Tema 3. Construcción de redes bayesianas

UGR

Granada, noviembre/diciembre 2024

Índice

1. *Determinación de la estructura*

Al construir un modelo de red bayesiana (u otro tipo de modelo) hemos de considerar el objetivo principal: **obtener estimaciones sobre la probabilidad de ocurrencia de ciertos eventos que no son directamente observables**

Por esta razón, los pasos a seguir deberían ser:

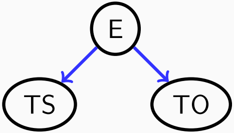
- identificar los eventos de interés o hipótesis. Estos eventos se agrupan en conjuntos mutuamente exclusivos para generar **variables hipótesis**
- identificar las fuentes que proporcionan información sobre las variables hipótesis: **variables informativas**
- establecer las relaciones entre las variables identificadas. Este paso puede requerir el uso de variables mediadoras

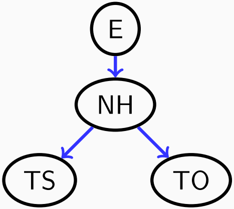
Las variables mediadoras:

- permiten expresar de forma más simple las relaciones entre las variables hipótesis e informativas
- su introducción agrega relaciones de independencia que simplifican el modelo

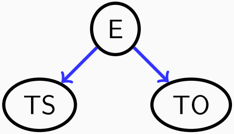
1. *Determinación de la estructura*

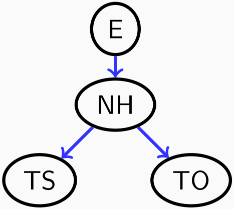
Imaginemos la situación de determinación del estado de embarazo (E), que influye en el resultado de dos test: de sangre (TS) y de orina (TO).





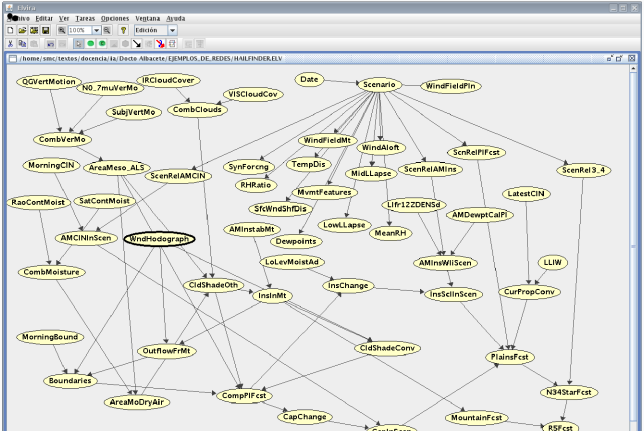
La primera aproximación sería considerar el modelo de la izquierda. pero en caso de conocer el estado de embarazo, entonces las variables asociadas a los test se hacen independientes. Y ocurre que el resultado de estos test puede depender de otras causas, además del embarazo. Por ejemplo, del nivel hormonal (**NH**). Esto justifica la introducción de esta variable mediadora.





Resulta más adecuado considerar el modelo de la derecha, en el que se introduce la variable **mediadora** **NH**. Ahora no se produce la situación anterior: si se conoce el valor de la variable **E** no se hacen independientes las variables de los test.

En otros problemas se agregan variables que representan **escenarios** y que definen situaciones típicas, relacionando valores de otras variables. Ejemplo: *A Bayesian System for forecasting severe weather. B. Abramson et al. 1996*. La variable está en la parte superior y se denomina **Scenario**.



Según el dominio, las variables pueden clasificarse en:

- **discretas**: cuentan con un conjunto finito de posibles valores
- **continuas**: el conjunto de posibles valores es infinito (valores reales)

Las variables continuas pueden convertirse a discretas usando **discretización**: particionando con intervalos de igual longitud, de igual frecuencia o agrupando valores similares... Seguro que habéis estudiado varias técnicas de discretización.

