

# Aprendizaje Automatico - Trabajo Practico 1

Gonzalo Castiglione - 49138

April 20, 2012

**Objetivo: Aprender a clasificar datos utilizando clasificadores bayesianos.**

## 1 Aprendizaje bayesiano

1. A continuación se muestra una matriz de probabilidades de gustos para cada tipo de oyente de la radio:

	J	V
P(1)	0.95	0.03
P(2)	0.05	0.82
P(3)	0.02	0.34
P(4)	0.2	0.92

La fila  $i$  indica la probabilidad que al grupo ( $J$  o  $V$ ) le guste el programa  $i$ .

Sea la condición  $c$  = El oyente disfruta del los programas 1 y 3, pero no de los programas 2 y 4.

Se pueden definen las siguientes variables a fin de simplificar las cuentas:

$x_j$  = Oyentes jóvenes que cumplen con  $c$ .

$x_v$  = Oyentes viejos que cumplen con  $c$ .

Con estás ultimas fórmulas recién definidas, resulta fácil ver que cualquier oyente que cumple con  $c$ , tiene que pertenecer al conjunto  $x_j$  o  $x_v$ . Por lo que se podría definir Además la varaible  $x_{Total} = x_j + x_v$  (notar que  $x_j \cap x_v = \phi$ )

Quedando así la siguiente fórmula:

$$P(J|c) = \frac{x_j}{x_{Total}} = \frac{x_j}{x_j + x_v} \quad (1)$$

En donde cada  $x$  se expresa como:

$$x_j = P(1|J) * P(2|J) * (1 - P(3|J)) * (1 - P(4|J))$$

$$x_v = P(1|V) * P(2|V) * (1 - P(3|V)) * (1 - P(4|V))$$

Sabiendo que el oyente cumple con  $c$  y reemplazando las fórmulas con los valores de la matriz de probabilidades, se obtiene que:

$$x_j = 0.95 * 0.02 * (1 - 0.02) * (1 - 0.2) \approx 0.015$$

$$x_v = 0.03 * 0.82 * (1 - 0.34) * (1 - 0.92) \approx 0.001$$

Finalmente, si reemplazamos estos últimos resultados en (1), resulta que

$$P(J|c) = \frac{0.015}{0.015 + 0.001} = 0.92$$

Es decir, hay un 92% de probabilidad que el oyente en cuestión sea joven.

## 2. Algoritmo $h_{MAP}$

Para cada hipótesis  $h$  de  $H$ , se calcula:  $P(h|D) = \frac{P(D|h)*P(h)}{P(D)}$

Se da como salida  $h_{MAP} = \max_{h \in H} P(h|D)$

Hipótesis a considerar:

$$x = (1, 0, 1, 1, 0)$$

$h_1$  = La persona es escoses

$h_2$  = La persona es inglesa.

Cálculo de probabilidad para cada hipótesis:

Se define la probabilidad de  $h$  como:  $P(h) = \frac{1}{|H|} = 0.5 = h_1 = h_2$

- $P(h_1|D) = \frac{1*0.5}{P(D)} = 0.5$
- $P(h_2|D) = \frac{1*0.5}{P(D)} = 0.5$

Con los datos observados el algoritmo  $h_{max}$  no es capaz de asegurar si  $x$  es escoses o inglés ya que ambos tienen la misma probabilidad de ocurrir.

## 3. Solución

	Maíz	Granola	Azucarados	Avena	Mayor a 60
1	1	0	0	0	1
2	1	0	0	1	1
3	1	1	1	1	1
4	0	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0

Se desea clasificar la instancia:  $x = (0, 1, 1, 0)$ .

fórmula del clasificador de Naive de Bayes

$$v_{NB} = \max_{v_j \in V} P(v_j) \prod_{i=0}^n P(a_i|v_j)$$

$$v_{NB} = P(v_j)P(\text{Maíz} = 0|v_j)P(\text{Granola} = 1|v_j)P(\text{Azucarado} = 1|v_j)P(\text{Avena} = 0|v_j)$$

$$P(\text{Mayor a 60} = 1) = \frac{4}{6} = 0.667$$

$$P(\text{Mayor a 60} = 0) = 1 - P(\text{Mayor a 60} = 1) = \frac{2}{6} = 0.333$$

$$P(\text{Maíz} = 0|\text{Mayor a 60} = 1) = \frac{1}{4} = 0.25$$

$$P(\text{Maíz} = 0|\text{Mayor a 60} = 0) = \frac{1}{2} = 0.5$$

$$P(\text{Granola} = 1|\text{Mayor a 60} = 1) = \frac{1}{4} = 0.25$$

$$P(\text{Granola} = 1|\text{Mayor a 60} = 0) = \frac{2}{2} = 1$$

$$P(\text{Azucarado} = 1|\text{Mayor a 60} = 1) = \frac{1}{4} = 0.25$$

$$P(\text{Azucarado} = 1|\text{Mayor a 60} = 0) = \frac{1}{2} = 0.5$$

$$P(\text{Avena} = 0|\text{Mayor a 60} = 1) = \frac{1}{4} = 0.25$$

$$P(\text{Avena} = 0|\text{Mayor a 60} = 0) = \frac{2}{2} = 1$$

Quedando las ecuaciones de las probabilidades de  $x$  de la siguiente manera:

Sea  $c$  = Mayor a 60.

Sea  $d$  = No es mayor a 60.

$$(1) - P(c)P(\text{Maíz} = 0|c)P(\text{Granola} = 1|c)P(\text{Azucarado} = 1|c)P(\text{Avena} = 0|c) = 0.667 * 0.25 * 0.25 * 0.25 * 0.25 = 2.6 \times 10^{-3}$$

$$(2) - P(d)P(\text{Maíz} = 0|d)P(\text{Granola} = 1|d)P(\text{Azucarado} = 1|d)P(\text{Avena} = 0|d) = 0.333 * 0.5 * 1 * 0.5 * 1 = 0.083$$

Por ser el resultado de (1) > (2), el algoritmo escoge  $v_{NB} = c$ , es decir, la persona es mayor a 60.

Además (por normalización), como las se cantidades anteriores suman uno, se puede calcular la probabilidad de que  $x$  ocurra realizando  $\frac{(1)}{(1)+(2)} \simeq 97\%$ .

Cabe aclarar que este resultado fue obtenido a partir de una muestra muy *pequeña*, por lo que su grado de certeza podría no ser aceptable.

4. Matriz con las mediciones realizadas por el psicólogo

	Rico	Casado	Saludable	Contenta?
1	1	1	1	1
2	0	0	1	1
3	1	1	0	1
4	1	0	1	1
5	0	0	0	0
6	1	0	0	0
7	0	0	1	0
8	0	1	0	0
9	0	0	0	0

Se desea calcular la probabilidad que la instancia:  $x = (0, 1, 1)$  este feliz con su vida.

(a) Cálculo de la probabilidad

$$v_{NB} = P(v_j)P(\text{Rico} = 0|v_j)P(\text{Casado} = 1|v_j)P(\text{Saludable} = 1|v_j)$$

$$P(\text{Contenta} = 1) = \frac{4}{9} = 0.444$$

$$P(\text{Contenta} = 0) = \frac{5}{9} = 0.556$$

$$P(\text{Rico} = 0|\text{Contenta} = 1) = \frac{1}{4} = 0.25$$

$$P(\text{Rico} = 0|\text{Contenta} = 0) = \frac{4}{5} = 0.8$$

$$P(\text{Casado} = 1|\text{Contenta} = 1) = \frac{2}{4} = 0.5$$

$$P(\text{Casado} = 1|\text{Contenta} = 0) = \frac{1}{5} = 0.2$$

$$P(\text{Saludable} = 1|\text{Contenta} = 1) = \frac{3}{4} = 0.75$$

$$P(\text{Saludable} = 1|\text{Contenta} = 0) = \frac{1}{5} = 0.2$$

Ecuaciones de las probabilidades de  $x$  de la siguiente manera:

Sea  $c$  = está contenta con su vida.

Sea  $d$  = No está contenta con su vida.

$$1 \Rightarrow P(c)P(\text{Rico} = 0|c)P(\text{Casado} = 1|c)P(\text{Saludable} = 1|c) = 0.444 * 0.25 * 0.5 * 0.75 = 0.042$$

$$2 \Rightarrow P(d)P(\text{Rico} = 0|d)P(\text{Casado} = 1|d)P(\text{Saludable} = 1|d) = 0.556 * 0.8 * 0.2 * 0.2 = 0.0178$$

El algoritmo retorna que la persona está contenta con una probabilidad de acierto de  $\frac{0.042}{0.042+0.0178} = 0.70$ . Es decir, un 70%.

(b) Sea una persona  $/x = (0, 1)$ . La probabilidad de que  $x$  este Contenta, está dada por:

$c$  = la persona está contenta.

$d$  = la persona no está contenta.

$$v_1 = P(c)P(\text{Rico} = 0|c)P(\text{Casado} = 1|c) = 0.444 * 0.25 * 0.5 = 0.055$$

$$v_2 = P(d)P(\text{Rico} = 0|d)P(\text{Casado} = 1|d) = 0.556 * 0.8 * 0.2 = 0.089$$

$$P(x_1) = \frac{0.089}{0.089+0.055} = 0.62 \Rightarrow 62\%$$

Por lo tanto, la probabilidad que una persona Pobre y Casada este contenta es del 62%

- (c) Una persona pobre, casada y saludable estaría definida por  $x = (0, 1, 1)$ . La probabilidad que  $x$  este contenta es de  $\frac{0.042}{0.042+0.012} = 0.78$   
 $c$  = la persona está contenta.  
 $d$  = la persona no está contenta.  
 $v_1 = P(c)P(\text{Rico} = 0|c)P(\text{Casado} = 1|c)P(\text{Saludable} = 1|c) = 0.444 * 0.25 * 0.5 * 0.75 = 0.042$   
 $v_2 = P(d)P(\text{Rico} = 0|d)P(\text{Casado} = 1|d)P(\text{Saludable} = 1|d) = 0.556 * 0.8 * 0.2 * 0.2 = 0.018$   
 $P(x_1) = \frac{0.042}{0.042+0.018} = 0.7 \Rightarrow 70\%$

## 5. Solución

- (a) Las medidas obtenidas a través del *matlab* son las siguientes:

		Largo cépalo	Ancho cépalo	Largo pétalo	Ancho pétalo
Max	Setosa	5.8000	4.4000	1.9000	0.6000
	Versicolor	7.0000	3.4000	5.1000	1.8000
	Virginica	7.9000	3.8000	6.9000	2.5000
Min	Setosa	4.3000	2.3000	1.0000	0.1000
	Versicolor	4.9000	2.0000	3.0000	1.0000
	Virginica	4.9000	2.2000	4.5000	1.4000
Media	Setosa	5.0060	3.4280	1.4620	0.2460
	Versicolor	5.9360	2.7700	4.2600	1.3260
	Virginica	6.5880	2.9740	5.5520	2.0260
Desvio	Setosa	0.3525	0.3791	0.1737	0.1054
	Versicolor	0.5162	0.3138	0.4699	0.1978
	Virginica	0.6359	0.3225	0.5519	0.2747
Rango	Setosa	1.5000	2.1000	0.9000	0.5000
	Versicolor	2.1000	1.4000	2.1000	0.8000
	Virginica	3.0000	1.6000	2.4000	1.1000

A continuación se presentará algunas de las mediciones de la tabla graficadas

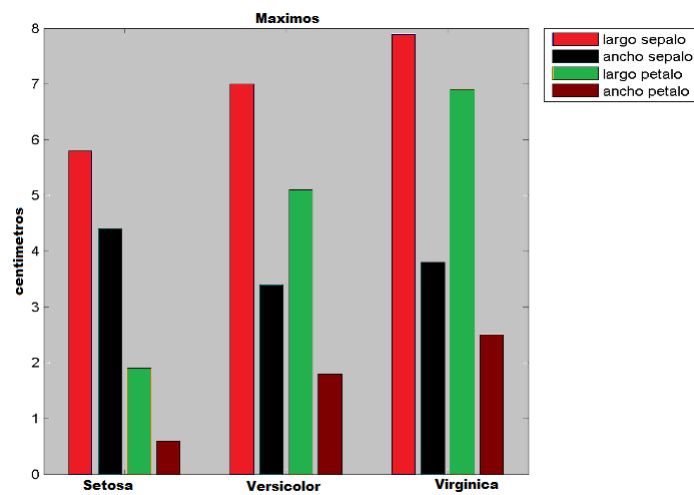


Figure 1: Máximos valores alcanzados por cada tipo de Lirio

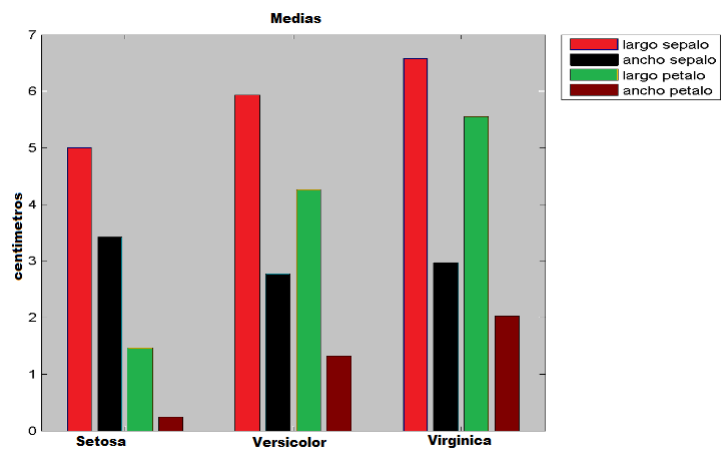


Figure 2: Medias alcanzadas por cada tipo de Lirio

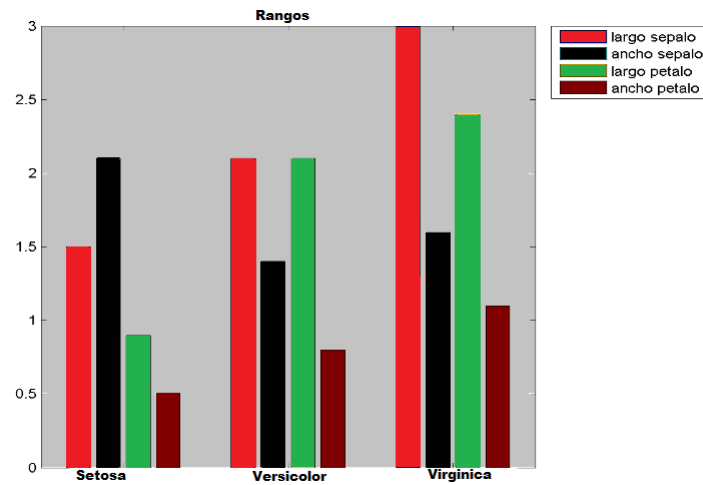


Figure 3: Rangos alcanzados por cada tipo de Lirio

(b) Gráfico de los valores:

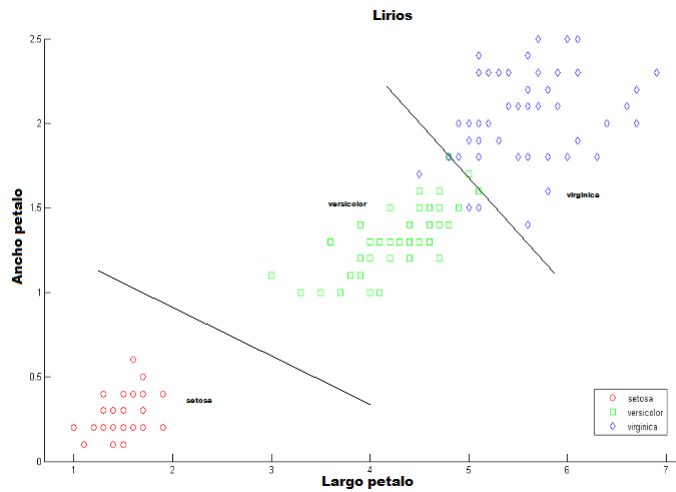


Figure 4: Posibles “rangos” de valores útiles para clasificar cada lirio a partir de las dimensiones de su pétalo

En el gráfico puede observarse que hay un umbral muy definido entre la clase *setosa* de Lirio frente a las otras dos. No siendo tan así para los tipos *versicolor* y *virginica*. Por lo que, a simple vista, se puede intuir que si hay errores de clasificación, es muy probable que sean entre estas dos especies.

(c) Asumiendo distribución gaussiana

i. Código

```
load fisheriris;  
NB = NaiveBayes.fit(meas,species);  
NB_Clases=NB.predict(meas);
```

i. Porcentaje de datos mal calculados = 4%

ii. Matriz de confusión

50	0	0
0	47	3
0	3	47

En la matriz de confusión, todos los valores que se encuentran fuera de la diagonal principal, son valores que se clasificaron mal. A partir de esta observación, se puede calcular el error cometido como  $\frac{\textit{incorrectos}}{\textit{total}} = \frac{3+3}{150} = 0.04$ . Un 4% de error.