Facultad de Ingeniería

Universidad Rafael Landívar
Inteligencia Artificial
Primer Semestre 2025

# Hoja de Trabajo 8

## Descripción

Esta actividad se divide en dos partes: la primera aborda un problema de **redes bayesianas**, y la segunda aplica el clasificador **Naïve Bayes** para la predicción de etiquetas de clase.

Serie 1. Redes Bayesianas (50 puntos)

**VER PDFs adicionales**

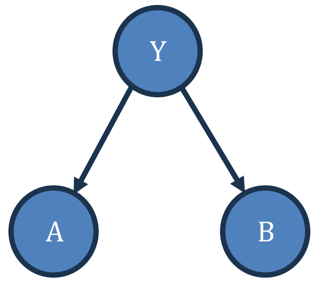
**<u>Serie 2. Naïve Bayes (50 puntos)</u>**

En este ejercicio, resolverá un caso particular de las redes bayesianas: el clasificador Naïve Bayes, que se utilizará para predecir la etiqueta de clase $Y$ en función de las variables de entrada $A$ y $B$. Todas las variables son binarias (0 o 1). En algunos casos, deberá aplicar conceptos de Máxima Verosimilitud y Suavizamiento de Laplace para estimar las distribuciones de probabilidad.

Datos de Entrenamiento:

Se tienen 10 ejemplos de entrenamiento:

| A | B | Y |
|---|---|---|
| 1 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 0 | 0 |
| 1 | 1 | 0 |
| 0 | 1 | 0 |
| 1 | 1 | 1 |
| 0 | 1 | 1 |
| 1 | 0 | 0 |
| 1 | 1 | 0 |
| 1 | 1 | 0 |




$$p(C)\prod_{i=1}^{n} p(F_i|C)$$

La estimación de Máxima Verosimilitud se basa en la frecuencia observada en los datos.

Preguntas:

1. **(20 puntos)** Calcule las CPTs necesarias para construir la tabla de probabilidades:

   1.1. Calcule la probabilidad marginal *P(Y)*:

$$P(Y=0)=\frac{6}{10}$$

$$P(Y=1)=\frac{4}{10}$$

| $Y$ | $P(Y)$ |
|---|---|
| 0 | $3/5$ |
| 1 | $2/5$ |

   1.2. Calcule la distribución condicional *P(A|Y)*:

   - *P(A=0|Y=0), P(A=1|Y=0)*
   - *P(A=0|Y=1), P(A=1|Y=1)*

| $A$ | $Y$ | $P(A\|Y)$ |
|---|---|---|
| 0 | 0 | $1/6$ |
| 1 | 0 | $5/6$ |
| 0 | 1 | $1/4$ |
| 1 | 1 | $3/4$ |

1.3. Calcule la distribución condicional *P(B|Y)*:

- *P(B=0|Y=0), P(B=1|Y=0)*
- *P(B=0|Y=1), P(B=1|Y=1)*

| $B$ | $Y$ | $P(B\|Y)$ |
|---|---|---|
| 0 | 0 | $1/3$ |
| 1 | 0 | $2/3$ |
| 0 | 1 | $1/4$ |
| 1 | 1 | $3/4$ |

2. (20 puntos) Clasificación con Naïve Bayes

Se sabe que *A=1* y *B=1*, identifique la etiqueta aplicando Naïve Bayes:

2.1. Prediga la clase más probable (Aplica *Maximun Likelihood*).

$$\begin{aligned} P(Y=0, A=1, B=1) &= P(Y=0)P(A=1|Y=0)P(B=1|Y=0) \\ &= (3/5)(5/6)(2/3) \\ &= 1/3 \end{aligned}$$

$$\begin{aligned} P(Y=1, A=1, B=1) &= P(Y=1)P(A=1|Y=1)P(B=1|Y=1) \\ &= (2/5)(3/4)(3/4) \\ &= 9/40 \end{aligned}$$

La clase predicha será Y=0

2.2. *P (Y=0 | A=1, B=1) (Normaliza)*

P(Y = 0)P(A = 1|Y = 0)P(B = 1|Y = 0)
= (0.6*5/6*4/6) / (0.6*5/6*4/6 + 0.4*3/4*3/4)
≈ 0.3333/0.5583 ≈ 0.597

2.3. *P (Y=1 | A=1, B=1) (Normaliza)*

$P(Y=1)\cdot P(A=1|Y=1)\cdot P(B=1|Y=1)$
= [0.4 * (3/4) * (3/4)] / [0.4*(3/4)*(3/4) + 0.6*(5/6)*(4/6)]
≈ 0.403

3. (10 puntos) Aplicación de Suavizamiento de Laplace
Lea el complemento al final de esta hoja de trabajo.
Para evitar probabilidades de cero, aplica el suavizamiento de Laplace con $k=2$ y recalcula la tabla $P(A|Y)$.
3.1 Calcule los nuevos valores de $P(A|Y)$ con Laplace Smoothing.

