E.T.S. INGENIERÍA
INFORMÁTICA

LENGUAJES Y
CIENCIAS DE LA
COMPUTACIÓN
UNIVERSIDAD DE MÁLAGA

APUNTES DE SISTEMAS INTELIGENTES II

Curso 2018-2019

Eva Millán

Departamento de Lenguajes y Ciencias de la Computación
ETSI Informática
Universidad de Málaga

Versión del 18 de febrero de 2019

TEMA 1. SISTEMAS BASADOS EN EL CONOCIMIENTO

Resultados de aprendizaje

Al finalizar el tema dedicado a los sistemas basados en el conocimiento, el estudiante deberá ser capaz de:

- Describir el concepto de sistema basado en el conocimiento (SBC).
- Conocer las partes de las que consta un SBC.
- Conocer la arquitectura propia de los SBC.
- Describir el funcionamiento de un motor de inferencias hacia atrás y un motor de inferencias hacia delante.

Contenidos

1.1 Introducción

Un sistema basado en el conocimiento (SBC) es un sistema informático capaz de emular las prestaciones de un experto humano en un área concreta de conocimiento especializado. Más concretamente, el sistema experto debe ser capaz de llevar a cabo las siguientes tareas:

- Aceptar las consultas que el usuario realice acerca de una situación dada del mundo real.
- Aceptar los datos proporcionados por el usuario acerca de esta situación, y solicitar otros datos que el sistema estime relevantes.
- Procesar esta información, en busca de una respuesta a la consulta planteada.
- Emitir la respuesta hallada, que debe ser análoga en la mayor parte de los casos a la respuesta que daría un experto humano.
- Justificar la respuesta finalmente emitida, siempre que el usuario así lo solicite.

Si es posible, es interesante dotar al SBC de un *sistema de aprendizaje*, que modificaría la base de conocimientos de modo automático en función de los casos que vaya recibiendo.

El diseño de un sistema basado en el conocimiento corre a cargo del *ingeniero de conocimiento*, que es una persona que estudia la forma en que un experto en cierta materia toma decisiones y traduce esta información de forma que un ordenador pueda emular el proceso.

1.2 Partes de un sistema basado en el conocimiento

Habitualmente, un sistema experto consta de las siguientes partes:

- La base de conocimientos, que contiene el conjunto de conocimientos expertos aplicables al dominio considerado del mundo real. Esta base permanece constante a lo largo del proceso de razonamiento y también -salvo cuando interviene el módulo de adquisición o el de aprendizaje- de una sesión a otra.
- El *motor de inferencias*, o máquina lógica, que implementa un algoritmo de manipulación de los conocimientos de forma que se alcancen soluciones a los problemas propuestos. Este elemento y el anterior constituyen el núcleo del sistema.
- La *memoria de trabajo*, que contiene los datos proporcionados o los objetivos propuestos por el usuario, y los resultados intermedios (datos deducidos o sub-objetivos generados) obtenidos por el sistema en su proceso de razonamiento. Obviamente, el contenido de este archivo va siendo diferente a lo largo de cada sesión del sistema.
- La *interfaz con el usuario (o los módulos de interfaz con los usuarios)*, que se encarga de presentar de forma comprensible las respuestas del sistema y de aceptar las entradas del usuario. Una parte fundamental es el servicio de explicaciones o de justificación, que debe ser capaz de exponer al usuario el proceso de razonamiento seguido por el sistema, de forma que se explique o justifique la respuesta proporcionada en la consulta.
- El *módulo de adquisición de conocimientos*, que permite la modificación de la base de conocimientos. Esta modificación puede ser realizada por el ingeniero del conocimiento junto con el experto, o automáticamente por el sistema.

En la Figura 1.1. se representan estos elementos, así como las interacciones entre ellos.

Figura 1. 1. Arquitectura de un sistema experto

El módulo que proporciona realmente su potencia al sistema es la *base de conocimientos*. En ella se almacena lo que necesita saber el experto -humano o informático- para realizar su tarea. Cada conocimiento es un cierto fragmento de pericia en la resolución de problemas.

Los formalismos más extendidos para representar el conocimiento en los sistemas expertos en la actualidad son las reglas y las redes bayesianas. En este tema de introducción vamos a describir someramente los sistemas basados en reglas

1.3 Sistemas basados en reglas

La *representación basada en reglas* considera que el conocimiento está constituido fundamentalmente por hechos y reglas. Los hechos son afirmaciones incondicionales acerca de algún aspecto del dominio del sistema. Las reglas son afirmaciones condicionales de la forma SI <combinación de hechos del caso> ENTONCES <acciones sobre los hechos del caso>. Sin embargo, para la *representación basada en redes bayesianas* la información debe estructurarse mediante el uso de variables y mecanismos causales entre ellas.

La representación basada en reglas estructura el conocimiento en *hechos y reglas*.

- Los *hechos* son afirmaciones incondicionales. Son también los componentes mínimos del sistema, o los datos con los que trabaja. Por ejemplo, "la temperatura del sensor s1 es alta".
- Las *reglas* son las estructuras fundamentales. Están formadas por dos elementos que se suelen escribir en las formas:

  <izq> → <der>

  SI <izq> **ENTONCES** <der>

  <izq> se llama también a veces *antecedente, premisa* o *condición*,
  <der> se llama *consecuente, conclusión* o *acción*.

<izq> y <der> están formados por uno o varios hechos, o bien por hechos generalizados (obtenidos sustituyendo constantes por variables o empleando operadores aritméticos y de comparación). Algunos ejemplos serían:

la temperatura es alta y la presión es alta → el peligro es grave.
la temperatura es X y X>100 → el peligro es grave.
el barómetro baja → el tiempo será lluvioso.

Las palabras NO, Y, O se emplean a menudo para agrupar los hechos del antecedente o del consecuente. No deben confundirse, sin embargo, con las correspondientes conectivas lógicas. Cada lenguaje o sistema define un significado diferente para estas palabras, implementando de esta forma diversas versiones del razonamiento por defecto, razonamiento no monótono, etc.

En los casos más sencillos, una base de conocimientos formada por hechos y reglas puede visualizarse como un hipergrafo llamado red de inferencias. Los nodos raíz corresponden a respuestas o hipótesis finales. Los nodos hoja corresponden a hechos primitivos que se consideran datos o evidencias y deben estar presentes en la memoria de trabajo o irse preguntando a lo largo de la sesión de consulta. Los restantes nodos corresponden a hechos o hipótesis intermedias que van siendo generadas por el sistema. Cada regla corresponde a un conector del hipergrafo, cuyo origen es <der> y cuyos hijos son los hechos de <izq>. Por ejemplo, consideremos una base de conocimiento con:

- 7 posibles hechos, representados por las proposiciones a, b, c, d, e, f, g.
- las tres reglas:

  a, b → c
  c → e
  b, f → g

La red de inferencia correspondiente se representa en la figura 1.2.

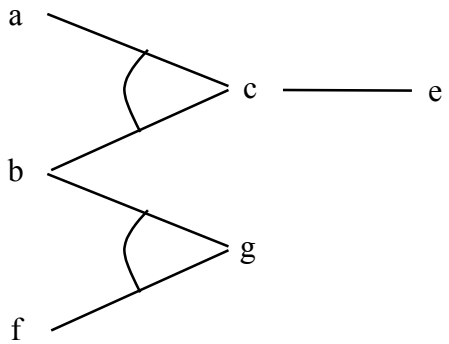


Figura 1. 2. Red de inferencias

Son respuestas finales e y g; datos o evidencias a, b y f; e hipótesis intermedia c.

1.4 Motores de inferencias en los sistemas basados en reglas

En los sistemas que representan el conocimiento mediante reglas, el motor de inferencias consiste en un algoritmo (algoritmo 1) que cíclicamente lleva a cabo estas operaciones:

- Determinar qué reglas se pueden aplicar
- Seleccionar una de ellas (resolución de conflictos)
- Ejecutar la regla seleccionada, modificando así la memoria de trabajo

Esto se repite hasta que la memoria de trabajo adopta cierta configuración que el motor de inferencias considera satisfactoria. Entonces el motor se para y se muestran los valores correspondientes. De este modo, el algoritmo para un motor de inferencias en general sería:

```
Inicializar(memoria-trabajo);
Mientras no configuración-final(memoria-trabajo)
      conjunto-reglas ← aplicables(memoria-trabajo);
      R ← resolver-conflictos(conjunto-reglas);
      memoria-trabajo ← aplicar(R, memoria-trabajo)
fin-mientras;
```

Algoritmo 1. Motor de inferencias

La descripción anterior es muy general. Para diseñar un motor de inferencias, es necesario especificar cuándo una regla es aplicable a la memoria de trabajo (*rule triggering*), cuál es el criterio de selección en caso de varias posibilidades (*resolución de conflictos*) y en qué consiste la ejecución de una regla (*rule firing*). Según se haga esto, tendremos dos grandes clases de motores:

- *con encadenamiento hacia delante o dirigidos por los datos* (*forward-chaining*, *data-driven*), en los que una regla se selecciona para su aplicación cuando su antecedente figura en la memoria de trabajo. A partir de ese momento, el proceso de razonamiento continúa, buscando soluciones del problema. Este esquema resulta apropiado para problemas cuyos dominios conlleven síntesis, como diseño, configuración, planificación, etc.
- *con encadenamiento hacia atrás o dirigidos por los objetivos* (*backward-chaining, goal-driven),* en los que se selecciona un objetivo y el sistema trata de comprobar su validez encontrando evidencias que lo apoyen. Este esquema encaja perfectamente para problemas de diagnóstico, que tienen un número pequeño de conclusiones que pueden ser extraídas, pero gran cantidad de datos iniciales. El sistema sólo pedirá datos en el momento que los necesite, y, cuando encuentre la respuesta, parará la ejecución o se dirigirá a otro objetivo. Es el esquema que se implementó en MICYN y E-MYCIN.

Los motores hacia delante (algoritmo 2) parten de ciertos hechos iniciales y consideran que una regla es aplicable cuando los hechos ya encontrados satisfacen su antecedente. Al ejecutar una regla, añaden a la memoria de trabajo, como nuevos hechos probados, lo que figura en su consecuente (que puede ser la modificación o supresión de un hecho de la memoria de trabajo). El proceso acaba cuando se añade a los hechos del caso un nuevo hecho considerado como final, o cuando ya no hay más reglas aplicables.

```
Mem-trabajo ← hechos-iniciales;
Mientras no configuración-final(mem-trabajo)
      conjunto-reglas ← match(mem-trabajo,antecedentes);
      R ← resolver-conflictos(conjunto-reglas);
      C ← consecuente(R);
      mem-trabajo ← mezclar(C, mem-trabajo)
fin-mientras;
```

Algoritmo 2. Algoritmo motor de inferencias hacia delante

Por el contrario, los motores hacia atrás (algoritmo 3) parten de un objetivo inicial y consideran que una regla está en correspondencia con la memoria de trabajo cuando su consecuente está compuesto por sub-objetivos presentes en ella. Al ejecutar una regla, sustituyen estos sub-objetivos por los que figuran en el antecedente. El proceso acaba cuando todos los sub-objetivos presentes son datos proporcionados por el usuario o hechos de la base de conocimientos.

```
mem-trabajo ← obj-iniciales;
Mientras no configuración-final(mem-trabajo)
      obj ← seleccionar(mem-trabajo);
      conjunto-reglas ← match(obj, consecuentes);
      R ← resolver-conflictos(conjunto-reglas);
      A ← antecedente(R);
      mem-trabajo ← mezclar(A, mem-trabajo)
fin-mientras;
```

Algoritmo 3. Motor de inferencias hacia atrás

Es posible diseñar motores de inferencias que funcionen parcialmente hacia delante y parcialmente hacia atrás. De este modo, y, partiendo de los datos del caso, se puede buscar aquellas hipótesis que son plausibles (con encadenamiento hacia delante). Una vez seleccionadas las hipótesis, y, mediante el encadenamiento hacia atrás, intentaremos demostrar cuáles de ellas son válidas pidiendo más información al usuario.

El momento adecuado para parar el proceso de razonamiento depende tanto del dominio en cuestión como del modelo de tratamiento de la incertidumbre que estemos usando y el efecto que queramos conseguir. Por ejemplo, en un dominio en que se pretenda clasificar un objeto y en el que el conjunto de clases posibles sea exhaustivo y excluyente, no tendrá sentido continuar, puesto que una vez asignado un objeto a una clase no puede pertenecer a otras. Sin embargo, si el sistema está determinando probabilidades de pertenecer a clases puede que solamente sea sensato parar cuando determinada clase alcance una probabilidad superior a cierto umbral prefijado. En dominios de diagnóstico, en los que posiblemente haya más de una avería, puede que tenga sentido continuar el proceso de inferencias hasta el final (o puede que no, si la hipótesis alcanzada se considera lo suficientemente plausible). Para decidir cuándo debe parar el motor, lo que hay que definir es qué se entiende por `configuración-final(memoria_trabajo)`. Podemos por ejemplo decir que es aquella que contiene al menos una hipótesis final (si queremos que el razonamiento pare en el momento que se alcance una conclusión), o aquella que contiene al menos una

hipótesis final con un grado que consideremos aceptable (un factor de certeza de 0.95; una probabilidad de 0.9), o aquella que hace que la agenda esté vacía. Esta definición determinará el momento en el que el motor de inferencias se detiene.

Ejemplo 1. *Funcionamiento de los motores de inferencia*

El funcionamiento de los motores de inferencia se explica mejor con un ejemplo sencillo. Supongamos un sistema que intenta diagnosticar una avería en un coche, y que dispone de las siguientes reglas:

SI el motor no se enciende y el motor recibe gasolina,
ENTONCES el problema es de las bujías
SI el motor no enciende y las luces no se encienden,
ENTONCES el problema es de la batería
SI el motor no se enciende y las luces encienden,
ENTONCES el problema es del arranque
SI hay gasolina, ENTONCES el motor recibe gasolina

Nuestro problema es averiguar qué le ocurre al coche, dados ciertos hechos que son directamente observables. Hay tres problemas posibles con el coche: bujías, batería, arranque. Supongamos que el conjunto de los datos iniciales es vacío.

Un sistema dirigido por los objetivos trataría de comprobar cada posible problema del coche. Supongamos que primero intenta ver si el problema es de las bujías. Entonces, OBJ = {problemas con la bujía?}. Buscaríamos en la base de conocimientos qué reglas hay que tengan "problemas con la bujía" en su consecuente. Solamente está la regla 1, así que los hechos contenidos en su antecedente se convierten en nuestros nuevos objetivos, es decir, OBJ = {motor recibe gasolina? motor enciende?}. Supongamos que el sistema intenta probar primero si el motor recibe gasolina. Este hecho es consecuente de la regla 4, luego el nuevo objetivo es entonces el antecedente de la regla 4, OBJ = {hay gasolina?}. Dado que no es el consecuente de ninguna regla, el sistema preguntará al usuario:

¿Hay gasolina?

Supongamos que la respuesta es que sí. El motor de inferencia añadiría al conjunto de datos del caso, como hecho ya probado, que hay gasolina, de forma que esta pregunta no vuelva a plantearse de nuevo.

Una vez comprobado este objetivo, el conjunto de objetivos a probar es OBJ={motor enciende?}. Dado que no es consecuente de ninguna regla, el sistema preguntará al usuario:

¿Enciende el motor?

Supongamos que la respuesta es que no. Tenemos entonces que el antecedente de la regla 1 se satisface, y por tanto se considera probado que el coche tiene un problema de batería. Dependiendo de cómo se haya implementado el sistema, algunos podrían paren en ese momento, mientras que otros podrían continuar razonando. Supongamos que hemos definido que la configuración final de la memoria de trabajo se alcanza cuando la agenda está vacía. En ese caso, intentaríamos probar la siguiente hipótesis, "problemas con la batería", que es el consecuente de la regla 2. Sus antecedentes se convierten en los nuevos objetivos. Como en los datos del caso figura que el motor no enciende, el objetivo ahora es probar si las luces encienden o no. Se preguntaría al usuario:

¿Encienden las luces?

Supongamos que la respuesta es sí. Se añade a los datos del caso, y se considera que no se ha podido demostrar si hay un problema de batería o no. El sistema intentaría ahora comprobar si hay problemas con el arranque, pero dados los datos del caso (el motor no funciona y las luces no encienden), el sistema no puede inferir nada acerca del sistema de arranque. La interacción completa con este sistema tan simple sería:

Sistema ¿Hay gasolina?
Usuario. Sí.
Sistema ¿Enciende el motor?
Usuario. No.
Sistema ¿Encienden las luces?
Usuario. Sí
Sistema. Creo que el problema es de las bujías.

Nótese que, en general, resolver problemas utilizando el encadenamiento hacia atrás conlleva un proceso de búsqueda en el espacio de las posibles formas de resolver un problema, comprobándolas todas sistemáticamente.

Otro enfoque consiste en utilizar el encadenamiento hacia delante. Se trabaja a partir de los datos iniciales del caso. Supongamos en nuestro ejemplo que los datos iniciales son: {hay gasolina, no enciende el motor, no encienden las luces}.

Se intenta ver qué reglas tienen estos hechos en sus antecedentes. En principio, la única regla aplicable es la regla 4, ya que es la única cuyo antecedente está en el conjunto de datos iniciales. Tras la aplicación de la regla 4, añadimos motor_recibe_gasolina a los datos del caso. Nos encontramos ahora con que es aplicable la regla 1, indicando que el problema es de las bujías.

1.5 Resolución de conflictos en sistemas basados en reglas

Al determinar las reglas aplicables, es posible que varias reglas se satisfagan simultáneamente en un paso del proceso. El motor debe decidir entonces cuál de estas reglas se ejecutará. Es la llamada *resolución de conflictos*. Si consideramos que cada combinación de hechos que se presenta en la memoria de trabajo representa un estado del problema que se resuelve, y que las reglas representan los operadores de transición entre estados, la cuestión se reduce a determinar la estrategia de búsqueda en el espacio de estados.

A) CRITERIOS ESTÁTICOS. La prioridad de cada regla queda determinada cuando se codifica la base de conocimientos. Esto se puede hacer de varias formas:

- *Orden textual de las reglas*. Las reglas anteriores en la base tienen prioridad sobre las posteriores.
- *Utilidad explícita de reglas*. Cada regla tiene asociado un valor numérico. Tienen prioridad las reglas con mayor valor.
- *Utilidad explícita de hechos*. Cada hecho tiene asociado un valor numérico. Tienen prioridad las reglas que añadan o modifiquen hechos de mayor valor. Si el valor numérico se refiere al tiempo que llevan los hechos que activan la regla en la lista de hechos, tenemos las estrategias LEX (menos tiempo) o MEA (más tiempo).

- *Especificidad*. Si el antecedente de R1 subsume al de R2, entonces R2 tiene prioridad sobre R1. O bien si el consecuente de R1 subsume al de R2, entonces R2 tiene prioridad.
- *Generalidad*. Lo contrario a lo anterior.

B) CRITERIOS DINÁMICOS U OPORTUNÍSTICOS. La prioridad se determina en cada paso del motor. Algunos de los criterios más comunes son:

- *Mínima espera.* Tienen prioridad las reglas que llevan menos tiempo en el conjunto de reglas aplicables.
- *Máxima espera.* Tienen prioridad las reglas que llevan más tiempo en el conjunto de reglas aplicables.

C) CRITERIOS DINÁMICOS MANIPULABLES. Los parámetros que sirven para determinar la prioridad se pueden modificar en cada paso del motor. Ello se consigue mediante el uso de meta-reglas, que se aplican en cada paso antes de cualquier regla. Por ejemplo, la meta-regla:

SI (s1 temperatura NO-CALCULADA) y (s2 temperatura alta)
ENTONCES UTILIDAD(s1 temperatura) $\leftarrow$ 100.

establece un valor muy alto para la utilidad del parámetro "temperatura" del sensor s1 siempre que en el sensor s2 la temperatura sea alta.

1.6 Ejercicios propuestos

Ejercicio 1.1. Supongamos un sistema con las siguientes reglas:

R1: SI el motor no se enciende y el motor recibe gasolina, ENTONCES el problema es de las bujías,
R2: SI el motor no enciende y las luces no se encienden, ENTONCES el problema es de la batería,
R3: SI el motor no se enciende y las luces encienden, ENTONCES el problema es del arranque,
R4: SI hay gasolina, ENTONCES el motor recibe gasolina.

Supongamos que los datos del caso son {hay gasolina, motor no enciende, luces encienden}. Describe el funcionamiento del motor de inferencias bajo los siguientes supuestos:

- Supuesto 1. Motor de inferencias hacia delante, resolución conflictos orden de reglas, configuración final cuando se demuestre una hipótesis
- Supuesto 2. Motor de inferencias hacia delante, resolución conflictos orden inverso de reglas, configuración final cuando no se puedan ejecutar más reglas
- Supuesto 3. Motor de inferencias hacia atrás, configuración final cuando se demuestre una hipótesis.

Ejercicio 1.2. (Adaptado de Winston, 1992). Consideremos las siguientes reglas:

R1: SI un animal tiene pelo o da leche, ENTONCES es mamífero
R2: SI un animal tiene plumas o vuela y pone huevos, ENTONCES es un ave
R3: SI un animal es mamífero y come carne, ENTONCES es carnívoro
R4: SI un animal tiene dientes puntiagudos y tiene garras y tiene ojos saltones ENTONCES es carnívoro
R5: SI un animal es mamífero y tiene pezuñas ENTONCES es un ungulado
R6: SI un animal es mamífero y rumia ENTONCES es un ungulado

R7: SI un animal es mamífero y es carnívoro y tiene color leonado y tiene manchas oscuras ENTONCES es un leopardo
R8: SI un animal es mamífero y es carnívoro y tiene color leonado y tiene rayas negras ENTONCES es un tigre
R9: SI un animal es ungulado y tiene cuello largo y tiene piernas largas y tiene manchas oscuras ENTONCES es una jirafa
R10: SI un animal es un ungulado y tiene rayas negras ENTONCES es una cebra
R11: SI un animal es ave y no vuela y tiene el cuello largo y tiene piernas largas y tiene color blanco y negro ENTONCES es un avestruz
R12: SI un animal es ave y no vuela y nada y tiene color blanco y negro, ENTONCES es un pingüino
R13: SI es un ave y vuela bien, ENTONCES es un albatros

Describe cómo funcionará el proceso de inferencias si:

a) Tenemos un motor de inferencias hacia delante y en la memoria de trabajo tenemos un animal llamado Robbie que vuela, pone huevos, y tiene cuello largo
b) Tenemos un motor de inferencias hacia detrás y en la memoria de trabajo tenemos un animal llamado Jimmy, que tiene pelo, dientes puntiagudos, garras, ojos saltones, rayas negras, color leonado, e intentamos probar que Jimmy es un tigre.

Ejercicio 1.3. (Tomado del libro de Castillo, Gutiérrez y Hadi, tema 2). Cuatro agentes secretos, Alberto, Luisa, Carmen y Tomás, están en uno de los cuatro países: Egipto, Francia, Japón y España. Se han recibido los siguientes telegramas de los agentes:

- De Francia: Luisa está en España.
- De España: Alberto está en Francia.
- De Egipto: Carmen está en Egipto.
- De Japón: Carmen está en Francia.

El problema radica en que no se sabe quién ha enviado cada uno de los mensajes, pero es conocido que Tomás miente (¿es un agente doble?) y que los demás agentes dicen la verdad. ¿Quién está en cada país?

1.7 Bibliografía

Castillo, E., Gutiérrez, J., & Hadi, A. (1997). *Sistemas Expertos y Modelos de Redes Probabilísticas.* Monografías de la Academia Española de Ingeniería.
Gómez, A., Juristo, N., Montes, C., & Pazos, J. (1997). *Ingeniería del Conocimiento.* Editorial Centro de Estudios Ramón Areces, S.A.

TEMA 2. REDES BAYESIANAS

Resultados de aprendizaje

Al finalizar este tema, los estudiantes deberán ser capaces de:

- Modelar una situación real con una red bayesiana (nodos, enlaces, parámetros).
- Entender el significado e importancia de las condiciones de independencia condicional.
- Aplicar algoritmos de propagación básicos.
- Predecir la evolución de las probabilidades en la red conforme se adquiere nueva evidencia.
- Manejar herramientas de implementación de redes bayesianas (GENIE).

Contenidos

2.1 Introducción: el método probabilístico clásico

Para introducir el uso del razonamiento Bayesiano en los sistemas basados en el conocimiento, comenzaremos con un pequeño ejemplo.

Ejemplo 2. 1. Modelo para un pequeño SBC basado en teoría de la probabilidad

Supongamos que estamos estudiando la relación entre las variables Edad, Obesidad, Hernia de hiato (Hernia), Indigestión y Vómitos en una población que consta de 20 individuos:

Individuos	Edad	Obesidad	Hernia	Indigestión	Vómitos
Individuo 1	Mayor_50	no	no	no	no
Individuo 2	Mayor_50	no	no	no	no
Individuo 3	Mayor_50	no	no	no	no
Individuo 4	Mayor_50	no	no	no	no
Individuo 5	Mayor_50	no	sí	no	sí
Individuo 6	Mayor_50	sí	sí	no	sí
Individuo 7	Mayor_50	sí	sí	no	sí
Individuo 8	Mayor_50	sí	sí	no	si
Individuo 9	Menor_50	no	no	no	no
Individuo 10	Menor_50	No	no	no	no
Individuo 11	Menor_50	no	no	no	no
Individuo 12	Menor_50	no	no	no	no
Individuo 13	Menor_50	no	no	no	no
Individuo 14	Menor_50	no	no	no	no
Individuo 15	Menor_50	no	no	no	no
Individuo 16	Menor_50	no	no	no	no
Individuo 17	Menor_50	no	no	sí	sí
Individuo 18	Menor_50	sí	no	no	sí
Individuo 19	Menor_50	sí	no	sí	sí
Individuo 20	Menor_50	sí	no	no	no

Como tenemos los datos de toda la población, es sencillo calcular la distribución de probabilidad conjunta, simplemente contando las frecuencias. De este modo, tendríamos que:

P(Edad= Mayor_50; Obesidad=no; Hernia=no; Indigestión=no, Vómitos=no) = 4/20
P(Edad= Mayor_50; Obesidad=no; Hernia=sí; Indigestión=no, Vómitos=sí) = 1/20
P(Edad= Mayor_50; Obesidad=sí; Hernia=sí; Indigestión=no, Vómitos=sí) = 3/20
P(Edad= Menor_50; Obesidad=no; Hernia=no; Indigestión=no, Vómitos=no) = 8/20
P(Edad= Menor_50; Obesidad=no; Hernia=no; Indigestión=si, Vómitos=si) = 1/20
P(Edad= Menor_50; Obesidad=sí; Hernia=no; Indigestión=no, Vómitos=sí) = 1/20
P(Edad= Menor_50; Obesidad=sí; Hernia=no; Indigestión=si, Vómitos=sí) = 1/20
P(Edad= Menor_50; Obesidad=sí; Hernia=no; Indigestión=no, Vómitos=no) = 1/20

Pero para dar la distribución conjunta completa, aún no hemos terminado. En efecto, como tenemos cinco variables binarias, la distribución conjunta consta de 2^5 valores, de los cuales ya hemos especificado 8. ¿Qué ocurre con el resto? Pues que todos valen cero, al no haber ningún caso.

Dada esta probabilidad conjunta, ahora podríamos utilizarla para calcular otras probabilidades: Por brevedad, renombremos cada variable con el nombre de su primera letra (en mayúscula). Así por ejemplo podríamos calcular las distribuciones marginales asociadas:

P(Edad) que sería:

- $P(E= Mayor\_50) = \sum_{O,H,I,V} P(E > Mayor > 50, O, H, I, V) = 8/20$
- $P(E= Menor\_50) = 1- P(E= Mayor\_50) = 12/20$

O también P(Edad, Hernia):

- $P(E = Mayor\_50, H=sí) = \sum_{O,I,V} P(E > Mayor > 50, O, H = sí, I, V) = 4/20$
- $P(E = Mayor\_50, H=no) = \sum_{O,I,V} P(E > Mayor > 50, O, H = no, I, V) = 4/20$
- $P(E = Menor\_50, H=sí) = 0$
- $P(E = Menor\_50, H=no) = 12/20$

Y también hacer inferencias, de dos tipos: diagnósticos y predicciones. Por ejemplo, una inferencia tipo diagnóstico sería calcular la probabilidad de que una persona que tenga (o no) como síntoma vómitos padezca una hernia P(Hernia/Vómitos):

- $P(H=si/V=si) = \frac{P(H=si, V=sí)}{P(V=si)} = 4/7$
- $P(H=si/V=no) = \frac{P(H=si, V=no)}{P(V=no)} = 0$
- $P(H=no/V=si) = 3/7$
- $P(H=no/V=no) = 1$

En cuanto a inferencias predictivas, podríamos por ejemplo calcular cuál es la probabilidad de que una persona obesa y con hernias vomite:

$$P(V=sí/O=sí, H=sí) = \frac{P(V=si, O=sí, H=si)}{P(O=si, H=si)} = \frac{3/20}{3/20} = 1$$

Como vemos, la distribución conjunta nos permite calcular cualquier distribución de interés. ¿Qué problema plantea entonces en uso de esta distribución? De un lado, el número de parámetros requeridos. En efecto, en este problema son 32 valores, pero en un problema en el que hubiese 10 variables binarias, son 2^{10}. Es decir, el número de parámetros crece de forma exponencial conforme aumenta el número de variables y sus posibles valores. De otro, este crecimiento exponencial afecta también a los cálculos que son necesarios realizar para calcular las inferencias. En un ejemplo pequeño como este son sencillos, pero en el caso general, el procedimiento es computacionalmente complejo. Por eso se hace necesario buscar otras alternativas viables a dicho modelo, como son las redes bayesianas.

Ejercicio 2.1. Considérense el siguiente conjunto de datos;

Tiempo (T)	Temperatura(ºC) (Te)	Humedad(%) (H)	Viento (V)	Jugar al tenis (JT)
soleado	29	65	falso	no
soleado	27	70	verdadero	no
cubierto	28	66	falso	sí
lluvioso	21	76	falso	sí
lluvioso	20	60	falso	sí
lluvioso	18	50	verdadero	no
cubierto	18	45	verdadero	sí
soleado	22	75	falso	no
soleado	21	50	falso	sí
lluvioso	24	60	falso	sí
soleado	24	50	verdadero	sí
cubierto	22	70	verdadero	sí
cubierto	27	55	falso	sí
lluvioso	22	71	verdadero	no

Se pide,

a) Discretiza la variable `Temperatura` en los siguientes valores: "≤20º"; "entre 20º y 26º";"≥26º".
b) Discretiza la variable `Humedad` en los siguientes valores: "< 60%"; "≥ 60%"
c) Calcula las siguientes probabilidades:
   - P(T=soleado, Te= ≥ 26º, H≥60%,V=falso, JT=no)
   - P(T=soleado, Te= ≥ 26º, H≥60%,V=falso)
   - P(JT=no/ T=soleado, Te= ≥ 26º, H≥60%,V=falso)
   - P(JT=no/Te= ≥ 26)
   - P(T=soleado)

2.2 Presentación intuitiva

Antes de presentar formalmente la teoría matemática de las redes bayesianas, explicaremos mediante ejemplos sencillos[1] el significado intuitivo de los conceptos que después introduciremos de un modo más formal.

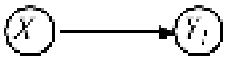
En una red bayesiana, cada nodo corresponde a una variable, que a su vez representa una entidad del mundo real. Por tanto, de aquí en adelante hablaremos indistintamente de nodos y variables, y los denotaremos con letras mayúsculas, como X. Utilizaremos la misma letra en minúscula, x, para referirnos a un valor cualquiera de la variable X. Los arcos que unen los nodos indican relaciones de influencia causal.

Una red bayesiana consta de nodos, enlaces y parámetros

Ejemplo 2. 2. La red* bayesiana *más simple

La red bayesiana no trivial más simple que podemos imaginar consta de dos variables, que llamaremos X e Y1, y un arco desde la primera hasta la segunda.

[1] Estos ejemplos están tomados de "Apuntes de razonamiento aproximado" de Francisco Javier Díez.



Para concretar el ejemplo, supongamos que X representa paludismo e Y_1 representa gota gruesa, que es la prueba más habitual para detectar la presencia de dicha enfermedad.

Cuando X sea una variable binaria, denotaremos por $+x$ la presencia de aquello a lo que representa y por $\neg x$ a su ausencia. Así, por ejemplo, en este caso $+x$ significará "el paciente tiene paludismo" y $\neg x$ "el paciente no tiene paludismo; $+y_1$ significará un resultado positivo del test de la gota gruesa y $\neg y_1$ un resultado negativo.

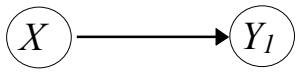
Los datos numéricos necesarios son:
- Probabilidades a priori de nodos sin padres
- Probabilidad condicionada de los demás

La información cuantitativa de una red bayesiana viene dada por:

- La probabilidad a priori de los nodos que no tienen padres.
- La probabilidad condicionada de los nodos con padres.

Por tanto, en nuestro ejemplo, los datos que debemos conocer son P(x) y P(y1/x).

Así, la red bayesiana completa sería:



$P(+x) = 0.003$ $\quad P(+y_1/+x) = 0.992$

$P(+y_1/\neg x) = 0.0006$

Veamos qué significado tienen en este caso estos valores:

- $P(+x) = 0.003$ indica que, a priori, un 0.3% de la población padece el paludismo. En medicina, esto se conoce como *prevalencia* de la enfermedad.
- $P(+y_1/+x) = 0.992$ indica que cuando hay paludismo, el test de la gota gruesa da positivo en el 99.2% de los casos. Esto se conoce como *sensibilidad* del test.
- $P(+y_1/\neg x) = 0.0006$ indica que, cuando no hay paludismo, el test de la gota gruesa da positivo en el 0.06% de los casos, y negativo en el 99.94%.A esta segunda probabilidad se la llama *especificidad* del test.

Asociados a un test tenemos dos parámetros:
- Sensibilidad (probabilidad de resultado positivo si enfermo)
- Especificidad (probabilidad de resultado negativo si no enfermo)

En medicina siempre se buscan las pruebas con mayor grado de sensibilidad y especificidad.

Alternativamente, se habla también de las *tasas de falsos positivos* (probabilidad de que el test de positivo si la persona no está enferma) y *tasas de falsos negativos* (probabilidad de test negativo cuando la persona está enferma).

Conociendo estos datos, podemos calcular:

Con la prevalencia, la sensibilidad y la especificidad, es posible calcular la probabilidad de que un paciente esté enfermo según el resultado de su test

a) La probabilidad a priori de Y1,

$P(+y_1) = P(+y_1/+x)\,P(+x) + P(+y_1/\neg x)\,P(\neg x) = 0.00357.$

$P(\neg y_1) = P(\neg y_1/+x)\,P(+x) + P(\neg y_1/\neg x)\,P(\neg x) = 0.99643.$

b) Las probabilidades a posteriori dada una evidencia observada e, P*(x) = P(x/e).

Supongamos que el test de la gota gruesa ha dado positivo. ¿Qué probabilidad hay ahora de que la persona padezca la enfermedad? Si la prueba tuviese fiabilidad absoluta, esta probabilidad sería del 100%. Pero como existe la posibilidad de que haya habido un falso positivo, buscamos P*(+x) = P(+x/+y1). Para calcularla, podemos aplicar el teorema de Bayes:

$$P^*(+x) = P(+x/+y_1) = \frac{P(+x)\ P(+y_1/+x)}{P(+y_1)} = \frac{0.003\,0.992}{0.00357} = 0.83263$$

Es decir, de acuerdo con el resultado de la prueba, hay un 83,2% de probabilidad de que el paciente tenga paludismo.

De la misma forma podríamos calcular *P(¬x):*

$$P^*(\neg x) = P(\neg x/+y_1) = \frac{P(\neg x)\ P(+y_1/\neg x)}{P(+y_1)} = \frac{0.997\,0.0006}{0.00357} = 0.16737$$

Que, por supuesto, es la probabilidad complementaria.

La expresión general del teorema de Bayes que hemos utilizado es:

$$P^*(x) = P(x/y) = \frac{P(x)\ P(y/x)}{P(y)}$$

Por razones que quedarán claras más adelante, vamos a reescribirla como;

$$P^*(x) = \alpha\ P(x)\ \lambda_{Y_i}(x)$$

Donde $\alpha = [P(y)]^{-1}$ y $\lambda_{Y_i}(x) = P(y/x)$.

Cuando una variable se instancia o cambia su probabilidad, informa a su padre a través del paso de un λ-mensaje. En base a este mensaje, el padre actualiza su probabilidad

Con la fórmula expresada de esta forma, queda claro que la probabilidad a posteriori de la variable *X* depende fundamentalmente de la probabilidad a priori de X (prevalencia de la enfermedad) y de la probabilidad condicionada de *Y* dado *X* (sensibilidad y especificidad del test), puesto que α juega simplemente el papel de una constante de normalización.

Utilizando esta nueva expresión, podemos repetir los cálculos:

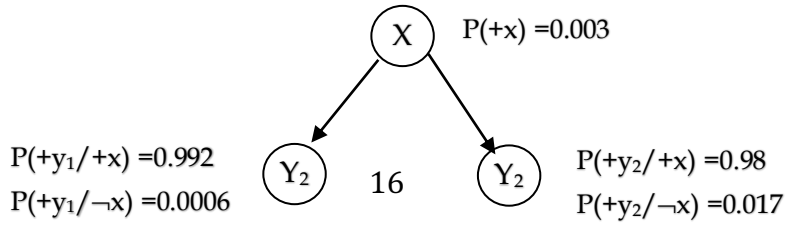
P(+x)* = α 0.003 0.992 = 0.00298 α.
P(¬x)* = α 0.997 0.0006 = 0.000598 α.

Y normalizando obtenemos el mismo resultado que antes.

Para el caso en que el test de la gota gruesa diese negativo, la probabilidad a posteriori de padecer paludismo se calcula con un procedimiento totalmente análogo.

Ejemplo 2. 3. **Una red bayesiana con tres nodos**

Supongamos que ampliamos el modelo anterior añadiendo un nuevo efecto del paludismo, la fiebre, que representaremos mediante la variable Y_2. La red bayesiana se modifica entonces y queda como se muestra en la figura:



Vemos aquí que, para el paludismo, la fiebre tiene menor especificidad que la gota gruesa. Así, este sencillo modelo tiene en cuenta que hay muchas otras causas que pueden producir fiebre.

Veamos qué tipo de conclusiones podemos extraer a partir de esta información.

a) Supongamos que e = {+y_2}. Entonces, podemos calcular como antes la probabilidad a posteriori de que el paciente tenga paludismo sabiendo que tiene fiebre:

   $P^*(+x) = P(+x/+y_2) = \alpha\ 0.003\ 0.98 = 0.00294\ \alpha = 0.148$
   $P^*(\neg x) = P(\neg x/+y_2) = \alpha\ 0.997\ 0.0017 = 0.016949\ \alpha = 0.852.$

   Como podemos observar hay solo un 0,148 de probabilidad de que tenga paludismo, resultado mucho más bajo que antes.

b) Supongamos que e = {+y_1, +y_2}. ¿Cuál es ahora $P^*(x) = P(x/ +y_1, +y_2)$?

   Para calcularla, usamos de nuevo el teorema de Bayes:

$$P^*(x) = P(x/y_1, y_2) = \frac{P(x)\ P(y_1, y_2 / x)}{P(y_1, y_2)}$$

   Pero ahora vemos que hay datos del problema que no conocemos, como $P(y_1,y_2)$ y $P(y_1,y_2/x)$. Para poder seguir nuestros cálculos, necesitamos realizar unas hipótesis adicionales que se llaman *hipótesis de independencia condicional*. En concreto, vamos a suponer que las variables Y_1 e Y_2 son independientes dados su padre común, X, es decir:

   $P(y_1,y_2/x) = P(y_1/x)\ P(y_2/x).$

   Si suponemos esto podremos continuar con los cálculos porque $P(y_1, y_2)$ se obtendrá como constante de normalización.

   ¿Qué significa aceptar esta hipótesis? Significa aceptar que, conocido que un paciente tiene paludismo, el hecho de que tenga fiebre o no, no depende de que el resultado del test de la gota gruesa sea positivo o negativo, lo cual parece razonable.

   Para seguir con la nueva formulación que introdujimos en el ejemplo 1, vamos a denotar por $\lambda(x)$ a $\lambda_{Y_1}(x)\ \lambda_{Y_2}(x)$. Entonces tendríamos que

   $P^*(x) = \alpha\ P(x)\ \lambda\ (x).$

Las hipótesis de independencia condicional permiten en este ejemplo realizar el cálculo de las probabilidades a posteriori

En nuestro ejemplo, $e = \{+y_1, +y_2\}$, y por tanto;

$\lambda(+x) = \lambda_{Y1}(+x)\ \lambda_{Y2}(+x) = 0.97216$
$\lambda(\neg x) = \lambda_{Y1}(\neg x)\ \lambda_{Y2}(\neg x) = 0.000102$

Por tanto,

$P^*(+x) = 0.9663$
$P^*(\neg x) = 0{,}0347$

Como era de esperar, cuando tenemos dos evidencias en favor del paludismo, la probabilidad resultante es mayor que la correspondiente a cada uno de ellos por separado.

c) En el caso en que tengamos un hallazgo a favor y otro en contra, podemos ponderar su influencia mediante estas mismas expresiones. Supongamos por ejemplo que el test de la gota gruesa ha dado negativo, pero que el paciente padece fiebre, esto es, e ={¬y1, +y2}, y calculemos la probabilidad de que el paciente padezca fiebre.

$\lambda(+x) = \lambda_{Y1}(+x)\ \lambda_{Y2}(+x) = P(\neg y1/+x)\ P(+y2/+x) = 0.008\ 0.98 = 0.00784$
$\lambda(\neg x) = \lambda_{Y1}(\neg x)\ \lambda_{Y2}(\neg x) = P(\neg y1/\neg x)\ P(+y2/\neg x) = 0.994\ 0{,}017 = 0{,}01699$

Vemos que hay más evidencia a favor de ¬x que a favor de +x, debido principalmente a la alta sensibilidad de la prueba de la gota gruesa. Al tener en cuenta además la probabilidad a priori de la enfermedad nos queda que,

$P^*(+x) = 0.0014$
$P^*(\neg x) = 0.9986$

d) Aún podemos extraer más información de este ejemplo. Supongamos ahora que la que tenemos un paciente con fiebre al que aún no hemos realizado el test de la gota gruesa, es decir, la evidencia considerada es $e = \{+y_2\}$. ¿Cuál es la probabilidad de que, al hacerle el test, el resultado sea positivo?

Buscamos ahora $P(y_1/+y_2)$. Por teoría elemental de probabilidad, sabemos que;

$$P^*(y_1) = P(y_1/y_2) = \sum_x P(y_1/x, y_2)\, P(x/y_2) = \sum_x P(y_1/x, y_2)\, \frac{P(x, y_2)}{P(y_2)}.$$

Aplicando la hipótesis de independencia condicional y llamando

$\pi_{Y1}(x) = P(x, y_2) = P(x)\ P(y_2/x)$
$\alpha = [P(y_2)]^{-1}$,

la expresión anterior nos queda:

$$P^*(y_1) = \alpha \sum_x P(y_1/x)\ \pi_{Y_1}(x).$$

Sustituyendo los valores numéricos de nuestro ejemplo, tenemos que,

$\pi_{Y_1}(+x) = P(+x)P(+y_2/+x) = 0.003\ \ 0.98 = 0.00294$

Cuando una variable se instancia o cambia su probabilidad, informa a su hijo a través del paso de un π-mensaje. En base a este mensaje, el hijo actualiza su probabilidad

$\pi_{Y_1}(\neg x) = P(\neg x)P(+y_2/\neg x) = 0.997 \ \ 0.017 = 0.01695$

Y, finalmente,

$P^*(+y_1) = \alpha\ [\pi_{Y1}(+x)\ P(+y_1/+x) + \pi_{Y1}(\neg x)\ P(+y_1/\neg x)] = 0.14715$
$P^*(\neg y_1) = \alpha\ [\pi_{Y1}(+x)\ P(\neg y_1/+x) + \pi_{Y1}(\neg x)\ P(\neg y_1/\neg x)] = 0.85285$

Resulta interesante comparar las expresiones utilizadas para calcular la probabilidad a priori $P(y_1)$ y la a posteriori $P^*(y_1)$. Para la primera, utilizábamos $P(x)$, ahora hemos utilizado $\pi_{Y1}(+x)$, que indica la probabilidad de x tras considerar la evidencia relativa a x *diferente* de Y_1.

De este modo observamos que la información que aporta el nodo Y_2 modifica la probabilidad de X, y, en consecuencia, también la de Y_1. El carácter simultáneamente ascendente y descendente del mecanismo de propagación es lo que nos permite utilizar la red tanto para realizar *inferencias abductivas* (cuál es el diagnóstico que mejor explica los hallazgos o síntomas) como *predictivas* (cuál es la probabilidad de obtener cierto resultado en el futuro). Un mismo nodo puede ser tanto fuente de información como objeto de predicción, dependiendo de cuáles sean los hallazgos disponibles y el objeto del diagnóstico.

En las redes bayesianas, la información circula hacia arriba y hacia abajo, permitiendo hacer inferencias tanto abductivas como predictivas

Terminada ya esta presentación intuitiva, vamos a introducir formalmente las redes bayesianas.

2.3 Definición formal de red bayesiana

Antes de definir formalmente las redes bayesianas, vamos a definir algunos conceptos de teoría de grafos y teoría de la probabilidad:

2.3.1 Definiciones previas

- **Arco.** Es un par ordenado (X, Y). Esta definición de arco corresponde a lo que en otros lugares se denomina arco dirigido. En la representación gráfica, un arco (X,Y) viene dado por una flecha desde X hasta Y.
- **Grafo dirigido.** Es un par $G = (N, A)$ donde N es un conjunto de nodos y A un conjunto de arcos definidos sobre los nodos.
- **Grafo no dirigido.** Es un par $G = (N,A)$ donde N es un conjunto de nodos y A un conjunto de arcos no orientados (es decir, pares no ordenados (X,Y)) definidos sobre los nodos.
- **Camino.** Es una secuencia ordenada de nodos $(X_{i_1}, \ldots, X_{i_r})$ tal que $\forall\ j = 1, \ldots, r\text{-}1$, ó bien el arco $X_j \rightarrow X_{j+1} \in A$ o bien el arco $X_{j+1} \rightarrow X_j \in A$.
- **Camino dirigido.** Es una secuencia ordenada de nodos $(X_{i_1}, \ldots, X_{i_r})$ tal que para todo $j = 1, \ldots, r\text{-}1$ el arco $X_j \rightarrow X_{j+1} \in A$.
- **Ciclo:** es un camino no dirigido que empieza y termina en el mismo nodo X.
- **Grafo acíclico:** es un grafo que no contiene ciclos.

- **Padre.** *X* es un *padre* de *Y* si y sólo si existe un arco $X \rightarrow Y$. Se dice también que *Y* es **hijo** de X. Al conjunto de los padres de *X* se representa como *pa(X)*, y al de los hijos de *X* por *S*(*X*).
- **Antepasado o ascendiente.** *X* es un *antepasado* o ascendiente de *Z* si y sólo si existe un camino dirigido de *X* a *Z*.
- **Conjunto ancestral** de un nodo *X* es un conjunto que contiene a *X* y a todos sus antepasados.
- **Descendiente.** *Z* es un *descendiente* de *X* si y sólo si *X* es un antepasado de *Z*. Al conjunto de los descendientes de *X* lo denotaremos por *de(X)*.
- **Variable proposicional** es una variable aleatoria que toma un conjunto exhaustivo y excluyente de valores. La denotaremos con letras mayúsculas, por ejemplo X, y a un valor cualquiera de la variable con la misma letra en minúscula, x.
- Dos variables X e Y son **independientes** si se tiene que P(X/Y) = P(X). De esta definición se tiene una caracterización de la independencia que se puede utilizar como definición alternativa: X e Y son independientes sí y sólo sí P(X,Y) = P(X) ·P(Y).
- Dos variables X e Y son **independientes** dado una tercera variable Z si se tiene que P(X/Y,Z) = P(X/Y). De esta definición se tiene una caracterización de la independencia que se puede utilizar como definición alternativa: X e Y son independientes dado Z sí y sólo sí P(X,Y/Z) = P(X/Z) ·P(Y/Z). También se dice que *Z separa condicionalmente* a *X* e *Y*.

2.3.2 Definición formal de red bayesiana

Una red bayesiana es:

- Un conjunto de variables proposicionales, V.
- Un conjunto de relaciones binarias definida sobre las variables de V, E.
- Una distribución de probabilidad conjunta sobre las variables de V.

tales que:

- *(V, E)* forman un grafo acíclico, conexo y dirigido *G*.
- *(G, P)* cumplen las *hipótesis de independencia condicional*, también llamadas de *separación direccional*, que se enuncian a continuación.

Los nodos de una red bayesiana deben ser variables proposicionales (toman un conjunto exhaustivo y excluyente de valores)

Hipótesis de independencia condicional.

Un grafo acíclico conexo y dirigido *G = (V, E)* y una distribución de probabilidad conjunta *P* definida sobre las variables del grafo se dice que cumplen las hipótesis de independencia condicional, si para toda variable *X* de *V* se tiene que el conjunto de los padres directos de *X*, que denotaremos por *pa(X) separa condicionalmente* a *X* de todo otro nodo Y de la red que no sea X, ni sus descendientes ni sus padres.

$$\forall X \in V \text{ y } \forall \mathbf{Y} \subset V - \{X \cup de(X) \cup pa(X)\} \text{ se tiene que } P(X/pa(X), Y) = P(X/pa(X)).$$

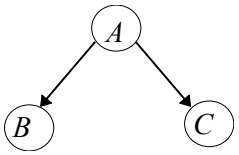
donde de(X) denota al conjunto de descendientes (directos e indirectos) de X.

Vamos a hacer un ejemplo en el que demostraremos que una red es una red bayesiana.

Las hipótesis de independencia condicional establecen que cada nodo debe ser independiente de los otros nodos de la red (salvo sus descendientes) dados sus padres

Ejemplo 2. 4. Comprobando si una red es bayesiana

Consideremos la red dada en la siguiente figura



$P(a_1) = 0.3$; $P(b_1/a_1) = 0.4$ $P(b_1/a_2) = 0.2$ $P(c_1/a_1) = 0.7$ $P(c_1/a_2) = 0.6$

En el que las variables que aparecen son binarias, junto con la siguiente distribución de probabilidad conjunta:

$P(a_1,b_1,c_1) = 0.084$ $P(a_1,b_1,c_2) = 0.036$
$P(a_1,b_2,c_1) = 0.126$ $P(a_1,b_2,c_2) = 0.054$
$P(a_2,b_1,c_1) = 0.084$ $P(a_2,b_2,c_1) = 0.056$
$P(a_2,b_1,c_2) = 0.336$ $P(a_2,b_2,c_2) = 0.224$

¿Es esto una red bayesiana?

En este caso, la condición de independencia condicional se satisface si C y B son independientes dado su padre común A, es decir, queda reducida a probar que P(B/A,C) = P(B/A). Nótese que esta red presenta la misma estructura que la del ejemplo 2, y la hipótesis que ahora tenemos que demostrar es la misma que tuvimos que hacer allí para poder seguir con los cálculos.

Para probar esto, tendríamos que ver que

$P(b_1/c_1, a_1) = P(b_1/c_2, a_1) = P(b_1/a_1)$
$P(b_1/c_1, a_2) = P(b_1/c_2, a_2) = P(b_1/a_2)$

Puesto que las restantes comprobaciones no sería necesario hacerlas ya que al ser B una variable binaria, las probabilidades relativas a b_2 son complementarias de éstas.

A modo de ejemplo vamos a comprobar una de ellas.

$$P(b_1/c_1, a_1) = \frac{P(a_1, b_1, c_1)}{P(a_1, c_1)} = \frac{P(a_1, b_1, c_1)}{P(a_1, b_1, c_1) + P(a_1, b_2, c_1)} = \frac{0.084}{0.084 + 0.126} = 0.4$$

$$P(b_1/a_1) = \frac{P(a_1, b_1)}{P(a_1)} = \frac{P(a_1, b_1, c_1) + P(a_1, b_1, c_2)}{P(a_1, b_1, c_1) + P(a_1, b_2, c_1) + P(a_1, b_1, c_2) + P(a_1, b_2, c_2)} =$$

$$= \frac{0.084 + 0.036}{0.084 + 0.126 + 0.036 + 0.054} = 0.4$$

Realizando las comprobaciones restantes veríamos que en este caso sí tenemos una red bayesiana.

En la estructura cola-con-cola la comunicación entre los hijos está abierta, y se cierra al conocer el valor del padre común

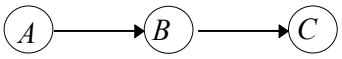
Para este tipo de estructura se dice que los enlaces convergen cola-con-cola en el nodo A. De este modo, conocer el valor del padre común (A) cierra la comunicación entre los hijos (B y C), es decir, una vez que se conoce el valor de A, conocer el valor que toma un hijo ya no aporta información sobre el valor que puede tomar el otro.

Este tipo de estructura es el que teníamos en el ejemplo 2, en el que el paludismo tenía influencia causal en el test de la gota-gruesa y en la fiebre. Así, antes de saber con seguridad si una persona padece paludismo, conocer el resultado del test de la gota gruesa cambia mi opinión acerca de si padece paludismo, y a su vez esto cambia la probabilidad de que tenga fiebre. Sin embargo, una vez que sabemos

que una persona padece paludismo, saber el resultado del test de la gota-gruesa ya no me aporta información sobre si tendrá o no fiebre.

Veamos otros ejemplos,

- Para la red;



Habría que comprobar que *B* separa condicionalmente a *A* de *C*.

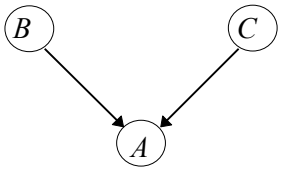
En este tipo de estructura se dice que los enlaces convergen *cola-con-cabeza* en B. Al conocer el valor de B, se *cierra la comunicación* entre el padre de B y el hijo de B.

En la estructura cola-con-cabeza, la comunicación entre los nodos raíz y hoja está abierta, y se cierra al conocer el valor nodo intermedio

Si por ejemplo tenemos que la metástasis (A) causa tumor cerebral (B) y jaquecas (C), a priori saber si tiene metástasis o no cambia mi opinión acerca de su probabilidad de desarrollar un tumor cerebral, lo que a su vez afecta a la probabilidad de padecer jaquecas. Sin embargo, una vez que se que la persona tiene un tumor cerebral, saber si ese tumor ha sido producido por una metástasis ya no afecta a la probabilidad de que la persona tenga jaquecas.

Otro ejemplo: supongamos que estamos haciendo un estudio sobre la bolsa. El estado de la bolsa ayer (A) influye en el estado de la bolsa hoy (B), que a su vez influye en el estado de la bolsa mañana (C). Por tanto, a priori el estado de la bolsa ayer tiene influencia causal en el estado de la bolsa mañana, pero esa influencia es sólo *a través* del estado de la bolsa hoy. En cuanto conocemos el valor de la bolsa hoy, la relación entre A y C se interrumpe, es decir, una variable ya no nos aporta información sobre otra. Los sistemas que presentan este comportamiento se dice que tienen la propiedad de Markov, que dice que, dado el presente, el futuro es independiente del pasado.

- Para la red;



Habría que comprobar que *B* es independiente de *C*.

En este tipo de estructura se dice que los enlaces convergen *cabeza-con-cabeza* en A. Conocer el valor del hijo común (A) abre la comunicación entre los padres (B y C), ya que conocer el valor de un padre cambia las probabilidades del otro.

En la estructura cabeza-con-cabeza la comunicación está cerrada, y se abre al conocer el valor del hijo común

De este modo, pensemos en el caso en que hay dos enfermedades que provocan el mismo síntoma (por ejemplo, tanto la gripe como una infección

de orina provocan fiebre). A priori, las dos enfermedades son independientes (a menos que estén relacionadas, en cuyo caso aparecería un enlace de una a otra). Sin embargo, una vez que sabemos que el paciente padece una las enfermedades, queda explicado el síntoma, y por tanto la probabilidad de que padezca la otra enfermedad disminuye. Del mismo modo, si sabemos con certeza que no padece una de las enfermedades, la probabilidad de que padezca la otra aumenta. Este efecto se conoce con el nombre de *explaining-away*, que podríamos traducir como *descartar/potenciar causas*.

El efecto *explaining-away* permite que conforme una de las posibles explicaciones cobra fuerza, las otras se vayan debilitando.

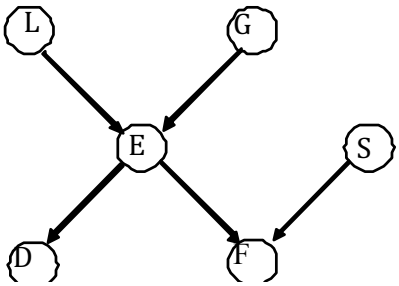
Como se puede ver, las comprobaciones de independencia necesarias dependen de la estructura de la red.

Ejemplo 2. 5.

Consideremos las siguientes variables binarias:

L = situación laboral
G = ganancias por inversiones
E = situación económica
S = salud
D = donaciones
F = felicidad.

Entre ellas existen las relaciones causales que se reflejan en esta red:



Especifica qué independencias condicionales deben cumplirse para que esta sea una red bayesiana.

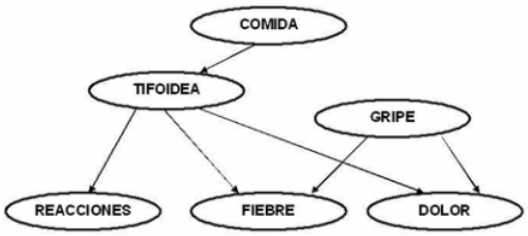
Aplicando las hipótesis de independencia condicional, tendríamos que:

- D independiente de F, S, G, L dado E
- F independiente de D, L, G dado {E, S}
- E independiente de S dados {L, G}
- S independiente de L, G, E, D
- L independiente de G y S
- G independiente de L y S

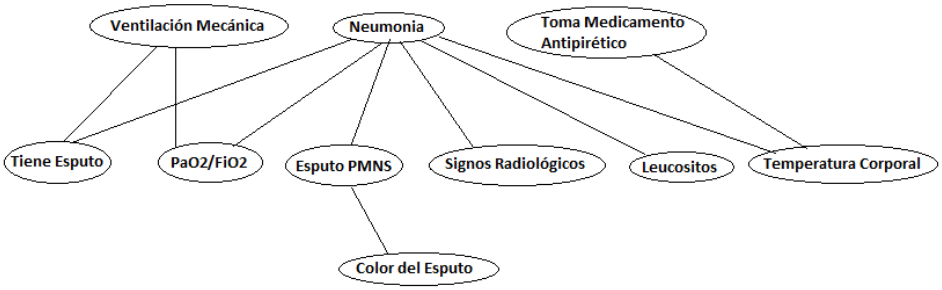
En la definición de red bayesiana, hemos partido de una distribución de probabilidad conjunta para las variables. Aparentemente, suponiendo que tuviésemos una red con N nodos y con variables binarias, haría falta conocer 2^N-1 valores. Sin embargo, las *condiciones de independencia condicional* permiten que no sea necesario conocer todos estos valores, puesto que, como veremos en el siguiente Teorema, la distribución de probabilidad conjunta se puede expresar

como producto de las distribuciones condicionadas de cada nodo, dados sus padres.

Ejercicio 2.2. Dadas las siguientes estructuras, dí que condiciones de independencia condicional deberían cumplir las variables para que las redes fuesen bayesianas.



(tomada de http://redbay.wikidot.com/start)



(tomada de http://ferminpitol.blogspot.com.es/2014/04/redes-bayesianas.html)

Teorema (Factorización de la probabilidad)

Dada una red bayesiana, su distribución de probabilidad puede expresarse como:

$$P(X_1, ..., X_n) = \prod_i P(X_i/pa(X_i)).$$

Demostración:
Es fácil construir una ordenación de las variables en la que los padres de cada nodo aparezcan siempre después de él. Supongamos por tanto que la ordenación $\{X_1, ..., X_n\}$ cumple dicha propiedad. Por tanto:

$$P(X_1, ..., X_n) = \prod_i P(X_i/X_{i+1},...,X_n)$$

Pero por la forma de escoger la ordenación, el conjunto $\{X_{i+1}, ..., X_n\}$ incluye a todos los padres de X, y, en consecuencia, la separación direccional nos dice que

$$P(X_i/X_{i+1},...,X_n) = P(X_i/pa(X_i))$$

Con lo que concluimos la demostración.

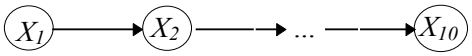
La importancia de este teorema es que nos permite describir una red bayesiana a partir de la probabilidad condicionada de cada nodo (o la probabilidad a priori en el caso de nodos sin padres) en lugar de dar la probabilidad conjunta, que,

- requiere un número de parámetros exponencial en el número de nodos.
- plantea el problema de verificar la separación direccional.

Si se cumplen las condiciones de independencia condicional, a partir de las probabilidades condicionadas es posible calcular la distribución coniunta

Sin embargo, el número de parámetros requerido para dar las probabilidades condicionadas es mucho menor (proporcional al número de nodos), nos permite reconstruir la distribución conjunta aplicando el teorema, y además, a la hora de pedirle estos valores al experto, son valores con pleno significado, como vimos en el ejemplo 1.

Por ejemplo, para la red bayesiana dada por:



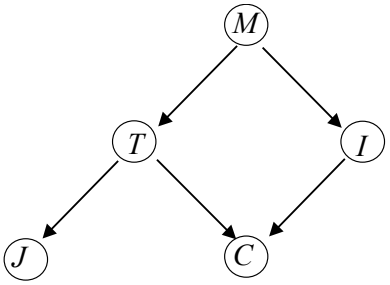
Suponiendo que todas las variables fuesen binarias, para dar la distribución conjunta habría que dar 2^{10}-1 valores, sin embargo, si construimos la distribución conjunta a partir de los 19 valores necesarios para dar las condicionadas, tendremos además asegurado que se satisfacen las hipótesis de independencia condicional.

Ejemplo 2. 6. Independencias, distribuciones condicionadas y distribución conjunta

En un sistema de diagnóstico médico, supongamos que tenemos la siguiente información:

- Metástasis (M) causa tumor cerebral (T) e incremento en los niveles de calcio (I).
- Tumor cerebral causa coma (C).
- Incremento en nivel de calcio causa coma.
- Tumor cerebral causa fuertes jaquecas (J)

Representamos dicha información en una siguiente red bayesiana:



¿Qué independencias implica la red?

- I es independiente de T ,J dado M. Así por ejemplo dado metástasis, el incremento de calcio no depende de si hay o no tumor cerebral.
- T independiente de I dado M.
- C independiente de M, J dados {T, I}
- J independiente M, I ,C dado T.

Según el teorema de factorización, los únicos datos que debemos pedir al experto son:

$P(m_1) = 0.2$

$P(i_1/m_1) = 0.8$	$P(i_1/m_2) = 0.2$
$P(t_1/m_1) = 0.2$	$P(t_1/m_2) = 0.05$
$P(c_1/i_1,t_1) = 0.8$	$P(c_1/i_1,t_2) = 0.9$
$P(c_1/i_2,t_1) = 0.7$	$P(c_1/i_2,t_2) = 0.05$
$P(j_1/t_1) = 0.8$	$P(j_1/t_2) = 0.6$

Y a partir de estas probabilidades podríamos construir la distribución conjunta. Por ejemplo:

$$P(m_1,i_2,t_1,j_2,e_2) = P(m_1)\cdot P(i_2/m_1)\cdot P(t_1/m_1)\cdot P(c_2/i_2,t_1)\cdot P(j_2/t_1) =$$
$$= 0.2\cdot 0.2\cdot 0.2\cdot 0.3\cdot 0.2 = 0{,}00048.$$

Con la distribución conjunta P así construida tenemos una red causal que representa a la perfección la noción humana de causalidad. Una vez obtenida la distribución conjunta, podemos a partir de ella obtener la distribución de probabilidad que queramos. Por ejemplo, si queremos saber P(M/C,J):

$$P(M/C=c_1, J=j_2) = \frac{P(M,c_1,j_2)}{P(c_1,j_2)} = \frac{\sum_{I,T} P(M,I,T,j_2,c_1)}{\sum_{M,I,T} P(M,I,T,j_2,c_1)}$$

Como vemos, una vez conocida la distribución conjunta es posible conocer la distribución de cualquier conjunto de variables, condicionadas o no condicionadas.

Sin embargo, este método de cálculo es computacionalmente muy costoso. Desde la aparición de las redes bayesianas, se han desarrollado varios algoritmos más eficientes de propagación de probabilidades. Aunque el procedimiento de cálculo exacto de las probabilidades es NP-duro (Cooper, 1987), hoy en día existen algoritmos exactos que se ejecutan en tiempo real y, para redes con estructuras muy complejas (elevado número de padres para cada nodo), algoritmos aproximados que proporcionan una buena estimación de las probabilidades en tiempos razonables. Además, se dispone de bibliotecas que permiten integrar estos algoritmos de forma sencilla en nuestras aplicaciones. En la sección 4 veremos en detalle el algoritmo exacto para redes con estructura de árbol, que es el caso más simple.

Ejercicio 2.3. Aplica el teorema de factorización de la probabilidad a las redes del Ejercicio 2.2.

2.4 Modelado con redes bayesianas

Una vez hemos definido las redes bayesianas, vamos a aprender a modelar problemas de la vida real utilizando este enfoque. En esta sección utilizaremos el siguiente ejemplo.

Ejemplo 2. 7. El ejemplo del estornudo

Una tarde, Juan va a visitar a su amigo Pablo. De repente, comienza a estornudar. Juan piensa que se ha resfriado, hasta que observa que los muebles de la casa están arañados. Entonces, especula con la posibilidad de que su amigo tenga un gato y sus estornudos se deban a una crisis de la alergia a los gatos que tiene diagnosticada.

Principalmente, los tipos de problemas que se suelen modelar con redes bayesianas son problemas de diagnóstico o problemas de predicción. El ejemplo de la alergia es un problema de diagnóstico, puesto que Juan intenta determinar la causa de sus estornudos.

2.4.1 Identificación de las variables

Es importante identificar las variables relevantes en el problema

En primer lugar, es importante estudiar el dominio para tener el grado máximo de conocimiento y comprensión sobre el problema que vamos a modelar. En la

mayoría de los casos reales, esto nos obligará a contar con expertos en el área, que deberán estar suficientemente interesados y motivados para que la colaboración tenga buenos frutos.

Una vez conocemos suficientemente el problema, el siguiente paso consiste en identificar las variables que son relevantes. Es importante centrarse sólo en aquellas variables que son de interés en el problema actual. Para ello, ayuda realizarse preguntas del tipo:

- ¿Cuál es la situación/problema que se plantea?
- ¿Qué posibles causas pueden explicar esta situación?
- ¿Qué otros factores pueden hacer que los problemas o causas ocurran, o impedir que ocurran?
- ¿De qué evidencia se dispone para soportar dichas causas, problemas o factores?

Veamos cómo aplicar esto a nuestro ejemplo. En nuestro caso, el problema parece ser que Juan está *estornudando*. Las causas posibles son que se ha *resfriado* o que tiene *rinitis*. La rinitis puede estar causada porque sus amigos tienen un *gato* y Juan es *alérgico* a los gatos. La evidencia que sugiere que sus amigos tienen un gato es que algunos muebles tienen *arañazos*. La información relevante de la situación está contenida en las seis palabras subrayadas. Ejemplo de información irrelevante en este caso es que Juan y Pablo son amigos o que Juan está visitando a Pablo.

En el caso general del modelado de problemas de diagnóstico, hay ciertos tipos de variables susceptibles de ser agrupados en clases. Si se aborda el problema teniendo estas clases en mente, el proceso de modelado resulta más sencillo. Hablaremos por tanto de estas clases.

Variables objetivo

Estas variables se usan para modelar los objetos de interés, es decir, aquellos objetos sobre lo que nos gustaría razonar. Las variables objetivo suelen utilizarse para modelar fenómenos *latentes*, es decir, fenómenos que no son directamente observables. En el ejemplo del estornudo, Juan piensa en dos alternativas: o bien se ha Resfriado o bien tiene Alergia. Ambos son ejemplos de variables objetivo ya que Juan está interesado en saber más sobre ellas (el estado en el que están o los valores que tienen). En diagnóstico médico, las enfermedades serían modeladas como variables objetivo.

Variables de observación

Las variables de observación se usan para modelar las formas indirectas que tenemos de medir las variables objetivo. También se denominan variables *de evidencia*. En el ejemplo del estornudo, Juan piensa que está bien hasta que empieza a estornudar. Sólo después de observarse a sí mismo estornudando se pregunta si está Resfriado. *Estornudar* sería una variable de observación. Otra podría ser *Arañazos*, porque Juan hace esa observación y la usa para razonar sobre la posibilidad de que exista un gato en casa (por el momento, no directamente observable). En el diagnóstico médico, los síntomas que muestra el paciente y los resultados de sus pruebas serían modeladas como observaciones. Algunas observaciones pueden ser obligatorias. Por ejemplo, en el diagnóstico médico un tipo específico de escáner puede ser requisito indispensable con el objetivo de detectar un posible cáncer.

Factores

Estas variables se usan para modelar los fenómenos que afectan a las variables objetivo. También se denominan *variables de contexto*. En el ejemplo del estornudo, la estación del año podría ser un factor que afecta al resfriado, pues es más probable que una persona se resfríe en invierno que en verano.

Los factores pueden dividirse en cuatro categorías, con respecto al tipo de influencia en las variables afectadas.

- *Promotores*. Si el factor promotor ocurre, la variable afectada será más probable (correlación positiva). Por ejemplo, fumar puede incrementar las probabilidades de tener un cáncer de pulmón.
- *Inhibidores*. Si el factor promotor ocurre, la variable afectada es menos probable (correlación negativa). Por ejemplo, practicar deporte puede disminuir las probabilidades de caer enfermo.
- *Requeridos*. Es indispensable que estos factores entren en acción para sea posible que ocurran las variables afectadas. Por ejemplo, para que una población específica de bacterias crezca se requiere que la temperatura esté por encima de un determinado nivel.
- Preventivos. Si el factor ocurre, la variable afectada no puede ocurrir. Por ejemplo, recibir la vacuna de la viruela a edades tempranas previene que se sufra esta enfermedad.
- Auxiliares. Son variables que se usan por conveniencia. Por ejemplo, para simplificar el proceso de modelado y especificación de parámetros.

Esta lista de tipos de variables no pretende ser exhaustiva, sino simplemente ilustrar diversos tipos existentes. La resumimos en la siguiente tabla.

Tipo de variable	Breve descripción
Objetivo	Modelan objetos de interés. No observables directamente.
Observación	Modelan la forma de medir variables objetivo. Pueden ser observadas directamente
Factor	Modelan fenómenos que afectan a otras variables del modelo.
Promotor Inhibidor Requerido Preventivo	La variable afectada es más probable cuando están presentes. La variable afectada es menos probable cuando están presentes. Si no entra en acción, no ocurre la variable afectada. Si entra en acción, no ocurre la variable afectada.
Auxiliares	Usadas por conveniencia (para simplificar el modelo)

Tabla A. Tipos de variables en modelado Bayesiano

2.4.2 Estados y valores

Hasta ahora hemos definido varios tipos de variables con respecto al papel que juegan en el modelado de un problema. Pero también podemos dividir las según su escala de medición:

Variables cualitativas:

Son las variables que expresan distintas cualidades, características o modalidad. Cada modalidad que se presenta se denomina atributo o categoría y la medición consiste en una clasificación de dichos atributos. Las variables cualitativas pueden ser *dicotómicas* cuando sólo pueden tomar dos valores posibles como sí y no, hombre y mujer, o *politómicas*, cuando pueden adquirir tres o más valores.

Variables cuantitativas:

Son las variables que se expresan mediante cantidades numéricas. Las variables cuantitativas además pueden ser *discretas*, que presentan separaciones o interrupciones en la escala de valores que puede tomar (por ejemplo, el número de hijos), o *continuas*, que pueden adquirir cualquier valor dentro de un rango especificado (por ejemplo, la edad).

A menudo conviene representar un fenómeno continuo en la naturaleza usando variables discretas. Para ello, las medidas continuas tienen que ser discretizadas. Esto puede hacerse proyectando la escala de valores continua en un conjunto finito de intervalos. Los valores que caigan en el mismo rango se considerarán como un mismo estado. Un ejemplo de discretización es modelar la variable temperatura con tres estados: bajo, medio, y alto.

A la hora de decidir si una entidad del mundo real debe ser una variable o un estado de una variable, debemos recordar que las variables de una red deben tomar un conjunto exhaustivo y excluyente de valores

La definición de variable proposicional tiene su importancia a la hora de modelar un problema con una red bayesiana, ya que deberemos tener en cuenta *que los nodos de la red son variables proposicionales y por tanto deben tomar un conjunto exhaustivo y excluyente de valores*.

De este modo, si por ejemplo estamos construyendo un sistema de diagnóstico médico en que las enfermedades posibles son *gripe, faringitis y alergia*, cada una de estas enfermedades será representada por una variable dicotómica diferente (que tomará valores *si/no*), ya que nada impide que un paciente padezca dos o más enfermedades a la vez. Es decir, al no conformar las enfermedades un conjunto exhaustivo y excluyente de variables, cada una de ellas debe ser modelada como una variable dicotómica y no como valores de una única variable.

Sin embargo, si estamos construyendo un sistema de clasificación de animales en el que hemos representado todas las posibilidades (mamífero, ave, reptil, pez, etc), debemos introducir una única variable, cuyos estados serán las diferentes clases consideradas (ya que un animal no puede ser a la vez un mamífero y un reptil). A veces, si no se está completamente seguro de que el conjunto considerado es exhaustivo, se puede añadir un estado indeterminado "otro" de modo que se cumplan todas las condiciones para que cada nodo contenga una variable proposicional.

En el ejemplo del estornudo consideraremos una variable por cada dato identificado como relevante (dado que las variables no son incompatibles entre sí), y que todas las variables son dicotómicas, de forma que cada una de las variables tomará el valor "*presente*" o "*ausente*". En problemas y aplicaciones más sofisticados se podría usar variables discretas (por ejemplo, la variable estornudo podría tomar los valores pocos, algunos, muchos) o variables continuas (por ejemplo, si tuviéramos una variable temperatura, podría tomar los valores que van desde 35,8 a 42,5). Pero evidentemente, cuantos más valores tomen las variables, más complicado será el modelo, así que a la hora de decidir los estados deberíamos tener presente qué grado de granularidad es realmente necesario en nuestra aplicación.

No debemos introducir en el modelo mayor nivel de detalle del realmente necesario para nuestra aplicación

2.4.3 Estructura

Después de definir las variables, el siguiente paso en la construcción de un modelo es definir su estructura. Esto lo hacemos conectando variables con arcos (también llamados enlaces). Como hemos visto, en las redes bayesianas los arcos son dirigidos. Cambiar la dirección de un arco cambia su significado.

La ausencia de un arco entre dos variables indica que no existen relaciones de dependencia directa entre ellas, sino a lo sumo a través de otras variables. En el ejemplo del estornudo las variables Estornudo y Gato no son directamente dependientes, por lo que no debería haber un arco que las una[2].

A la hora de definir los arcos, debemos reflejar en ellos las relaciones de *influencia causal* entre las variables.

La presencia de un arco indica una *relación de influencia causal* entre dos variables. Por ejemplo, la *presencia* o *ausencia* de una enfermedad tiene influencia que el resultado de las pruebas sea *positivo* o *negativo* (*Figura 2*). En el ejemplo del estornudo, `Gato` tiene influencia causal en `Arañazos`, y por tanto debe existir un arco entre estas dos variables.

Una alternativa a la dirección causal es la dirección de diagnóstico (Figura 2.1). La dirección de diagnóstico surge de aplicar reglas de diagnóstico del tipo "SI el paciente tiene tos, entonces el paciente tiene gripe". En este caso, la variable de observación precede a la variable objetivo. Utilizar este tipo de relaciones para modelado en redes bayesianas no es en sí un error, pero la dificultad añadida que supone hace que frecuentemente se usen de un modo incorrecto.

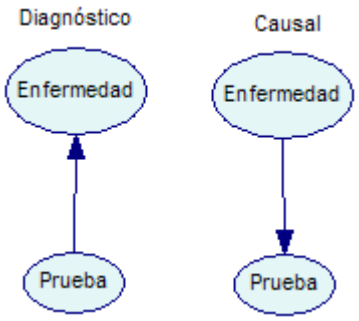


Figura 3.1. Posibles direcciones de los arcos.

El marco de trabajo de las redes bayesianas no impone que los arcos se construyan siempre en la dirección causal. Pero según Pearl, "*los patrones de independencia plasmados en un grafo acíclico dirigido son típicos de organizaciones causales*" (Pearl, 1998). También, Druzdel y Simon explican que "*Se pueden usar muchos modelos equivalentes para representar el mismo sistema … pero se prefieren los modelos que utilicen las direcciones causales, ya que minimizan el número de arcos en el grafo, aumentando la claridad de los modelos y ofreciendo ventajas computacionales*" (Druzdel y Simon, 1993).

