# 7. Probabilistic Models

## 7.1 Probability review

### Conditional Probability

Conditional probability is a measure of the probability of an event (some particular situation occurring) given that another event has occurred. The conditional probability of an event A occurs given the event B is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Exercise:

| | Hombres | Mujeres | Niños | |
|---|---|---|---|---|
| Clase Alta | 17 | 36 | 42 | 95 |
| Clase Media o Baja | 5 | 12 | 44 | 61 |
| | 22 | 48 | 86 | 156 |

1. What is the probability of randomly selecting a person from upper class?
2. What is the probability of randomly selecting a person from upper class given that he is a man?

### Independence

Two events are independent, statistically independent, if the occurrence of one does not affect the probability of occurrence of the other. Examples:

Independent events:

- Throw coins, the result of coins is independent of each other.
- The probability of rain is independent of passing the math test.

Dependent events:

- The probability of rain depends on the quantity of clouds.
- The hair color depends on age.

The events A and B are independent if and only if:

$$P(A \cap B) = P(A)P(B)$$

**Bayes' theorem**

It describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

**Total probability**

$$P(A) = \sum_i P(A|B_i)P(B_i)$$

Exercises:

1. En la sala de pediatría de un hospital, el 60% de los pacientes son niñas. De los niños el 35% son menores de 24 meses. El 20% de las niñas tienen menos de 24 meses. Un pediatra que ingresa a la sala selecciona un infante al azar.
a. Determine el valor de la probabilidad de que sea menor de 24 meses.
b. Si el infante resulta ser menor de 24 meses. Determine la probabilidad que sea una niña.

2. Un doctor dispone de tres equipos electrónicos para realizar ecografías. El uso que le da a cada equipo es de 25% al primero, 35% el segundo en y 40% el tercero. Se sabe que los aparatos tienen probabilidades de error de 1%, 2% y 3% respectivamente. Un paciente busca el resultado de una ecografía y observa que tiene un error. Determine la probabilidad de que se ha usado el primer aparato.

## 7.2 Naive Bayes

Naive Bayes is a classifier based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features.

$$\hat{y} = \arg\max_{y} \log P(y) + \sum_{i=1}^{d} \log P(x_i|y)$$

Prove

| | |
|---|---|
| Appling the Bayes' theorem to calculate the probability of a sample belongs to class y based on its input values: $$P(y|x_1,\dots,x_d) = \frac{P(y)P(x_1,\dots,x_d|y)}{P(x_1,\dots,x_d)}$$ Using the naive independence assumption: $$P(y|x_1,\dots,x_n) = \frac{P(y)P(x_1|y)\dots P(x_d|y)}{P(x_1,\dots,x_d)} = \frac{P(y)\prod_{i=1}^{d} P(x_i|y)}{P(x_1,\dots,x_d)}$$ The idea is to calculate the probability of each class $P(y|x_1,\dots,x_n)$ and select the highest one. $$\hat{y} = \arg\max_{y}\ P(y|x_1,\dots,x_n)$$ $$\hat{y} = \arg\max_{y} \frac{P(y)\prod_{i=1}^{d} P(x_i|y)}{P(x_1,\dots,x_d)}$$ | In this case, the term $P(x_1,\dots,x_d)$ is the same for all the classes, so, it can be removed. $$\hat{y} = \arg\max_{y} P(y)\prod_{i=1}^{d} P(x_i|y)$$ The multiplication of several probalities can generate a numerical problem, for that reason, the log function is used: $$\hat{y} = \arg\max_{y} log\left(P(y)\prod_{i=1}^{d} P(x_i|y)\right)$$ $$\hat{y} = \arg\max_{y} \log P(y) + \sum_{i=1}^{d} \log P(x_i|y)$$ |

### Gaussian Naive Bayes (continuous variables)

For continuous variables, it is simple to assume that variables follows a Gaussian distribution. In this case:

$$\hat{y} = \arg\max_{y} \log P(y) - \frac{1}{2}\sum_{i=1}^{d} \log\left(2\pi\sigma_{yi}^2\right) - \frac{1}{2}\sum_{i=1}^{d} \frac{\left(x_i - \mu_{yi}\right)^2}{\sigma_{yi}^2}$$

