

Práctica 4

Modelos Gráficos Probabilísticos

Vicente Gras Mas

Introducción

El objetivo de esta práctica es experimentar con Modelos Gráficos Probabilísticos, también conocidos como Redes Bayesianas. Para ello utilizaremos el toolkit BNT para matlab.

Haremos uso de un ejemplo simple llamado "Sprinkler" el cual es una red bayesiana simple que tiene información sobre la relación entre Nublado, que un aspersor esté encendido, que llueva y que esté el césped húmedo. Luego vamos a explorar las capacidades de aprendizaje principales que ofrece BNT: aprendizaje con datos completos y aprendizaje con datos incompletos mediante Esperanza Maximización (EM).

Por último haremos un par de ejercicios sobre lo que se ha visto en la práctica, que son los que recoge este boletín.

Ejercicio A

En el ejemplo "Sprinkler", repetir los procesos de aprendizaje a partir de datos completos y de datos incompletos explicados en la Sec.2.3, usando un número mucho mayor de muestras de aprendizaje (por ejemplo, $n_{\text{Muestras}} = 1000$), así como permitiendo un mayor número de iteraciones. Comentar los resultados obtenidos.

Con los resultados obtenidos de los scripts de matlab DatosCompletos.m y DatosIncompletos.m hemos desarrollado las siguientes tablas:

Datos Completos

Nº Muestr	1000		5000		10000		15000		20000	
Original										
f t	f	t	f	t	f	t	f	t	f	t
1.0 0.0	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000	0.0000
0.1 0.9	0.1085	0.8915	0.0966	0.9034	0.1013	0.8987	0.1063	0.8937	0.1081	0.8919
0.1 0.9	0.0905	0.9095	0.1041	0.8959	0.1048	0.8952	0.1059	0.8941	0.1069	0.8931
0.01 0.99	0.0250	0.975	0.0150	0.985	0.0108	0.989	0.0129	0.987	0.0130	0.987

En la tabla anterior podemos ver la probabilidad de que el césped este húmedo, siendo sus precedentes que esté lloviendo o los aspersores activados.

Como podemos observar, con un número alto de muestras (1000), ya se parece mucho a la original, se pueden apreciar pocas variaciones con el aumento del número de iteraciones.

Se puede apreciar una mejora en la probabilidad de que el césped no esté húmedo y los aspersores no estén encendidos y no llueva, cuando hacemos la prueba con 1000 muestras obtenemos una probabilidad de 0.250, que se asemeja a la original, pero aún tiene vagaje (0.01), con el incremento del número de muestras, se va aproximando hasta que tiene un margen de error menor y se asemeja mas al original, no obstante una ejecución con un número de muestras de 1000 podría ser válido ya que es muy próximo al original.

Datos Incompletos

Para este método se ha hecho un experimento con 100, 1000 y 2000 elementos de muestra para ver como se comporta, en la siguiente tabla recogemos los resultados:

NºMuestras		100		1000		2000	
Original							
f	t						
1.0	0.0	0.9996	0.0004	0.9988	0.0012	0.9993	0.0007
0.1	0.9	0.0340	0.9660	0.2952	0.7048	0.2559	0.7441
0.1	0.9	0.0071	0.9929	0.1496	0.8504	0.1401	0.8599
0.01	0.99	0.1329	0.8671	0.0601	0.9399	0.0631	0.9369

Al haber información oculta, hay probabilidades que le cuesta acercarse a la original como puede ser la que el césped no esté húmedo y los aspersores estén apagados y esté lloviendo, que cuanto mayor es el número de muestras, más se asemeja al original, necesitaría muchísimas muestras para poder acercarse a las probabilidades de la original con más exactitud.

Al permitir mayor número de iteraciones, no hemos notado cambio, ya que converge con pocas iteraciones.