Utilizar relaciones causales conduce a modelos más sencillos de especificar y entender

Para ilustrar el motivo por el que se prefiere la dirección causal, vamos a construir una RB causal para el ejemplo de los estornudos. Una vez definida, calcularemos la red equivalente con los enlaces invertidos para mostrar que el modelo obtenido es bastante más difícil de interpretar.

Vamos ahora a desarrollar el modelo del ejemplo del estornudo. Para ello vamos a asignar a cada información que hemos considerado relevante en nuestro modelo una variable o nodo, y los vamos a ir incluyendo en una red que construiremos de modo incremental:

[2] A veces será difícil juzgar cual es la causa y cual el efecto. Por ejemplo, pensemos en el caso de una persona que no come y padece anorexia nerviosa. Podemos pensar que anorexia es la causa de que no coma, pero también que la anorexia ha venido provocada por no comer.

En este problema, la situación a la que queremos ofrecer una explicación es a los estornudos de Juan. Por ello, introducimos en la red un nodo que llamaremos `Estornudo:`

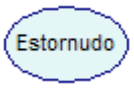


Figura 3.2. Primer nodo para la RB del estornudo.

Las causas posibles del estornudo son `Resfriado` o `Rinitis`. La añadimos al modelo y ahora la red es:

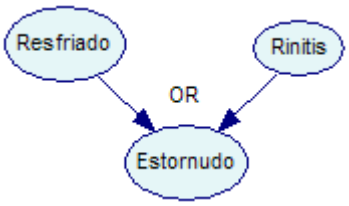


Figura 3.3. Añadiendo causas a la RB del estornudo.

Entonces Juan observa `Arañazos` en los muebles:

Figura 3.4. Añadiendo fuentes de evidencia a la RB del estornudo.

Juan empieza a pensar que sus amigos pueden tener un `Gato`, lo que puede causar los `Arañazos`:

Figura 3.5. Añadiendo explicaciones a la RB del estornudo.

Lo que causa que Juan piense que él está estornudando podría ser debido a la `Rinitis`, causada por la `Alergia` y la presencia del `Gato`:

Figura 3.6. La RB final del problema del estornudo

De esta forma, hemos construido de modo incremental el modelo de RB para el ejemplo del estornudo.

2.4.4 Parámetros

El último paso en el proceso de modelado es *especificar parámetros*. Como se explicó antes, basta con proporcionar las probabilidades a priori de los nodos raíz y las probabilidades condicionales del resto de los nodos

Existen varias alternativas para obtener los parámetros necesarios de una red:

- Especificación directa de los parámetros, normalmente contando con la ayuda de expertos. Este procedimiento es ciertamente costoso.
- Aprendizaje a partir de bases de datos, que obviamente depende de la existencia de dicha base de datos. De ser así, se tienen dos opciones; a) aprendizaje de los parámetros, si se dispone de la estructura; b) aprendizaje estructural, en el que es posible aprender tanto la estructura como los parámetros.; y
- Combinar especificación y aprendizaje. Por ejemplo, contar con expertos que nos ayuden a especificar la estructura, aprender los parámetros y disponer de expertos de nuevo para supervisar el modelo obtenido. Normalmente, esta última alternativa ofrece lo mejor de cada caso.

Usaremos el ejemplo del resfriado para ilustrar el proceso de especificación de parámetros. La notación que usaremos será la siguiente: el nombre de la variable indicará su presencia, y el nombre precedido de un símbolo ~ indicará su ausencia.

Para los nodos raíz, necesitamos proporcionar las probabilidades anteriores, que son las probabilidades a priori o en ausencia de información. Por ejemplo, necesitamos proporcionar la probabilidad del nodo `Gato`. Para ello, podemos pensar en la proporción de familias que tienen gato. Si disponemos información disponible sobre esto (por ejemplo, estadísticas) podemos usarlas, en otro caso, tendremos que hacer uso de una estimación razonable. Vamos a considerar que un 20% de las familias tienen gato. Entonces la probabilidad anterior de que una familia tenga un gato es 0.2.

Cuando hay dos causas que por separado e independientemente entre ellas pueden causar el mismo síntoma, tenemos una relación tipo OR

Para aquellos nodos que tienen padres, necesitamos proporcionar probabilidades condicionadas. Por ejemplo, necesitamos proporcionar la probabilidad de *estornudo dado resfriado y rinitis*. Para ello, necesitamos pensar en la clase de relación entre estas tres variables, que en este caso es una relación tipo OR, es decir, tanto el resfriado como la rinitis pueden causar el estornudo. Si

tenemos ambas se incrementan las probabilidades de estornudo. Un ejemplo de probabilidades que funcionarían para este caso es:

P(estornudo/resfriado,rinitis) = 0.99
P(estornudo/resfriado,~rinitis) = 0.85
P(estornudo/~resfriado,rinitis) = 0.9
P(estornudo/~restriado,~rinitis) = 0.01

Como se puede observar, las primeras tres probabilidades están cercanas a 1, mientras que la última está cercana a 0. Este modelo simple permite expresar varias cosas diferentes: la primera es que cuando hay más de una causa presente es más probable que el efecto ocurra (0.99 comparado a 0.85 y 0.9); la segunda es que la relación entre algunas causas es más fuerte que entre otras (por ejemplo, en este caso las probabilidades indican que la relación entre `rinitis` y `estornudo` es de alguna manera más fuerte que la relación entre `resfriado` y `estornudo`); la tercera es que, incluso cuando las dos causas están presentes, algo extraño puede ocurrir y la persona puede no estar estornudando (esta es la razón por la que la probabilidad es 0.99 y no 1). Los números 1 y 0 también se podrían usar para definir las probabilidades, indicando de este modo que no queremos modelar la incertidumbre presente en las relaciones.

Cuando hay dos causas que necesitan actuar conjuntamente para provocar cierto síntoma, tenemos una relación tipo AND

El uso de modelos canónicos puede simplificar la especificación de parámetros. El lector interesado puede encontrar más información al respecto en (Diez, 2001).

Si procedemos de la misma forma con el resto de nodos podemos terminar el proceso de modelado. La figura siguiente muestra la RB completa, con nodos, enlaces y probabilidades. Nótese que, en el caso de la `rinitis`, las probabilidades condicionales han sido definidas para modelar una relación tipo AND (la `rinitis` ocurre sólo cuando el gato y la alergia están presentes a la vez, excepciones aparte). En lo que sigue, para indicar las relaciones AND y OR, incluiremos una etiqueta en el modelo gráfico para dejar claro el tipo de relación existente entre los nodos.

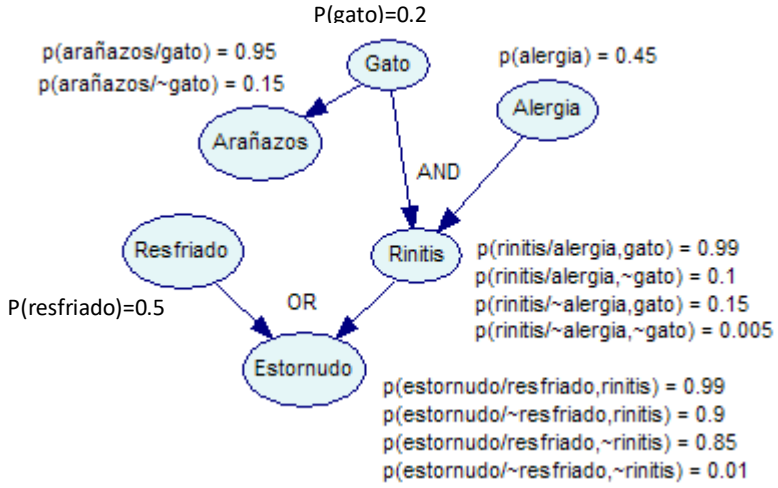


Figura 3.7. RB y parámetros para el ejemplo del resfriado.

2.4.5 Estructuras equivalentes

Como se explicó antes, desarrollar el modelo según las relaciones de causalidad simplifica el trabajo tanto en la deducción de la estructura como de los parámetros.

Cualquier enlace de la RB puede siempre ser invertido para desarrollar un modelo equivalente. Sin embargo, en el proceso de inversión se añaden nuevos enlaces. La figura siguiente muestra la red bayesiana para el ejemplo del estornudo, en la que se han invertido todos los enlaces, y los nuevos enlaces que ha sido necesario crear para que el comportamiento de la red sea equivalente. Asimismo, muestra a modo de ejemplo las probabilidades que se obtienen para uno de los nodos (el nodo gato). Como se puede observar, el modelo es mucho más complicado de interpretar y las probabilidades condicionadas más difíciles de estimar.

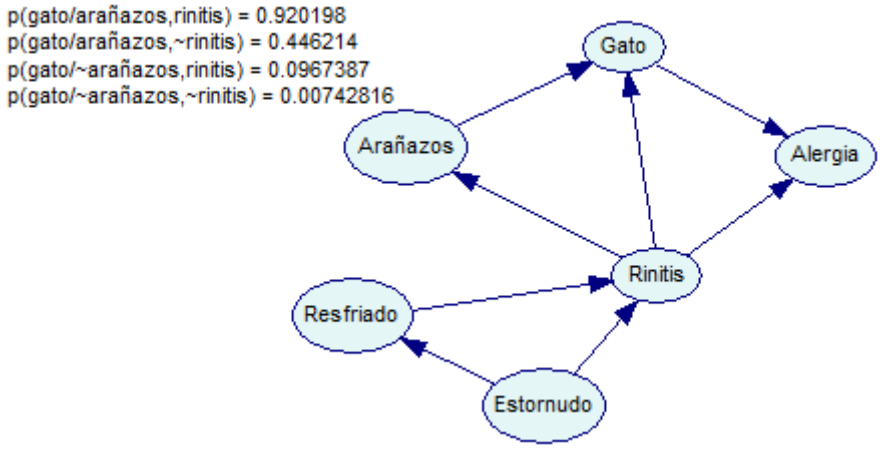


Figura 3.8. RB equivalente para el ejemplo del estornudo, con los enlaces invertidos

El conjunto de reglas de diagnóstico que explica este modelo es el siguiente:

- SI rinitis, ENTONCES arañazos
- SI rinitis, ENTONCES gato
- SI rinitis, ENTONCES alergia
- SI arañazos, ENTONCES gato
- SI refriado, ENTONCES rinitis
- SI estornudos, ENTONCES resfriado
- SI estornudos, ENTONCES rinitis
- SI gato, ENTONCES alergia

Como se puede ver, tanto el conjunto de reglas como el de probabilidades son de interpretación más compleja que el modelo causal.

Ejercicio 2.4. (un problema de diagnóstico). Considera la siguiente situación: Los padres de Luisito, que acaba de cumplir un año, deciden llevarlo al pediatra porque vomita con cierta frecuencia. Con el pediatra sostienen la siguiente conversación:

Pediatra -. Denme toda la información que consideren que puede ser relevante.
Madre-. El otro día Luisito estaba resfriado. Vomitó el biberón de la noche, creo que por culpa de los mocos, ya que había muchos en el vómito. Otras veces parece que vomita por una pequeña indigestión.
Padre-. Además, creo que debe saber que mi hermano es celíaco (Aclaración: la celiaquía es una intolerancia al gluten, que poco a poco hace que se destruya el vello intestinal. Los vómitos son uno de sus síntomas más relevantes. Se cree que tiene cierta componente hereditaria).
Pediatra-. ¿Y la dieta de Luisito incluye gluten?
Ambos-. Sí, desde hace unos meses.

Plantea este problema de diagnóstico mediante una red bayesiana

Ejercicio 2.5. (un problema de clasificación). En el planeta Zyx se pueden encontrar varias clases de animales, llamemos a estas clases Wurros, Hobexas y Wackas. Todos tienen un tamaño muy pequeño, y sus pieles son o bien escamosas o bien están cubiertas de suave pelo. Además, una observación atenta ha permitido deducir lo siguiente:

- Todos los Wurros tienen 5 ó 6 patas. Su color es rojizo, y tienen la piel peluda y suave.
- El número de patas de las Hobexas es un entero que varía uniformemente entre 4 y 6, ambos inclusive. Su piel es escamosa.
- En cuanto a las Wackas, tienen 4 ó 5 patas, y ofrecen a la vista una tonalidad casi siempre azul, pero a veces (20% de los casos) rojiza.
- Los animales que tienen un número impar de patas cojean siempre. Los animales que tienen un número par de patas cojean sólo cuando tienen alguna anomalía (malformación congénita, heridas, etc.), lo cual ocurre en el 10% de los casos para los animales de 4 patas, y en el 20% para los de seis.

Plantea el problema de la clasificación de animales de Zyx mediante una red bayesiana

Ejercicio 2.6. El problema de Monty Hall. A un concursante del concurso televisivo *Let´s Make a Deal* se le pide que elija una puerta entre tres (todas cerradas), y su premio consiste en llevarse lo que se encuentra detrás de la puerta elegida. Se sabe que una de ellas oculta un coche, y las otras dos tienen una cabra. Una vez que el concursante ha elegido una puerta y le comunica al público y al presentador su elección, el presentador (que conoce en que puerta está el premio) abre una de las otras puertas y muestra una cabra. En este momento se le da la opción al concursante de quedarse con la puerta que eligió inicialmente o bien cambiar de puerta. ¿Debe el concursante mantener su elección original o escoger la otra puerta? Modela la situación con una red bayesiana.

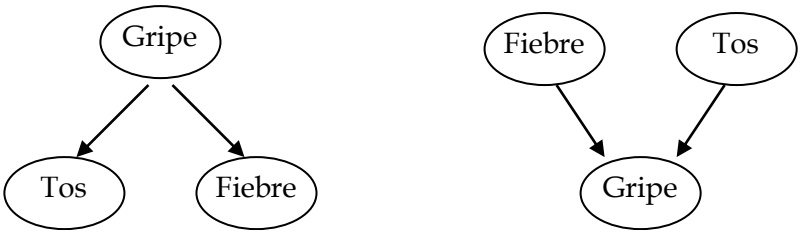
Práctica 1. Introducción a GeNIe. El objetivo de esta práctica es familiarizarse con la herramienta e implementar los modelos desarrollados para los ejercicios 2.4 a 2.7

2.4.6 Algunos trucos para el modelado

Hay una serie de trucos que en la práctica pueden resultar de utilidad. Mencionaremos algunos de ellos.

Verificar las independencias que supone el modelo

Una vez definido el modelo, para asegurarnos de que es correcto debemos comprobar si las relaciones entre las variables reflejan adecuadamente las dependencias e independencias existentes. Por ejemplo, para definir la relación entre las variables gripe, tos y fiebre tenemos dos posibles modelos:



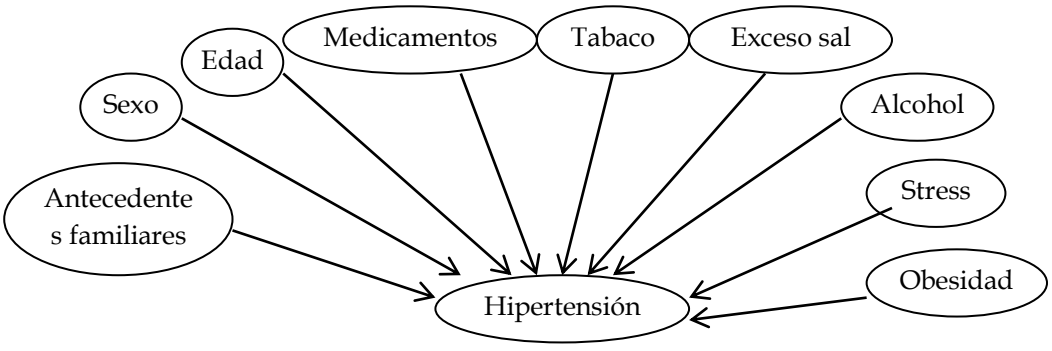
- En el modelo A, Fiebre y Tos son dependientes a priori, pero independientes dado gripe
- En el modelo B, Fiebre y Tos son independientes a priori, pero dependientes dado gripe (explaining-away).

Vemos que en este caso el modelo que mejor refleja las independencias/dependencias que ocurren en la vida real es el A.

Introducir nodos intermedios para reducir la complejidad

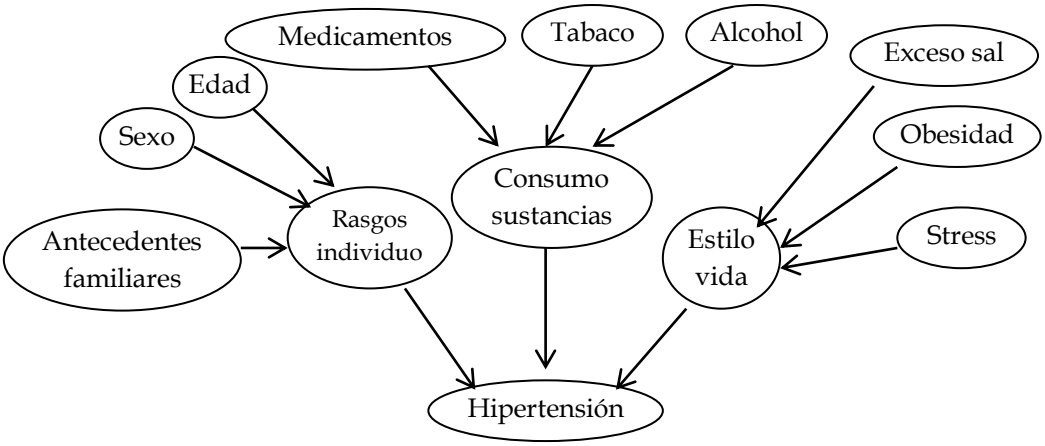
Sabemos ya que los parámetros necesarios a la hora de definir una red son las probabilidades condicionadas de cada nodo dados sus padres. Por ello, el número de probabilidades necesarias para cada nodo es exponencial en el número de padres. De esta forma, cuando un nodo tiene muchos padres una forma de reducir la complejidad del modelo es introducir nodos intermedios que agrupen a varios padres. De este modo, no sólo reduciremos el número de parámetros necesarios, sino también el tiempo de ejecución de los algoritmos de propagación de probabilidades.

Veamos un ejemplo. Consideremos el siguiente ejemplo: Supongamos que estamos construyendo un modelo para la hipertensión arterial, de la que un estudio ha identificado las siguientes causas: sexo, edad, antecedentes familiares, consumo de alcohol, tabaco, o ciertos medicamentos) excesiva ingesta de sal, obesidad, estrés. La red bayesiana para modelar esta situación sería:



Como vemos, en este modelo la hipertensión arterial tiene nueve padres, de modo que, en el caso de que los nodos sean binarios, para dar la probabilidad condicionada de la hipertensión necesitamos 2^9 valores.

Para simplificar el modelo, podemos intentar agrupar las causas del siguiente modo: edad, sexo y antecedentes familiares se pueden considerar rasgos propios del individuo; alcohol, tabaco y consumo de medicamentos se pueden agrupar en consumo de productos adictivos, mientras que exceso de sal en la dieta, estrés y obesidad pueden considerarse hábitos de vida. De este modo, si consideramos razonable este agrupamiento, el modelo sería:



En la que ahora para el nodo `hipertensión` necesitamos 23 parámetros, igual que para los nodos `rasgos, individuo, productos adictivos y estilo de vida.` Es decir, que mediante esta técnica, hemos reducido el número de parámetros necesarios de 512 a 32.

Uso de modelos canónicos

Hemos visto que, en el caso general, el número de parámetros requerido para especificar la probabilidad condicional de un nodo crece exponencialmente con el número de padres, lo que plantea problemas tanto de obtención y almacenamiento de datos como de tiempos de computación de los algoritmos. (por ejemplo, para un nodo binario con 10 padres binarios, la tabla de probabilidad condicional tiene 2^{10} = 2.048 parámetros).

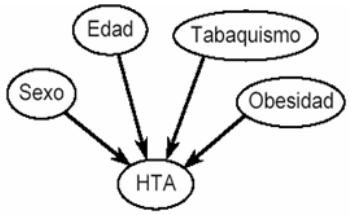
Por ello, es conveniente buscar modelos simplificados de interacción entre variables. Pearl los llama modelos *canónicos* por su aplicabilidad a diferentes dominios. Los más utilizados son el modelo NOISY-OR y NOISY-ADD.

En el caso de la puerta OR, es necesario realizar las siguientes hipótesis:

1. Cada una de las causas, por sí misma, puede producir el efecto, y basta que una de las causas actúe para que el efecto esté presente;
2. Cuando todas las causas están ausentes, el efecto está ausente;
3. No hay interacción entre las causas.

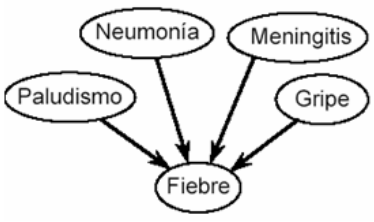
Para ilustrar situaciones en las que estas hipótesis son razonables y otras en las que no lo son, pensemos en los siguientes ejemplos:

Consideremos los siguientes factores, que tienen influencia causal en que una persona desarrolle hipertensión:



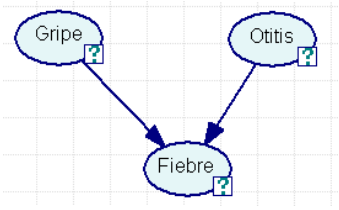
En este caso no conviene utilizar una puerta OR, pues este modelo no representa a *causas directas* que *por sí solas* pueden provocar hipertensión, sino a *factores* que contribuyen a desarrollarla.

Sin embargo, en este otro modelo:



Cada una de las enfermedades puede provocar el síntoma *por sí sola*, aunque no aparezcan las demás. En este caso sí podríamos utilizar la puerta OR, pues se cumplen las hipótesis.

Estudiaremos este modelo de interacción causal a través de un ejemplo sencillo. Supongamos la siguiente red bayesiana:



Bajo estas hipótesis, para construir las probabilidades necesarias para el modelo bastaría con dar las probabilidades de que cada una de las causas provoque el efecto por separado (que denotaremos por cx). Sea por ejemplo:

$c_g = P(+f/+g)=0.8$
$c_o = P(+f/+o)=0.6$

Y en ese caso tendríamos que:

$P(+f/+g, +o)= 0.8 + 0.2*0.6 = 0.92$
$P(+f/\neg g, +o)= 0.6$
$P(+f/+g, \neg o)= 0.8$
$P(+f/\neg g, \neg o)= 0$

Las probabilidades de que no se manifieste el síntoma pueden calcularse como complementarias de éstas, o aplicando la siguiente expresión:

$$P(\neg x/c_1, c_2) = \prod_{i \in T_U} (1 - c_i)$$

Donde T_u representa el conjunto de causas de U que están presentes. Así, tenemos que:

$P(\neg f/+g, +o)= 0.2 * 0.4 = 0.08$
$P(\neg f/\neg g, +o)= 0.4$
$P(\neg f/+g, \neg o)= 0.2$
$P(\neg f/\neg g, \neg o)= 1$

Supongamos que en el modelo queremos incluir que es posible que otras causas no determinadas provoquen también fiebre, introduciendo así cierto ruido en el mismo. Bajo las hipótesis anteriormente mencionadas, es posible construir las probabilidades condicionadas necesarias en el modelo a partir de unos parámetros

básicos, concretamente las probabilidades de que cada causa provoque el efecto por separado y un factor de ruido r, que expresa que otras condiciones no presentes en el modelo podrían provocar la fiebre. Para continuar con el ejemplo, demos unos valores a estos parámetros:

c_g = P(+f/+g, ¬o, ¬r)=0.8
c_o = P(+f/¬g, +o, ¬r)=0.6
r = p(+f/¬g, ¬o)=0.01

Veamos cómo se calculan en este caso las probabilidades condicionadas necesarias. En primer lugar, tenemos que:

P(+f/+g, +o, ¬r)= P(+f/+g, ¬o, ¬r) + P(¬f/+g, ¬o, ¬r) P(+f/¬g, +o, ¬r) = 0.8 + 0.2*0.6 =0,92

Por la hipótesis 2:

P(+f/¬g, ¬o, ¬r) = 0

Las otras probabilidades se calculan mediante:

P(+f/+g, +o)= P(+f/+g, +o, ¬r) + P(¬f/+g, +o, ¬r) P(+f/¬g, ¬o) = 0,92 + 0,08*0.01 = 0,9208
P(+f/+g, ¬o)= P(+f/+g, ¬o, ¬r) + P(¬f/+g, ¬o, ¬r) P(+f/¬g, ¬o) = 0,8 + 0.2*0.01 = 0,802
P(+f/¬g, +o)= P(+f/¬g, +o, ¬r) + P(¬f/¬g, +o, ¬r) P(+f/¬g, ¬o) = 0,6 + 0.4*0.01 = 0,604

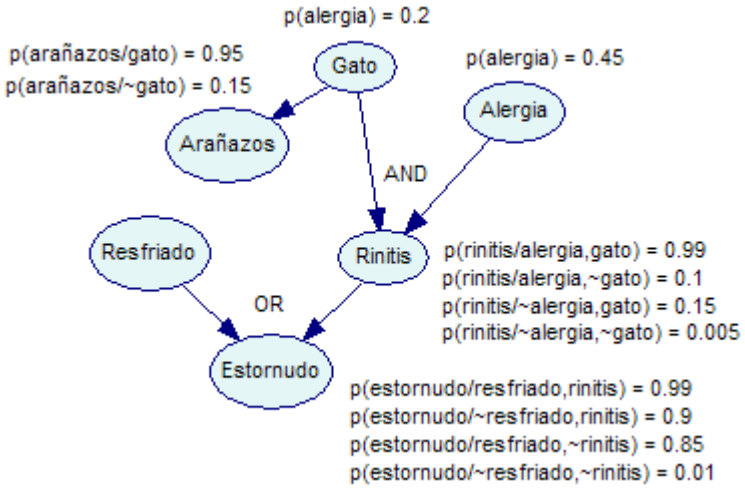
En general, si en un modelo tenemos un efecto X y U_1, …, U_n son las causas posibles, si denotamos por c_i a la probabilidad que tiene cada causa de producir el efecto por separado y por q_i a la probabilidad complementaria (q_i = 1-c_i), entonces:

$$P(\neg x/u_1, ..., u_n) = \prod_{i \in T_U} q_i$$

Existen otros modelos de interacción causal que también están implementados en GeNIe:

- Noisy MAX: La generalización de la puerta OR al caso en que los las variables son multivaluadas. En este caso se considera que el efecto aparece a partir de cierto valor de la variable causa, y se procede de igual modo.
- Noisy ADDER: Según la ley de DeMorgan, el opuesto del Noisy-OR.

Ejercicio 2.7. Consideremos de nuevo el ejemplo del resfriado:



Se pide:

a) Estudia qué relaciones de dependencia condicional deben darse para que la red sea bayesiana, y razona si dichas condiciones resultan adecuadas.
b) ¿Cuántos valores sería necesario especificar para dar la distribución conjunta?
c) ¿Cómo podemos calcular la probabilidad conjunta a partir de las condicionadas?
d) Aplicando el teorema de factorización, explica cómo se calcularía la probabilidad de rinitis alérgica dado que Pablo tiene un gato y que Juan está estornudando (no es necesario realizar los cálculos).

Ejercicio 2.8. Juan y Luisa llegan un día a casa y observan que el coche no está en el garaje, con lo cual piensan que se lo han robado. Cuando Juan está a punto de llamar a la policía, Luisa le dice que no llame, ya que es probable que haya sido María (su hija adolescente) la que haya cogido el coche sin permiso: Juan le pregunta qué le hace pensar eso, y Luisa responde que, además de que el coche de María está en el taller, esa misma mañana María recibió una misteriosa llamada telefónica, lo cual indica que quizás tuviera una cita importante para la que necesitara el coche.

a) Modela la situación con una red bayesiana (nodos, enlaces y parámetros)
b) ¿Qué independencias/dependencias entre las variables de la red implican las hipótesis de independencia condicional?
c) Si suponemos ciertas las hipótesis de independencia condicional, ¿cuántas probabilidades sería necesario especificar? Dar estos valores de una forma coherente con el sentido común.
d) Si no podemos suponer las hipótesis de independencia condicional, ¿qué probabilidades deberíamos pedir al experto? ¿Cuántos valores son, en total?
e) Indica cómo calcularías la probabilidad de que hayan robado el coche sabiendo que el coche no está en el garaje y María ha recibido una llamada.

Ejercicio 2.9. Juan está en la parada del autobús de la línea 20, y el autobús se está retrasando. Juan piensa que puede que haya retenciones de tráfico, pero también puede ser que el autobús haya sufrido una avería o que hayan suspendido el servicio de la línea por las obras del metro. El servicio de una

línea se suspende cuando hay obras que la afectan y hay otras líneas en servicio que pueden utilizar los usuarios para sus desplazamientos.

a) Modela esta situación con una red bayesiana (nodos, enlaces y parámetros)
b) Estudia las relaciones de independencia condicional que se dan en esta red bayesiana
c) Explica cómo se puede calcular la distribución de probabilidad conjunta a partir de las condicionadas.
d) Indica cómo calcularías la probabilidad de que la línea haya sido suspendida dado que el autobús se está retrasando y que hay obras que afectan al recorrido de la línea 20.

Ejercicio 2.10. La policía está intentando establecer un modelo que permita razonar sobre los accidentes de tráfico causados por una pérdida de control del vehículo del conductor. Esta pérdida de control suele venir provocada por un error humano, una carretera resbaladiza, un fallo mecánico o un exceso de velocidad. El error humano suele deberse a una distracción del conductor y una capacidad de reacción mermada por alguna circunstancia (consumo de sustancias o cansancio). La carretera puede estar resbaladiza por vertido de sustancias o por las condiciones atmosféricas.

a) Representa este problema mediante una red bayesiana.
b) Establece unos parámetros para la red acordes con el sentido común.
c) Explica cómo se puede calcular la distribución de probabilidad conjunta a partir de las condicionadas.
d) Ha ocurrido un accidente en el que se ha determinado que la carretera estaba en buenas condiciones y el conductor ha triplicado la tasa de alcohol permitida. Indica como calcularías la probabilidad de que el accidente se deba a un error humano.

Práctica 2. Modelos canónicos El objetivo de esta práctica es conocer los modelos canónicos y su implementación en GeNIe

Tarea. Modelado capítulo House. El objetivo de esta práctica es acercarnos a lo que sería el modelado de un problema real y su implementación en GeNIe

Una vez definido formalmente el concepto de red bayesiana e introducido el proceso de modelado, pasamos a explicar cómo tiene lugar el proceso de razonamiento utilizando redes bayesianas. Para ello sólo vamos a ver el caso más simple, que es el algoritmo de propagación de probabilidades para el caso de redes con forma de árbol. Este caso nos servirá de ejemplo ilustrativo. Para el caso general y algoritmos aproximados, disponemos del software GeNIe.

2.5 Algoritmos de propagación de probabilidades

En esta sección veremos el algoritmo de propagación para el caso en que la red tenga forma de árbol, y un ejemplo de funcionamiento del caso general.

2.5.1 Algoritmo para el caso de redes con forma de árbol[3]

El teorema de factorización proporciona un primer método de actualización de las probabilidades dada la evidencia disponible, puesto que a partir de las probabilidades condicionadas es posible obtener la probabilidad conjunta, y a partir de ésta aplicando la definición de probabilidad condicionada y marginalizando la distribución conjunta sobre el conjunto de las variables de interés es posible conseguir P(X/Y) donde X es cualquier subconjunto de V e Y cualquier conjunto de evidencias. Sin embargo, este procedimiento es computacionalmente costoso. Se han desarrollado muchos otros algoritmos más eficientes para el cálculo de las probabilidades, de los cuales sólo vamos a explicar en profundidad el algoritmo para redes con forma de árbol. El algoritmo consta de dos fases:

Fase de inicialización

En esta fase se obtienen las probabilidades a priori de todos los nodos de la red, obteniendo un estado inicial de la red que denotaremos por S_0.

Fase de actualización

Cuando una variable se instancia, se actualiza el estado de la red, obteniéndose las probabilidades a posteriori de las variables de la red basadas en la evidencia considerada, adoptando la red un estado que denotaremos por S_1.

Este paso se repite cada vez que una variable se instancia, obteniéndose los sucesivos estados de la red.

La idea principal en la que se basa el algoritmo es la siguiente:

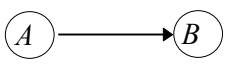
Cada vez que una variable se instancia o bien cuando actualiza su probabilidad, informa a sus nodos vecinos mediante el paso de lo que llamaremos mensajes, de la siguiente forma:

- La variable envía a su padre un mensaje, que llamaremos el λ-mensaje, para informarle de que ha cambiado su valor/probabilidad.
- La variable envía a todos sus hijos un mensaje, que llamaremos el π-mensaje, para informarlos de que ha cambiado su valor/probabilidad.

Así, la información se va propagando por la red tanto en sentido ascendente como descendente.

Estos mensajes asignan a cada variable unos valores que llamaremos λ-valor y π-valor. Multiplicando estos valores obtendremos las probabilidades a posteriori de cada una de las variables de la red.

Tanto los valores como los mensajes son vectores de números. Por ejemplo, supongamos que tenemos el arco:



en el que la variable *A* toma tres valores posibles a_1, a_2, a_3, y la variable *B* toma dos, b_1 y b_2, tendríamos que:

- Si *B* se instancia, enviará un λ-mensaje a *A*,

[3] Tomado de (Neapolitan, 1990)

$\lambda_B(A) = (\lambda_B(a_1), \lambda_B(a_2), \lambda_B(a_3))$.

- Si *A* se instancia, enviará un π-mensaje a *B*,

  $\pi_B(A) = (\pi_B(a_1), \pi_B(a_2), \pi_B(a_3))$.

En función de esos mensajes, tendremos un λ-valor y π-valor para *A*,

$\lambda(A) = (\lambda(a_1), \lambda(a_2), \lambda(a_3))$
$\pi(A) = (\pi(a_1), \pi(a_2), \pi(a_3))$

Y también un λ-valor y π-valor para B,

$\lambda(B) = (\lambda(b_1), \lambda(b_2))$
$\pi(B) = (\pi(b_1), \pi(b_2))$

Multiplicando los valores y normalizando, obtendremos las probabilidades asociadas a *A* o a *B*, según sea el caso.

Pasamos entonces a describir el algoritmo.

Fórmulas de cálculo de λ y π-mensajes, λ y π-valores y probabilidades P*:

e) Si *B* es un hijo de *A*, *B* tiene *k* valores posibles y *A* *m* valores posibles, entonces para *j=1,2,...,m*, el λ-mensaje de *B* a *A* viene dado por;

$$\lambda_B(a_j) = \sum_{i=1}^{k} P(b_i / a_j)\lambda(b_i).$$

f) Si *B* es hijo de *A* y *A* tiene m valores posibles, entonces para *j=1,2,...,m*, el π-mensaje de *A* a *B* viene dado por;

$$\pi_B(a_j) = \begin{cases} \pi(a_j) \prod_{\substack{c \in s(A) \\ c \neq B}} \lambda_c(a_j) & \text{si A no ha sido instanciada (*)} \\ 1 & \text{si } A = a_j \\ 0 & \text{si } A \neq a_j. \end{cases}$$

donde *s(A)* denota al conjunto de hijos de *A*.

(*) Esta fórmula es válida en todos los casos. Otra fórmula de aplicación más sencilla, pero sólo es válida cuando todas las probabilidades $P^*(a_i)$ son no nulas, es $P(a_j)/ \lambda_B(a_j)$. Proporciona un π-mensaje distinto (pero proporcional al de la otra fórmula) e iguales probabilidades a posteriori.

g) Si *B* tiene *k* valores posibles y *s(B)* es el conjunto de los hijos de *B*, entonces para *i=1,2,...,k*, el λ-valor de *B* viene dado por;

$$\pi(b_i) = \begin{cases} \prod_{C \in s(B)} \lambda_C(b_i) & \text{si B no ha sido instanciada} \\ 1 & \text{si } B = b_i \\ 0 & \text{si } B \neq b_i. \end{cases}$$

h) Si *A* es padre de *B*, *B* tiene *k* valores posibles y *A* tiene *m* valores posibles, entonces para *i=1,2,...,k*, el π-valor de *B* viene dado por;

$$\pi(b_i) = \sum_{j=1}^{m} P(b_i / a_j)\, \pi_B(a_j).$$

i) Si *B* es una variable con *k* posibles valores, entonces para *i = 1,2,...,k* la probabilidad a posteriori basada en las variables instanciadas se calcula como:

$P^*(b_i) = \alpha\, \lambda(b_i)\, \pi(b_i)$.

ALGORITMO:

1. Inicialización

A. Inicializar todos los λ-mensajes y λ-valores a 1.
B. Si la raíz *A* tiene m posibles valores, entonces para *j = 1,...,m*, sea

$\pi(a_j) = P(a_j)$

C. Para todos los hijos *B* de la raíz *A*, hacer

Enviar un nuevo π-mensaje a *B* usando la fórmula 2.

(En ese momento comenzará un flujo de propagación debido al procedimiento de actualización C).

Cuando una variable se instancia o una variable recibe un λ o π-mensaje, se usa uno de los siguientes procedimientos de actualización;

2. Actualización

A. Si una variable *B* se instancia a un valor b_j, entonces

BEGIN

A.1. Inicializar $P^*(b_j) = 1$ y $P^*(b_i) = 0$, para todo $i \neq j$.
A.2. Calcular $\lambda(B)$ usando la fórmula 3.
A.3. Enviar un nuevo λ-mensaje al padre de *B* usando la fórmula 1.
A.4. Enviar nuevos π–mensajes a los hijos de *B* usando la fórmula 2.

END

B. Si una variable *B* recibe un nuevo λ-mensaje de uno de sus hijos y la variable *B* no ha sido instanciada todavía, entonces,

BEGIN

B.1. Calcular el nuevo valor de $\lambda(B)$ usando la fórmula 3.
B.2. Calcular el nuevo valor de $P^*(B)$ usando la fórmula 5.
B.3. Enviar un nuevo λ-mensaje al padre de *B* usando la fórmula 1.
B.4. Enviar nuevos π–mensajes a los otros hijos de *B* usando fórmula 2.

END.

C. Si una variable *B* recibe un nuevo π-mensaje de su padre y la variable *B* no ha sido instanciada todavía, entonces,

BEGIN

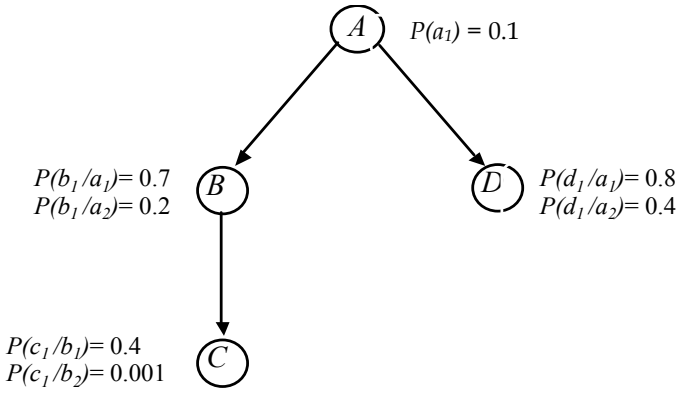
C.1. Calcular el nuevo valor de $\pi(B)$ usando la fórmula 4.
C.2. Calcular el nuevo valor de $P^*(B)$ usando la fórmula 5.
C.3. Enviar nuevos π–mensajes a los hijos de *B* usando fórmula 2.

