

Redes Bayesianas

Sesión3

5 de octubre de 2023

1. Introducción

En esta última sesión dedicada a las redes bayesianas profundizaremos en el funcionamiento de los algoritmos de inferencia aproximada, también veremos brevemente como se pueden aprender redes bayesianas de forma automática a partir de datos. Todo esto con las herramientas que nos proporciona OpenMarkov.

2. Análisis de los algoritmos de inferencia aproximada

En este apartado vamos a usar el menú “Herramientas - Exportar datos de propagación estocástica”. En la ventana que nos aparece al seleccionar este menú, Figura 1, podemos elegir entre los dos algoritmos de inferencia aproximada que hay en OpenMarkov y el número de muestras que se generarán.

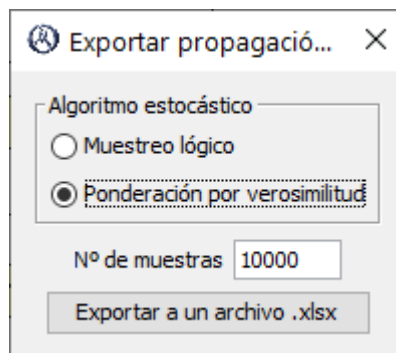


Figura 1: Ventana para generar datos de la propagación de los algoritmos de inferencia aproximada.

Este proceso toma como base la red en el estado actual. Esto quiere decir que se realizará la propagación del último caso de evidencia que hemos introducido, si hay alguno.

Vamos a analizar, para cada algoritmo, sobre la red *asia* y 10000 muestras, los resultados sin evidencia y otro caso en el que la evidencia es “*Has tuberculosis*”=*yes* y “*Positive X-ray*”=*yes*. En el archivo *resultadosPropagacion.zip* se encuentran los ficheros producidos, los usaremos para que todos veamos los mismos números ya que es un proceso estocástico y ahora no podemos controlar la generación de números aleatorios. En cualquier caso, utilizando los mismos parámetros debemos obtener unos resultados muy similares.

Empecemos con el muestreo estocástico. El fichero Excel con los resultados tiene dos hojas, en la primera se presentan cada una de las muestras generadas. Aquí, cada valor de las variables



se representa según el orden en que están definidos, empezando por 0. Para esta red el primer valor es *no* y el segundo es *yes*. También tenemos una columna con el peso de cada muestra, aunque sabemos que en el muestreo estocástico, o si no hay evidencia, todas las muestras pesan lo mismo y por tanto este valor no es significativo ahora. Observa que las variables están en orden topológico.

En la segunda hoja del fichero de resultados se da la estimación de las probabilidades de cada variable en base a las frecuencias observadas en la hoja 1. En esta segunda hoja también se muestra el valor exacto de cada probabilidad para usarlo como referencia.

Ejercicio 1. ¿Cuál es el error que se produce en esta estimación en media para las 8 probabilidades estimadas? ¿Es un error asumible?

Ejercicio 2. Abre ahora el fichero *asia - Likelihood weighting - sin evidencia*. Analiza de igual forma el error de la estimación. ¿Es una estimación mejor o peor que la del muestreo estocástico? ¿Se debe esta diferencia a las propiedades de cada algoritmo?

Usando la funcionalidad de Excel para filtrar filas según el valor de algunas columnas y para mostrar contadores o sumas de las celdas seleccionadas, podemos usar la primera hoja de estos ficheros de resultados para calcular probabilidades conjuntas de varias variables.

Ejercicio 3. ¿Cuál es la probabilidad $P(\text{Smoker?} = \text{no}, \text{Has bronchitis} = \text{yes})$? ¿Cuál es la probabilidad $P(\text{Dyspnoea?} = \text{yes}, \text{Has lung cancer} = \text{no})$?