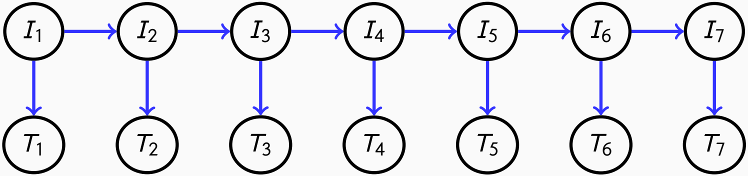
Índice

La leche de una vaca puede estar infectada. Para detectar esta situación se puede realizar un test con resultados positivo - negativo. El test no es perfecto y puede producir falsos positivos y negativos.

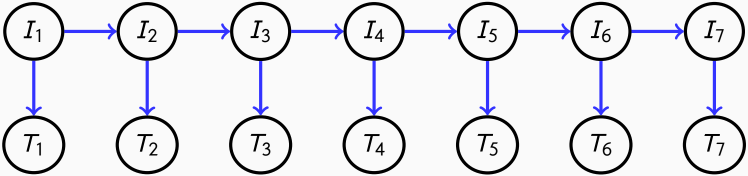
- hipótesis: variable I(nfección), con valores posibles sí, no
- variable informativa: T(test), valores posibles positivo, negativo
- relación entre ellas: de infección a test

$$I \longrightarrow T$$

El estado de la leche puede cambiar de un día a otro: las vacas enfermas sanan y al revés. El granjero realiza el test todos los días. Tras una semana dispone de los resultados de todos los tests. Considerando toda la información disponible podemos modelizar el paso del tiempo y usar la información del pasado para tomar nuevas decisiones.

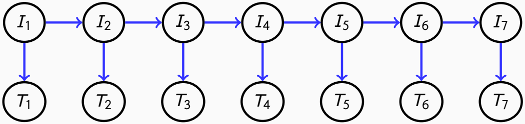


Este tipo de modelos que consideran el paso del tiempo se denominan **redes bayesianas dinámicas**

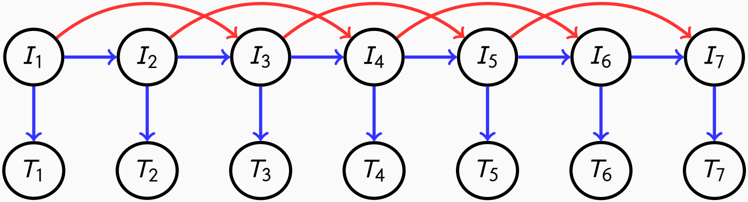


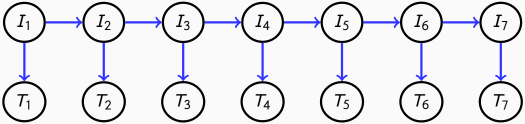
Los supuestos de este modelo son (derivados de las relaciones de **d-separación**):

- propiedad de Markov: si conocemos el presente, entonces el pasado no tiene influencia en el futuro. Más formalmente: $I_{i-1} \perp I_{i+1}|I_i$ (donde $\perp$ representa la d-separación)
- dos nodos de test están d-separados si se dispone de información de alguno de los nodos de infección entre ellos: $T_i \perp T_{i+1}|I_k, k = i \text{ o } k = i+1$

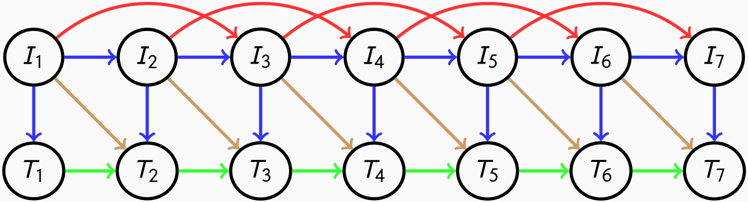


La primera condición, $I_{i-1} \perp I_{i+1} | I_i$, no siempre se cumple en la realidad. Las enfermedades suelen tener un trascurso natural a lo largo de varios días. Por ejemplo, si hoy tengo gripe y estaba sano ayer, entonces lo normal es que mañana también esté enfermo (es decir, el pasado sirve par saber el estado futuro). Por esta razón, el modelo siguiente se ajusta más a la realidad:

