un porcentaje de error del 21%.



Ejercicios



Si calculamos la probabilidad a posteriori de que el cuadro pertenezca al pintor P_1 suponiendo que a priori ambos tienen las mismas probabilidades de ser el autor, tenemos

$$\begin{aligned} P(P_1 / \mathbf{x}) &= \frac{1}{1 + \exp(-\frac{1}{2}(D_{P_2}^2 - D_{P_1}^2))} = \frac{1}{1 + \exp(-\frac{1}{2}(0.81 - 0.52))} = \\ &= \frac{1}{1.86} = 0.54 \end{aligned}$$

Esta probabilidad nos indica que al clasificar la obra como perteneciente al pintor P_1 existe mucha incertidumbre en la decisión, ya que las probabilidades de pertenencia a cada uno de los pintores son muy semejantes, 0.54 frente a $1 - 0.54 = 0.46$



INTRODUCCIÓN A LOS MODELOS COMPUTACIONALES:

CUARTO CURSO DEL GRADO DE ING. INFORMÁTICA EN COMPUTACION

Análisis Discriminante Lineal

GRACIAS POR SU ATENCIÓN

César Hervás-Martínez
Grupo de Investigación AYRNA

**Departamento de Informática y Análisis
Numérico**
Universidad de Córdoba
Campus de Rabanales. Edificio Einstein.
Email: chervas@uco.es

2018-2019



APRENDIZAJE AUTOMÁTICO: TERCER CURSO DEL GRADO DE ING. INFORMÁTICA EN COMPUTACION

Aprendizaje estadístico: Teoría de la Información

César Hervás-Martínez
Grupo de Investigación AYRNA

**Departamento de Informática y Análisis
Numérico**
Universidad de Córdoba
Campus de Rabanales. Edificio Einstein.
Email: chervas@uco.es

2018-2019



INTRODUCCION

TEORIA DE LA INFORMACION ESTADISTICA

1 Cantidad de información

2 Entropía de una variable

3 Divergencia de Kullback–Leibler

4 Cantidad de Información Mútua



CANTIDAD DE INFORMACIÓN

Sea una urna con 9 bolas negras y 1 bola blanca. Se efectúan extracciones sin reemplazamiento y sean $A=\{\text{sacar una bola blanca}\}$ y $B=\{\text{sacar una bola negra}\}$ dos sucesos con probabilidades $P(A)=1/10$ y $P(B)=9/10$

Se saca una bola blanca. El suceso A proporciona una alta información, ya que la incertidumbre sobre la siguiente extracción desaparece, puesto que $P(B/A)=9/9=1$ y $P(A/A)=0$

Se saca una bola negra. El suceso B proporciona una información pequeña, ya que la incertidumbre acerca de la siguiente extracción se mantiene puesto que $P(A/B)=1/9$ y $P(B/B)=8/9$



Cantidad de información como medida de reducción de la incertidumbre



Al lanzar un dado si nos dicen que ha salido:
un numero menor que 2, suceso A, tenemos mas información (reduce mas la incertidumbre) que si nos dicen que ha salido un numero múltiplo de 2, suceso B, puesto que si las probabilidades de los sucesos que salga la puntuación i-ésima son equiprobables, esto es $P(E_i)=1/6$, entonces

$P(E1/A)=1$, $P(E2/A)=0$, $P(E3/A)=0$, $P(E4/A)=0$, $P(E5/A)=0$, $P(E6/A)=0$;

mientras que

$P(E1/B)=0$, $P(E2/B)=1/3$, $P(E3/B)=0$, $P(E4/B)=1/3$, $P(E5/B)=0$,
 $P(E6/B)=1/3$



CANTIDAD DE INFORMACIÓN

Sea X una variable aleatoria con posibles valores x_1, \dots, x_n y probabilidades asociadas $p(x_1), \dots, p(x_n)$, definimos la **Cantidad de Información** como

$$I(x_i) = -\log_2 p(x_i)$$

Si $p(x_i) \cong 1$, entonces $I(x_i) \cong 0$

Si $p(x_i) \cong 0$, entonces $I(x_i) \cong \infty$

Cuanto mas probable es un suceso menor cantidad de información aporta



ENTROPÍA DE UNA VARIABLE



Sea X una variable aleatoria con posibles valores x_1, \dots, x_n y probabilidades asociadas $p(x_1), \dots, p(x_n)$, definimos

Y sea $I(X)$ la variable aleatoria cantidad de información asociada a X , con posibles valores $I(x_1), \dots, I(x_n)$ y probabilidades asociadas $p(x_1), \dots, p(x_n)$.

