Redes bayesianas

Ramon Sangüesa i Solé

PID_00165733

UOC
Universitat Oberta de Catalunya
www.uoc.edu

Índice

Introducción

Las redes bayesianas son un modelo relativamente reciente, aunque empieza a tener muchas aplicaciones. Propuesto inicialmente por Pearl, el modelo de redes bayesianas es una representación que combina los aspectos cualitativos y cuantitativos de las relaciones entre los atributos (variables) de un dominio de una manera bastante intuitiva.

Lectura complementaria

Encontraréis la presentación del modelo de redes bayesianas en la obra siguiente:

J. Pearl (1988). *Probabilistic Reasoning in Intelligent Systems, Networks of Plausible Inference*. Morgan Kaufmann.

Su representación en forma de grafo y la sólida base estadística que lo sustenta lo hace relativamente fácil de entender y utilizar. En particular, la forma en que describe las relaciones de dependencia o influencia entre variables y el desarrollo de una serie de algoritmos de propagación bien probados para actualización, explicación y predicción mediante distribuciones de probabilidad hace que, además de ser un buen modelo descriptivo de un dominio, permita efectuar predicciones y encontrar explicaciones a situaciones nuevas. En este sentido, es un modelo bastante más versátil que los que hemos visto en otros módulos.

Objetivos

Tras haber trabajado los materiales didácticos de este módulo, el estudiante habrá alcanzado los objetivos siguientes:

1. Conocer las características principales de las redes bayesianas como representación de las dependencias entre los atributos de un dominio.

2. Aprender cómo se pueden utilizar las redes bayesianas para diferentes tareas de *data mining*: representación de asociaciones, predicción y explicación.

1. ¿Qué son las redes bayesianas?

Supongamos que describimos un dominio (un conjunto de datos) mediante una serie de atributos $X_1, \ldots, X_n$.

Una **red bayesiana** es un grafo dirigido acíclico cuyos nodos representan las variables X_i del dominio (atributos) donde cada variable es independiente del resto de las variables X_i, X_j del dominio dados sus predecesores directos.

Un **grafo dirigido acíclico** es un tipo de grafo en el que la dirección de los enlaces es relevante y en el que nunca puede suceder que en un camino entre dos nodos, el nodo inicial o el final, estén repetidos.

Un **enlace** entre dos variables X_i, X_j del dominio representa una asociación directa entre las dos; es decir, X_i influye sobre X_j.

Significado de las redes bayesianas

Una red bayesiana representa las influencias entre las variables de un dominio gráfica y probabilísticamente al mismo tiempo.

La influencia entre dos variables que son extremos de un enlace, tales que X_i es el origen del enlace y X_j, el extremo, está cuantificada por la distribución condicional de probabilidad de las dos variables implicadas: $P(X_i|X_j)$.

Ejemplo del viaje por carretera

Aclaremos con un ejemplo las propiedades de las redes bayesianas, que de entrada pueden parecer bastante abstractas.

Supongamos que queremos describir las relaciones que determinan lo que es importante a la hora de representar el coste y la duración de un viaje por carretera.

Las variables de interés son las siguientes:

- **Tipo de carretera**, que adopta los valores {'Autopista', 'Autovía', 'Nacional', 'Comarcal', 'Pista'}.
- **Tipo de vehículo**, que toma los valores {'Deportivo', 'Utilitario', 'Familiar'}.
- **Velocidad media en kilómetros por hora**, que adopta valores entre 0 y 150.
- **Coste de alquiler**, que puede tomar los valores {'Alto', 'Bajo', 'Medio'}.
- **Duración del viaje en horas**, que adopta valores entre 0 y 100.
- **Distancia del recorrido en km**, que puede tomar valores entre 0 y 5.000.

Dado este conjunto de atributos, esperamos que se produzcan las relaciones siguientes:

a) El coste del alquiler estará muy relacionado con el tipo de vehículo alquilado.
b) La velocidad media dependerá del tipo de coche que conduzcamos y del tipo de carretera.
c) La duración del viaje estará en función de la velocidad media y de la distancia por recorrer.

Esperamos que la duración del viaje dependa del tipo de vehículo y del tipo de carretera, pero si conocemos la velocidad media, estos dos factores pierden influencia frente a la velocidad media y la distancia en kilómetros. La velocidad media cubre la duración, la "protege" de la influencia de otros factores (carretera y vehículo). De hecho, es lo mismo que decir que la duración del viaje es independiente del tipo de carretera y del tipo de vehículo, si conocemos la velocidad media que se ha llevado y la distancia que había que recorrer. La variable *Velocidad media* hace que la variable *Duración* sea independiente de las varia-

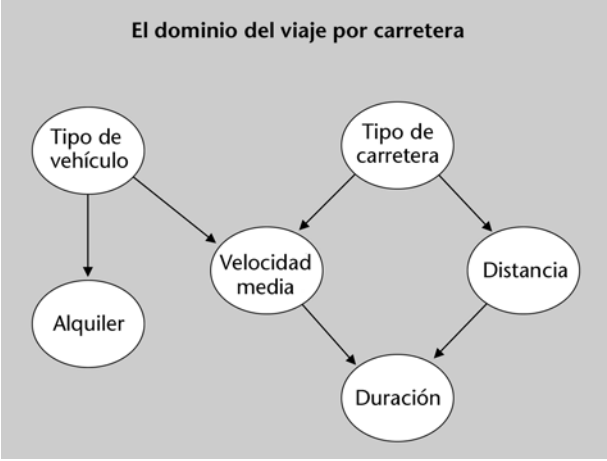
Lectura complenmentaria

Encontraréis una explicación detallada del ejemplo del viaje por carretera en la obra siguiente:

J.F. Huete (1995). *Aprendizaje de redes de creencia. Modelos no probabilísticos*. Tesis doctoral. Departamento de Ciencias de la Computación e Inteligencia Artificial. Universidad de Granada.

bles *Tipo de carretera* y *Tipo de coche*. Técnicamente, deberíamos decir que la variable *Duración* es condicionalmente independiente de *Tipo de carretera* y *Tipo de vehículo*, dada la *Velocidad media*.

Aquí tenemos una representación gráfica que registra el conocimiento de sentido común que acabamos de expresar sobre este dominio:



Esta estructura, por construcción y propiedades de las redes bayesianas, garantiza que se representen las propiedades de independencia condicional que hemos mencionado.

Hay un criterio gráfico, llamado *criterio de d-separación*, que permite "leer" las relaciones de independencia condicional de las diferentes variables directamente de un grafo como el que se refleja en el ejemplo del viaje por carretera.

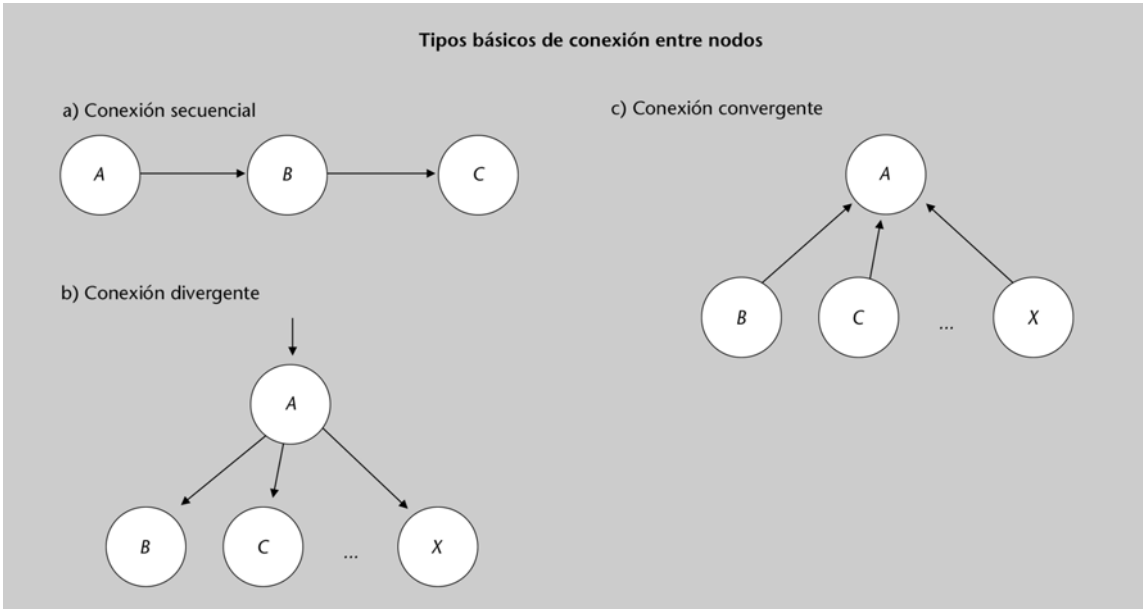
Lectura complemenraria

Hallaréis el criterio de *d*-separación en la obra siguiente:

J. Pearl (1988). *Probabilistic Reasoning in Intelligent Systems, Networks of Plausible Inference*. Morgan Kaufmann.

1.1. Relaciones cualitativas en las redes bayesianas: *d*-separación y modelos de dependencias

Para comprender mejor el criterio de *d*-separación vamos a ver cómo podemos interpretar algunos tipos básicos de conexiones entre nodos de una red. En la figura siguiente representamos estos tipos básicos de conexión:



Los tipos básicos de conexión entre nodos son los que mencionamos a continuación:

a) Conexión secuencial. Representada por el esquema **a** en la figura anterior. El atributo *A* tiene influencia sobre el *B* que, al mismo tiempo, influye sobre el *C*. Si contamos con alguna evidencia acerca de cuál es el valor de *A* (por ejemplo, porque hayamos observado su valor o porque *A* tenga cierta probabilidad de tomar uno determinado), podemos modular la certeza sobre los valores que puede adoptar *B* y propagar esa influencia a *C*. Por otra parte, la evidencia que tengamos sobre los valores de *C* influirá en nuestra certeza acerca de los valores que puede tomar *A* por medio de *B*. Ahora bien, si conocemos el valor de la variable *B*, entonces el camino entre *A* y *C* queda bloqueado. *A* y *C* se hacen independientes; *A* y *C* están *d*-separadas por *B*.

b) Conexión divergente. En la situación **b** de la figura anterior, la influencia puede pasar por todos los descendientes de *A*, a menos que conozcamos el valor de *A*. Decimos que *B, C, …, X* están *d*-separados, siendo conocido el valor de *A*.

c) Conexión convergente. En la situación **c** de la figura, si ignoramos totalmente el valor de *A* aparte de que se puede inferir uno a partir de los valores conocidos de los padres *B, C, …, X*, los padres son independientes entre sí; es decir, el conocimiento del valor de uno de los padres no posee influencia alguna en los valores que pueden tomar los demás.

Estos tres casos cubren todas las formas de transmisión de evidencia por medio de una variable. Siguiendo estas tres reglas, es posible decidir para cualquier par de variables de la red si son dependientes o no.

Se dice que dos variables *X* e *Y* en una red bayesiana están *d*-separadas por otra variable *Z* si para todos los caminos entre *X* e *Y* existe una variable intermedia *Z* tal que su conexión es secuencial o divergente y el estado de *Z* es conocido, o bien la conexión es convergente y ni *Z* ni ninguno de los descendientes de *Z* han recibido ninguna evidencia.

Podemos extender esta definición a conjuntos de variables en lugar de variables individuales.

Es interesante darse cuenta de que con este criterio podemos extraer el modelo de independencias asociado a una red bayesiana.

El **modelo de independencias** es un conjunto formado por una colección de aserciones del tipo "el conjunto de variables *X* es independiente de *Y* conocido *Z*".

Lectura recomendada

Podéis consultar los procedimientos para construir una red bayesiana a partir de un modelo de independencias en la obra siguiente:

J. Pearl; G. Rebane (1987). "The Recovery of Causal Poly-Trees from Statistical Data". *Uncertainty in Artificial Intelligence* (vol. 3, pág. 222-228).

Denotamos las aserciones del modelo de independencias con $I(X|Z|Y)$. Inversamente, hay procedimientos para construir una red que, dado un modelo de independencias, devuelven la red bayesiana correspondiente.

El problema es que, si tenemos un modelo de independencias y una red bayesiana definida sobre el mismo dominio, podemos encontrar las tres situaciones siguientes:

a) Todas las relaciones de independencia que hay en el modelo pueden detectarse en el grafo por medio de *d*-separación. Decimos que el grafo es un *D-map* del modelo de dependencias.

b) Todas las relaciones de independencia que se pueden detectar en el grafo mediante *d*-separación están presentes en el modelo de dependencias. Entonces decimos que el grafo es un *I-map* del modelo de dependencias.

c) En el grafo sólo se localizan las relaciones de dependencia del modelo y en el modelo sólo aparecen las relaciones de dependencia del grafo. En tal caso decimos que el grafo es un *P-map* o *Perfect map* del modelo.

Un mismo modelo de dependencias puede ser representado por varios grafos. En cuanto al *data mining*, nos interesa tener en cuenta lo siguiente:

1) Una base de datos definida sobre un conjunto de atributos $X_1, \ldots, X_n$ permite extraer un conjunto de dependencias.

2) Un método de construcción de redes bayesianas debe asegurar que el grafo que resulte de la aplicación sobre un conjunto de datos ha de ser, como mínimo, un *I-map* del conjunto de dependencias existente e, idealmente, un *P-map*.

Ya veremos más adelante cómo afecta todo esto al diseño de métodos de construcción de redes bayesianas. Ahora es preciso que conozcamos otras propiedades de este tipo de modelo.

Podéis ver los métodos de construcción de redes bayesianas a partir de datos en el apartado 2 de este módulo.

1.2. Relaciones cuantitativas en las redes bayesianas: probabilidades condicionales

Las relaciones cualitativas (estructura de conexiones del grafo, modelo de dependencias implícito en las relaciones de *d*-separación) de una red bayesiana vienen complementadas por relaciones cuantitativas que corresponden a las distribuciones de probabilidad condicional existentes entre los diferentes nodos que guardan una relación de parentesco directo.

Si tenemos un dominio X_1, …, X_n, y sobre ese dominio definimos una distribución de probabilidad $P(X_1,\ldots,X_n)$, recordemos que dos variables X_i, X_j son independientes dada una tercera, Z, si se cumple la relación siguiente:

$$P(X_i \mid X_j,Z) = P(X_i \mid Z) \text{ si } P(X_i,X_j) > 0$$

Para las relaciones de independencia condicional en una red bayesiana podemos ver que la distribución de probabilidad conjunta $P(X_1, \ldots, X_n)$ factoriza de la manera siguiente:

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P(X_i \mid pa_i)$$

donde pa_i es el conjunto de antecesores directos (padres) de la variable X_i. Lo único que hemos hecho es aplicar la regla de la cadena en probabilidades, que establece lo siguiente:

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P(X_i \mid X_1,\ldots,X_{i-1})$$

En efecto, hemos aprovechado la relación de orden entre las variables que establecen las conexiones de parentesco. Debemos tener en cuenta que para cada variable X_i, su conjunto de padres, $pa_i \subseteq \{X_1, \ldots, X_n\}$ hace que X_i y $\{X_1, \ldots, X_{i-1}\}$ se conviertan en independientes y precisamente sabemos que podemos expresar una probabilidad conjunta $P(X_1, \ldots, X_n)$ como producto de las probabilidades marginales $P(X_1)$, …, $P(X_n)$ cuando las variables son independientes entre sí.

Éste es uno de los puntos fuertes de las redes bayesianas en tanto modelos de representación del conocimiento. En efecto, sin esta característica que nos permite factorizar la distribución conjunta en función de la estructura de la red, para n variables necesitaríamos expresar la distribución conjunta en términos de 2^n distribuciones de probabilidad de variables para especificar la misma distribución. En cambio, ahora queda reducido a las probabilidades condicionales que hay que especificar entre padres e hijos. Puesto que en un grafo dirigido acíclico, el número de enlaces posibles para n nodos es $n(n-1)$, la reducción es lo suficientemente significativa. Las redes bayesianas suministran una representación compacta.

Finalmente, para especificar las características de las distribuciones de probabilidad condicional presentes en la red necesitamos unos cuantos parámetros estadísticos, tantos como combinaciones de valores de variables x_i^j para cada variable X_i y para cada una de las configuraciones de padres posibles para cada variable x_i^j haya. Es decir, debemos especificar los parámetros:

$$\theta_{ijk} = P(X_i = x_i^k \mid pa_i)$$

Desde el punto de vista del *data mining*, las propiedades cuantitativas también se pueden utilizar para extraer la red o las redes bayesianas "ocultas" dentro de un conjunto de datos. En efecto, una base de datos definida sobre X_1, ..., X_n tiene asociada una distribución de probabilidad $P(X_1,...,X_n)$ y la red bayesiana que queremos obtener también representa una factorización de la misma distribución de probabilidad.