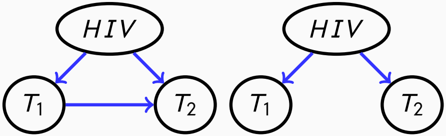




La segunda condición, $T_i \perp T_{i+1} | I_k, k = i \;\; \text{o} \;\; k = i+1$, implica que el fallo en un test no depende del funcionamiento del test previo. Sin embargo, si puede haber lotes defectuosos de pruebas, esto podría no ser así. Para considerar este efecto deberíamos usar:



En relación a los test, es buena idea considerar que no hay test perfectos y asignar probabilidades a ocurrencia de falsos positivos y negativos. Además, habría que considerar la causa del posible fallo. Si la causa de fallo no depende del propio test puede ocurrir que su repetición produzca el mismo resultado (modelo de la izquierda). Si la causa del fallo depende del propio test, sería más adecuado el modelo de la derecha



Índice

Esta mañana he sentido irritación de garganta (**IG**). Puede ser el principio de un resfriado o de anginas. Si las anginas son fuertes pueden producir placas. Ambas enfermedades pueden producir también fiebre

- hipótesis: resfriado sí y no; infección de anginas leve, moderada y fuerte. Se agrupan en dos variables, llamadas **R**(resfriado) y **A**(nginas)
- información: presencia o no de fiebre, irritación de garganta y placas . Se usa una variable para cada síntoma (**F**, **IG**, **P**), con posibles valores sí y no.
- el resfriado puede producir fiebre e irritación de garganta. Las anginas pueden producir los tres síntomas

Indicar las relaciones de independencia que se cumplen en el modelo propuesto.

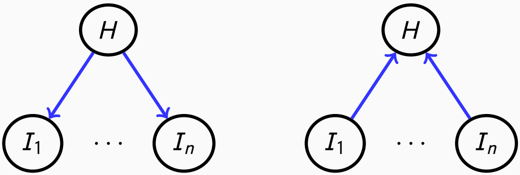
Índice

En un juego de poker el jugador recibe tres cartas y hay dos turnos de descarte. En la primera pueden descartarse todas las que quiera. En la segunda ronda solo pueden descartarse dos cartas como máximo. Estamos interesados en estimar el juego del oponente tras dos descartes. Consideramos las siguientes variables:

- O_0: cartas del oponente al inicio
- O_1: cartas del oponente tras el primer descarte
- O_2: cartas del oponente tras el segundo descarte
- D_1: cartas pedidas en el primer descarte
- D_2: cartas pedidas en el segundo descarte

Índice

Se trata de un modelo clásico usado para diagnóstico y que pese a su simplicidad ha demostrado ser muy útil en diferentes ámbitos (por ejemplo, para clasificación). En él se asume que las variables de información son independientes dado que se conoce el valor de la variable hipótesis.



Usando esta hipótesis, la asignación de valores de probabilidad resulta más sencilla si se considera el modelo de la izquierda (puede generarse el modelo de la derecha mediante cálculo si es necesario.....)