Prove

| | |
|---|---|
| The objective is to define the term $P(x_i|y)$ in the general Naïve Bayes equation: $$\hat{y} = \arg\max_{y} \log P(y) + \sum_{i=1}^{d} \log P(x_i|y)$$ If it is assumed that variables follows a Gaussian distribution: $$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{\left(x_i - \mu_{yi}\right)^2}{2\sigma_{yi}^2}\right)$$ Applying the logarithm function: $$\log P(x_i|y) = -\log\left(\sqrt{2\pi\sigma_{yi}^2}\right) - \frac{\left(x_i - \mu_{yi}\right)^2}{2\sigma_{yi}^2}$$ $$\log P(x_i|y) = -\frac{1}{2}\log\left(2\pi\sigma_{yi}^2\right) - \frac{\left(x_i - \mu_{yi}\right)^2}{2\sigma_{yi}^2}$$ | But, the Naïve Bayes equation needs the sum of all variables log probabilities $\sum_{i=1}^{d} \log P(x_i|y)$ , so: $$\sum_{i=1}^{d} \log P(x_i|y) = \sum_{i=1}^{d} -\frac{1}{2}\log\left(2\pi\sigma_{yi}^2\right) - \frac{\left(x_i - \mu_{yi}\right)^2}{2\sigma_{yi}^2}$$ $$\sum_{i=1}^{d} \log P(x_i|y) = -\frac{1}{2}\sum_{i=1}^{d} \log\left(2\pi\sigma_{yi}^2\right) - \frac{1}{2}\sum_{i=1}^{d} \frac{\left(x_i - \mu_{yi}\right)^2}{\sigma_{yi}^2}$$ |

Where the parameters are calculated base on the training data.

## Bernoulli Naive Bayes (Text – binary variables)

It is the Naive Bayes version where the all features are binary.

$$\hat{y} = \arg\max_{y} \log P(y) + \log \sum_{i=1}^{d} x_i * P(x_i = 1|y) + \log \sum_{i=1}^{d} (1 - x_i) * P(x_i = 0|y)$$

Prove

| | |
|---|---|
| The objective is to define the term $P(x_i\|y)$ in the general Naïve Bayes equation: $$\hat{y} = \arg\max_{y} \log P(y) + \sum_{i=1}^{d} \log P(x_i\|y)$$ If it is assumed that variables are binary, we have two probabilities: $$P(x_i = 1\|y), P(x_i = 0\|y)$$ $P(x_i = 1\|y)$ is the number of times that $x_i$ is equals to 1 in the entries of class y divided by the number of samples. $$P(x_i = 1\|y) = \frac{\sum_{ij=1}^{n} x_{ij}}{N}$$ $$P(x = 0\|y) = 1 - P(x = 1\|y)$$ We can write that $P(x_i\|y)$ as follows, it activates $P(x_i = 1\|y)$ if $x_i = 1$ and $P(x = 0\|y)$ if $x_i = 0$. $$P(x_i\|y) = P(x_i = 1\|y)^{x_i} * P(x = 0\|y)^{1-x_i}$$ | Applying the logarithm function: $$\log P(x_i\|y) = \log\left(P(x_i = 1\|y)^{x_i} * P(x = 0\|y)^{1-x_i}\right)$$ But, the Naïve Bayes equation needs the sum of all variables log probabilities $\sum_{i=1}^{d} \log P(x_i\|y)$, so: $$\sum_{i=1}^{d} \log P(x_i\|y) = \sum_{i=1}^{d} \log\left(P(x_i = 1\|y)^{x_i} * P(x = 0\|y)^{1-x_i}\right)$$ $$\sum_{i=1}^{d} \log P(x_i\|y) = \sum_{i=1}^{d} \log P(x_i = 1\|y)^{x_i} + \sum_{i=1}^{d} \log P(x = 0\|y)^{1-x_i}$$ $$\sum_{i=1}^{d} \log P(x_i\|y) = \sum_{i=1}^{d} x_i \log P(x_i = 1\|y) + \sum_{i=1}^{d} (1 - x_i) \log P(x = 0\|y)$$ |

Example: Bag of words and Naive Bayes

Training:

Sentence 1: Today is wonderful ☺
Sentence 2: I am happy because AI is wonderful
Sentence 3: I have test ☹

Bag of words

| Sentence | Today | is | wondeful | I | am | happy | because | AI | have | test | ☺ | ☹ | Class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 2 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| 3 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | -1 |

Probabilities $P(x_i = 1|y)$ (counting only the sentences where the variable appears)

| $y/x_i$ | Today | is | wondeful | I | am | happy | because | AI | have | test | ☺ | ☹ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.5 | 1 | 1 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0 | 0 | 0.5 | 0 |
| -1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |

Test:

Sentence: I am happy ☺

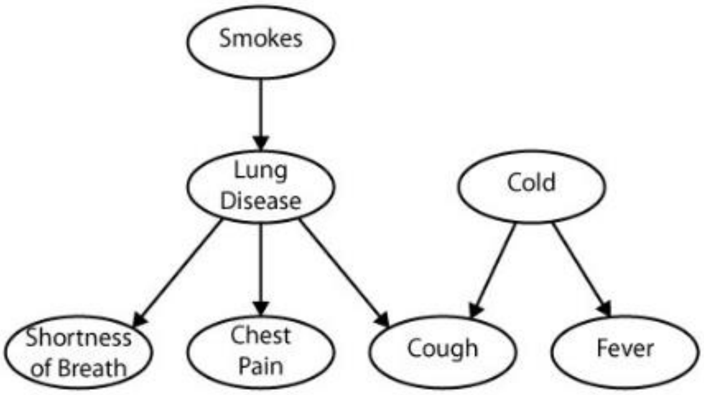
$$
\begin{aligned}
P(y=1) = \log P(y=1) &+ \log p(today = 0|y=1) + \log p(is = 0, y = 1) + \log p(wonderful = 0|y=1) \\
&+ \log p(I = 1, y = 1) + \log p(am = 1|y=1) + \log p(happy = 1|y=1) \\
&+ \log p(because = 0|y=1) + \log p(AI = 0, y = 1) + \log p(have = 0|y=1) \\
&+ \log p(test = 0, y = 1) + \log p(\text{☺} = 1|y=1) + \log p(\text{☹} = 0, y = 1)
\end{aligned}
$$

$$
\begin{aligned}
P(y=0) = \log P(y=0) &+ \log p(today = 0|y=0) + \log p(is = 0, y = 0) + \log p(wonderful = 0|y=0) \\
&+ \log p(I = 1, y = 0) + \log p(am = 1|y=0) + \log p(happy = 1|y=0) \\
&+ \log p(because = 0|y=0) + \log p(AI = 0, y = 0) + \log p(have = 0|y=0) \\
&+ \log p(test = 0, y = 0) + \log p(\text{☺} = 1|y=0) + \log p(\text{☹} = 0, y = 0)
\end{aligned}
$$

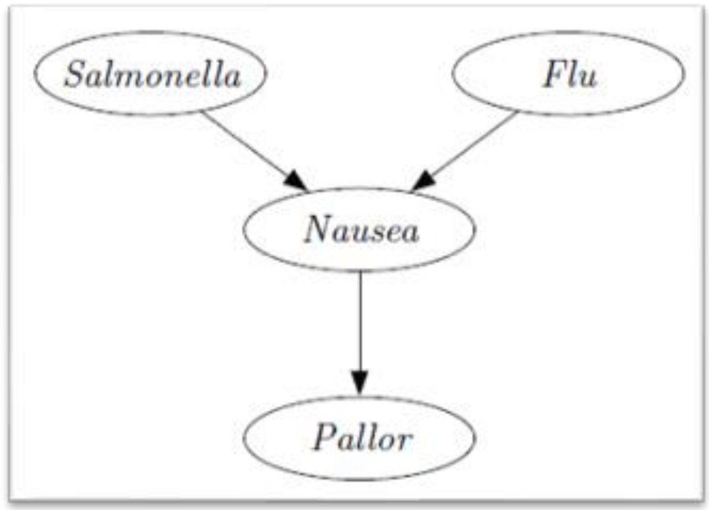
Hint: In text (and in other contexts), it is recommended to add a small value to the frequencies in order to avoid probabilities with value 0.

