Aprendizaje Automatizado

Tarea 4: Gráficos Probabilísticos

PCIC - UNAM

20 de mayo de 2020

Diego de Jesús Isla López
(*dislalopez@gmail.com*)
(*diego.isla@comunidad.unam.mx*)

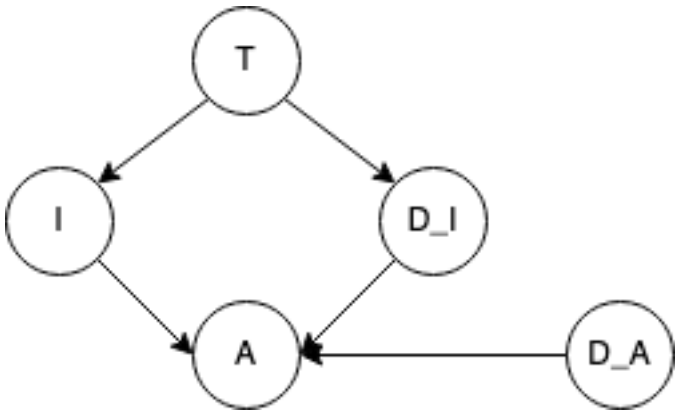
Ejercicio 1

Se tienen las siguientes variables:

Variable	Dominio
A (alarma)	{0,1}
D_A (alarma defectuosa)	{0,1}
D_I (indicador defectuoso)	{0,1}
I (lectura del indicador)	$\mathbb{Z}$
T (temperatura real)	$\mathbb{Z}$

Primer modelo

La gráfica para el primer modelo es la siguiente:



La probabilidad conjunta del modelo queda expresada como:

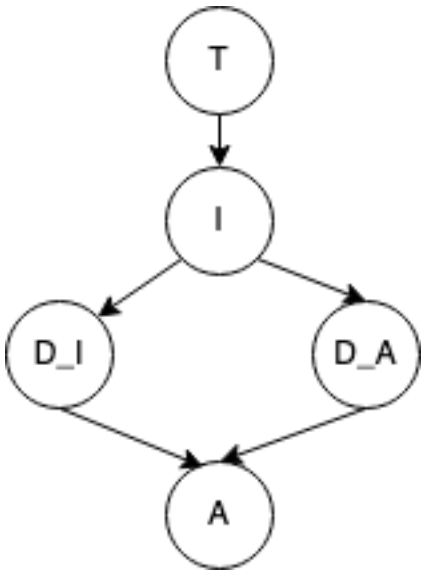
$$P(T,I,D_I,A,D_A) = P(T)\cdot P(I|T)\cdot P(A|I)\cdot P(D_I|T)\cdot P(A|D_A,D_I,I)\cdot P(D_A) \qquad (1)$$

En este modelo se busca representar el efecto de la temperatura sobre las variables I y D_I, siendo A influenciada por estas dos además de D_A.

Calculando el número de variables tenemos:
- $P(T) = 100$
- $P(D_A) = 2$
- $P(I|T) = 100 * 100 = 10000$
- $P(D_I|T) = 2 * 100 = 200$
- $P(A|D_A, D_I, I) = 2 * 100 * 2 * 2 = 800$
- Total = 11102

Segundo modelo

La gráfica para el segundo modelo es la siguiente:



La probabilidad conjunta del modelo queda expresada como:

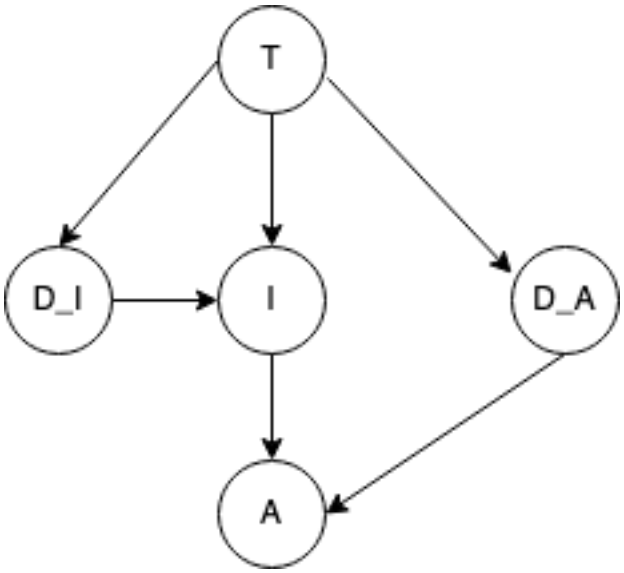
$$P(T, I, D_I, A, D_A) = P(T) \cdot P(I|T) \cdot P(D_A|I) \cdot P(D_I|I) \cdot P(A|D_A, D_I) \tag{2}$$

