# Aprendizaje Automatico - Trabajo Practico 1

Gonzalo Castiglione - 49138

April 20, 2012

**Objetivo: Aprender a clasificár datos utilizando clasíficadores bayesianos.**

## 1 Aprendizaje bayesiano

1. A continuación se muestra una matríz 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.

Reaalizando un análists previo de los resultados, se ve que la mayoría de los jovenes disfrutan de escuchar los programas 1 y 3 y por el otro lado, la mayoría de los viejos no disfrutan de los progrmas 1 y 3. Por lo que, a simple vista, debería esperarse un resultado en donde haya un probabilidad muy grande que oyente sea joven.

Sea P(J) = probabilidad que el oyente sea joven = 1 - P(V) (ya que para esta estación de radio, no existe ningúna otra categoría de oyentes)

Por el teorema de *Bayes*, se tiene la fórmula:

$$P(J|c) = \frac{P(c|J)P(J)}{P(c)}$$

A partir de esta fórmula, se realiza el cálculo de $v_{NB}$

P($c$|J) = P(p1 = 1|J) * P(p3 = 1|J) * P(p2 = 0|J) * P(p4 = 0|J) $\simeq 0.1444$

Para el cálculo de P($c$), se puede aplicar el teorema de la probabilidad total,

$$P(c) = P(c|J) * P(J) + P(c|V) * P(V) \quad (1)$$

En esta fórmula faltaría calcular la proabilidad P($c$|V) (de manera análoga a P($c$|J))

P($c$|V) = P(p1 = 1|V) * P(p3 = 1|V) * P(p2 = 0|V) * P(p4 = 0|V) $\simeq 0.0016$

Reemplazando estos últimos valores en (1), se obtiene que P($c$) $\simeq 0.016$

Y reemplazando nuevamente todos los valores calculados en la fórmula de *Bayes*. Queda finalmente que P(J|$c$)$\simeq 0.9$.

Hay un 90% de probabilidad que el oyente sea joven.

2. Algormitmo $h_{MAP}$

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

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

   Hipótesis a consderar:

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

   $h_e$ =La persona es escoses y siempre se cumplen los datos observados.

   $h_i$ =La persona es inglésa y siempre se cumplen los datos observados.

   Cálculo de probabilidad para cada hipótesis:

$$P(X) = \sum_i P(X|Ai)P(Ai) \quad (2)$$

   $P(h_e) = (1 * \frac{1}{13}) * 5 + (\frac{1}{2} * \frac{2}{13}) * 2 = 7/13 = 0.54$

   $P(h_i) = (1 * \frac{1}{13}) * 4 + (\frac{1}{2} * \frac{2}{13}) * 2 = 6/13 = 0.46$

   (Notar que la suma de ambas hipótesis es 1, y como son los únicos casos que pueden ocurrir, esta bien)

   Quedando así $P(h_e|D) = P(h_i|D) = \frac{1}{13}$

   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 clasificár la instancia: $x = (0, 1, 1, 0)$.

fórmula del clasíficador de Naive de Bayes

$$v_{NB} = \underset{v_j \varepsilon V}{max} P(v_j) \prod_{i=0}^{n} P(a_i|v_j)$$

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

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

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

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

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

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

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

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

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

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

P(Avena = 0|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(Maíz = 0|$c$)P(Granola = 1|$c$)P(Azucarado = 1|$c$)P(Avena = 0|$c$) = $0.667 * 0.25 * 0.25 * 0.25 * 0.25 = 2.6x10^{-3}$

(2) - P($d$)P(Maíz = 0|$d$)P(Granola = 1|$d$)P(Azucarado = 1|$d$)P(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 mayoy 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 ser aceptable.

4. Matríz 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(Rico $= 0|v_j$)P(Casado $= 1|v_j$)P(Saludable $= 1|v_j$)

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

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

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

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

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

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

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

P(Saludable $= 1|$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.

$1 \Longrightarrow$ P($c$)P(Rico $= 0|c$)P(Casado $= 1|c$)P(Saludable $= 1|c$) $= 0.444 * 0.25 * 0.5 * 0.75 = 0.042$

$2 \Longrightarrow$ P($d$)P(Rico $= 0|d$)P(Casado $= 1|d$)P(Saludable $= 1|d$) $= 0.556 * 0.8 * 0.2 * 0.2 = 0.0178$

El algorimto 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(Rico $= 0|c$)P(Casado $= 1|c$) $= 0.444 * 0.25 * 0.5 = 0.055$

$v_2$= P($d$)P(Rico $= 0|d$)P(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 estária 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.

