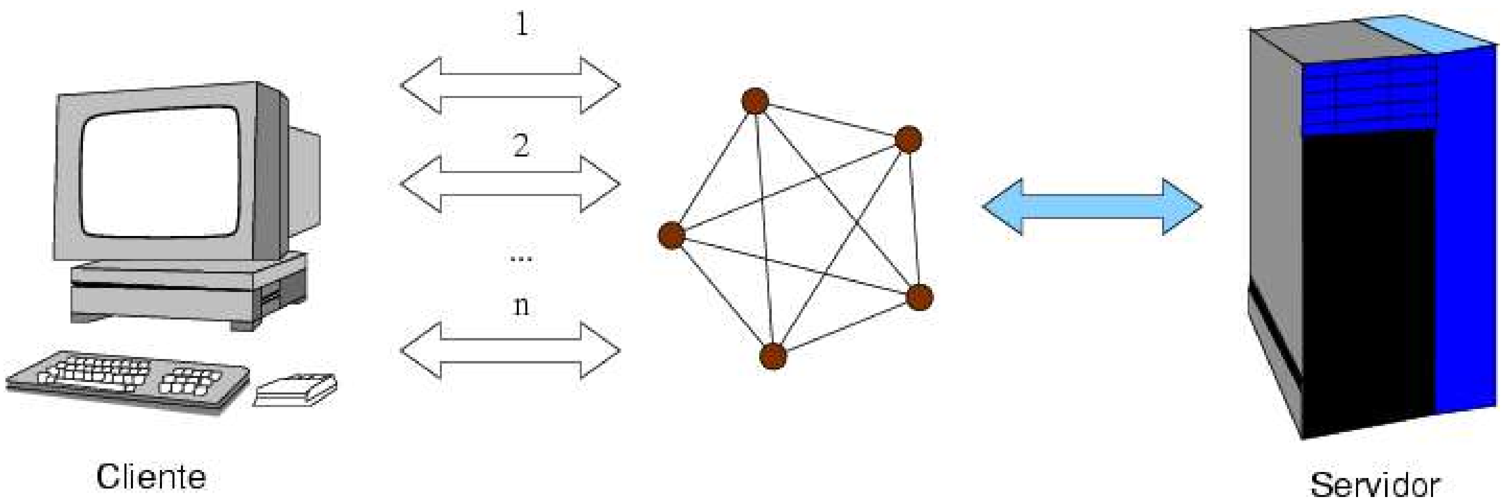
Cálculo de probabilidades. Experiencia, resultados y sucesos

Situación de partida: un cliente se conecta con un servidor remoto a través de una red de comunicaciones. El proceso consiste en *n* intentos de conexión en un periodo determinado. El resultado de cada intento depende de si se ha encontrado un camino por la red hasta el servidor, y si el servidor está en marcha y responde.



Existe incertidumbre en esta experiencia debido a dos factores:

- Puede que el servidor esté parado
- Puede que un intento dado no encuentre un camino por la red

Así que, como el resultado es incierto, vamos a enumerar el *conjunto de todos los resultados posibles* (Ω, o conjunto fundamental); usemos n=2:

- (s, c_1, c_2)
- $(s, c_1, \neg c_2)$
- $(s, \neg c_1, c_2)$
- $(s, \neg c_1, \neg c_2)$
- $(\neg s, c_1, c_2)$
- $(\neg s, c_1, \neg c_2)$
- $(\neg s, \neg c_1, c_2)$
- $(\neg s, \neg c_1, \neg c_2)$

entendiendo que $\underline{s}$ significa que el servidor responde (por tanto, $\underline{\neg s}$ que no está en marcha o no responde), y $\underline{c_i}$ que la i-ésima petición del cliente ha encontrado línea por la red (y $\underline{\neg c_i}$ que no ha llegado hasta el servidor).

Un subconjunto de Ω puede tener un interés particular: por ejemplo, todos los resultados observables si el servidor funciona. A un conjunto como este de resultados le llamamos suceso o acontecimiento.