Este modelo es similar al anterior, pero en este caso I influye directamente sobre D_A y D_I y estas a su vez influyen directamente en A.
Calculando el número de variables tenemos:
- $P(T) = 100$
- $P(I|T) = 100 * 100 = 10000$
- $P(D_I|I) = 2 * 100 = 200$
- $P(D_A|I) = 2 * 100 = 200$

- $P(A|D_A, D_I) = 2 * 2 * 2 = 8$
- Total = 10508

Tercer modelo

La gráfica para este modelo es la siguiente:



La probabilidad conjunta del modelo queda expresada como:

$$P(T, I, D_I, A, D_A) = P(T) \cdot P(D_I|T) \cdot P(I|D_I T) \cdot P(D_A|T) \cdot P(A|I, D_A) \tag{3}$$

En este modelo se representa la influencia de la temperatura sobre la lectura del indicador, además de la posibilidad de que tanto la alarma y el indicador estén defectuosos, influyendo sobre la alarma a través de la lectura del indicador y la posibilidad de defecto de la alarma.
Calculando el número de variables tenemos:

- $P(T) = 100$
- $P(I|D_I, T) = 100 * 100 * 2 = 20000$
- $P(D_I|T) = 2 * 100 = 200$
- $P(D_A|T) = 2 * 100 = 200$
- $P(A|I, D_A) = 2 * 100 * 2 = 400$
- Total: 20900

Ejercicio 2

Para cada caso:

$T \perp\!\!\!\perp F|D$

Se tienen los caminos $\{T,O,C,F\}$ y $\{T,O,D,B,F\}$.

Para el camino $\{T,O,C,F\}$ tenemos las tripletas $\{T,O,C\}$ y $\{O,C,F\}$. La tripleta $\{T,O,C\}$ es activa ya que es una causa común de la que deriva la variable observada D. La tripleta $\{O,C,F\}$ representa una cadena causal y dado que ninguna de sus variables es observada, se considera como activa.

El camino $\{T,O,D,B,F\}$ es bloqueado dado que contiene a la variable observada, sin embargo ninguna de las tripletas del camino anterior resultan afectadas.

Por lo tanto, se tiene que $T \not\perp\!\!\!\perp F|D$.

$C \perp\!\!\!\perp B|F$

Se tienen los caminos $\{C,F,B\}$ y $\{C,O,D,B\}$.

El camino $\{C,F,B\}$ es inactivo ya que es una causa común derivada de la variable observada F.

El camino $\{C,O,D,B\}$ tiene las tripletas $\{C,O,D\}$ y $\{O,D,B\}$. La tripleta $\{C,O,D\}$ está activa ya que representa una cadena causal donde no hay variables observadas. Por su parte, la tripleta $\{O,D,B\}$ representa un efecto común donde ninguna variable es observada, por lo que se considera inactiva. Ya que solo una de las tripletas está activa, el camino se considera inactivo.

Dado que no existen caminos activos, se sigue que $C \perp\!\!\!\perp B|F$

$A \perp\!\!\!\perp F|C$

Se tienen los caminos $\{A,T,O,C,F\}$ y $\{A,T,O,D,B,F\}$
Para el camino $\{A,T,O,C,F\}$ se tienen las tripletas $\{A,T,O\}$ y $\{O,C,F\}$. La tripleta $\{A,T,O\}$ es una cadena causal que no tiene variables observadas, por lo que está activa. La tripleta $\{O,C,F\}$ es una cadena causal donde se observa la variable C por lo cual es inactiva. Por lo tanto, este camino es inactivo.

El camino $\{A,T,O,D,B,F\}$ cuenta con las tripletas $\{A,T,O\}$, $\{O,D,B\}$ y $\{D,B,F\}$. La tripleta $\{O,D,B\}$ es inactiva ya que representa una causa común sin variables observadas.
Como ambos caminos son inactivos, se sostiene que $A \perp\!\!\!\perp F|C$.

$$A \perp\!\!\!\perp F | C, D$$

Se tienen nuevamente los caminos $\{A, T, O, C, F\}$ y $\{A, T, O, D, B, F\}$.

El camino $\{A, T, O, D, B, F\}$ es activo puesto que todas las tripletas son activas. Las tripletas $\{A, T, O\}$ y $\{T, O, D\}$ son cadenas causales sin variables observadas. La tripleta $\{D, B, F\}$ también representa una cadena causal activa, aunque contiene a la variable observada D, ya que esta se encuentra al final de la cadena.

