APR: Práctica sobre Modelos Gráficos Probabilísticos

Roselló Beneitez, N. U.; Roselló Oviedo, M.

6 de Enero de 2020

Índice

Índice de figuras

1. Descripción de la práctica

En esta práctica se ha hecho uso de la librería *BNT* (*Bayes Net Toolbox*) implementada en *matlab* con el fin de trabajar con redes bayesianas, un subconjunto de los modelos gráficos visto en teoría.

En primer lugar, se han hecho experimentos con una red de nodos booleanos que representa un césped que puede estar húmedo por causa del riego y/o la lluvia, disponiendo también de un nodo representando si el cielo se encuentra nublado.

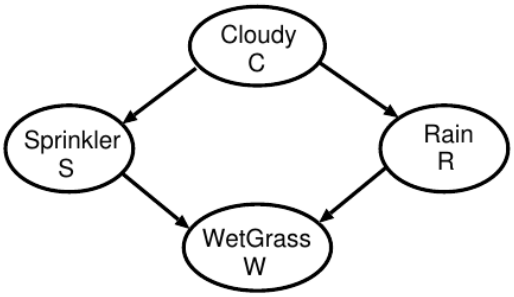

Figura 1: Red bayesiana para el modelo *Sprinkler*.

Para el segundo ejercicio, hemos trabajado con la red vista en los ejemplos de la teoría representando el diagnóstico de cáncer de pulmón y/o disnea en un paciente sometido a rayos X que puede ser o no fumador y puede haber sufrido o no de polución.

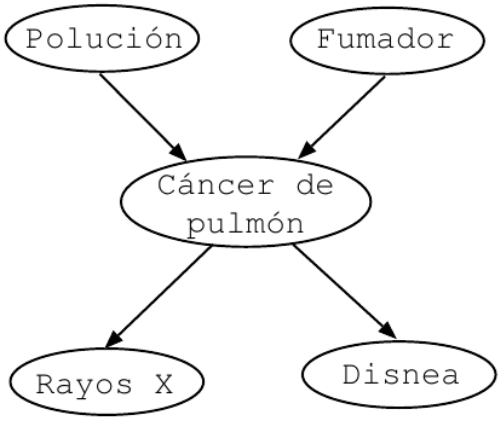

Figura 2: Red bayesiana para el modelo *Cáncer de pulmón*.

2. Ejercicio A

En este ejercicio se han comparado los resultados del aprendizaje a partir de datos completos (izquierda) e incompletos (derecha):

```
Datos completos
W:
1 1 : 1.0000 0.0000
2 1 : 0.0556 0.9444
1 2 : 0.0435 0.9565
2 2 : 0.0000 1.0000
S:
1 : 0.5532 0.4468
2 : 0.9057 0.0943
R:
1 : 0.7234 0.2766
2 : 0.2264 0.7736
C:
1 : 0.4700
2 : 0.5300
```

```
Datos incompletos
W:
1 1 : 0.9998 0.0002
2 1 : 0.0136 0.9864
1 2 : 0.1177 0.8823
2 2 : 0.0048 0.9952
S:
1 : 0.5556 0.4444
2 : 0.9999 0.0001
R:
1 : 0.5518 0.4482
2 : 0.2445 0.7555
C:
1 : 0.5425
2 : 0.4575
```

Figura 3: Resultados con datos completos e incompletos para la red bayesiana del modelo *Sprinkler*.

Los resultados de los dos conjuntos de datos son muy parecidos. La única diferencia significativa observable (mayor al 10 %) está en los valores de *lluvioso* (R) para el valor *falso* (no llueve): cuando no está nublado, la probabilidad de que no llueva es del 55.18 % en los datos incompletos, mientras que esta misma probabilidad es del 72.34 % para los datos completos.

Planteamos dos hipótesis para justificar este fenómeno: o bien se han ocultado muchos valores de *lluvioso* dado un valor *falso* de *nublado* (C), en cuyo caso la red bayesiana no sabría si clasificar las muestras en verdadero o falso; o bien se han ocultado muchos de los valores *nublado* = *falso* y algunos de los valores *nublado* = *verdadero*, por lo que los valores de *lluvioso* dado *nublado* = *falso* podrían tomarse como la distribución *a priori* de *lluvioso*.

