Sistemas Inteligentes

Escuela Técnica Superior de Informática

Universitat Politècnica de València

Tema B2T1
Razonamiento probabilístico

Índice

Índice

Introducción

Considérese la acción y pregunta siguientes:

$A_t =$ SALIR AL AEROPUERTO t MINUTOS ANTES DEL VUELO

¿ME PERMITE A_{90} LLEGAR A TIEMPO?

Es difícil decidir a partir de una respuesta del tipo:

A_{90} ME PERMITE LLEGAR A TIEMPO
SI NO HAY ATASCOS Y NO HAY PINCHAZOS Y . . .
MUCHAS OTRAS CONDICIONES DIFÍCILES DE GARANTIZAR

Es más fácil decidir a partir de una respuesta probabilística:

$P(A_{90}$ ME PERMITE LLEGAR A TIEMPO$) = 0.9999$

Índice

Representación probabilística

El conocimiento se representa mediante la distribución de probabilidad conjunta de las variables aleatorias de interés.

Ejemplo del dentista: conocimiento para diagnosticar caries

Variables aleatorias de interés:

$$Dolor: \quad D \in \{0,1\}$$

$$Caries: \quad C \in \{0,1\}$$

$$Hueco: \quad H \in \{0,1\}$$

Representación:

$$P(D=d, C=c, H=h)$$

d	c	h	P
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108
	Suma:		1.000

Nota: $\sum_x P(x) = 1$ y $0 \leq P(x) \leq 1$

Índice

Inferencia probabilística

A partir de la distribución conjunta podemos calcular la probabilidad de cualquier *suceso* (*proposición*) mediante aplicación de:

La regla suma:

$$P(x) = \sum_y P(x, y)$$

La regla producto:

$$P(x, y) = P(y \mid x)\, P(x)$$

Inferencia probabilística: ejemplo del dentista

Probabilidad de observar caries y hueco (a la vez):

$$P(c=1,h=1)=\sum_{d=0,1} P(d,c=1,h=1)=0.180$$

d	c	h	P
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

Probabilidad de observar hueco:

$$P(h=1)=\sum_{d=0,1}\sum_{c=0,1} P(d,c,h=1)=0.200$$

d	c	h	P
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

Probabilidad de observar caries tras observar hueco:

$$P(c=1\mid h=1)=\frac{P(c=1,h=1)}{P(h=1)}=\frac{0.180}{0.200}=0.900$$

Índice

Independencia probabilística

Decimos que x e y son ***independientes*** si:

$$P(x,y) = P(x)\,P(y) \quad \text{ó} \quad P(x \mid y) = P(x) \quad \text{ó} \quad P(y \mid x) = P(y)$$

La independencia de variables puede establecerse por conocimiento experto.

Ejemplo del dentista: $Tiempo:\ T \in \{sol, nubes, lluvia, nieve\}$

$$P(t,d,c,h) = P(t)\,P(d,c,h)$$

Reducimos el número de probabilidades a almacenar:

32 probabilidades vs $4+8$ probabilidades

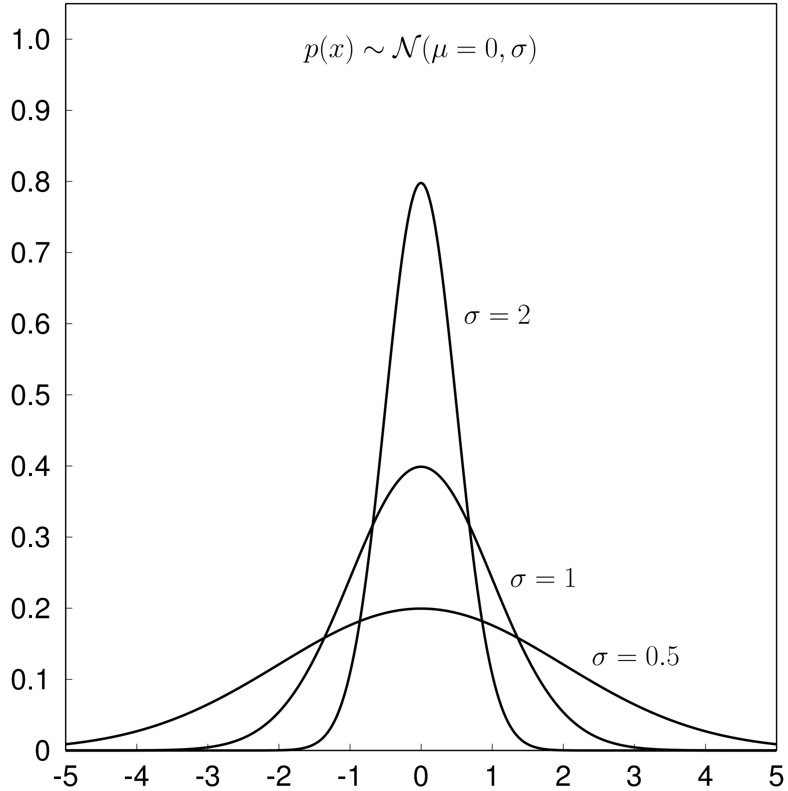
Índice

Variables continuas

También se suelen emplear variables continuas caracterizadas mediante funciones de *densidad de probabilidad:*

$$p(x) \geq 0 \quad \text{para todo } x \qquad \text{y} \qquad \int p(x)\, dx = 1$$

Ejemplo: la distribución normal



$$p(x) \sim \mathcal{N}(\mu, \sigma^2)$$

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\,\sigma^2}\right)$$

$$P(x \in [\mu \pm 2\sigma]) = 0.95$$

Índice

Teorema de Bayes

El *teorema de Bayes* permite actualizar nuestro conocimiento sobre una hipótesis y tras la observación de nueva evidencia x:

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

Ejemplo del dentista: comprobemos la hipótesis $c = 1$

d	c	h	P
0	0	0	0.576
0	0	1	0.008
0	1	0	0.144
0	1	1	0.072
1	0	0	0.064
1	0	1	0.012
1	1	0	0.016
1	1	1	0.108

Probabilidad de observar caries:

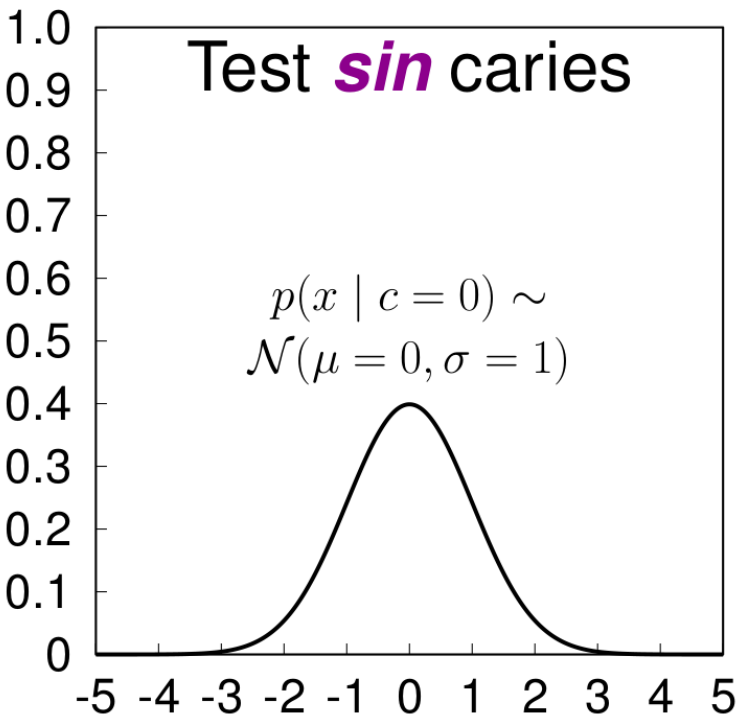
$$P(c = 1) = \sum_{d=0,1} \sum_{h=0,1} P(d, c = 1, h) = 0.340$$