Dado que existe al menos un camino activo, se sigue que $A \not\perp\!\!\!\perp F | C, D$

Ejercicio 3

Escribe la distribución conjunta de la red bayesiana en función de las probabilidades condicionales

La distribución conjunta de la red queda expresada como:

$$P(E, F, S, V, C, D) = P(E) \cdot P(F|E) \cdot P(S|E) \cdot P(V|F, S) \cdot P(C|S) \cdot P(D|V) \tag{4}$$

Si un paciente es llevado al doctor, usando un paquete de software calcula la probabilidad de que no tenga ébola

Para este ejercicio se utiliza la biblioteca pgmpy. El código fuente se encuentra adjunto a este documento.

El resultado encontrado es:

Ebola	P(Ebola)
Ebola = T	0.0752
Ebola = F	0.9248

Por lo tanto, la probabilidad de que el paciente no tenga ébola dado que fue llevado al doctor es de 92.48 %.

Convierte la red bayesiana en un modelo gráfico no dirigido (campo aleatorio de Markov) y dibújalo. Captura tantas relaciones de independencia condicional como sea posible

El modelo queda expresado por el siguiente campo aleatorio de Markov:

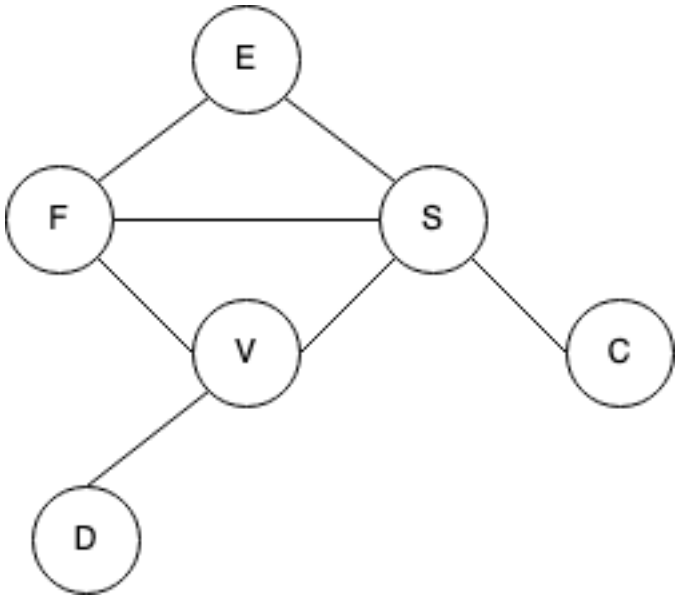


Figura 1: Representación del problema mediante un campo aleatorio de Markov

Las probabilidades condicionales que podemos encontrar en el modelo son:

- Si S es observada, podemos ver que C queda desconectada de la gráfica, teniendo así:
  1. $E \perp\!\!\!\perp C|S$
  2. $F \perp\!\!\!\perp C|S$
  3. $V \perp\!\!\!\perp C|S$
  4. $D \perp\!\!\!\perp C|S$
- Si V es observada, podemos ver que D queda desconectada de la gráfica, teniendo así:
  1. $E \perp\!\!\!\perp D|V$
  2. $F \perp\!\!\!\perp D|V$
  3. $S \perp\!\!\!\perp D|V$
  4. $C \perp\!\!\!\perp D|V$
- Finalmente si S y F son observadas, podemos ver que E y C quedan desconectadas de la gráfica, teniendo así:
  1. $V \perp\!\!\!\perp C|F,S$
  2. $D \perp\!\!\!\perp C|F,S$
  3. $V \perp\!\!\!\perp E|F,S$
  4. $D \perp\!\!\!\perp E|F,S$
  5. $E \perp\!\!\!\perp C|F,S$
  6. $C \perp\!\!\!\perp E|F,S$

Debido a una campaña de concientización de la salud, las personas son alentadas a visitar la clínica en caso de que tengan fiebre. Esto incrementa la cantidad de visitas de personas con fiebre sin importar el estado de cualquier otra variable.

1. ¿Qué probabilidades condicionales en la red se modifican debido a este cambio y en qué sentido?

Hay probabilidades que aumentarán y otras que disminuirán. Las probabilidades que aumentan son:

- $P(V = \texttt{verdadero} | F = \texttt{verdadero}, S = \texttt{verdadero})$
- $P(V = \texttt{verdadero} | F = \texttt{verdadero}, S = \texttt{falso})$

Esto es debido a que las visitas aumentarán independientemente de la presencia de sangrado en el paciente. Por ende, disminuirá la probabilidad de que el paciente no asista a la clínica a pesar de tener fiebre ($P(V = \texttt{falso} | F = \texttt{verdadero}, S)$).