Se define la **entropía de Shannon** (1948), $H(X)$, de una variable aleatoria discreta X como la esperanza matemática de la variable aleatoria asociada $I(X)$

$$H(X) = E(I(X)) = -\sum_{i=1}^n p(x_i) \cdot \log_2 p(x_i)$$

Si $p(x_i) = 0$, la indeterminación $p(x_i) \cdot \log_2 p(x_i)$ se resuelve asignándole el valor 0



ENTROPÍA DE UNA VARIABLE



Sea X variable aleatoria de Bernoulli de parámetro p , $B(p)$

La función de probabilidad es

$$P(X = x) = p^x (1 - p)^{1-x} \quad \text{para } x \in \{0, 1\}$$

$$H(X) = E(l(X)) = -\sum_{i=1}^n p(x_i) \cdot \log_2 p(x_i)$$

$$H(X) = -p \log_2 p - (1 - p) \log_2 (1 - p)$$

Si $p = 0,50$; $H(X) = -0,5 \log_2 0,5 - 0,5 \log_2 0,5 = 1$

Si $p = 0,60$; $H(X) = -0,6 \log_2 0,6 - 0,4 \log_2 0,4 = 0,97$

Si $p = 0,90$ urna con 9 bolas negras y 1 blanca

$$H(X) = -0,9 \log_2 0,9 - 0,1 \log_2 0,1 = 0,468$$



ENTROPÍA DE UNA VARIABLE

Se verifica:

$$0 \leq H(X) \leq \log_2 n$$

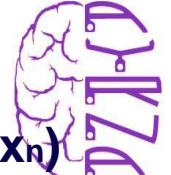
$$H(X) = 0 \Leftrightarrow \exists x_i \text{ con } p(x_i) = 1$$

Esto significa que la variable aleatoria es singular y toma un solo valor con probabilidad 1. Por otra parte

**Si X es variable aleatoria uniforme discreta, es decir
 $P(X = x_i) = 1/n$ para todo $i = 1, \dots, n$, entonces
 $H(X) = \log_2 n$**



ENTROPÍA CONDICIONADA



Sea X una v.a. con valores x_1, \dots, x_n y con probabilidades $p(x_1), \dots, p(x_n)$

Sea Y una v.a. con valores y_1, \dots, y_m y con probabilidades $p(y_1), \dots, p(y_m)$

Definimos (X, Y) como una v.a. bidimensional con $(x_1, y_1), \dots, (x_1, y_m), \dots, (x_n, y_1), \dots, (x_n, y_m)$ y con probabilidades $p(x_1, y_1), \dots, p(x_1, y_m), \dots, p(x_n, y_1), \dots, p(x_n, y_m)$

Definimos $X|Y = y_j$ como una v.a. condicionada con $p(x_1|y_j), \dots, p(x_n|y_j)$

Entonces la Entropía de la v.a. bidimensional conjunta (X, Y) es:

$$H(X, Y) = - \sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log_2 p(x_i, y_j)$$

Y la Entropía de la v.a. X condicionada al valor $Y = y_j$

$$H(X | Y = y_j) = - \sum_{i=1}^n p(x_i | y_j) \log_2 p(x_i | y_j)$$

Y la Entropía de la v.a. X condicionada a la v.a. Y

$$H(X | Y) = \sum_{j=1}^m p(y_j) H(X | Y = y_j) = - \sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log_2 p(x_i | y_j)$$



ENTROPÍA DE UNA VARIABLE



Ley de entropías totales: $H(X,Y) = H(X) + H(Y|X)$

$$H(X) + H(Y | X) = -\sum_{i=1}^n p(x_i) \log_2 p(x_i) - \sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log_2 p(y_j | x_i)$$