Veamos ahora que ocurre cuando propagamos una evidencia no nula analizando los ficheros *asia - Logic sampling - evidencia Tub_Xray* y *asia - Likelihood weighting - evidencia Tub_Xray*. Lo primero que podemos observar es que la fila de algunas muestras para el método del muestreo estocástico aparecen en rojo. Estas son aquellas que no concuerdan con la evidencia seleccionada y que como sabemos serán descartadas. Muchas lo serán porque la probabilidad de la evidencia es baja, alrededor del 1%. En el caso del muestreo con ponderación de la verosimilitud todas las muestras están en verde ya que se fijan los valores de las variables que forman parte de la evidencia y, ahora sí, la columna *Weight* es importante para hacer la estimación.

Ejercicio 4. Analiza de nuevo el error de la estimación entre los dos algoritmos con esta evidencia. ¿Qué se puede decir de esta diferencia?

Ejercicio 5. Comprueba en unas pocas filas que el peso que se indica para el algoritmo de ponderación de la verosimilitud se corresponde con lo esperado en función de los valores de cada variable.

3. Aprendizaje de redes

Hasta ahora hemos visto como construir las redes bayesianas de forma manual. Para esta tarea se necesita conocer bien el problema que queremos modelar, o acceso a expertos, pero también se asume en la práctica que el número de variables y relaciones no es muy alto, de lo contrario el proceso de construcción de la red no será manejable.

Como los problemas de interés suelen ser complejos y de un tamaño considerable y además es difícil encontrar expertos que puedan dar cada detalle necesario de la red, la alternativa es generar esta red a partir de datos que se han podido recoger sobre el problema en cuestión. En este apartado vamos a ver como podemos hacer esto con OpenMarkov.

En primer lugar vamos a simular un conjunto de datos sobre los que después aprender una red. Con OpenMarkov podemos hacer esto con el menú “Herramientas - Generador de BBDDs”. Una vez hecho esto aparece la ventana para configurar los casos a generar, como se muestra en la Figura 2. Hay dos elementos a configurar, el primero es la red que se toma como base que puede ser la red cargada con la que estamos trabajando o una red que se cargará de un fichero. El segundo es el número de casos de entre las opciones que se ofrecen en la lista desplegable, de

100 a 100000 casos. Cuando pulsemos el botón “Generar” se nos pedirá el nombre, la ubicación y el formato del fichero de salida. Hay tres formatos en modo texto, como csv, arff y dbc, o el formato xls de Excel.



Figura 2: Ventana para generar casos simulados a partir de una red Bayesiana.

Usaremos en los siguientes pasos el fichero *asia10k.csv* que se ha creado sobre la red *asia.pgmx* con 10000 casos.

Para hacer el aprendizaje vamos al menú “Herramientas - Aprendizaje” y nos aparecerá una ventana como se muestra en la Figura 3.