$S = \{(s, c_1, c_2),(s, c_1, \neg c_2),(s, \neg c_1, c_2),(s, \neg c_1, \neg c_2)\}$

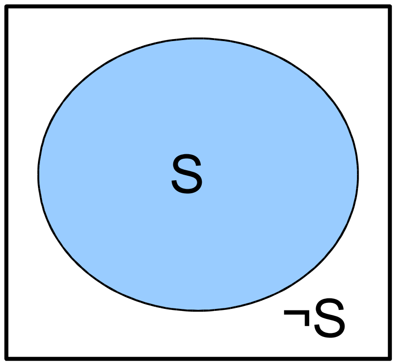
También son sucesos $\neg s$, c_1, $\neg c_2$, Ω (seguro) o $\varnothing$ (imposible).

Cualquier subconjunto de Ω (si este es discreto) es un suceso, por ejemplo: "conectar más de la mitad de *n* intentos". Por tanto, existen 2^N sucesos distintos, donde N es el número de resultados diferentes observables (si n=2, N=8, o $N=2^{n+1}$).

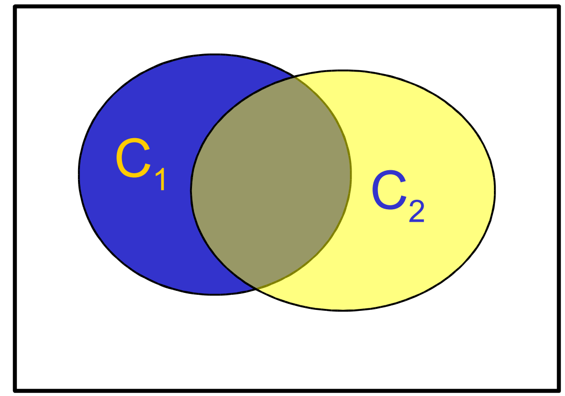
Los resultados son directamente observables. En cambio, un suceso se realiza o no (dependiendo de si el resultado de la experiencia está en el suceso o no).

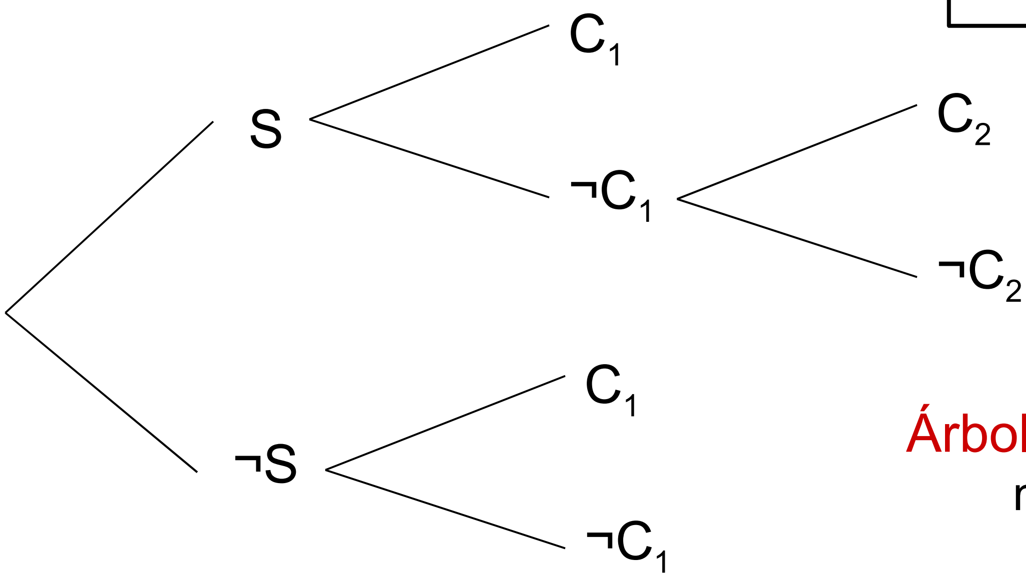
Como los sucesos son conjuntos, todas las operaciones de los conjuntos (unión, intersección, complementario) se pueden aplicar, y el resultado es otro suceso.

