

Aprendizaje bayesiano

Índice

- Aprendizaje con redes bayesianas
 - Aprendizaje paramétrico con datos completos
 - Aprendizaje paramétrico con datos incompletos. Algoritmo EM
- El clasificador Naive Bayes
- Validación
 - El problema del sobreajuste
 - Métodos de validación
 - Medidas de rendimiento
 - La matriz de confusión
 - Medidas numéricas
 - La curva ROC

Aprendizaje con redes bayesianas

El aprendizaje con redes bayesianas consiste en, a partir de una base de datos preexistente, construir de modo automático una red bayesiana que lo represente.

En el aprendizaje con redes bayesianas, se distinguen dos tipos de situaciones:

- *Aprendizaje paramétrico.* Se supone dada una base de datos de casos, y la estructura de la red (nodos y enlaces). A partir de los datos y la estructura, se aprenden los parámetros (probabilidades a priori de los nodos sin padres, probabilidades condicionadas de cada nodo dados sus padres).
- *Aprendizaje estructural.* En este caso se parte de una base de datos de casos y se aprende de ellos tanto los parámetros como la estructura. Se puede no poner ningún tipo de restricción a la estructura o también partir de información previa que prohíba o fuerce algunos enlaces.

Función de verosimilitud

Llamamos **modelo** y lo representamos por $\vec{\theta}$ a una asignación de valores para las probabilidades que se quiere estimar.

Dado un conjunto \mathcal{O} de n observaciones O_1, \dots, O_n , y un modelo $\vec{\theta}$, definimos **la verosimilitud del modelo** dadas las observaciones como:

$$L(\vec{\theta} / \vartheta) = P(O_1, \dots, O_n / \vec{\theta}) = \prod_i P(O_i / \vec{\theta})$$

Es posible demostrar que el modelo de máxima verosimilitud se puede calcular asignando a cada parámetro la frecuencia relativa calculada a partir de las n observaciones.

En la práctica, se maximiza el logaritmo, o su promedio:

$$l(\vec{\theta} / \vartheta) = \log(L(\vec{\theta} / \vartheta)) = \sum_i \log(P(O_i / \vec{\theta}))$$

$$\hat{l}(\vec{\theta} / \vartheta) = \frac{1}{n} \sum_i \log(P(O_i / \vec{\theta}))$$

Las tres funciones alcanzan sus máximos en los mismos valores

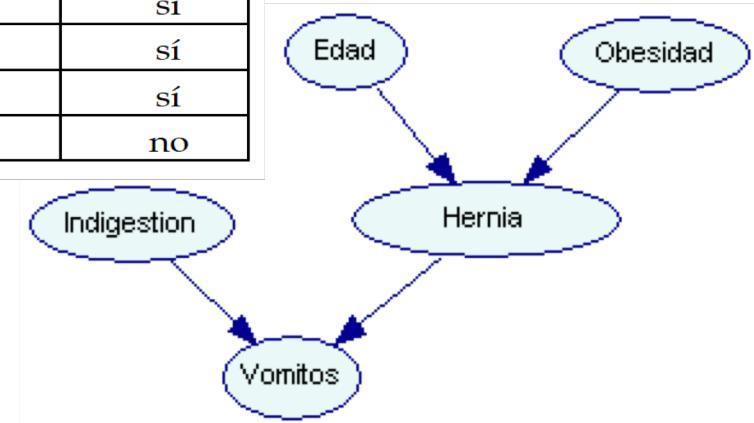
Además, como $0 \leq L \leq 1$, se tiene que l y $\hat{l} \leq 0$

Ejemplo: aprendizaje paramétrico datos completos

Individuos	Edad	Obesidad	Hernia	Indigestión	Vómitos
Individuo 1	Mayor_50	no	no	no	no
Individuo 2	Mayor_50	no	no	no	no
Individuo 3	Mayor_50	no	no	no	no
Individuo 4	Mayor_50	no	no	no	no
Individuo 5	Mayor_50	no	sí	no	sí
Individuo 6	Mayor_50	sí	sí	no	sí
Individuo 7	Mayor_50	sí	sí	no	sí
Individuo 8	Mayor_50	sí	sí	no	sí
Individuo 9	Menor_50	no	no	no	no
Individuo 10	Menor_50	no	no	no	no
Individuo 11	Menor_50	no	no	no	no
Individuo 12	Menor_50	no	no	no	no
Individuo 13	Menor_50	no	no	no	no
Individuo 14	Menor_50	no	no	no	no
Individuo 15	Menor_50	no	no	no	no
Individuo 16	Menor_50	no	no	no	no
Individuo 17	Menor_50	no	no	sí	sí
Individuo 18	Menor_50	sí	no	no	sí
Individuo 19	Menor_50	sí	no	sí	sí
Individuo 20	Menor_50	sí	no	no	no

Ejemplo: aprendizaje paramétrico datos completos

Individuos	Edad	Obesidad	Hernia	Indigestión	Vómitos
Individuo 1	Mayor_50	no	no	no	no
Individuo 2	Mayor_50	no	no	no	no
Individuo 3	Mayor_50	no	no	no	no
Individuo 4	Mayor_50	no	no	no	no
Individuo 5	Mayor_50	no	sí	no	sí
Individuo 6	Mayor_50	sí	sí	no	sí
Individuo 7	Mayor_50	sí	sí	no	sí
Individuo 8	Mayor_50	sí	sí	no	si
Individuo 9	Menor_50	no	no	no	no
Individuo 10	Menor_50	no	no	no	no
Individuo 11	Menor_50	no	no	no	no
Individuo 12	Menor_50	no	no	no	no
Individuo 13	Menor_50	no	no	no	no
Individuo 14	Menor_50	no	no	no	no
Individuo 15	Menor_50	no	no	no	no
Individuo 16	Menor_50	no	no	no	no
Individuo 17	Menor_50	no	no	sí	sí
Individuo 18	Menor_50	sí	no	no	sí
Individuo 19	Menor_50	sí	no	sí	sí
Individuo 20	Menor_50	sí	no	no	no



Ejemplo: aprendizaje paramétrico datos completos

Las probabilidades que debemos aprender son:

- $P(\text{edad})$, $P(\text{obesidad})$, $P(\text{indigestión})$, (3 parámetros)
- $P(\text{vómitos}/\text{indigestión}, \text{hernia})$, (4 parámetros)
- $P(\text{hernia}/\text{edad}, \text{obesidad})$ (4 parámetros)

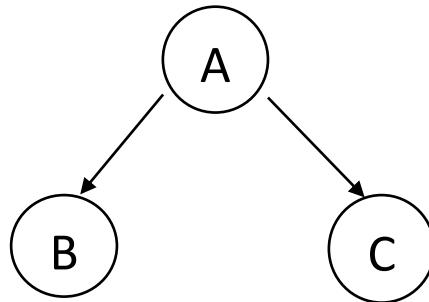
Son en total 11 valores. A modo de ejemplo, calculemos dos de ellos:

$$P(\text{edad}=menor \text{ de } 50) = \frac{n^o \text{ de casos (edad} = \text{Menor_50})}{n^o \text{ total de casos}} = 12/20 = 0,6$$

$$P(\text{vómitos}=no/\text{indigestión}=no, \text{hernia}=no) = \frac{n^o \text{ casos (vómitos}=no,\text{indigestión}=no,hernia=no)}{n^o \text{ casos (indigestión}=no,hernia=no)} = 13 / 14 = 0.92857$$

Aprendizaje paramétrico con datos incompletos

Alternativa 1. Eliminar la observación incompleta



	A	B	C
O1	+a	+b	+c
O2	+a	+b	¬c
O3	+a	¬b	¬c
O4	¬a	+b	¬c
O5	¬a	¬b	+c
O6	+a	¬b	

$$\begin{aligned}\theta_1 &= P(+a) = 3/5 \\ \theta_2 &= P(+b/+a) = 2/3 \\ \theta_3 &= P(+b/¬a) = 1/2 \\ \theta_4 &= P(+c/+a) = 1/3 \\ \theta_5 &= P(+c/¬a) = 1/2\end{aligned}$$

$$P(O1 / \vec{\theta}) = P(+a) * P(+b/+a) * P(+c/+a) = (3/5) * (2/3) * (1/3) = 0,1333$$

$$P(O2 / \vec{\theta}) = P(+a) * P(+b/+a) * P(¬c/+a) = (3/5) * (2/3) * (2/3) = 0,2667$$

$$P(O3 / \vec{\theta}) = P(+a) * P(¬b/+a) * P(¬c/+a) = (3/5) * (1/3) * (2/3) = 0,1333$$

$$P(O4 / \vec{\theta}) = P(¬a) * P(+b/¬a) * P(¬c/¬a) = (2/5) * (1/2) * (1/2) = 0,1$$

$$P(O5 / \vec{\theta}) = P(¬a) * P(¬b/¬a) * P(+c/¬a) = (2/5) * (1/2) * (1/2) = 0,1$$

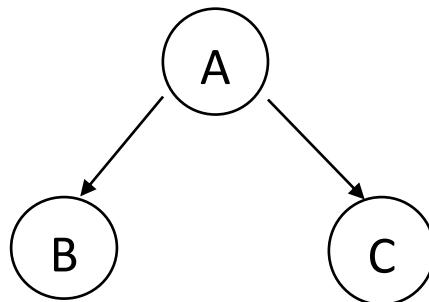
$$P(O6 / \vec{\theta}) = P(+a) * P(¬b/+a) * P(¬c/+a) + P(+a) * P(¬b/+a) * P(+c/+a) = 0,1333 + (3/5) * (1/3) * (1/3) = 0,2$$

Se tiene un valor de $p = -1,927694$

Esta medida indica la bondad de este modelo y permitirá compararlo con otros

Aprendizaje paramétrico con datos incompletos

Alternativa 2. Sustituir el valor incompleto por la moda



	A	B	C
O1	+a	+b	+c
O2	+a	+b	¬c
O3	+a	¬b	¬c
O4	¬a	+b	¬c
O5	¬a	¬b	+c
O6	+a	¬b	¬c

$$P(01 / \vec{\theta}) = P(+a) * P(+b / +a) * P(+c / +a) = 0,083$$

$$P(02 / \vec{\theta}) = P(+a) * P(+b / +a) * P(\neg c / +a) = 0,25$$

$$P(03 / \vec{\theta}) = P(+a) * P(\neg b / +a) * P(\neg c / +a) = 0,25$$

$$P(04 / \vec{\theta}) = P(\neg a) * P(+b / \neg a) * P(\neg c / \neg a) = 0,083$$

$$P(05 / \vec{\theta}) = P(\neg a) * P(\neg b / \neg a) * P(+c / \neg a) = 0,083$$

$$P(06 / \vec{\theta}) = P(+a) * P(\neg b / +a) * P(\neg c / +a) = 0,25$$

Se tiene un valor de $p = -1,935605$.

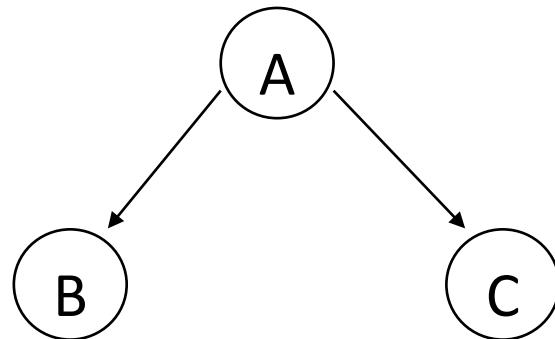
Al ser p menor que el anterior, este modelo es peor que el de la alternativa 1

$$\begin{aligned}\theta_1 &= P(+a) = 4/6 \\ \theta_2 &= P(+b / +a) = 1/2 \\ \theta_3 &= P(+b / \neg a) = 1/2 \\ \theta_4 &= P(+c / +a) = 1/4 \\ \theta_5 &= P(+c / \neg a) = 1/2\end{aligned}$$

Aprendizaje paramétrico con datos incompletos

Alternativa 3. Vamos a aplicar el algoritmo EM para estimar el modelo de máxima verosimilitud para el problema que estábamos considerando

Red:



	A	B	C
O1	+a	+b	+c
O2	+a	+b	-c
O3	+a	-b	-c
O4	-a	+b	-c
O5	-a	-b	+c
O6	+a	-b	

Partimos de un modelo arbitrario:

$$\theta_{0,1} = P(+a) = 0.5$$

En este caso $p = -1,9639$

$$\theta_{0,2} = P(+b/+a) = 0.5$$

$$\theta_{0,3} = P(+b/\neg a) = 0.5$$

Esta estimación $\vec{\theta}_0$ del modelo es peor que la obtenida en las alternativas 1 y 2.

$$\theta_{0,4} = P(+c/(+a) = 0.5$$

$$\theta_{0,5} = P(+c/(\neg a)= 0.5$$

Aprendizaje paramétrico con datos incompletos

Fase expectation:

Vamos a calcular suponiendo que el modelo es $\vec{\theta}_0$, cual sería el valor esperado de la variable C en la observación incompleta:

$$P(+c/+a, \neg b) = P(+c/+a) = 0.5$$

$$P(\neg c/+a, \neg b) = 1 - P(+c/+a) = 0.5$$

A	B	C
+a	+b	+c
+a	+b	$\neg c$
+a	$\neg b$	$\neg c$
$\neg a$	+b	$\neg c$
$\neg a$	$\neg b$	+c
+a	$\neg b$	$P(+c/+a, \neg b) = 0.5$
		$P(\neg c/+a, \neg b) = 0.5$

Aprendizaje paramétrico con datos incompletos

Fase maximization:

A partir de este conjunto de observaciones calculamos el nuevo modelo:

$$\theta_{1,1} = P(+a) = 4/6 = 0,6667$$

$$\theta_{1,2} = P(+b/+a) = 2/4 = 0,5$$

$$\theta_{1,3} = P(+b/\neg a) = 1/2 = 0,5$$

$$\theta_{1,4} = P(+c/+a) = (1 + 0,5)/4 = 0,375$$

$$\theta_{1,5} = P(+c/\neg a) = \frac{1}{2} = 0,5$$

En este caso $p = -1,8808$

Es decir, esta estimación $\vec{\theta}_1$ del modelo es mejor que las tres obtenidas hasta el momento (su verosimilitud es mayor).

Aprendizaje paramétrico con datos incompletos

Continuamos iterando:

Fase expectation:

Calculamos el valor esperado de la variable C en la observación incompleta:

$$P(+c/+a, \neg b) = P(+c/+a) = 0.375$$

$$P(\neg c/+a, \neg b) = 1 - P(+c/+a) = 0.625$$

A	B	C
+a	+b	+c
+a	+b	$\neg c$
+a	$\neg b$	$\neg c$
$\neg a$	+b	$\neg c$
$\neg a$	$\neg b$	+c
+a	$\neg b$	$P(+c/+a, \neg b) = 0.375$ $P(\neg c/+a, \neg b) = 0.625$

Aprendizaje paramétrico, datos incompletos

Fase maximization:

Calculamos el nuevo modelo:

$$\theta_{2,1} = P(+a) = 4/6 = 0,6667$$

$$\theta_{2,2} = P(+b/+a) = 2/4 = 0,5$$

$$\theta_{2,3} = P(+b/\neg a) = 1/2 = 0,5$$

$$\theta_{2,4} = P(+c/+a) = (1 + 0,375)/4 = 0,34375$$

$$\theta_{2,5} = P(+c/\neg a) = \frac{1}{2} = 0,5$$

En este caso $p = -1,8791$

Es decir, esta estimación $\vec{\theta}_2$ del modelo es mejor que las tres obtenidas hasta el momento (su verosimilitud es mayor).

Aprendizaje paramétrico, datos incompletos

Continuamos iterando:

Fase expectation:

$$P(+c/+a, \neg b) = P(+c/+a) = 0.34375$$

$$P(\neg c/+a, \neg b) = 1 - P(+c/+a) = 0.6562$$

Fase maximization:

Calculamos el nuevo modelo:

$$\theta_{3,1} = P(+a) = 4/6 = 0,6667$$

$$\theta_{3,2} = P(+b/+a) = 2/4 = 0,5$$

$$\theta_{3,3} = P(+b/\neg a) = 1/2 = 0,5$$

$$\theta_{3,4} = P(+c/+a) = (1 + 0,34375)/4 = 0,3359$$

$$\theta_{3,5} = P(+c/\neg a) = 1/2 = 0,5$$

En este caso $p = -1,8790$

Es decir, esta estimación del modelo es mejor que las obtenidas hasta el momento (su verosimilitud es mayor).

El algoritmo finalizaría cuando veamos que el valor de p aumenta menos que una cantidad dada

Algoritmo EM

Entrada: conjunto de datos (con valores incompletos) + Estructura red
Salida: estimación de los parámetros de la red - conjunto de valores para $\vec{\theta}$

Pasos:

1. $\vec{\theta} \leftarrow \vec{\theta}_0;$
2. $l \leftarrow \hat{l}(\vec{\theta}_0);$
3. Repetir:
 4. Para cada dato desconocido $d \in \vartheta;$
 5. estimar d mediante $\vec{\theta};$
 6. $\vec{\theta} \leftarrow EMV(\vartheta);$
 7. $l' \leftarrow \hat{l}(\vec{\theta})$
 8. $\Delta l = l' - l;$
 9. $l \leftarrow l';$
10. Hasta que $\Delta l \leq \varepsilon$

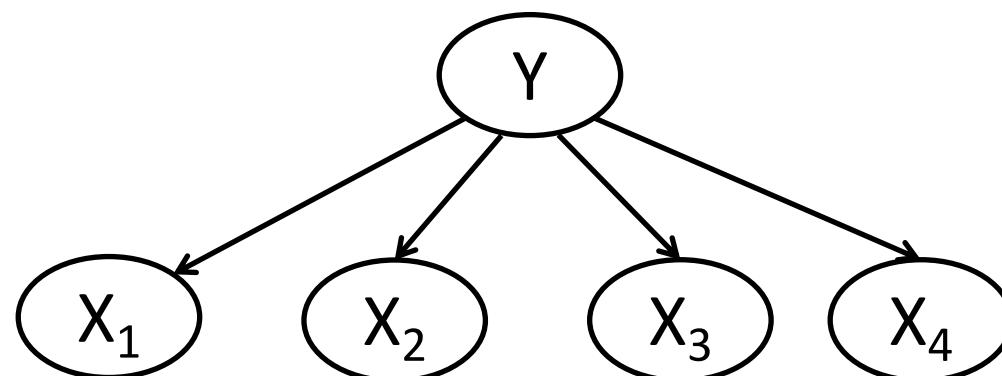
El clasificador Naive Bayes

- Se desea clasificar un objeto, dada una serie de rasgos.
- Llamemos Y a la clase, y \mathbf{X} al vector de rasgos.
- Los clasificadores probabilísticos determinan la probabilidad de que un nuevo ejemplo, \mathbf{x}_q , pertenezca a una cierta clase, y , es decir, $P(y|\mathbf{x}_q)$.
- El objetivo es encontrar un estimador h que $y_q = h(\mathbf{x}_q)$.
Este estimador nos ayudará a determinar la clase a la que pertenece el nuevo ejemplo.
- Dada esa función h , tendremos que la clase estimada para \mathbf{x}_q será aquella clase $y_q = y$ tal que $P(y|\mathbf{x}_q)$ sea máxima. En notación matemática se escribe así:

$$y_q = \arg \max_{\{y \in V\}} P(y|\mathbf{x}_q)$$

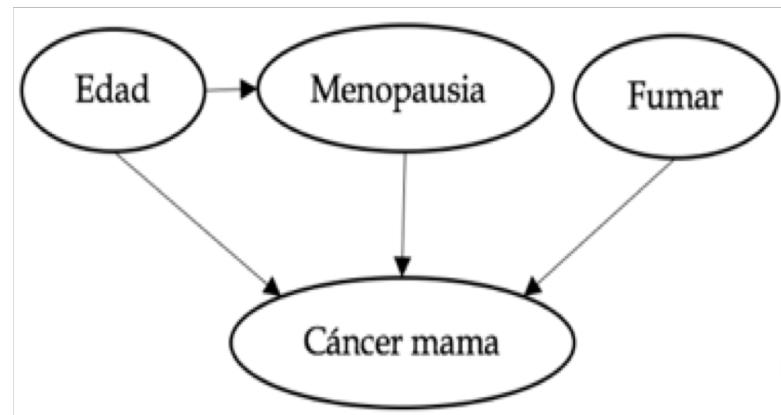
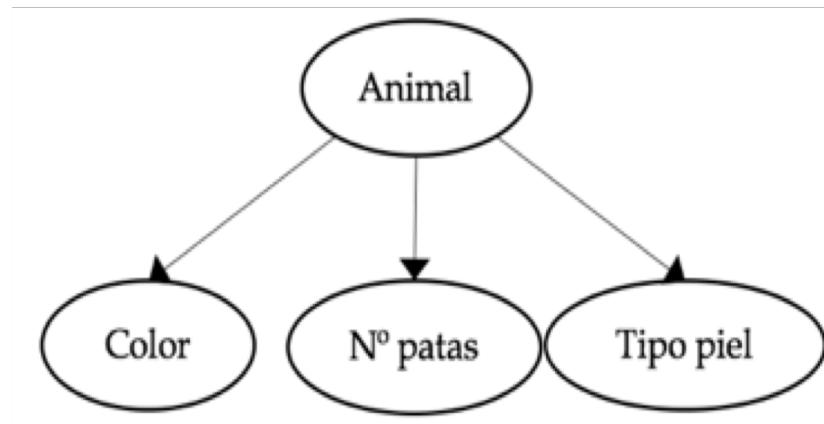
Hipótesis Naive

La hipótesis Naive consiste en suponer que los rasgos son independientes dado la clase, lo cual no siempre tiene por qué ser cierto. Es decir, un clasificador Naive Bayes supondrá que la estructura es:



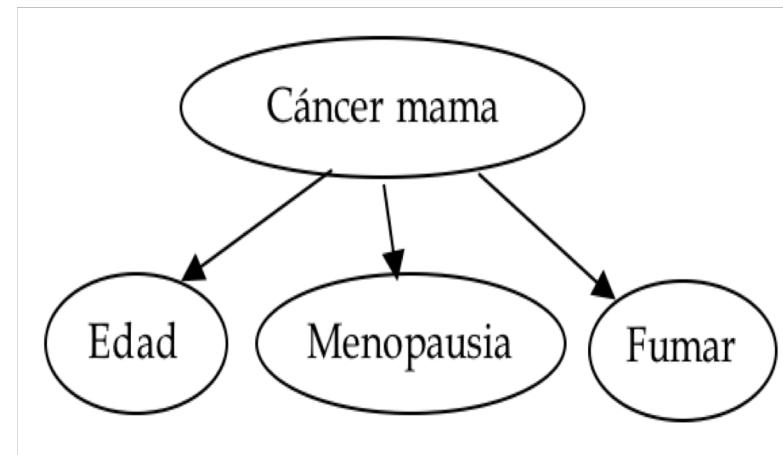
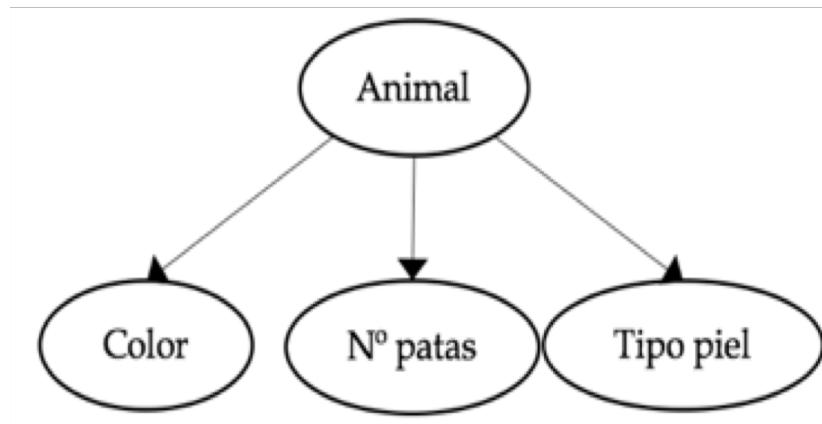
Hipótesis Naive Bayes

A veces es cierta, a veces no



Hipótesis Naive Bayes

Para el clasificador Naive Bayes:



Clasificador Naive Bayes

Paso 1.

A partir de los datos, obtener estimadores para $P(y)$ y para cada uno de los valores posibles de los rasgos, $P(x_d/y)$.

Como estimador de $P(x_d/y)$, se utiliza el llamado suavizado de Laplace:

$$P(x_d/y) \sim \frac{n + mp}{n * + m}$$

donde:

- n es el número de ejemplos de la clase y con x_d ,
- n^* es el número de ejemplos de entrenamiento de clase y ,
- m y p son factores correctores (se utilizan para evitar sesgos producidos por la base de datos). Si la variable X tiene r valores posibles, suele utilizarse $m=r$ y $p=1/r$

Clasificador bayesiano

Paso 2:

Calcular la probabilidad de la clase para el nuevo ejemplo, es decir:

$$P(y | \mathbf{x}) = \frac{P(y)P(\mathbf{x} | y)}{\sum_{v \in V} P(v)P(\mathbf{x} | v)}$$

* (el denominador será una constante de normalización α)

Para realizar este cálculo, aplicamos la hipótesis Naive:

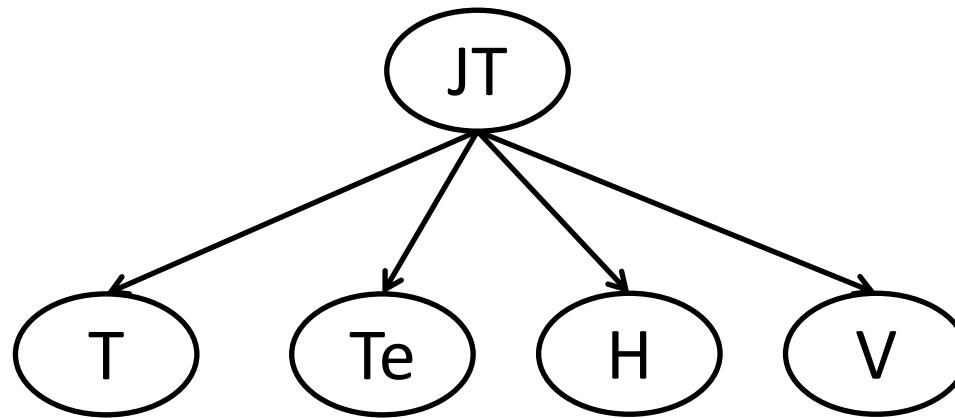
$$P(\mathbf{x} | y) = \prod_{d=1}^D P(x_d | y)$$

Asignaremos el nuevo ejemplo \mathbf{x} a la clase de mayor probabilidad

Clasificador Naive Bayes: ejemplo 1

Día	Cielo (T)	Temperatura (Te)	Humedad (H)	Viento (V)	Jugar al tenis (JT)
D1	soleado	fría	normal	débil	sí
D2	soleado	suave	normal	fuerte	sí
D3	cubierto	fría	normal	fuerte	sí
D4	cubierto	cálido	alta	débil	sí
D5	cubierto	cálido	normal	débil	sí
D6	cubierto	suave	alta	fuerte	sí
D7	Lluvia	suave	alta	débil	sí
D8	Lluvia	suave	normal	débil	sí
D9	Lluvia	fría	normal	débil	sí
D10	soleado	cálido	alta	débil	no
D11	soleado	cálido	alta	fuerte	no
D12	soleado	suave	alta	débil	no
D13	lluvia	fría	normal	fuerte	no
D14	lluvia	suave	alta	fuerte	no

Clasificador Naive Bayes: ejemplo 1



Nuevo ejemplo, $E = \{ T=\text{soleado}, T=\text{fría}, H=\text{alta}, V=\text{fuerte} \}$

Queremos saber si $JT = \text{sí}$, o $JT = \text{no}$

Para ello calculamos:

- $P(JT=\text{sí}/E)$
- $P(JT=\text{no}/E)$

Y asignamos el nuevo ejemplo a la clase con mayor probabilidad

Clasificador Naive Bayes: ejemplo 1

Paso 1.

En base a los datos, calculamos un estimador de las probabilidades de la clase y de cada variable dada la clase para E (consideramos las frecuencias de aparición sin factor corrector, o sea, $m=0$):

$$P(JT=\text{sí}) = 9/14$$

$$P(T=\text{fría}/JT=\text{sí}) = 3/9$$

$$P(C=\text{soleado}/JT=\text{sí}) = 2/9$$

$$P(H=\text{alta}/JT=\text{sí}) = 3/9 = 1/3$$

$$P(V=\text{fuerte}/JT=\text{sí}) = 3/9 = 1/3$$

$$P(JT=\text{no}) = 5/14$$

$$P(T=\text{fría}/JT=\text{no}) = 1/5$$

$$P(C=\text{soleado}/JT=\text{no}) = 3/5$$

$$P(H=\text{alta}/JT=\text{no}) = 4/5$$

$$P(V=\text{fuerte}/JT=\text{no}) = 3/5$$

Clasificador Naive Bayes: ejemplo 1

Paso 2.

Calculamos ahora $P(JT=sí/E)$ y $P(JT=no/E)$. Utilizamos la expresión

$$P(JT/E) = \frac{P(JT,E)}{P(E)} = \frac{P(JT) P(E/JT)}{P(E)}$$

- $P(JT=sí, E) = P(JT=sí)*P(\{C=sol, T=fría, H=alta, V=fuer\}/JT=sí) =$
 - $=P(JT=sí)*P(C=sol/JT=sí)*P(T=fría/JT=sí)*P(H=alta/JT=sí)*P(V=fuer/JT=sí)=$ $=\frac{9}{14} * \frac{1}{3} * \frac{2}{9} * \frac{1}{3} * \frac{1}{3} = \frac{2}{378} = 0,0053$
- $P(JT=no/E) = P(JT=no)*P(C=soleado,T=fría,H=alta,V=fuerte/JT=no)=$ $\frac{5}{14} * \frac{1}{5} * \frac{3}{5} * \frac{4}{5} * \frac{3}{5} = \frac{36}{1750} = 0,0206$

Vemos que es más probable $P(JT=no/E)$, así que asignaríamos el nuevo ejemplo a “no jugar al tenis”

Clasificador Naive Bayes, ejemplo 2

M1	send us your password	spam
M2	send us your review	ham
M3	review your password	ham
M4	review us	spam
M5	send your password	spam
M6	send us your account	spam

Nuevo e-mail, con texto “review your account”. ¿Será SPAM?

Definimos el *vocabulario* de nuestro ejemplo:

	P (password)	R (review)	(S) send	U (us)	Y (your)	A (account)
M1 (spam)	1	0	1	1	1	0
M2 (ham)	0	1	1	1	1	0
M3 (ham)	1	1	0	0	1	0
M4 (spam)	0	1	0	1	0	0
M5 (spam)	1	0	1	0	1	0
M6 (spam)	0	0	1	1	1	1
Mensaje nuevo (?)	0	1	0	0	1	1

Clasificador Naive Bayes, ejemplo 2

	P (password)	R (review)	(S) send	U (us)	Y (your)	A (account)
M1 (spam)	1	0	1	1	1	0
M2 (ham)	0	1	1	1	1	0
M3 (ham)	1	1	0	0	1	0
M4 (spam)	0	1	0	1	0	0
M5 (spam)	1	0	1	0	1	0
M6 (spam)	0	0	1	1	1	1
Mensaje nuevo (?)	0	1	0	0	1	1

Paso 1. A partir de la base de datos, estimamos $P(y)$ y $P(x_d/y)$

	spam	ham
a priori	2/3	1/3
password=1	2/4	1/2
review=1	1/4	2/2
send=1	3/4	1/2
us=1	3/4	1/2
your=1	3/4	2/2
account=1	1/4	0/2

Dominaría la clasificación,
utilizamos el corrector de Laplace

Clasificador Naive Bayes, ejemplo 2

Paso 1.

Aplicamos la corrección de Laplace, con valores $p=0.5$ y $m=2$ y obtenemos:

	spam	ham
password=1	$(2+1)/(4+2)$	$(1+1)/(2+2)$
review=1	$(1+1)/(4+2)$	$(2+1)/(2+2)$
send=1	$(3+1)/(4+2)$	$(1+1)/(2+2)$
us=1	$(3+1)/(4+2)$	$(1+1)/(2+2)$
your=1	$(3+1)/(4+2)$	$(2+1)/(2+2)$
account=1	$(1+1)/(4+2)$	$(0+1)/(2+2)$

Clasificador Naive Bayes, ejemplo 2

Paso 2:

	spam	ham
a priori	2/3	1/3
password=1	1/2	1/2
review=1	1/3	3/4
send=1	2/3	1/2
us=1	2/3	1/2
your=1	2/3	3/4
account=1	1/3	1/4

	spam	ham
-	-	-
password=0	1/2	1/2
review=0	2/3	1/4
send=0	1/3	1/2
us=0	1/3	1/2
your=0	1/3	1/4
account=0	2/3	3/4

Calculamos ahora:

$$P(\text{spam}/\text{mensaje nuevo}) = \alpha * \frac{2}{3} * \frac{1}{2} * \frac{1}{3} * \frac{1}{3} * \frac{1}{3} * \frac{2}{3} * \frac{1}{3} = \alpha * (2/729) = 0,0027 \alpha$$

$$P(\text{ham}/\text{mensaje nuevo}) = \alpha * \frac{1}{3} * \frac{1}{2} * \frac{3}{4} * \frac{1}{2} * \frac{1}{2} * \frac{3}{4} * \frac{1}{4} = \alpha * (3/512) = 0,0058 \alpha$$

El mensaje será clasificado como "ham"

Validación

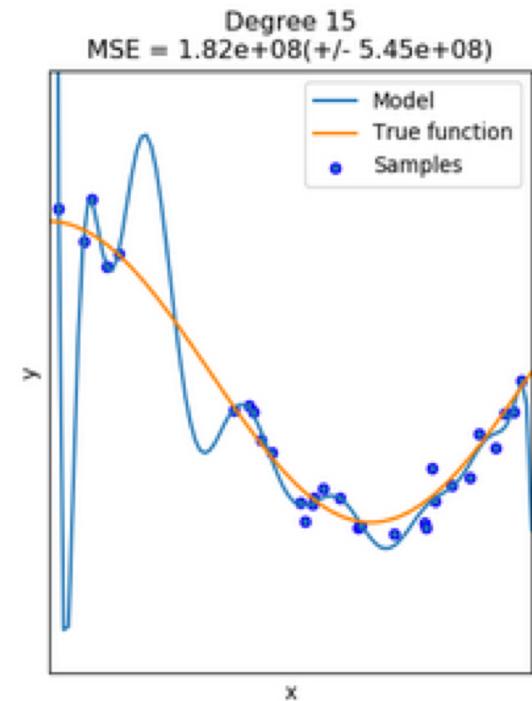
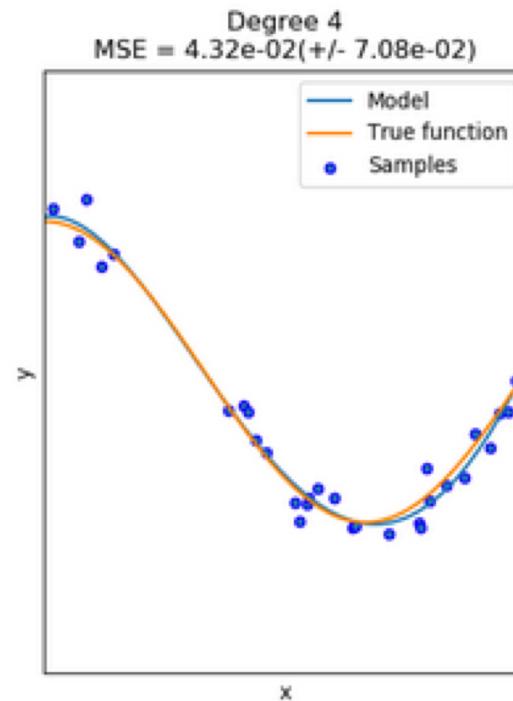
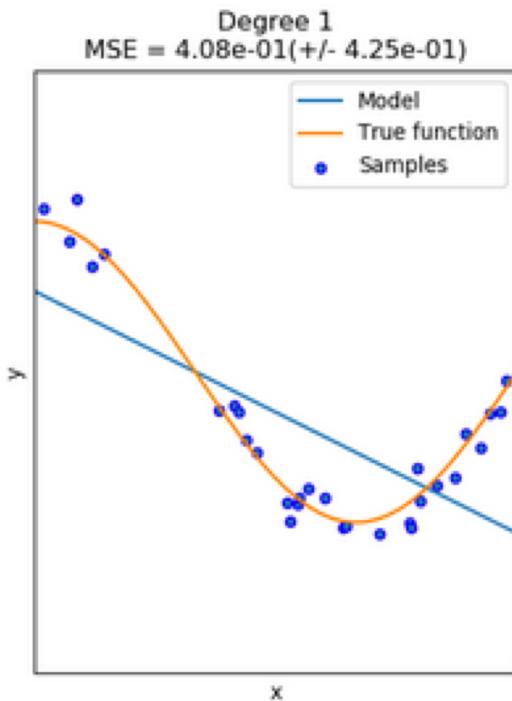
Una vez que se obtiene un modelo para un determinado problema, resulta útil poder evaluar su rendimiento,

- para saber si las inferencias que realicemos con el modelo son fiables,
- para poder comparar unos modelos con otros.

Dependiendo del tamaño del conjunto de datos, se puede dividir el mismo en tres partes:

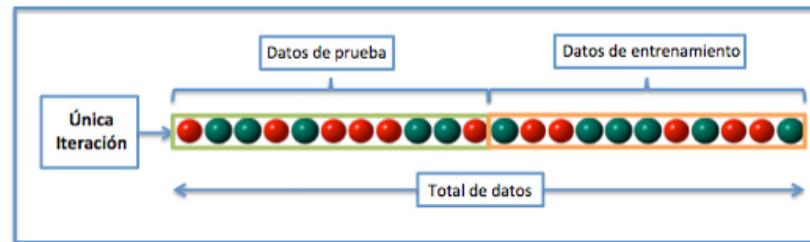
- El **conjunto de entrenamiento** (training set), que es el que se proporciona a los algoritmos de aprendizaje para obtener los estimadores \hat{y}_i .
- El **conjunto de validación** (validation set), que se utiliza para elegir el mejor estimador entre los obtenidos
- El **conjunto de pruebas** (test set), que se utiliza para evaluar el rendimiento del estimador elegido mediante una serie de medidas

Problemas de sobreajuste e infrajuste

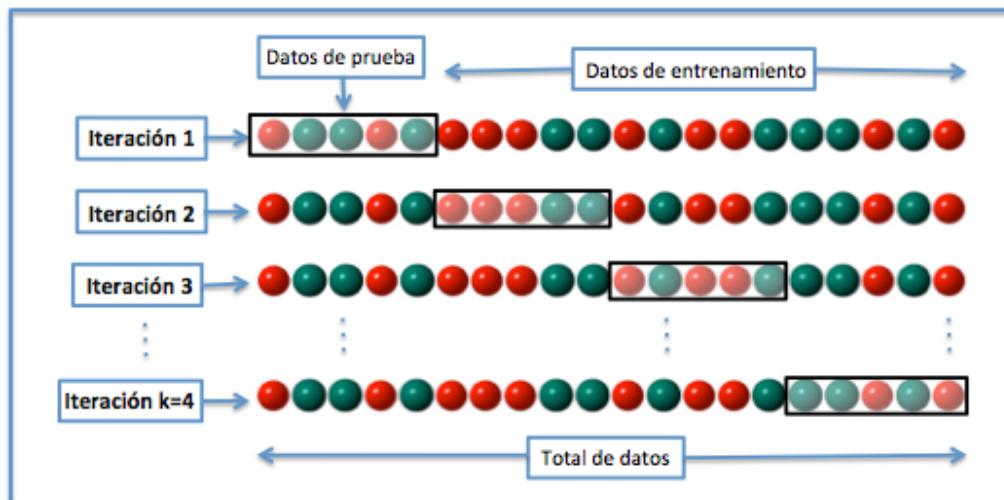


Técnicas de validación

Método de retención (hold-out)



Validación cruzada (k-fold cross validation)



Medidas de rendimiento: matriz de confusión

La matriz de confusión es una matriz cuadrada de tamaño s , donde s es el número de clases posibles. Su elemento (i,j) se define como:

$c_{ij} = \text{nº de ejemplos de la clase } j \text{ que se estimó que pertenecían a la clase } i$

		Clase real			
		y_1	...	y_s	
Clase estimada	y_1	c_{11}	...	c_{1s}	
	
	y_s	c_{s1}	...	c_{ss}	

Medidas numéricas de rendimiento

- **Verdaderos positivos (VP)** = c_{11}
- **Verdaderos negativos (VN)** = c_{22}
- **Falsos positivos o errores de tipo 1 (FP)** = c_{12}
- **Falsos negativos o errores de tipo 2 (FN)** = c_{21}
- **Precisión (accuracy)** =
$$\frac{VP+VN}{VP+VN+FP+FN}$$
- **Tasa o razón de verdaderos positivos (sensibilidad)** =
$$\frac{VP}{VP+FN}$$
- **Tasa o razón de falsos positivos** =
$$\frac{FP}{FP+VN}$$
.
- **Tasa o razón de verdaderos negativos (especificidad)** =
$$\frac{VN}{VN+FN}$$
.