El proceso completo puede esquematizarse de la siguiente forma:

- las posibles causas pueden englobarse en una variable hipótesis H, con probabilidad a priori $P(H)$
- para todas las variables de información I_i se determina la distribución de probabilidad $P(I_i|H)$
- para cualquier combinación de las variables informativas $I_1 \ \dots \ I_n$, el cálculo del producto

$$P(I_1 \ \dots \ I_n|H) = P(I_1|H) \ \dots \ P(I_n|H)$$

- la probabilidad a posteriori de H puede calcularse ahora:

$$P(H|I_1 \ \dots \ I_n) = \mu P(H)P(I_1 \ \dots \ I_n|H) =$$

$$\mu P(H) \prod_{i=1}^{n} p(I_i|H)$$

donde μ es una constante de normalización:

$$\mu = \frac{1}{P(I_1 \ \ldots. \ I_n)}$$

- lo interesante de este modelo radica en su simplicidad, ya que la complejidad es lineal en el número de variables informativas
- este modelo suele funcionar bien aunque se viole el supuesto de independencia entre las variables informativas dado el valor de las variables hipótesis
- esto se debe a que en muchos problemas de diagnóstico el objetivo consiste en identificar la causa más probable y normalmente este cálculo no se verá afectado por el supuesto de independencia condicional

Índice

En una granja de cría de caballos se dan las siguientes relaciones:

- del cruce de Ann y Brian ha nacido Dorothy
- del cruce de Brian y Cecily ha nacido Eric
- Dorothy y Fred son los padres de Henry
- Eric y Gwenn son padres de Irene
- Ann es la madre de Fred y Gwenn, pero sus padres no están relacionados
- el potro John nace de Henry e Irene

John sufre un trastorno hereditario producido por un gen recesivo. Para evitar que nazcan más caballos con problemas, Henry e Irene se descartan para nuevos cruces. Se desea determinar la probabilidad de que el gen esté presente en el resto de caballos.

La única variable informativa es John. Los genotipos posibles son:

- aa - enfermo
- aA - portador
- AA - no portador

Las variables hipótesis son las que corresponden al resto de caballos. Para todos los caballos excepto John también se sabe que su genotipo no puede ser aa porque se habría puesto de manifiesto la enfermedad.

Para poder tratar con Fred y Gwenn es necesario incluir variables X y Y que representan a los padres de estos caballos, como variables mediadoras, y se asume que estos caballos no están enfermos.

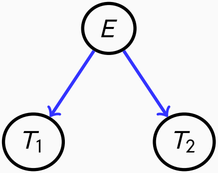
Índice

Un lenguaje con dos símbolos a y b se transmite mediante un canal. Cada palabra se delimita mediante el símbolo c. El canal puede introducir ruido. Se transmite una palabra de 5 letras. Se trata de construir un modelo para este problema

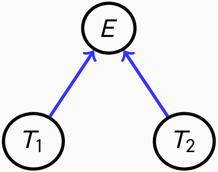
- hay 5 variables hipótesis T_1 .... T_5 con estados a y b y 5 variables informativas R_1 .... R_5 con estados a, b y c (que representan lo que realmente se ha recibido)
- también hay una relación T_i a T_{i+1}, que codifica que se sabe que algunos pares de símbolos son más frecuentes que otros

Índice

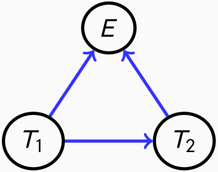
Ya se ha indicado que la estructura de una red bayesiana no tiene por qué reflejar obligatoriamente relaciones causales. El único requisito que debe cumplirse tiene que ver con la *d-separación*. De todas formas, conviene tratar de representar relaciones causales si están realmente presentes. Consideremos una enfermedad E y dos test T_1 y T_2



Cuando se hace diagnóstico, se suele razonar en dirección opuesta y suele ser normal considerar el siguiente modelo en su lugar



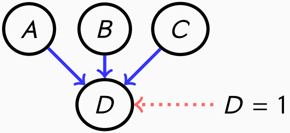
Sin embargo, este modelo implica que T_1 y T_2 son independientes entre sí, lo que no es cierto viendo el primer modelo sobre este problema. Para evitar esta situación habría que complicar el modelo agregando enlaces adicionales:



Índice

Índice

Puede ocurrir que el modelo contenga relaciones entre varias variables A, B y C, por ejemplo, pero no es posible asignar dirección a las relaciones entre estas variables. Esta dificultad puede resolverse a veces considerando una nueva variable mediadora D.