## 7.3 Bayesian Networks

Una Red Bayesiana es un grafo donde cada nodo representa una variable y las ligas representan dependencia entre las variables; es probabilístico porque la forma en calcular dependencias o inferencias se hace en base a probabilidad. Ejemplos:








Una Red Bayesiana está formada por:

- Nodos, cada nodo representa una variable discreta.
  **No puede haber dos nodos con la misma variable,** cada variable aparece sólo una vez en la red.
- Liga, cada conexión representa dependencia entre variables.
- Cada nodo está asociado a una probabilidad condicional P(Xi|Padres(Xi)).

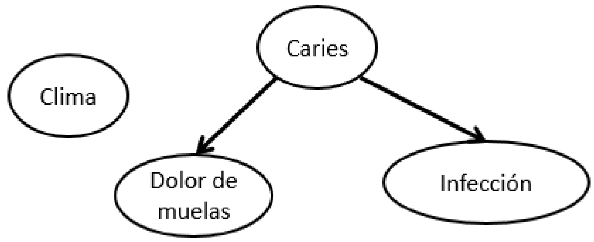
Nota: El grafo no debe tener ciclos.

Los Redes Bayesianas nos pueden servir para modelar y predecir en un conjunto de datos como por ejemplo:

- **Enfermedades y síntomas**, donde cada variable podría ser una enfermedad o un síntoma y las relaciones son las dependencias entre ellas.
- **Predicciones**, por ejemplo si quisiéramos predecir si hay lluvia o no, debemos crear una Red Bayesiana con variables como: temperatura, velocidad del viento, húmedad y las dependencias entre ellas.
- **Sistemas expertos**, si queremos determinar una falla de un auto, en una Red Bayesiana se pueden representar las diferentes causas o efectos que determinan diferentes tipos de fallas.

Ejemplo 1:

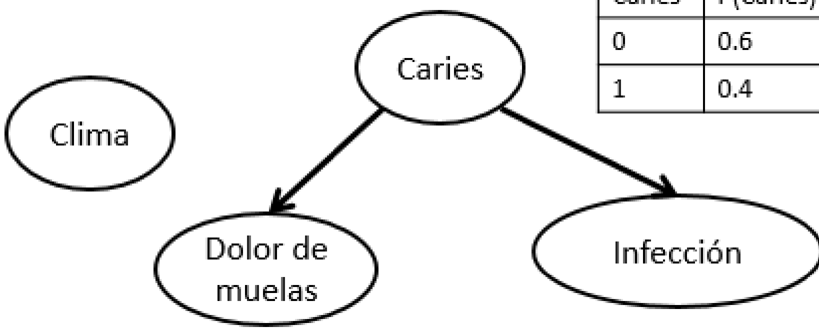
Una Red Bayesiana que represente el hecho de que si una persona tiene caries, puede generar una infección y dolor de muelas, pero esto no depende del clima, sería la siguiente:




Más aún, cada nodo está asociado a una probabilidad, como se muestra en la siguiente figura:

| Caries | P(Caries) |
|---|---|
| 0 | 0.6 |
| 1 | 0.4 |

| Clima | P(Clima) |
|---|---|
| Soleado | 0.6 |
| Nublado | 0.3 |
| Lluvioso | 0.1 |




| Caries | Dolor | P(Dolor\|Caries) |
|---|---|---|
| 0 | 0 | 0.8 |
| 0 | 1 | 0.2 |
| 1 | 0 | 0.3 |
| 1 | 1 | 0.7 |

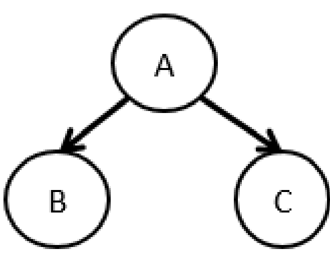
| Caries | Inf | P(Inf\|Caries) |
|---|---|---|
| 0 | 0 | 0.85 |
| 0 | 1 | 0.15 |
| 1 | 0 | 0.1 |
| 1 | 1 | 0.9 |

En una Red Bayesiana la probabilidad de la ocurrencia de un evento está dada por:

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} P\left(X_i \middle| Padres(X_i)\right)$$

Ejemplo:

Si tenemos la siguiente Red Bayesiana, y queremos calcular P(A=0,B=1,C=0)




| A | P(A) |
|---|---|
| 0 | 0.6 |
| 1 | 0.4 |

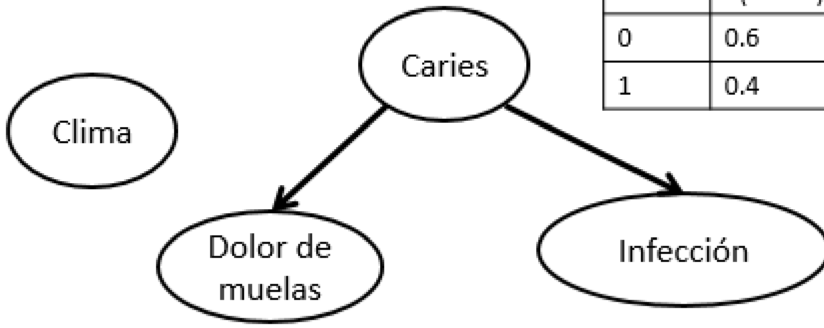
| A | B | P(B\|A) |
|---|---|---|
| 0 | 0 | 0.8 |
| 0 | 1 | 0.2 |
| 1 | 0 | 0.4 |
| 1 | 1 | 0.6 |

| A | C | P(C\|A) |
|---|---|---|
| 0 | 0 | 0.7 |
| 0 | 1 | 0.3 |
| 1 | 0 | 0.9 |
| 1 | 1 | 0.1 |

$$P(A=0, B=1, C=0) = P(A=0) * P(B=1|A=0) * P(C=0|A=0) = (0.6) * (0.2) * (0.7) = 0.084$$

**ACTIVIDAD:** Calcule las probabilidades en las siguientes Redes Bayesianas:

| Clima | P(Clima) |
|---|---|
| Soleado | 0.6 |
| Nublado | 0.3 |
| Lluvioso | 0.1 |




| Caries | P(Caries) |
|---|---|
| 0 | 0.6 |
| 1 | 0.4 |

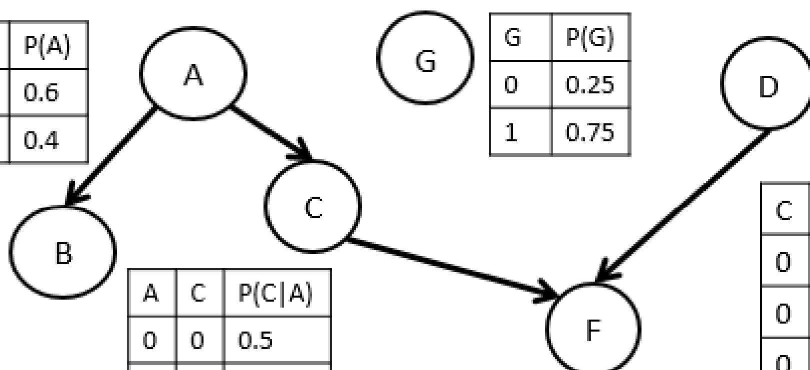
| Caries | Dolor | P(Dolor\|Caries) |
|---|---|---|
| 0 | 0 | 0.8 |
| 0 | 1 | 0.2 |
| 1 | 0 | 0.3 |
| 1 | 1 | 0.7 |

| Caries | Inf | P(Inf\|Caries) |
|---|---|---|
| 0 | 0 | 0.85 |
| 0 | 1 | 0.15 |
| 1 | 0 | 0.1 |
| 1 | 1 | 0.9 |

P(Clima=Nublado, Caries=1, Dolor=1, Infección=1) =

P(Clima=Soleado, Caries=1, Dolor=0, Infección=0) =

| A | P(A) |
|---|---|
| 0 | 0.6 |
| 1 | 0.4 |




| G | P(G) |
|---|---|
| 0 | 0.25 |
| 1 | 0.75 |

| D | P(D) |
|---|---|
| 0 | 1 |
| 1 | 0 |

| A | B | P(B\|A) |
|---|---|---|
| 0 | 0 | 0.3 |
| 0 | 1 | 0.7 |
| 1 | 0 | 0.4 |
| 1 | 1 | 0.6 |

| A | C | P(C\|A) |
|---|---|---|
| 0 | 0 | 0.5 |
| 0 | 1 | 0.5 |
| 1 | 0 | 0.8 |
| 1 | 1 | 0.2 |

| C | D | F | P(F\|C,D) |
|---|---|---|---|
| 0 | 0 | 0 | 0.1 |
| 0 | 0 | 1 | 0.9 |
| 0 | 1 | 0 | 0.3 |
| 0 | 1 | 1 | 0.7 |
| 1 | 0 | 0 | 0.6 |
| 1 | 0 | 1 | 0.4 |
| 1 | 1 | 0 | 0.2 |
| 1 | 1 | 1 | 0.8 |

$$P(A=1, B=1, C=1, D=0, F=1, G=1) =$$
$$P(A=0, B=1, C=0, D=0, F=1, G=1) =$$
$$P(A=1, B=0, C=0, D=1, F=1, G=1) =$$

## Modelado de las Redes Bayesianas

Algunas Redes Bayesianas pueden modelarse por simple intuición, para hacer esto debemos seguir los siguientes pasos:

1. Identificar las variables involucradas
2. Definir las relaciones y causalidades entre las variables

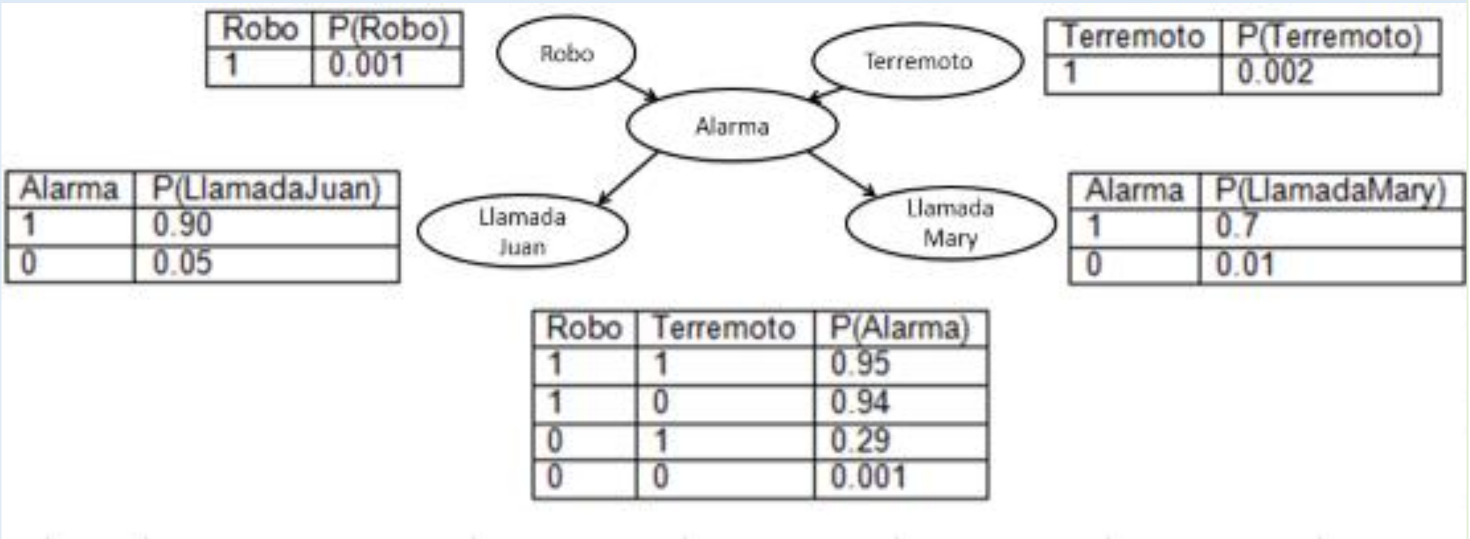
Ejemplo 1:

Pedro tiene un sistema de alarma instalado en su casa. Es muy factible detectar un robo, pero la alarma responde también cuando hay terremotos menores. Pedro tiene dos vecinos, Juan y Mary, quienes han prometido llamarlo cuando escuchen la alarma. Juan siempre llama cuando escucha la alarma, pero a veces confunde el timbre del teléfono con la alarma (en este caso también llama). Mary escucha música fuerte y algunas veces no escucha la alarma.

Ejemplo 2:

Supongamos que hay dos eventos los cuales pueden causar que la hierba esté húmeda: que el rociador esté activado o que esté lloviendo. También supongamos que la lluvia tiene un efecto directo sobre el uso del rociador (usualmente cuando llueve el rociador se encuentra apagado).

**ACTIVIDAD:** Suponiendo los modelos y probabilidades de las siguientes Redes Bayesianas, calcule las probabilidades:




| Robo | P(Robo) |
|---|---|
| 1 | 0.001 |

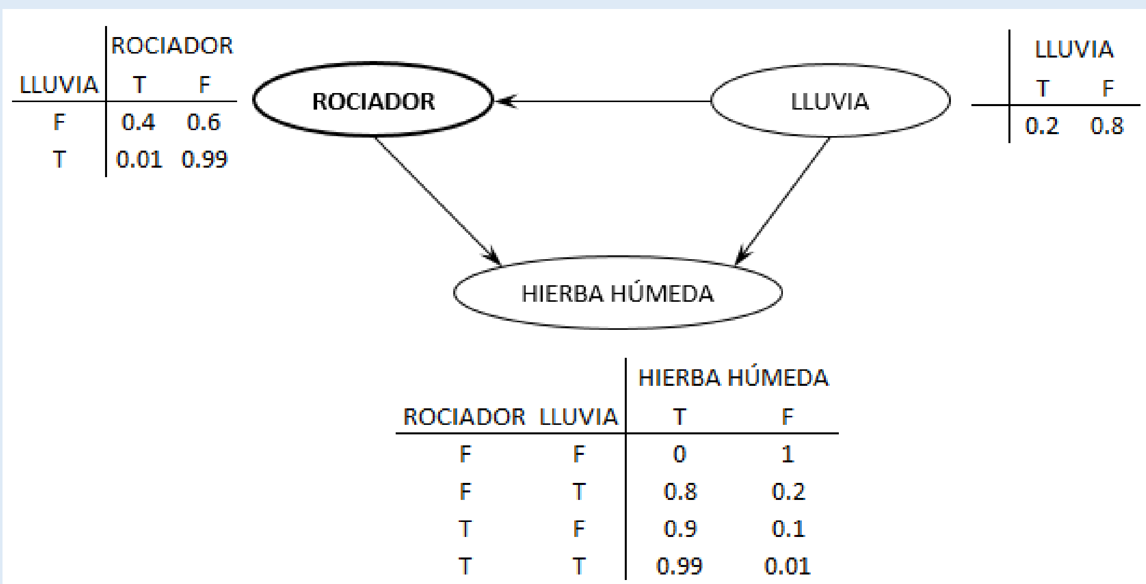
| Terremoto | P(Terremoto) |
|---|---|
| 1 | 0.002 |

| Alarma | P(LlamadaJuan) |
|---|---|
| 1 | 0.90 |
| 0 | 0.05 |

| Alarma | P(LlamadaMary) |
|---|---|
| 1 | 0.7 |
| 0 | 0.01 |

| Robo | Terremoto | P(Alarma) |
|---|---|---|
| 1 | 1 | 0.95 |
| 1 | 0 | 0.94 |
| 0 | 1 | 0.29 |
| 0 | 0 | 0.001 |

1. ¿Cuál es la probabilidad de tener un robo, sin terremoto, que suene la alarma y que llamen Juan y Mary?
2. ¿Cuál es la probabilidad de tener un robo, sin terremoto, que suene la alarma y que ninguno de los vecinos llame?
3. ¿Cuál es la probabilidad de tener un robo, sin terremoto, que no suene la alarma y que ninguno de los vecinos llame?




| LLUVIA | ROCIADOR T | ROCIADOR F |
|---|---|---|
| F | 0.4 | 0.6 |
| T | 0.01 | 0.99 |

| LLUVIA T | LLUVIA F |
|---|---|
| 0.2 | 0.8 |

| ROCIADOR | LLUVIA | HIERBA HÚMEDA T | HIERBA HÚMEDA F |
|---|---|---|---|
| F | F | 0 | 1 |
| F | T | 0.8 | 0.2 |
| T | F | 0.9 | 0.1 |
| T | T | 0.99 | 0.01 |

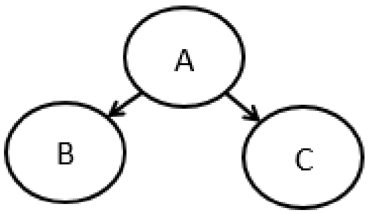
¿Cuál es la probabilidad de que llueva, el rociador este apagado y la hierba este húmeda?

