Aprendizaje Automático

Beltrán Domínguez, Victoria
Ge, Xiangyang

Memoria práctica 4: Redes bayesianas

El objetivo de esta práctica es experimentar con Modelos Gráficos Probabilísticos, también conocidos como Redes Bayesianas. Para ello utilizaremos el toolkit "BNT" para matlab. Así, la práctica nos va guiando en las primeras páginas sobre cómo poder utilizar el entorno en matlab, así como las diferentes funciones que existen y la forma de abordar los diferentes problemas mediante un ejemplo llamado *Sprinkler*.

El primer ejercicio que nos plantea la práctica es sobre el ejemplo *Sprinkler*. En este, se pide repetir los procesos de aprendizaje a partir de datos completos y de datos incompletos explicados en la Sec.2.3 de la práctica, usando un número mucho mayor de muestras de aprendizaje (por ejemplo, nMuestras = 1000), así como permitiendo un mayor número de iteraciones. Además, hemos de comentar los resultados obtenidos.

Para conseguir un número mucho mayor de muestras de aprendizaje permitiendo un mayor número de iteraciones cambiamos estas líneas de código:

```
% creacion de las muestras aleatorias
semilla=0;
rng(semilla);
nMuestras=1000;
muestras=cell(N,nMuestras);
```

Código para conseguir usar un número mayor de muestras de aprendizaje

```
% aprender los parametros
maxIter = 200;
eps = 1e-4;
rng(semilla);
[redEM2, trazaLogVer] = learn_params_em(motorEM, muestrasS, maxIter, eps);
```

Código para conseguir usar un número mayor de iteraciones (sprinkler incompleto)

En la siguiente tabla se muestra el resultado que querríamos obtener:

Tabla de probabilidades originales			
C (Cloudy)	S (Sprinkler)	R (Rain)	W (WetGrass)
1: 0.5 2: 0.5	1: 0.5 0.5 2: 0.9 0.1	1: 0.8 0.2 2: 0.2 0.8	1 1: 1.0 0.0 2 1: 0.1 0.9 1 2: 0.1 0.9 2 2: 0.01 0.99

Los resultados obtenidos para sprinkler con los datos completos probando sólo diferentes números de muestras (porque al conocer todos los datos, el número de iteraciones sería irrelevante) son los siguientes:

Aprendizaje con los datos completos				
N° muestras	C (Cloudy)	S (Sprinkler)	R (Rain)	W (WetGrass)
50	1: 0.38 2: 0.62	1: 0.68 0.32 2: 0.87 0.13	1: 0.68 0.32 2: 0.26 0.74	1 1: 1.0 0.0 2 1: 0.0 0.0 1 2: 0.04 0.96 2 2: 0.0 0.0
200	1: 0.49 2: 0.51	1: 0.49 0.51 2: 0.89 0.11	1: 0.82 0.18 2: 0.18 0.82	1 1: 1.0 0.0 2 1: 0.13 0.87 1 2: 0.09 0.91 2 2: 0.06 0.94
500	1: 0.49 2: 0.51	1: 0.52 0.48 2: 0.91 0.09	1: 0.82 0.18 2: 0.15 0.85	1 1: 1.0 0.0 2 1: 0.11 0.89 1 2: 0.11 0.89 2 2: 0.05 0.95
1000	1: 0.49 2: 0.51	1: 0.5 0.5 2: 0.91 0.09	1: 0.81 0.19 2: 0.17 0.83	1 1: 1.0 0.0 2 1: 0.11 0.89 1 2: 0.09 0.91 2 2: 0.03 0.97
2000	1: 0.5 2: 0.5	1: 0.49 0.51 2: 0.91 0.09	1: 0.81 0.19 2: 0.19 0.81	1 1: 1.0 0.0 2 1: 0.1 0.9 1 2: 0.1 0.9 2 2: 0.02 0.98
5000	1: 0.5 2: 0.5	1: 0.5 0.5 2: 0.9 0.1	1: 0.8 0.2 2: 0.19 0.81	1 1: 1.0 0.0 2 1: 0.1 0.9 1 2: 0.1 0.9 2 2: 0.01 0.99