**Para** $Y=0$**:**

- $P(A=0|Y=0) = \frac{1+2}{6+4} = \frac{3}{10} = 0.3$
- $P(A=1|Y=0) = \frac{5+2}{6+4} = \frac{7}{10} = 0.7$

**Para** $Y=1$**:**

- $P(A=0|Y=1) = \frac{1+2}{4+4} = \frac{3}{8} = 0.375$
- $P(A=1|Y=1) = \frac{3+2}{4+4} = \frac{5}{8} = 0.625$

| $A$ | $Y$ | $P(A\|Y)$ |
|---|---|---|
| 0 | 0 | 3/10 |
| 1 | 0 | 7/10 |
| 0 | 1 | 3/8 |
| 1 | 1 | 5/8 |

Facultad de Ingeniería

**Entrega:**

- Quiz en el portal:
  - Responder antes la HT
  - 10 minutos para contestar el quiz
- Habilitado a partir del 1 Abril 2025
- Fecha máxima para responder: Domingo 6/04/2025 (Antes de la medianoche)

## Complemento:

La **fórmula de suavizado de Laplace** en el contexto del **Clasificador Naïve Bayes** se usa para evitar probabilidades nulas cuando una categoría no ha aparecido en los datos de entrenamiento.

$$P(Característica|clase) = \frac{Frecuencia(Característica|clase) + k}{Total\ de\ características\ en\ la\ clase\ +\ kV}$$

Donde:

- *k* es el parámetro de suavizado (si k = 1, se trata del suavizado de Laplace clásico).
- *V* es el número total de valores distintos que puede tomar la característica.

Este ajuste permite usar valores distintos de *k* para controlar la cantidad de suavizado aplicado. Este método evita que la probabilidad condicional sea cero cuando una combinación no aparece en los datos de entrenamiento, lo que podría hacer que la probabilidad total de una clase se vuelva cero en el cálculo de **Naïve Bayes**.

1. **¿Cual es la probabilidad de que el paciente presente los síntomas, padezca ambas enfermedades y sea afectado por el gen G?**

   Lo que se solicita es calcular
   +s y +a y +b y +g o sea:

   **P(+s,+a, +b, +g)**

   Dada las tablas:

| $\mathbb{P}(B)$ | |
|---|---|
| $+b$ | 0.4 |
| $-b$ | 0.6 |

| $\mathbb{P}(G)$ | |
|---|---|
| $+g$ | 0.1 |
| $-g$ | 0.9 |

   Se sabe que:
   P(+b) = 0.4
   P(+g) = 0.1

   para P(+a) está condicionada por la evidencia g+ por tanto P(+a|+g) en la tabla:

| $\mathbb{P}(A\|G)$ | | |
|---|---|---|
| $+g$ | $+a$ | 1.0 |
| $+g$ | $-a$ | 0.0 |
| $-g$ | $+a$ | 0.1 |
| $-g$ | $-a$ | 0.9 |

   P(+a|+g) = 1.0

   Para P(+s) está condicionada por la evidencia a+ y b+ por tanto P(+s|+a,+b) en la tabla:

| $\mathbb{P}(S\|A,B)$ | | | |
|---|---|---|---|
| $+a$ | $+b$ | $+s$ | 1.0 |
| $+a$ | $+b$ | $-s$ | 0.0 |
| $+a$ | $-b$ | $+s$ | 0.9 |
| $+a$ | $-b$ | $-s$ | 0.1 |
| $-a$ | $+b$ | $+s$ | 0.8 |
| $-a$ | $+b$ | $-s$ | 0.2 |
| $-a$ | $-b$ | $+s$ | 0.1 |
| $-a$ | $-b$ | $-s$ | 0.9 |