END.

Para ilustrar el algoritmo de propagación de probabilidades, vamos a utilizar el siguiente ejemplo:

Ejemplo 2. 8.

Supongamos que un señor piensa que su esposa le está siendo infiel. La red bayesiana que se construye para evaluar esta posibilidad es la siguiente:



donde la variable *A* se refiere a si la esposa está engañando al marido o no, la variable *B* se refiere a si la esposa cena con otro hombre o no, la variable C se refiere a si la esposa es vista cenando con otro hombre o no, y la variable D se refiere a si en el domicilio se reciben llamadas telefónicas sospechosas o no. Supondremos que la letra minúscula con el subíndice uno representa a la afirmación del hecho, y la minúscula con el subíndice 2, a la negación.

En primer lugar, vamos a calcular las probabilidades a priori de cada una de las variables de la red. Para hacer esto, también podríamos calcular la probabilidad conjunta como producto de las condicionadas, y luego sumar los términos necesarios para obtener cada probabilidad, pero este método no es computacionalmente factible en la mayoría de los casos. Vamos pues a hacerlo con el algoritmo.

Inicialización

A. Ponemos todos los λ-mensajes y λ-valores a 1.

B. Hacemos $\pi(a_j) = P(a_j)$, para $j = 1,2$. π(A)= (0.1, 0.9).

C. *A* envía un mensaje a su hijo, *B*,

$$\pi_B(a_1) = \pi(a_1)\lambda_D(a_1) = 0.1$$
$$\pi_B(a_2) = \pi(a_2)\lambda_D(a_2) = 0.9$$

B toma entonces nuevos π-valores;

$$\pi(b_1) = P(b_1/a_1)\,\pi_B(a_1) + P(b_1/a_2)\,\pi_B(a_2) = 0.7\ 0.1 + 0.2\ 0.9 = 0.25$$
$$\pi(b_2) = P(b_2/a_1)\,\pi_B(a_1) + P(b_2/a_2)\,\pi_B(a_2) = 0.75$$

Y con ellos y con los λ-valores de B, se obtienen las probabilidades:

$P(b_1) = \alpha\ *0.25*\ 1 = 0.25.$
$P(b_2) = \alpha\ *0.75\ *1 = 0.75.$

Ahora, C recibe un π-mensaje por ser hijo de *B*:

$\pi_C(b_1) = \pi(b_1) = 0{,}25$
$\pi_C(b_2) = \pi(b_2) = 0{,}75$

Y actualiza su π-valor:

$\pi(c_1) = P(c_1/b_1)\ \pi_C(b_1) + P(c_1/b_2)\ \pi_C(b_2) = 0.4\ 0.25 + 0.001\ 0.75 = 0.10075$
$\pi(c_2) = P(c_2/b_1)\ \pi_C(b_1) + P(c_2/b_2)\ \pi_C(b_2) = 0.89925.$

A partir de ellos, calculamos las probabilidades de C, multiplicando por los λ-valores y normalizando:

$P(c_1) = 0.10075.$
$P(c_2) = 0.89925.$

El mismo procedimiento se repite para D, y obtenemos el estado inicial S_0 de la red causal:

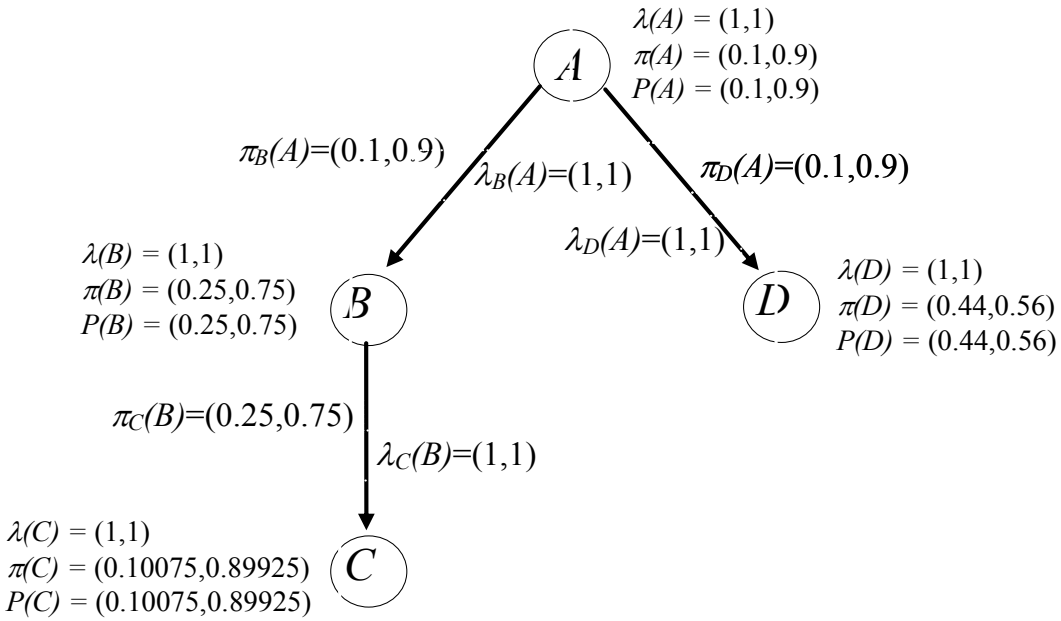


Figura 3.9. Estado S_0 de la red.

Supongamos ahora que nos informan de que la esposa ha cenado con otro, es decir, conocemos ahora con certeza que $B = b_1$.

Esta información se irá transmitiendo por la red, haciendo que las probabilidades a priori de los nodos, *P(X)* cambien a las probabilidades a posteriori, $P^*(X) = P(X/B = b_1)$. En este caso, al ser la evidencia aportada a favor de la hipótesis que queremos probar, lo lógico será que todas estas probabilidades aumenten. En el momento que una variable se actualiza, comienza un flujo de propagación por la red, que en este caso es el siguiente:

- B informa a su padre mediante un λ-mensaje.
- B informa a su hijo mediante un π-mensaje.
- A su vez, A va a informar a su hijo, D, mediante un π-mensaje.

Tras el paso de estos mensajes, todas las variables van a actualizar sus λ y π-valores y sus probabilidades.

Veamos entonces cómo se efectúa la actualización con el algoritmo;

Actualización:

Actualización de B:

A.1. Calculamos ahora la probabilidad a posteriori de *B*, conocido que ha tomado el valor b_1, que evidentemente será;

$P^*(b_1) = 1.$
$P^*(b_2) = 0.$

A.2. Calculamos $\lambda(B)$;

$\lambda(b_1) = 1.$
$\lambda(b_2) = 0.$

A.3. Enviamos un λ-mensaje al padre de *B, A*

$\lambda_B(a_1) = P(b_1/a_1)\lambda(b_1) + P(b_2/a_1)\lambda(b_2) = 0.7\ 1 + 0.3\ 0 = 0.7$
$\lambda_B(a_2) = 0.2$

A.4. Enviamos un π-mensaje al hijo de *B, C*

$\pi_C(b_1) = 1$ puesto que *B* ha sido instanciada a b_1.
$\pi_C(b_2) = 0$ puesto que *B* ha sido instanciada a b_1.

Ahora, al haber recibido A y C nuevos mensajes, tienen que actualizar sus valores;

Actualización de C:
Al recibir *C* un π-mensaje, se dispara el procedimiento de actualización *C*;

C.1. El π-valor de C cambia,

$\pi(c_1) = P(c_1/b_1)\pi_C(b_1) + P(c_1/b_2)\ \pi_C(b_2) = 0.4.$
$\pi(c_2) = 0.6$

C.2. Calculamos la nueva probabilidad de *C*

$P^*(c_1) = 0.4\ \alpha = 0.4$
$P^*(c_2) = 0.6\ \alpha = 0.6$

C.3. No es necesario puesto que *C* no tiene hijos.

Actualización de A. Al recibir *A* un λ-mensaje, se dispara el procedimiento de actualización *B*;

B.1. Actualizamos el λ-valor

$\lambda(a_1) = \lambda_B(a_1)\ \lambda_D(a_1) = 0.7$
$\lambda(a_2) = \lambda_B(a_2)\ \lambda_D(a_2) = 0.2$

B.2. En base al λ-valor, calculamos la probabilidad a posteriori;

$P^*(a_1) = \alpha\ \ 0.7\ \ 0.1 = 0.07\ \alpha = 0.28$
$P^*(a_2) = \alpha\ \ 0.2\ 0.9 = 0.18\ \alpha = 0.72.$

Como era de esperar, la probabilidad de que la esposa sea infiel ha aumentado, porque la evidencia aportada es a favor de la hipótesis.

B.3. *A* no tiene padre.

B.4. A envía un π-mensaje a su hijo, *D*,

$\pi_D(a_1) = \pi(a_1)\ \lambda_B(a_1) = 0.1\ 0.7 = 0.07.$
$\pi_D(a_2) = \pi(a_2)\ \lambda_B(a_2) = 0.9\ 0.2 = 0.18.$

Actualización de D:

Ahora la variable D, debido a la recepción del π-mensaje, comienza el proceso de actualización C

C.1. El π-valor de D cambia,

$\pi(d_1) = 0.128$
$\pi(d_2) = 0.122$

C.2. Calculamos la nueva probabilidad de D

$P^*(d_1) = 0.512$
$P^*(d_2) = 0.488$

C.3. No es necesario puesto que D no tiene hijos.

Así, tras la instanciación de B a b1, la red queda;

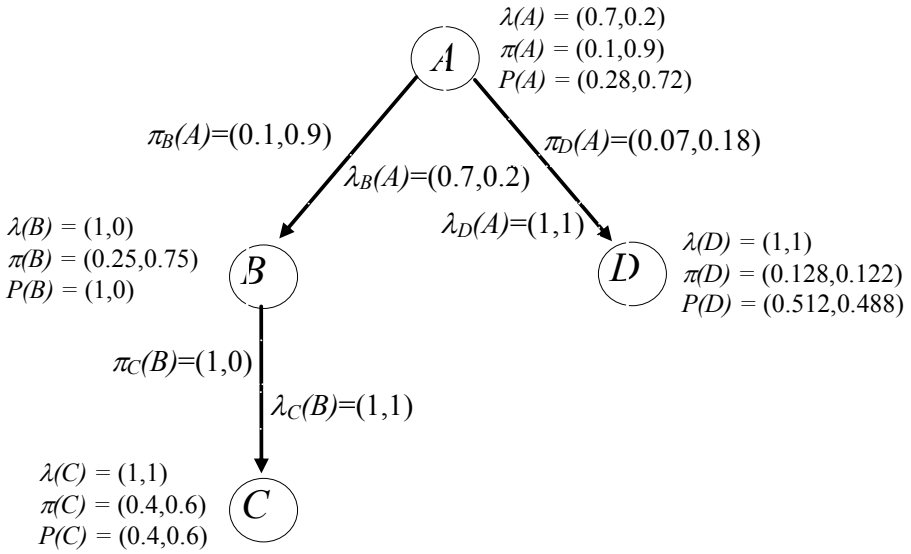


Figura 3.10. Estado S1 de la red.

Supongamos ahora que tenemos la información de que no se han recibido llamadas telefónicas extrañas en el domicilio, es decir, que sabemos que D ha tomado el valor d2.

Nuevamente se iniciará el algoritmo que propagará esta información por la red:

- *D* enviará un λ-mensaje a su padre, *A*,
- *A* enviará un π-mensaje a su hijo, *B*.

Pero ahora, al estar B inicializada, el algoritmo se parará ahí, puesto que P(B) = (1, 0), y no podemos permitir que nada cambie ya estos valores. Así, en la ejecución del algoritmo, las variables que ya han sido inicializadas son extremos muertos, donde la propagación se para (en el caso de la propagación en árboles).

Hacemos pues el paso A de actualización para la variable D,

Actualización:

Actualización de D

A.1. Calculamos ahora la probabilidad a posteriori de *D*,

$P^*(d_1) = 0.$
$P^*(d_2) = 1.$

A.2. Calculamos $\lambda(D)$;

$\lambda(d_1) = 0.$
$\lambda(d_2) = 1.$

A.3. Enviamos un λ-mensaje al padre de *D, A*

$\lambda_D(a_1) = P(d_1/a_1)\lambda(d_1) + P(d_2/a_1)\lambda(d_2) = 0.8\ 0 + 0.2\ 1 = 0.2$
$\lambda_D(a_2) = 0.6$

A.4. No se hace puesto que D no tiene hijos.

Actualización de A

B.1. Calculamos $\lambda(A)$

$\lambda(a_1) = \lambda_B(a_1)\ \lambda_C(a_1) = 0.7\ 0.2 = 0.14$
$\lambda(a_2) = \lambda_B(a_1)\ \lambda_C(a_2) = 0.2\ 0.6 = 0.12$

B.2. Calculamos la probabilidad actualizada de A

$P^*(a_1) = \alpha\ \ 0.014 = 0.1148$
$P^*(a_2) = \alpha\ \ 0.108 = 0.8852$

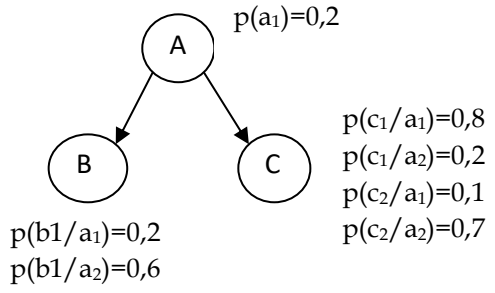
Ahora la probabilidad de a_1 se ha reducido, puesto que la evidencia aportada es *en contra* de a_1.

B.3. A no tiene padre.

B.4. Este paso no se realiza pues B está ya instanciado.

Tras estos cálculos se obtiene un estado de la red, S_2. Este estado es el mismo que obtendríamos si procesásemos la información al revés, es decir, si instanciásemos primero la variable D al valor d_2, y después la variable B al valor b_1.

Ejercicio 2.11. Consideremos la siguiente red bayesiana, donde los nodos A y B son binarios y el nodo C toma tiene tres posibles valores c1, c2 y c3:



Calcula las probabilidades de los valores del nodo A y del nodo C, dado que B toma el valor b_1.

Ejercicio 2.12. Supongamos que construimos un pequeño sistema experto Bayesiano para diagnosticar una enfermedad. Para ello definimos tres variables:

E = presencia de la enfermedad, que toma los valores

e_1= la persona padece la enfermedad
e_2= la persona no padece la enfermedad

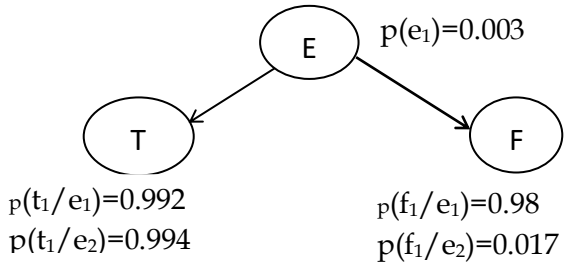
T = resultado de un test indicativo de dicha enfermedad, que toma los valores

t_1= el resultado del test es positivo.
t_2= el resultado del test es negativo.

F = presencia de fiebre en el enfermo, que toma los valores

f_1= el enfermo tiene fiebre.
f_2= el enfermo no tiene fiebre.

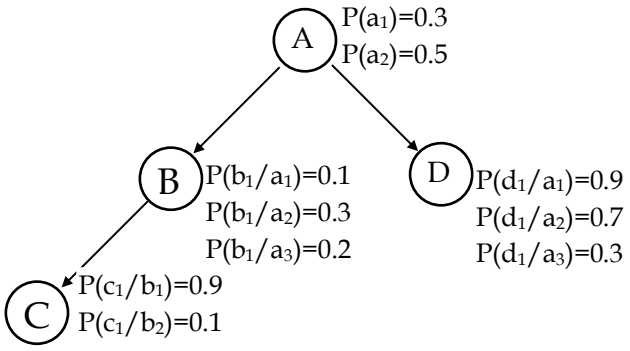
Entre estas variables establecemos las relaciones de influencia causal que se dan en la siguiente figura, donde también se indican los valores de los parámetros necesarios:



Se pide;

a) Inicializar la red bayesiana.
b) Actualizar la red bayesiana sabiendo que el enfermo tiene fiebre.
c) Implementa la red en GeNIe y comprueba que los resultados obtenidos son correctos

Ejercicio 2.13. Consideremos la siguiente red bayesiana:



donde las variables proposicionales son:

A = edad, que toma los valores a_1= joven, a_2= medio, a_3= viejo.
B = nivel de ingresos, que toma los valores b_1= alto, b_2= bajo.
C = nivel de vida, que toma los valores c_1= bueno, c_2= malo.
D = salud, que toma los valores d_1= buena, d_2= mala.

Se pide:

a) Inicializar la red bayesiana.
b) Actualizar las probabilidades sabiendo que la variable nivel de ingresos ha tomado el valor bajo.
c) Implementa la red en GeNIe y comprueba que los resultados obtenidos son correctos

Ejercicio 2.14. Cuando el conducto C funciona correctamente, la cámara A está libre de aire (casi siempre, pero hay un 5% de ocasiones en que tiene aire) y la presión en la cámara B suele ser normal (salvo un 2% de las veces, en las que es elevada). Pero cuando el conducto C se obstruye (lo cual ocurre en el 10% de las ocasiones) casi siempre en la cámara A hay aire (salvo en un 5% de las ocasiones) y la presión en la cámara B es elevada (salvo un 5% de las veces). El exceso de presión de B es indicado por el sensor SB, que presenta un 10% de falsos positivos y un 20% de falsos negativos. La presencia de aire en A es indicada por el sensor SA, que presenta un 15% de falsos positivos y un 15% de falsos negativos. Se pide:

a) Elabora un modelo de red bayesiana para la situación descrita e inicialízalo.
b) Supóngase que el sensor SA arroja una lectura positiva. Actualizar las probabilidades de los otros nodos de la red

Implementa la red en GeNIe y comprueba que los resultados obtenidos son correctos

2.5.2 Ejemplo de funcionamiento del caso general

Como ya hemos mencionado, existen muchos algoritmos diferentes para el caso general, tanto exactos como aproximados, que quedan fuera del alcance de este curso. Pero para ilustrar su funcionamiento vamos a ver un ejemplo de razonamiento en el caso general, utilizando el software GeNIe. Como modelo utilizaremos el desarrollado anteriormente en el ejemplo del estornudo.

El primer paso es realizar el *proceso de inicialización* en el que se calculan todas las probabilidades a priori de todos los nodos de la red

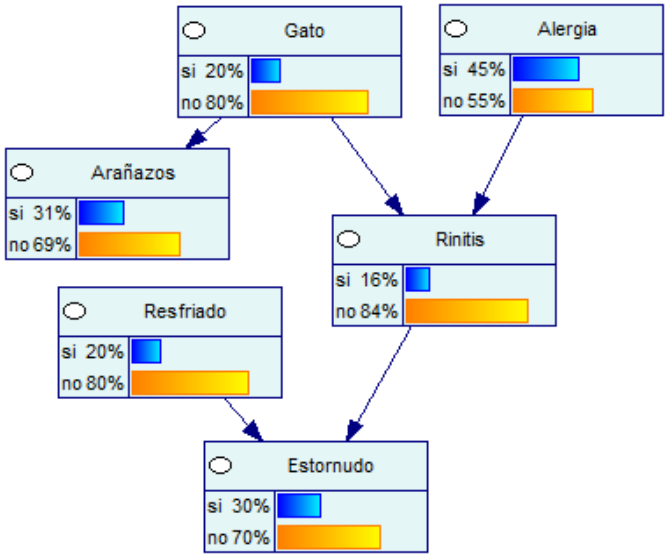


Figura 3.11. Estado S_0 de la red para el ejemplo del estornudo

Cuando Juan empieza a estornudar, las probabilidades se actualizan para contabilizar esta información.

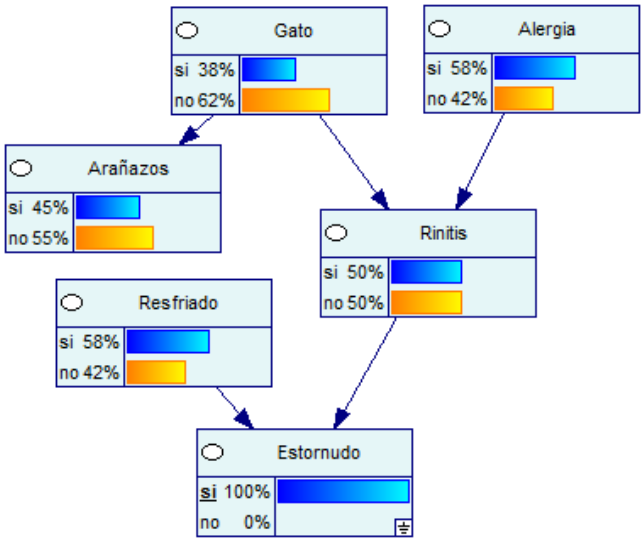


Figura 3.12. Estado S_1 de la red, tras conocer que Juan estornuda

Nótese que ahora las probabilidades de los estados positivos de todos los nodos de la red son mayores. Esto es así porque la evidencia disponible "Juan está estornudando" apoya el estado positivo de todas las variables. Recuerde que para el nodo estornudo, la única relación de independencia era que `estornudo` es independiente de `gato`, `alergia` y `arañazos` dado `rinitis` y `resfriado` (a priori, todos los nodos son dependientes de estornudo, y por eso cambian todas las probabilidades).

La siguiente evidencia disponible es que "Juan ve arañazos". Las probabilidades actualizadas se pueden observar en la siguiente figura:

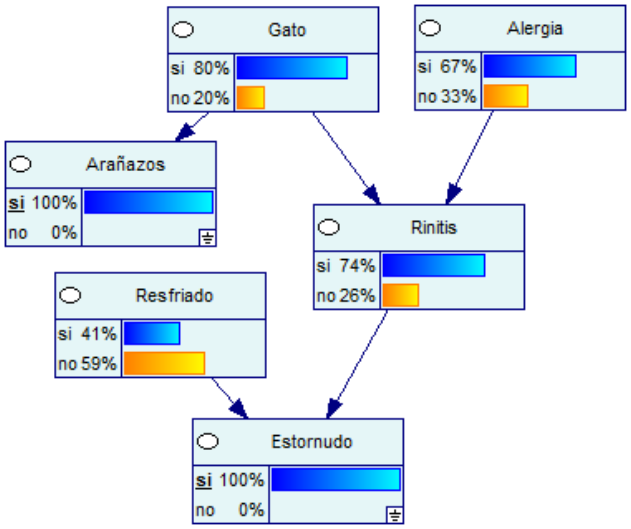


Figura 3.13. Estado S_2 de la red del ejemplo del estornudo

Como puede observarse, ahora la probabilidad de `gato` se ha incrementado porque la evidencia `arañazos` favorece la presencia de `gato`. Al aumentar la probabilidad de `gato`, aumentan también las probabilidades de `rinitis` y `alergia`. La probabilidad de `resfriado` disminuye porque el `estornudo` puede explicarse ahora por la `rinitis` (efecto explaining-away).

Esta información puede presentarse también resumida en una tabla:

	Resfriado=sí	Rinitis=sí	Alergia=sí	Arañazos=sí	Gato=sí	Estornudo=sí
e_1 = {estornudo=sí}	↑	↑	↑	↑	↑	1
e_2 = {arañazos=sí}	↓	↑	↑	1	↑	1

En los siguientes ejercicios debes utilizar el software GeNIe para adquirir una noción intuitiva de cómo cambian las probabilidades de los nodos de la red (aumentan, disminuyen o permanecen igual) según se adquiere la nueva evidencia.

Ejercicio 2.15. Para el Ejemplo 2. 7, explica la evolución de las probabilidades de la red conforme se va adquiriendo nueva evidencia (la evidencia se adquiere tal como se explica en la descripción de la situación), bajo los siguientes supuestos

- Juan no sabe si es alérgico a los gatos o no
- Juan no es alérgico a los gatos
- Juan es alérgico a los gatos

Ejercicio 2.16. Sea la situación descrita en el Ejercicio 2.9. Supongamos ahora que el autobús sigue retrasándose. Juan consulta su teléfono móvil, y en la prensa local se informa de que esta mañana han empezado las obras del metro en zonas que afectan al recorrido de la línea 11. Respecto al estado del tráfico, la radio menciona que en las zonas que atraviesa la línea 11 el tráfico es fluido. De repente, en la carretera de enfrente, Juan ve a un autobús de la línea 11 circulando en sentido contrario. Identifica cuáles son las evidencias disponibles para razonar en esta situación, y explica cómo van evolucionando las probabilidades de los nodos de la red tras ir procesando esta evidencia (en el mismo orden en que Juan la recibe).

Ejercicio 2.17. Sea la situación descrita en el Ejercicio 2.10. Supongamos que utilizamos la red diseñada para analizar un caso de accidente por pérdida de control, en el que se ha determinado que: la carretera estaba en buenas condiciones, el vehículo no ha tenido errores mecánicos pero el conductor ha dado una tasa de alcohol en sangre 3 veces superior a la permitida. Razona cómo evolucionan las probabilidades de cada uno de los nodos de la red (si aumentan, disminuyen o no cambian) tras la introducción de cada una de estas evidencias (en el orden mencionado).

Práctica 4. Inferencias con GeNIe. El objetivo de esta práctica es, dada una red bayesiana, aprender a realizar inferencias sobre los nodos de la red conforme se adquiere nueva evidencia. También aprenderás a utilizar GeNIe como herramienta de diagnóstico

2.6 Bibliografía

Castillo, E., Gutiérrez, J., & Hadi, A. (1997). *Sistemas Expertos y Modelos de Redes Probabilísticas.* Monografías de la Academia Española de Ingeniería.
Díez, F. (2005). *Apuntes de Razonamiento Aproximado.* Universidad Nacional de Educación a Distancia.
Gómez, A., Juristo, N., Montes, C., & Pazos, J. (1997). *Ingeniería del Conocimiento.* Editorial Centro de Estudios Ramón Areces, S.A.

Jensen, F. (1996). *An Introduction to Bayesian Networks.* UCL Press.
Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models.* MIT Press.
Neapolitan, R. (1990). *Probabilistic Reasoning in Expert Systems: Theory and Algorithms.* Wiley-Interscience.
Neapolitan, R. (2009). *Probabilistic Methods for Bioinformatics.* Morgan Kauffman.
Russell, S., & Norvig, P. (2004). *Inteligencia Artificial: Un enfoque moderno.* Prentice-Hall.

TEMA 3. APRENDIZAJE BAYESIANO

Resultados de aprendizaje

Al finalizar este tema, el estudiante deberá ser capaz de:

- Aplicar diversos algoritmos de aprendizaje bayesianos e interpretar sus resultados

Contenidos

3.1 Aprendizaje con redes bayesianas

El aprendizaje con redes bayesianas consiste en, a partir de una base de datos preexistente, construir de modo automático una red bayesiana que lo represente. En el aprendizaje con redes bayesianas, se distinguen dos tipos de situaciones:

- *Aprendizaje paramétrico.* Se supone dada una base de datos de casos, y la estructura de la red (nodos y enlaces). A partir de los datos y la estructura, se aprenden los parámetros (probabilidades a priori de los nodos sin padres, probabilidades condicionadas de cada nodo dados sus padres).
- *Aprendizaje estructural*. En este caso se parte de una base de datos de casos y se aprende de ellos tanto los parámetros como la estructura. Se puede no poner ningún tipo de restricción a la estructura o también partir de información previa que prohíba o fuerce algunos enlaces.

En general, existirá una fase previa de pre-procesamiento de datos. En esta fase se pueden llevar a cabo diversas opciones:

- Depuración de datos. Es posible que en la base de datos existan valores atípicos (outliers) o valores perdidos (missing values). Habrá que decidir si

se descartan, se pre-procesan o se aceptan tal como aparecen (como veremos, los algoritmos habituales de aprendizaje de parámetros pueden aplicarse a datos de entrada con valores perdidos).
- Discretización de valores. Muchos de los algoritmos de aprendizaje de redes bayesianas son de aplicación sólo al caso de variables discretas, por lo que en el caso de que alguna de las variables de nuestra base de datos sea continua será necesario discretizarla. Para ello existen varios métodos (ancho uniforme, frecuencia uniforme, jerárquico, etc.).

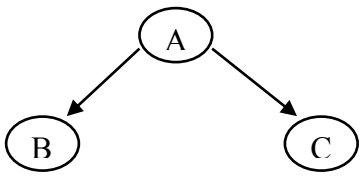
A continuación, exponemos los métodos básicos para el aprendizaje de parámetros (sección 3.1.1) y mencionamos brevemente algunos aspectos del aprendizaje estructural (sección 3.1.2).

3.1.1 Aprendizaje paramétrico

Vamos a distinguir dos casos: el aprendizaje paramétrico con datos completos y el aprendizaje paramétrico con datos incompletos

3.1.1.1 Aprendizaje paramétrico con datos completos

El aprendizaje paramétrico de redes bayesianas se basa en la idea de máxima verosimilitud. Supongamos una red de la que conocemos su estructura, pero no alguno de sus parámetros. Llamaremos modelo a una asignación de valores a los parámetros desconocidos. Por ejemplo, sea la red:



Para definir la red, sería necesario conocer cinco valores:

P(+a), P(+b/+a), P(+b/¬a), P(+c/(+a), P(+c/(¬a).

Supongamos que no conocemos ninguno de ellos. Un modelo $\vec{\theta}$ es una asignación de valores para las probabilidades que se quiere estimar. En el ejemplo, $\vec{\theta}$ es un vector (θ_1, θ_2, ...,θ_5), es decir, una asignación de valores para las probabilidades anteriormente descritas.

Dado un conjunto ϑ de n observaciones O_1, ..., O_n, la verosimilitud de un modelo $\vec{\theta}$ (que en adelante denotaremos $L(\vec{\theta}/\vartheta)$) se define como <u>*la probabilidad de que se den dichas observaciones, supuesto que los parámetros toman los valores del modelo*</u>. Si suponemos que las observaciones son independientes:

$$L(\vec{\theta}/\vartheta) = P(O_1,...,O_n/\vec{\theta}) = \prod_i P(Oi/\vec{\theta}))$$

Por ejemplo, consideremos el caso en que lanzamos una moneda al aire. El modelo en este caso es un solo parámetro, θ_1=*P(cara)*. Supongamos que se han lanzado dos veces la moneda, con resultados cara y cruz. La verosimilitud del modelo es entonces:

$$L(\theta_1/cara, cruz) = P(cara/(P(cara)=\theta_1)*P(cruz/(P(cara)=\theta_1)=\theta_1(1-\theta_1)$$

Nos interesa hallar el modelo de máxima verosimilitud, por lo que buscamos el θ_1 que maximiza la función $\theta_1(1-\theta_1)$. Sabemos que dicha función alcanza su máximo para un valor de $\theta_1=1/2$.

Ejercicio 3.1. Calcular el modelo de máxima verosimilitud si el conjunto de observaciones es: a) cara, cara; b) cruz, cruz.

En general, se puede probar que el modelo de máxima verosimilitud para una red bayesiana de la que tenemos n observaciones se halla asignando a cada parámetro la frecuencia relativa calculada a partir de las n observaciones. En la práctica, en lugar de maximizar la función de verosimilitud L se maximiza su logaritmo, l, que es más sencillo ya que nos transforma el productorio en un sumatorio (y ambas funciones alcanzan su máximo en el mismo punto, al ser el logaritmo una función estrictamente creciente).

$$l(\vec{\theta} / \vartheta) = \log(L(\vec{\theta} / \vartheta)) = \sum_i \log(P(0i / \vec{\theta})$$

O la función promedio del logaritmo, $\hat{l}$, que también alcanza el máximo en el mismo punto:

$$\hat{l}(\vec{\theta} / \vartheta) = \frac{1}{n} \sum_i \log(P(0i / \vec{\theta}))$$

.

Se demuestra que las tres funciones alcanzan sus máximos para los mismos valores de $\vec{\theta}$. Además, como $0 \leq L \leq 1$, se tiene que l y $\hat{l} \leq 0$.

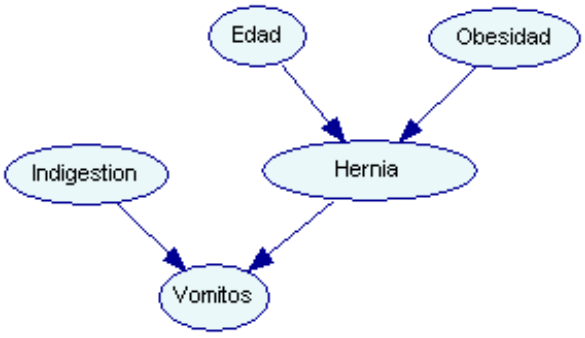
Veamos un ejemplo.

Ejemplo 3. 1. *Cálculo de un modelo basado en redes bayesianas a partir de una base de datos*

Tenemos una base de datos con 20 casos, de los cuales se tiene registrado información sobre cinco variables: edad (con valores menor de 50, mayor o igual a 50); y obesidad, hernia, indigestión y vómitos (binarios con valores sí/no).

Individuos	Edad	Obesidad	Hernia	Indigestión	Vómitos
Individuo 1	Mayor_50	no	no	no	no
Individuo 2	Mayor_50	no	no	no	no
Individuo 3	Mayor_50	no	no	no	no
Individuo 4	Mayor_50	no	no	no	no
Individuo 5	Mayor_50	no	sí	no	sí
Individuo 6	Mayor_50	sí	sí	no	sí
Individuo 7	Mayor_50	sí	sí	no	sí
Individuo 8	Mayor_50	sí	sí	no	si
Individuo 9	Menor_50	no	no	no	no
Individuo 10	Menor_50	no	no	no	no
Individuo 11	Menor_50	no	no	no	no
Individuo 12	Menor_50	no	no	no	no
Individuo 13	Menor_50	no	no	no	no
Individuo 14	Menor_50	no	no	no	no
Individuo 15	Menor_50	no	no	no	no
Individuo 16	Menor_50	no	no	no	no
Individuo 17	Menor_50	no	no	sí	sí
Individuo 18	Menor_50	sí	no	no	sí
Individuo 19	Menor_50	sí	no	sí	sí
Individuo 20	Menor_50	sí	no	no	no

Consultando con un equipo de médicos la relación existente entre estas variables, han proporcionado la siguiente estructura:



Por lo tanto, las probabilidades que debemos aprender son: P(edad), P(obesidad), P(indigestión), P(vómitos/indigestión, hernia), P(hernia/edad, obesidad). Sabemos que en total son 1 + 1 + 1 + 4 + 4, es decir, en este caso el modelo es un vector de 11 componentes.

Para tener la estimación de máxima verosimilitud de estos datos de acuerdo con las observaciones, sabemos que tenemos que utilizar las frecuencias de aparición relativa en las veinte observaciones que tenemos. Podemos calcular como ejemplo dos de esos 11 valores:

- P(edad=menor de 50) = $\frac{\text{nº de casos (edad = Menor\_50)}}{\text{nº total de casos}}$ = 12/20 = 0,6
- P(vómitos=no/indigestión=no, hernia=no)= $\frac{\text{nº casos (vómitos=no,indigestión=no,hernia=no)}}{\text{nº casos (indigestión=no,hernia=no)}}$ = = 13 /14 = 0.92857

Ejercicio 3.2. Calcular las otras distribuciones como ejercicio.

Aprendizaje paramétrico con datos completos en GENIE

A continuación, vamos a describir en una serie de pasos cómo realizar aprendizaje paramétrico con GENIE:

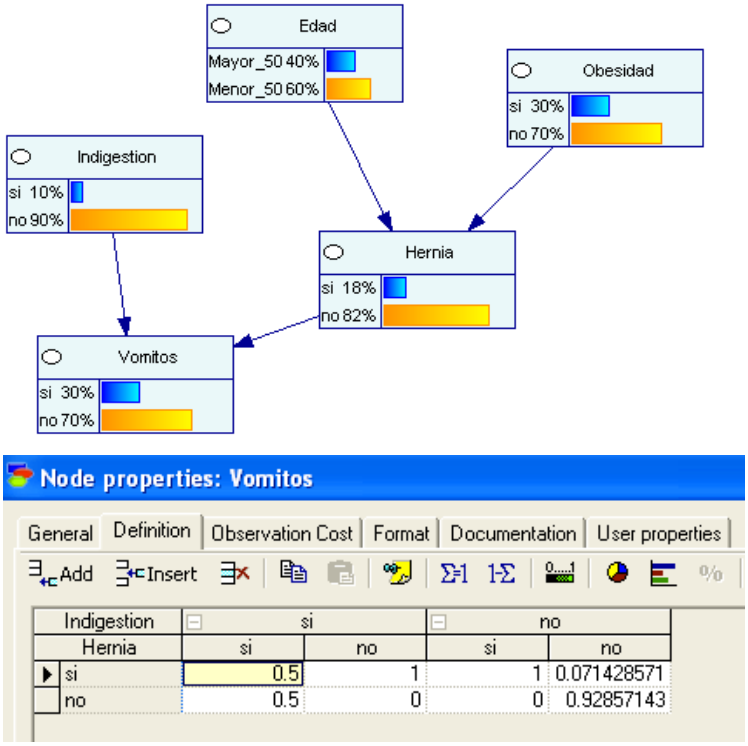
PASO 1. Introducir los datos en un fichero Excel y guardarlos con la extensión csv.

PASO 2. Abrir el fichero de datos "nombre.csv" desde la opción "Open Data File" de GeNIe

PASO 3. Crear la estructura de la red en GeNIe. Los identificadores de los nodos deben ser los mismos que los nombres de las columnas del fichero de datos (caso de que no lo sean, GeNIe nos abrirá un menú para que podamos establecer correspondencias). Las probabilidades no se completan, pues vamos a aprenderlas (por defecto estarán todas a 0.5, que es la distribución inicial uniforme).

PASO 4. En el menú Data, utilizamos la opción "Learn Parameters". Activamos "randomize initial parameters" y desactivamos "enable relevance". La red aparecerá ya con las probabilidades estimadas cargadas.

En el siguiente gráfico vemos las probabilidades a priori aprendidas por GeNIe (en la red) y las condicionadas del nodo vómito, con el procedimiento descrito anteriormente:



Cuando no partimos de un modelo "en blanco", sino que tenemos ya ciertas estimaciones o conjeturas sobre los valores de algunos parámetros, es frecuente representar estas conjeturas como "casos ficticios" adicionales. Veamos un ejemplo:

Supongamos que tenemos una moneda que creemos que está equilibrada. Sin embargo, la hemos lanzado al aire 10 veces y hemos obtenido 9 caras. Pero queremos tener en cuenta nuestra creencia de que la moneda está equilibrada. En este caso, lo que haríamos sería introducir en el conjunto de datos un número de tiradas que elijamos (por ejemplo, 90 tiradas) y estimar que, para esta moneda:

$$P_{est}(\text{cara}) = \frac{9+45}{10+90} = \frac{54}{100} = 0{,}54$$

Es decir, combinamos la experiencia del experto con el resultado del experimento: si la experiencia del experto es que de cada m lanzamientos n resultan ser caras, pero al realizar el experimento con una moneda concreta se obtiene que de s lanzamientos a han resultado ser caras, la probabilidad de cara se estimaría mediante:

$$P_{est}(\text{cara}) = \frac{a+n}{s+m}$$

De modo que por ejemplo si el número de lanzamientos del experimento es mucho mayor (por ejemplo, pongamos 900 caras de 1000 lanzamientos), la probabilidad estimada de cara sería:

$$P_{est}(\text{cara}) = \frac{900+45}{1000+90} = \frac{945}{1090} = 0{,}86$$

3.1.1.2 Aprendizaje paramétrico con información incompleta

Vamos ahora a estudiar cómo se puede hacer el aprendizaje de los parámetros cuando a algunas observaciones les faltan los valores de algunas de las variables. Existen varias alternativas para tratar este caso. Entre ellas, las más frecuentes son:

- Alternativa 1. Eliminar los casos en los que falta el valor de alguna variable (se suele aplicar si el conjunto de datos es suficientemente numeroso).
- Alternativa 2. Sustituir el valor perdido por algún valor concreto (por ejemplo: media, mediana, moda, etc.)
- Alternativa 3. Considerar el valor más probable, dado los valores que han tomado las otras variables.

La primera alternativa puede resultar adecuada si la base de datos es muy numerosa, pero, si no es el caso, estamos perdiendo información.

La segunda alternativa no suele dar buenos resultados, pues no tiene en cuenta el valor de otras variables.

La tercera alternativa es la que general suele dar mejores resultados (y la que podemos encontrar implementada en GENIE). La idea es muy simple: se calculan los parámetros de la red bayesiana que mejor se ajusta a los datos existentes, y, utilizando dicha red, se estiman las probabilidades de cada uno de los valores de los datos desconocidos. Es el conocido algoritmo heurístico EM (Expectation Maximization), que alterna pasos de "Expectation" (se calculan los valores esperados en función del modelo) con "Maximization" (se calcula el modelo que maximiza la verosimilitud, es decir, el modelo basado en las frecuencias relativas), dados los valores conocidos y los estimados en el paso anterior. A continuación, describimos el algoritmo:

Algoritmo EM

Entrada:

Conjunto de datos (con algunos valores incompletos) + Estructura red

Salida:

Estimación de los parámetros de la red (conjunto de valores para $\vec{\theta}$)

Pasos:

1. $\vec{\theta} \leftarrow \vec{\theta}_0$;
2. $l \leftarrow \hat{l}(\vec{\theta}_0)$;
3. Repetir:
4. Para cada dato desconocido $d \in \vartheta$;
5. estimar d mediante $\vec{\theta}$;
6. $\vec{\theta} \leftarrow EMV(\vartheta)$;
7. $l' \leftarrow \hat{l}(\vec{\theta})$
8. $\Delta l = l' - l$;
9. $l \leftarrow l'$;
10. Hasta que $\Delta l \leq \varepsilon$

Describamos más detalladamente las instrucciones de cada línea

Línea 1. Se asigna una distribución inicial al modelo. Dicha distribución inicial se puede hacer o bien utilizando estimaciones de expertos (si existen). En ausencia total de información, se comienza con un modelo uniforme.

Línea 2. Se calcula el valor promedio del logaritmo de la función de verosimilitud de dicho modelo. Sea l dicho valor promedio.

Línea 3. Repetir:

Línea 4. Para cada dato desconocido d de las observaciones:

Línea 5. Estimamos d a partir de los valores actuales de $\vec{\theta}$ (fase *expectation*)

Línea 6. Estimamos un nuevo valor para $\vec{\theta}$ aproximando a partir de las frecuencias relativas esperadas (fase *maximization*)

Línea 7. Calculamos el valor promedio del logaritmo de la función de verosimilitud del nuevo modelo calculado para $\vec{\theta}$, llamémosle l'

Línea 8. Calculamos $\Delta l = l' - l$

Línea 9. Asignamos a l el valor de l'

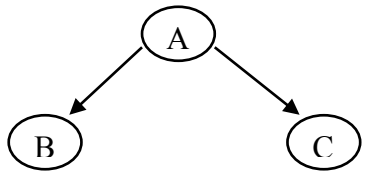
Línea 10. Repetimos los pasos 4 al 9 hasta que Δl sea menor que cierto valor prefijado

La principal limitación de este algoritmo es que puede caer en óptimos locales, por lo que los valores finales obtenidos pueden depender de la inicialización.

Ejemplo 3. 2. *Ejemplo de cálculo de parámetros*

Supongamos que tenemos la red bayesiana anterior, junto con un conjunto de seis observaciones:

Red:



Observaciones:

	A	B	C
O1	+a	+b	+c
O2	+a	+b	¬c
O3	+a	¬b	¬c
O4	¬a	+b	¬c
O5	¬a	¬b	+c
O6	+a	¬b	

Vamos a calcular los modelos para los parámetros con cada una de las alternativas.

Alternativa 1. Para estimar la distribución de probabilidad suprimimos el caso con valores desconocidos, es decir, vamos a utilizar los valores en color negro de la siguiente tabla:

	A	B	C
O1	+a	+b	+c
O2	+a	+b	¬c
O3	+a	¬b	¬c
O4	¬a	+b	¬c
O5	¬a	¬b	+c
O6	+a	¬b	

θ_1 = P(+a) = 3/5
θ_2 = P(+b/+a) = 2/3
θ_3 = P(+b/¬a) = 1/2
θ_4 = P(+c/+a) = 1/3
θ_5 = P(+c/¬a)= 1/2

La probabilidad de cada una de las observaciones dada esta distribución $\vec{\theta}$ sería:

P(01/ $\vec{\theta}$) = P(+a)*P(+b/+a)*P(+c/+a) = (3/5)*(2/3)*(1/3) = 0,1333
P(02/ $\vec{\theta}$) = P(+a)*P(+b/+a)*P(¬c/+a) = (3/5)*(2/3)*(2/3) = 0,2667
P(03/ $\vec{\theta}$) = P(+a)*P(¬b/+a)*P(¬c/+a) = (3/5)*(1/3)*(2/3) = 0,1333
P(04/ $\vec{\theta}$) = P(¬a)*P(+b/¬a)*P(¬c/¬a) = (2/5)*(1/2)*(1/2) = 0,1
P(05/ $\vec{\theta}$) = P(¬a)*P(¬b/¬a)*P(+c/¬a) = (2/5)*(1/2)*(1/2) = 0,1
P(06/ $\vec{\theta}$) = P(+a)*P(¬b/+a)*P(¬c/+a) + P(+a)*P(¬b/+a)*P(+c/+a) = 0,1333 + (3/5)*(1/3)*(1/3) = 0,2

Si ahora calculamos (de acuerdo con la fórmula dada) el promedio del logaritmo neperiano de dichos valores $\hat{l}(\vec{\theta})$ (llamémosle en adelante p, para simplificar) obtenemos que:

p = -1,927694

Alternativa 2. Consistiría en asignar un valor a los datos perdidos, teniendo en cuenta los otros valores de esa variable en las observaciones. Como en este caso las variables no son numéricas, podemos por ejemplo reemplazarlo por la moda (valor más probable). Es decir, trabajaríamos ahora con las siguientes observaciones:

	A	B	C
O1	+a	+b	+c
O2	+a	+b	¬c
O3	+a	¬b	¬c
O4	¬a	+b	¬c
O5	¬a	¬b	+c
O6	+a	¬b	¬c

Y tendríamos que:

θ_1 = P(+a) = 4/6 = 0,6667
θ_2 = P(+b/+a) = 1/2
θ_3 = P(+b/¬a) = 1/2
θ_4 = P(+c/+a) = 1/4
θ_5 = P(+c/¬a)= 1/2

La probabilidad de cada una de las observaciones dada esta distribución $\vec{\theta}$ sería:

P(01/ $\vec{\theta}$) = P(+a)*P(+b/+a)*P(+c/+a) = 0,083
P(02/ $\vec{\theta}$) = P(+a)*P(+b/+a)*P(¬c/+a) = 0,25
P(03/ $\vec{\theta}$) = P(+a)*P(¬b/+a)*P(¬c/+a) = 0,25

$P(04/\vec{\theta}) = P(\neg a)*P(+b/\neg a)*P(\neg c/\neg a) = 0{,}083$

$P(05/\vec{\theta}) = P(\neg a)*P(\neg b/\neg a)*P(+c/\neg a) = 0{,}083$

$P(06/\vec{\theta}) = P(+a)*P(\neg b/+a)*P(\neg c/+a) = 0{,}25$

Si ahora calculamos (de acuerdo con la fórmula dada) el promedio del logaritmo neperiano de dichos valores p, obtenemos que:

p = -1,935605[1].

Al ser este promedio menor que en la alternativa 1, vemos que la estimación obtenida para $\vec{\theta}$ con la alternativa 2 es peor que la obtenida con la alternativa 1.

Alternativa 3. Vamos a aplicar el algoritmo EM para estimar el modelo de máxima verosimilitud. Partimos de un modelo arbitrario en el que a todas las probabilidades asignamos el valor 0.5:

$\theta_{0,1} = P(+a) = 0.5$
$\theta_{0,2} = P(+b/+a) = 0.5$
$\theta_{0,3} = P(+b/\neg a) = 0.5$
$\theta_{0,4} = P(+c/(+a) = 0.5$
$\theta_{0,5} = P(+c/(\neg a) = 0.5$

En este caso p = -1,9639. Es decir, esta estimación $\vec{\theta}_0$ del modelo es peor que la obtenida en las alternativas 1 y 2.

Continuamos iterando, y para ello vamos a vamos a calcular el valor esperado de la variable C en la observación incompleta (fase *expectation*). Será:

$P(+c/+a, \neg b) = P(+c/+a) = 0.5$
$P(\neg c/+a, \neg b) = 1- P(+c/+a) = 0.5$

Añadimos esta información al conjunto de observaciones:

A	B	C
+a	+b	+c
+a	+b	¬c
+a	¬b	¬c
¬a	+b	¬c
¬a	¬b	+c
+a	¬b	P(+c/+a, ¬b) = 0.5 P(¬c/+a, ¬b) = 0.5

Y ahora recalculamos a partir de este conjunto de observaciones el nuevo modelo (fase *maximization*):

$\theta_{1,1} = P(+a) = 4/6 = 0{,}6667$
$\theta_{1,2} = P(+b/+a) = 2/4 = 0{,}5$
$\theta_{1,3} = P(+b/\neg a) = 1/2 = 0{,}5$
$\theta_{1,4} = P(+c/+a) = (1 + 0{,}5)/4 = 0{,}375$
$\theta_{1,5} = P(+c/\neg a) = ½ = 0{,}5$

[1] Si resolvemos este problema con GeNIe vemos que el valor del indicador utilizado para medir la bondad de la estimación es p=11,6136. El motivo es que GeNIe utiliza la suma en lugar del promedio, por lo que dicho valor puede obtenerse multiplicando 1,9356 por el número de casos (6).

En este caso p = -1,8808. Es decir, esta estimación $\vec{\theta}_1$ del modelo es mejor que las tres obtenidas hasta el momento (su verosimilitud es mayor).

Continuamos iterando, calculando el valor esperado de la variable C en la observación incompleta. Será:

P(+c/+a, ¬b) = P(+c/+a) = 0.375
P(¬c/+a, ¬b) = 1- P(+c/+a) = 0.625

Añadimos ahora esta información al conjunto de observaciones

A	B	C
+a	+b	+c
+a	+b	¬c
+a	¬b	¬c
¬a	+b	¬c
¬a	¬b	+c
+a	¬b	P(+c/+a, ¬b) = 0.375 P(¬c/+a, ¬b) = 0.625

Vemos que como sólo cambian los valores de la variable desconocida, el único valor que cambia en los nuevos modelos es el cuarto. En el caso del modelo segundo, tendremos ahora que todos los valores son iguales que en el primero, excepto:

$\theta_{2,4}$ = P(+c/(+a) = (1 + 0,375)/4 = 0,34375

En este caso p = -1,8791. Es decir, esta estimación $\vec{\theta}_2$ del modelo es mejor que todas las obtenidas hasta ahora.

Continuamos iterando:

P(+c/+a, ¬b) = P(+c/+a) = 0.34375
P(¬c/+a, ¬b) = 1- P(+c/+a) = 0.6562

En el nuevo modelo ahora cambiaría el valor:

$\theta_{3,4}$ = P(+c/(+a) = (1 + 0,34375)/4 = 0,3359

En este caso p = -1,8790[2]. Es decir, esta estimación $\vec{\theta}_3$ del modelo es mejor que todas las obtenidas hasta ahora.

Supongamos que la condición de parada fuese que la variación en estos promedios sea menor o igual que 0.001. Entonces el algoritmo pararía, y devolvería como estimación el modelo $\vec{\theta}_3$

El algoritmo EM permite también estimar valores de variables ocultas, es decir, de aquellas que no se conoce ninguno de sus valores. Esto resulta muy útil cuando se piensa que en el modelo existe una variable oculta que puede explicar las demás.

Ejercicio 3.3. Repite el mismo ejercicio de los apuntes a partir del mismo conjunto de observaciones, si la estructura de la red es cabeza-con-cola.

[2] En GeNIe el algoritmo también se detiene en esta distribución, con un valor de p de -11,2739.

Red:

Observaciones:

	A	B	C
O1	+a	+b	+c
O2	+a	+b	¬c
O3	+a	¬b	¬c
O4	¬a	+b	¬c
O5	¬a	¬b	+c
O6	+a	¬b	

Ejercicio 3.4. En el mismo ejemplo anterior, y, a partir del modelo obtenido con la alternativa 1, realiza una iteración del algoritmo EM.

3.1.2 Aprendizaje estructural

En el aprendizaje estructural los nodos y enlaces no se suponen dados, sino que es la misma estructura de la red la que se aprende a partir de los datos. Hay dos enfoques básicos, basados respectivamente en pruebas de independencia y en búsqueda en el espacio de grafos. En los métodos basados en pruebas de independencia se parte de la misma definición de red bayesiana a partir del concepto de independencia condicional.

Recordemos que, en una red bayesiana, una variable es independiente de todas las demás (salvo sus descendientes) dados sus padres. Por tanto, podemos seguir el siguiente procedimiento ingenuo: para cada par de variables X_i, X_j y cada conjunto de variables Z, estimar a partir de los datos si Xi y Xj son independientes dado Z. De esta manera sabremos qué arcos deben estar presentes y qué arcos deben estar ausentes en la red.

Así planteado, el método es muy ineficiente (calcule el lector cuántos tests de independencia habrá que realizar con, por ejemplo, 20 variables) y requiere bases de datos demasiado grandes; pero haciendo alguna suposición adicional sobre el máximo número de padres e hijos de cada nodo, es posible diseñar métodos que resultan aplicables en la práctica. El método más conocido de esta familia es el algoritmo PC de Spirtes, Glymour y Scheines (1993).

En los métodos basados en búsqueda se realiza una búsqueda en el espacio de Grafos Acíclicos Dirigidos (GADs), buscando el GAD que maximice cierta medida de ajuste. Una medida muy usada es el BIC o Bayesian Information Criterion" definido por

$$BIC(G/\vartheta) = \log(\vartheta/G,\vec{\theta}) - \frac{d}{2}\log m$$

donde:

- G es la estructura candidata.
- ϑ son las observaciones.
- $\vec{\theta}$ es la estimación de máxima verosimilitud de los parámetros de G dado ϑ
- d es el tamaño de G (número de parámetros)
- m es el tamaño de O.

Es decir, el BIC es la mejor verosimilitud (logarítmica) que puede obtenerse con la estructura candidata, menos un término que penaliza la complejidad de la estructura.

Pero el espacio de DAGs es muy grande: por ejemplo, si hay 10 variables, el número de posibles DAGs es 4,2*1018. Por ello hay que emplear algoritmos incompletos. Un algoritmo básico de búsqueda local parte de un DAG inicial aleatorio (o, mejor, basado en cierta estimación previa) y tiene como operadores de búsqueda añadir un arco, eliminar un arco e invertir un arco. En cada paso se aplicaría el operador que maximice el incremento del ajuste; si ninguno lo incrementa, se devolvería el DAG como solución, es decir, como estructura de la red. Este procedimiento básico lleva en general a máximos de la función de ajuste que son meramente locales, por lo que habría que complicarlo (búsqueda iterada, recocido simulado, algoritmos genéticos, etc.) Este tipo de algoritmos son los que podemos encontrar en GENIE.

En general, como ya se dijo, el aprendizaje estructural no se realiza de forma puramente automática, sino de forma interactiva: el ingeniero del conocimiento debe guiar la búsqueda, imponiendo la existencia de ciertos arcos que corresponden a relaciones causales conocidas, prohibiendo o descartando otros que no tienen sentido, etc.

3.2 El clasificador Naive-Bayes

Supongamos que deseamos clasificar un objeto en una clase Y, dado un conjunto de rasgos, representados por vectores $\mathbf{X}_i$, i=1,…, n. Supongamos que tenemos un conjunto de ejemplos ($\mathbf{x}_1$, y_1), ..., ($\mathbf{x}_n$, y_n). Supongamos que tenemos un nuevo ejemplo $\mathbf{x}_q$ del que nos interesa conocer el valor y_q.

El objetivo del aprendizaje supervisado es encontrar una función h que, dado el nuevo punto $\mathbf{x}_q$ sea capaz de dar un *estimador* para $y_q=h(\mathbf{x}_q)$

- Si el rango de la función h es continuo, se llama un problema de *regresión*.
- Por el contrario, si el rango de la función es discreto, tenemos un problema de *clasificación*. Si el número de clases es 2, diremos que es un problema de *clasificación binario*, en caso contrario se dice que es un problema de *clasificación multiclase*.

De este modo, para resolver un problema de clasificación, debemos construir una *función estimadora h* (que llamaremos *hipótesis*). El estimador tiene un *buen rendimiento* si *generaliza* bien, es decir, si es capaz de predecir correctamente el valor de y para los ejemplos nuevos. De acuerdo con el *principio de la navaja de Occam*[3], se busca el estimador más simple posible.

Los clasificadores probabilísticos son aquellos que determinan la probabilidad de que un nuevo ejemplo dado $\mathbf{x}_d$ pertenezca a cada una de las clases y, es decir, $P(y \mid \mathbf{x}_q)$. En ese caso, la clase estimada y_q se calculará como la clase más probable, es decir,

$$y_q = y \in V \text{ tal que } P(y/x_q) \text{ es máximo} = \arg\max_{\{y \in V\}} P(y \mid x_q)$$

Uno de los clasificadores más utilizados es el clasificador Naive-Bayes. El nombre "*naive*" se utiliza porque se basa en una hipótesis "*ingenua*". Esta hipótesis supone independencia condicional de los rasgos dada la clase, lo cual es cierto sólo

[3] La navaja de Ockham es un principio metodológico y filosófico atribuido a Guillermo de Ockham (1280-1349), según el cual, "*en igualdad de condiciones, la explicación más sencilla suele ser la correcta*".

en algunos casos. Por ejemplo, consideremos los dos siguientes problemas de clasificación:

- Clasificar los animales en función de rasgos como el color, el número de patas y el tipo de piel
- Clasificar pacientes femeninas según si tienen cáncer de mama o no, atendiendo a rasgos tales como su edad, si tienen situación de menopausia y si fuman.

Supongamos que representamos estos problemas en forma de red bayesiana:

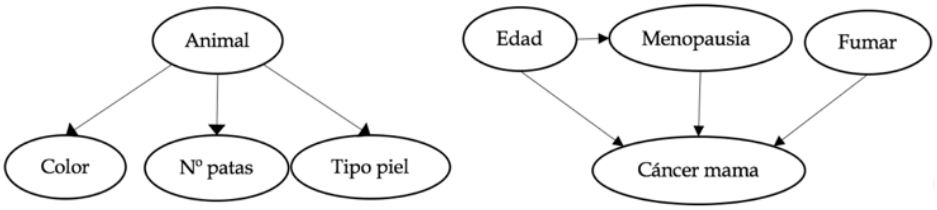
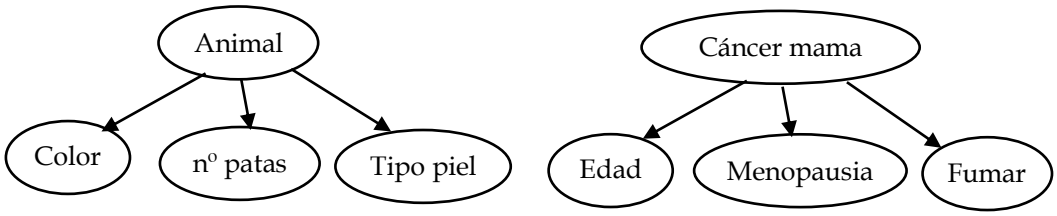


Figura 3.5. Dos problemas de clasificación

Vemos que en el primer ejemplo sí se dan las hipótesis de independencia condicional de los rasgos dada la clase, mientras que en el segundo no se dan.

El clasificador Naive Bayes va a tratar estos dos ejemplos con la misma estructura de red, es decir, trabajará con los siguientes grafos:



Aunque como sabemos, el segundo no se corresponde con la situación real.

Pese a esta suposición ingenua, el clasificador Naive-Bayes es muy popular, por su simplicidad y por tener un rendimiento comparable al de otros métodos más complejos como las redes neuronales y los árboles de decisión. Se ha aplicado con éxito en conjuntos de entrenamiento de tamaño medio o grande, en problemas de diagnóstico, clasificación de documentos, desarrollo de filtros anti SPAM, etc. Ilustremos su funcionamiento con un ejemplo antes de describir la técnica.

Ejemplo 3. 3. *Funcionamiento del clasificador Naive Bayes*

Consideremos la siguiente base de datos:

Día	Cielo (T)	Temperatura (Te)	Humedad (H)	Viento (V)	Jugar al tenis (JT)
D1	soleado	fría	normal	débil	sí
D2	soleado	suave	normal	fuerte	sí
D3	cubierto	fría	normal	fuerte	sí
D4	cubierto	cálido	alta	débil	sí
D5	cubierto	cálido	normal	débil	sí

D6	cubierto	suave	alta	fuerte	sí
D7	Lluvia	suave	alta	débil	sí
D8	Lluvia	suave	normal	débil	sí
D9	Lluvia	fría	normal	débil	sí
D10	soleado	cálido	alta	débil	no
D11	soleado	cálido	alta	fuerte	no
D12	soleado	suave	alta	débil	no
D13	lluvia	fría	normal	fuerte	no
D14	lluvia	suave	alta	fuerte	no

En este ejemplo consideramos que los rasgos son el tiempo, la temperatura, la humedad y el viento, y la clase es Jugar al Tenis. Para aplicar el modelo de Naive Bayes, consideraremos que los rasgos son independientes dado la clase (por ejemplo, consideraremos que la temperatura de un día no tiene nada que ver con que el día sea soleado o llueva, hipótesis que en este caso sabemos que no es cierta).

Tenemos ahora un ejemplo nuevo que querremos clasificar, en el que el tiempo es soleado, la temperatura fría, la humedad alta y el viento es fuerte, y nos interesa saber si en este caso el valor más probable de la variable JT será sí o no. Es decir, si llamamos E a la evidencia disponible E = {T=soleado, T=fría, H=alta, V=fuerte}, para asignar este nuevo ejemplo a la clase correspondiente lo que hacemos es calcular P(JT=sí/E) y P(JT=no/E) y asignaremos el nuevo ejemplo al valor de la clase que haga que se maximice esa probabilidad.

Lo primero que tenemos que hacer calcular un estimador de las probabilidades de la clase y de cada variable dada la clase para los datos del nuevo ejemplo:

P(JT=sí) = 9/14	P(JT=no) = 5/14
P(C=soleado/JT=sí) = 2/9	P(C=soleado/JT=no) = 3/5
P(T=fría/JT=sí) = 3/9 =1/3	P(T=fría/JT=no) = 1/5
P(H=alta/JT=sí) = 3/9=1/3	P(H=alta/JT=no) = 4/5
P(V=fuerte/JT=sí) = 3/9=1/3	P(V=fuerte/JT=no) = 3/5

Ahora, para clasificar el nuevo ejemplo, tenemos que determinar si es mayor P(JT=sí/E) o P(JT=no/E). Para ello necesitamos calcular P(JT/E). Si aplicamos la definición de probabilidad condicionada tenemos que:

$$P(JT/E) = \frac{P(JT,E)}{P(E)} = \frac{P(JT)\,P(E/JT)}{P(E)}$$

En este caso, sabemos que P(E) juega el papel de una constante de normalización, y como en realidad lo que nos interesa saber para clasificar el ejemplo es si es mayor P(JT=sí/E) o P(JT=no/E), podemos limitarnos a comprobar si es mayor P(JT=sí,E) o P(JT=no,E).

Calculemos en primer lugar la probabilidad del valor positivo:

P(JT=sí, E) = P(JT=sí)* P(C=soleado, T=fría, H=alta, V=fuerte/JT=sí)[4]=
P(JT=sí)*P(C=soleado/JT=sí)*P(T=fría/JT=sí)*P(H=alta/JT=sí)*
P(V=fuerte/JT=sí) = $\frac{9}{14} * \frac{2}{9} * \frac{1}{3} * \frac{1}{3} * \frac{1}{3} = \frac{2}{378}$ = 0,0053

De igual modo,

[4] Por la hipótesis *naive*, se supone la independencia de los rasgos dada la clase y por tanto P(C=soleado, T=fría, H=alta, V=fuerte/JT=sí) =
P(JT=sí)*P(C=soleado/JT=sí)*P(T=fría/JT=sí)*P(H=alta/JT=sí)* P(V=fuerte/JT=sí)

P(JT=no/E) = P(JT=no)*P(C=soleado,T=fría,H=alta,V=fuerte/JT=no)=
$\frac{5}{14} * \frac{3}{5} * \frac{1}{5} * \frac{4}{5} * \frac{3}{5} = \frac{36}{1750}$ = 0,0206.

De modo vemos que es más probable el valor JT =no, de modo que asignaríamos el nuevo ejemplo a esa clase.

Sin embargo, podría darse el caso que haya conjuntos de datos en los que para cierto valor de la clase haya valores de rasgos que no se observen. Por ejemplo, supongamos la base de datos fuese:

Día	Cielo (T)	Temperatura (Te)	Humedad (H)	Viento (V)	Jugar al tenis (JT)
D1	soleado	fría	normal	débil	sí
D2	soleado	suave	normal	fuerte	sí
D3	cubierto	fría	normal	fuerte	sí
D4	cubierto	cálido	alta	débil	sí
D5	cubierto	cálido	normal	débil	sí
D6	cubierto	suave	alta	fuerte	sí
D7	lluvia	suave	alta	débil	sí
D8	lluvia	suave	normal	débil	sí
D9	lluvia	fría	normal	débil	sí
D10	soleado	cálido	alta	débil	no
D11	lluvia	cálido	alta	fuerte	no
D12	lluvia	suave	alta	débil	no
D13	lluvia	fría	normal	fuerte	no
D14	lluvia	suave	alta	fuerte	no

Tendríamos que la estimación para P(C=soleado/JT=no) = 0. Este término dominará al clasificador, pues siempre que llegase un nuevo ejemplo en el que C fuese soleado, P(JT=no/E) sería siempre 0 y lo clasificaría como JT=sí. Para evitar este problema se utilizan técnicas como el llamado *Laplace smoothing* (*suavizado de Laplace o corrección de Laplace),* que además de tener en cuenta la frecuencia de aparición en la tabla de los datos considera también estimaciones previas de la probabilidad (*p*) y la naturaleza de los datos (m), como se detallará más adelante.

3.2.1 Descripción general del clasificador Naive Bayes

Supongamos que tenemos un problema de clasificación, y sea Y la variable clase.

Paso 1. En primer lugar, a partir del conjunto de datos que tenemos, para cada uno de los rasgos X se calcula un estimador para sus probabilidades, dada la clase P(**X**=$\mathbf{x}_d$/Y=y). Por simplicidad, escribiremos P(x_d/y). Para calcular dicho estimador, llamemos

n = número de ejemplos de la clase y,
n* = número de ejemplos de la clase y en las que **X** toma el valor x_d

Entonces, la forma más simple de estimar el valor que necesitamos es considerar el número de casos posibles dividido por el número de casos favorables, es decir:

$$\mathrm{P}(\mathrm{x}_d/\mathrm{y}) \sim \frac{n}{n*}$$

Este estimador se conoce como ***el estimador de máxima verosimilitud***. Es el modelo más sencillo, pero como inconveniente nos encontramos con que se necesita una base de datos grande para obtener una buena estimación (como

hemos visto en el ejemplo anterior). Para paliar esos inconvenientes se propone la *corrección (o suavizado) de Laplace*: si partimos de una distribución de probabilidad inicial p a la que damos una importancia m[5], la estimación de la probabilidad a posteriori de X se construye del siguiente modo:

$$\mathrm{P}(\mathbf{x}_d/\mathrm{y}) \sim \frac{n + mp}{n * + m}$$

Paso 2: Una vez estimadas las probabilidades a posteriori de los rasgos, podemos calcular la probabilidad de cada valor de la clase para un nuevo ejemplo[6]:

$$P(y/\mathbf{x}) = \frac{P(\mathrm{y})P(\mathbf{x}/\mathrm{y})}{\sum_v P(\mathrm{v})P(\mathbf{x}/v)}$$

Y para poder realizar este cálculo, realizamos la hipótesis *ingenua (naive)*, es decir, suponemos independencia condicional de la clase dados los rasgos, lo que nos permite calcular P(x/y) mediante la expresión:

$$P(\boldsymbol{x}/y) = \prod_i^s P(x_d/y)$$

Una vez calculadas las probabilidades de cada una de las posibles clases, asignaríamos el nuevo ejemplo a la clase más probable.

Ejemplo 3. 4. *Clasificación de mensajes de SPAM*

Supongamos que tenemos una base de datos con 5 mensajes, que están etiquetados como spam o como no-spam (ham):

M1	send us your password	spam
M2	send us your review	ham
M3	review your password	ham
M4	review us	spam
M5	send your password	spam
M6	send us your account	spam

Recibimos un nuevo e-mail, con el texto "review your account", y queremos clasificarlo como spam o no spam. Veamos cómo se haría.

El vocabulario de estos mensajes está formado por seis palabras, por lo que vamos a utilizar como rasgos seis variables aleatorias que representan si cada una de estas palabras aparece (1) o no aparece (0). Por tanto, los mensajes se podrían codificar como:

[5] Normalmente, si la variable X tiene r valores posibles, se suele utilizar $m=r$ y $p=1/r$

[6] Nótese que, como en otras ocasiones, podemos considerar el denominador de esta fracción como una constante de normalización, y, como en este contexto lo que nos interesa es asignar el ejemplo a la clase que tiene mayor probabilidad, podemos ignorar esta constante.

	P (password)	R (review)	(S) send	U (us)	Y (your)	A (account)
M1 (spam)	1	0	1	1	1	0
M2 (ham)	0	1	1	1	1	0
M3 (ham)	1	1	0	0	1	0
M4 (spam)	0	1	0	1	0	0
M5 (spam)	1	0	1	0	1	0
M6 (spam)	0	0	1	1	1	1
Mensaje nuevo (?)	0	1	0	0	1	1

Queremos utilizar el modelo Naive-Bayes para construir un clasificador. Para ello calculamos las probabilidades a priori de la clase P(y), y tenemos que: P(spam)=2/3; P(ham)=1/3.

Calculamos ahora el estimador $P(x_d/y)$, que será:

	spam	ham
a priori	**2/3**	**1/3**
password=1	2/4	1/2
review=1	**1/4**	**2/2**
send=1	3/4	1/2
us=1	3/4	1/2
your=1	**3/4**	**2/2**
account=1	**1/4**	**0/2**

	spam	ham
-	-	-
password=0	**2/4**	**1/2**
review=0	3/4	0/2
send=0	**1/4**	**1/2**
us=0	**1/4**	**1/2**
your=0	1/4	0/2
account=0	3/4	0/2

Para clasificar el nuevo mensaje, tenemos que calcular la probabilidad de que un mensaje nuevo que contenga esas palabras sea spam y la probabilidad de que no lo sea, es decir, P(spam/mensaje nuevo) y P(ham/mensaje nuevo). En la tabla, aparecen en azul los valores que hemos de multiplicar en cada columna (los correspondientes al mensaje nuevo). De modo que tendríamos que:

P(spam/mensaje nuevo) = α∗P(spam)*P((P=0,R=1,S=0,U=0,Y=1,A=1)/spam) =

$$= \alpha * \frac{2}{3} * \frac{2}{4} * \frac{1}{4} * \frac{1}{4} * \frac{1}{4} * \frac{3}{4} * \frac{1}{4} = \alpha * \frac{1}{1024}$$

P(ham/mensaje nuevo) = α∗P(ham)*P((P=0,R=1,S=0,U=0,Y=1,A=1)/ham)=

$$= \alpha * \frac{1}{3} * \frac{1}{2} * \frac{2}{2} * \frac{1}{2} * \frac{1}{2} * \frac{2}{2} * \frac{0}{2} = 0$$

Normalizando el resultado, tendríamos que P(spam/review, your, account)=1 y P(ham/review, your, account)=0, de forma que el mensaje sería clasificado como "spam" con probabilidad 1 (debido a que contiene una palabra que nunca antes había aparecido en un mensaje "ham").

Si aplicamos la corrección de Laplace con valores p=1/2 y m=2[7] las probabilidades estimadas serían:

[7] *r*=2 puesto que cada una de las variables rasgos son binarias, toman valores "aparece" y "no aparece"

	spam	ham
password=1	(2+1)/(4+2)	(1+1)/(2+2)
review=1	(1+1)/(4+2)	(2+1)/(2+2)
send=1	(3+1)/(4+2)	(1+1)/(2+2)
us=1	(3+1)/(4+2)	(1+1)/(2+2)
your=1	(3+1)/(4+2)	(2+1)/(2+2)
account=1	(1+1)/(4+2)	(0+1)/(2+2)

Es decir;

	spam	ham
a priori	**2/3**	**1/3**
password=1	1/2	1/2
review=1	**1/3**	**3/4**
send=1	2/3	1/2
us=1	2/3	1/2
your=1	**2/3**	**3/4**
account=1	**1/3**	**1/4**

	spam	ham
-	-	-
password=0	**1/2**	**1/2**
review=0	2/3	1/4
send=0	**1/3**	**1/2**
us=0	**1/3**	**1/2**
your=0	1/3	1/4
account=0	2/3	3/4

P=0,R=1,S=0,U=0,Y=1,A=1

Calculamos ahora las probabilidades:

$$P(\text{spam}/\text{mensaje nuevo}) = \alpha * \frac{2}{3} * \frac{1}{2} * \frac{1}{3} * \frac{1}{3} * \frac{1}{3} * \frac{2}{3} * \frac{1}{3} = \alpha * (2/729) = 0{,}0027\alpha$$

$$P(\text{ham}/\text{mensaje nuevo}) = \alpha * \frac{1}{3} * \frac{1}{2} * \frac{3}{4} * \frac{1}{2} * \frac{1}{2} * \frac{3}{4} * \frac{1}{4} = \alpha * (3/512) = 0{,}0058\ \alpha$$

En este caso vemos que la probabilidad mayor corresponde a que el mensaje no sea spam.

Para ver un video con otro ejemplo resuelto (en el contexto de clasificación de textos, y, explicado por el profesor de la Universidad de Stanford Francisco Lacobelli), véase https://www.youtube.com/watch?v=EGKeC2S44Rs

Ejercicio 3.5. Clasificación de comentarios de películas[8]

Consideremos el siguiente conjunto de datos relativos a opiniones de usuarios sobre películas:

Just plain boring	+
Entirely predictable and lacks energy	-
No surprises and very few laughs	+
Very powerful	-
The most fun film of the summer	+

Utilizando un modelo Naive-Bayes con la corrección de Laplace con p= ½ y m=2, clasificar como positivo o negativo el comentario "predictable with no fun"

3.2.2 Validación de modelos

Una vez que se obtiene un modelo para un determinado problema, resulta útil poder evaluar su rendimiento, para saber si las inferencias que realicemos con el

[8] Tomado de https://web.stanford.edu/~jurafsky/slp3/6.pdf (es el ejemplo resuelto en la sección 6.3).

modelo son fiables y también para poder comparar unos modelos con otros y saber cuál predice mejor.

Para ello se usan técnicas de validación de modelos. La más simple consiste en evaluar la bondad de las predicciones de un modelo sobre un conjunto de datos de que contenga ejemplos sobre los cuales realizar predicciones, que permitan comparar los resultados de dichas predicciones con los valores reales que aparezcan en los ejemplos del conjunto de prueba, de modo que podamos analizar cuántas veces el modelo ha acertado el valor real, cuántas se ha equivocado, que tipo de errores ha cometido, etc. En esta sección vamos a definir diversas medidas que se utilizan para evaluar el rendimiento de un modelo.

En la práctica, si se dispone de un conjunto de datos suficientemente grande, se puede dividir en dos conjuntos: el conjunto de entrenamiento de los datos o *training set*, con los que aprenderemos el modelo, y el conjunto de pruebas o *test set,* con el que aplicaremos el modelo obtenido y obtendremos las medidas de rendimiento. Lo que ocurre es que en ocasiones el conjunto de datos del que se dispone no es suficientemente grande y es necesario emplear algunas técnicas para sacarle el máximo partido.

Un problema que ocurre frecuentemente a la hora de generar los modelos que describan nuestros datos es que a veces el modelo se *sobreajusta* a los datos. En esta sección hablaremos en primer lugar del problema del sobreajuste del modelo a los datos, para a continuación describir algunas de las técnicas de validación que se utilizan para paliar dicho problema. Finalizaremos con la presentación de algunas de las medidas de rendimiento más habituales.

3.2.2.1 Sobreajuste (overfitting) e infrajuste (underfitting)

El problema del *sobreajuste* ocurre cuando el modelo se ajusta en exceso al conjunto de datos que se ha utilizado para entrenarlo. Un modelo sobre ajustado es un modelo estadístico que contiene más parámetros que los que justifican los datos. En esencia, el problema se produce porque en el modelo se ha incluido información de variaciones residuales, tales como valores atípicos (*outliers*, ruido, etc).