Ejercicio B

En la pagina 6.10 del tema 6 de teora sobre modelos grafos probabilisticos podemos ver una sencilla red bayesiana para diagnostico de cancer de pulmon. Se pide:

1. Desarrollar un script matlab que implemente dicha red usando BNT

El desarrollo de el implemento de la red se puede encontrar en los scripts EjDiagnostico.m y ExplProb.m, donde se hace lo siguiente:

Declaramos los nodos y ponemos las aristas.

```
Plcn = Polución, Fum = Fumador, RX = Rayos X, Dis = Disnea.  
N = 5; Plcn = 1; Fum = 2; CdP = 3; RX = 4; Dis = 5;  
grafo = zeros(N, N);  
grafo(Plcn,CdP) = 1;  
grafo(Fum,CdP) = 1;  
grafo(CdP,RX) = 1;  
grafo(CdP,Dis) = 1;
```

Asignamos pesos a cada nodo, al nodo RX le asignamos peso 3 ya que puede tomar 3 valores.

```
nodosDiscretos = 1:N;  
tallaNodos = 2*ones(1,N);  
tallaNodos(1,4) = 3  
redB = mk_bnet(grafo, tallaNodos,'discrete', nodosDiscretos);
```

Asignamos tablas a cada nodo.

```
redB.CPD{Plcn} = tabular_CPD(redB, Plcn, [0.9 0.1]);  
redB.CPD{Fum} = tabular_CPD(redB, Fum, [0.7 0.3]);  
redB.CPD{CdP} = tabular_CPD(redB, CdP, [0.999 0.97 0.95 0.92 0.001 0.03 0.05  
0.08]);  
redB.CPD{RX} = tabular_CPD(redB, RX, [0.80 0.10 0.10 0.2 0.1 0.7]);  
redB.CPD{Dis} = tabular_CPD(redB, Dis, [0.7 0.35 0.3 0.65]);
```

2. Cual es la probabilidad de que un paciente no fumador no tenga cancer de pulmón si la radiografía ha dado un resultado negativo pero sufre disnea?

Para resolver este problema hemos hecho un script (EjDiagnostico.m) donde con la red bayesiana generada, hemos hecho lo siguiente:

```
evidencia = cell(1,N);  
evidencia{Fum} = 1;  
evidencia{RX} = 1;  
evidencia{Dis} = 2;  
[motor, logVerosimi] = enter_evidence(motor, evidencia);  
m = marginal_nodes(motor, CdP);  
m.T'
```

Queremos saber la probabilidad de que un paciente no tenga cáncer de pulmón dado que no es fumador, en la radiografía ha dado negativo y sufre disnea, por tanto, añadimos las evidencias que tenemos y creamos el motor, con el motor, indicamos que queremos saber los posibles valores de tener cáncer de pulmón (CdP) y ejecutamos, obteniendo:

0.9989 0.0011

Donde podemos ver de manera trivial que la probabilidad de no tener cáncer en este caso es de 0.9989.

3. Cual es la explicación mas probable de que un paciente sufra cáncer de pulmón?

Para saber esta explicación, en el script ExplProb.m se ha hecho lo siguiente:

```
evidencia = cell(1,N);  
evidencia{CdP} = 2;  
[motor, logVerosimi] = enter_evidence(motor, evidencia);  
[explMaxProb, logVver] = calc_mpe(motor, evidencia);  
explMaxProb
```

Hemos tomado como evidencia que el paciente tiene cáncer y hemos usado `calc_mpe` para saber que estado de los nodos son los mas probables, obteniendo tras su ejecución la siguiente solución:

`{[1]}` `{[2]}` `{[2]}` `{[3]}` `{[2]}`

El 2 en la tercera posición nos indica que el paciente tiene cáncer, como hemos indicado en el código anterior, y nos muestra el estado más probable de cada nodo, dónde nos dice que lo más probable es que no haya polución, el paciente sea fumador, las radiografías muestran cáncer, y sufre de disnea, por lo que podemos decir que la explicación más probable que podemos dar al cáncer del paciente es que se ha desarrollado por ser fumador.