Ahora

$$\begin{aligned} -\sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log_2 p(y_j | x_i) &= -\sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log_2 \frac{p(x_i, y_j)}{p(x_i)} \\ &= -\sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log_2 p(x_i, y_j) + \sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log_2 p(x_i) \end{aligned}$$

pero

$$\sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log_2 p(x_i) = \sum_{i=1}^n p(x_i) \log_2 p(x_i)$$

luego

$$\begin{aligned} H(X) + H(Y | X) &= -\sum_{i=1}^n p(x_i) \log_2 p(x_i) - \sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log_2 p(x_i, y_j) \\ &\quad + \sum_{i=1}^n p(x_i) \log_2 p(x_i) = -\sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log_2 p(x_i, y_j) = H(X, Y) \end{aligned}$$



ENTROPÍA DE UNA VARIABLE



Si X e Y son variables aleatorias independientes, esto es, si $p(x_i, y_j) = p_1(x_i) p_2(y_j)$ entonces:

$$H(X|Y) = H(X)$$

$$H(Y|X) = H(Y)$$

$$H(X, Y) = H(X) + H(Y)$$



DIVERGENCIA DE KULLBACK–LEIBLER

Mide la distancia entre dos distribuciones de probabilidad – una de las cuales actúa como referencia , por ejemplo p es la probabilidad “a priori” y q la probabilidad “a posteriori”– definidas sobre la misma variable aleatoria X , se denomina también “entropía relativa” o “entropía cruzada”. Se puede interpretar como el incremento de información necesaria para para cambiar la distribución “a priori” p en una distribución “a posteriori” q . Para una variable discreta x .

$$KL(p \parallel q) = D_{K-L}(p, q) = \sum_{i=1}^n q(x_i) \log_2 \frac{q(x_i)}{p(x_i)}$$



DIVERGENCIA DE KULLBACK–LEIBLER

Se supone que para cualquier x ,
si $p(x)=0$ entonces $q(x)=0$
y también que $0 \cdot \log(0/p)=0$.

Además

$$D_{K-L}(p, q) \geq 0$$

$$D_{K-L}(p, q) = 0 \Leftrightarrow p(x_i) = q(x_i), \forall i=1, \dots, n$$

Ejemplo.- Cálculo de la media de la divergencia de K-L
para un conjunto de clasificadores (ensemble)

Supongamos que $DP(x)$ es la distribución a priori sobre un
conjunto de etiquetas de clase Ω y utilizamos $d_{i,j}(x)$
como el estimador de la probabilidad



CANTIDAD DE INFORMACIÓN MÚTUA

En teoría de la probabilidad, y en teoría de la información, la información mutua de dos v. a.s. X , Y es una cantidad que mide la dependencia mutua de las dos variables, es decir, mide la reducción de la incertidumbre (entropía) de una variable aleatoria, X , debido al conocimiento del valor de otra variable aleatoria Y .

$$I(X,Y) = H(X) - H(X|Y)$$



CANTIDAD DE INFORMACIÓN MÚTUA



Ejercicio.-

Consideremos dos monedas: La A en la cual la probabilidad de cara es $1/2$, y la B en la cual la probabilidad de cara es igual a 1. Se elige una moneda al azar, se lanza dos veces y se anota el numero de caras obtenidas.

Definimos dos variables aleatorias.

X denota la moneda escogida, con valores A con probabilidad $1/2$ y B con probabilidad $1/2$

Y denota el número de caras obtenidas, con valores 1 (sacar X en A y F en B) y 2 (sacar F en A y F en B) y con probabilidades $1/2 \times 1$ y $1/2 \times 1$

$$H(X) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1; \text{ y de igual manera } H(Y)=1$$

$$H(X | Y) = -\sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log_2 p(x_i | y_j)$$

$$I(X, Y) = H(X) - H(X|Y) = 1 - 0,4509 (\text{¿¿??}) = 0,5491$$



CANTIDAD DE INFORMACIÓN MÚTUA



$$H(X | Y) = \sum_{j=1}^m p(y_j) H(X | Y = y_j) = - \sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log_2 p(x_i | y_j)$$