Por el contrario, el problema del *infrajuste* ocurre cuando un modelo estadístico no es capaz de capturar la estructura subyacente en un conjunto de datos. Un modelo infrajustado es modelo en el que no se encuentran algunos parámetros o términos que deberían aparecer en un modelo correcto. El problema se presentaría por ejemplo en el caso de tener datos no lineales e intentar ajustarlos con un modelo lineal.

En la siguiente figura se ilustran problemas de ambos tipos, junto con un modelo correcto:

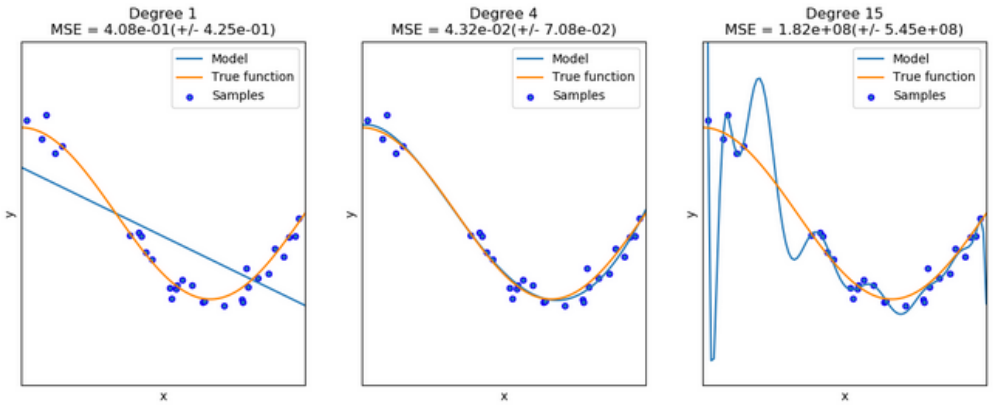


Figura 3. 6. Ejemplos de infrajuste y sobreajuste[9]

A la izquierda vemos que cuando utilizamos una recta para modelar estos datos (regresión lineal) el modelo está infrajustado. A la derecha utilizamos un polinomio de grado 15, que se ajusta casi perfectamente a los datos, pero que como se ve en la figura no recoge correctamente la estructura de los datos. En el centro aparece un polinomio de grado cuatro, que se ajusta bastante bien al conjunto de datos y a su estructura.

3.2.2.2 Métodos de validación

Para validar un clasificador y obtener medidas de su eficacia, se realiza un proceso de evaluación del mismo. en el que el conjunto de ejemplos disponibles se suele dividir en tres partes:

- ***El conjunto de entrenamiento*** (*training set*), que es el que se proporciona a los algoritmos de aprendizaje para obtener los estimadores h_i.
- ***El conjunto de validación*** *(validation set),* que se utiliza para elegir el mejor estimador entre los obtenidos. En el ejemplo presentado en la figura, para evaluar cuál de los tres modelos es el mejor, calcularíamos el error cuadrático medio (Mean Squared Error, MSE) sobre el conjunto de datos que hayamos conservado para hacer la validación, en la figura se observa que el menor error cuadrático medio es el del polinomio de grado 4.
- ***El conjunto de pruebas*** *(test set),* que se utiliza para evaluar el rendimiento del estimador elegido, mediante una serie de medidas que explicaremos en la siguiente sección.

Hay diversas técnicas para construir estos conjuntos:

- *Método de retención (hold-out method)*: el conjunto de ejemplos se subdivide aleatoriamente en dos conjuntos: un conjunto de entrenamiento d_0 que permite aprender el modelo, y un conjunto de pruebas d_1 (normalmente de menor tamaño) que permite evaluar el rendimiento del modelo. La siguiente figura ilustra el funcionamiento de este método.

[9] Tomado de http://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html

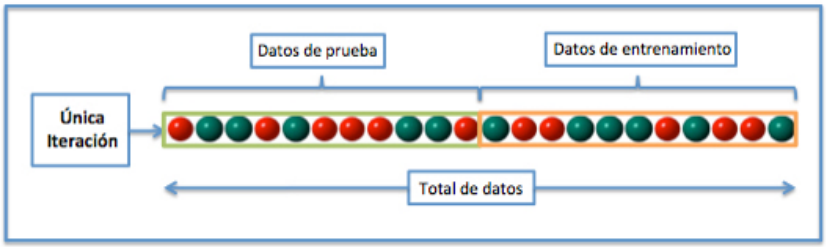


Figura 3.7. Método de retención[10]

El problema de este método es que la evaluación puede depender en gran medida de cómo es la división entre los datos de entrenamiento y los datos de prueba, y por lo tanto puede variar mucho en función de cómo se realice esta división. Para evitar este problema, surgen los métodos de validación cruzada.

- *Validación cruzada de k iteraciones (k-fold cross validation)*: se divide el conjunto de datos en *k* subconjuntos del mismo tamaño y se realizan *k* tandas de aprendizaje, de modo que cada subconjunto sirva una vez de conjunto de entrenamiento y *k-1* como conjunto de pruebas. Se elige el mejor clasificador según el rendimiento en la validación, y se calcula su rendimiento sobre los conjuntos de prueba. El esquema que representa la división del conjunto en cada iteración sería:

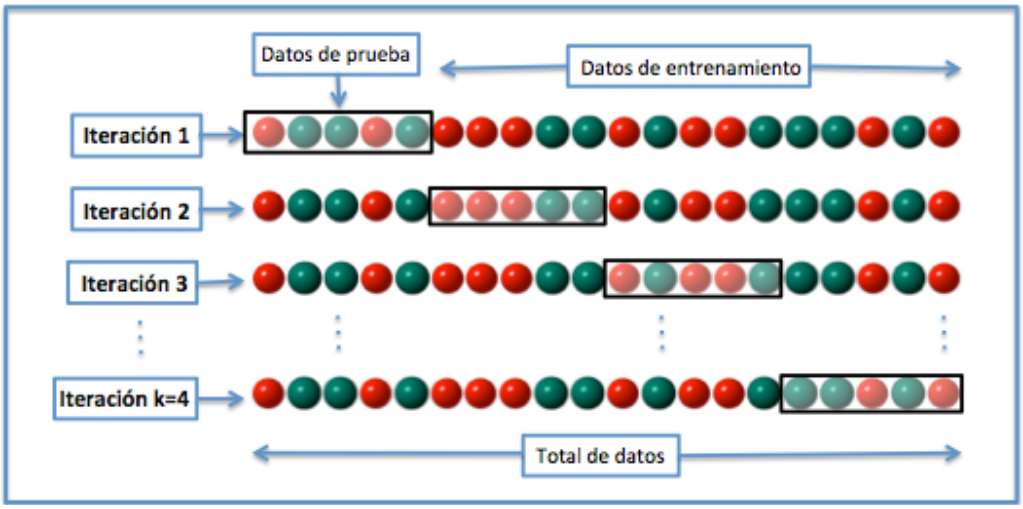


Figura 3.7. División del conjunto de datos en cada iteración (4-fold cross validation)

El funcionamiento del algoritmo se describe en la siguiente figura:

[10] Atribución de las figuras relativas a métodos para construir los conjuntos: Joan.domenech91 (Own work) [CC BY-SA 3.0 (https://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

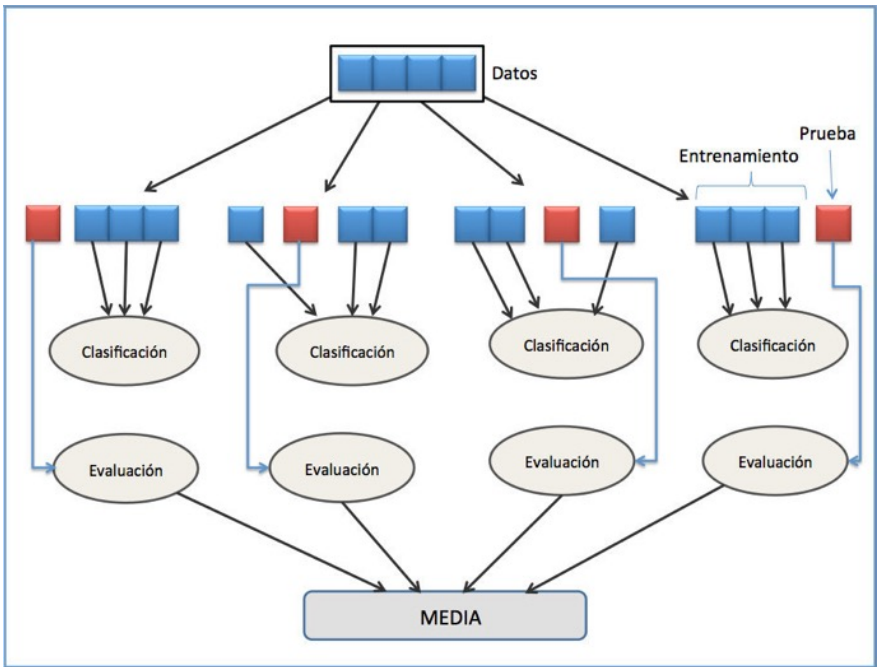


Figura 3.8. Esquema de validación cruzada con 4 iteraciones (4-fold cross validation)

Así, si por ejemplo estamos construyendo un modelo Naive Bayes, en cada iteración del algoritmo obtendríamos un clasificador diferente, cuyo rendimiento evaluaríamos con el conjunto de pruebas. En función de ese rendimiento nos quedaríamos con el clasificador mejor. El mismo proceso podríamos hacer, construyendo una red bayesiana. Podríamos después utilizar un conjunto de pruebas diferente para poder comparar el rendimiento del clasificador Naive Bayes con la red bayesiana.

En la práctica es habitual utilizar sólo los conjuntos de entrenamiento y validación, incluso en algunas referencias se habla del conjunto de validación o conjunto de pruebas. Pero si se dispone de los datos suficientes, se pueden considerar los tres conjuntos.

3.2.2.3 Medidas de rendimiento de un clasificador

Para evaluar el rendimiento del algoritmo, se utilizan diversas medidas, que definimos a continuación.

3.2.2.3.1 Matriz de confusión

Supongamos que tenemos un problema de clasificación en el que n clases posibles, C_1, ..., C_n , y para este problema tenemos un modelo clasificador. Aplicamos dicho modelo a un conjunto de datos de prueba. Definimos la matriz de confusión como una matriz tamaño nxn, en la que cada uno de sus elementos c_{ij} se define como:

c_{ij}= número de ejemplos de la clase j que se estimó que pertenecían a la clase i

De esta forma, en la diagonal de la matriz se representan los aciertos, mientras los otros elementos representan el número y tipo de los fallos cometidos en la clasificación.

3.2.2.3.2 Medidas numéricas

Para problemas de clasificación binaria, se definen asimismo varias medidas numéricas. La primera de ellas es la precisión (*accuracy*), que se define como el

número de ejemplos que se predicen correctamente, dividido por el número total de ejemplos. Esta medida toma un valor comprendido entre 0 y 1. A mayor valor de la medida, mejor es la clasificación (1 significaría clasificación perfecta).

En el caso de problemas de clasificación binaria, se definen también otro tipo de medias. Si llamamos clase 1 a la clase positiva, y clase 2 a la clase negativa, podemos definir:

- verdaderos positivos (VP) = c_{11}
- verdaderos negativos (VN) = c_{22}
- falsos positivos o errores de tipo 1 (FP) = c_{12}
- falsos negativos o errores de tipo 2 (FN) = c_{22}

Nótese que, en problemas de clasificación binaria, se tiene que la precisión puede calcularse mediante la expresión:

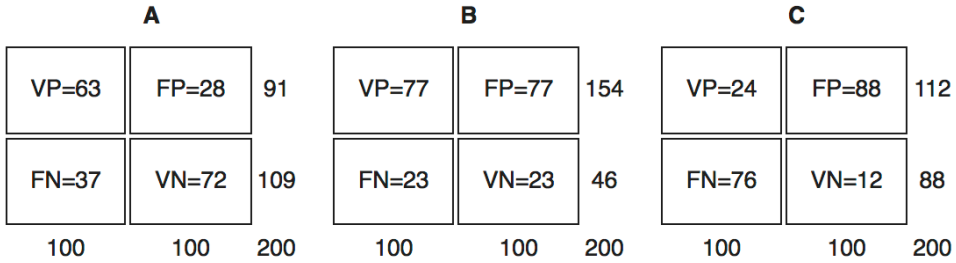
$$Accuracy = \frac{VP+VN}{VP+VN+FP+FN}$$

A partir de estos valores, pueden definirse también otras medidas, por ejemplo:

- ***Tasa o razón de verdaderos positivos***, también llamada sensibilidad, que se define como la proporción de positivos que se identifican correctamente como tales y se calcula mediante $\frac{VP}{VP+FN}$.
- ***Tasa o razón de falsos positivos***, que se define como la proporción de negativos que se identifican incorrectamente como positivos, y se calcula mediante $\frac{FP}{FP+VN}$.
- ***Tasa o razón de verdaderos negativos*** (también llamada ***especificidad***), que mide el porcentaje de negativos que se identifican correctamente como tales, es decir *especificidad* = $\frac{VN}{VN+FN}$.

Ejemplo 3. 5. *Obtención de medidas de rendimiento de clasificadores*[11]

Supongamos que, para un problema de clasificación binaria, se han generado tres modelos, A, B, C, que evaluados sobre el mismo conjunto de pruebas (formado por 200 ejemplos), han arrojado los resultados expresados en las siguientes matrices de confusión:



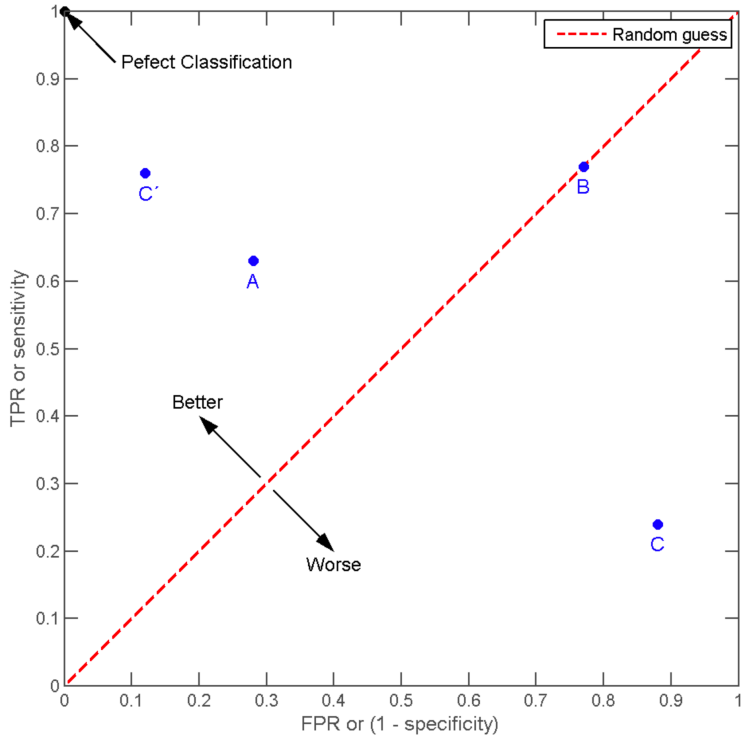
Si calculamos para cada uno de los clasificadores la precisión, sensibilidad, y especificidad, tenemos que:

	A	B	C
Tasa de Verdaderos positivos o (sensibilidad)	0.23	0.77	0.24
Tasa de Falsos positivos	0.28	0.77	0.88

[11] Ejemplo, datos y gráficos tomados de https://es.wikipedia.org/wiki/Curva_ROC

Tasa de Verdaderos negativos (especificidad)	0.62	0.23	0.12
Precisión	0.68	0.50	0.18

Para comparar cuál de estos modelos nos ha dado un mejor resultado dibujamos los puntos correspondientes a las tasas de falsos positivos (abcisa) frente a la tasa de verdaderos positivos (ordenada), como se muestra en el siguiente gráfico:



La clasificación perfecta se obtendría si la abcisa es 0 (ningún falso positivo) y la ordenada es 1 (todos los positivos son verdaderos). Contra más cercano al punto de coordenadas esté el punto representado, mejor es el clasificador. Vemos que A es claramente mejor que B y C. El punto B se sitúa sobre la diagonal, porque tiene tantos verdaderos positivos como falsos positivos, en realidad, si la mitad de las veces acierta y la mitad se equivoca, es tan buen método clasificador como lanzar una moneda al aire y clasificar como positivo si sale cara. El bajo rendimiento del método B se puede ver también en el valor de su precisión, que es un 50%. El peor método es el método C, con una alta tasa de falsos positivos y una tasa muy baja de verdaderos positivos, y una precisión de sólo el 18%. Pero a partir del método C, podríamos construir el mejor clasificador de todos, si simplemente clasificamos el ejemplo al contrario que lo indicado por el método C, obtendríamos el clasificador C'. De este modo a la hora de elegir un clasificador entre los tres, deberíamos fijarnos no sólo en su cercanía al clasificador ideal (0,1), sino en el que maximiza la distancia respecto a la diagonal. Si dicho punto está por encima de la diagonal, será el mejor clasificador, y si está por debajo, a partir de él podremos construir el mejor clasificador, simplemente invirtiendo el resultado que arroje.

3.2.2.3.3 La curva ROC

También es frecuente utilizar la curva ROC (*Receiving operating characteristic*). Para un problema de clasificación binaria, y un clasificador que proporciona una

probabilidad de pertenecer a la clase (como es el caso de los modelos bayesianos), a la hora de clasificar cada ejemplo hay varias alternativas

- Clasificar el ejemplo asignándole el valor de la clase y (1 o 0) con mayor probabilidad,
- Definir un valor umbral τ para asignar un ejemplo a la clase positiva. Es decir, La clase será y = 1 sí y solo sí $P(y=1/x_q) > \tau$

Así por ejemplo, si establecemos un umbral $\tau = 0.8$, todas las instancias cuya probabilidad de pertenecer a la clase y=1 sea mayor de 0.8 serán clasificadas como positivas, y el resto como negativas. Al aumentar de este modo el nivel de exigencia para clasificar a un ejemplo como positivo, obtendremos más verdaderos positivos, pero también más falsos positivos. Es decir, a mayor valor de τ, mayor será la tasa de verdaderos positivos y falsos positivos (más se acercará el punto al punto de coordenadas (1,1).

Si dibujamos todos los puntos obtenidos para diferentes valores del umbral, obtenemos la llamada curva ROC, que muestra la evolución de ambas tasas. En la siguiente figura vemos un ejemplo de curva ROC.

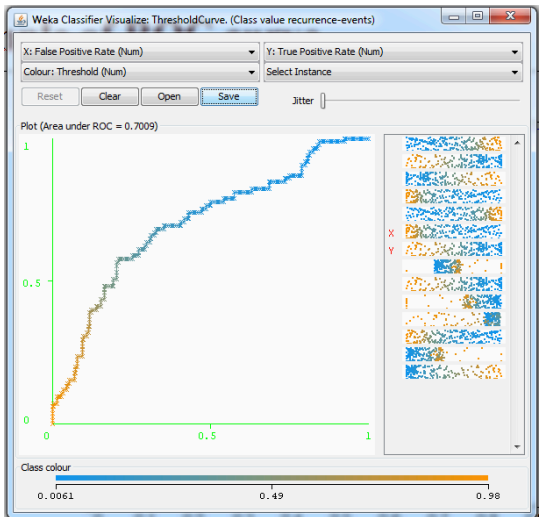


Figura 3.7 Ejemplo de curva ROC

Para comparar dos modelos utilizando una curva ROC, se suele usar el área bajo la curva ROC (AUC). En la siguiente figura mostramos diversas curvas ROC, con su valor AUC:

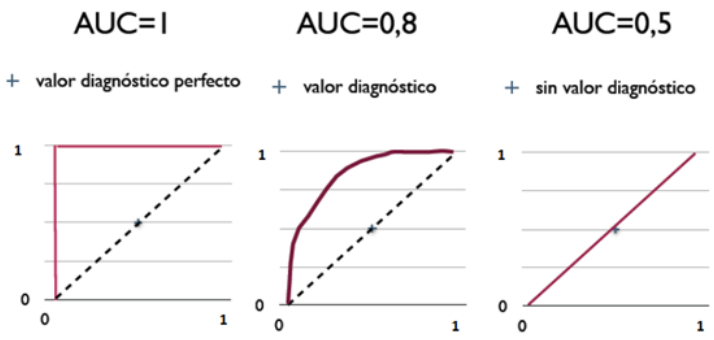


Figura 3.8 Ejemplos de curva ROC y sus áreas

Se debe elegir siempre la prueba que presente un mayor valor bajo la curva. En medicina, se suele considerar que una prueba con AUC mayor de 0,75 es buena, y si el AUC es mayor de 0,97 es una prueba excelente.

Práctica 6. Aprendizaje con GeNIe. El objetivo de esta práctica es aprender técnicas aprendizaje bayesiano y generación de casos.

3.3 Bibliografía

Jurafsky, D., & Martin, J. H. *Speech and language processing* (Vol. 3). London: Pearson, 2014.
Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.
Lozano, T & Kaelbling, L. Apuntes del curso "Techniques in Artificial Intelligence: MITT Open Courseware, Massachusetts Institute of Technology (https://bit.ly/2JRiCSB) 2002.

TEMA 4. INTRODUCCIÓN AL RAZONAMIENTO DIFUSO

Resultados de aprendizaje

Al finalizar este tema, los estudiantes deberán ser capaces de:

- Modelar una situación real con un sistema de razonamiento difuso (conjuntos, relaciones, reglas de inferencia).
- Conocer y aplicar las diferentes alternativas:
  - para definir y modificar conjuntos difusos,
  - para operaciones con conjuntos difusos,
  - para razonar con reglas difusas, y
  - para codificar y decodificar.
- Manejar herramientas de implementación de lógica difusa.

Contenidos

4.1 Introducción[1]

El modelo ideal del razonamiento (humano o mecánico) es el razonamiento absolutamente exacto, en el sentido de que tanto las reglas como los hechos que se manejan son ciertos y precisos. No obstante, en el mundo real se suele razonar con información que es incierta (no estamos completamente seguros de que sea verdadera) e imprecisa (no manejamos valores completamente definidos). Por ejemplo, si decimos Es casi seguro que Juan es hermano de Pedro estamos manejando un hecho incierto; por otra parte, si decimos Juan es bastante alto estamos manejando un hecho impreciso. Un sistema razonador capaz de emular las capacidades del sentido común debe ser capaz de realizar inferencias que manejen este tipo de información. A grandes rasgos, podemos clasificar las fuentes de imprecisión e incertidumbre en tres grupos:

- Deficiencias de la información: información incompleta o errónea
- Características del mundo real: información imprecisa o indeterminismo del mundo real

[1] Introducción tomada de los apuntes de razonamiento aproximado de Francisco Javier Díez

- Deficiencias del modelo: modelo incompleto o inexacto

Veamos algunos ejemplos de estas situaciones en el contexto de la medicina[2]:

- *Información incompleta*. En muchos casos, no toda la información relativa al caso está disponible. Por ejemplo, pensemos en un paciente que acude a una consulta médica. Es frecuente que no se disponga de su historia clínica completa, y que además el paciente sea incapaz de recordar cómo se ha desarrollado la enfermedad. En otros casos, las limitaciones prácticas impiden contar con toda la información relevante (pensemos por ejemplo en el caso de pruebas médicas de elevado coste material) pero es necesario tomar una decisión con la información que se posee, aunque ésta sea muy limitada.
- *Información errónea.* La información suministrada puede ser errónea: el paciente describe incorrectamente sus síntomas, intenta deliberadamente engañar al médico, el diagnóstico anterior (contenido en su historia clínica) es erróneo, las pruebas de laboratorio han dado falsos positivos o falsos negativos, etc.
- *Información imprecisa.* Hay datos que son difícilmente cuantificables. Buenos ejemplos son síntomas como el dolor o la fatiga.
- *Mundo real no determinista.* A diferencia de las máquinas mecánicas o eléctricas, cuyo funcionamiento se rige por leyes deterministas, los profesionales de la medicina comprueban a diario que cada ser humano es un mundo, en que las leyes generales no siempre resultan aplicables. Muchas veces las mismas causas producen efectos diferentes en distintas personas, sin que haya ninguna explicación aparente. Por ello, el diagnóstico médico debe estar siempre abierto a admitir la aleatoriedad y las excepciones.
- *Modelo incompleto.* Por un lado, hay muchos fenómenos médicos cuya causa aún se desconoce. Por otro, es frecuente la falta de acuerdo entre los expertos de un mismo campo. Finalmente, aunque toda esta información estuviera disponible, sería imposible, por motivos prácticos, incluirla en un sistema basado en el conocimiento.
- *Modelo inexacto.* Por último, todo modelo que trate de cuantificar la incertidumbre, por cualquiera de los métodos que existen, necesita incluir un elevado número de parámetros: por ejemplo, en el caso de las redes bayesianas, necesitamos especificar todas las probabilidades a priori y condicionales. Sin embargo, una gran parte de esta información no suele estar disponible, por lo que debe ser estimada de forma subjetiva. Es deseable, por tanto, que el método de razonamiento empleado pueda tener en cuenta las inexactitudes del modelo.

Para modelar la incertidumbre se dispone de las redes bayesianas, para modelar la imprecisión se utiliza la lógica difusa. La lógica difusa[3], como su nombre indica, es una lógica alternativa a la lógica clásica que pretende introducir un grado de vaguedad en las cosas que califica. En el mundo real existe mucho conocimiento no-perfecto, es decir, conocimiento vago, impreciso, incierto, ambiguo, inexacto, o probabilístico por naturaleza. El razonamiento y

[2] Francisco Javier Díez. Apuntes de razonamiento aproximado. http://www.ia.uned.es/fjdiez/libros/razaprox.html, 2009.

[3] A la hora de traducir el término inglés fuzzy, hay principalmente dos alternativas: borroso y difuso. Aunque en alguna bibliografía se habla aún de Lógica Borrosa o Teoría de los conjuntos borrosos, se utiliza más el término difuso.

pensamiento humano frecuentemente conlleva información de este tipo, probablemente originada de la inexactitud inherente de los conceptos humanos y del razonamiento basado en experiencias similares, pero no idénticas a experiencias anteriores.

El problema principal surge de la poca capacidad de expresión de la lógica clásica. Supongamos por ejemplo que tenemos un conjunto de personas que intentamos agrupar según su altura, clasificándolas en *altas* o *bajas*. La solución que presenta la lógica clásica es definir un umbral de pertenencia (por ejemplo, un valor que todo el mundo considera que, de ser alcanzado o superado, la persona en cuestión puede llamarse *alta*). Si dicho umbral es 1.80, todas las personas que midan 1.80 o más serán *altas*, mientras que las otras serán *bajas*. Según esta manera de pensar, alguien que mida 1.79 será tratado igual que otro que mida 1.50, ya que ambos han merecido el calificativo de *bajas*. Sin embargo, si dispusiéramos de una herramienta para caracterizar las alturas de forma que las transiciones fueran suaves, estaríamos reproduciendo la realidad mucho más fielmente.

Los problemas asociados al razonamiento con conceptos vagos o difusos son puestos de manifiesto por la antigua paradoja del calvo, que traducida a términos económicos podríamos expresar así: una persona que sólo tiene dos céntimos de euro es sumamente pobre, indudablemente; ahora bien, si a una persona que es sumamente pobre le damos un céntimo de euro, sigue siendo sumamente pobre; aplicando esta regla repetidamente, llegamos a la conclusión de que una persona que tiene mil millones de euros es sumamente pobre. La solución a esta paradoja es que el concepto de "pobre" o "sumamente pobre" no tiene un límite completamente definido, sino que a medida que le damos a esa persona un céntimo de euro tras otro, hasta llegar a los mil millones de euros (en el supuesto de que tuviéramos esa cantidad de dinero), el grado de pobreza va disminuyendo paulatinamente: no hay un único céntimo de euro que le haga pasar de ser pobre a ser rico.

Asimismo, no hay un valor cuantitativo que defina el término joven. Para algunas personas, 25 años es joven, mientras que para otros, 35 es joven. Incluso el concepto puede ser relativo al contexto. Un presidente de gobierno o de 35 años es joven, mientras que un futbolista no lo es. Hay sin embargo cosas que están claras: una persona de 1 año es joven, mientras que una de 100 años no lo es. Pero una persona de 35 años tiene algunas posibilidades de ser joven (que normalmente dependen del contexto).

Para representar este hecho, definiremos el conjunto difuso *joven* de modo que cada uno de sus elementos pertenezca a él con cierto grado (posibilidad). De un modo más formal, un conjunto difuso A se caracteriza por una función de pertenencia:

$$\mu_A : U \rightarrow [0,1]$$

que asocia a cada elemento x de U un número $\mu_A(x)$ del intervalo [0,1], que representa el grado de pertenencia de x al conjunto difuso A. A U se le llama *universo de discurso*. Por ejemplo, el término difuso *joven* puede definirse mediante el conjunto difuso siguiente:

Edad	Grado de Pertenencia
≤25	1.0
30	0.8
35	0.6
40	0.4

45	0.2
≥50	0

Es decir, la función de pertenencia del conjunto difuso *joven* viene dada por:

$\mu_A(x) = 1$ si $x \leq 25$, $\mu_A(30) = 0.8$, ..., $\mu_A(x) = 0$ si $x \geq 50$.

Que podemos representar en la siguiente gráfica:

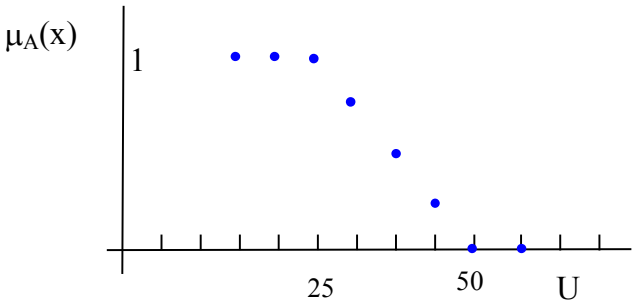


Figura 2.1. Función de pertenencia del conjunto difuso joven

Si el universo de discurso es continuo, tendremos funciones de pertenencia continuas:

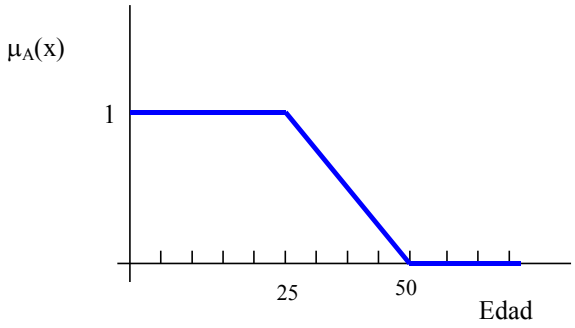


Figura 2.2. Función de pertenencia de joven si U es continuo

En general, si una función de pertenencia se da especificando los valores correspondientes a un conjunto discreto de elementos del universo de discurso, el valor asociado al resto de los elementos se obtiene por interpolación (utilizando la ecuación de la recta que une los dos puntos).

El origen del interés actual por la teoría de conjuntos difusos se debe a un artículo publicado por Lofti Zadeh en 1.965. En la actualidad es un campo de investigación muy importante, tanto por sus implicaciones matemáticas o teóricas como por sus aplicaciones prácticas. Prueba de esta importancia es el gran número de revistas internacionales (Fuzzy Sets and Systems, IEEE Transactions on Fuzzy Systems..) congresos (FUZZ-IEEE, IPMU, EUSFLAT, ESTYLF...) y libros (Klir & Yuan, 1995) (Ross, 1995), (Kruse, 1994), (McNeill, 1994), (Mohammd, 1993), (Pedrycz, 1998) dedicados al tema.

¿En qué situaciones es útil aplicar la lógica difusa?

- En procesos complejos, si no existe un modelo de solución sencillo.
- Cuando haya que introducir la experiencia de un operador "experto" que se base en conceptos imprecisos.
- Cuando ciertas partes del sistema a controlar son desconocidas y no pueden

medirse de forma fiable (con errores posibles).
- Cuando el ajuste de una variable puede producir el desajuste de otras.
- En general, cuando se quieran representar y operar con conceptos que tengan imprecisión o incertidumbre.

Por esta razón, son frecuentes las aplicaciones de la lógica difusa a:

- *Control de sistemas*: Control de tráfico, control de vehículos (helicópteros...), control de compuertas en plantas hidroeléctricas, centrales térmicas, control en máquinas lavadoras, control de metros (mejora de su conducción, precisión en las paradas y ahorro de energía), ascensores...
- *Predicción y optimización*: Predicción de terremotos, optimizar horarios...
- *Reconocimiento de patrones y Visión por ordenador*: Seguimiento de objetos con cámara, reconocimiento de escritura manuscrita, reconocimiento de objetos, compensación de vibraciones en la cámara, sistemas de enfoque automático...
- *Sistemas de información o conocimiento*: Bases de datos, sistemas expertos...

4.2 Teoría de conjuntos difusos

4.2.1 Teoría de conjuntos clásica (conjuntos nítidos)

Los conjuntos "clásicos", los conjuntos que estamos acostumbrados a manejar en la matemática, se suelen denominar "conjuntos nítidos" (en inglés, *crisp sets*) en la terminología de lógica difusa. Los conjuntos surgen de forma natural por la necesidad del ser humano de clasificar objetos y formar conceptos. Por ejemplo, si pensamos en los productos de alimentación, podemos considerar varios conjuntos: el conjunto de las frutas, el conjunto de las verduras, el de las carnes, el de los pescados, etc.

Un conjunto nítido A puede definirse de varias formas:

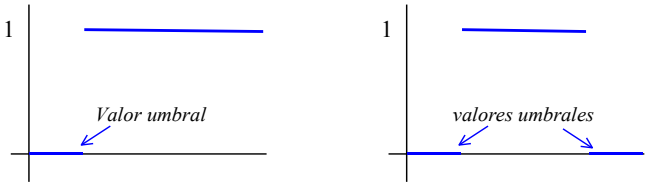
- A veces damos una lista de todos los elementos de A. Así por ejemplo, podemos definir el conjunto frutas como frutas = {manzana, pera, plátano, naranja} (definición extensional del conjunto).
- En otras ocasiones damos una característica que define los elementos de A (definición *intensional* del conjunto). Esto se puede hacer dando directamente la definición (por ejemplo, Fruto = Producto del desarrollo del ovario de una flor después de la fecundación); o bien delimitando un subconjunto dentro de un conjunto ya definido (por ejemplo, Frutas = Fruto comestible).
- Alternativamente, podríamos definir previamente un universo de discurso U formado por todos los objetos que podríamos tener en cuenta, pertenezcan o no a nuestro conjunto. En el ejemplo de las frutas, U podría ser el conjunto de todos los alimentos. A continuación, definimos una función $\mu_A : U \rightarrow \{0, 1\}$, dada para todo $x \in U$ mediante:

Conjuntos nítidos

$$\mu_A(x) = \begin{cases} 0 \; si \; x \notin A \\ 1 \; si \; x \in A \end{cases}$$

En nuestro ejemplo, el conjunto nítido `Frutas` se definiría como un conjunto A tal que μ_A(manzana)=1, μ_A(sardina)=0, etc.

Gráficamente, si alineamos los elementos de U a lo largo de una recta, μ_A sería una función tipo escalón centrada en el valor/valores umbral/umbrales de decisión.



Si se utilizan funciones de pertenencia, la forma de representar el vacío y el conjunto universal será:

- El vacío ∅ es un conjunto tal que, para todo x de U, μ(x)=0
- El conjunto universal Ω es un conjunto tal que, para todo x de U, μ(x)=1

4.2.2 Conjuntos difusos y variables lingüísticas

Conjuntos difusos, variables y valores lingüísticos

En los conjuntos difusos relajamos la restricción de que la función de pertenencia valga ó 0 ó 1, y dejamos que tome valores en el intervalo [0,1]. La necesidad de trabajar con conjuntos difusos, como ya hemos mencionado, surge del hecho que de que hay conceptos que no tienen límites claros. Por ejemplo: ¿una persona que mide 1.80 es alta? ¿Una temperatura de 15 grados es baja? Vemos que, a diferencia de lo que ocurre en el caso de las frutas (no hay vaguedad, un alimento o bien es una fruta o bien no lo es), en otras situaciones nos vemos obligados a tratar con ella. Y para representar la vaguedad del conjunto F correspondiente a un concepto permitimos que el "grado de pertenencia" a F de un elemento tome valores intermedios entre 0 (no pertenece en absoluto) y 1 (pertenece por completo).

Veamos algunas definiciones útiles:

- Llamaremos ***variable lingüística*** a aquella noción o concepto que vamos a calificar de forma difusa. Por ejemplo: la altura, la edad, el error, la variación del error... Le aplicamos el adjetivo "lingüística" porque definiremos sus características mediante el lenguaje hablado.
- Llamaremos ***universo de discurso*** U al rango de valores que pueden tomar los elementos que poseen la propiedad expresada por la variable lingüística. En el caso de las variables lingüísticas, el universo de discurso vendrá dado por el rango de posibles valores que puede tomar el atributo denotado por la variable lingüísticas. Siempre consideraremos que este rango es un intervalo cerrado de la recta real. Por ejemplo, en el caso de la variable lingüística "altura", si estamos considerando personas adultas normales, y medimos en metros, el universo de discurso podría ser el intervalo [1.4, 2.3].
- Llamamos ***valor lingüístico*** a una expresión en lenguaje natural que denota un valor que puede tomar una variable lingüística. Un repertorio de valores lingüísticos para una misma variable expresa una clasificación o división difusa del universo de discurso y por tanto del conjunto de objetos a los que se puede atribuir la variable lingüística. En el caso de la altura, podríamos por ejemplo considerar los valores lingüísticos bajo, mediano y alto, clasificando así los posibles valores de la altura y por tanto clasificando también a las personas. Cada valor lingüístico vendrá´ definido por un conjunto difuso, es decir, por una función de pertenencia.
- Llamaremos ***conjunto difuso*** a un valor lingüístico junto a una función de pertenencia. El valor lingüístico es el "nombre" del conjunto, y la función de pertenencia se define como aquella aplicación que asocia a cada elemento del universo de discurso el grado con que pertenece al conjunto difuso. Decimos que un conjunto es *nítido* si su función de pertenencia toma valores en {0,1}, y *difuso* si toma valores en [0,1].

- Dado un conjunto difuso A, se define como ***alfa-corte*** de A, al conjunto de elementos que pertenecen al conjunto difuso A con grado mayor o igual que alfa, es decir:

$$A_\alpha = \{ x \in U \,/\, \mu_A(x) \geq \alpha \}$$

- Se define como ***alfa corte estricto*** al conjunto de elementos con grado de pertenencia estrictamente mayor que alfa, es decir:

$$A_{\bar{\alpha}} = \{ x \in U \,/\, \mu_A(x) > \alpha \}$$

- Se define como ***soporte*** de un conjunto difuso A, al conjunto nítido de elementos que tienen grado de pertenencia estrictamente mayor que 0, o sea, al alfa-corte estricto de nivel 0.

$$\text{Soporte}(A) = \{ x \in U \,/\, \mu_A(x) > 0 \}$$

Se define como **núcleo** de un conjunto difuso A, al conjunto nítido de elementos que tienen grado de pertenencia 1. (alfa-corte de nivel 1)

$$\text{Núcleo}(A) = \{ x \in U \,/\, \mu_A(x) = 1 \}$$

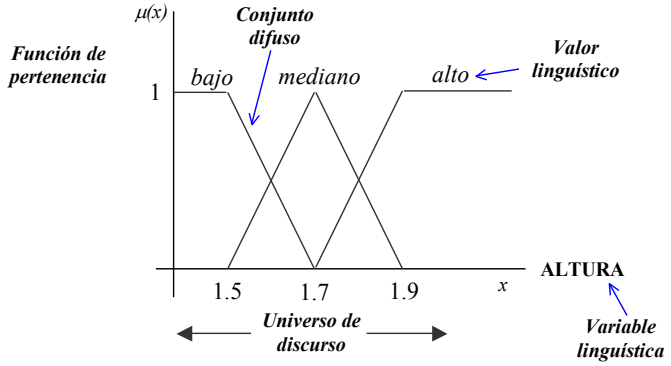
- Se define la ***altura*** de un conjunto difuso A como el valor más grande de su función de pertenencia.
- Se dice que un conjunto difuso está ***normalizado*** si y solo si su núcleo contiene algún elemento (o alternativamente, si su altura es 1), es decir:

$$\exists\, x \in U \qquad \mu_A(x) = 1$$

- El elemento x de U para el cual $\mu_A(x) = 0.5$ se llama el **punto de cruce**.
- Un conjunto difuso cuyo soporte es un único punto x de U y tal que la función de pertenencia de x es 1 (es decir, el soporte coincide con el núcleo y tienen un único punto) se llama un **conjunto difuso unitario** (singleton).
- Dados dos conjuntos difusos A y B, se dice que A está incluido en B ($A \subseteq B$) si para todo x de U se tiene que $\mu_A(x) \leq \mu_B(x)$. La inclusión es estricta si la desigualdad es estricta.

Ejemplo 4.1. *Aplicando las definiciones básicas en conjuntos difusos*

Consideremos la variable lingüística "*Altura de los seres humanos*", que toma valores en el universo de discurso U = [1.4, 2.50]. Vamos a hacer una clasificación difusa de los seres humanos en tres conjuntos difusos (o valores lingüísticos): *bajos*, *medianos* y *altos*.



En esta ilustración hemos dibujado 3 conjuntos difusos sobre la variable lingüística altura, cuyos valores lingüísticos asociados son bajo, mediano y alto respectivamente. Las funciones de pertenencia son de tipo L para bajo, Lambda o

Triángulo para el mediano y Gamma para el alto. Más adelante aclararemos el motivo por el cual usamos estos nombres (que únicamente determinan qué forma tendrán las funciones de pertenencia). De este modo si Luis mide 1.80 metros, la lógica difusa nos dice que es un 0.2 mediano y un 0.8 alto. De este modo expresamos que mientras un elemento puede estar dentro de un determinado conjunto, puede no cumplir las especificaciones de dicho conjunto al cien por cien (por ejemplo, en el caso de Luis, a la vista del resultado podríamos afirmar que es poco mediano y más bien alto.

En este ejemplo, dado el conjunto difuso mediano tenemos que:

- El alfa-corte 0.5 es el intervalo [1.6,1.8]
- El alfa corte estricto 0.5 es el intervalo (1.6, 1.8)
- El soporte es (1.5, 1.9)
- El núcleo es 1.7
- Es un conjunto difuso normalizado
- Tiene dos puntos de cruce: 1.6 y 1.8

Ejemplo 4.2. Definiciones básicas en conjuntos difusos discretos

Consideremos los posibles resultados de la tirada de un dado:

U = {1, 2, 3, 4, 5, 6}

Definamos un conjunto nítido sobre U. Por ejemplo, podríamos definir un conjunto enumerando sus elementos A ={4, 5, 6}, o describirlo por una propiedad ("valores de una tirada mayores que 3"). Definamos ahora un conjunto difuso A. Para cada $x \in U$, deberemos dar su grado de pertenencia a A. Por ejemplo, podemos escribir una tabla como esta:

Edad	Grado de Pertenencia
1	0
2	0
3	0.3
4	0.6
5	0.9
6	1

¿Podríamos describir este conjunto por una propiedad? Quizá´ diríamos que A está constituido por las "tiradas altas" del dado, concepto difuso que aparecería al "difuminar" o "emborronar" el concepto nítido anterior ("ser mayor que 3").

Apliquemos las definiciones anteriores al conjunto A:

- El soporte de A es {3, 4, 5, 6}.
- El núcleo de A es {6}.
- Su alfa-corte estricto a nivel 0.6 es {5, 6}.
- Su alfa-corte a nivel 0.6 es {4, 5, 6}.
- Su altura es 1, y por ello es un conjunto difuso normalizado.

▹

La notación habitual para los conjuntos difusos es la definida por Lofti Zadeh, que es la siguiente: sea A un conjunto difuso definido sobre el universo U:

$$A=\{(x, \mu_A(x)) / x \in U\}$$

que indica que A está formado por todos los pares ordenados x y el resultado de la función de pertenencia para todo elemento u dentro del universo de discurso U. Para denotar el conjunto difuso A:

- si el universo es discreto: $\sum_U \mu_A(x)/x$
- si el universo es continuo: $F = \int_u \mu_A(x)/x$

¡Cuidado con esta notación! El sumatorio o la integral pierden su significado habitual, En lógica difusa quieren simbolizar una mera enumeración de tuplas. La barra tampoco indica una fracción, sino que simplemente separa los dos elementos de la tupla. Así por ejemplo el conjunto difuso discreto "Tirada alta del dado" podría definirse como:

$$F = \{ 0/1 + 0/2 + 0.3/3 + 0.6/4 + 0.9/5 + 1/6\}$$

La parte derecha de la tupla indica el elemento y la parte izquierda el grado de pertenencia.

Los conjuntos difusos y las funciones de pertenencia pueden emplearse de dos formas posibles:

a) Para estimar grados de pertenencia a un conjunto. Por ejemplo, si nos dicen que una persona mide 170 cm, ¿en qué grado es una persona alta?

b) Para expresar *posibilidades* en una situación en la que se dispone de información incompleta. Por ejemplo, si nos dicen que una persona es mediana, ¿cuál será su altura? En este caso la función de pertenencia m puede interpretarse como una *distribución de posibilidad* que nos indica la preferencia sobre los valores que una variable de valor desconocido puede tomar.

De este modo vemos que la principal diferencia entre la teoría de conjuntos clásica y la difusa es que mientras que los valores de la función de pertenencia de un conjunto nítido son siempre 0 o 1, la función de pertenencia de un conjunto difuso toma valores en todo el intervalo [0,1][4]. Además, al contrario de los conjuntos nítidos, que pueden definirse de varias formas, los conjuntos difusos vienen siempre definidos por su función de pertenencia. Veamos qué tipos de funciones de pertenencia se usan más habitualmente en la lógica difusa.

Ejercicio 4.1. Sea U = [0, 24] $\subset$ R. Consideremos un día de verano. Dar definiciones razonables de los siguientes conjuntos difusos: la mañana; la tarde; la noche; la madrugada; sobre las 2 de la tarde.

Ejercicio 4.2. Para cada una de las siguientes funciones, indicar razonadamente si podría ser la función de pertenencia de un conjunto difuso sobre R y, en caso afirmativo, dar una descripción verbal intuitiva de dicho conjunto:

a) la función sen(x).
b) la función |sen(x)|.
c) la función escalón unitario u(x), u(x) = 0 si x < 0, en otro caso u(x) = 1.
d) la función impulso unitario δ(x), u(x) = ∞ si x = 0, en otro caso u(x) = 0.
e) f(x), si f(x) es una función de densidad de probabilidad.
f) F (x), si F (x) es una función de distribución de probabilidad.

[4] Se suele normalizar el grado de pertenencia máximo a 1.

Ejercicio 4.3. Sea U = {a, b, c, d, e} y sea C un conjunto difuso sobre definido sobre U, mediante C = 0,3/a + 0,7/b + 0,1/e.

g) Calcular el núcleo, el soporte y la altura de C.
h) ¿Cuál es el mayor α tal que Cα (no estricto) = soporte(C)?
i) ¿Cuál es el mayor α tal que Cα (no estricto) = núcleo(C)?

Ejercicio 4.4. Sea U = {1, 2, 3, 4, 5, 6} y sea C un conjunto difuso sobre U, del que conocemos los siguientes α-cortes no estrictos:

Ejercicio 4.5. $C\square = \begin{cases} \{1,3,4,6\} \text{ si } 0 < \alpha \leq 0{,}3 \\ \{1,3,6\} \text{ si } 0{,}3 < \alpha \leq 0{,}8 \\ \{1,6\} \text{ si } 0{,}8 < \alpha \end{cases}$

Indica todos los conjuntos difusos C que satisfacen estas condiciones.

Ejercicio 4.6. Sea U = {1, 2, 3, 4, 5, 6}, y sea D un conjunto difuso sobre U del que conocemos los siguientes α-cortes no estrictos:

$$C\alpha = \begin{cases} \{1,3,4,6\} \text{ si } 0 < \alpha \leq 0{,}3 \\ \{1,3,6\} \text{ si } 0{,}3 < \alpha \leq 0{,}8 \\ \{1,5,6\} \text{ si } 0{,}8 < \alpha \end{cases}$$

Indica todos los conjuntos difusos C que satisfacen estas condiciones.

Ejercicio 4.7. Supónganse definidos sobre N los siguientes conjuntos:

- "unos cuantos": {0,5/3, 1/4, 1/5, 0,5/6}
- "poco mayor que 1": {1/2, 0,8/3, 0,5/4, 0,2/5}
- F3: el resultado de sumar unas cuantas veces un número poco mayor que 1.

Razona cuáles serían el núcleo y el soporte de F3.

Ejercicio 4.8. Una relación difusa entre dos conjuntos (nítidos) U y V es un conjunto difuso sobre U ×V. Dar definiciones razonables de las siguientes relaciones difusas:

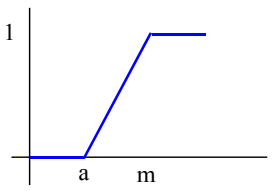
- "x disfruta estudiando y", donde x es el lector e y es una de las asignaturas que cursa.
- "x esta´ cerca de y", donde x, y ∈ {Nerja, Marbella, Fuengirola, Bilbao}.
- "x es mucho mayor que y", donde x, y ∈ N.

4.2.3 Funciones de pertenencia

Aunque en principio cualquier función sería válida para definir conjuntos difusos, en la práctica hay ciertas funciones típicas que siempre se suelen usar, tanto por la facilidad de computación que su uso conlleva como por su estructura lógica para definir su valor lingüístico asociado. Las funciones más comunes son:

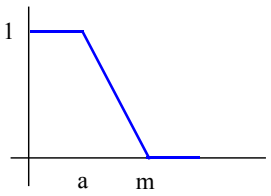
- Función GAMMA (Γ):

$$\mu(x) = \begin{cases} 0 & \text{para } x \leq a \\ \dfrac{x-a}{m-a} & \text{para } a < x < m \\ 1 & \text{para } x \geq m \end{cases}$$



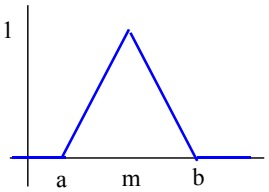
- Función L:

Puede definirse simplemente como 1 menos la función GAMMA



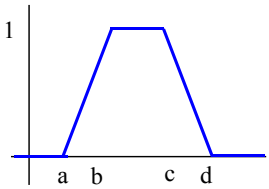
- Función LAMBDA o triangular:

$$\mu(x) = \begin{cases} 0 & \text{para } x \le a \\ \dfrac{x-a}{m-a} & \text{para } a < x \le m \\ \dfrac{b-x}{b-m} & \textit{para } m < x \le b \\ 0 & \text{para } x > b \end{cases}$$



- Función PI o trapezoidal:

$$\mu(x) = \begin{cases} 0 & \text{para } x \le a \\ \dfrac{x-a}{b-a} & \text{para } a < x \le b \\ 1 & \text{para } b < x \le c \\ \dfrac{d-x}{b-c} & \textit{para } c < x \le d \\ 0 & \text{para } x > d \end{cases}$$

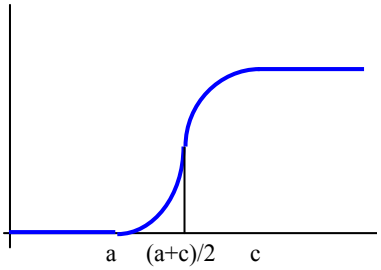


Las funciones L y GAMMA se usan para calificar valores lingüísticos extremos, tales como *bebé* o *anciano*, respectivamente. Las funciones PI y LAMBDA se usan para describir valores intermedios (como *joven*, de *mediana edad*, *maduro*). Su principal diferencia reside en que la función PI implica un margen de tolerancia alrededor del valor que se toma como más representativo del valor lingüístico asociado al conjunto difuso.

También se pueden utilizar otras funciones que no sean lineales a trozos. Por ejemplo:

- Función s, definida mediante:

$$\mu(x) = \begin{cases} 0 & \text{para } x \le a \\ 2\left(\dfrac{x-a}{c-a}\right)^2, & \text{para } a \le x \le \dfrac{a+} \\ 1-2\left(\dfrac{x-a}{c-a}\right)^2, & \text{para } \dfrac{a+c}{2} \le x \le \\ 1 & \text{para } x \ge c \end{cases}$$

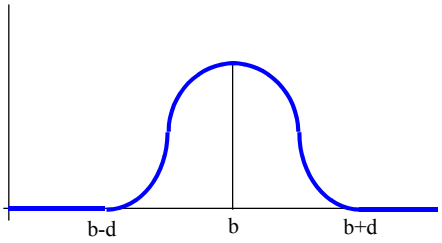


- Función z, que es la función opuesta:

$\mu_Z(x) = 1\mu_S(x)$

- Función P, definida mediante:

$$\mu(x) = \begin{cases} \mu_S(x) & \text{para } x \leq b \\ \mu_Z(x) & \text{para } x > b \end{cases}$$



4.2.4 Etiquetas lingüísticas

Tradicionalmente, se han utilizado modificadores de los conjuntos difusos a los que llamamos ***etiquetas lingüísticas***, equivalentes a lo que en lenguaje natural serían los adverbios. La interpretación en el modelo difuso de estos enunciados consiste en la composición de la función de pertenencia con una operación aritmética simple. Por ejemplo, es habitual considerar como interpretación del adverbio *muy* la operación de elevar al cuadrado la función de pertenencia original.

$$\mu_{\text{MUY A}}(x) = (\mu_A(x))^2$$

De este modo, si por ejemplo el grado de pertenencia de una persona a la clase *alto* es 0.6, el grado de pertenencia a la clase *muy alto* es 0.36.

Etiquetas lingüísticas

Del mismo modo, es habitual considerar como una interpretación de "*algo*" la operación de extraer la raíz cuadrada de la función de pertenencia general.

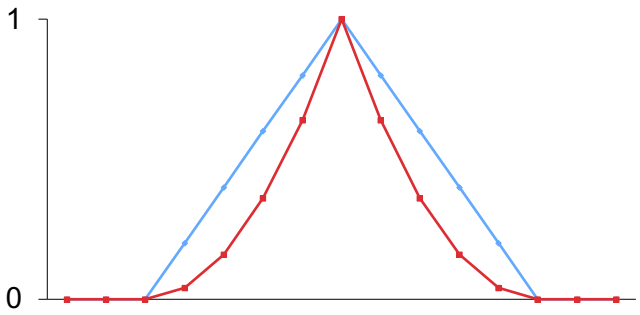
$$\mu_{\text{ALGO A}}(x) = \sqrt{\mu_A(x)}$$

Así, si el grado de pertenencia de una persona a la clase alto es 0.5, el grado de pertenencia a la clase algo alto es de 0.707. También se podría utilizar una raíz cúbica, etc. Cada implementador puede decidir entre las diferentes opciones, que normalmente se incluyen en las herramientas de implementación para control difuso.

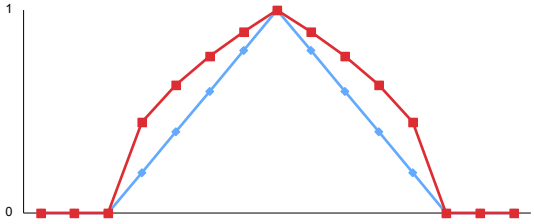
Existe todo un catálogo de posibles adverbios y sus modificadores asociados, pero las modificaciones que más usualmente se aplican a un conjunto difuso son las siguientes:

- *Normalización*, al convertir un conjunto difuso no normalizado en uno normalizado (dividiendo por la altura del conjunto).
- *Concentración*, al componer con una función tipo $f(y)=y^p$, con $p>1$. El efecto es que la función de pertenencia toma valores más pequeños, centrándose en los valores mayores.

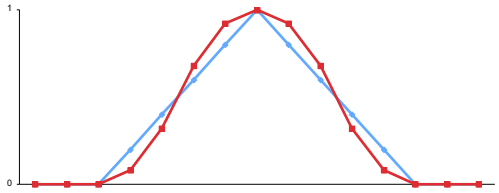
  El efecto de aplicar la concentración puede verse en la siguiente figura (la función de pertenencia base es la azul, y la modificada la roja):



- *Dilatación*, al componer con una función tipo $f(y)=y^p$ con $0<p<1$ (o también con $2y-y^2$). El efecto es el contrario a la concentración.



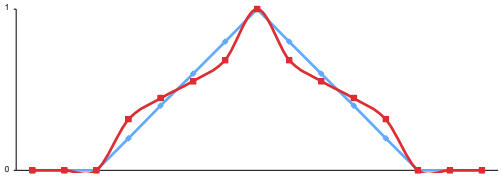
- *Intensificación del contraste*. Se disminuyen los valores menores a 1/2 y se aumentan los mayores. Componemos con una función del tipo:donde p>1. Normalmente se suele poner p=2 (a mayor p, mayor intensificación). El efecto es:



- Difuminación. Efecto contrario al anterior.

Se compone con la función:

Cuyo efecto es:



Los adverbios o modificadores pueden componerse entre sí, obteniendo múltiples combinaciones para representar enunciados complejos como "Juan es

mucho más que alto". Las herramientas de implementación de control difuso suelen tener varios modificadores predefinidos, como por ejemplo:

Nombre del modificador	Descripción del modificador
not	1-y
very (muy)	y^2
somewhat (algo)	$y^{1/3}$
more-or-less (más o menos)	$y^{1/2}$
extremely (extremadamente)	y^3

y también admite la definición de nuevos modificadores por el usuario.

4.2.5 Determinación de las funciones de pertenencia

Cuando tenemos que diseñar un sistema basado en razonamiento difuso, el primer paso será la determinación de los conjuntos difusos que emplearemos, es decir, de sus funciones de pertenencia. La determinación de éstas no es una tarea trivial y se han propuesto varios métodos para realizarla.

- ***Extracción del conocimiento humano.*** Un primer grupo de métodos su pone que los conjuntos que se van a manejar corresponden a conceptos más o menos presentes en la mente de *expertos* humanos. En este caso, habrá que diseñar protocolos experimentales que extraigan fiablemente de los sujetos este conocimiento.

  El método más obvio es el de *valoración directa*. Supongamos que deseamos determinar la función de pertenencia del conjunto *varones adultos altos*; entonces seleccionamos cierto conjunto de varones de muestra y le pedimos a un "evaluador" que proporcione un número 0 a 1 que indique si cada uno de ellos es alto o no lo es. Una variante es la *valoración inversa*: seleccionamos valores μi entre 0 y 1 y le pedimos al evaluador que indique los varones de los que se puede decir que son altos con grado μi.

  Alternativamente, podemos aplicar el método de la *votación*. Ahora seleccionamos un conjunto de evaluadores y a cada uno de ellos le presentamos los elementos de la muestra. Para cada elemento X de la muestra, cada evaluador deberá decidir si pertenece o no pertenece al conjunto (es decir, dará una respuesta nítida, 0/1). El grado de pertenencia de X será la proporción de res puestas positivas. Nótese que en este método estamos interpretando el grado de pertenencia de X al conjunto de "varones altos" como la probabilidad de que un observador diga que X es alto.

  Normalmente se supone además que las funciones de pertenencia tienen una forma conocida y solo hay que determinar sus parámetros. Por ejemplo, podemos suponer que son lineales a trozos o cuadráticas a trozos (estos son los casos definidos más arriba), gaussianas, logísticas, etc. Entonces ajustaremos los datos obtenidos de los evaluadores mediante alguna técnica estadística (mínimos cuadrados, por ejemplo).

- ***Aprendizaje automático.*** También es posible considerar conjuntos difusos que no provienen de la "cabeza" de los seres humanos, sino que se aprenden interactuando directamente con el entorno. Por ejemplo, si se dispone de un conjunto de muestras ya clasificadas de forma binaria (pertenece/no pertenece al conjunto C), es posible aplicar métodos basados en redes neuronales para generalizar el concepto C ejemplificado por estas muestras, e interpretar la salida de la red neuronal con entrada X como el grado de

pertenencia de X a C.

Aún más, los métodos de agrupamiento difuso que veremos más adelante pueden verse como una forma de aprendizaje no supervisado de conjuntos difusos.

Ejercicio 4.9. Para implementar un sistema difuso de control de una máquina, queremos definir la variable lingüística "número de revoluciones". Tras hablar con el experto encargado de su manejo, hemos extraído la siguiente información:

- El número es ciertamente "*alto*" si se superan las 1000 r.p.m. (a 1200 r.p.m. la máquina se rompe). Puede empezarse a decir que es alto cuando se alcanzan las 800 r.p.m.
- El número es ciertamente "*bajo*" si es inferior a 400 r.p.m. Si es inferior a 500 r.p.m, es "*algo bajo*".

Empleando funciones de pertenencia de tipo L, Π_r y Γ,

a) Representa gráficamente toda la información relativa a esta variable lingüística (universo de discurso y valores lingüísticos).
b) Idem, si además se supone que los valores que no son bajos ni altos se califican como "normales".

Ejercicio 4.10. Queremos saber qué entienden los habitantes de Kakastán por "ser rico". El bien que los kakastaníes usan para acumular su riqueza es la plata. Hemos averiguado la plata que atesoran 8 personas y para cada uno de ellos hemos preguntado a otros 100 individuos si lo considerían rico. Las respuestas obtenidas se resumen en esta tabla:

Arrobas de plata	0	5	10	15	20	25	30	35
% de síes	0	0	0	16,7	50	82,3	100	100

Suponiendo que el valor lingüístico "rico" viene dado por una función μ lineal a trozos de tipo Γ, definirlo de forma que se ajuste óptimamente a los datos recogidos.

Ejercicio 4.11. Realizar de nuevo el Ejercicio 4.9, pero ahora empleando funciones de tipo Z, Π y S. Para cada valor lingüístico, calcular la máxima diferencia (en valor absoluto) entre los valores de las funciones de pertenencia definidos ahora y los definidos en el ejercicio 7.

Ejercicio 4.12. Consideremos la variable y los valores definidos en el ejercicio 7. Se pide:

- Representa gráficamente los valores "*no bajo*", "*muy alto*", "*algo bajo*" y "*más o menos mediano*".
- Indica en qúe punto $x \in U$ difieren más $\mu_{bajo}(x)$ y $\mu_{algo\ bajo}(x)$, y cuál es el valor de esta diferencia máxima.

4.2.6 Operaciones elementales con conjuntos difusos

Al igual que en la teoría clásica de conjuntos, sobre los conjuntos difusos podemos definir las operaciones de unión, intersección, complementario, etc.

4.2.6.1 Complementario

Dado un conjunto A, el conjunto complementario de A está formado por los elementos del universo que no pertenecen a A. En el caso difuso, este conjunto

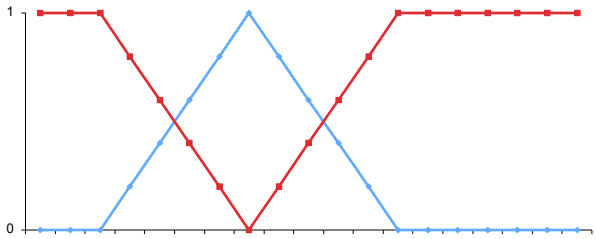
vendrá definido por una función de pertenencia que se calcula para cada elemento a partir de su pertenencia al conjunto A. Es decir:

$$\mu_{\overline{A}}(x) = c(\mu_A(x))$$

siendo c una función c: [0,1] → [0,1] que, dado el grado de pertenencia al conjunto A, nos da el grado de pertenencia al conjunto complementario de A. A esta función c desde un punto de vista intuitivo deben exigírseles las siguientes características:

- c1. concordancia con el caso nítido c(1) = 0 y c(0) = 1
- c2. estrictamente decreciente ∀ a,b∈ [0,1] a>b ⇒ c(a) < c(b)
- c3. involución ∀ a∈ [0,1] c(c(a)) = a

En general se considera como función del complementario a c(α) = 1 α. Así, para el conjunto difuso definido por una función triangular (por ejemplo, el conjunto difuso *mediano*) su complemento sería:

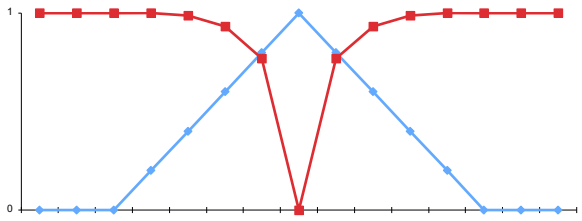


aunque también existen otras variantes que cumplen las propiedades antes citadas como:

- Complementario de Yager $c_w(a) = (1\ aw)1/w$ w∈ [0, ∞]

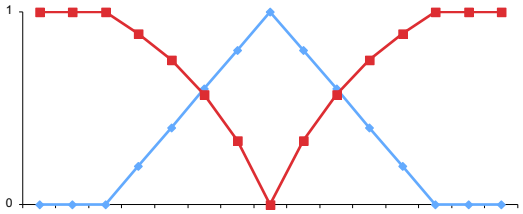
Para una función triangular y con w=2, tendríamos:

Funciones complemento



clase de complementarios de Sugeno $c_l(a) = \frac{1-\alpha}{1-\lambda\alpha}$ l∈ [0,1]

para λ =1/2:



4.2.6.2 Intersección

En teoría de conjuntos clásica, se considera que un elemento pertenece al conjunto intersección de dos conjuntos si pertenece a ambos. En el caso difuso el problema consiste en determinar el grado de pertenencia al conjunto intersección, conocido el grado de pertenencia a cada uno de los conjuntos originales. Supongamos:

$$\mu_{A\cap B}(x) = i(\mu_A(x), \mu_B(x))$$

donde:

$i : [0,1] \times [0,1] \rightarrow [0,1]$

análogamente al caso anterior, imponemos las siguientes condiciones:

$\forall a,b,g \in [0,1]$

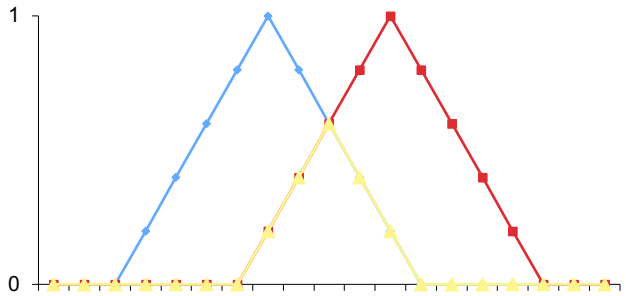
- **i1.** concordancia con el caso nítido $i(0,1) = i(0,0) = i(1,0) = 0;\ i(1,1) = 1$
- **i2.** conmutatividad $i(a,b) = i(b,a)$
- **i3.** asociatividad $i(a,i(b,g)) = i(i(a,b),g)$
- **i4.** identidad $(a,1) = a$
- **i5.** monotonía si $a \leq a'$ y $b \leq b'$, entonces $i(a,b) \leq i(a', b')$

Si se verifican los axiomas anteriores ([0,1],i) tiene estructura de semigrupo abeliano con elemento neutro. Las funciones i que verifican esta propiedad se llaman dentro de la teoría de conjuntos difusos *normas triangulares* (t-normas).

Las t-normas usadas más habitualmente son las siguientes:

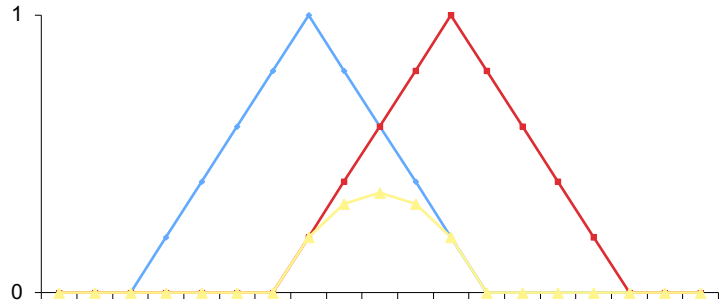
- t-norma del mínimo $i_{min}(a,b) = \min(a,b)$

Por ejemplo si consideramos dos funciones tipo triangular (*niño*, *adolescente*), la t-norma del mínimo sería:



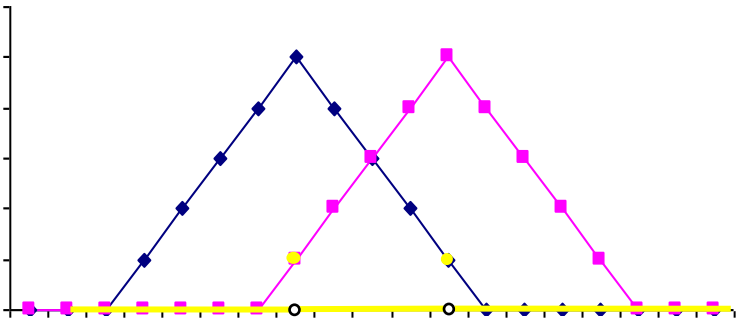
T-normas o funciones para la intersección difusa

- t norma del producto $i^*(a,b) = a * b$



- t-norma del producto drástico

$$i_{inf}(\alpha,\beta) = \begin{cases} \alpha & \text{si } \beta = 1 \\ \beta & \text{si } \alpha = 1 \\ 0 & \text{en otro caso} \end{cases}$$



Aunque no siempre se puede decir que una t-norma es mayor que otra, se puede demostrar que toda t-norma verifica las siguientes desigualdades:

$\forall a,b \in [0,1]$ $\quad i_{inf}(a,b) \le i(a,b) \le i_{min}(a,b)$,

es decir, que la menor t-norma es la t-norma del producto drástico y la mayor t-norma es la norma del mínimo.

4.2.6.3 Unión

Al igual que en el caso anterior podemos declarar una axiomática intuitiva para la unión de dos conjuntos difusos. Sea:

$$\mu_{A\cup B}(x) = u(\mu_A(x), \mu_B(x))$$

Donde es una función:

$u : [0,1] \times [0,1] \rightarrow [0,1]$

que debe verificar:

$\forall a, b, c \in [0,1]$

- **u1.** concordancia con el caso nítido $\quad u(0,1) = u(1,1) = u(1,0) = 1;\ u(0,0) = 0$
- **u2.** conmutatividad $\quad u(a,b) = u(b,a)$
- **u3.** asociatividad $\quad u(a,u(b,c)) = u(u(a,b),c)$

- **u4.** identidad ($A \cup \varnothing = A$) $u(a,0) = a$
- **u5.** monotonía Si $a \leq a'$ $b \leq b'$, entonces $u(a,b) \leq u(a', b')$

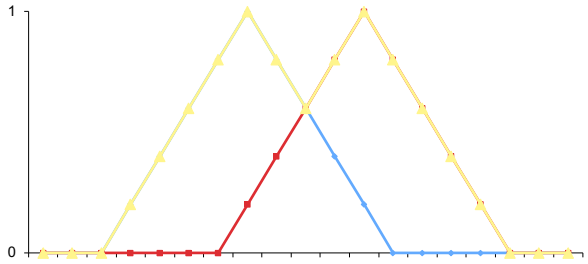
Además, sería deseable que se mantuvieran también las siguientes propiedades:

- **u6.** Leyes de De Morgan $u(a,b) = c(i(c(a),c(b))$
  $i(a,b) = c(u(c(a),c(b))$

Que nos permiten calcular el grado de la unión en función de los grados del complementario y la intersección. A las funciones que verifiquen estas seis propiedades se las llama *conormas triangulares* (t-conormas).

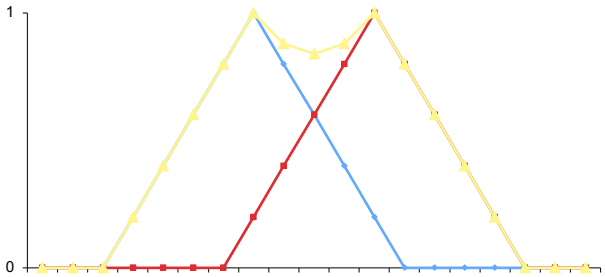
Considerando la función de complementación $c(a) = 1 - a$, las t-conormas correspondientes a las t-normas anteriores son:

- t-conorma del máximo $u_{max}(a,b) = max(a,b)$



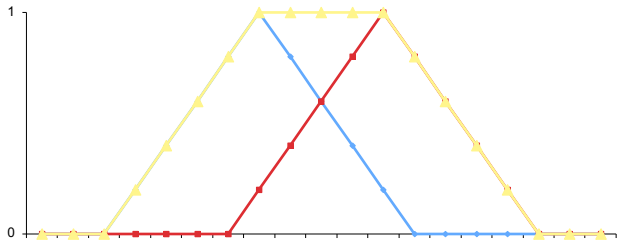
T-conormas o funciones para la unión difusa

- t-conorma de la suma $u^*(a,b) = a + b\ a * b$



- t-conorma de la suma drástica

$$u_{sup}(\alpha,\beta) = \begin{cases} \alpha & \text{si } \beta = 0 \\ \beta & \text{si } \alpha = 0 \\ 1 & \text{en otro caso} \end{cases}$$



análogamente al caso de la intersección, se puede demostrar que cualquier t-conorma verifica las desigualdades:

$\forall a,b \in [0,1] \qquad u_{max}(a,b) \le u(a,b) \le u_{sup}(a,b)$

Es decir, que la menor t-conorma es la del máximo y la mayor t-conorma la suma drástica.

Pero las condiciones que exigimos a la unión y a la intersección no garantizan en general que se cumplan las siguientes condiciones:

$\forall a,b,c \in [0,1]$

I1: Idempotencia $(A \cap A = A)$ $\quad i(a,a) = a$
I1: Distributividad $(A \cap (B \cup C)) = \ldots$ $\quad i(a,u(b,c)) = u(i(a,b),i(a,c))$
U1 : Idempotencia $(A \cup A = A)$ $\quad u(a,a) = a$
U2 : Distributividad $(A \cup (B \cap C)) = \ldots$ $\quad u(a,i(b,c)) = i(u(a,b),u(a,c))$

propiedades que sólo verifica la t-norma del mínimo y su t-conorma del máximo.

Podríamos también definir el conjunto vacío y el conjunto universal. El concepto de conjunto vacío corresponde al de aquel conjunto que no contiene ningún elemento. Por tanto, parece adecuado definirlo en la teoría de conjuntos difusos como:

$$\forall\, x \in X \quad \mu_{\varnothing}(x) = 0$$

y consiguientemente el conjunto universal se definiría como:

$$\forall\, x \in X \quad \mu_{X}(x) = 1$$

Conjuntos vacío y universal

Pero asumiendo estas definiciones no se verifican en la teoría de conjuntos difusos algunos famosos teoremas de la teoría de conjuntos clásica, como:

$$A \cap \overline{A} = \varnothing$$
$$A \cup \overline{A} = X$$

que se conocen como el principio de contradicción y del tercio excluso, respectivamente (lógica aristotélica).

Si tomamos por ejemplo el conjunto difuso "joven" es fácil comprobar que no se cumplen ninguno de los dos principios.

Sin embargo, es posible definir una t-norma y una t-conorma que satisfagan esto (la t-norma del producto acotado y la t-norma de la suma acotada), aunque entonces no se satisfarán las propiedades I1, I2, U1, U2.

Ejercicio 4.13. Supóngase que definimos la intersección y la unión de conjuntos difusos mediante la t-norma del mínimo y su correspondiente t-conorma. Sea U = {a, b, c, d, e} y sean C, D, E conjuntos difusos sobre U, C = 0,3/a + 0,7/b + 0,1/e, D = 1/a + 0,5/b + 0,1/e, E = 0,5/b + 1/c + 0,2/e. Ejercicio 4.1. Calcular: a) $\overline{C}$; b) C ∪ D; c) C ∩ D; d) C ∩ D ∩ E; d) C ∪ $\overline{C}$; e) C ∩ $\overline{C}$; f)$\overline{D \cup E}$; g)$\overline{D} \cap \overline{E}$.

Ejercicio 4.14. Supóngase que definimos ahora la intersección y la unión de conjuntos difusos mediante la t-norma del producto y su correspondiente t conorma. Realizar de nuevo las operaciones que se piden en el ejercicio 11 y comparar los resultados.

Ejercicio 4.15. Consideremos de nuevo el Ejercicio 4.9. Supóngase que definimos la intersección y la unión de conjuntos difusos mediante la t-norma del mínimo y su correspondiente t-conorma. Representar gráficamente los siguientes valores lingüísticos: a) ni normal ni alto; b) muy alto o muy bajo; c) bajo y normal; d) no muy bajo.

Ejercicio 4.16. Supóngase que definimos la intersección y la unión de conjuntos difusos mediante la t-norma del mínimo y su correspondiente t-conorma. Sean dos conjuntos difusos cualesquiera A, B sobre U. Definamos la posibilidad (incondicional) de A, Pos(A), como Pos(A) = altura(A), y la necesidad (incondicional) de A, Nec(A), como Nec(A) = 1 − Pos($\overline{A}$))

Demuestra que se cumplen las siguientes propiedades:

a) $0 \leq Nec(A) \leq Pos(A) \leq 1$
b) $Pos(A \cup B) = \max(Pos(A), Pos(B))$
c) $Pos(A) + Pos(\overline{A}) \geq 1$
d) $Pos(A) + Pos(\overline{A}) = 1$ sí y sólo sí para todos x, $y \in U$ $\mu_A(x) = \mu_A(y)$.
e) $Nec(A) + Nec(\overline{A}) \leq 1$
f) $Nec(A) + Nec(A) = 1$ sí y sólo sí si A = U.

Ejercicio 4.17. Supóngase que definimos la intersección y la unión de conjuntos difusos mediante la t-norma del mínimo y su correspondiente t-conorma. De mostrar que para cualesquiera conjuntos difusos A, B se cumplen las siguientes propiedades de la altura:

a) $altura(A \cup B) = \max(altura(A), altura(B))$
b) $altura(A \cap B) \leq \min(altura(A), altura(B))$
c) $altura(A \cup \overline{A}) \geq 0.5$
d) $altura(A \cap \overline{A}) \leq 0.5$

¿Se siguen cumpliendo si consideramos la t-norma del producto?

Ejercicio 4.18. La t-norma de Lukasiewicz se define por medio de la función

$i_1(\alpha, \beta) = máx(0, \alpha + \beta - 1)$

y su correspondiente t-conorma se define por:

$u_1(\alpha, \beta) = mín(1, \alpha + \beta)$

a) Demostrar que se trata realmente de una t-norma y su t-conorma.
b) Estudiar si para esta t-norma y t-conorma se cumplen las propiedades si guientes:
   - Idempotencia de la unión (para todo conjunto difuso A, $A \cup \overline{A} = A$).
   - Distributividad de la unión respecto a la intersección.
   - Contradicción (para cualquier conjunto difuso A, $A \cap \overline{A} = \varnothing$).

- Tercio excluso (para cualquier conjunto difuso A, $A \cup \overline{A} = U$).

4.3 Razonamiento difuso

La teoría de conjuntos difusos nos permite representar hechos y relaciones vagas (imprecisas). Se entiende por razonamiento difuso el proceso de realizar inferencias a partir de hechos y relaciones difusas, así como la combinación de evidencias difusas y la actualización de la precisión de las creencias. Vamos a definir a continuación los elementos básicos de la lógica difusa.

4.3.1 Hechos difusos

4.3.1.1 Hechos difusos simples

Un hecho difuso simple o ***proposición difusa simple*** es aquella que asigna un valor a una variable difusa, por ejemplo: "la estatura de Pepe es *mediana*" o "la velocidad es *normal*". Una proposición difusa tiene por tanto asociado un conjunto difuso A (el valor lingüístico asignado, "mediana" en este caso) y su correspondiente función de pertenencia μ_A definida sobre los elementos del universo de discurso $u \in U$.

Para cada valor $u \in U$, el valor de verdad $\mu_P(u)$ de la proposición difusa p dada por "*u es A* " se identifica con el grado de pertenencia $\mu_A(u)$ de u al conjunto A.

4.3.1.2 Hechos difusos compuestos

Un *hecho difuso compuesto* o *proposición difusa compuesta* es aquella que se obtiene mediante la agrupación de dos o más proposiciones difusas simples, que pueden haber sido modificadas o no antes de la agrupación. Para agrupar proposiciones difusas simples podemos utilizar las conectivas Y y O, y para modificar una proposición difusa simple podemos utilizar el NO. Así por ejemplo, dados los hechos difusos simples p = "la velocidad es *normal*" y q = "el objeto está *cerca*" podemos construir proposiciones difusas del tipo:

$\neg p$	"la velocidad NO es *normal*"
$p \wedge q$	"la velocidad es *normal*" Y "el objeto está *cerca*"
$p \vee q$	"la velocidad es *normal*" O "el objeto está *cerca*"

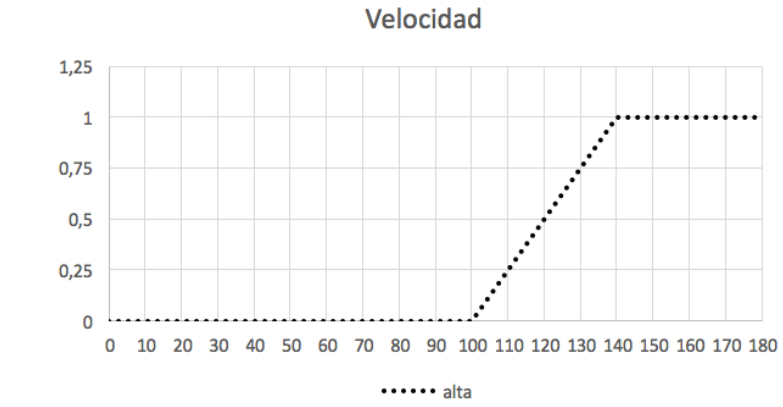
Los hechos difusos compuestos pueden considerarse también como proposiciones, es decir, podemos darles valores de verdad. Para ello hay que definir para cada conectiva el correspondiente operador lógico difuso. Estos operadores difusos se definen de forma análoga a como se definieron las operaciones entre conjuntos (complemento, unión e intersección).

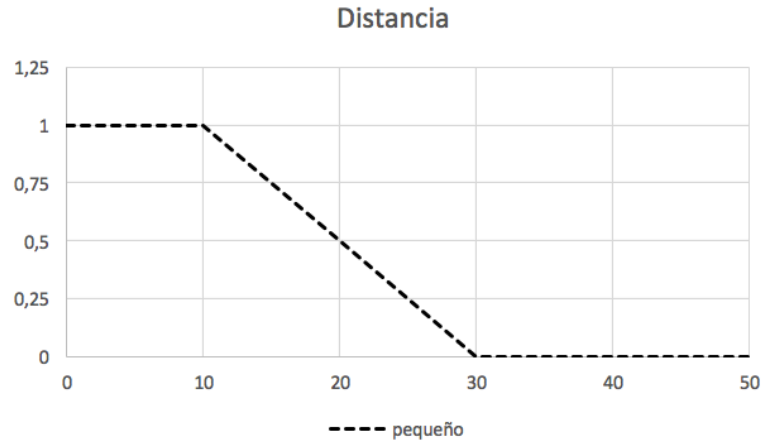
Más concretamente, sean p y q dos proposiciones difusas, y A y B los conjuntos difusos que intervienen en ellas, con funciones de pertenencia μ_A y μ_B definidas respectivamente sobre universos de discurso U y V. Entonces, los operadores lógicos pueden definirse mediante:

- NO ($\neg p$) vendrá definida por la función de pertenencia al complemento de A. Por tanto, en el caso habitual, el valor de verdad de *u es NO A* vendrá dado por una función de pertenencia tipo complemento de A, por ejemplo $\mu_{\neg A}(u) = 1 - \mu_A(u)$.

- Y ($p \wedge q$) vendrá definida por una función de pertenencia tipo intersección. Por ejemplo, el valor de verdad de $p \wedge q$ podría ser $\mu_{A \wedge B}(u,v) = \min(\mu_A(u), \mu_B(v))$
- O ($p \vee q$) vendrá definida por una función de pertenencia tipo unión. Por ejemplo el valor de verdad de $p \vee q$ podría ser $\mu_{A \vee B}(u,v) = \max(\mu_A(u), \mu_B(v))$.

Ejemplo 4.3. Supongamos que la variable velocidad tiene como universo de discurso V el intervalo [0,180], y que el valor *alto* es un conjunto Γ con núcleo [140,180] y soporte [100,180]. Supongamos además que la variable distancia tiene como universo de discurso D el intervalo [0,50] y que el valor *pequeño* es un conjunto L con núcleo [0,10] y soporte [0,30]. Estas definiciones se representan en las siguientes gráficas





Si el valor exacto de la velocidad es v = 130, y el valor exacto de la distancia es d = 20, tendremos los siguientes valores de verdad:

La velocidad es alta	0,75
La velocidad no es alta	0,25
La distancia es pequeña	0,50
La velocidad es alta y la distancia pequeña	0,75
La velocidad es alta o la distancia pequeña	0,50

4.3.1.3 Correspondencia entre hechos difusos

¿Qué ocurriría si el valor conocido de la velocidad o de la distancia no es nítido, sino también difuso? Por ejemplo, podemos saber únicamente que la velocidad es "aproximadamente 100". En este caso, hay que añadir un criterio adicional para determinar el valor de verdad que debe atribuirse a "la velocidad es alta",

supuesto que es "aproximadamente 100"; o, dicho de otra forma, el valor de la correspondencia entre "la velocidad es alta" y "la velocidad es aproximadamente 100". La correspondencia entre la premisa "v es A" y el hecho "v es B" suele definirse como *la posibilidad de A dado B*, P os(A | B):

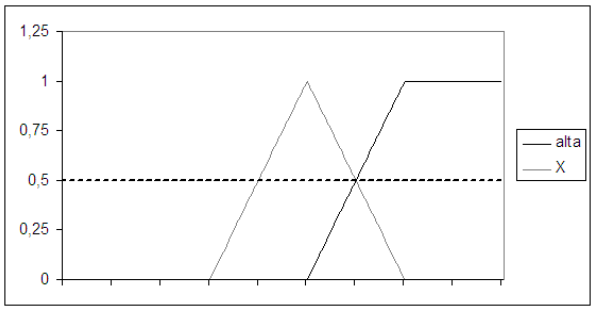
$$Pos(A \mid B) = Pos(B \mid A) = \max(A \cap B)$$

La medida dual de la posibilidad es la necesidad de A dado B, Nec(A | B), que se define como 1 menos la posibilidad del complementario de A dado B, o sea,

$$Nec(A \mid B) = 1 - Pos(\overline{A} \mid B) = 1 - \max((\overline{A} \cap B)$$

Con la medida de la posibilidad, podemos dar un valor $\alpha \in [0, 1]$ a cualquier proposición difusa en la que aparecen las variables u, v si conocemos sus valores difusos A', B', como muestra el siguiente ejemplo.

Ejemplo 4.4 Supongamos la variable velocidad y su valor *alta*, definidos como en el ejemplo 4.3. Supongamos ahora que el valoractual de la velocidad se define mediante un conjunto X de tipo Lambda con soporte [60,140] y núcleo [100,100], tal como se muestra en la siguiente figura



Gráfica de la variable velocidad

Tendremos los siguientes valores para la posibilidad de las respectivas proposiciones, supuesto que sabemos que la velocidad es X:

La velocidad es alta	0,50
La velocidad no es alta	1

▷

4.3.2 Implicaciones difusas

Estamos ya en condiciones definir lo que significa una implicación. Para ello, tenemos que asignar una función de pertenencia a una agrupación antecedente consecuente del tipo p→q. Definir el significado de la implicación nos permitirá razonar con reglas del tipo:

SI "la velocidad es *normal*"
ENTONCES "la fuerza de frenado debe ser *moderada*"

Esta función de pertenencia será del tipo:

$\mu_{p \to q}$: $U \times V \to [0,1]$

$$(u,v) \rightarrow \mu_{p\rightarrow q}(u,v)$$

Al definir la relación de implicación surge una cuestión importante ¿qué se quiere representar mediante la relación de implicación? La cuestión es fundamental porque las relaciones de implicación son la base del razonamiento basado en reglas. Para ello tenemos varias posibilidades, que describimos a continuación.

4.3.2.1 Implicación de Mamdani

La primera posibilidad es dar a la implicación el significado de relaciones causa-efecto normalmente utilizadas en los sistemas basados en conocimiento. La función más utilizada actualmente en problemas de control difuso y simulación, y fue propuesta por Mamdani y se corresponde con la siguiente función.

IMPLICACIÓN DE MAMDANI: $p\rightarrow q \equiv A\wedge B \Rightarrow \mu_{p\rightarrow q}(u,v) = \min(\mu_A(u), \mu_B(v))$

Para Mamdani, el grado de verdad de $p\rightarrow q$ es idéntico al de la proposición A y B. Podríamos justificar esto diciendo que, para Mamdani, una condición tan sólo resulta cierta cuando el antecedente es cierto y el consecuente también.

4.3.2.2 Implicación tipo max-prod

En este caso se utiliza la siguiente función:

IMPLICACIÓN MAX-PROD: $\mu_{p\rightarrow q}(u,v) = \text{prod}(\mu_A(u), \mu_B(v))$

4.3.2.3 Implicaciones basadas en la lógica clásica

En este caso, utilizamos funciones inspiradas en las equivalencias lógicas. Por ejemplo, en lógica clásica tenemos la equivalencia $p\rightarrow q \equiv \neg p\vee q$ (interpretación de Kleene-Dienes). De este modo, la función de pertenencia asociada a la regla "Si A entonces B", donde A y B son conjuntos difusos sería:

$\mu_{p\rightarrow q}(u,v) = \max(1-\mu_A(u), \mu_B(v))$

En lógica clásica también tenemos la equivalencia $p\rightarrow q \equiv \sim(p\wedge(\sim q))$, que conduciría a la siguiente definición:

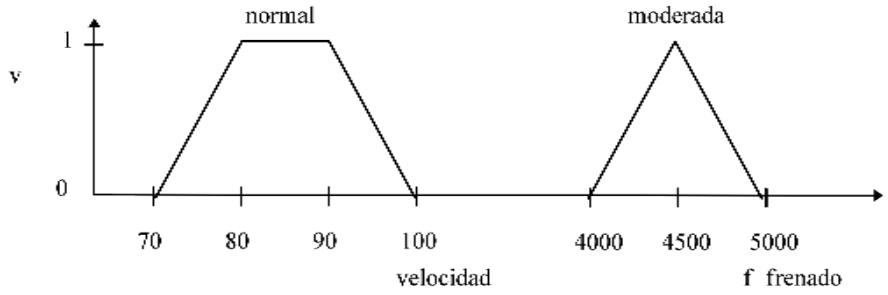
$\mu_{p\rightarrow q}(u,v) = 1 - \min[\mu_A(u), 1-\mu_B(v)]$

Ambas funciones son equivalentes. Aunque la implicación lógica difusa es interesante desde el punto de vista teórico, conduce a una formulación inadecuada para muchas aplicaciones de sistemas basados en conocimiento, que representan las relaciones causa-efecto de un modo no consistente plenamente con la lógica.

A continuación, vamos a ver ejemplos de inferencias realizadas con estas implicaciones, tanto en el caso de que los hechos iniciales sean difusos como que sean nítidos.

4.3.3 Ejemplos de aplicación

Para los ejemplos utilizaremos la siguiente situación: supongamos que tenemos dos variables lingüísticas, *velocidad* y *fuerza de frenado*, y los conjuntos difusos *A = normal* y *B= moderada* representados en la siguiente figura:



Supongamos asimismo que tenemos la regla:

p→q ≡ "SI la velocidad es normal, ENTONCES la fuerza de frenado es moderada".

Sobre este mismo ejemplo vamos a ver dos situaciones diferentes: en primer lugar, presentaremos un ejemplo en el que se realiza una inferencia difusa a partir de una premisa nítida, y a continuación un ejemplo en que se realiza una inferencia difusa a partir de una premisa difusa.

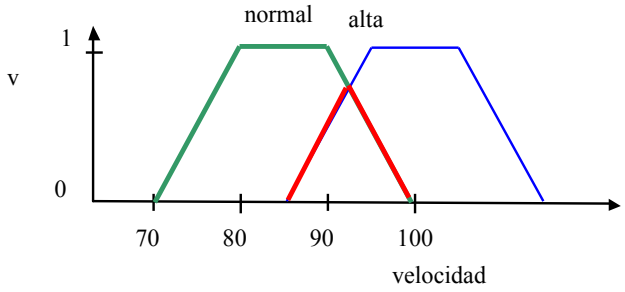
4.3.3.1 Ejemplo: inferencia difusa con hechos difusos

Supongamos que tenemos la situación descrita y el hecho del que disponemos es p* = "la velocidad es *alta*".

Tendríamos en primer lugar que establecer la correspondencia del conjunto difuso alta con el conjunto difuso normal, para determinar z como:

donde $z = \max(\min(\mu_{A^*}(u), \mu_A(u)))$

Así:

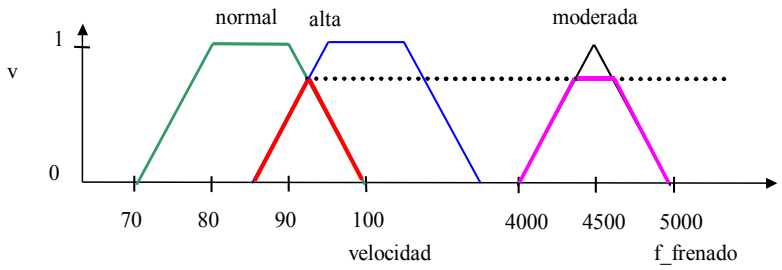


Por lo que z = 0.75

Vamos a aplicar las diferentes implicaciones para obtener q*, la conclusión de este sistema de razonamiento.

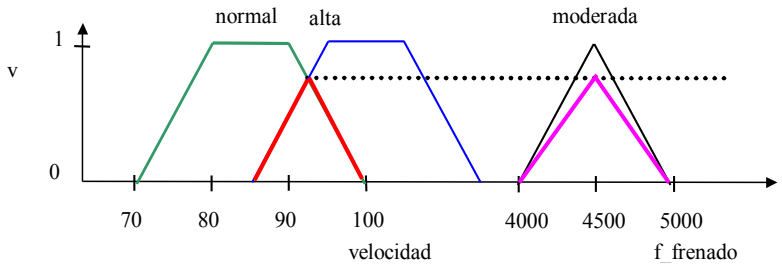
- Si aplicamos la implicación max-min o Mamdani, q* se obtendría mediante la función de pertenencia:

$$\mu_B^*(v) = \min(z, \mu_B(v))$$



- Si aplicamos la implicación max-prod:

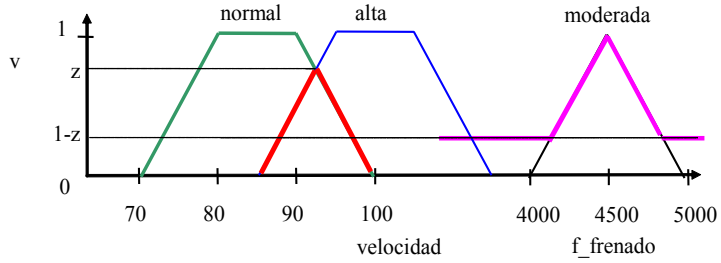
$$\mu_B^*(v) = prod(z, \mu_B(v))$$



- Si utilizamos las implicaciones lógicas:

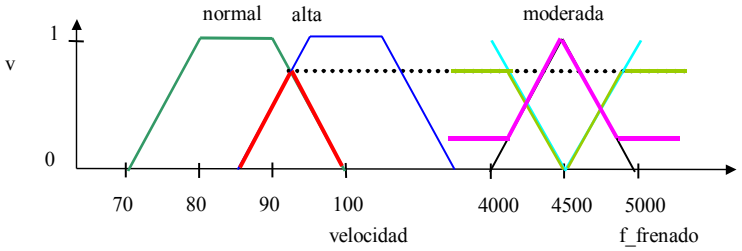
$\mu_B^*(v) = \max (1-z, \mu_B(v))$

cuyo resultado es:



o

$\mu_B^*(v) = 1-\min(z, 1-\mu_B(v))$



4.3.3.2 Ejemplo: Inferencia difusa con hechos nítidos.

Vamos a suponer que tenemos una regla difusa del tipo:

Si p ENTONCES q
y un valor de entrada nítido p*.

La conclusión será un hecho difuso q*, del cual queremos saber su función de pertenencia.

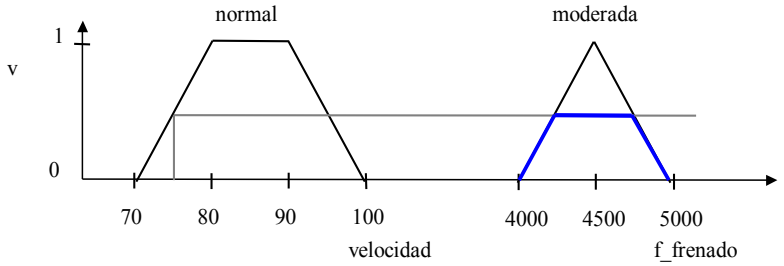
Vemos que el valor de la función de pertenencia para el hecho nítido p es $\mu_A^*(75) = 0.5$. La función de pertenencia asociada a la regla es $\mu_{p\rightarrow q}(x,y) = \min(\mu_A(u), \mu_B(v))$. El resultado de la inferencia será una proposición difusa q* con su correspondiente conjunto difuso B* asociado, que vendrá dado por la función de pertenencia $\mu_B^*(y)$.

La única diferencia con el caso nítido es la forma de escoger el valor z, que en este caso se calcula simplemente como $z=\mu_A(x)$, donde x es el valor nítido del que dispongamos, en nuestro ejemplo 75. La inferencia se hace entonces con cualquiera de las alternativas vistas en el apartado anterior, por ejemplo:

Inferencia tipo max-min (implicación de Mamdani):

$\mu_B^*(y) = \min(\mu_A(75), \mu_B(y))$

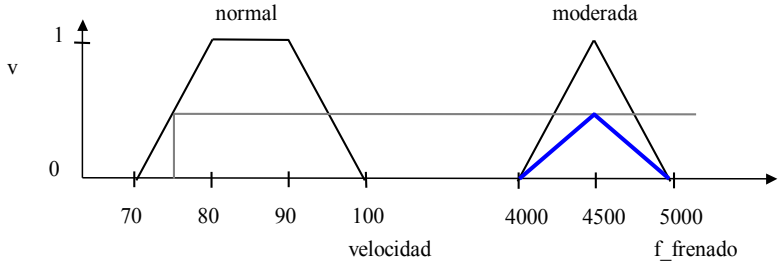
Es decir, que el resultado de la inferencia sería "velocidad es moderada*", donde la función de pertenencia del conjunto difuso moderada* es la representada en negrita en la siguiente figura:



Inferencia tipo max-prod:

$\mu_B^*(y) = \text{prod}(\mu_A(75), \mu_B(y))$

Cuyo resultado es el representado en la siguiente figura:



Igual se haría en el caso de que la implicación se interprete como una implicación lógica.

4.3.4 Decodificación (transformación de un conjunto difuso en un valor nítido)

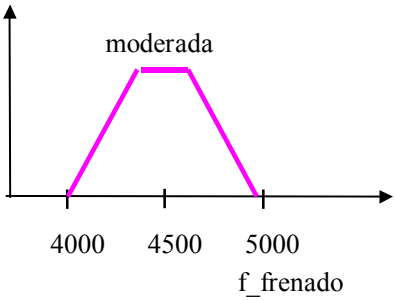
La principal aplicación de los sistemas de razonamiento difuso es el control de dispositivos, que normalmente precisan de una salida nítida (acción de control). Así por ejemplo en el ejercicio anterior podemos querer saber qué fuerza de frenado que debemos aplicar si la velocidad es *alta*. Existen diversas alternativas para transformar un valor difuso en nítido (proceso que en inglés se llama *defuzzification* y en español podríamos llamar decodificación. Para ello se han propuesto diversos métodos, siendo los más empleados los siguientes:

Métodos basados en los valores de máxima pertenencia.

Para decodificar, se elige el valor que tiene el grado máximo de pertenencia, es decir, el valor más posible. En el caso de que haya varios valores que tienen grado máximo de pertenencia, el empate puede rompers:

- seleccionando el mínimo de ellos (método *smallest of maximum*, SOM);
- seleccionando el mayor (método *largest of maximum*, LOM)
- seleccionando el valor medio (método*mean of máximum*, MOM)

De este modo, si por ejemplo la variable de salida es la representada por el cojunto difuso:



Vemos que la función de pertenencia tiene varios máximos: todos los valores entre 4250 y 4750. Si decodificamos el conjunto con la estrategia SOM, nos quedaríamos con el más pequeño de los máximos (4250); si decodificamos el conjunto con la estrategia LOM, nos quedaríamos con el mayor de los máximos (4750), y si decodificamos con la estrategia MOM, nos quedaríamos con el valor medio de los máximos (4500).

b) Método del centroide (*Center of Gravity*, COG). Sea A un conjunto difuso con universo de discurso U, y soporte S. El valor nítido a que se toma para representar A será la abcisa del centroide de A, es decir, en el caso discreto sería:

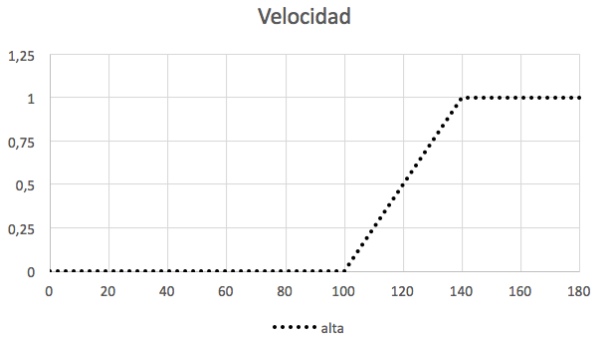
$$a_{centroide} = \frac{\sum_{u \in S} u \, \mu_A(u)}{\sum_{u \in S} \mu_A(u)}$$

Y, en el caso continuo:

$$a_{centroide} = \frac{\int_S u \, \mu_A(u) du}{\int_S \mu_A(u) du}$$

Si el conjunto A es simétrico respecto a un eje vertical $u=u_1$, se puede demostrar que el método COG coincide con el método MOM.

Ejemplo: Supongamos que la variable velocidad tiene como universo de discurso V el intervalo [0,180], y que el valor *alta* es un conjunto Γ con núcleo [140,180] y soporte [100,180], tal como se representa en la siguiente figura:



El resultado obtenido con cada técnica de defuzzificación sería:

- SOM a= 140
- LOM a = 180
- MOM a = 160
- COG a= 166.44

Puede observarse que como en este caso el conjunto no es simétrico, el centro de gravedad del mismo está desplazado a la derecha respecto al valor que da la media de los máximos. Nótese también que en el paso de decodificación se pierde mucha información acerca de la forma del conjunto, y además se obtienen

diferentes resultados según la técnica aplicada, por lo que, siempre que sea posible, es preferible tratar con el conjunto difuso en lugar de con su valor decodificado.

4.3.5 Acumulación de evidencia

Cuando se han aplicado varias reglas que establecen conjuntos difusos diferentes para una variable, es necesario acumular la evidencia obtenida tras la aplicación de cada una de esas reglas. Para ello lo que se hace es unir los conjuntos difusos resultantes, con cualquiera de las funciones tipo unión (las descritas en el apartado 4.2.6.3 o cualquier otra del amplio catálogo existente para representar uniones difusas).

4.3.6 Encadenamiento de reglas

En el caso de que en el problema aparezcan reglas encadenadas (por ejemplo, si p entonces q, y si q, entonces r, se procede aplicando la primera regla para obtener el conjunto difuso resultante que representará a q*, y después aplicando la segunda regla con la premisa difusa q*. No es conveniente decodificar q* debido a la pérdida de información que supone el proceso de decodificación.

4.3.7 Ejercicios de control difuso

Con estas definiciones estamos ya en condiciones de abordar los problemas de control difuso. Según hemos visto, para cada problema de control difuso, habrá cinco decisiones que tomar:

- Función que se va a utilizar para la unión difusa (t-conorma del máximo, de la suma, de la suma drástica, etc.)
- Función que se va a utilizar para la intersección (t-conorma del mínimo, del producto, del producto drástico, etc.)
- Función que se va a utilizar para la implicación difusa (Mamdani, max-prod, basadas en la lógica, etc.)
- Función que se va a utilizar para acumular la evidencia (las de tipo unión)
- Método de decodificación (SOM, LOM, MOM, COG).

Obviamente, dependiendo de la elección en cada una de estas opciones, el resultado podría ser diferente.

Ejercicio 4.19. La abuela María prepara sus deliciosas galletas caseras de forma artesanal desde hace más de 40 años. El toque secreto de la receta consiste en hornearlas cuidadosamente hasta que toman su característico color dorado. Durante este delicado proceso la abuela María observa periódicamente las galletas y ajusta la temperatura del horno de forma adecuada:

R1. Si las galletas están un poco crudas, entonces la temperatura del horno debe ser alta.
R2. Si las galletas están medio hechas, entonces la temperatura del horno debe ser media.
R3. Si las galletas están doraditas, entonces la temperatura del horno debe ser baja.

Tras diversas entrevistas con la abuela se han podido establecer los siguientes conjuntos difusos sobre un índice cromático especial (0 = galleta cruda; 10 = galleta chamuscada) y la temperatura del horno:

Índice cromático galletas:	Temperatura del horno (^{0}C):
un poco crudas: (1/4, 0.5/6, 0/7)	baja: (0/150, 1/160, 1/180, 0/190)
medio hechas (0/3, 1/5, 1/6, 0/8)	media: (0/170, 1/190, 1/210, 0/230)

doraditas: (0/5, 1/7, 1/8, 0/9) alta: (0/210, 1/220, 1/240, 0/250)

Supóngase que se interpretan las reglas R1, R2 y R3 anteriores como implicaciones de Mamdani y se construye un sistema para control automático de la temperatura del horno basado en reglas con encadenamiento hacia delante. Suponiendo que en cierto momento el índice cromático de las galletas es 6, se pide:

a) Trazar gráficamente la ejecución del sistema, mostrando el resultado producido por cada regla y el conjunto difuso resultante correspondiente a la temperatura.
b) ¿Cuál será el valor de temperatura aplicado al horno si se utiliza la técnica de la media de los máximos para obtener valores nítidos?

Ejercicio 4.20. Considérese un sistema con las siguientes reglas, interpretadas como implicaciones de Mamdani:

R1. Si la temperatura es alta entonces la presión es elevada.
R2. Si la temperatura es baja entonces la presión es baja.
R3. Si la presión es baja entonces la entrada de combustible debe ser grande
R4. Si la presión es elevada entonces la entrada de combustible debe ser pequeña

Considérense los siguientes conjuntos difusos:

Temperatura(ºC):

baja = (0/0 .2/30 .8/40 1/50 .7/60 .2/70 0/80)
alta = (0/50 .3/60 .8/70 1/80 1/90 .5/100 0/110)

Presión(bar):

baja = (0/0 .4/200 .8/400 1/600 1/800 .8/1000 .4/1200 0/1400)
elevada = (0/1000 .2/1200 .4/1400 .8/1600 1/1800 1/1900 0.5/2000 0/2200)

Entrada combustible(litros/hora):

pequeña = (0/0 .6/1 1/2 1/3 .4/4 0/5) grande = (0/4 .5/5 1/6 .5/7 0/8)

Si la temperatura actual es 60 ºC, determinar el valor para la entrada de combustible empleando la técnica del primer valor máximo para transformar valores difusos en nítidos

Ejercicio 4.21. Agapito lleva 8 años encargado del control de la turbina GG-35 y es ya una autoridad en su manejo. Agapito ha reconocido que para controlar la turbina solamente se fija en el ruido que produce y en un sensor de temperatura, concretamente:

R1. Si el nivel de ruido es normal y la temperatura es alta, entonces establece una velocidad suave.
R2. Si el nivel de ruido es normal y la temperatura no es alta, entonces establece una velocidad moderada.
R3. Si el nivel de ruido es bajo, entonces establece una velocidad alta.
R4. Si la velocidad es suave, la fuerza de frenado debe ser normal.
R5. Si la velocidad es moderada, la fuerza de frenado debe ser alta.
R6. Si la velocidad es alta, la fuerza de frenado debe ser alta.

Tras diversas entrevistas con Agapito se han elaborado los siguientes conjuntos difusos para los valores del nivel de ruido, temperatura, velocidad y fuerza de frenado:

Nivel de ruido (sobre una escala de 0 a 12):

bajo (0/1, 1/3, 1/5, 0/7)
normal (0/5, 1/7, 1/9, 0/11)

Temperatura (sobre una escala de 20o a 100o C):

alta (0/40, 1/60, 0/80)

Velocidad (sobre una escala de 0 a 100 rpm): suave (0/10, 1/30, 0/50)

moderada (0/30, 1/50, 0/70)
alta (0/20, 0.5/30, 0.5/40, 1/50, 0.5/60, 0.5/70, 0/80)

Fuerza de frenado (en un índice especial que varía de 0 a 5)

normal (0/1, 1/3, 0/5)
alta (0/3, 1/4, 1/5)

Suponiendo que en cierto momento el nivel de ruido es 5,5 y la temperatura de 30o C, se pide trazar gráficamente el proceso de razonamiento de Agapito, mostrando el resultado producido por cada regla, el conjunto difuso resultante correspondiente a velocidad de la turbina y el conjunto difuso correspondiente a la fuerza de frenado. ¿Cuál será el valor de la fuerza de frenado si se utiliza la técnica de la media de los valores máximos para obtener valores nítidos?

Ejercicio 4.22. La variable lingüística "presión" tiene como universo de discurso [20, 100] y se definen sobre ella los valores "alta" como el conjunto trapezoidal de soporte [60, 100] y núcleo [80, 100]; y "media", como el conjunto trapezoidal de soporte [20, 80] y núcleo [40, 60]. La variable "grado de peligro" tiene como universo de discurso [0, 1] y se definen los valores "grave" como el conjunto trapezoidal de soporte [0,6, 1] y núcleo [0,8, 1]; y "medio", como el conjunto trapezoidal de soporte [0,2, 0,8] y núcleo [0,4, 0,6]. Las reglas que se van a aplicar son

R1: Si la presión es alta, entonces el grado de peligro es grave.
R2: Si la presión es media, entonces el grado de peligro es medio.

Se tomará como valor nítido representativo de un conjunto difuso el de máximo grado de pertenencia y, si hay varios, su promedio.

Se pide dar un valor nítido para "grado de peligro" cuando "presión" es

a) el valor nítido 60
b) el valor nítido 70

Ejercicio 4.23. Para implementar un sistema de control difuso hemos hablado con diversos expertos, que nos han dicho lo siguiente:

Si la máquina va despacio, hay que apretar mucho el botón, y si la máquina va deprisa, hay que apretarlo poco; en otro caso, hay que apretarlo lo normal. La máquina debe ir más o menos a 1000 rpm; cuando va a 500 rpm, podemos decir que va *realmente* despacio, y cuando va a 1500 rpm, que va *realmente* deprisa. La fuerza ejercida sobre el botón en condiciones normales es de 10 N, la mínima posible de 5 N y la máxima posible de 15 N; si ejercemos solamente 6 N, es *realmente* poco, y si ejercemos 14 N, es *realmente* mucho.

Se pide:

a) Definir de manera razonable las variables y valores lingüísticos necesarios, en función de la información que nos han dado los expertos.

b) ¿Qué fuerza debe ejercerse sobre el botón cuando la máquina va a 750 rpm? Dar el valor difuso, y el valor nítido resultante de aplicar el procedimiento de la media de los máximos.

Ejercicio 4.24. La variable lingüística In tiene como universo de discurso [1000, 1600] y se definen sobre ella los valores "bajo" como el conjunto de tipo Λ de soporte [1000, 1400] y "alto", como el conjunto de tipo Λ de soporte [1200, 1600]. La variable lingüística Out tiene como universo de discurso [100, 400] y se definen sobre ella los valores "negativo" como el conjunto de tipo Λ de soporte [100, 300] y "positivo", como el conjunto de tipo Λ de soporte [200, 400]. Las reglas que se van a aplicar son

Si *In* es bajo, entonces *Out* es bajo. Si *In* es alto, entonces *Out* es alto.

Se pide dar la expresión analítica y gráfica de la relación entre los posibles valores de la variable de entrada In y la salida calculada por el sistema,

a) Cuando la entrada es un valor nítido x que no se difumina y para nitidificar los conjuntos difusos se usa la técnica COG o del centroide.
b) Cuando la entrada es un valor nítido x que no se difumina y para nitidificar los conjuntos difusos se usa la técnica MOM.
c) Cuando la entrada es un valor x que se difumina y se transforma en un conjunto de tipo Λ de soporte [x − 20, x + 20], y para nitidificar los conjuntos difusos se usa la técnica MOM.

Ejercicio 4.25. Considérese el siguiente sistema de razonamiento difuso para el ajuste del grado de apertura del diafragma y la velocidad de obturación de una cámara digital:

R1. Si la luminosidad es poca, el grado de apertura debe ser alto
R2. Si la luminosidad es media, el grado de apertura debe ser medio
R3. Si la luminosidad es intensa, el grado de apertura debe ser bajo
R4. Si el grado de apertura es alto o medio, la velocidad debe ser rápida
R5. Si el grado de apertura es alto, la velocidad debe ser lenta

Las variables lingüísticas y sus valores vienen definidos mediante:

Variable Apertura (Universo de discurso: de 2 a 16 f)

alto (1/2, 0/4)
medio (0/2, 1/4, 0/16)
bajo (0/8, 1/16)

Variable Logvelocidad (Universo de discurso: de -10 a 6 logs)

rápida (1/-4, 0/2)
lenta (0/-4, 1/2)

Variable Luminosidad (Universo de discurso: de 500 a 3000 candelas)

poca (1/500, 0/1000)
media (0/500, 1/1000, 1/1500, 0/2000)
intensa (0/1500, 1/2000, 1/2500)
lenta (0/-4, 1/2)

Supongamos que la luminosidad actual se define como el siguiente conjunto difuso:

(0/1500, 1/2000, 0/2500).

Calcular la velocidad de obturación y el grado de apertura del diafragma, utilizando la media de los máximos como técnica de decodificación. ¿En qué sentido cambiaría la respuesta si empleáramos la técnica del centroide?

4.4 Bibliografía

Francisco Javier Díez. Apuntes de razonamiento aproximado. http://www.ia.uned.es/fjdiez/libros/razaprox.html, 2009.

Mohammed Jamshidi, Nader Vadiee, y Timothy Ross. Fuzzy logic and control: software and hardware applications. Prentice Hall, 1993.

George Klir & Bo Yuan. (1995). Fuzzy Sets and Fuzzy Logic. Theory and Applications. pper Saddle River, NJ: Prentice Hall.

R. Kruse, J. Gebhardt, y F. Klawonn. Foundations of Fuzzy Systems. Wiley, Chichester, 1994.

E. H. Mamdani. Advances in the linguistic synthesis of fuzzy controllers. International Journal of Man-Machine Studies, 8:669–678, 1976.

E. H. Mamdani y S. Assilian. An experiment in linguistic synthesis with a fuzzy logic controller. International Journal of Man-Machine Studies, 7:1–13, 1975.

F. Martin McNeill y Ellen Thro. Fuzzy Logic: A practical approach. Academic Press, 1994.

Swarup Medasani, Jaeseok Kim, y Raghu Krishnapuram. An overview of membership function generation techniques for pattern recognition. International Journal of Approximate Reasoning, 19:391–417, 1998.

Witold Pedrycz y Fernando Gomide. An Introduction to Fuzzy Sets: Analysis and Design. MIT Press, Cambridge, MA., 1998.

Thimoty Ross. (1995). Fuzzy Logic with Engineering Applications. New York: McGraw-Hill.

M. Sugeno y G. T. Kang. Structure and identification of fuzzy model. Fuzzy Sets and Systems, 28:15–33, 1988.

T. Takagi y M. Sugeno. Fuzzy identification of systems and its applications to modelling and control. IEEE Transactions on Systems, Man and Cybernetics, 15:116–132, 1985.

I. B. Turksen. Measurement of membership functions and their acquisition.

Fuzzy Sets and Systems, 40:5–38, 1991.

Lotfi A. Zadeh. Fuzzy sets. Information and Control, 8:338–353, 1965.

H.-J. Zimmermann y P. Zysno. Quantifying vagueness in decision models. European Journal of Operational Research, 22:148–158, 1985.