## Inferencia

En algunas ocasiones es necesario inferir una probabilidad marginal, para esto es necesario realizar una sumatoria sobre las variables que no se infieren.

Ejemplo:

Calcular la probabilidad de C




| A | P(A) |
|---|---|
| 0 | 0.6 |
| 1 | 0.4 |

| A | B | P(B\|A) |
|---|---|---|
| 0 | 0 | 0.8 |
| 0 | 1 | 0.2 |
| 1 | 0 | 0.4 |
| 1 | 1 | 0.6 |

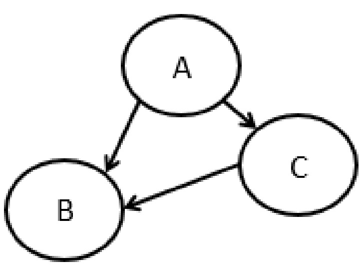
| A | C | P(C\|A) |
|---|---|---|
| 0 | 0 | 0.7 |
| 0 | 1 | 0.3 |
| 1 | 0 | 0.9 |
| 1 | 1 | 0.1 |

$$P(C=0) = \sum_a \sum_b P(A=a, B=b, C=0)$$
$$= P(A=0, B=0, C=0) + P(A=0, B=1, C=0)$$
$$+ P(A=1, B=0, C=0) + P(A=1, B=1, C=0)$$
$$= (.6 * .8 * .7) + (.6 * .2 * .7) + (.4 * .4 * .9) + (.4 * .6 * .9) = 0.78$$

$$P(C=1) = \sum_a \sum_b P(A=a, B=b, C=1)$$
$$= (.6 * .8 * .3) + (.6 * .2 * .3) + (.4 * .4 * .1) + (.4 * .6 * .1) = 0.22$$

| C | P(C) |
|---|---|
| 0 | 0.78 |
| 1 | 0.22 |

**ACTIVIDAD:** Calcule:




| A | P(A) |
|---|---|
| 0 | 0.7 |
| 1 | 0.3 |

| A | C | B | P(B\|A,C) |
|---|---|---|---|
| 0 | 0 | 0 | 0.1 |
| 0 | 0 | 1 | 0.9 |
| 0 | 1 | 0 | 0.3 |
| 0 | 1 | 1 | 0.7 |
| 1 | 0 | 0 | 0.6 |
| 1 | 0 | 1 | 0.4 |
| 1 | 1 | 0 | 0.2 |
| 1 | 1 | 1 | 0.8 |

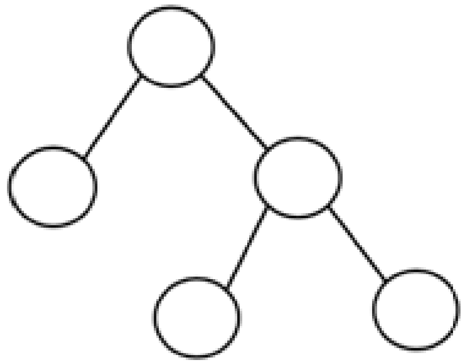
| A | C | P(C\|A) |
|---|---|---|
| 0 | 0 | 0.8 |
| 0 | 1 | 0.2 |
| 1 | 0 | 0.5 |
| 1 | 1 | 0.5 |

1. ¿Cuál es la probabilidad de B?
2. ¿Cuál es la probabilidad de A dado B?
3. ¿Cuál es la probabilidad de C?

## Creación de modelos a partir de datos de entrenamiento

**Árbol de dependencias (Chou-Liu)**

El Árbol de Dependencias busca modelar la distribución de una tabla de datos con un árbol no dirigido, donde cada enlace implica dependencia entre variables.



Nota.- Cada variable puede tener sólo UN padre

Pare crear un Árbol de dependencias se realiza lo siguiente:

1. Se crea la tabla de Informaciones mutuas.
2. Se elige como raíz uno de los nodos que forman la pareja con mayor información mutua y se crea el árbol con estos dos nodos
3. Repetir hasta que no existan variables sin conectar
   - Buscar un par de variables que: una de ellas este en la red bayesiana y la otra no, y tengan la mayor información mutua a partir de esas características.
   - Conectar la variable que no está en la red a la variable que está en la red.