   P(+s|+a,+b)= 1.0
   Por tanto:
   P(+s,+a, +b, +g) = P(+b) * P(+g) * P(+a) * P(+s)
   P(+s,+a, +b, +g) = P(+b) * P(+g) * P(+a|+g) * P(+s|+a,+b) = **0.4*0.1*1.0*1.0 = 0.04**

2. **¿Cual es la probabilidad de que el paciente tenga la enfermedad A?**

Lo que se solicita es calcular P(+a)

Esta tiene dos posibles enfoques para resolverlo:

a) eliminación de variables donde se identifican los factores involucrados:

P(A|G)  P(S|A,B), y para eliminar las variables se necesita P(G) y P(B)

se filtra los factores donde ocurre +a
P(+a|G)
+g  +a   1.0
−g  +a   0.1

P(S|+a,B)
+a  +b  +s   1.0
+a  +b  −s   0.0
+a  −b  +s   0.9
+a  −b  −s   0.1

Esta última tabla cumple con la condición de Factor II en la eliminacion de variables:

$P(W|cold)$

| T | W | P |
|---|---|---|
| cold | sun | 0.4 |
| cold | rain | 0.6 |

Condicional simple (suma 1) $P(Y|x)$

Por tanto P(S|+a,B) = 1, lo cual se puede demostrar por marginalización hacia +a, eliminando S y B.

Por lo que únicamente para trabajar se necesita:
P(G) * P(+a|G)

| $\mathbb{P}(G)$ | |
|---|---|
| $+g$ | 0.1 |
| $-g$ | 0.9 |

x
P(+a|G)
+g  +a   1.0
−g  +a   0.1

= Probabilidad conjunta P(+a,G)

| | P(+a,G) | |
|---|---|---|
| +g | +a | 0.1 |
| −g | +a | 0.09 |

Se marginaliza +a y se obtiene:

**P(+a) = 0.19**

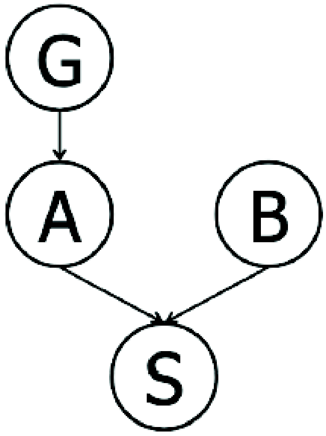
b) Por probabilidad total:

**P(+a)** = P(+a| +g)P(+g) + P(+a| −g)P(−g)=(1.0)(0.1) + (0.1)(0.9) = **0.19**

**3. ¿Cual es la probabilidad de que el paciente tenga la enfermedad A dado que ya tiene la enfermedad B?**

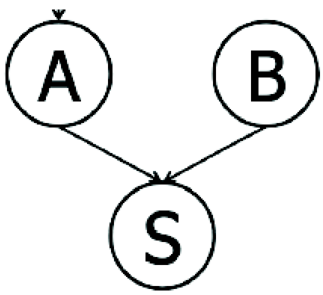
Lo que se solicita es calcular

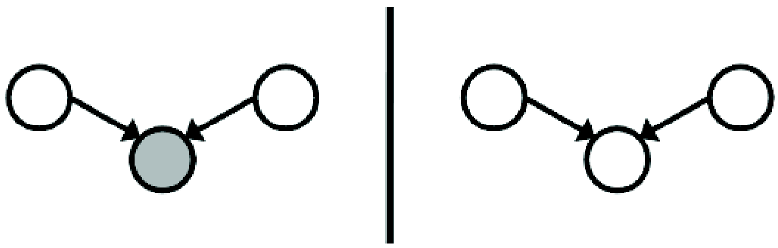
**P(+a|+b)**




A y B son dos variables no inmediatamente conectadas, por tanto puede existir independencia de variables (A⊥B), la cual debe ser demostrada. Lo cual implica calcular si la triada involucrada es activa o inactiva.

Se conoce B ya que la evidencia indica que ocurrió +b por lo que se tiene una triada de la forma:



Dado que la evidencia es B y no S, entonces cumple con el segundo criterio por lo que es una triada inactiva y por tanto independiente, entonces A⊥B.