A menudo usamos representaciones gráficas para visualizar mejor el proceso de una experiencia indeterminista (por costumbre, empleamos símbolos en mayúscula para denotar sucesos):



Diagramas de Venn





Árboles de sucesos (no se muestra completo)

Es habitual trabajar con resultados agrupados de acuerdo con una partición. Una partición es un conjunto de sucesos A_i tales que son disjuntos y su unión cubre totalmente Ω. Repartimos todos los resultados de modo que cada uno pertenezca a uno y sólo un suceso A_i.

Obsérvese que, cada vez que desplegamos un nodo del árbol en varios subárboles, estamos utilizando una partición como criterio para efectuar la división.

- ¿responde el servidor o no?
- ¿hay un camino en el primer intento, o no lo hay?
- $C_1, C_2, \ldots, C_n$ **no** es una partición

Ejemplo: una partición basada en el número de veces que se ha contactado con el servidor. A_0: no se ha conseguido contactar, A_1: sólo se ha contactado una vez, …, A_n: se ha contactado en todos los intentos.

Compruebe que se cumplen las siguientes condiciones:

$A_0 \cup A_1 \cup \ldots \cup A_n = \Omega$; $A_i \cap A_j = \emptyset$, si $i \neq j$.

Probabilidad de un suceso.

Para poder determinar la frecuencia con la que esperamos que un suceso se realice, necesitamos saber algunas condiciones de la experiencia. ¿Con qué frecuencia el servidor no está a punto? ¿Cuán a menudo no encontramos un camino por la red? ¿Tienen relación estas dos componentes?

En nuestro caso, se supondrán las siguientes hipótesis:

- El servidor falla 1 de cada 10 veces (al azar)
- Los fallos atribuibles a la red son 1 de cada 5 (al azar)
- En un proceso de *n* intentos, el servidor no cambia de estado: o responde o no responde, y esto es independiente de la red
- Los *n* intentos de conexión efectuados por el cliente son independientes unos de otros.

Esta información nos permitirá calcular la probabilidad de cualquier suceso (incluidos los llamados *sucesos elementales: los sucesos formados por un solo resultado*).

La probabilidad se entiende como un número entre 0 y 1 que se aplica a un suceso A: P(A). Se define de manera que involucra a todos los sucesos de la experiencia que se estudia, en concreto:

- la probabilidad de la unión de sucesos disjuntos debe ser igual a la suma de las probabilidades respectivas, y
- $P(\Omega) = 1$

Ejemplos. $P(S) = 9/10$; $P(\neg C_1) = 1/5$

El concepto de independencia se aplica a dos (o más) sucesos conjuntamente: A y B son independientes $\Leftrightarrow P(A \cap B)=P(A)\cdot P(B)$.

Ejemplo. Realizamos un solo intento; ¿cuál es la probabilidad de contactar con el servidor?

P(el servidor funciona $\cap$ se encuentra línea) = $P(S \cap C_1)$ =

$P(S)\cdot P(C_1) = 9/10\cdot 4/5 = 0.72$

por ser independientes

Tomemos n=2, y veamos cómo se asigna P() a cada uno de los sucesos elementales (con una notación más simple, en números binarios: un 0 para **s** quiere decir que el servidor no responde, un 0 para **c1** que en el primer intento no se halló un camino por la red, etc):

s	c1	c2	P(S) | P(¬S)	P(C1) | P(¬C1)	P(C2) | P(¬C2)	P()
0	0	0	0.1	0.2	0.2	0.004
0	0	1	0.1	0.2	0.8	0.016
0	1	0	0.1	0.8	0.2	0.016
0	1	1	0.1	0.8	0.8	0.064
1	0	0	0.9	0.2	0.2	0.036
1	0	1	0.9	0.2	0.8	0.144
1	1	0	0.9	0.8	0.2	0.144
1	1	1	0.9	0.8	0.8	0.576
					suma:	**1**

A partir de estos datos, podemos determinar que la probabilidad de conectar al menos una vez (T_2) es 0.144+0.144+0.576=0.864.