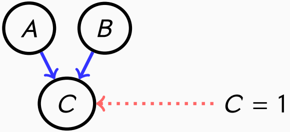


Imaginemos que las variables describen posibles configuraciones y D posee dos valores 0 y 1 y representa el valor asignado a las configuraciones, de la forma:

$$P(D = 1|A, B, C) = R(A, B, C)$$
$$P(D = 0|A, B, C) = 1 - R(A, B, C)$$

Como ejemplo, consideremos el caso en que A y B deben estar en el mismo estado; C es ahora la variable mediadora y la relación asociada a C sería:

C	$C=1$		$C=0$	
B	$B=1$	$B=0$	$B=1$	$B=0$
$A=1$	1	0	0	0
$A=0$	0	0	0	1



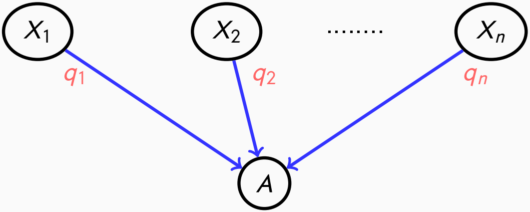
Índice

Imaginemos una variable A con varios padres. Entonces hay que especificar una distribución para cada configuración $\boldsymbol{c}$ de los padres: $P(A|\boldsymbol{c})$. Podría ocurrir:

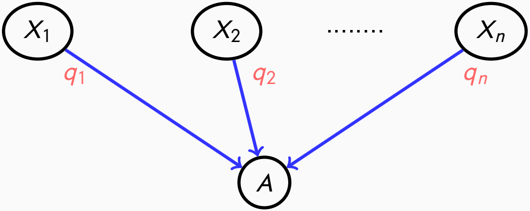
- que algunas configuraciones sean difíciles de considerar o bien haya pocos datos para obtener las probabilidades asociadas
- puede que sea fácil asignar $P(A|X_i)$ (para algunos de los padres), pero difícil de asignar la relación conjunta $P(A|X_1 \ldots. X_n)$

En estos casos conviene disponer de estrategias que permitan reducir el número de distribuciones a especificar. Una de ellas se denomina **noisy-or** y puede definirse de la siguiente forma:

- sean X_1 .... X_n las variables binarias que indican las causas que pueden producir un determinado efecto A
- cada evento $X_i = 1$ es susceptible de causar por sí mismo el efecto, a menos que un inhibidor lo impida, con probabilidad q_i



$$P(A = 0|X_i = 1) = q_i$$



Se asume que todos los inhibidores actúan de forma independiente:

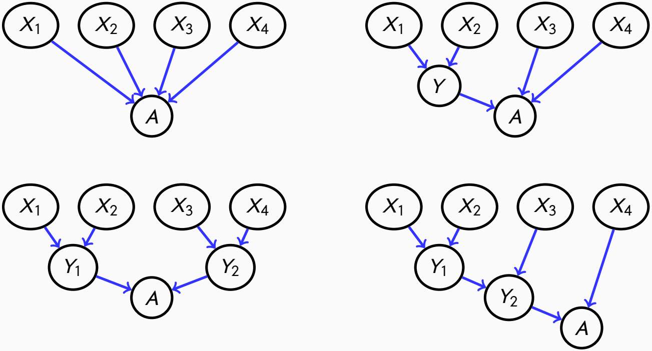
$$P(A = 0|X_1 = 1, \; .... \; X_j = 1) = \prod_{i \in \boldsymbol{Y}} q_i$$

donde $\boldsymbol{Y}$ es el conjunto de índices para las variables con estado igual a 1. Por ejemplo:

$$P(A = 0|X_1 = 1, X_2 = 1, X_3 = 0, \; .... \; X_n = 0) = q_1 q_2$$
$$P(A = 1|X_1 = 1, X_2 = 1, X_3 = 0, \; .... \; X_n = 0) = 1 - q_1 q_2$$