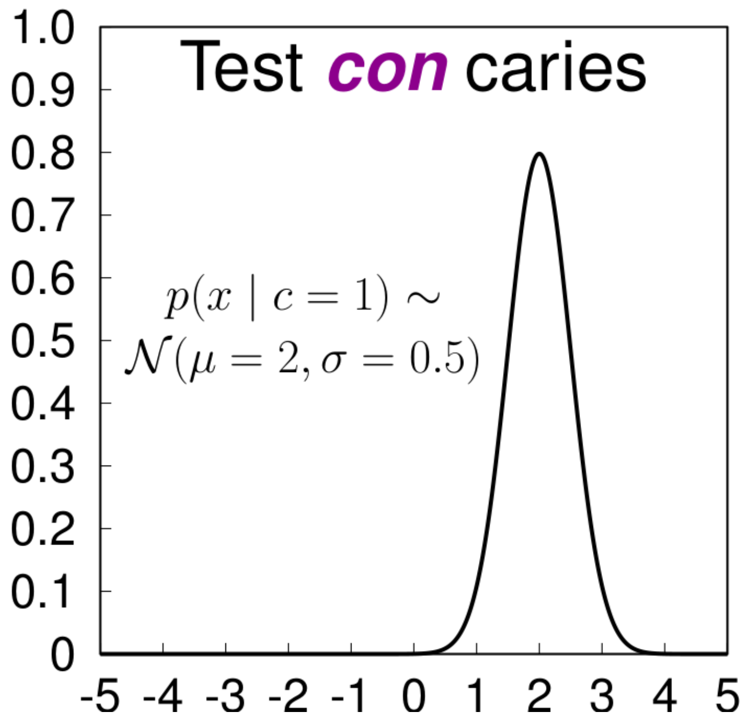
Anteriomente inferimos que, si se observa hueco:

$$P(c = 1 \mid h = 1) = P(c = 1)\frac{P(h = 1 \mid c = 1)}{P(h = 1)} = 0.900$$

Ejemplo del dentista (cont.):

Sea x una variable continua que mide el resultado de un test de saliva para el diagnóstico de caries. Se tiene:





Si realizamos el test y obtenemos $x=2$, por Bayes tenemos:

$$P(c=1 \mid x=2) = P(c=1)\,\frac{p(x=2 \mid c=1)}{p(x=2)} = 0.340\,\frac{0.798}{0.307} = 0.884$$

Ejercicio

Problema de clasificación de flores en 3 clases (c={Setosa, Versicolor, Virgínica}) en base al número de pétalos $y = \{3, 4, 5\}$.

Ejemplos

c	y
Setosa	3
Setosa	3
Setosa	4
Setosa	4
Setosa	5
Versicolor	3
Versicolor	3
Versicolor	3
Versicolor	3
Versicolor	4
Virgínica	4
Virgínica	4
Virgínica	4
Virgínica	5
Virgínica	5

c	Setosa	Versicolor	Virgínica
$P(c)$			

Prob. cond.	$P(y = 3 \mid c)$	$P(y = 4 \mid c)$	$P(y = 5 \mid c)$
Setosa			
Versicolor			
Virgínica			

Prob. a posteriori	$P(c \mid y = 3)$	$P(c \mid y = 4)$	$P(c \mid y = 5)$
Setosa			
Versicolor			
Virgínica			

Solución Ejercicio

Problema de clasificación de flores en 3 clases (c={Setosa, Versicolor, Virgínica}) en base al número de pétalos $y = \{3, 4, 5\}$.

Ejemplos

c	y
Setosa	3
Setosa	3
Setosa	4
Setosa	4
Setosa	5
Versicolor	3
Versicolor	3
Versicolor	3
Versicolor	3
Versicolor	4
Virgínica	4
Virgínica	4
Virgínica	4
Virgínica	5
Virgínica	5

c	Setosa	Versicolor	Virgínica
$P(c)$	5/15	5/15	5/15

Prob. cond.	$P(y = 3 \mid c)$	$P(y = 4 \mid c)$	$P(y = 5 \mid c)$
Setosa	2/5	2/5	1/5
Versicolor	4/5	1/5	0
Virgínica	0	3/5	2/5

Prob. a posteriori	$P(c \mid y = 3)$	$P(c \mid y = 4)$	$P(c \mid y = 5)$
Setosa	0.33	0.33	0.33
Versicolor	0.67	0.17	0.00
Virgínica	0.00	0.50	0.67

- Dada una flor desconocida de 4 pétalos, ¿a qué clase pertenece?
- ¿Cuál es la decisión de clasificación d de mínimo error dado que $y = 4$?

Índice

Teoría de la decisión

- *Simplificación*: Coste 0 (decisión *acertada*) ó 1 (decisión *errónea*).

Probabilidad de error si se toma la decisión d dado el evento y:

$$P_d(\text{error} \mid y) = 1 - P(d \mid y)$$

Teoría de la decisión

- *Simplificación*: Coste 0 (decisión *acertada*) ó 1 (decisión *errónea*).

Probabilidad de error si se toma la decisión d dado el evento y:

$$P_d(\text{error} \mid y) = 1 - P(d \mid y)$$

Prob. a posteriori	$P(c \mid y = 3)$	$P(c \mid y = 4)$	$P(c \mid y = 5)$
Setosa	0.33	0.33	0.33
Versicolor	0.67	0.17	0.00
Virgínica	0.00	0.50	0.67

y	3	4	5
$P_{\text{seto}}(\text{error} \mid y)$			
$P_{\text{vers}}(\text{error} \mid y)$			
$P_{\text{virg}}(\text{error} \mid y)$			

Teoría de la decisión

- *Simplificación*: Coste 0 (decisión *acertada*) ó 1 (decisión *errónea*).

Probabilidad de error si se toma la decisión d dado el evento y:

$$P_d(\text{error} \mid y) = 1 - P(d \mid y)$$

Prob. a posteriori	$P(c \mid y = 3)$	$P(c \mid y = 4)$	$P(c \mid y = 5)$
Setosa	0.33	0.33	0.33
Versicolor	0.67	0.17	0.00
Virgínica	0.00	0.50	0.67

y	3	4	5
$P_{\text{seto}}(\text{error} \mid y)$	0.67	0.67	0.67
$P_{\text{vers}}(\text{error} \mid y)$	0.33	0.83	1.00
$P_{\text{virg}}(\text{error} \mid y)$	1.00	0.50	0.33

Teoría de la decisión

- *Simplificación*: Coste 0 (decisión *acertada*) ó 1 (decisión *errónea*).

Probabilidad de error si se toma la decisión d dado el evento y:

$$P_d(\text{error} \mid y) = 1 - P(d \mid y)$$

Teoría de la decisión

- *Simplificación*: Coste 0 (decisión *acertada*) ó 1 (decisión *errónea*).

Probabilidad de error si se toma la decisión d dado el evento y:

$$P_d(\text{error} \mid y) = 1 - P(d \mid y)$$

Minimización de la probabilidad de error para cada y:

$$P(\text{error} \mid y) = \min_{d \in \mathcal{D}} P_d(\text{error} \mid y) \;=\; 1 - \max_{d \in \mathcal{D}} P(d \mid y) \quad \forall y \in \mathcal{Y}$$

Es decir, para cada y la probabilidad de error se minimiza si se toma la decisión de mayor probabilidad (a posteriori).

Teoría de la decisión

- *Simplificación*: Coste 0 (decisión *acertada*) ó 1 (decisión *errónea*).

Probabilidad de error si se toma la decisión d dado el evento y:

$$P_d(\text{error} \mid y) = 1 - P(d \mid y)$$

Minimización de la probabilidad de error para cada y:

$$P(\text{error} \mid y) = \min_{d \in \mathcal{D}} P_d(\text{error} \mid y) \;=\; 1 - \max_{d \in \mathcal{D}} P(d \mid y) \quad \forall y \in \mathcal{Y}$$

Es decir, para cada y la probabilidad de error se minimiza si se toma la decisión de mayor probabilidad (a posteriori).

y	3	4	5
$P_{\text{seto}}(\text{error} \mid y)$	0.67	0.67	0.67
$P_{\text{vers}}(\text{error} \mid y)$	0.33	0.83	1.00
$P_{\text{virg}}(\text{error} \mid y)$	1.00	0.50	0.33
$P(\text{error} \mid y)$	0.33	0.50	0.33

Teoría de la decisión

- *Simplificación*: Coste 0 (decisión *acertada*) ó 1 (decisión *errónea*).

Probabilidad de error si se toma la decisión d dado el evento y:

$$P_d(\text{error} \mid y) = 1 - P(d \mid y)$$

Minimización de la probabilidad de error para cada y:

$$P(\text{error} \mid y) = \min_{d \in \mathcal{D}} P_d(\text{error} \mid y) \;=\; 1 - \max_{d \in \mathcal{D}} P(d \mid y) \quad \forall y \in \mathcal{Y}$$

Teoría de la decisión

- *Simplificación*: Coste 0 (decisión *acertada*) ó 1 (decisión *errónea*).

Probabilidad de error si se toma la decisión d dado el evento y:

$$P_d(\text{error} \mid y) = 1 - P(d \mid y)$$

Minimización de la probabilidad de error para cada y:

$$P(\text{error} \mid y) = \min_{d \in \mathcal{D}} P_d(\text{error} \mid y) \;=\; 1 - \max_{d \in \mathcal{D}} P(d \mid y) \quad \forall y \in \mathcal{Y}$$

Regla de decisión de mínimo riesgo de error o de Bayes:

$$\hat{d}(y) = \operatorname*{argmax}_{d \in \mathcal{D}} P(d \mid y) \quad \forall y \in \mathcal{Y}$$

Teoría de la decisión

- *Simplificación*: Coste 0 (decisión *acertada*) ó 1 (decisión *errónea*).

Probabilidad de error si se toma la decisión d dado el evento y:

$$P_d(\text{error} \mid y) = 1 - P(d \mid y)$$

Minimización de la probabilidad de error para cada y:

$$P(\text{error} \mid y) = \min_{d \in \mathcal{D}} P_d(\text{error} \mid y) \;=\; 1 - \max_{d \in \mathcal{D}} P(d \mid y) \quad \forall y \in \mathcal{Y}$$

Regla de decisión de mínimo riesgo de error o de Bayes:

$$\hat{d}(y) = \operatorname*{argmax}_{d \in \mathcal{D}} P(d \mid y) \quad \forall y \in \mathcal{Y}$$

Prob. a posteriori	$P(c \mid y = 3)$	$P(c \mid y = 4)$	$P(c \mid y = 5)$
Setosa	0.33	0.33	0.33
Versicolor	0.67	0.17	0.00
Virgínica	0.00	0.50	0.67

y	3	4	5
$\hat{d}(y)$	Vers	Virg	Virg

Teoría de la decisión

- *Simplificación*: Coste 0 (decisión *acertada*) ó 1 (decisión *errónea*).

Probabilidad de error si se toma la decisión d dado el evento y:

$$P_d(\text{error} \mid y) = 1 - P(d \mid y)$$

Minimización de la probabilidad de error para cada y:

$$\forall y \in \mathcal{Y}: \quad P(\text{error} \mid y) = \min_{d \in \mathcal{D}} P_d(\text{error} \mid y) \; = \; 1 - \max_{d \in \mathcal{D}} P(d \mid y)$$

Regla de decisión de mínimo riesgo de error o de Bayes:

$$\hat{d}(y) = \operatorname*{argmax}_{d \in \mathcal{D}} P(d \mid y) \quad \forall y \in \mathcal{Y}$$

Mínimo riesgo global:

$$P(\text{error}) = \sum_{y \in \mathcal{Y}} P(\text{error}, y) = \sum_{y \in \mathcal{Y}} P(\text{error} \mid y)\, P(y)$$

Solución Ejercicio

Problema de clasificación de flores en 3 clases (c={Setosa, Versicolor, Virgínica}) en base al número de pétalos $y = \{3, 4, 5\}$.

Ejemplos

c	y
Setosa	3
Setosa	3
Setosa	4
Setosa	4
Setosa	5
Versicolor	3
Versicolor	3
Versicolor	3
Versicolor	3
Versicolor	4
Virgínica	4
Virgínica	4
Virgínica	4
Virgínica	5
Virgínica	5

c	Setosa	Versicolor	Virgínica
$P(c)$	5/15	5/15	5/15

Prob. cond.	$P(y=3 \mid c)$	$P(y=4 \mid c)$	$P(y=5 \mid c)$
Setosa	2/5	2/5	1/5
Versicolor	4/5	1/5	0
Virgínica	0	3/5	2/5

Prob. a posteriori	$P(c \mid y=3)$	$P(c \mid y=4)$	$P(c \mid y=5)$
Setosa	0.33	0.33	0.33
Versicolor	0.67	0.17	0.00
Virgínica	0.00	0.50	0.67

y	3	4	5
$P_{\text{seto}}(\text{error} \mid y)$	0.67	0.67	0.67
$P_{\text{vers}}(\text{error} \mid y)$	0.33	0.83	1.00
$P_{\text{virg}}(\text{error} \mid y)$	1.00	0.50	0.33
$P(\text{error} \mid y)$	0.33	0.50	0.33
$\hat{d}(y)$	Vers	Virg	Virg
$P(y)$	0.40	0.40	0.20
$P(\text{error})$		0.40 (40 %)	

Ejercicio

Un problema clásico de decisión consiste en clasificar flores de la familia *Iris* en tres clases; *setosa, versicolor* y *virgínica*, en base a los tamaños de sus pétalos y sépalos (y).

Para ello se han calculado sendos histogramas de las superficies de los pétalos de una muestra de 50 flores de cada clase. Normalizando estos histogramas, se ha estimado la siguiente distribución de tamaños de pétalos para cada clase (c):

	tamaño de los pétalos en cm^2											
$P(y \mid c)$	$<$1	1	2	3	4	5	6	7	8	9	10	$>$10
Setosa	0.90	0.10	0	0	0	0	0	0	0	0	0	0
Versicolor	0	0	0	0.20	0.30	0.32	0.12	0.06	0	0	0	0
Virgínica	0	0	0	0	0	0	0.08	0.12	0.24	0.14	0.20	0.22

Asumiendo que las clases son equiprobables, calcular:

a) Las probabilidades a posteriory $P(c \mid y),\ c \in \{\text{SETO}, \text{VERS}, \text{VIRG}\}$, para una flor cuyo tamaño de pétalos es $y =$7 cm^2

b) La decisión óptima de clasificación de esta flor y la probabilidad de que dicha decisión sea errónea

c) La mejor decisión y la correspondiente probab. de error para tamaños de pétalos $1, 2, \ldots, 10$ cm^2

d) La mínima probabilidad de error de decisión esperada para cualquier flor Iris; es decir, $P(\text{error})$

e) Repetir los calculos anteriores, asumiendo que las probabilidades a priori son:
$P(\text{SETO}) = 0.3,\ P(\text{VERS}) = 0.5,\ P(\text{VIRG}) = 0.2$