En general, se calcula fácilmente recurriendo al complementario:

$$P(T_n) = 1-P(\neg S \cup (S \cap \neg C_1 \cap \neg C_2 \ldots \cap \neg C_n))$$

Interpretación: no conectarse nunca es el suceso complementario a conectarse alguna vez, y equivale a:

•o el servidor no funciona, …

•… o sí funciona, pero todos los intentos de hallar un camino han fallado

Se ve que ambos son disjuntos; así que podemos reescribir la probabilidad de la unión como una suma de probabilidades:

$$1-[P(\neg S) + P(S \cap \neg C_1 \cap \neg C_2 \ldots \cap \neg C_n)] = 1-[0.1+0.9 \cdot 0.2^n]$$

Como se puede comprobar, si *n* aumenta, la probabilidad de conectar tiende a 0.9 (concretamente a P(S)), lo cual es lógico.

La idea de sucesos independientes está ligada a la de sus probabilidades, es decir, a la frecuencia con que esperamos su realización. A y B son independientes cuando la probabilidad de A es la misma, indiferentemente de lo que haya ocurrido con B (o viceversa).

Para profundizar en el aspecto anterior, introducimos el concepto de probabilidad condicional: P(A|B), que se lee como "probabilidad de observar A teniendo en cuenta que se ha realizado B" (O sencillamente, "probabilidad de A condicionada por B").

Supongamos que se han realizado dos intentos de conexión, sin éxito (no sabemos si por causa de la red o del servidor). ¿Qué probabilidad tenemos de conectarnos en un tercer intento?

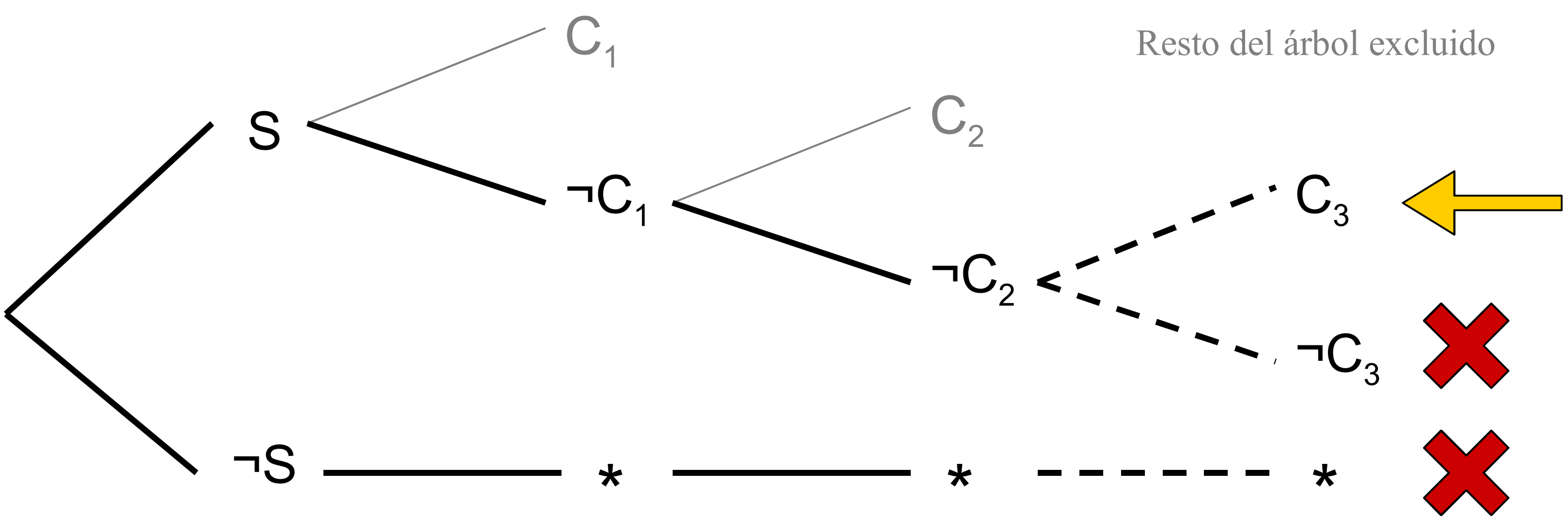
Es decir, ¿esa información parcial que tenemos de la experiencia nos va a cambiar algo la probabilidad del suceso buscado, respecto de si partimos de un conocimiento nulo a priori?