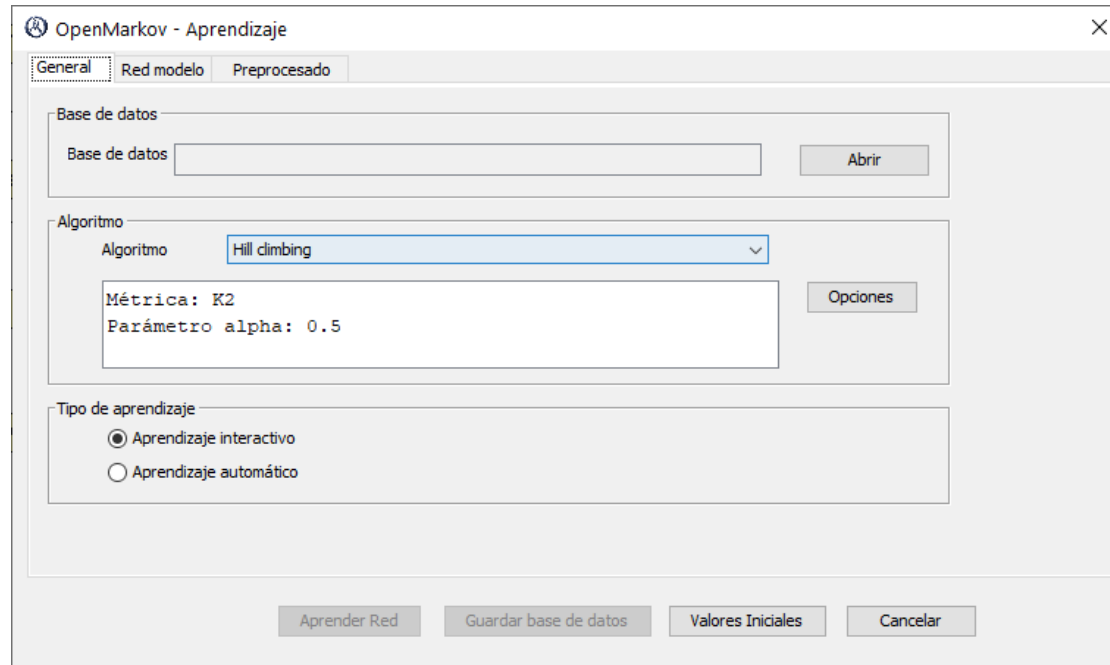


Figura 3: Ventana para configurar el aprendizaje automático de redes.

Por el momento nos quedamos con la pestaña “General”. Aquí debemos seleccionar en primer lugar el fichero con los casos, *asia10k.csv*. Después podemos elegir el método de aprendizaje. No es necesario profundizar ahora en estos algoritmos y nos quedamos la búsqueda *Hill climbing* y el heurístico llamado *K2*. Por último vamos a seleccionar la opción de aprendizaje totalmente automático. Con estas opciones se obtiene la red que se muestra en la Figura 4. Aunque es

una red generada de forma automática contiene todos los componentes de una red bayesiana y, por ejemplo, también se puede realizar inferencia sobre este modelo. Sin embargo es importante tener en cuenta que en la versión actual de OpenMarkov es necesario guardar la red aprendida y volverla a abrir para poder pasar al modo de inferencia.

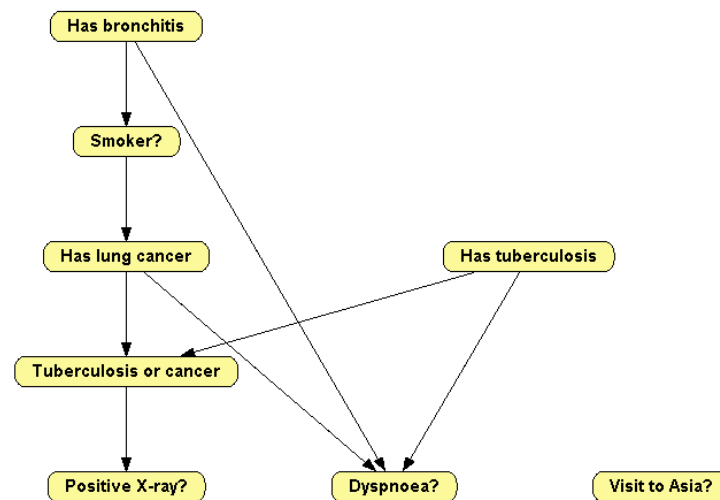


Figura 4: Red aprendida con la primera configuración de opciones.

Esta red no se parece mucho a la original, hay enlaces que no esperábamos y otros que desaparecen. Al generar una red a partir de datos, Opemmarkov intenta posicionar los nodos según un orden topológico. Esto es bueno en general para poder ver la estructura de la red, pero a veces puede resultar engañoso cuando se colocan varios nodos en vertical, porque no se podrá distinguir de que nodo salen los enlaces y a donde llegan. Por esto es bueno, una vez aprendido el modelo, desplazar con el ratón los nodos para ver si hay enlaces superpuestos.

