Manejo de Incertidumbre

Eduardo Morales, Enrique Sucar

INAOE

Contenido

Introducción

- Los sistemas inteligentes deben ser capaces de representar y razonar con incertidumbre.
- Existen varias causas de incertidumbre que tienen que ver con la información, el conocimiento y la representación:
- **Información:** incompleta, poco confiable, ruido, ...
- **Conocimiento:** impreciso, contradictorio, ...
- **Información:** no adecuada, falta de poder descriptivo

Ejemplos

- Diagnóstico médico e industrial
- Predicción financiera
- Exploración minera / petrolera
- Interpretación de imágenes
- Reconocimiento de voz
- Monitoreo / control de procesos industriales
- Robótica
- Modelado del usuario
- Bioinformática

Efectos de Incertidumbre

- Se pierden varias propiedades de los sistemas que no tienen incertidumbre, basados en lógicas o reglas, lo cual hace el manejo de incertidumbre más complejo:
- Modularidad
- Monotonicidad

Modularidad

- Un sistema de reglas es modular, ya que para saber la verdad de una regla sólo tiene que considerarla a ésta, sin importar el resto del conocimiento
- Pero si hay incertidumbre ya no puedo considerar la regla por si sola, debo tomar en cuenta otras reglas

Monotónico

- Un sistema es monotónico si al agregar nueva información a su base de datos, no se alteran las conclusiones que seguían de la base de datos original
- Si hay incertidumbre ya no puedo considerar que la certeza en una hipótesis ya no puede cambiar, debo tomar en cuenta otras reglas que involucren a dicha hipótesis

Técnicas No Numéricas

- Lógicas no-monotónicas
- Sistemas de mantenimiento de verdad (TMS, ATMS)
- Teoría de endosos

Técnicas Numéricas

- Empíricas (MYCIN, Prospector)
- Métodos aproximados
- Lógica difusa
- Teoría de Dempster-Shafer
- Probabilísticas - Redes Bayesianas

Axiomas

1. $P(A)$ is a continuous monotonic function in $[0, 1]$.
2. $P(A, B \mid C) = P(A \mid C)P(B \mid A, C)$ (product rule).
3. $P(A \mid B) + P(\neg A \mid B) = 1$ (sum rule).

Conditional Probabilty

- $P(H \mid B)$ conditioned only on the background B is called a *prior* probability;
- Once we incorporate some additional information D we call it a *posterior* probability, $P(H \mid D, B)$
- The conditional probability can be defined as (for simplicity we omit the background): $P(H \mid D) = P(H, D)/P(D)$

Bayes Rule

From the product rule we obtain:

$$P(D, H \mid B) = P(D \mid H, B)P(H \mid B) = P(H \mid D, B)P(D \mid B) \tag{1}$$

From which we obtain:

$$P(H \mid D, B) = \frac{P(H \mid B)P(D \mid H, B)}{P(D \mid B)} \tag{2}$$

This last equation is known as the *Bayes rule*

The term $P(H \mid B)$ is the *prior* and $P(D \mid H, B)$ is the *likelihood*

Independence

- In some cases the probability of H is not influenced by the knowledge of D, so it is said that H and D are *independent*, therefore $P(H, D \mid B) = P(H \mid B)$
- The product rule can be simplified to: $P(A, B \mid C) = P(A \mid C)P(B \mid C)$

Conditional Independence

- If two propositions are independent given only the background information they are *marginally* independent; however if they are independent given some additional evidence, E, then they are *conditionally* independent: $P(H, D \mid B, E) = P(H \mid B, E)$
- Example: H represents the proposition *watering the garden*, D the *weather forecast* and E *raining*

Chain Rule

- The probability of a conjunction of N propositions, that is $P(A_1, A_2, \ldots, A_N \mid B)$, is usually called the *joint* probability
- If we generalize the product rule to N propositions we obtain what is known as the *chain* rule: $P(A_1, A_2, \ldots, A_N \mid B) = P(A_1 \mid A_2, A_3, \ldots, A_N, B)P(A_2 \mid A_3, A_4, \ldots, A_N, B) \cdots P(A_N \mid B)$
- Conditional independence relations between the propositions can be used to simplify this product