Para seguir este razonamiento, cabe mencionar que si se oculta el valor de *nublado* para todas las muestras (que son equiprobables), la probabilidad de *lluvioso* = *verdadero* sería igual (o bastante parecida, como es el caso) a la probabilidad de *lluvioso* = *falso*.

Ambos razonamientos se basan en el hecho de que faltan muestras que tengan los valores de *nublado* = *falso* o de *lluvioso* = *falso*, pudiendo darse también una combinación de ambos.

3. Ejercicio B

El script *matlab* para la red de diagnóstico de cáncer de pulmón se adjuntará, comentado y listo para su ejecución, junto con esta memoria. Mediante su utilización, se ha podido responder a las siguientes cuestiones:

- ¿Cuál es la probabilidad de que un paciente no fumador no tenga cáncer de pulmón si la radiografía ha dado un resultado negativo pero sufre disnea?

  El resultado obtenido por *matlab* tras programar la red e insertar las evidencias es:

  $$P\left(C = n | F = n, X = n, D = s\right) = 0.9989 = 99.89\,\%$$

  Podemos comprobar dicho resultado manualmente. A partir de la estructura de la red sabemos que:

  $$P(P, F, C, X, D) = P(P)\,P(F)\,P(C \mid P, F)\,P(X \mid C)\,P(D \mid C)$$

  Por tanto, se puede deducir:

  $$\begin{aligned} P(C \mid F, X, D) &= \frac{P(C, F, X, D)}{P(F, X, D)} \\ &= \frac{\sum\limits_{p\prime \in \{b,a\}} P(P = p\prime)\,P(F)\,P(C \mid P = p\prime, F)\,P(X \mid C)\,P(D \mid C)}{\sum\limits_{c \in \{n,p\}} \sum\limits_{p\prime \in \{b,a\}} P(P = p\prime)\,P(F)\,P(C = c \mid P = p\prime, F)\,P(X \mid C = c)\,P(D \mid C = c)} \\ &= \frac{P(X \mid C)\,P(D \mid C) \sum\limits_{p\prime \in \{b,a\}} P(P = p\prime)\,P(C \mid P = p\prime, F)}{\sum\limits_{c \in \{n,p\}} \left(P(X \mid C = c)\,P(D \mid C = c) \sum\limits_{p\prime \in \{b,a\}} P(P = p\prime)\,P(C = c \mid P = p\prime, F) \right)} \end{aligned}$$

  Aplicado a nuestro caso particular:

  $$\begin{aligned} &P\left(C = n | F = n, X = n, D = s\right) \\ &= \frac{P(X = n \mid C = n)\,P(D = s \mid C = n) \sum\limits_{p\prime \in \{b,a\}} P(P = p\prime)\,P(C = n \mid P = p\prime, F = n)}{\sum\limits_{c \in \{n,p\}} \left(P(X = n \mid C = c)\,P(D = s \mid C = c) \sum\limits_{p\prime \in \{b,a\}} P(P = p\prime)\,P(C = c \mid P = p\prime, F = n) \right)} \\ &= \frac{0.8 \cdot 0.3 \cdot (0.9 \cdot 0.999 + 0.1 \cdot 0.95)}{0.1 \cdot 0.65 \cdot (0.9 \cdot 0.001 + 0.1 \cdot 0.05) + 0.8 \cdot 0.3 \cdot (0.9 \cdot 0.999 + 0.1 \cdot 0.95)} \\ &= 0.998395 = 99.84\,\% \approx 99.89\,\% \end{aligned}$$

- ¿Cuál es la explicación más probable de que un paciente sufra cáncer de pulmón?

  $$\{(1)\} \quad \{(2)\} \quad \{(2)\} \quad \{(3)\} \quad \{(2)\}$$

  Lo cual se traduce en: polución **baja**, fumador **sí**, cáncer **positivo**, rayos X **positivo** y disnea **sí**. El log-verosimilitud de esta explicación es de -5.0925, esto es, una probabilidad de $e^{-5.0925} = 0.0061 = 0.61\,\%$.

4. Conclusiones

Las redes bayesianas nos permiten representar gráficamente y de una forma compacta distribuciones de probabilidad conjunta. Su capacidad de aprendizaje e inferencia nos manifiesta su utilidad como herramienta potencial para múltiples aplicaciones como pueden ser las predicciones o el diagnóstico médico (como hemos visto con estos ejemplos); así como para otros campos como pueden ser la visión por computador o el procesado de lenguajes naturales.