No obstante, es más sencillo comparar el resultado del aprendizaje con la red original si los nodos están en la misma posición en la ventana. Esto se puede conseguir indicando en la segunda pestaña de la ventana de aprendizaje, “Red modelo”, la red que queremos usar para extraer esa posición de los nodos. Bien seleccionamos la opción de una red cargada si esta red estaba en la ventana activa, o bien podemos indicar el fichero que contiene esta red. De esta forma la red que se obtiene es la misma que antes pero es más fácil identificar las diferencias. Aunque ya sabemos que ciertas relaciones de dependencia o independencia se pueden reflejar con distintas estructuras de una red.

Ejercicio 6. Identifica relaciones de independencia en la red *asia* original que no están presentes en la red de la Figura 5.

En OpenMarkov también es posible realizar un aprendizaje que combina el proceso automático con información de expertos en el problema. Esto se consigue seleccionando *Aprendizaje interactivo* como el tipo de aprendizaje en la pestaña “General”. Haciendo esto se nos presentará una nueva ventana como se muestra en la Figura 6.

Aquí podemos ver que operaciones se pueden realizar sobre la red, que, dado que se empieza con una red sin enlaces, inicialmente solo son operaciones para añadir enlaces entre nodos. Lo interesante es que cada operación tiene asociada una valoración que se hace en base a los datos de entrada.

En el caso de la Figura 6 podemos ver como las dos primeras operaciones tienen la misma motivación. Esto es porque cualquiera de ellas generaría una red con las mismas relaciones. Es decir, para el algoritmo de aprendizaje no es posible distinguir que enlace es mejor. Sin embargo, para un experto la respuesta puede estar clara si se conoce la causalidad, ¿la disnea causa bronquitis o es la bronquitis la que causa la disnea?

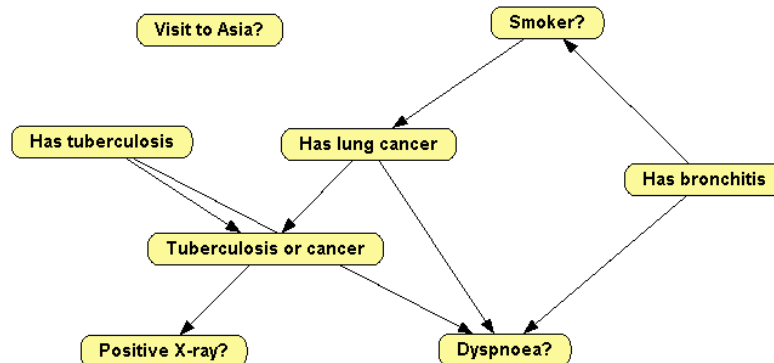



Figura 5: Red aprendida manteniendo la posición de los nodos.


Aprendizaje interactivo
✕

Descripción de la edición	Motivación
Add link: Dyspnoea? --> Has bronchitis	2500.71
Add link: Has bronchitis --> Dyspnoea?	2500.71
Add link: Has lung cancer --> Tuberculosis or cancer	1867.60
Add link: Tuberculosis or cancer --> Has lung cancer	1867.44
Add link: Tuberculosis or cancer --> Positive X-ray?	1550.62
Add link: Positive X-ray? --> Tuberculosis or cancer	1550.11
Add link: Has lung cancer --> Positive X-ray?	1277.29
Add link: Positive X-ray? --> Has lung cancer	1276.62

☒ Sólo permitidas
☒ Sólo positivas

Bloquear edición

Mostrar bloqueadas

Aplicar edición

Deshacer

Rehacer

Completar fase

Finalizar

Figura 6: Ventana para seleccionar los pasos para construir la red.

Con este modo de aprendizaje los expertos no tendrían por qué identificar todas y cada una de las relaciones de dependencia/independencia, sino solo aquellas que en base a los datos no se muestran muy claras. Si se continúa aplicando en cada paso la operación que nos parece más adecuada veremos que en la mayoría de los casos la diferencia en la valoración de las mismas es muy grande y basta con seleccionar la primera para llegar a un buen modelo final.