2. Describe cualquier otro efecto que esto tenga en la proporción de personas con complicaciones que visiten la clínica. Menciona exactamente qué probabilidades condicionales usaste para llegar a tu conclusión.

Observando el modelo, la aparición de complicaciones depende directamente de la presencia de sangrado en el paciente. Esto es:

- $P(C = \texttt{verdadero} | S = \texttt{verdadero}) = 0.75$
- $P(C = \texttt{verdadero} | S = \texttt{verdadero}) = 0.1$

De esta manera podemos ver que la diferencia es significativa para los casos en los que existe sangrado y los que no. Sin embargo, se espera que las visitas aumenten debido a la campaña, la cual está basada en la presencia de fiebre. Dado que los demás síntomas son irrelevantes para la campaña, se espera que aumente el número de pacientes que presentan tanto fiebre como sangrado.

Asume que alguien que no tiene fiebre va al doctor, ¿qué relación de independencia condicional existe en la distribución que no puede ser descubierta a través del grafo solamente?

Poniendo el enunciado en términos de independencia condicional, buscamos demostrar que:

$$E \perp\!\!\!\perp V | F = \texttt{falso}, D = \texttt{verdadero} \tag{5}$$

Observando el gráfico, es posible ver que entre E y V) existen dos caminos, los cuales consisten de una sola tripleta cada uno: $\{E, F, V\}$ y $\{E, S, V\}$. Ambos caminos son cadenas causales; de esta manera el camino $\{E, F, V\}$ es inactivo puesto que la variable F es

observada y está en el medio. Por su parte, el camino $\{E, S, V\}$ es activo puesto que no contiene ninguna variable observada. Entonces, se concluye que no es posible demostrar este caso de independencia mediante el modelo gráfico.

Usando marginalización tenemos que:

$$\begin{aligned} P(E, V | F = \texttt{falso}, D = \texttt{verdadero}) = & P(E | F = \texttt{falso}, D = \texttt{verdadero}) \\ & \cdot P(V | \texttt{falso}, D = \texttt{verdadero}) \end{aligned}$$

Esto es, $\forall e, v \in \{\texttt{verdadero}, \texttt{falso}\}$:

$$\begin{aligned} P(e, v | \texttt{falso}, D = \texttt{verdadero}) = & P(e | F = \texttt{falso}, D = \texttt{verdadero}) \\ & \cdot P(v | \texttt{falso}, D = \texttt{verdadero}) \end{aligned}$$

Utilizando pgmpy encontramos la solución:

Ebola	P(Ebola)
Ebola = T	0.0670
Ebola = F	0.9330

Visita	P(Visita)
Visita = T	1
Visita = F	0

Ebola,Visita	P(Ebola,Visita)
Ebola = T,Visita = T	0.0670
Ebola = F, Visita = T	0
Ebola = T, Visita = F	0.9330
Ebola = F, Visita = F	0

Problema opcional

Sabemos que la probabilidad condicional está expresada por:

$$P(E, F, S, V, C, D) = P(E) \cdot P(F|E) \cdot P(S|E) \cdot P(V|F, S) \cdot P(C|S) \cdot P(D|V)$$

De este modo, las variables ocultas para este problema son F, S, C y D. Primero obtenemos las tablas de valores para cada variable:

E	P(E)
E = T	0.01
E = F,	0.99

Tabla 1: Valores para E

E,F	P(F|D)
E = verdadero, F = verdadero	0.6
E = verdadero, F = falso	0.4
E = falso, F = verdadero	0.1
E = falso, F = falso	0.9

Tabla 2: Valores para F|E

E,S	P(S|E)
E = verdadero, S = verdadero	0.8
E = verdadero, S = falso	0.2
E = falso, S = verdadero	0.05
E = falso, S = falso	0.95

Tabla 3: Valores para S|E

F,S,V	P(V|F,S)
F = verdadero, S = verdadero, V = verdadero	0.8
F = verdadero, S = verdadero, V = falso	0.2
F = verdadero, S = falso, V = verdadero	0.5
F = verdadero, S = falso, V = falso	0.5
F = falso, S = verdadero, V = verdadero	0.7
F = falso, S = verdadero, V = falso	0.3
F = falso, S = falso, V = verdadero	0
F = falso, S = falso, V = falso	1

Tabla 4: Valores para V|F,S

S,C	P(C|S)
S = verdadero, C = verdadero	0.75
S = verdadero, C = falso	0.25
S = falso, C = verdadero	0.1
S = falso, C = falso	0.9

Tabla 5: Valores para C|S

V,D	P(D|V)
V = verdadero, D = verdadero	0.6
V = verdadero, D = falso	0.4
V = falso, D = verdadero	0
V = falso, D = falso	1

Tabla 6: Valores para D|V

Marginalizando las variables ocultas, la probablidad se expresa como:

$$P(V, E = \text{verdadero}) = \sum_f \sum_s \sum_c \sum_d [P(E = \text{verdadero}) P(f|E = \text{verdadero}) P(s|E = \text{verdadero}) P(V|f,s) P(c|s) P(d|V)]$$

Reformulando, se obtiene:

$$P(V, E = \text{verdadero}) = P(E = \text{verdadero})[\sum_f \sum_s [P(f|E = \text{verdadero}) P(s|E = \text{verdadero}) P(V|f,s) \sum_c [P(c|s)] \sum_d [P(d|V)]]]$$

Para eliminar D obtenemos $F_d(V) = \sum_d [P(d|V)]$:

V	$F_d(V)$
verdadero	0.6 + 0.4 = 1
falso	0 + 1 = 1

Tabla 7: Valores para $F_d(V)$

En este paso, la ecuación queda:

$$P(V, E = \texttt{verdadero}) = \\ P(E = \texttt{verdadero})[\sum_f \sum_s [P(f|E = \text{verdadero})P(s|E = \text{verdadero})P(V|f,s) \sum_c [P(c|s)]F_d(V)]]$$

Para eliminar C obtenemos $F_c(s) = \sum_c [P(c|s)]$:

S	$F_c(s)$
verdadero	0.75 + 0.25 = 1
falso	0.1 + 0.9 = 1

Tabla 8: Valores para $F_c(s)$

En este paso, la ecuación queda:

$$P(V, E = \text{verdadero}) = \\ P(E = \text{verdadero})[\sum_f [P(f|E = \text{verdadero}) \sum_s [P(s|E = \text{verdadero})P(V|f,s)F_c(s)]F_d(V)$$

Para eliminar S obtenemos $F_s(V, f) = \sum_s [P(f|E = \text{verdadero})P(s|E = \text{verdadero})P(V|f,s)F_c(s)]$.
Construyendo $F_s(V, f)$:

S	V	F	$P(s\|E = \text{verdadero})P(V\|f,s)F_c(s)$
verdadero	verdadero	verdadero	0.64
verdadero	verdadero	falso	0.56
verdadero	falso	verdadero	0.16
verdadero	falso	falso	0.24
falso	verdadero	verdadero	0.1
falso	verdadero	falso	0
falso	falso	verdadero	0.1
falso	falso	falso	0.2

Tabla 9: Construcción $F_s(V, f)$

Así, calculamos:

V	F	$F_s(V,f)$
verdadero	verdadero	0.74
verdadero	falso	0.56
falso	verdadero	0.26
falso	falso	0.44

Tabla 10: Valores para $F_s(V,f)$

En este paso, la ecuación queda:

$$P(V,E=\texttt{verdadero}) = P(E=\texttt{verdadero})[\sum_f [P(f|E=\text{verdadero})F_s(V,f)]F_d(V)]$$

Para eliminar F obtenemos $F_f(V) = \sum_f [P(f|E=\text{verdadero})F_s(V,f)]$. Construyendo, tenemos:

F	V	$P(f\|E=\texttt{verdadero})F_s(V,f)$
verdadero	verdadero	0.444
verdadero	falso	0.156
falso	verdadero	0.224
falso	falso	0.176

Tabla 11: Construcción para $F_f(V)$

Así, el valor para $F_f(V)$ queda:

V	$F_f(V)$
verdadero	0.668
falso	0.332

Tabla 12: Valores para $F_f(V)$

Finalmente la ecuación queda:

$$P(V,E=\texttt{verdadero}) = P(E=\texttt{verdadero})[F_f(V)F_d(V)] \tag{6}$$

Entonces calculamos la probabilidad:

V	$P(E = \texttt{verdadero})[F_f(V)F_d(V)]$
verdadero	0.00668
falso	0.00332

Tabla 13: Valores para $P(V, E = \texttt{verdadero})$

Así, se tiene:

$$P(V = \text{verdadero}|E = \text{verdadero}) = \frac{0.00668}{0.00668 + 0.00332} = 0.668$$

$$P(V = \text{falso}|E = \text{verdadero}) = \frac{0.00332}{0.00668 + 0.00332} = 0.332$$