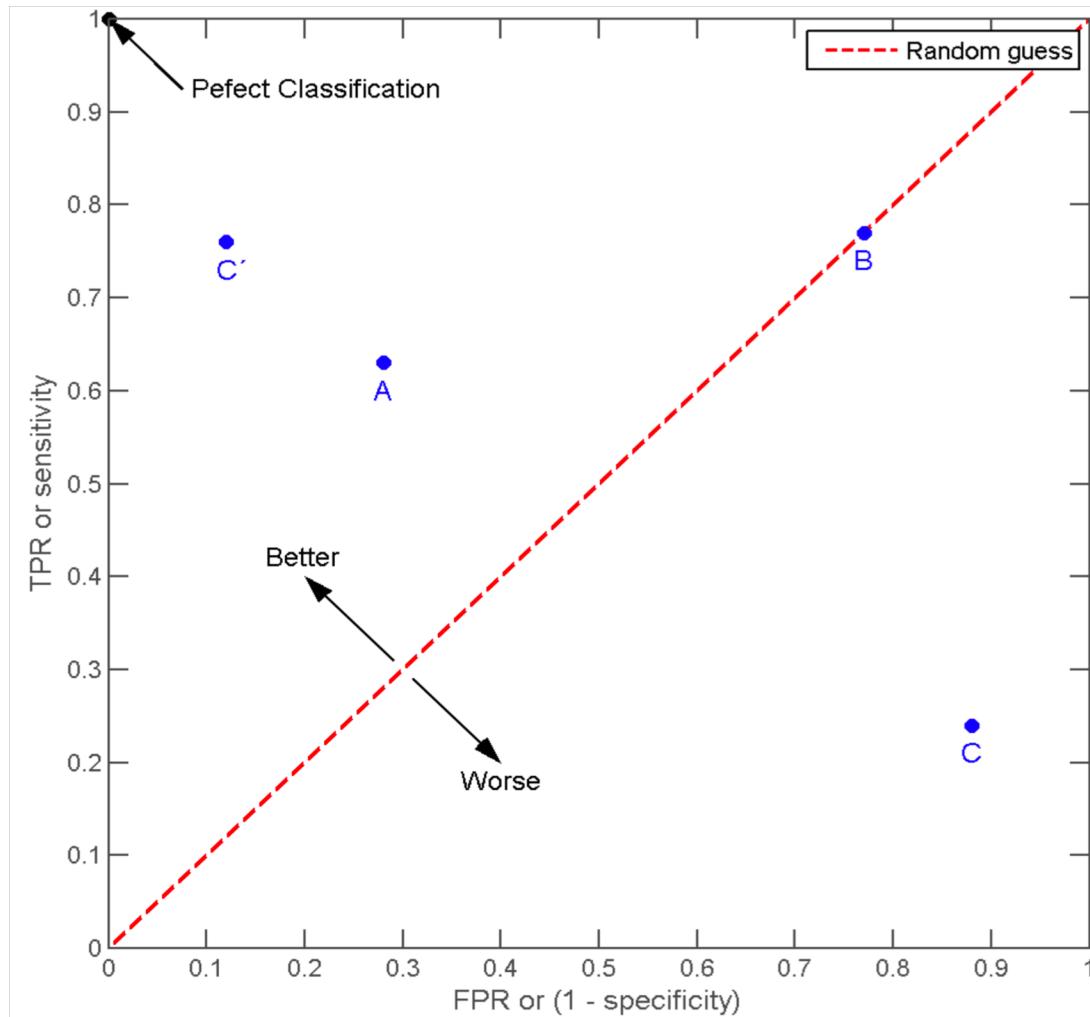
Ejemplo. Medidas de rendimiento de clasificadores

Tres modelos

A		B		C	
VP=63	FP=28	91	VP=77	FP=77	154
FN=37	VN=72	109	FN=23	VN=23	46
100	100	200	100	100	200

	A	B	C
Tasa de Verdaderos positivos o (sensibilidad)	0.63	0.77	0.24
Tasa de Falsos positivos	0.28	0.77	0.88
Tasa de Verdaderos negativos (especificidad)	0.62	0.23	0.12
Precisión	0.68	0.50	0.18

ESPACIO ROC

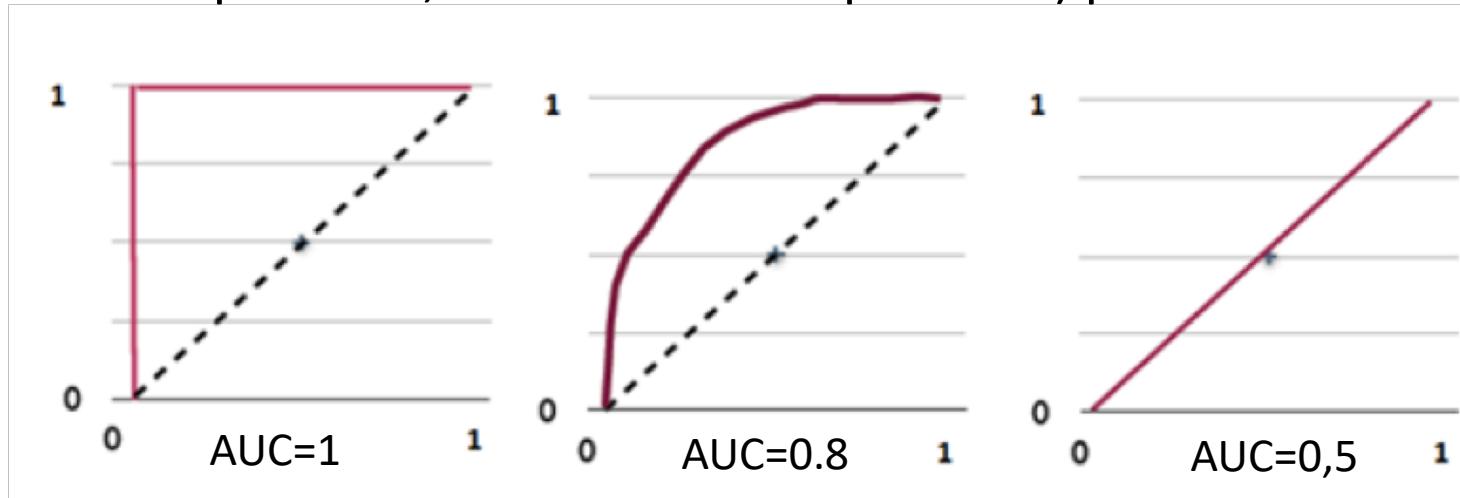


Curva ROC y AUC (Area under curve ROC)

Se define un valor umbral τ para asignar un ejemplo a la clase positiva. La clase será $y = 1$ sí y solo sí $P(y=1/x_q) > \tau$

La curva ROC se obtiene dibujando los pares

(tasa falsos positivos, tasa verdaderos positivos) para cada valor de τ



AUC es el área bajo la curva ROC:

- Si $AUC > 0,75 \rightarrow$ prueba buena
- Si $AUC > 0,97 \rightarrow$ prueba excelente