Entonces, la probabilidad conjunta cuando ambas variables son independientes es:

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

por lo que P(A,B)=P(A)×P(B)

siguiendo el teorema de bayes se tiene la expresión:
P(A|B)=P(A,B) / P(B)

que se puede reescribir basado en la premisa de independencia de variables como:
$P(A|B)=(P(A)\times P(B)) / P(B)$

por lo que $P(A|B)=P(A)$ y $P(+a|+b)=P(+a)$

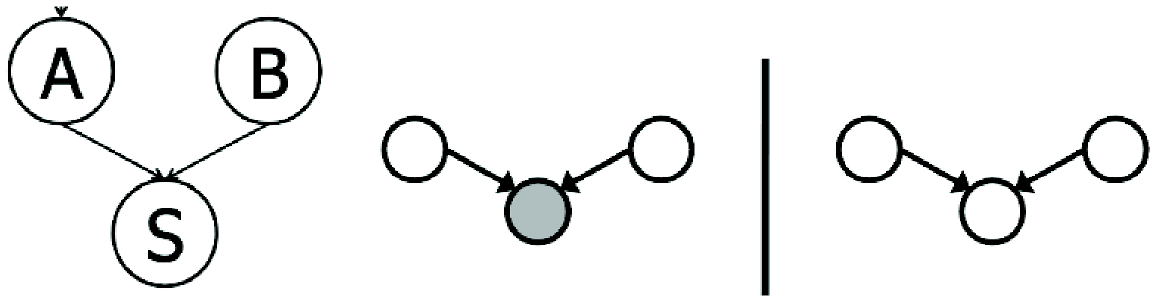
siendo el resultado dado que se conoce $P(+a)$ de incisos anteriores:
**$P(+a|+b)$=0.19**

4. **¿Cual es la probabilidad de que un paciente tenga la enfermedad A dado que tiene los síntomas S y la enfermedad B?**

Lo que se solicita es calcular

**P(+a | +s, +b)**

Nuevamente al igual que el ejercicio 3, se tiene la triada




Pero dada la evidencia +s y +b ahora se cumple con el primer criterio por lo que la triada es activa y por tanto no independiente.

Una vez definida la no independencia, se procede con el teorema de bayes:

P(+a| +s, +b) = **P(+a, +s, +b)** / P(+s, +b)

Para obtener la probabilidad conjunta **P(+a, +s, +b)** se puede utilizar el método de eliminación de variables:
Se identifican los factores involucrados reduciendo:
P(G) P(+a | G) P(+s| +a, +b)  P(+b)

teniendo
P(+a|G)
+g +a 1.0
−g +a 0.1

P(S|+a,+b)
+a +b +s 1.0
+a +b −s 0.0

P(+b)
+b  0.4

| $\mathbb{P}(G)$ | |
|---|---|
| $+g$ | 0.1 |
| $-g$ | 0.9 |

P(+a|G) x P(G) = P(+a,G)

P(+a,G)
+g +a 0.1
−g +a 0.09

marginalizando +a
P(+a)=0.19

entonces:
P(+a) x P(S|+a,+b) x P(+b) =  P(+a, +s, +b)

P(+a, +s, +b)  = 0.19 x 1.0 x 0.4 = 0.076

por lo tanto continuando con el teorema de bayes ahora se sabe que:
P(+a| +s, +b) = P(+a, +s, +b) / P(+s, +b)
P(+a| +s, +b) = 0.076 / **P(+s, +b)**

para calcular **P(+s, +b)** se hace la probabilidad total:
P(+s, +b) = P(+a,+s,+b)+P(-a,+s,+b)

Se sabe que P(+a,+s,+b) = 0.076 por tanto:
P(+s, +b) = 0.076 + P(−a)P(+b)P(+s|−a,+b)

P(−a)= 1 - P(+a) = 0.81
P(+b) = 0.4
P(+s|−a,+b) es según la tabla

| $\mathbb{P}(S\|A,B)$ | | | |
|---|---|---|---|
| $+a$ | $+b$ | $+s$ | 1.0 |
| $+a$ | $+b$ | $-s$ | 0.0 |
| $+a$ | $-b$ | $+s$ | 0.9 |
| $+a$ | $-b$ | $-s$ | 0.1 |
| $-a$ | $+b$ | $+s$ | 0.8 |
| $-a$ | $+b$ | $-s$ | 0.2 |
| $-a$ | $-b$ | $+s$ | 0.1 |
| $-a$ | $-b$ | $-s$ | 0.9 |

P(+s|−a,+b) = 0.8

por lo que:

P(+s, +b) = 0.076 + (0.81x0.4x0.8) = 0.076 + 0.2592 = **0.3352**

entonces:
P(+a| +s, +b) = P(+a, +s, +b) / P(+s, +b)
**P(+a| +s, +b)** = 0.076 / 0.3352 = **0.2267**

5. **¿Cual es la probabilidad de que un paciente fue afectado por el gen G dado que presenta la enfermedad A?**

Lo que se solicita es calcular
**P(+g | +a)**

Usando teorema de bayes se sabe que:

P(+g | +a) = P(+g,+a)/P(+a)

De ejercicios anteriores se sabe que: P(+a) = 0.19 entonces:

P(+g | +a) = P(+g,+a)/0.19

la probabilidad conjunta se puede escribir como: P(+g,+a)= P(+g)P(+a|+g) entonces:

P(+g | +a) = P(+g)P(+a|+g)/0.19

dadas las tablas:

| $\mathbb{P}(G)$ | |
|---|---|
| $+g$ | 0.1 |
| $-g$ | 0.9 |

| $\mathbb{P}(A\|G)$ | | |
|---|---|---|
| $+g$ | $+a$ | 1.0 |
| $+g$ | $-a$ | 0.0 |
| $-g$ | $+a$ | 0.1 |
| $-g$ | $-a$ | 0.9 |

se puede concluir que:
P(+g) = 0.1
P(+a|+g) = 1.0

entonces:
P(+g | +a) = P(+g)P(+a|+g)/0.19
P(+g | +a) = (0.1x1.0)/0.19
**P(+g | +a)** = **0.5263**

6. **¿Cual es la probabilidad de que un paciente fue afectado por el gen G dado que presenta la enfermedad B?**

Lo que se solicita es calcular

**Resolviendolo por el camino largo:

**P(+g | +b)**

Usando teorema de bayes se sabe que:
**P(+g | +b) = P(+g,+b) / P(+b)**

**P(+b)** se sabe por las tablas de la red bayesiana que
P(+b) = 0.4

entonces:

P(+g | +b) = P(+g,+b) / P(+b)
P(+g | +b) = P(+g,+b) / 0.4

P(B) P(S|A,B) P(A | G) P(G)
P(+b) P(S|A,+b) P(A | +g) P(+g)

P(A | +g) y P(S|A,+b) esto se sabe que es 1, por la premisa

$P(W|cold)$

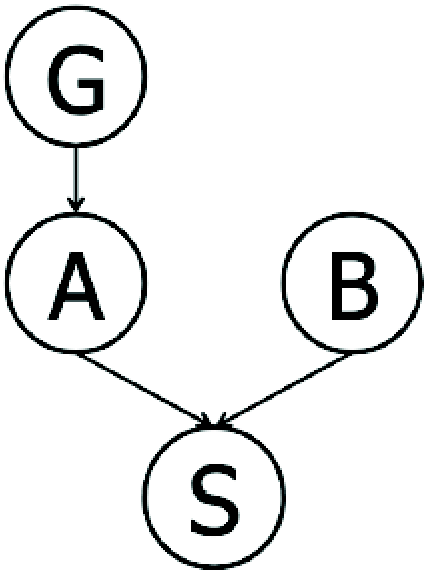
| T | W | P |
|---|---|---|
| cold | sun | 0.4 |
| cold | rain | 0.6 |

Condicional simple (suma 1) $P(Y|x)$

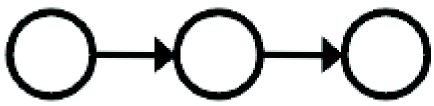
P(+b) P(S|A,+b) P(A | +g) P(+g)
entonces
P(+g,+b) =0.4 x 1.0 x 1.0x0.1 = 0.04
P(+g | +b) = P(+g,+b) / 0.4
P(+g | +b) = 0.04 / 0.4
**P(+g | +b)** = **0.1**

Resolviendo por independencia de variables :D (camino corto y el mejor)

Dado el camino entre G y B




Sabiendo que la evidencia es +b, se tiene la triada
G -> A -> S



La triada es activa

luego se tiene la triada A -> S <- B donde B es la evidencia se tiene

la triada es inactiva

Para decir que un conjunto de tríadas son activas, TODAS las triadas deben ser activas, se tiene una triada inactiva por lo que G y B son independientes, entonces:

**P(+g | +b) = P(+g)**
**P(+g | +b) = 0.1**