10000	1: 0.5 2: 0.5	1: 0.5 0.5 2: 0.9 0.1	1: 0.8 0.2 2: 0.2 0.8	1 1: 1.0 0.0 2 1: 0.1 0.9 1 2: 0.1 0.9 2 2: 0.01 0.99

Después de obtener las distintas distribuciones de probabilidad dependiendo del número de muestras, se aprecia que cuando mayor sea este número, más se aproxima a la distribución original. Entre 1000-2000 muestras podemos ver que empieza a no haber mucha diferencia en las probabilidades de los datos y entre 5000-10000 las probabilidades son casi idénticas, lo cual quizás no sea necesario tantas muestras.

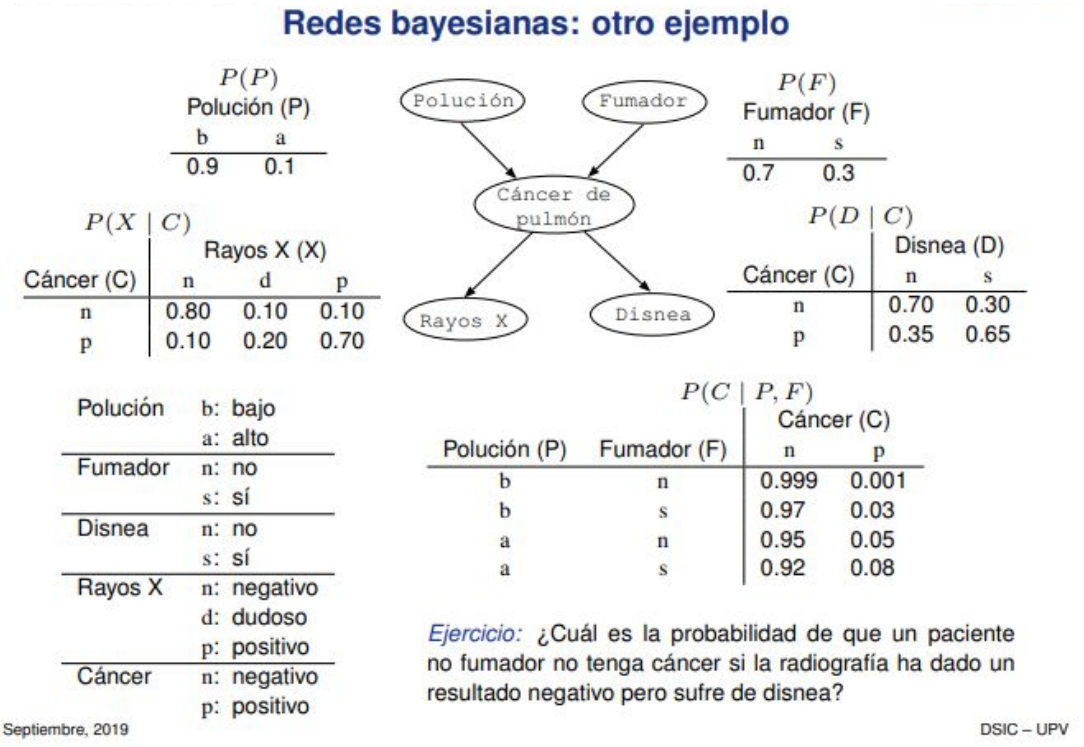
Los resultados obtenidos para sprinkler con los datos incompletos probando diferentes números de muestras y números de máximas iteraciones son los siguientes:

Aprendizaje con los datos incompletos					
N° muestras	N° máx. iteraciones	C (Cloudy)	S (Sprinkler)	R (Rain)	W (WetGrass)
50	100	1: 1.0 2: 0.0	1: 0.73 0.27 2: 0.99 0.01	1: 0.43 0.57 2: 0.0 1.0	1 1: 1.0 0.0 2 1: 0.0 1.0 1 2: 0.09 0.91 2 2: 0.01 0.99
50	200	1: 1.0 2: 0.0	1: 0.73 0.27 2: 0.99 0.01	1: 0.43 0.57 2: 0.0 1.0	1 1: 1.0 0.0 2 1: 0.0 1.0 1 2: 0.09 0.91 2 2: 0.01 0.99
500	100	1: 1.0 2: 0.0	1: 0.69 0.31 2: 1.0 0.0	1: 0.46 0.54 2: 0.0 1.0	1 1: 1.0 0.0 2 1: 0.24 0.76 1 2: 0.15 0.85 2 2: 0.1 0.9
500	200	1: 1.0 2: 0.0	1: 0.69 0.31 2: 1.0 0.0	1: 0.46 0.54 2: 0.0 1.0	1 1: 1.0 0.0 2 1: 0.24 0.76 1 2: 0.15 0.85 2 2: 0.1 0.9
1000	100	1: 1.0 2: 0.0	1: 0.7 0.3 2: 1.0 0.0	1: 0.42 0.58 2: 0.0 1.0	1 1: 1.0 0.0 2 1: 0.3 0.7 1 2: 0.15 0.85

					2 2: 0.06 0.94
1000	200	1: 1.0 2: 0.0	1: 0.7 0.3 2: 1.0 0.0	1: 0.42 0.58 2: 0.0 1.0	1 1: 1.0 0.0 2 1: 0.3 0.7 1 2: 0.15 0.85 2 2: 0.06 0.94
5000	100	1: 1.0 2: 0.0	1: 0.71 0.29 2: 1.0 0.0	1: 0.45 0.55 2: 0.0 1.0	1 1: 1.0 0.0 2 1: 0.28 0.72 1 2: 0.17 0.83 2 2: 0.04 0.96
5000	200	1: 1.0 2: 0.0	1: 0.71 0.29 2: 1.0 0.0	1: 0.45 0.55 2: 0.0 1.0	1 1: 1.0 0.0 2 1: 0.28 0.72 1 2: 0.17 0.83 2 2: 0.04 0.96

En el caso de los datos incompletos, podemos observar que al aumentar el número de las iteraciones no afecta a los datos, lo que significa que nunca se llega a la iteración 100 en el proceso para este particular caso. También podemos presenciar que la cantidad de datos incompletos es demasiado grande para intentar predecir correctamente la distribución de probabilidad original.

La última tarea que nos propone la práctica consiste en desarrollar un script de matlab que implementa una red bayesiana para diagnóstico de cáncer de pulmón usando BNT, siguiendo el modelo descrito en la imagen que se encuentra a continuación:



Para ello, hemos desarrollado un nuevo script (cancer.m) para introducir los datos correctamente mediante los pasos explicados del boletín de la práctica. Cabe destacar que la única diferencia es que aun siendo todos nodos discretos, el nodo X (Rayos X) presenta tres variables aleatorias discretas.

Una vez implementado, finalmente quedaría contestar a estas preguntas:

- ¿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?

  Al tener la red bayesiana definida, calcularemos esta probabilidad mediante inferencia. Para ello, tenemos que introducir las evidencias, que en este caso son:

  1. F=1 (No fumador)
  2. X=1 (Rayos X negativo)
  3. D=2 (Sí disnea)

  Con esto obtenemos la probabilidad condicional de no tener cáncer de pulmón (C=1) es 99.89%.

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

  En este caso sólo necesitamos introducir la evidencia C=2 (tener cáncer de pulmón) y el sistema nos devuelve la explicación más probable.

  1. P=1 (baja polución)
  2. F=2 (sí fumador)
  3. C=2 (sí cáncer)
  4. X=3 (rayos X positivo)
  5. D=2 (sí disnea)