El problema puede plantearse como vemos a continuación:

a) Hay que estimar los parámetros θ_{ijk} de la distribución de probabilidad implícita en los datos. En principio, no es necesario conocer esta distribución.

b) Es preciso que encontremos la red bayesiana que se adecua a la distribución caracterizada por estos parámetros. Por norma general, será preciso elegir entre todas las redes posibles aquella que se ajuste mejor a la distribución implícita en los datos. Es necesario, pues, disponer de una **medida de ajuste**.

Normalmente hay que realizar una doble búsqueda: en el espacio de parámetros y en el espacio de estructuras que se avienen con los parámetros. Asimismo, es preciso asegurar que la red recuperada es, como mínimo, un *I-map*. No es un problema simple, pero la utilidad práctica de los modelos recuperados merece la pena.

Podéis ver los *I-map* en el subapartado 1.1 de este módulo.

Hay varios algoritmos que permiten efectuar dos operaciones básicas fundamentadas en la propagación de valores de probabilidad. Veamos brevemente cuáles son estas operaciones que permite hacer una red bayesiana.

1.2.1. Operaciones sobre una red bayesiana

Las operaciones mínimas que permite hacer una red bayesiana son la **predicción** y la **explicación**. Ambas se basan en la propagación de evidencias dentro de la red.

Los **algoritmos de propagación de evidencias** se encargan de ver dentro de una red bayesiana cómo afecta la evidencia de que una variable X_i adopte un valor x_i determinado a los valores que pueden tomar el resto de las variables.

Propagación de evidencias

La propagación de evidencias se basa en un viejo teorema de la estadística, el teorema de Bayes, que indica cómo se actualizan los valores de las variables al acumular evidencias nuevas.

Para efectuar estos cálculos es necesario actualizar las diferentes distribuciones de probabilidad. Para ello, necesitamos conocer los elementos siguientes:

a) La distribución *a priori* de las variables raíz (las que carecen de antecesor o predecesor).

b) Las distribuciones condicionales entre una variable y sus padres que ya están codificadas en la misma red.

La explicación concreta de los algoritmos de propagación supera los objetivos de esta asignatura. Ahora nos basta con saber que tales algoritmos se basan en una extensión muy hábil de la relación entre las probabilidades *a priori* y *a posteriori* conocida como *teorema de Bayes*.

El **teorema de Bayes** establece que, dada una hipótesis (H) y una evidencia (E), se verifica la relación siguiente:

$$P(H \mid E) = \frac{P(E|H)P(H)}{P(E)}$$

A priori, la hipótesis puede tener cierta distribución de probabilidad.

Ejemplo de propagación de evidencias

Reanudamos el ejemplo inicial del viaje por carretera. Sin conocer más detalles, podríamos saber que la probabilidad de que el viaje dure tres horas es del 50%.

Si tenemos la hipótesis de que la duración del viaje por carretera es de tres horas y contamos con la evidencia de que la carretera es una nacional y el coche, un utilitario, esperamos que la probabilidad *a posteriori* de que el viaje dure tres horas, teniendo en cuenta que sabemos que el coche es un utilitario y la carretera es nacional, sea diferente de la que teníamos antes de reunir estas evidencias (o no, si las variables son independientes).

En cualquier caso, éste no es momento ni lugar adecuado para entrar en más detalles ni sobre el teorema de Bayes ni sobre los algoritmos de propagación y actualización de evidencias en redes bayesianas. Ya hemos indicado dónde podéis encontrar referencias más extensas y detalladas al respecto.

Pondremos un ejemplo con la intención de aclarar las posibilidades y el interés que puede tener la realización de consultas sobre las redes.

Supongamos una red bayesiana extraída de la base de datos de Hyper-Gym. Utilizaremos una herramienta de construcción de redes bayesianas a partir de datos (BKD, *Bayesian Knowledge Discoverer*).

Lecturas complementarias

Hallaréis la explicación concreta de los algoritmos de propagación en las obras que mencionamos a continuación. La tercera obra presenta la exposición más actualizada. La última obra apuntada es una referencia excelente en cuanto a la multitud de algoritmos de propagación exactos y aproximados. Encontraréis la referencia completa en el apartado de bibliografía del módulo.

J. Pearl (1988).

R.E. Neapolitan (1990).

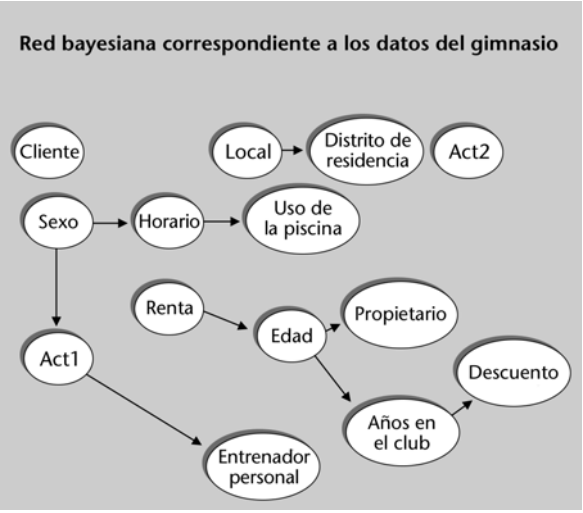
F.V. Jensen (1996).

E. Castillo; M. Gutiérrez; A. Hadi (1997).

Podéis ver el ejemplo del viaje por carretera en el inicio de este apartado.

Interés de las redes bayesianas

El interés de las redes bayesianas reside en la variedad y potencia de las operaciones que permiten realizar.



De entrada, sólo por inspección visual, el esquema de la red ya nos dice bastantes cosas. Por ejemplo, que los siguientes atributos no tienen mucha relación con el resto:

- *Cliente*. Si se trata del identificador de cliente (diferente para cada observación), no es de extrañar que no se pueda establecer ninguna relación con el resto de las variables. De hecho, habría sido más normal no considerar esta variable en el estudio.

- *Local y Distrito de residencia*. No parecen influir en el resto de las variables ni influirse mutuamente, pero entre ambas hay una fuerte relación. Podemos interpretar que conocer a qué local asiste un cliente nos informa de cuál es su distrito de residencia.

Actividad

1.1. Intentad extraer vosotros mismos las relaciones de independencia condicional expresadas por la estructura de la red que acabáis de ver en la figura anterior.

Vamos a ver qué podemos hacer con esta red:

1) De entrada, el modelo nos indica las diferentes distribuciones de probabilidad condicional.

Distribuciones de probabilidad condicional de las variables del modelo

Explicitamos la distribución de probabilidad condicional entre las variables *Sexo* y *Horario* en la tabla siguiente:

		Horario	
		Mañana	Tarde
Sexo	Hombre	0,770	0,230
	Mujer	0,551	0,449

A continuación vemos la tabla correspondiente a las variables *Actividad 1* y *Sexo*:

		Actividad 1				
		Aeróbic	TBC	Yoga	*Stretch*	*Steps*
Sexo	Hombre	0,358	0,477	0,023	0,066	0,077
	Mujer	0,737	0,156	0,036	0,019	0,053

La relación entre *Entrenador personal* y la primera actividad que desarrolla el cliente tiene la tabla de probabilidad condicional siguiente:

		Actividad 1				
		Aeróbic	TBC	Yoga	*Stretch*	*Steps*
Entrenador	No	0,959	0,979	0,262	0,986	0,990
	Sí	0,041	0,021	0,738	0,014	0,010

2) El modelo también nos da las probabilidades *a priori* de todas las variables.

Probabilidades *a priori* de todas las variables del modelo

Fijémonos en la distribución *a priori* de la variable *Entrenador personal*. La distribución *a priori* de este atributo en la base de datos indica que es un servicio mayoritariamente no solicitado: el 90,5% de las observaciones tienen el valor 'No' para este atributo y el 0,5% restante tienen el valor 'Sí'. Si ahora observamos que uno de los clientes es una mujer (Evidencia = 1), e introducimos esta observación en el modelo, podemos ver si hay cambios. En efecto, el valor de la distribución *a posteriori* de *Entrenador personal* ha descendido hasta el 94%. Es un cambio poco espectacular pero lo suficientemente interesante si tenemos en cuenta que la relación entre la variable *Sexo* y la variable *Entrenador personal* está mediatizada por la actividad principal (*Act1*) que se desarrolla.

3) También se puede realizar otros tipos de consulta, como ver qué cambios tienen lugar después de actualizar dos variables simultáneamente (por ejemplo, *Sexo* y *Actividad*).

4) Otra característica interesante de las redes bayesianas es la que incorporan los algoritmos de explicación.

Los **algoritmos de explicación** devuelven la configuración de variables y valores más probable a partir del valor observado de una o más variables.

Algoritmos de explicación

Si vemos que alguien ha solicitado un entrenador personal, el algoritmo de explicación nos devuelve el conjunto, {(*Hombre* = 'Sí'), (*Act1* = 'TBC'), (*Renta* = 3.000.000 – 10.000.000)}. Es decir, el algoritmo nos indica que la causa más probable de que alguien solicite un entrenador personal es que sea un hombre que practica TBC como actividad principal y tenga una renta alta. No aparecen otras variables porque no contribuyen con la suficiente evidencia para explicar esta observación. Este tipo de proceso es muy interesante en problemas de diagnosis que relacionan síntomas y causas: dado un valor observado para un síntoma, nos retorna el conjunto de causas más probables.

Lectura complementaria

Para una introducción fácil a los métodos de propagación podéis consultar la obra siguiente:

E. Charniak (1991). "Bayesian Networks without Tears". *AI Magazine* (vol. 12, núm. 4, pág. 50-63).

Y ya para finalizar, las redes bayesianas también se pueden utilizar para llevar a cabo tareas de clasificación.

Así pues, podemos considerar que son un modelo un tanto costoso de construir, pero bastante útil.

2. Métodos de construcción de redes bayesianas a partir de datos

Evidentemente, la manera más directa de construir una red bayesiana es a partir del conocimiento de un experto que indique cuáles son los atributos relevantes, qué relación hay entre éstos y que nos diga cuál es su fuerza de asociación. Sin embargo, nos interesa conocer cómo podemos extraer automáticamente una red bayesiana a partir de un conjunto de datos que describe un dominio.

Podemos exponer el **problema de la construcción de una red bayesiana** a partir de un conjunto de datos de la manera siguiente: dado un conjunto de datos, el problema de la construcción de una red bayesiana consiste en extraer la topología de la red que se encuentra implícitamente representada y su distribución de probabilidad correspondiente.

En una aproximación ingenua, podemos pensar que se trata de ir probando varias combinaciones de modelos construidos a fuerza de conectar nodos diferentes y en direcciones también distintas. La magnitud del número de modelos posibles que se pueden obtener nos impide este tipo de aproximación. Robinson calculó el número de grafos dirigidos acíclicos posibles que se pueden extraer de un conjunto de n nodos (correspondientes a las n variables $X_1, ..., X_n$ del dominio) con la ayuda de esta impresionante fórmula recursiva:

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} f(n-i)$$

con $f(0) = 1$ y $f(1) = 1$. Si calculáis un poco podréis ver que sólo con nueve nodos, ya estamos en el orden de los millares de millones de posibles grafos. Si tenéis en cuenta la cantidad de variables que se necesitan para describir un dominio real, podréis ver que no se trata de un espacio de búsqueda precisamente pequeño.

Lectura complementaria

Encontraréis la deducción de la fórmula recursiva de Robinson en la obra siguiente:

R.W. Robinson (1977). "Counting Unlabeled Acyclic Graphs. Lecture Notes in Statistics, 622". En: C. Little (ed.), pág. 28-43. Springer-Verlag.

Por lo tanto, es necesario recurrir a una estrategia de búsqueda heurística. Hay que establecer alguna medida de calidad que nos sirva para explorar este espacio de búsqueda tan grande. Existen tres tipos de métodos:

- basados en las propiedades de la distribución de probabilidad,
- basados en propiedades de independencia,
- híbridos.

Los más conocidos y utilizados son los primeros son los únicos que describiremos con cierto detalle.

Lecturas complementarias

Para un tratamiento más detallado de los métodos basados en propiedades de independencia e híbridos, consultad las obras siguientes. Encontraréis la referencia completa en el apartado de bibliografía del módulo.

R. Sangüesa; U. Cortés (1997).

J.F. Huete (1995).

L.M. De Campos; J.F. Huete (1997).

2.1. Métodos basados en propiedades de la distribución de probabilidad

Podéis ver la propiedad de factorización de la distribución de probabilidad conjunta en el subapartado 1.2 de este módulo didáctico.

Recordemos que la propiedad de factorización de la distribución de probabilidad conjunta de las redes bayesianas nos permite expresar la distribución que corresponde a un modelo de red bayesiana que se está construyendo en un momento dado del proceso de búsqueda como un multiplicatorio de las probabilidades de cada nodo condicionadas a sus padres.

Por lo tanto, a lo largo del proceso de construcción de una red a partir de los datos, en todo momento tenemos una disparidad entre la distribución conjunta (expresada como un multiplicatorio que corresponde a la estructura padres-hijos de la red construida hasta ese momento) y la distribución que suponemos que hay en los datos, que admite la forma de multiplicación de probabilidades marginales.

En consecuencia, se trata de encontrar aquella estructura de red bayesiana cuya distribución de probabilidad conjunta (expresada como un multiplicatorio que sigue la estructura padres-hijos) sea la más próxima a la que suponemos implícita en los datos.

Se trata de un problema típico de extracción de modelos a partir de datos, que admite varias formas de ataque. Las medidas de ajuste se pueden expresar mediante los criterios siguientes:

a) Criterios basados en entropía. Se trata de encontrar la red con una entropía cruzada o divergencia de Kullback-Leibler menor.

b) Criterios basados en estimación bayesiana. Se trata de encontrar una estructura y un conjunto de distribuciones que presenten la máxima probabilidad *a posteriori* (MAP) dados los datos existentes.

MAP es el acrónimo de la expresión *máxima probabilidad a posteriori.*

c) Criterios basados en el principio MDL. Trata de encontrar el modelo que tiene la mínima codificación, dados los datos, y que permite codificar los datos con la mínima longitud de descripción.

Podéis ver la MDL en el subapartado 5.2.1 del módulo "Agregación" de este módulo.

Todos estos métodos han derivado alguna forma de establecer la calidad global de la red en construcción durante el proceso de búsqueda según sus componentes y han reducido las medidas de calidad globales a expresiones en función de las relaciones entre las variables y sus padres. No olvidemos que esto es posible gracias a la propiedad de factorización de las redes. En general, todas estas medidas poseen una forma parecida a la expresión siguiente:

$$Calidad(Red \mid Datos) = \sum_{X_i} Calidad(X_i \mid pa_i, Datos)$$

2.1.1. Métodos basados en la entropía

El método más antiguo para construir una estructura parecida a una red bayesiana fue propuesto por Chow y Liu. Este método hace uso de la entropía y la información mutuas para construir la red bayesiana. Nosotros presentaremos el método Kutato, que deriva en parte de aquel método.

Lecturas complementarias

Encontraréis el método propuesto por Chow y Liu en la primera de las obras que mencionamos a continuación, y el método que seguimos aquí en la segunda.

C. Chow; S. Liu (1968). "Approximating Discrete Probability Distributions with Dependence Trees". *IEEE Transactions on Information Theory* (núm. 14, pág. 462-467).

E.H. Herskovitz; G.F. Cooper (1990). "Kutato: an Entropy-Driven System for the Construction of Probabilistic Expert Systems from Data". *Proceedings of the sixth conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers.

La **entropía** puede considerarse una medida de la cantidad de información presente en una distribución de probabilidad o como una medida de la incertidumbre asociada a una variable. Este concepto también se ve al hablar de árboles de decisión, en concreto al explicar cómo decide ID3 sobre la homogeneidad de una partición, así como al hablar de los métodos de discretización. Recordemos su definición:

$$H(X) = -\sum_{i=1}^{r} P(x_i)\log_2 P(x_i)$$

En esta expresión, r denota el número de valores posibles que puede tomar la variable X.

Sabemos que la entropía cumple la propiedad de ser una medida positiva que llega a su mínimo cuando la incertidumbre de la distribución es mínima y viceversa.

La **entropía conjunta** puede definirse como:

$$H(X,Y) = -\sum_{i,j=1}^{r_x,r_y} P(x_i,y_j)\log_2 P(x_i,y_j)$$

donde r_x y r_y son las cardinalidades de los conjuntos de valores que pueden tomar X e Y, respectivamente. Esta definición se puede extender a un conjunto de n variables $X_1, \ldots, X_n$.

La **entropía condicional de X dado que $Y = y_j$** es la entropía de la distribución condicional $P(X|Y = y_j)$:

$$H(X|Y = y_j) = -\sum_{i}^{r_x} P(x_i|Y = y_j)\log_2 P(x_i|Y = y_j)$$

La **entropía condicional de X con respecto a Y** es la media del valor de la entropía de X con respecto a los valores de y:

$$H(X|Y) = -\sum_{j}^{r_Y} P(y_j)\left[\sum_{i=1}^{r_X} P(x_i|Y = y_j)\log_2 P(x_i|Y = y_j)\right] =$$
$$= -\sum_{i,j=1}^{r_X,r_Y} P(x_i|y_j)\log_2 P(x_i|y_j)$$

La **regla de la cadena** relaciona la entropía conjunta y la condicional de la manera siguiente:

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

La **información mutua** entre dos variables X e Y mide la media de la reducción en la incertidumbre sobre X que provoca el hecho de tener información sobre el valor de Y y viceversa. De la misma manera podemos decir que la información mutua mide la cantidad de información media que Y aporta sobre X o también el grado de restricción que una variable aporta sobre la otra:

$$I(X;Y) = H(X) - H(X|Y) = \sum_{i,j=1}^{r_X,r_Y} P(x_i,y_j)\log_2\frac{P(x_i|y_j)}{P(x_i)P(y_j)}$$

La **entropía cruzada** o **divergencia de Kullback-Leibler** entre dos distribuciones de probabilidad P y P'es:

$$D_{KL}(P|P') = \sum_{i}^{r_X} P(x_i)\log_2\frac{P(x_i)}{P'(x_i)}$$

El algoritmo de Chow y Liu construye un grafo en forma de árbol en el que cada rama conecta las variables con información mutua máxima. El mérito consistió en demostrar que este método de construcción siempre encuentra el árbol con divergencia mínima. Por tanto, este método recupera la estructura que más información aporta y la distribución de probabilidad de la información que es más parecida a la que se halla implícita en los datos.

Herskovitz y Cooper diseñaron un método para recuperar la estructura de una red bayesiana dado un conjunto de datos que utilizaba la entropía como medida de calidad. El método considera que la red de entropía mínima es la más informativa. La entropía de la red siempre es mayor que la de la distribución del conjunto de datos.

Lectura complementaria

Encontraréis el algoritmo de Herskovits y Cooper en la obra siguiente:

E.H. Herskovitz; G.F. Cooper (1990). "Kutato: an Entropy-Driven System for the Construction of Probabilistic Expert Systems from Data". *Proceedings of the sixth conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers.

La entropía para una red bayesiana B_s se calcula como la suma de la entropía condicional de cada una de las variables X_i, dados sus padres pa_i.

$$H(B_S) = \sum_{i=1}^{n}\left(\sum_{j}^{q_i} P(pa_i^j)\sum_{k}^{r_i} P(x_i^k|pa_i^j)\log_2 P(x_i^k|pa_i^j)\right)$$

Esta fórmula calcula la entropía de la red teniendo en cuenta los factores siguientes:

- r_i es el número de valores que puede tomar la variable X_i.
- x_i^k representa el valor k-ésimo que puede adoptar la variable X_i.
- q_i es el número de configuraciones de padres de X_i, pa_i posibles.

- pa_i^j es la *j*-ésima configuración de padres presente en la base de datos.

El **algoritmo Kutato** empieza con un grafo formado por todos los nodos sin ninguna conexión entre sí. A cada paso añade el arco que produce la estructura con la mínima entropía. Se detiene cuando la estructura B_s alcanza un nivel de entropía suficientemente bajo.

A continuación presentamos el algoritmo que sigue el método Kutato. En la entrada del algoritmo hay una base de datos *D* sobre un conjunto de variables $\{X_1, \ldots, X_n\}$, un valor límite de entropía inferior, α, y un orden sobre las variables. En este caso, el método sigue los pasos siguientes:

1) Construir una estructura sobre $\{X_1, \ldots, X_n\}$ y suponer que las variables son marginalmente independientes {grafo inconexo}.

2) $\beta = H(B_s)$ {calcular la entropía de la red}.

3) Repetir hasta que $\beta \leq \alpha$ el bucle siguiente:

a) Seleccionar un enlace tal que:
- No crea ningún ciclo.
- Es el que crea la nueva estructura B_s con entropía mínima.
- Relaciona las variables *X* e *Y* de manera que *X* es anterior en el orden definido.

b) Dar la orientación $X \Rightarrow Y$.

A continuación, comentaremos el método que los mismos autores propusieron más tarde y que se basa en inferencia bayesiana.

El método K2

La intuición que subyace en el método Kutato es encontrar la estructura más probable dado el conjunto de observaciones recogido en la base de datos iniciales y alguna información sobre las distribuciones *a priori* de las estructuras posibles (esta información se reduce a suponer que todas las estructuras de red bayesiana son igualmente probables o siguen una distribución normal).

La idea se puede expresar de nuevo mediante el teorema de Bayes. Supongamos que queremos encontrar la red bayesiana B_S, donde *S* denota el par (B_S, B_P) y B_P son las distribuciones de probabilidad condicionales asociadas

a la estructura. La probabilidad de la red, dados los datos D, se puede expresar de esta manera:

$$P(B_S|D) = \frac{P(B_S,D)}{P(D)}$$

En realidad, nos interesa comparar redes posibles entre sí; es decir, dadas dos redes posibles B_{S1} y B_{S2}, debemos comparar la razón:

$$\frac{P(B_{S_1}|D)}{P(B_{S_2}|D)} = \frac{\frac{P(B_{S_1},D)}{P(D)}}{\frac{P(B_{S_2},D)}{P(D)}} = \frac{P(B_{S_1},D)}{P(B_{S_2},D)}$$

Así pues, el método K2 utiliza la probabilidad de una estructura procedente de los datos para aproximarse a la probabilidad condicional correspondiente. Pero calcular esta probabilidad no es tan sencillo, incluso teniendo en cuenta las simplificaciones siguientes:

1) Los atributos que aparecen en la base de datos son discretos.

2) Las observaciones dentro de una base de datos son independientes entre sí dada una estructura de red bayesiana.

3) No hay observaciones a las que les falte algún valor.

4) Antes de observar los datos D no tenemos ninguna preferencia con respecto a las probabilidades numéricas que se asignan a la red bayesiana.

Todas estas simplificaciones permiten derivar un método heurístico para encontrar la red que maximiza la probabilidad *a posteriori*, dados los datos. El proceso de derivación de este proceso es bastante complicado y no lo reproducimos aquí. La fórmula final del método heurístico, denotado por g, es bastante notable:

$$g(X_i,pa_i) = \prod_{j=1}^{q_i} \frac{(r_i-1)!}{(N_{ij}-r_i-1)!} \prod_{k=1}^{r_i} N_{ijk}!$$

donde cada factor de la expresión tiene el significado siguiente:

- pa_i es el conjunto de padres de X_i.
- q_i es el número de configuraciones diferentes que aparecen en la base de datos para los padres de X_i; es decir, las diferentes asignaciones de valores que los padres toman en la base de datos.
- r_i es el número de valores que puede adoptar la variable X_i.
- N_{ijk} denota el número de observaciones en las que la variable X_i toma en la base de datos el valor k-ésimo de entre los que puede tomar, y la configu-

Lectura complementaria

Si estáis interesados en la deducción de la expresión heurística g, podéis consultar la obra siguiente:

G.F. Cooper; E.H. Herskovitz (1992). "A Bayesian Method for the Induction of Probabilistic Networks from Data". *Machine Learning* (núm. 9, pág. 309-347).

ración de valores de los padres es la *j*-ésima de entre las que hay en la base de datos.

- N_{ij} es la suma de configuraciones posibles para cada valor de X_i.

El algoritmo es muy sencillo desde un punto de vista conceptual. Empieza por una variable sin padres y se le van añadiendo. Se trata de seleccionar en todo momento la variable que, junto con los padres de la variable considerada, maximiza el valor de *g*. Para facilitar la selección de las variables que se conectan entre sí, se declara un orden inicial entre las variables. Además, se impone un número máximo de padres por variable. Cuando al añadir un nuevo padre a la variable no incrementa su probabilidad, el algoritmo deja de añadir padres a una variable y pasa a considerar otra. El algoritmo K2 parte de los elementos siguientes: una base de datos sobre un dominio de variables $\{X_1, \ldots, X_n\}$, un orden entre las variables y un número máximo de padres por variable, u. El algoritmo consiste los pasos siguientes:

1) Construir una estructura B_s y suponer que todas las variables son marginalmente independientes (es decir, crear un grafo inconexo).

2) Ejecutar el bucle siguiente:

para $i := 1$ **hasta** n **hacer**
  $pa_i := \emptyset$
  $P_{anterior} := g(X_i, pa_i)$
  $OK :=$ verdadero
    mientras OK y $(|pa_i| < u)$ **hacer**
      Sea z la variable anterior a X, $z \notin pa_i$ tal que maximiza $g(X_i, pa_i \cup \{z\})$
      $P_{actual} := g(X_i, pa_i \cup \{z\})$
      si $P_{actual} > P_{anterior}$ **entonces**
        $P_{anterior} := P_{actual}$
        $pa_i := pa_i \cup \{z\}$
      sino
        $OK :=$ falso
      fsi
    fmientras
fpara

2.1.2. Métodos basados en el principio de la mínima longitud de descripción

Ya hemos comentado al hablar de procesos de discretización cuál era la lógica que seguía el principio de la mínima longitud de descripción. En este caso, se trata de poder turbar una red bayesiana a partir de un conjunto de observaciones tales que se minimice la codificación de la red y de los datos, dada la red.

Podéis ver el principio de mínima longitud de descripción en el subapartado 2.1.2 del módulo "Clasificación: árboles de decisión" de esta asignatura.

Los **métodos basados en el principio de la mínima longitud de descripción** consisten, en primer lugar, en encontrar una manera de codificar una red bayesiana y, una vez conocida la red de la que se supone que se han derivado los datos, codificarlos.

Codificación de la red

Codificaremos la red entendida como la combinación de la estructura B_S y la lista de probabilidades condicionales asociadas a cada nodo, B_P. Es necesario, pues, codificar la estructura y la lista de probabilidades condicionales.

Supongamos que hay n variables en el conjunto de datos $\{X_1, \ldots, X_n\}$. Para una variable X_i, representada como un nodo dentro del grafo, que tiene $|pa_i|$ padres diferentes, necesitamos $|\mathrm{pa}_i|\log_2(n)$ bits para codificar la lista de sus padres. En total, en toda la red necesitamos la cantidad de bits determinada por la expresión siguiente:

$$\sum_{i=1}^{n} |pa_i|\log_2(n)$$

Para calcular cuántos bits se necesitan para codificar las probabilidades condicionales de cada nodo X_i hay que multiplicar el número de bits necesarios para codificar el valor numérico de cada probabilidad condicional por el número total de probabilidades condicionales. El número de bits viene dado por la expresión siguiente:

$$\sum_{i=1}^{n} d(r_i - 1)q_i$$

donde, como siempre, r_i es el número de valores que puede tomar la variable X_i y q_i es el número de configuraciones que toman sus padres. Finalmente, d es el número de bits necesarios para codificar un valor numérico (los valores que adoptan los diferentes valores de la probabilidad de la distribución condicional).

La suma total de la codificación de la estructura y de las distribuciones condicionales es la que da como resultado la cantidad siguiente:

$$\sum_{i=1}^{n} |pa_i|\log_2(n) + \sum_{i=1}^{n} d(r_i - 1)q_i$$

La codificación de los datos se hace considerando un modelo formado por la estructura más las distribuciones condicionales de probabilidad. Tal como se explica, al hablar por primera vez del MDL, la codificación se hace utilizando el algoritmo de Huffman.

Podéis ver el algoritmo de Huffman en el subapartado 5.2.1 del módulo "Agregación" de esta asignatura.

El **algoritmo de Lam y Bacchus** utiliza un método basado en el principio MDL para construir redes bayesianas a partir de datos. Este algoritmo utiliza la medida de información mutua entre una variable y sus padres con el fin de seleccionar las variables que se conectan en un momento dado. A continuación presentamos la medida de la información mutua entre una variable X_i y sus padres pa_i.

$$I(X_i;pa_i) = -\sum_{i=1}^{r_X,pa_i} P(x_i,pa_i)\log_2\frac{P(x_i|pa_i)}{P(x_i)P(pa_i)}$$

Como demostraron Chow y Liu en el caso de los grafos estructurados como árboles, si se encuentra el árbol de expansión maximal con pesos iguales a la información mutua entre cada nodo (por ejemplo, con el algoritmo de Kruskal) se consigue la distribución que tiene la mínima divergencia de Kullback con respecto a los datos. Lam y Bacchus hicieron una demostración parecida para el caso de una variable con más de un padre (lo cual no ocurre en los árboles).

Ahora bien, si utilizásemos la medida de información mutua, se generaría un grafo con demasiadas conexiones entre padres e hijos. Aquí es donde entra en juego el principio MDL. En efecto, utilizando el criterio de la información mutua obtenemos una distribución de probabilidad para la red que es la más próxima a la distribución de probabilidad de los datos. Esta distribución permite obtener la codificación de longitud mínima.

El **algoritmo de Lam y Bacchus** calcula la información mutua entre todas las variables. Mantiene "abiertas" varias redes al mismo tiempo. A cada una le añade el arco con máxima información mutua. Después, calcula la longitud de descripción y siempre se queda con la red que tiene descripción mínima.

Los métodos de construcción de redes bayesianas constituyen un área con mucha actividad. Se han diseñado otros métodos nuevos que permiten trabajar con valores continuos, variables ocultas (que no se han reflejado en la base de datos) y variables para las que faltan valores en más de una observación (Ramoni, 1998). También es interesante la línea de trabajo en creación de métodos que permiten desarrollar algoritmos incrementales (Roure, 1999; Friedman, 1997), así como aquellos que permiten introducir alguna forma de conocimiento *a priori* (Castelo, 1998). Para una revisión exhaustiva, podéis consultar: Buntine, 1996; Heckerman, 1996; y, como más reciente, Sangüesa, 1997. Varias herramientas comerciales se basan en redes bayesianas que incorporan algún nivel de aprendizaje (Hugin) y herramientas integradas en sistemas comerciales de *data mining* (Castelo, 1997).

Lectura complementaria

Hallaréis el algoritmo de Lam y Bacchus en la obra siguiente:

W. Lam; F. Bacchus (1993). "Learning Bayesian Belief Networks, an Approach Based on the MDL Principle". *Computational Intelligence* (vol. 10, núm. 4, pág. 269-293).

3. Clasificación con redes bayesianas

La clasificación mediante la relación de la probabilidad *a priori* y *a posteriori* de una observación una vez conocida su etiqueta de clase es un método bastante antiguo, pero muy eficaz.

Lectura complementaria

Encontraréis el método de clasificación de redes bayesianas cuando se conoce la etiqueta de clase en la obra siguiente:

R.O. Duda; P.E. Hart (1973). *Pattern Classification and Science Analysis*. Nueva York: John Wiley & Sons.

Las suposiciones que siguen los clasificadores bayesianos son las siguientes:

a) El conjunto de las etiquetas de clase es exhaustivo y sus elementos, mutuamente excluyentes. Lo que acabamos de apuntar significa que no hay más clases que las que aparecen en las observaciones y las clases no se solapan. Por ejemplo, en el caso del gimnasio, los valores de la variable *Sexo,* de la variable *Horario* o de la variable *Entrenador personal* cumplen los citados requisitos.

b) Si se conoce la etiqueta de clase, mejor dicho, si se conoce la clase a la que pertenece un grupo de observaciones, entonces los atributos son condicionalmente independientes entre sí.

La última suposición ha hecho que los métodos de clasificación bayesianos hayan recibido el calificativo de "ingenuos", porque resulta bastante curioso que dentro de una clase determinada los valores que toman los atributos de las observaciones que pertenecen a dicha clase puedan hacer que éstos sean independientes. Los métodos de agregación* tratan de encontrar, precisamente los grupos de observaciones con una elevada influencia mutua entre variables. Por este motivo, durante un tiempo determinado, los métodos bayesianos ("ingenuos") han sido relegados y sólo se han tenido en cuenta como "vara de medida" para comparar otros métodos en apariencia más sofisticados.

*** Por ejemplo, Autoclass o COBWEB.**

El método para construir un clasificador de este tipo es bastante sencillo. Se trata de encontrar las probabilidades de aparición de los valores de cada atributo independientemente de los otros atributos, dada la clase. Cuando llega una observación nueva, sólo hay que tener en cuenta sus valores y calcular qué valor de clase es más probable una vez conocidas las probabilidades de todos los atributos (que se han estimado a partir de los datos en el paso anterior).

Encontrar la probabilidad de clase *a posteriori*, es decir, una vez conocidos los valores de la observación *a posteriori* no es más que aplicar el teorema de Bayes. Más sencillo, imposible.

Podéis ver el ejemplo de las lentes de contacto en el subapartado 2.1. del módulo "Clasificación: árboles de decisión" de esta asignatura.

Ejemplo de construcción de un clasificador con una red bayesiana

Tomemos un ejemplo muy sencillo para ver cómo puede funcionar el método "ingenuo" de clasificación con redes bayesianas. Aquí tenemos la tabla del ejemplo de las lentes de

contacto. La variable de clase es *Recomendación,* que adopta tres valores: lentes 'blandas', lentes 'duras' o 'ningún' tipo de lente.

Edad	Diagnóstico	Astigmatismo	Lágrima	Recomendación
Joven	Miope	No	Reducida	Ninguna
Joven	Miope	No	Normal	Blandas
Joven	Miope	Sí	Reducida	Ninguna
Joven	Miope	Sí	Normal	Duras
Joven	Hipermétrope	No	Reducida	Ninguna
Joven	Hipermétrope	No	Normal	Blandas
Joven	Hipermétrope	Sí	Reducida	Ninguna
Joven	Hipermétrope	Sí	Normal	Duras
Prepresbicia	Miope	No	Reducida	Ninguna
Prepresbicia	Miope	No	Normal	Blandas
Prepresbicia	Miope	Sí	Reducida	Ninguna
Prepresbicia	Miope	Sí	Normal	Duras
Prepresbicia	Hipermétrope	No	Reducida	Ninguna
Prepresbicia	Hipermétrope	No	Normal	Blandas
Prepresbicia	Hipermétrope	Sí	Reducida	Ninguna
Prepresbicia	Hipermétrope	Sí	Normal	Ninguna
Presbicia	Miope	No	Reducida	Ninguna
Presbicia	Miope	No	Normal	Ninguna
Presbicia	Miope	Sí	Reducida	Ninguna
Presbicia	Miope	Sí	Normal	Duras
Presbicia	Hipermétrope	No	Reducida	Ninguna
Presbicia	Hipermétrope	No	Normal	Blandas
Presbicia	Hipermétrope	Sí	Reducida	Ninguna
Presbicia	Hipermétrope	Sí	Normal	Ninguna

En la tabla siguiente resumimos los datos anteriores con el nombre de cada atributo, el valor y las diferentes configuraciones que pueden aparecer para cada valor del atributo de clasificación, que es el atributo *Recomendación* (lentes duras, blandas o ningún tipo de lente de contacto).

Edad				Diagnóstico				Astigmatismo				Lágrima			
	Ninguna	Blandas	Duras		Ninguna	Blandas	Duras		Ninguna	Blandas	Duras		Ninguna	Blandas	Duras
Joven	4	2	2	Miope	7	2	3	Sí	8	0	4	Normal	3	5	4
Prepresbicia	5	2	1	Hipermétrope	8	3	1	No	7	5	0	Reducida	12	0	0
Presbicia	6	1	1												

En total, hay cuatro observaciones en las que la recomendación fue 'Duras'; cinco, en las que fue 'Blandas' y quince, en las que la recomendación fue 'Ninguna'. Por lo tanto, la tabla de frecuencias es la que vemos a continuación:

Edad				**Diagnóstico**				**Astigmatismo**				**Lágrima**			
	Ninguna	**Blandas**	**Duras**		**Ninguna**	**Blandas**	**Duras**		**Ninguna**	**Blandas**	**Duras**		**Ninguna**	**Blandas**	**Duras**
Joven	4/15	2/5	2/4	Miope	7/15	2/5	3/4	Sí	8/15	0/0	4/4	Normal	3/15	5/5	4/4
Prepresbicia	5/15	2/5	1/4	Hipemétrope	8/15	3/5	1/4	No	7/15	5/5	0/4	Reducida	12/15	0/5	0/4
Presbicia	6/15	1/5	1/4												

Supongamos que aparece un paciente con la configuración de valores siguiente:

{'Joven', 'Hipermétrope', 'No astigmatismo', 'Lágrima normal'}.

¿Qué le recomendamos? En otras palabras, ¿a qué clase pertenece? Aplicamos el teorema de Bayes:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

Aquí, nuestra hipótesis es la clase y la evidencia, los valores que muestra este paciente para cada atributo. Teniendo en cuenta la independencia que hay entre los diferentes atributos, podemos expresarlo de esta forma:

$$P(Blandas|E) =$$

$$= \frac{P(Joven|Blandas)P(Hipermétrope|Blandas)P(No\ astigmatismo|Blandas)P(Lágrima\ Normal|Blandas)P(Blandas)}{P(E)} =$$

$$= \frac{\frac{2}{5}\times\frac{3}{5}\times\frac{5}{5}\times\frac{5}{24}}{P(E)}$$

$$P(Ninguna|E) =$$

$$= \frac{P(Joven|Ninguna)P(Hipermétrope|Ninguna)P(No\ astigmatismo|Ninguna)P(Lágrima\ Normal|Ninguna)P(Ninguna)}{P(E)} =$$

$$= \frac{\frac{4}{15}\times\frac{8}{15}\times\frac{7}{15}\times\frac{12}{15}\times\frac{4}{24}}{P(E)}$$

Normalizando, obtenemos: $P(Duras|E) = 0$, $P(Blandas|E) = 0{,}85$ y $P(Ninguna|E) = 0{,}13$.

Por lo tanto, un clasificador de este tipo recomendaría que nuestro paciente llevase lentes de contacto blandas. En otras palabras, predeciría que la clase que le corresponde es 'Blandas', o utilizando otra expresión igualmente equivalente, lo clasificaría dentro de la clase *Recomendación* = 'Blandas'.

La sorpresa es que los métodos bayesianos "ingenuos" dan unos resultados de clasificación muy buenos, en el sentido de que aseguran una precisión más elevada.

Lectura complementaria

Respecto a la precisión de los modelos bayesianos, podéis consultar la obra siguiente:

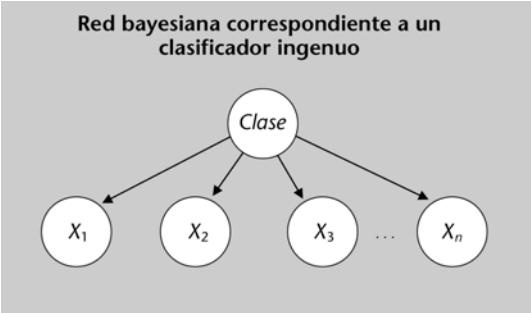
N. Friedman; Goldsmidt. (1997). "Sequential Update of Bayesian Network Structures". *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers.

Existen más factores que explican el interés suscitado por estos métodos. Principalmente, podemos decir que son métodos muy sencillos. No hay que efectuar una búsqueda en un espacio descomunal, sino sólo llevar un cálculo de recuento bastante fácil de realizar.

A continuación, veremos cómo debemos aplicar las redes bayesianas a un problema de clasificación. Como hemos dicho, una de las suposiciones básicas del

método bayesiano "ingenuo" consiste en suponer que los atributos son independientes entre sí, puesto que conocemos la clase.

Entonces, la red que tomaremos se parecerá más a un árbol que a un grafo general. En efecto, la variable *Clase* ocupa el nodo de ese árbol y los atributos están conectados a dicho nodo como hijos. Gráficamente, la situación es la que vemos a continuación:



La construcción de la red es bastante sencilla. En principio, para cada variable del conjunto inicial de variables $\{X_1, \ldots, X_n, C\}$, donde C representa la variable de clasificación, sólo hay que estimar las respectivas probabilidades condicionales y construir su red.

A la hora de la verdad, no se trata solamente de tener en cuenta la fiabilidad de las estimaciones según el tamaño de la base de datos y de asegurar que en efecto se cumpla que cada atributo lo es condicionalmente de los demás dada la clase.

Lectura recomendada

Para una revisión exhaustiva de métodos de clasificación bayesianos en general y los métodos correspondientes que utilizan redes bayesianas, podéis consultar la obra siguiente:

S. Acid. (1999). *Métodos de aprendizaje de redes bayesianas. Aplicación a la clasificación.* Tesis doctoral. Departamento de Ciencias de la Computación e Inteligencia Artificial. Universidad de Granada.

Resumen

Las redes bayesianas son un modelo que recoge la influencia entre los atributos de un dominio y la ponderan mediante las distribuciones de probabilidad condicionales entre los diferentes pares de variables que se pueden conectar en una relación padre-hijo dentro de un grafo dirigido acíclico.

Las redes bayesianas representan relaciones estructurales y cuantitativas al mismo tiempo.

El principal interés de estos modelos es que permiten efectuar operaciones de predicción y explicación con la misma representación.

Los métodos para construir redes bayesianas pueden basarse en las propiedades de independencia condicional de las diferentes estructuras o bien tener en cuenta las propiedades de las distribuciones implícitas en cada estructura.

Los métodos que se basan en las propiedades de las distribuciones se dividen, a su vez, en métodos basados en información, métodos bayesianos y métodos basados en el principio MDL.

Las redes también se pueden utilizar para clasificar operando bajo un principio parecido al de los clasificadores bayesianos “ingenuos”.

Actividades

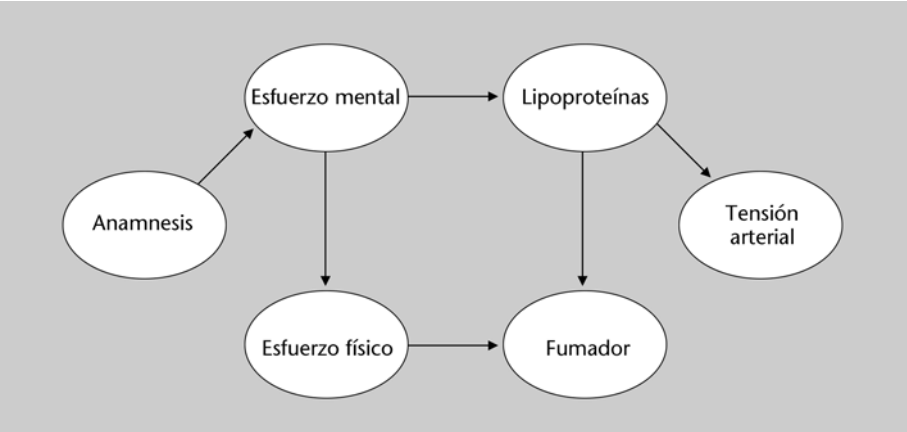
1. Acceded a la dirección de Internet que se da al margen y comparad las especificaciones de los diferentes *softwares* orientados a construir redes bayesianas.

Para hacer la actividad 1, acceded a la dirección *http://www.kdnuggets.com.*

2. Para el problema que os habíais propuesto en la actividad 1 del módulo "Extracción de conocimiento a partir de datos" de esta asignatura, ¿os sirven las redes bayesianas? ¿Qué método creéis que resultaría más conveniente?

Ejercicios de autoevaluación

1. Decid qué variables se pueden considerar condicionalmente independientes si tenemos la red bayesiana siguiente:



2. Dado el conjunto de datos que presentamos a continuación, clasificad las observaciones aplicando el método bayesiano "ingenuo". La variable de clasificación es *Entrenador personal*.

Cliente	Sexo	Renta	Edad	Años en el club	Entrenador personal
1	Mujer	6.000.000	40	2	No
4	Hombre	3.200.000	35	6	No
6	Mujer	0	30	3	No
7	Hombre	4.000.000	28	4	No
11	Hombre	10.000.000	60	10	Sí
14	Hombre	1.500.000	67	4	No
221	Mujer	0	32	3	Sí
61	Mujer	4.000.000	41	6	Sí
18	Mujer	0	32	4	No
19	Hombre	3.000.000	37	3	No
20	Hombre	2.800.000	32	3	No
21	Mujer	0	32	4	No
81	Mujer	3.200.000	33	2	No
84	Hombre	2.000.000	67	4	No
343	Mujer	0	20	1	Sí
31	Mujer	4.000.000	30	1	No

Bibliografía

Acid, S. (1999). *Métodos de aprendizaje de redes bayesianas. Aplicación a la clasificación*. Tesis doctoral. Departamento de Ciencias de la Computación e Inteligencia Artificial. Universidad de Granada.

Buntine,W. (1996). "A Guide to the Literature on Learning Probabilistic Networks from Data". *IEEE Transactions on Knowledge and Data Engineering* (núm. 8, págs. 195-210).

Castelo, R. (1997). *Bayesian Networks in Data Surveyor*. Proyecto de final de carrera. Facultad de Informática de Barcelona. Universidad Politécnica de Cataluña y Centrum voor Wiskunde en Informática. Amsterdam.

Castelo, R.; Siebes, A. (1998). "Priors on Networks Structures. Biasing the Search for Bayesian Networks". En: Sangüesa, R.; Cortés, U. (eds.). *Proceedings of the First Workshop on Causal Networks: from Inference to Data Mining. Sixth International Iberoamerican Conference on Artificial Intelligence*. Lisboa: IBERAMIA-98.

Castillo, E.; Gutiérrez, M.; Hadi, A. (1997). *Expert Systems and Probabilistic Network Models*. Springer Verlag.

Charniak, E. (1991). "Bayesian Networks without Tears". *AI Magazine* (vol. 12, núm. 4, págs. 50-63).

Chow, C.; Liu, S. (1968). "Approximating Discrete Probability Distributions with Dependence Trees". *IEEE Transactions on Information Theory* (núm. 14, págs. 462-467).

Cooper, G.F.; Herskovitz, E.H. (1992). "A Bayesian Method for the Induction of Probabilistic Networks from Data". *Machine Learning* (núm. 9, págs. 309-347).

De Campos, L.M.; Huete, J.F. (1997). "On the Use of Independence Relationships for Learning Simplified Belief Networks". *International Journal of Intelligent Systems* (vol. 12, núm. 7, págs. 495-522).

Duda, R.O.; Hart, P.E. (1973). *Pattern Classification and Science Analysis*. Nueva York: John Wiley & Sons.

Friedman, N.; Goldsmidt (1997). "Sequential Update of Bayesian Network Structures". *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers.

Heckerman, D. (1996). "Bayesian Networks for Knowledge Discovery". En: U. Fayyad; G. Piatetsky-Shapiro; P. Smyth; U. Uthurusamy (eds.). *Advances in Knowledge Discovery and Data Mining* (págs. 273-306). Menlo Park (CA, EE.UU.): AAAI Press.

Herskovitz, E.H.; Cooper, G.F. (1990). "Kutato: an Entropy-Driven System for the Construction of Probabilistic Expert Systems from Data". *Proceedings of the sixth conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers.

Huete, J.F. (1995). *Aprendizaje de redes de creencia. Modelos no probabilísticos*. Tesis doctoral. Departamento de Ciencias de la Computación e Inteligencia Artificial. Universidad de Granada.

Jensen, F.V. (1996). *An Introduction to Bayesian Networks*. UCL Press.

Lam, W.; Bacchus, F. (1993). "Learning Bayesian Belief Networks, an Approach Based on the MDL Principle". *Computational Intelligence* (vol. 10, núm. 4, págs. 269-293).

Neapolitan, R.E. (1990). *Probabilistic Reasoning in Expert Systems*. Nueva York: John Wiley & Sons.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems, Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, J.; Rebane, G. (1987). "The Recovery of Causal Poly-Trees from Statistical Data". *Uncertainty in Artificial Intelligence* (vol. 3, págs. 222-228).

Ramoni, M.; Sebastiani, P. (1998). "Parameter Estimation in Bayesian Networks from Incomplete Data". *Intelligent Data Analysis Journal* (núm. 2).

Robinson, R.W. (1977). "Counting Unlabeled Acyclic Graphs. Digraphs", Combinatorial Mathematics V. En: C.H.C. Little (ed.). *Springer Lecture Notes in Math. 622* (págs. 28-43). Springer-Verlag.

Roure, J.; Sangüesa, R. (1999). *A Survey on Incremental Methods for Bayesian Network Learning. Informe de Investigación LSI-99-42-R*. Departamento de Lenguajes y Sistemas Informáticos. Universidad Politécnica de Cataluña. Barcelona.

Sangüesa, R.; Cortés, U. (1997). "Learning Causal Networks from Data: a Survey and a New Algorithm for Learning Possibilistic Networks from Data". *AI Communications* (núm. 19, págs. 1-31).