Se define P(A|B), si P(B)>0, como: $\boxed{\dfrac{P(A \cap B)}{P(B)}}$

En nuestro ejemplo, lo escribiremos así,

$$P(S \cap C_3 \mid \neg T_2) = \frac{P(S \cap C_3 \cap \neg T_2)}{P(\neg T_2)}$$

En la práctica, condicionar por un suceso B significa que reducimos a B el conjunto fundamental, y las probabilidades de los sucesos deben considerar sólo los nuevos resultados posibles.



$S \cap C_3 \cap \neg T_2 = S \cap C_3 \cap [\neg S \cup (S \cap \neg C_1 \cap \neg C_2)] = S \cap \neg C_1 \cap \neg C_2 \cap C_3$

Su probabilidad: $0.9 \cdot 0.2 \cdot 0.2 \cdot 0.8 = 0.0288$

$$P(S \cap C_3 \mid \neg T_2) = \frac{0.0288}{0.136} = 0.2118$$

Vemos que fallar en dos intentos y acertar en el tercero (y el servidor en funcionamiento) es un suceso de baja ocurrencia (menos del 3%). Pero se compensa cuando se compara con el suceso que condiciona ($\neg T_2$), tampoco muy frecuente.

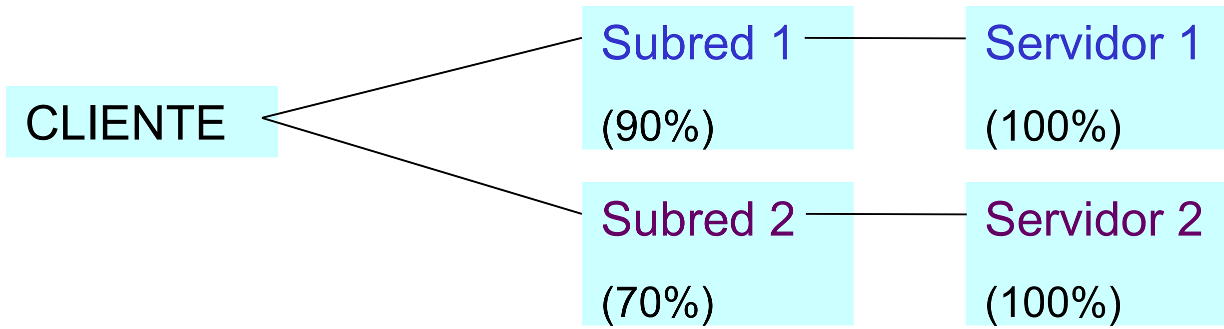
La probabilidad "bruta" de conectar en el tercer intento es $P(S \cap C_3)=0.72$. No conectar en los dos intentos previos es una mala señal: ¡baja nuestras posibilidades a un 21%!

Algunas propiedades:

- A y B son independientes $\Leftrightarrow$ $P(A|B)=P(A)$ $\Leftrightarrow$ $P(B|A)=P(B)$.
- En cualquier caso, $P(A \cap B) = P(A|B)\cdot P(B) = P(B|A)\cdot P(A)$.
- De la anterior se deduce la Fórmula de Bayes:

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

UN CASO NUEVO.



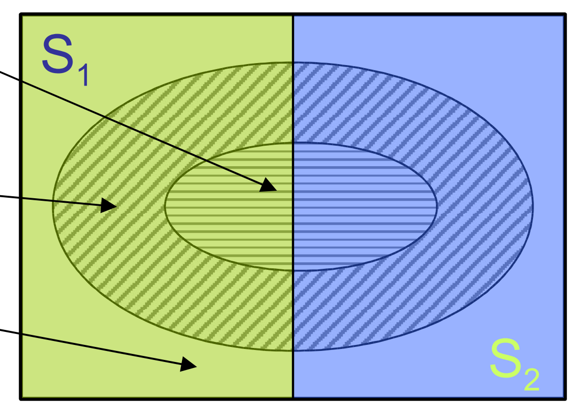
Servidor al azar (50-50). *n* intentos sobre la misma subred.

Tomemos n=2. ¿Cómo calcular la probabilidad de obtener 0, 1, 2 conexiones? Suponiendo que se han conseguido k conexiones, ¿con qué probabilidad hemos sido atendidos por el servidor i ?