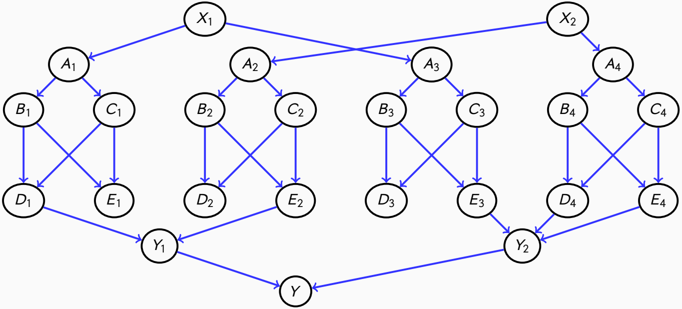
Índice

Sea $X_1, \ldots. X_n$ una lista de variables que causan el efecto A. Se desea asignar probabilidades a la distribución $P(A|X_1, \ldots. X_n)$, lo que podría ser muy difícil de conseguir. Para manejar esta situación suele aplicarse una técnica conocida como **divorcio de padres**, introduciendo variables mediadoras y reduciendo el conjunto de padres de la variable A.



Índice

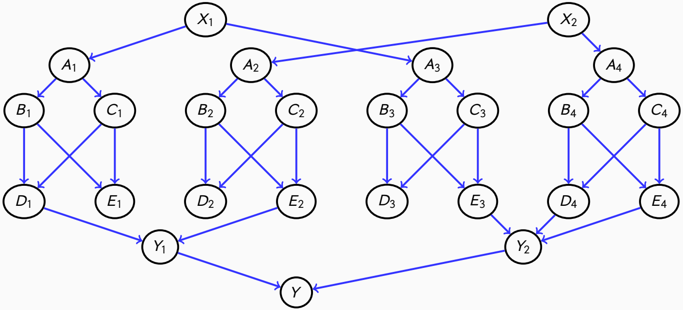
En redes complejas suele haber copias de fragmentos casi idénticos que se repiten varias veces- Supongamos el modelo siguiente:

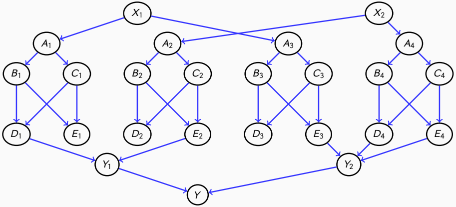


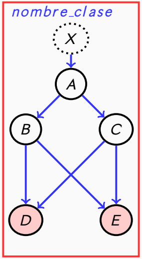
Se asume que:

- X_1 y X_2 tienen el mismo conjunto de estados
- que la distribución de probabilidad de las variables A_1, A_2, A_3 y A_4 es idéntica
- lo mismo ocurre con las distribuciones de B_i, C_i, D_i y E_i

Asumiendo esas condiciones, se observa que la red tiene 4 copias del mismo fragmento, que contiene las variables A_i, B_i, C_i, D_i y E_i. Esto puede aprovecharse, creando una plantilla genérica para el fragmento repetido, que se denomina **clase**, siguiendo la terminología propia de orientación a objetos. Cada uso del patrón se denominará **objeto** o **instancia**







- para poder especificar la distribución de probabilidad de la variable A se incluye el nodo artificial X, con el mismo conjunto de estados que X_1 y X_2
- los nodos D y E se corresponden con la parte del objeto que es accesible desde el exterior, ya que pueden ser padres de variables fuera del objeto.
- los nodos A, B y C son transparentes para el resto del modelo

Haciendo uso de la clase definida antes el modelo completo quedaría de la siguiente forma: