Xarxes bayesianes

Ramon Sangüesa i Solé

P03/05054/01039

UOC
www.uoc.edu

Índex

Introducció

Les xarxes bayesianes són un model relativament recent, però que comença a tenir moltes aplicacions. Proposat inicialment per Pearl, aquest model és una representació que combina els aspectes qualitatius i quantitatius de les relacions entre els atributs (variables) d'un domini de manera força intuïtiva.

La representació en forma de graf que presenten i la sòlida base estadística que hi ha al darrere, les fa relativament fàcils d'entendre i utilitzar. En particular, la manera en què descriuen les relacions de dependència o influència entre variables i el desenvolupament d'una sèrie d'algorismes de propagació ben provats per a actualització, explicació i predicció mitjançant distribucions de probabilitat, fa que, a més de ser un bon model descriptiu d'un domini, permetin d'efectuar prediccions i trobar explicacions a situacions noves. En aquest sentit són models força més versàtils que la resta de models que s'han vist en altres mòduls.

Lectura complementària

Trobareu la presentació del model de xarxes bayesianes en l'obra següent:

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

Objectius

Després de treballar els materials didàctics en aquest mòdul, l'estudiant haurà assolit els objectius següents:

1. Conèixer les característiques principals de les xarxes bayesianes com a representació de les dependències entre els atributs d'un domini.

2. Aprendre com es poden fer servir les xarxes bayesianes per a diverses tasques de mineria de dades: representació d'associacions, predicció i explicació.

1. Què són les xarxes bayesianes?

Suposem que descrivim un domini (un conjunt de dades) mitjançant una sèrie d'atributs $X_1, \ldots, X_n$.

Una **xarxa bayesiana** és un graf dirigit acíclic on els nodes representen les variables X_i del domini (atributs), on cada variable és independent de totes les altres variables X_i, X_j del domini donats els seus predecessors directes.

Un **graf dirigit acíclic** és un tipus de graf en el qual la direcció dels enllaços és rellevant i en el qual mai no es pot produir que en un camí entre dos nodes, el node inicial o el final estiguin repetits.

Un **enllaç** entre dues variables X_i, X_j del domini representa una associació directa entre les dues. És a dir, X_i influeix sobre X_j.

Significat de les xarxes bayesianes

Una xarxa bayesiana representa les influències entre les variables d'un domini gràficament i probabilísticament alhora.

La influència existent entre dues variables que són extrems d'un enllaç, tals que X_i és l'origen de l'enllaç i X_j n'és l'extrem, està quantificada per la distribució condicional de probabilitat de les dues variables implicades: $P(X_i|X_j)$.

Exemple del viatge per carretera

Aclarim amb un exemple les propietats de les xarxes bayesianes, que poden semblar massa abstractes de bon començament.

Suposem que volem descriure les relacions que determinen allò que és important per a representar el cost i la durada d'un viatge per carretera.

Les variables d'interès són les següents:

- **Tipus de carretera**, que pren els valors {'Autopista', 'Autovia', 'Nacional', 'Comarcal', 'Pista'}.
- **Tipus de vehicle**, que pren els valors {'Esportiu', 'Utilitari', 'Familiar'}.
- **Velocitat mitjana en quilòmetres per hora**, que pren valors entre 0 i 150.
- **Cost de lloguer**, que pot prendre els valors {'Alt', 'Baix', 'Mitjà'}.
- **Durada del viatge en hores**, que pren valors entre 0 i 100.
- **Distància del recorregut en km**, que pot prendre valors entre 0 i 5.000.

Donat aquest conjunt d'atributs, esperem que s'esdevinguin les relacions següents:

a) El cost del lloguer estarà molt relacionat amb el tipus de vehicle llogat.

b) La velocitat mitjana dependrà del tipus de cotxe que portem i del tipus de carretera.

c) La durada del viatge dependrà de la velocitat mitjana i de la distància que caldrà recórrer.

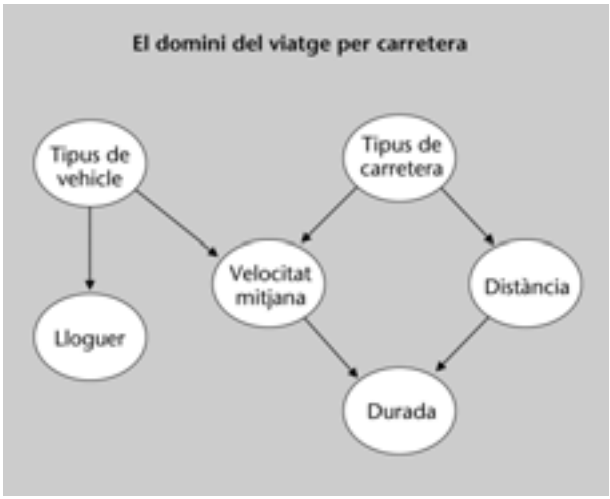
Lectura complementària

Trobareu una explicació detallada de l'exemple del viatge per carretera en l'obra següent:

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

Fixeu-vos que esperem que la durada del viatge depengui del tipus de vehicle i del tipus de carretera, però si coneixem la velocitat mitjana, aquests dos factors perden influència davant la velocitat mitjana i la distància en quilòmetres. La velocitat mitjana cobreix la durada, la "protegeix" de la influència d'altres factors (carretera i vehicle). De fet, és el mateix que dir que la durada del viatge és independent del tipus de carretera i del tipus de vehicle, si coneixem la velocitat mitjana que s'ha portat i la distància que calia recórrer. La variable *Velocitat mitjana* fa que la variable *Durada* sigui independent de les variables *Tipus de carretera* i *Tipus de cotxe*. Tècnicament hauríem de dir que la variable *Durada* és condicionalment independent de *Tipus de carretera* i *Tipus de vehicle*, donada la *Velocitat mitjana*.

Aquí tenim una representació gràfica que recull el coneixement de sentit comú que acabem d'expressar sobre aquest domini:



Aquesta estructura, per construcció i propietats de les xarxes bayesianes, assegura que es representen les propietats d'independència condicional que hem esmentat.

Hi ha un criteri gràfic, anomenat *criteri de d-separació*, que permet de "llegir" les relacions d'independència condicional de les diferents variables directament d'un graf com el que es reflecteix en l'exemple del viatge per carretera.

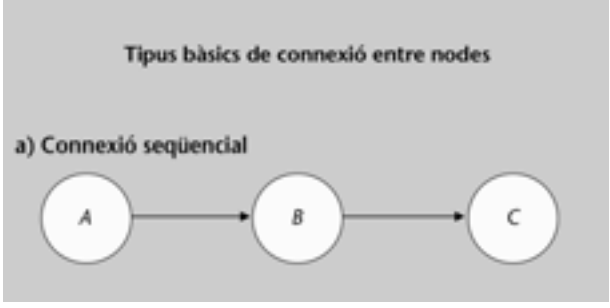
Lectura complementària

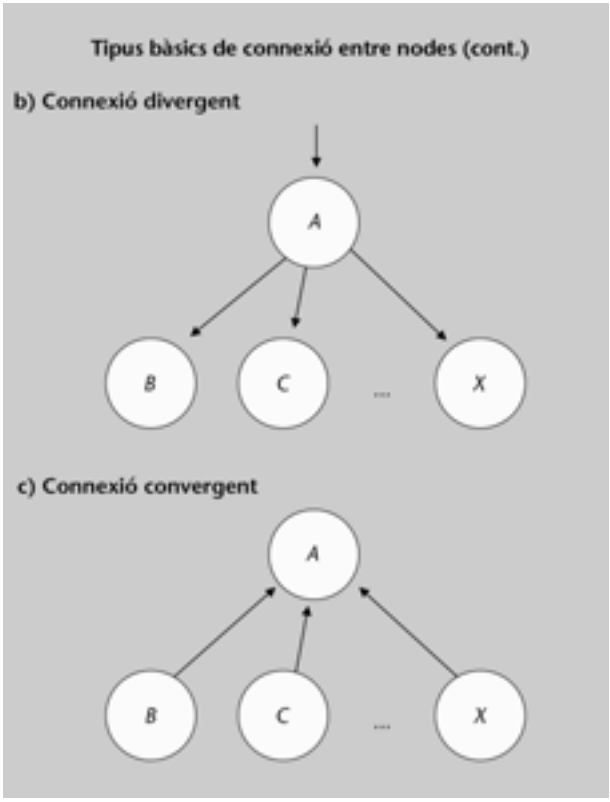
Trobareu el criteri de *d*-separació en l'obra següent:

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

1.1. Relacions qualitatives en les xarxes bayesianes: *d*-separació i models de dependències

Per a comprendre millor el criteri de *d*-separació, veurem com podem interpretar alguns tipus bàsics de connexions entre nodes d'una xarxa. En la figura següent representem aquests tipus bàsics de connexió:





Els tipus bàsics de connexió entre nodes són els que esmentem a continuació:

a) Connexió seqüencial: representada per l'esquema **a)** en la figura anterior. L'atribut *A* té influència sobre el *B* que, alhora, té influència sobre el *C*. Si tenim evidències sobre quin és el valor de *A* (per exemple, perquè n'hem observat el valor o perquè *A* té una certa probabilitat de prendre'n un de determinat), llavors podem modular la certesa sobre els valors que pot prendre *B,* i aleshores propagar aquesta influència sobre *C*. D'altra banda, l'evidència que tinguem sobre els valors de *C* influirà en la certesa que tenim sobre els valors que pot prendre *A* a través de *B*. Ara bé, si coneixem el valor de la variable *B*, llavors el camí entre *A* i *C* queda bloquejat: *A* i *C* es fan independents, *A* i *C* estan *d*-separades per *B*.

b) Connexió divergent: en la situació **b)** de la figura anterior, la influència pot passar per tots els descendents de *A,* a menys que coneguem el valor de *A*. Diem que *B*, *C*, …, *X* estan *d*-separats, conegut el valor de *A*.

c) Connexió convergent: en la situació **c)** de la figura, si no coneixem res del valor de *A,* a part del fet que se'n pot inferir un a partir dels valors coneguts dels pares *B*, *C*, …, *X*, llavors els pares són independents entre si; és a dir, el coneixement del valor d'un dels pares no té cap influència en els valors que en poden prendre els altres.

Aquests tres casos cobreixen totes les formes de transmissió d'evidència a través d'una variable. Seguint aquestes tres regles és possible decidir per a qualsevol parell de variables de la xarxa si són dependents o no.

Es diu que dues variables X i Y en una xarxa bayesiana estan *d*-separades per una altra variable Z si per a tots els camins entre X i Y hi ha una variable intermèdia Z tal que la seva connexió és seqüencial o divergent i l'estat de Z és conegut, o bé la connexió és convergent i ni Z ni cap dels descendents de Z no han rebut cap evidència.

Aquesta definició es pot estendre a conjunts de variables en comptes de variables individuals.

És interessant adonar-se que amb aquest criteri es pot extreure el model d'independències associat a una xarxa bayesiana.

El **model d'independències** és un conjunt format per una col·lecció d'assercions del tipus "el conjunt de variables X és independent de Y conegut Z".

Denotem les assercions del model d'independències amb $I(X|Z|Y)$. Inversament, hi ha procediments per a construir una xarxa que, donat un model d'independències, retornen la xarxa bayesiana corresponent.

El problema és que, donat un model d'independències i una xarxa bayesiana definida sobre el mateix domini, poden passar tres coses:

a) Totes les relacions d'independència que estan presents en el model es poden detectar en el graf per mitjà de *d*-separació. Diem que el graf és un *D-map* del model de dependències.

b) Totes les relacions d'independència que es poden detectar en el graf mitjançant *d*-separació estan presents en el model de dependències. Llavors diem que el graf és un *I-map* del model de dependències.

c) En el graf només es localitzen les relacions de dependència del model i en el model només apareixen les relacions de dependència del graf. Llavors diem que el graf és un *P-map* o *Perfect map* del model.

Per a un mateix model de dependències hi pot haver diversos grafs que les representin. Per a la mineria de dades el que ens interessa tenir en compte és el següent:

1) Una base de dades definida sobre un conjunt d'atributs $X_1, \ldots, X_n$ permet d'extreure un conjunt de dependències.

2) Un mètode de construcció de xarxes bayesianes ha d'assegurar que el graf que resulti de l'aplicació sobre un conjunt de dades ha de ser, com a mínim, un *I-map* del conjunt de dependències existent i, idealment, un *P-map*.

Lectura recomanada

Consulteu els procediments per a construir una xarxa bayesiana a partir d'un model d'independències en l'article següent:

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

Ja veurem més endavant com afecta això el disseny de mètodes de construcció de xarxes bayesianes. Ara hem de conèixer altres propietats d'aquest tipus de model.

Vegeu els mètodes de construcció de xarxes bayesianes a partir de dades en l'apartat 2 d'aquest mòdul.

1.2. Relacions quantitatives en les xarxes bayesianes: probabilitats condicionals

Les relacions qualitatives (estructura de connexions del graf, model de dependències implícit en les relacions de *d*-separació) d'una xarxa bayesiana són complementades per relacions quantitatives, que corresponen a les distribucions de probabilitat condicional existents entre els diversos nodes que tenen una relació de parentiu directe.

Si tenim un domini $X_1, \ldots, X_n$ sobre el qual definim una distribució de probabilitat $P(X_1,\ldots,X_n)$, recordem que dues variables X_i, X_j són independents, donada una tercera, Z, si es compleix la relació següent:

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

Per les relacions d'independència condicional existents en una xarxa bayesiana podem veure que la distribució de probabilitat conjunta $P(X_1,\ldots,X_n)$ es pot factoritzar de la manera següent:

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

on pa_i és el conjunt d'antecessors directes (pares) de la variable X_i. L'única cosa que hem fet és aplicar la regla de la cadena en probabilitats que estableix el següent:

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

En efecte, hem aprofitat la relació d'ordre entre les variables que estableixen les connexions de parentiu. Hem de tenir en compte que per a cada variable X_i, el seu conjunt de pares, $pa_i \subseteq \{X_1, \ldots, X_n\}$, fa que X_i i $\{X_1, \ldots, X_{i-1}\}$ es converteixin en independents i, precisament, sabem que podem expressar una probabilitat conjunta $P(X_1, \ldots, X_n)$ com a producte de les probabilitats marginals $P(X_1), \ldots, P(X_n)$ quan les variables són independents entre si.

Aquest és un dels punts forts de les xarxes bayesianes com a models de representació del coneixement. En efecte, sense aquesta característica que ens permet de factoritzar la distribució conjunta tenint en compte l'estructura de la xarxa, per a n variables necessitaríem expressar la distribució conjunta en termes de 2^n distribucions de probabilitat de variables per a especificar la mateixa distribució. En canvi, ara queda reduït a les probabilitats condicionals que s'han

d'especificar entre pares i fills. Com que en un graf dirigit acíclic, el nombre d'enllaços possibles per a n nodes és $n(n-1)$, la reducció és prou significativa. Les xarxes bayesianes forneixen una representació compacta.

Finalment, cal tenir en compte que per a especificar les característiques de les distribucions de probabilitat condicional presents en la xarxa ens calen uns quants paràmetres estadístics, tants com tantes combinacions de valors de variables x_i^j per a cada variable X_i i per a cadascuna de les configuracions de pares possibles per a cada variable x_i^j hi hagi. És a dir, s'han d'especificar els paràmetres següents:

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

Des del punt de vista de la mineria de dades, les propietats quantitatives també es poden utilitzar per a extreure la xarxa o les xarxes bayesianes "ocultes" dins un conjunt de dades. En efecte, una base de dades definida sobre $X_1, \ldots, X_n$ té associada una distribució de probabilitat $P(X_1, \ldots, X_n)$ i la xarxa bayesiana que volem obtenir també representa una factorització de la mateixa distribució de probabilitat.

El problema es pot plantejar llavors de la manera següent:

a) Cal estimar els paràmetres θ_{ijk} de la distribució de probabilitat implícita en les dades. En principi, no hem de conèixer aquesta distribució.

b) Hem de trobar la xarxa bayesiana que s'adiu amb la distribució caracteritzada per aquests paràmetres. Normalment, caldrà escollir entre les xarxes possibles la que s'ajusti millor a la distribució implícita en les dades. Cal, doncs, tenir una **mesura d'ajust**.

Per tant, el que cal fer, en general, és una doble cerca: en l'espai de paràmetres i en l'espai d'estructures que s'adiuen amb els paràmetres. A més, cal assegurar que la xarxa recuperada és, com a mínim, un *I-map*. No és un problema petit, però la utilitat pràctica dels models recuperats paga la pena.

Vegeu els *I-map* en el subapartat 1.1 d'aquest mòdul.

Hi ha diversos algorismes que permeten d'efectuar dues operacions bàsiques basades en la propagació de valors de probabilitat. Vegem molt breument quines són aquestes operacions que permet fer una xarxa bayesiana.

1.2.1. Operacions sobre una xarxa bayesiana

Les operacions mínimes que permet de fer una xarxa bayesiana són la **predicció** i l'**explicació**. Totes dues es basen en la propagació d'evidències dins de la xarxa.

Els **algorismes de propagació d'evidències** s'encarreguen de veure dins una xarxa bayesiana com afecta l'evidència que una variable X_i pren un valor x_i determinat els valors que poden prendre la resta de variables.

Propagació d'evidències

La propagació d'evidències es basa en un vell teorema de l'estadística, el teorema de Bayes, que indica com s'actualitzen els valors de les variables en acumular evidències noves.

Per a efectuar aquests càlculs cal actualitzar les diferents distribucions de probabilitat. Necessitem conèixer els elements següents:

a) La distribució *a priori* de les variables arrel (les que no tenen cap antecessor o predecessor).

b) Les distribucions condicionals entre una variable i els seus pares, que ja estan codificades en la mateixa xarxa.

Lectures complementàries

Trobareu l'explicació concreta dels algorismes de propagació en les obres que esmentem a continuació. La tercera obra en presenta l'exposició més actualitzada. L'última obra esmentada és una referència excel·lent per a la munió d'algorismes de propagació exactes i aproximats existents. En trobareu la referència completa en l'apartat de bibliografia del mòdul.

J. Pearl (1988).

R.E. Neapolitan (1990).

F.V. Jensen (1996).

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

L'explicació concreta dels algorismes de propagació supera els objectius d'aquest curs. Ara només ens cal saber que aquests algorismes es basen en una extensió molt hàbil de la relació existent entre les probabilitats *a priori* i *a posteriori* coneguda com el *teorema de Bayes*.

El **teorema de Bayes** estableix que, donada una hipòtesi (H) i una evidència (E), es verifica la relació següent:

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

A priori, la hipòtesi pot tenir una certa distribució de probabilitat.

Exemple de propagació d'evidències

Reprenem l'exemple inicial del viatge per carretera. Sense saber res més, podríem conèixer que la probabilitat que el viatge duri tres hores és del 50%.

Si tenim la hipòtesi que la durada del viatge per carretera és de 3 hores i tenim l'evidència que la carretera és una nacional i el cotxe, un utilitari, esperem que la probabilitat *a posteriori* que el viatge duri 3 hores, atès el fet que sabem que el cotxe és un utilitari i la carretera és nacional, sigui diferent de la que teníem abans d'aplegar aquestes evidències (o no, si les variables són independents).

Vegeu l'"Exemple del viatge per carretera" a l'inici d'aquest apartat.

Ara i aquí no és el moment d'entrar en més detalls ni sobre el teorema de Bayes ni sobre els algorismes de propagació i actualització d'evidències en xarxes bayesianes. Ja hem remarcat on se'n poden trobar referències més extenses i detallades.

Posarem un exemple per tal d'aclarir les possibilitats i l'interès de fer consultes sobre les xarxes.

Interès de les xarxes bayesianes

L'interès de les xarxes bayesianes rau en la varietat i potència d'operacions que permeten de fer.

Suposem que tenim una xarxa bayesiana extreta de la base de dades d'HyperGym. Utilitzarem una eina de construcció de xarxes bayesianes a partir de dades (BKD, *bayesian knowledge discoverer*).

És interessant adonar-se que, d'entrada, només per inspecció visual, l'esquema de la xarxa ja ens diu força coses. Per exemple, els atributs següents no tenen gaire relació amb la resta:

- *Client*: si tenim en compte que es tracta de l'identificador de client (diferent per a cada observació) no és estrany que no es pugui establir cap relació amb la resta de variables. De fet, hauria estat més normal no considerar aquesta variable en l'estudi.

- *Local* i *Districte de residència*: no semblen influir en la resta de variables ni influir-se mútuament, però entre les dues hi ha una relació forta. Es pot interpretar que conèixer a quin local assisteix un client ens dóna informació respecte a quin és el seu districte de residència.

Activitat

1.1. Proveu d'extreure vosaltres mateixos les relacions d'independència condicional expressades per l'estructura de la xarxa que acabeu de veure en la figura anterior.

Ens interessa veure què podem fer amb aquesta xarxa:

1) D'entrada, el model ens diu les diferents distribucions de probabilitat condicional existents.

Distribucions de probabilitat condicional de les variables del model

La distribució de probabilitat condicional entre les variables *Sexe* i *Horari* és la que s'explicita en la taula següent:

		Horari	
		Matí	**Tarda**
Sexe	**Home**	0,770	0,230
	Dona	0,551	0,449

Entre les variables *Activitat* i *Sexe*, la taula corresponent és la que veiem a continuació:

		Activitat				
		Aeròbic	**TBC**	**Ioga**	***Stretch***	***Steps***
Sexe	**Home**	0,358	0,477	0,023	0,066	0,077
	Dona	0,737	0,156	0,036	0,019	0,053

La relació entre *Entrenador personal* i la primera activitat que desenvolupa el client té la taula de probabilitat condicional següent:

		Activitat				
		Aeròbic	**TBC**	**Ioga**	***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 model també ens dóna les probabilitats *a priori* de totes les variables.

Probabilitats *a priori* de totes les variables del model

Fixem-nos en la distribució *a priori* de la variable *Entrenador personal* en la base de dades, que indica que és un servei majoritàriament no demanat: el 90,5% de les observacions tenen el valor 'No' per a aquest atribut i el 0,5% restant té el valor 'Sí'. Si ara observem que un dels clients és una dona (Evidència = 1), i introduïm aquesta observació en el model, podem veure si hi ha canvis. Efectivament, el valor de la distribució *a posteriori* d'*Entrenador personal* ha baixat fins al 94%. És un canvi poc espectacular, si voleu, però prou interessant si tenim en compte que la relació entre la variable *Sexe* i la variable *Entrenador personal* està mediatitzada per l'activitat principal (*Act1*) que es desenvolupa.

3) També es poden fer altres tipus de consulta, com veure quins canvis hi ha després d'actualitzar simultàniament dues variables (per exemple, *Sexe* i *Activitat*).

4) Una altra característica interessant de les xarxes bayesianes és la que incorporen els algorismes d'explicació.

Els **algorismes d'explicació** retornen la configuració de variables i valors més probable a partir del valor observat d'una o més variables.

Algorismes d'explicació

Si veiem que algú ha demanat un entrenador personal, l'algorisme d'explicació ens torna el conjunt {(*Home* = 'Sí'), (*Act1* = 'TBC'), (*Renda* = 3.000.000 – 10.000.000)}. És a dir, l'algorisme ens diu que la causa més probable que algú demani un entrenador personal és que sigui un home que practica TBC com a activitat principal i tingui una renda alta. Fixeu-vos que no apareixen altres variables perquè no contribueixen amb prou evidència a explicar aquesta observació. Aquest tipus de procés és molt interessant en problemes de diagnosi que relacionen símptomes i causes: donat un valor observat per a un símptoma, ens retorna el conjunt de causes més probables.

Finalment, les xarxes bayesianes també es poden utilitzar per a portar endavant tasques de classificació.

Per tant, podem considerar que són un model un tant costós de construir, però força útil.

Lectura complementària

Per a una introducció fàcil als mètodes de propagació podeu consultar l'article següent:

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

2. Mètodes de construcció de xarxes bayesianes a partir de dades

Evidentment la manera més directa de construir una xarxa bayesiana és a partir del coneixement d'un expert que ens indiqui quins són els atributs rellevants i quina relació presenten entre si, i que ens digui quina és la força d'associació. Però ens interessa conèixer de quina manera podem extreure automàticament una xarxa bayesiana a partir d'un conjunt de dades que descriu un domini.

El **problema de la construcció d'una xarxa bayesiana** a partir d'un conjunt de dades es pot expressar tal com presentem a continuació.

> Donat un conjunt de dades, el problema de la construcció d'una xarxa bayesiana consisteix a extreure la topologia de la xarxa que hi està implícitament representada i la seva distribució de probabilitat corresponent.

En una aproximació ingènua es pot pensar que es tracta d'anar provant diverses combinacions de models, construïts a còpia de connectar nodes diferents i en direccions diferents. La magnitud del nombre de models possibles que es poden obtenir ens barra aquesta mena d'aproximació. Robinson va calcular el nombre de grafs dirigits acíclics possibles que es poden extreure d'un conjunt de n nodes (que corresponen a les n variables $X_1, \ldots, X_n$ del domini) amb l'ajuda d'aquesta impressionant fórmula recursiva:

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

amb $f(0) = 1$ i $f(1) = 1$. Si calculeu una mica podreu veure que, només amb nou nodes, ja estem en l'ordre dels milers de milions de possibles grafs. Si teniu en compte la quantitat de variables que cal per a descriure un domini real, podreu veure que no es tracta d'un espai de cerca precisament petit.

Lectura complementària

Trobareu la deducció de la fórmula recursiva de Robinson en l'obra següent:

R.W. Robinson (1977). "Counting Unlabeled Acyclic Graphs. A: C. Little (ed.). Lectures Notes in Mathematics 622: Combinatorial Mathematics (pàg. 28-43). Nova York: Springer-Verlag.

Per tant, cal recórrer a una estratègia de cerca heurística. Cal establir alguna mesura de qualitat que ens serveixi per a explorar aquest espai de cerca tan gran. Hi ha tres tipus de mètodes:

- Mètodes basats en les propietats de la distribució de probabilitat.
- Mètodes basats en propietats d'independència.
- Mètodes híbrids.

Lectures complementàries

Per a un tractament més aprofundit dels mètodes basats en propietats d'independència i híbrids, consulteu les obres següents. En trobareu la referència completa en l'apartat de bibliografia del mòdul.

R. Sangüesa; U. Cortés (1997).

J.F. Huete (1995).

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

Els més coneguts i emprats són els primers i són els únics que descriurem amb un cert detall.

2.1. Mètodes basats en propietats de la distribució de probabilitat

Recordem que la propietat de factorització de la distribució de probabilitat conjunta de les xarxes bayesianes ens permet d'expressar la distribució que correspon a un model de xarxa bayesiana que es construeix en un moment donat del procés de cerca com un multiplicatori de les probabilitats de cada node condicionades als seus pares.

Vegeu la propietat de factorització de la distribució de probabilitat conjunta en el subapartat 1.2 d'aquest mòdul didàctic.

Per tant, en el procés de construcció d'una xarxa a partir de les dades, en tot moment tenim una disparitat entre la distribució conjunta (expressada com un multiplicatori que correspon a l'estructura pares-fills de la xarxa construïda fins aquell moment) i la distribució que suposem que hi ha en les dades, que sí que admet la forma de multiplicació de probabilitats marginals.

Per tant, es tracta de trobar l'estructura de xarxa bayesiana de manera que la seva distribució de probabilitat conjunta (expressada com un multiplicatori que segueix l'estructura pares-fills) sigui la més propera a la que suposem que està implícita en les dades.

Aquest és un problema típic d'extracció de models a partir de dades i admet diverses formes d'atac. Les mesures d'ajust es poden expressar mitjançant els criteris següents:

a) Criteris basats en l'entropia: es tracta de trobar la xarxa amb una entropia creuada o divergència de Kullback-Leibler més petita.

b) Criteris basats en estimació bayesiana: es tracta de trobar l'estructura i el conjunt de distribucions amb la màxima probabilitat *a posteriori* (MAP), donades les dades existents.

MAP és la sigla de l'expressió *màxima probabilitat a posteriori*.

c) Criteris basats en el principi MDL: cerca trobar el model que té la mínima codificació, donades les dades, i que permet de codificar les dades amb la mínima longitud de descripció.

Vegeu la MDL en el subapartat 5.2.1 del mòdul "Agregació (*clustering*)" d'aquesta assignatura.

Tots aquests mètodes han derivat alguna manera d'establir la qualitat global de la xarxa en construcció durant el procés de cerca segons els seus components, i han reduït les mesures de qualitat globals a expressions segons les relacions entre les variables i els seus pares. Recordem que això és possible gràcies a la propietat de factorització de les xarxes. En general, totes aquestes mesures tenen una forma semblant a l'expressió següent:

$$Qualitat(Xarxa|Dades) = \sum_{X_i} Qualitat(X_i|pa_i, Dades).$$

2.1.1. Mètodes basat en l'entropia

El mètode més vell per a construir una estructura semblant a una xarxa bayesiana va ser proposat per Chow i Liu. Aquest mètode fa ús de l'entropia i la informació mútues per a construir la xarxa bayesiana. Nosaltres presentarem el mètode Kutató, que deriva en part d'aquell mètode.

Lectures complementàries

Trobareu el mètode proposat per Chow i Liu en la primera de les obres que esmentem a continuació, i el mètode que seguim aquí en la segona.

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

L'**entropia** es pot considerar com una mesura de la quantitat d'informació present en una distribució de probabilitat o com una mesura de la incertesa associada a una variable. Aquest concepte també es veu en parlar d'arbres de decisió, en concret en explicar com ID3 decideix sobre l'homogeneïtat d'una partició, i també en parlar dels mètodes de discretització. Recordem-ne la definició:

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

En aquesta expressió, r denota el nombre de valors possibles que pot prendre la variable X.

Sabem que l'entropia compleix la propietat de ser una mesura positiva que arriba al seu mínim quan la incertesa de la distribució és mínima i viceversa.

L'**entropia conjunta** es pot definir de la manera següent:

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

on r_x i r_y són les cardinalitats dels conjunts de valors que poden prendre X i Y, respectivament. Aquesta definició es pot estendre a un conjunt de n variables $X_1, \ldots, X_n$.

L'**entropia condicional de X, atès que $Y = y_j$**, és l'entropia de la distribució condicional $P(X|Y = y_j)$:

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

L'**entropia condicional de X respecte a Y** és la mitjana del valor de l'entropia de X respecte als valor 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 l'entropia conjunta i la condicional, de la manera següent:

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

La **informació mútua** entre dues variables X i Y mesura la mitjana de la reducció a la incertesa sobre X que provoca el fet de tenir informació sobre el valor de Y i viceversa. De la mateixa manera es pot dir que la informació mútua mesura la quantitat d'informació mitjana que Y aporta sobre X o també el grau de restricció que una variable aporta sobre l'altra:

$$I(X;Y) = H(X) - H(X|Y) =$$

$$= \sum_{i,j=1}^{r_X,r_Y} P(x_i,y_j)\log_2\frac{P(x_i|y_j)}{P(x_i)P(y_j)}.$$

L'**entropia creuada** o **divergència de Kullback-Leibler** entre dues distribucions de probabilitat P i P' és la següent:

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

L'algorisme de Chow i Liu construeix un graf en forma d'arbre, on cada branca connecta les variables amb informació mútua màxima. El mèrit va ser demostrar que aquest mètode de construcció sempre troba l'arbre amb divergència mínima. Per tant, aquest mètode recupera l'estructura que més informació aporta i la distribució de probabilitat de la informació que és més semblant a la que hi ha implícita en les dades.

Herskovitz i Cooper van dissenyar un mètode per a recuperar l'estructura d'una xarxa bayesiana, donat un conjunt de dades que feia servir l'entropia com a mesura de qualitat. El mètode considera que la xarxa d'entropia mínima és la més informativa. L'entropia de la xarxa sempre és més gran que la de la distribució del conjunt de dades.

Lectura complementària

Trobareu l'algorisme de Herskovits i Cooper en l'obra següent:

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.

L'entropia per a una xarxa bayesiana B_S es calcula com la suma de les entropies condicionals de cadascuna de les variables X_i, donats els seus pares pa_i.

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

Aquesta fórmula calcula l'entropia de la xarxa tenint en compte els factors següents:

- r_i és el nombre de valors que pot prendre la variable X_i.
- x_i^k representa el valor k-èsim que pot prendre la variable X_i.
- q_i és el nombre de configuracions de pares de X_i, pa_i possibles.
- pa_i^j és la j-èsima configuració de pares present en la base de dades.

L'**algorisme Kutato** comença amb un graf format per tots els nodes sense cap connexió entre si. A cada pas afegeix l'arc que produeix l'estructura amb la mínima entropia. S'atura quan l'estructura B_s assoleix un nivell d'entropia prou baix.

A continuació presentem l'algorisme que segueix el mètode Kutato. A l'entrada de l'algorisme hi ha una base de dades D sobre un conjunt de variables $\{X_1, \ldots, X_n\}$, un valor límit d'entropia inferior, α, i un ordre sobre les variables. Aleshores, el mètode segueix els passos següents:

1) Construir una estructura sobre $\{X_1, \ldots, X_n\}$ i suposar que les variables són marginalment independents {graf inconnex}.

2) $\beta = H(B_S)$ {calcular l'entropia de la xarxa}.

3) Repetir fins que $\beta \leq \alpha$ el bucle següent:

a) Seleccionar un enllaç tal que:

- No crea cap cicle.
- És el que crea la nova estructura B_s amb entropia mínima.
- Enllaça les variables X i Y de manera que X és anterior en l'ordre definit.

b) Donar l'orientació $X \Rightarrow Y$.

A continuació comentarem el mètode que els mateixos autors van proposar posteriorment, basat en la inferència bayesiana.

El mètode K2

La intuïció que hi ha darrere del mètode Kukato és trobar l'estructura més probable, donat el conjunt d'observacions recollit en la base de dades inicials, i alguna informació sobre les distribucions *a priori* de les estructures existents possibles (aquesta informació es redueix a suposar que totes les estructures de xarxa bayesiana són igualment probables o segueixen una distribució normal).

La idea es pot expressar mitjançant el teorema de Bayes un altre cop. Suposem que volem trobar la xarxa bayesiana B_S, on S denota el parell (B_S, B_P) i B_P són les distribucions de probabilitat condicionals associades a l'estructura. La probabilitat de la xarxa, donades les dades D, es pot expressar d'aquesta manera:

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

En realitat, ens interessa comparar xarxes possibles entre si; és a dir, donades dues xarxes possibles B_{S1} i B_{S2}, n'hem de comparar la raó:

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

Per tant, el mètode K2 fa l'aproximació utilitzant la probabilitat d'una estructura procedent de les dades per aproximar-se a la probabilitat condicional corresponent. Ara bé, calcular aquesta probabilitat no és pas tan senzill, fins i tot tenint en compte les simplificacions següents:

1) Els atributs que apareixen en la base de dades són discrets.

2) Les observacions dins una base de dades són independents entre si donada una estructura de xarxa bayesiana.

3) No hi ha observacions a les quals falti algun valor.

4) Abans d'observar les dades D no tenim cap preferència respecte a les probabilitats numèriques que cal assignar a la xarxa bayesiana.

Lectura complementària

Si esteu interessats en la deducció de l'expressió heurística *g*, podeu consultar l'article següent:

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

Totes aquestes simplificacions permeten de derivar un mètode heurístic per trobar la xarxa que maximitza la probabilitat *a posteriori*, donades les dades. El procés de derivació d'aquest procés és força complicat i no el reproduïm aquí. La fórmula final del mètode heurístic, denotat per *g*, és prou notable:

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

on cada factor de l'expressió té el significat següent:

- pa_i és el conjunt de pares de X_i.
- q_i és el nombre de configuracions diferents que apareixen en la base de dades per als pares de X_i; és a dir, les diferents assignacions de valors que els pares prenen en la base de dades.
- r_i és el nombre de valors que pot prendre la variable X_i.
- N_{ijk} denota el nombre d'observacions en què la variable X_i pren en la base de dades el valor k-èsim d'entre els que pot prendre, i la configuració de valors dels pares és la j-èsima, d'entre les que hi ha en la base de dades.
- N_{ij} és la suma de configuracions possibles per a cada valor de X_i.

Afortunadament, l'algorisme és conceptualment molt senzill. Comença per una variable sense pares i se n'hi van afegint. Es tracta de seleccionar en tot moment la variable que, conjuntament amb els pares de la variable que es considera en un moment donat, maximitza el valor de g. Per a facilitar la selecció de les variables que cal connectar entre si, es declara un ordre inicial entre les variables. A més s'imposa un nombre màxim de pares per variable. Quan el fet d'afegir un nou pare a la variable no pot contribuir a incrementar-ne la probabilitat, l'algorisme deixa d'afegir pares a una variable i passa a considerar-ne una altra.

L'algorisme K2 parteix dels elements següents: una base de dades sobre un domini de variables $\{X_1, \ldots, X_n\}$, un ordre entre les variables i un nombre màxim de pares per variable, u. Aleshores, l'algorisme consisteix a efectuar els passos següents:

1) Construir una estructura B_s i suposar que totes les variables són independents marginalment (és a dir, crear un graf inconnex).

2) Executar el bucle següent:

Per a $i := 1$ **fins a** n **fer**
  $pa_i := \emptyset$
  $P_{anterior} := g(X_i, pa_i)$
  $OK :=$ cert
  Mentre OK i $(|pa_i| < u)$ **fer**
    Sigui z la variable anterior a X, $z \notin pa_i$
    tal que maximitza $g(X_i, pa_i \cup \{z\})$
    $P_{actual} := g(X_i, pa_i \cup \{z\})$
    si $P_{actual} > P_{anterior}$ **aleshores**
      $P_{anterior} := P_{actual}$
      $pa_i := pa_i \cup \{z\}$
    sinó
      $OK :=$ fals
    fsi
  fmentre
fper

2.1.2. Mètodes basats en el principi de la mínima longitud de descripció

Ja hem comentat en parlar de processos de discretització quina era la lògica que seguia el principi de la mínima longitud de descripció. En aquest cas, es tracta de poder trobar un xarxa bayesiana a partir d'un conjunt d'observacions tals que es minimitzi la codificació de la xarxa i de les dades, donada la xarxa.

Vegeu el principi de mínima longitud de descripció en el subapartat 2.3.2 del mòdul "Classificació: arbres de decisió" d'aquesta assignatura.

Els **mètodes basats en el principi de la mínima longitud de descripció** consisteixen, en primer lloc, a trobar una manera de codificar una xarxa bayesiana i, un cop coneguda la xarxa d'on se suposa que s'han derivat les dades, codificar-les.

Codificació de la xarxa

Codificarem la xarxa entesa com la combinació de l'estructura B_S i la llista de probabilitats condicionals associades a cada node, B_P. Cal, doncs, codificar l'estructura i la llista de probabilitats condicionals.

Suposem que hi ha n variables en el conjunt de dades, $\{X_1, \ldots, X_n\}$. Per a una variable X_i, representada com un node dins el graf, que té $|pa_i|$ pares diferents, necessitem $|pa_i| \log_2(n)$ bits per a codificar la llista dels seus pares. En total, per a tota la xarxa necessitem la quantitat de bits determinada per aquesta expressió:

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

Per a calcular quants bits calen per a codificar les probabilitats condicionals de cada node X_i cal multiplicar el nombre de bits necessaris per a codificar el valor numèric de cada probabilitat condicional i el nombre total de probabilitats condicionals. El nombre de bits és representat per l'expressió següent:

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

on, com sempre, r_i és el nombre de valors que pot prendre la variable X_i i q_i és el nombre de configuracions que prenen els seus pares. Finalment, d és el nombre de bits necessaris per a codificar un valor numèric (els valors que prenen els diversos valors de la probabilitat de la distribució condicional).

Per tant, la suma total de la codificació de l'estructura i de les distribucions condicionals és la que dóna la quantitat següent:

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

La codificació de les dades es fa tenint en compte que tenim un model format per l'estructura més les distribucions condicionals de probabilitat. Tal com s'explica, en parlar per primera vegada de l'MDL, la codificació es fa utilitzant l'algorisme de Huffman.

Vegeu l'algorisme de Huffman en el subapartat 5.2.1 del mòdul "Agregació (*clustering*)" d'aquesta assignatura.

L'**algorisme de Lam i Bacchus** fa servir un mètode basat en el principi MDL per a construir xarxes bayesianes a partir de dades. Aquest algorisme fa servir

la mesura d'informació mútua entre una variable i els seus pares per a seleccionar les variables que cal connectar en un moment donat. La mesura de la informació mútua entre una variable X_i i els seus pares pa_i és la següent:

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

Tal com van demostrar Chow i Liu per al cas dels grafs estructurats com a arbres, si es troba l'arbre d'expansió maximal amb pesos iguals a la informació mútua entre cada node (per exemple, amb l'algorisme de Kruskal) s'aconsegueix la distribució que té la mínima divergència de Kullback respecte a les dades. Lam i Bacchus van fer una demostració semblant per al cas en què una variable té més d'un pare (cosa que no passa en els arbres).

Ara bé, si s'utilitzés la mesura d'informació mútua, es generaria un graf amb massa connexions entre pares i fills. Aquí és on entra en joc el principi MDL. En efecte, utilitzant el criteri de la informació mútua obtenim una distribució de probabilitat per a la xarxa que és la més propera a la distribució de probabilitat de les dades. Aquesta distribució és la que permet d'obtenir la codificació de longitud mínima.

L'**algorisme de Lam i Bacchus** calcula la informació mútua entre totes les variables. Manté "obertes" al mateix temps diverses xarxes. A cadascuna li afegeix l'arc amb màxima informació mútua. Després calcula la longitud de descripció i sempre es queda amb la xarxa que té descripció mínima.

Els mètodes de construcció de xarxes bayesianes són una àrea amb molta activitat. Se n'han dissenyat de nous que permeten de treballar amb valors continus, variables ocultes (que no s'han reflectit en la base de dades) i variables per a les quals manquen valors per a més d'una observació (Ramoni, 1998). També és interessant la línia de treball en creació de mètodes que permeten de desenvolupar algorismes incrementals (Roure, 1999 i Friedman, 1997), i també els que permeten d'introduir alguna forma de coneixement *a priori* (Castelo, 1998). Per a una revisió exhaustiva consulteu Buntine, 1996 i Heckerman, 1996 i, com a més recent; Sangüesa, 1997. Hi ha diverses eines comercials basades en xarxes bayesianes que incorporen algun nivell d'aprenentatge (Hugin) i eines integrades a sistemes comercials de mineria de dades (Castelo, 1997).

Lectura complementària

Trobareu l'algorisme de Lam i Bacchus en l'article següent:

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

3. Classificació amb xarxes bayesianes

La classificació mitjançant el concepte de relacionar la probabilitat *a priori* i *a posteriori* d'una observació, coneguda la seva etiqueta de classe, és un mètode força antic, però molt eficaç.