Cjto. de resultados asociados a "**2 conexiones**" (X_2)

Cjto. de resultados asociados a "**1 conexión**" (X_1)

Cjto. de resultados restante: "**0 conexiones**" (X_0)



Nótese que X_0, X_1 o X_2 pueden ser expresados como unión de conjuntos disjuntos (ya que $\{S_1, S_2\}$ es una *partición*):

$$X_i = (X_i \cap S_1) \cup (X_i \cap S_2); \quad \Rightarrow$$

$$P(X_i) = P(X_i \cap S_1) + P(X_i \cap S_2) = P(X_i | S_1) \cdot P(S_1) + P(X_i | S_2) \cdot P(S_2)$$

(Ley de probabilidades totales)

Se aplica cuando disponemos de una partición, y la probabilidad del suceso de interés es sencilla de obtener si está condicionado por un elemento cualquiera de la partición (suponiendo que se ha observado tal circunstancia).

En el ejemplo, suponiendo que utilizamos o bien el servidor 1, o bien el servidor 2.

Cálculos:

$P(X_0 | S_1) = 0.1 \cdot 0.1 \quad = 0.01$

$P(X_0 | S_2) = 0.3 \cdot 0.3 \quad = 0.09$

$P(X_1 | S_1) = 0.1 \cdot 0.9 + 0.9 \cdot 0.1 \quad = 0.18$

$P(X_1 | S_2) = 0.3 \cdot 0.7 + 0.7 \cdot 0.3 \quad = 0.42$

$P(X_2 | S_1) = 0.9 \cdot 0.9 \quad = 0.81$

$P(X_2 | S_2) = 0.7 \cdot 0.7 \quad = 0.49$

$P(X_0) = 0.01 \cdot 0.5 + 0.09 \cdot 0.5 = 0.05$

$P(X_1) = 0.18 \cdot 0.5 + 0.42 \cdot 0.5 = 0.30$

$P(X_2) = 0.81 \cdot 0.5 + 0.49 \cdot 0.5 = 0.65$

cero conexiones, 1 de cada 20 veces

una conexión, 3 de cada 10 veces

dos conexiones, 13 de cada 20 veces

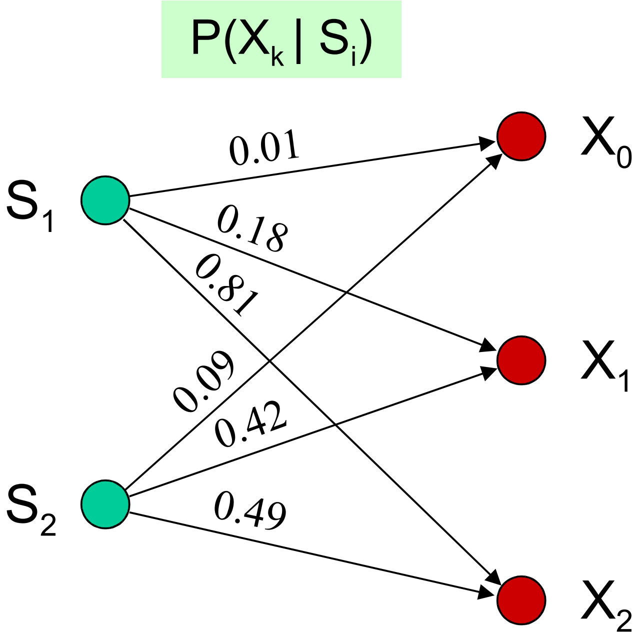
Combinando la fórmula de Bayes con la ley de probabilidades totales (y una partición $\{A_i\}$ adecuada) se obtiene el teorema de Bayes:

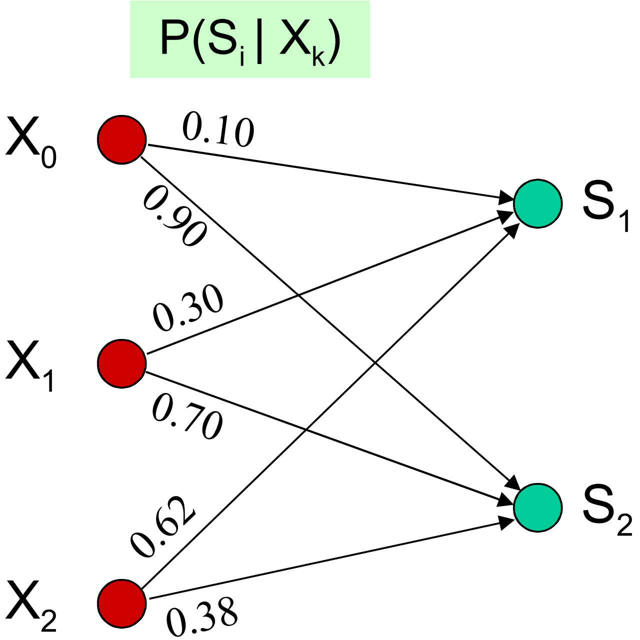
$$\boxed{P(A_i \mid B_k) = \frac{P(B_k \mid A_i) \cdot P(A_i)}{\sum_j P(B_k \mid A_j) \cdot P(A_j)}}$$

Cálculos:	*i=1*	*i=2*	**P(Xk)** $=\Sigma$		*i=1*	*i=2*
P(X0 | Si)	0.01	0.09	0.05	**P(Si | X0)**	0.10	0.90
P(X1 | Si)	0.18	0.42	0.30	**P(Si | X1)**	0.30	0.70
P(X2 | Si)	0.81	0.49	0.65	**P(Si | X2)**	0.62	0.38
P(Si)	0.5	0.5				

Si se han conseguido dos conexiones, creemos que hemos empleado el segundo servidor con probabilidad 62%.

El teorema de Bayes transforma unas probabilidades condicionadas en otras.





Conociendo el servidor que empleamos, llegamos a la probabilidad de obtener un número determinado de conexiones.

Sabiendo el número de conexiones conseguidas, llegamos a calcular la probabilidad de haber utilizado determinado servidor.

Ejercicio

Las compras de los clientes de un hipermercado permiten analizar posteriormente sus preferencias, y así descubrir que un 70% de ellos consumen refrescos de cola, un 60% adquieren patatas fritas, pero un 25% no adquiere ni uno ni otro producto.

Relacione las expresiones siguientes, y determine sus probabilidades:

- comprar refresco y patatas fritas
- comprar refresco si no compra patatas
- no comprar refresco pero sí patatas
- no comprar patatas cuando compra refresco
- no comprar patatas y sí refresco
- comprar refresco sabiendo que ha comprado patatas

- $P(\neg Pf \mid Rc)$
- $P(Rc \cap Pf)$
- $P(Rc \mid Pf)$
- $P(Rc \mid \neg Pf)$
- $P(\neg Rc \cap Pf)$
- $P(Rc \cap \neg Pf)$

Ejercicio

El parque informático de una empresa consta de 18 ordenadores de menos de 1 año, 36 de 1 a 2 años, 28 de 2 a 3, 14 de 3 a 4, y 4 con más de 4 años pero menos de 5. Se escogen dos ordenadores al azar (y diferentes) para su examen.

- Enumere el conjunto de resultados de la experiencia descrita
- Calcula la probabilidad de los sucesos elementales
- Exprese los siguientes acontecimientos, y halle su probabilidad:
  - los dos ordenadores tienen menos de un año
  - los dos ordenadores tienen la misma antigüedad
  - la diferencia entre los dos es de un año
  - la suma de sus tiempos es mayor o igual que 4 (asuma que menos de 1 año es 0, de 1 a 2 años es 1, etc.)
  - la suma es mayor de 2 y menor que 5

Ejercicio

Se prueba un procedimiento heurístico para determinar si un archivo está infectado por virus. Se determina que cuando hay infección, en un 95% de las veces aparece un cierto fragmento de código Z, mientras que si no hay infección este porcentaje desciende al 10%. Se admite que uno de cada 100 archivos está infectado. Si cierto archivo examinado presenta el código Z, ¿qué probabilidad tenemos que esté infectado?

Antes de responder,

- enumere el conjunto de resultados Ω
- asigne la probabilidad correspondiente a cada suceso elemental
- dibuje el diagrama de Venn y el árbol de sucesos correspondientes

Halle también la probabilidad de que un archivo sin código Z esté libre de virus.

Finalice el problema justificando si le parece eficaz la heurística.