$v_1 = \mathrm{P}(c)\mathrm{P}(\text{Rico} = 0|c)\mathrm{P}(\text{Casado} = 1|c)\mathrm{P}(\text{Saludable} = 1|c) = 0.444 * 0.25 * 0.5 * 0.75 = 0.042$

$v_2 = \mathrm{P}(d)\mathrm{P}(\text{Rico} = 0|d)\mathrm{P}(\text{Casado} = 1|d)\mathrm{P}(\text{Saludable} = 1|d) = 0.556 * 0.8 * 0.2 * 0.2 = 0.018$

$\mathrm{P}(x_1) = \frac{0.042}{0.042+0.018} = 0.7 \Rightarrow 70\%$

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 presental algunas de las mediciones de la tabla graficadas

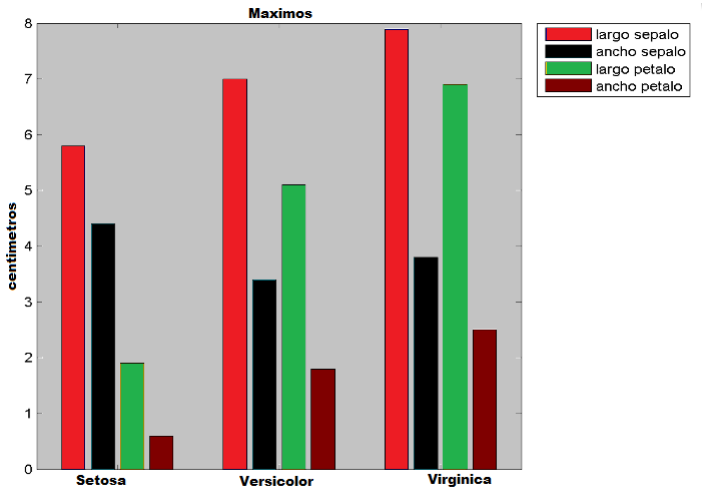


Figure 1: Máximos valores alcanzados por cada tipo de Lirio

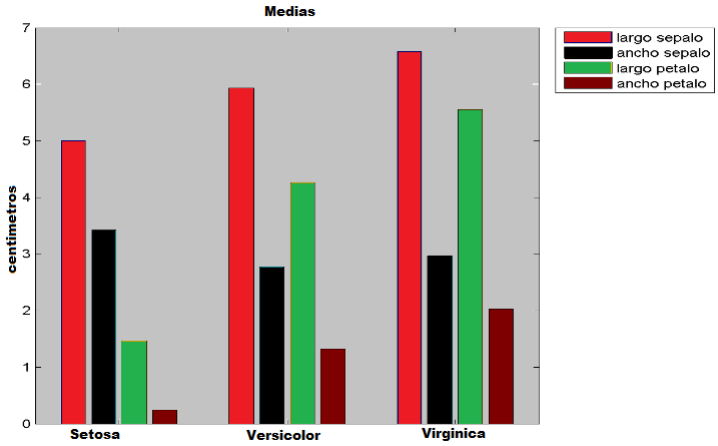


Figure 2: Medias alcanzadas por cada tipo de Lirio

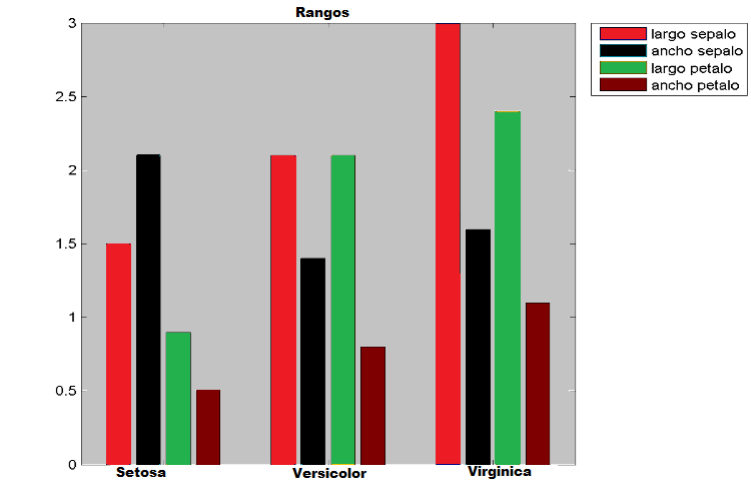


Figure 3: Rangos alcanzados por cada tipo de Lirio

(b) Gráfico de los valores:

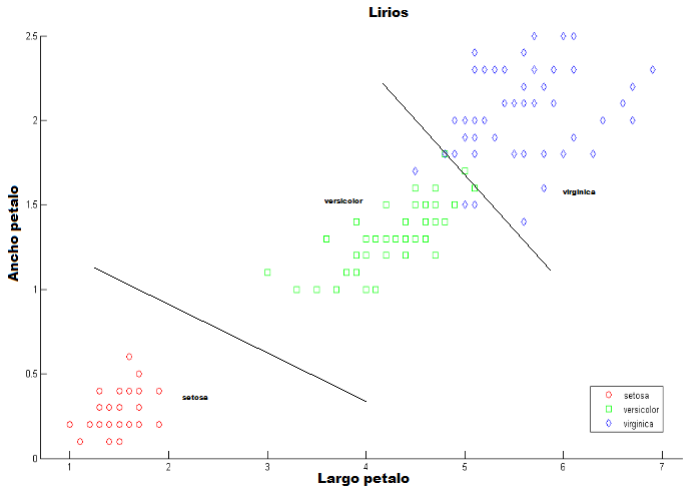


Figure 4: Posibles "rangos" de valores útiles para clasificar cada lirio a partir de las dimesiones de su pétalo

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 *virgínica*. Por lo que, a simple vista, se puede intuir que si hay errores de clasificacion, 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. Matríz de confusión

| | | |
|---|---|---|
| 50 | 0 | 0 |
| 0 | 47 | 3 |
| 0 | 3 | 47 |

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