Lectura complementària

Trobareu el mètode de classificació de xarxes bayesianes coneguda l'etiqueta de classe en l'obra següent:

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

Les suposicions que segueixen els classificadors bayesians són les següents:

a) El conjunt d'etiquetes de classe és exhaustiu i els seus elements mútuament excloents: això significa que no hi ha més classes que les que apareixen en les observacions i que les classes no s'encavalquen. Per exemple, en el cas del gimnàs, els valors de la variable *Sexe* o de la variable *Horari* o de la variable *Entrenador personal* compleixen aquests requisits.

b) Si es coneix l'etiqueta de classe o, millor dit, si es coneix la classe a la qual pertanyen un grup d'observacions, llavors els atributs són condicionalment independents entre si.

La darrera suposició ha fet que els mètodes de classificació bayesians hagin rebut el qualificatiu d'"ingenus" perquè resulta força curiós que dins una classe determinada els valors que prenen els atributs de les observacions que hi pertanyen puguin fer que aquests siguin independents. Els mètodes d'agregació* cerquen precisament trobar els grups d'observacions en què la influència mútua entre les variables és molt alta. Per aquest motiu, durant un cert temps els mètodes bayesians ("ingenus") han estat relegats i només s'han tingut en compte com a "vara de mesura" per a comparar altres mètodes aparentment més sofisticats.

*** Per exemple, Autoclass o COBWEB.**

El mètode per a construir un classificador d'aquesta mena és força senzill. Es tracta de trobar les probabilitats d'aparició dels valors de cada atribut independentment dels altres atributs, donada la classe. Quan arriba una observació nova només cal tenir en compte els seus valors i calcular quin valor de classe és més probable un cop conegudes les probabilitats de tots els atributs (que s'han estimat a partir de les dades en el pas anterior).

Trobar la probabilitat de classe *a posteriori* –és a dir, un cop coneguts els valors de l'observació *a posteriori*– no és més que aplicar el teorema de Bayes. Més senzill, impossible.

Exemple de construcció d'un classificador amb una xarxa bayesiana

Prenem un exemple ben fàcil per a veure com pot funcionar el mètode "ingenu" de classificació amb xarxes bayesianes. Aquí tenim la taula de l'exemple de les lents de contacte. La variable de classe és *Recomanació,* que pren tres valors: lents 'Toves', lents 'Dures' o 'Cap' tipus de lent.

Vegeu l'exemple de les lents de contacte en el subapartat 2.3 del mòdul "Classificació: arbres de decisió" d'aquesta assignatura.

Edat	Diagnòstic	Astigmatisme	Llàgrima	Recomanació
Jove	Miop	No	Reduïda	Cap
Jove	Miop	No	Normal	Toves

Edat	Diagnòstic	Astigmatisme	Llàgrima	Recomanació
Jove	Miop	Sí	Reduïda	Cap
Jove	Miop	Sí	Normal	Dures
Jove	Hipermetrop	No	Reduïda	Cap
Jove	Hipermetrop	No	Normal	Toves
Jove	Hipermetrop	Sí	Reduïda	Cap
Jove	Hipermetrop	Sí	Normal	Dures
Prepresbícia	Miop	No	Reduïda	Cap
Prepresbícia	Miop	No	Normal	Toves
Prepresbícia	Miop	Sí	Reduïda	Cap
Prepresbícia	Miop	Sí	Normal	Dures
Prepresbícia	Hipermetrop	No	Reduïda	Cap
Prepresbícia	Hipermetrop	No	Normal	Toves
Prepresbícia	Hipermetrop	Sí	Reduïda	Cap
Prepresbícia	Hipermetrop	Sí	Normal	Cap
Presbícia	Miop	No	Reduïda	Cap
Presbícia	Miop	No	Normal	Cap
Presbícia	Miop	Sí	Reduïda	Cap
Presbícia	Miop	Sí	Normal	Dures
Presbícia	Hipermetrop	No	Reduïda	Cap
Presbícia	Hipermetrop	No	Normal	Toves
Presbícia	Hipermetrop	Sí	Reduïda	Cap
Presbícia	Hipermetrop	Sí	Normal	Cap

En la taula següent resumim les dades anteriors amb el nom de cada atribut, el valor i les diverses configuracions que poden aparèixer per a cada valor de l'atribut de classificació, que és l'atribut *Recomanació* (lents dures, toves o cap tipus de lent de contacte).

Edat				Diagnòstic				Astigmatisme				Llàgrima			
	Cap	Toves	Dures		Cap	Toves	Dures		Cap	Toves	Dures		Cap	Toves	Dures
Jove	4	2	2	Miop	7	2	3	Sí	8	0	4	Normal	3	5	4
Prepresbícia	5	2	1	Hipermetrop	8	3	1	No	7	5	0	Reduïda	12	0	0
Presbícia	6	1	1												

En total hi ha quatre observacions en què la recomanació va ser 'Dures'; cinc en què va ser 'Toves' i quinze en què la recomanació va ser 'Cap'. Per tant, la taula de freqüències és la que veiem a continuació:

Edat				Diagnòstic				Astigmatisme				Llàgrima			
	Cap	Toves	Dures		Cap	Toves	Dures		Cap	Toves	Dures		Cap	Toves	Dures
Jove	4/15	2/5	2/4	Miop	7/15	2/5	3/4	Sí	8/15	0/0	4/4	Normal	3/15	5/5	4/4
Prepresbícia	5/15	2/5	1/4	Hipemetrop	8/15	3/5	1/4	No	7/15	5/5	0/4	Reduïda	12/15	0/5	0/4
Presbícia	6/15	1/5	1/4												

Suposem que apareix un pacient amb la configuració de valors següent:

{'Jove', 'Hipermetrop', 'No-astigmatisme', 'Llàgrima normal'}

Què li recomanem? En altres paraules, a quina classe pertany? Apliquem el teorema de Bayes:

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

Aquí, la nostra hipòtesi és la classe, i l'evidència, els valors que mostra aquest pacient per a cada atribut. Donada la independència que hi ha entre els atributs, podem expressar-ho d'aquesta manera:

$$P(Toves|E) =$$

$$= \frac{P(Jove|Toves)P(Hipermètrop|Toves)P(No\ astigmatisme|Toves)P(Llàgrima\ normal|Toves)P(Toves)}{P(E)} =$$

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

$$P(Cap|E) =$$

$$= \frac{P(Jove|Cap)P(Hipermètrop|Cap)P(No\ astigmatisme|Cap)P(Llàgrima\ normal|Cap)P(Cap)}{P(E)} =$$

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

Normalitzant, obtenim el següent: $P(Dures|E) = 0$, $P(Toves|E) = 0{,}85$ i $P(Cap|E) = 0{,}13$.

Per tant, un classificador d'aquesta mena recomanaria que el nostre pacient portés lents de contacte toves. En altres paraules, prediria que la classe que li correspon és 'Toves', o en una altra expressió igualment equivalent, el classificaria dins la classe *Recomanació* = 'Toves'.

La sorpresa és que els mètodes bayesians "ingenus" donen uns resultats de classificació molt bons en el sentit que asseguren una precisió més alta.