Algunas soluciones: a) 0.0, 0.33, 0.67; b) Virgínica, 0.33; d) 0.05 (5 %) e.a) 0.0, 0.55, 0.44; e.b) Versicolor, 0.44; e.d) 0.04 (4 %)

Solución

Las clases son equiprobables $P(C = \textsc{Seto}) = P(C = \textsc{Vers}) = P(C = \textsc{Virg}) = \frac{1}{3}$.

a) Aplicamos las regla de Bayes:

$$\begin{aligned}
P(c \mid Y = 7) &= \frac{P(Y = 7 \mid c)\, P(c)}{P(Y = 7)} \\
&= \frac{P(Y = 7 \mid c)\, P(c)}{\sum_{c \in \{\textsc{Seto},\textsc{Vers},\textsc{Virg}\}} P(Y = 7, c)} \\
&= \frac{P(Y = 7 \mid c)\, P(c)}{\sum_{c \in \{\textsc{Seto},\textsc{Vers},\textsc{Virg}\}} P(Y = 7 \mid c) P(c)}
\end{aligned}$$

$$P(C = \textsc{Seto} \mid Y = 7) = \frac{P(Y = 7 \mid C = \textsc{Seto})\, P(C = \textsc{Seto})}{\sum_{c \in \{\textsc{Seto},\textsc{Vers},\textsc{Virg}\}} P(Y = 7 \mid c) P(c)}$$

Solución

b) La decisión óptima sería la regla de decisión de Bayes:

$$\hat{d}(y) = \operatorname*{argmax}_{d \in \mathcal{D}} P(d \mid y) \quad \forall y \in \mathcal{Y}$$

aplicado a nuestro caso sería:

$$\hat{c}(y) = \operatorname*{argmax}_{c \in \mathcal{C}} P(c \mid Y = 7)$$
$$= \operatorname*{argmax}_{c \in \{\text{SETO}, \text{VERS}, \text{VIRG}\}} P(c \mid Y = 7)$$

La probabilidad de error se define como:

$$P(\text{error} \mid y) = 1 - \max_{d \in \mathcal{D}} P(d \mid y) \quad \forall y \in \mathcal{Y}$$

en nuestro caso:

$$P(\text{error} \mid Y = 7) = 1 - \max_{c \in \{\text{SETO}, \text{VERS}, \text{VIRG}\}} P(c \mid Y = 7)$$

Solución

c) Repetimos los cálculos realizados para $Y = 7$ para el resto de valores de y. Existe un caso particular cuando $Y = 2$, ya que

$$P(Y = 2 \mid c) = 0 \quad \forall c$$

En este caso cualquier decisión (clase) es óptima y la probabilidad de error es cero.

d) Aplicamos directamente la definición de probabilidad de error:

$$P(\text{error}) = \sum_{y \in \mathcal{Y}} P(\text{error} \mid y)\, P(y)$$

donde

$$\begin{aligned} P(y) &= \sum_{c \in \{\text{SETO}, \text{VERS}, \text{VIRG}\}} P(y, c) \\ &= \sum_{c \in \{\text{SETO}, \text{VERS}, \text{VIRG}\}} P(y \mid c)\, P(c) \end{aligned}$$

Índice

Bibliografía

[1] A.N. Abdallah. The Logic of Partial Information. Springer Verlag, 1995.

[2] B.G. Buchanan, E.H. Shortliffe (editires): Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. Addison Wesley, 1984. (También en http://aitopics.net/RuleBasedExpertSystems).

[3] J.F. Baldwin. Fuzzy sets and expert systems. Wiley, 1985.

[4] R.O. Duda, D.G. Stork, P.E. Hart. Pattern Classification. Wiley, 2001.

[5] S. Russell, P. Norvig. Artificial Intelligence: A Modern Approach. Pearson, third edition, 2010.

El material de este tema se basa principalmente en [5].