$$P(Y = 2|X = A) = 1/4 ; P(Y = 1|X = A) = 1/2 ; P(Y = 0|X = A) = 1/4;$$

$$P(Y = 2|X = B) = 1; P(Y = 1|X = B) = 0; P(Y = 0|X = B) = 0$$

$$P(X = A, Y = 0) = 1/8; P(X = B, Y = 0) = 0; P(X = A, Y = 1) = 1/4 ;$$

$$P(X = B, Y = 1) = 0; P(X = A, Y = 2) = 1/8 ; P(X = B, Y = 2) = 1/2$$

$$P(Y = 0) = 1/8; P(Y = 1) = 1/4; P(Y = 2) = 5/8$$

$$P(X = A|Y = 0) = 1; P(X = B|Y = 0) = 0; P(X = A|Y = 1) = 1;$$

$$P(X = B|Y = 1) = 0; P(X = A|Y = 2) = 1/5 ; P(X = B|Y = 2) = 4/5$$

$$H(X|Y = 2) = -P(X|Y=2) \cdot \log_2 P(X|Y=2) = -(1/5)(4/5) \cdot \log_2(4/25)$$

$$H(X|Y = 2) = 0,7215; H(X|Y = 1) = 0; H(X|Y = 0) = 0$$

$$H(X|Y) = P(Y = 0) \cdot H(X|Y = 0) + P(Y = 1) \cdot H(X|Y = 1) + P(Y = 2) \cdot H(X|Y = 2) \\ = 5/8 \cdot 0,7215 = 0,4509$$

$$I(X, Y) = H(X) - H(X|Y) = 1 - 0,4509 = 0,5491$$



CANTIDAD DE INFORMACIÓN MÚTUA

$$I(X, Y) = H(X) - H(X|Y)$$

$$I(X, Y) = -\sum_{i=1}^n p(x_i) \log_2 p(x_i) + \sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log_2 p(x_i | y_j)$$

Pero la segunda sumatoria es

$$\begin{aligned} -\sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log_2 p(y_j | x_i) &= -\sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log_2 \frac{p(x_i, y_j)}{p(y_j)} \\ &= -\sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log_2 p(x_i, y_j) + \sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log_2 p(y_j) \end{aligned}$$

luego

$$\begin{aligned} I(X, Y) &= \sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \left[\log_2 p(x_i | y_j) - (\log_2 p(x_i) + \log_2 p(y_j)) \right] \\ &= \sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log_2 \frac{p(x_i, y_j)}{p(x_i) p(y_j)} \end{aligned}$$



CANTIDAD DE INFORMACIÓN MÚTUA

En el caso continuo, reemplazamos la suma con una integral definida doble :

$$I(X;Y) = \iint_{Y,X} p(x, y) \log\left(\frac{p(x, y)}{P(x)p(y)}\right) dx dy$$



CANTIDAD DE INFORMACIÓN MÚTUA



Se verifica:

$$I(X, Y) = I(Y, X)$$

$$I(X, Y) = D_{K-L}(p(x, y), p(x) \cdot p(y))$$

$$\begin{aligned} I(X, Y | Z) &= \sum_{k=1}^r p(z_k) I(X, Y | Z = z_k) = \\ &= \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^r p(x_i, y_j, z_k) \log_2 \frac{p(x_i, y_j | z_k)}{p(x_i | z_k) p(y_j | z_k)} \end{aligned}$$

$$I(X, Y | Z) = H(X | Z) + H(Y | Z) - H(X, Y | Z)$$

$I(X, Y | Z) = 0$. Sii X e Y son condicionalmente independientes dado Z

X e Y son condicionalmente independientes dado Z Sii $p(x|y, z) = p(x|z)$ para todo x, y, z



APRENDIZAJE AUTOMÁTICO: TERCER CURSO DEL GRADO DE ING. INFORMÁTICA EN COMPUTACION

Aprendizaje estadístico: Teoría de la Información

GRACIAS POR SU ATENCIÓN

César Hervás-Martínez
Grupo de Investigación AYRNA

Departamento de Informática y Análisis
Numérico
Universidad de Córdoba
Campus de Rabanales. Edificio Einstein.
Email: chervas@uco.es

2018-2019