Lectura complementària

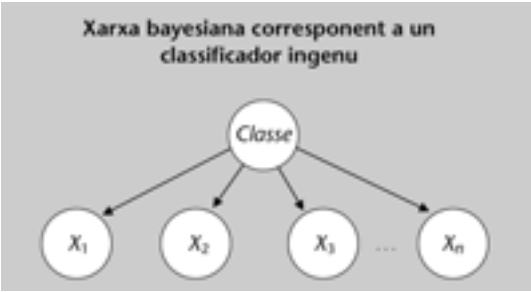
Quant a la precisió dels models bayesians, consulteu l'obra següent:

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

Una segona raó per al revifament de l'interès en aquests mètodes és causada per altres factors. Principalment es pot dir que són mètodes molt senzills. No cal fer una cerca en cap espai descomunal, sinó tan sols portar un càlcul de comptatge prou fàcil de fer.

A continuació veurem com hem d'aplicar les xarxes bayesianes a un problema de classificació. Com hem dit, una de les suposicions bàsiques del mètode bayesià "ingenu" consisteix a suposar que els atributs són independents entre si, atès que coneixem la classe.

Aleshores, la xarxa que prendrem s'assemblarà més a un arbre que no pas a un graf general. En efecte, la variable classe ocupa el node d'aquest arbre i els atributs estan connectats a aquest node com a fills. Gràficament la situació es veu a continuació:



La construcció de la xarxa és força senzilla. En principi, per a cada variable del conjunt inicial de variables, $\{X_1, \ldots, X_n, C\}$, on C representa la variable de classificació, només cal estimar les respectives probabilitats condicionals i construir-ne la xarxa.

En la realitat, les coses són un xic més complicades que haver de tenir en compte la fiabilitat de les estimacions segons la grandària de la base de dades i d'assegurar que, efectivament, es compleixi que cada atribut ho és condicionalment dels altres, donada la classe.

Lectura recomanada

Per a una revisió exhaustiva de mètodes de classificació bayesians en general i els mètodes corresponents que utilitzen xarxes bayesianes, consulteu l'obra següent:

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

Resum

Les xarxes bayesianes són un model que recull la influència entre els atributs d'un domini, i la ponderen mitjançant les distribucions de probabilitat condicionals entre els diversos parells de variables que es poden connectar en una relació pare-fill dins un graf dirigit acíclic.

Les xarxes bayesianes representen relacions estructurals i quantitatives al mateix temps.

El principal interès d'aquests models és que permeten d'efectuar operacions de predicció i explicació amb la mateixa representació.

Els mètodes per a construir xarxes bayesianes es poden basar en les propietats d'independència condicional de les diverses estructures o bé tenir en compte les propietats de les distribucions implícites en cada estructura.

Els mètodes que es basen en les propietats de les distribucions es divideixen al seu torn en mètodes basats en informació, mètodes bayesians i mètodes basats en el principi MDL.

Les xarxes també es poden fer servir per a classificar, operant sota un principi semblant al dels classificadors bayesians "ingenus".

Activitats

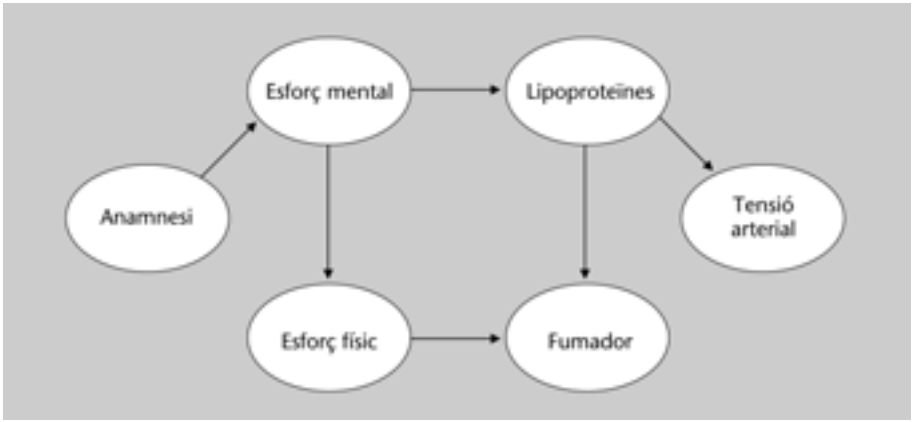
1. Accediu a l'adreça d'Internet que es dóna al marge i compareu les especificacions dels diversos programaris adreçats a construir xarxes bayesianes.

Per a fer l'activitat 1, accediu a l'adreça http://www.kdnuggets.com.

2. Per al problema que us havíeu proposat en l'activitat 1 del mòdul "Extracció de coneixement a partir de dades" d'aquest curs, us serveixen les xarxes bayesianes? Quin mètode creieu que us resultaria més convenient?

Exercicis d'autoavaluació

1. Digueu quines variables es poden considerar independents condicionalment, donada la xarxa bayesiana següent:



2. Donat el conjunt de dades que presentem a continuació, classifiqueu les observacions aplicant el mètode bayesià "ingenu". La variable de classificació és *Entrenador personal*.

Client	Sexe	Renda	Edat	Anys al club	Entrenador personal
1	Dona	6.000.000	40	2	No
4	Home	3.200.000	35	6	No
6	Dona	0	30	3	No
7	Home	4.000.000	28	4	No
11	Home	10.000.000	60	10	Sí
14	Home	1.500.000	67	4	No
221	Dona	0	32	3	Sí
61	Dona	4.000.000	41	6	Sí
18	Dona	0	32	4	No
19	Home	3.000.000	37	3	No
20	Home	2.800.000	32	3	No
21	Dona	0	32	4	No
81	Dona	3.200.000	33	2	No
84	Home	2.000.000	67	4	No
343	Dona	0	20	1	Sí
31	Dona	4.000.000	30	1	No

Bibliografia

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

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

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

Castelo, R. (1997). *Bayesian Networks in Data Surveyor*. Projecte de final de carrera. Facultat d'Informàtica de Barcelona / Universitat Politècnica de Catalunya / Centrum voor Wiskunde en Informatica. Amsterdam.

Castelo, R.; Siebes, A. (1998). "Priors on Networks Structures. Biasing the Search for Bayesian Networks". A: R. Sangüesa; U. Cortés (ed.). *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, pàg. 50-63).

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

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

Duda, R.O.; Hart, P.E. (1973). *Pattern Classification and Scence Analysis*. Nova 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". A: U. Fayyad; G. Piatetsky-Shapiro; P. Smyth; U. Uthurusamy (ed.). *Advances in Knowledge Discovery and Data Mining* (pàg. 273-306). Menlo Park (Califòrnia): 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*. Tesi doctoral. Universidad de Granada: Departamento de Ciencias de la Computación e Inteligencia Artificial.

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, pàg. 269-293).

Neapolitan, R.E. (1990). *Probabilistic Reasoning in Expert Systems*. Nova 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-Tress from Statistical Data". *Uncertainty in Artificial Intelligence* (vol. 3, pàg. 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. A: C. Little (ed.). Lectures Notes in Mathematics 622: Combinatorial Mathematics V (pàg. 28-43). Nova York: Springer-Verlag.

Roure, J.; Sangüesa, R. (1999). *A Survey on Incremental Methods for Bayesian Network Learning*. Informe de recerca LSI-99-42-R. Universitat Politècnica de Catalunya: Departament de Llenguatges i Sistemes Informàtics. 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, pàg. 1-31).