Total Probability

- Consider a partition, $B = \{B_1, B_2, ...B_n\}$, on the sample space Ω, such that $\Omega = B_1 \cup B_2 \cup ... \cup B_n$ and $B_i \cap B_j = \emptyset$
- A is equal to the union of its intersections with each event $A = (B_1 \cap A) \cup (B_2 \cap A) \cup ... \cup (B_n \cap A)$
- Then:

$$P(A) = \sum_i P(A \mid B_i)P(B_i) \tag{3}$$

- Given the total probability rule, we obtain Bayes theorem:

$$P(B \mid A) = \frac{P(B)P(A \mid B)}{\sum_i P(A \mid B_i)P(B_i)} \tag{4}$$

Probabilidad en Sistemas Expertos

- Sean $H = h_1, \ldots, h_n$ el conjunto de n posibles hipótesis y $E = e_1, \ldots, e_m$, m posibles evidencias. En general lo que queremos encontrar es la h^* más probable dado E.
- Se tiene que calcular $P(h \mid E)$ para cada subconjunto de hipótesis y seleccionar la de mayor probabilidad utilizando el teorema de Bayes.
- Dos opciones básicas: exhaustivo o independiente

Exhuativo

$$P(h_i \mid e_{j1} \ldots e_{jk}) = \frac{P(e_{j1} \ldots e_{jk} \mid h_i) P(h_i)}{\sum_{j=1}^{n} P(e_{j1} \ldots e_{jk} \mid h_j) P(h_j)}$$

- Se requiere calcular las probabilidades condicionales $P(E \mid h_i)$ para cada combinación de evidencias (en general no se pueden calcular de sus componentes individuales).
- Esto implica que se tienen que conocer un número exponencial de probabilidades!

Independiente

$$P(h_i \mid e_{j1} \ldots e_{jk}) = \frac{P(e_{j1} \mid h_i) \ldots P(e_{jk} \mid h_i) P(h_i)}{\sum_{l=1}^{n} P(e_{j1} \mid h_l) \ldots P(e_{jk} \mid h_l) P(h_l)}$$

- Las evidencias son condicionalmente independientes
- Con esto, sólo se requieren $m \times n$ probabilidades condicionales y $n - 1$ probabilidades a priori.

Análisis

- En la opción 1 el sistema se vuelve demasiado complejo, mientras que la opción 2 puede no ser realista para muchas aplicaciones.
- Una alternativa es buscar un compromiso entre ambos extremos, esto se logra mediante las redes Bayesianas (que veremos más adelante)
- Dados estas limitaciones de los enfoques básicos, en los primeros sistemas expertos se propusieron técnicas *ad hoc* para el manejo de incertidumbre.

Técnicas Empíricas

- Son técnicas empíricas o ad-hoc orientadas a resolver aplicaciones específicas y sin un fuerte fundamento teórico.
- Las más conocidas son las que corresponden a dos de los primeros sistemas expertos:
- PROSPECTOR (exploración minera)
- MYCIN (diagnóstico de enfermedades infecciosas en la sangre)

MYCIN

- Las técnicas desarrolladas en MYCIN y Prospector son similares, ambas consideran sistemas basados en reglas a los que se les adicionan Factores de Certeza o Probabilidades Subjetivas, respectivamente:

Si: se observa cierta evidencia E, Entonces: se concluye cierta hipótesis H con probabilidad (certeza, ...) P

- Veremos brevemente el método de MYCIN.

Factores de Certeza

- MYCIN define un Factor de Certeza que se asocia a cada regla y cada evidencia, y se definen un conjunto de reglas para combinar estos factores.
- Los factores de certeza están en el rango $[-1, +1]$

Reglas de Combinación

- Propagación o reglas en serie:
- $CF(h, e') = CF(h, e) \times max(0, CF(e, e'))$
- Conjunción (AND) y disjunción (OR) de evidencias:
- $CF(e_1 and e_2) = min[CF(e_1), CF(e_2)]$
- $CF(e_1 or e_2) = max[CF(e_1), CF(e_2)]$

Reglas de Combinación

- Co-conclusión o reglas en paralelo:

$$CF(h, e_1' \text{ co } e_2') = \begin{cases} CF(h,e_1') + CF(h,e_2')(1 - CF(h,e_1')) \\ \quad \text{if } CF(h,e_i') > 0, i = 1,2 \\ \dfrac{CF(h,e_1') + CF(h,e_2')}{1 - \min\{|CF(h,e_1')|, |CF(h,e_2')|\}} \\ \quad \text{if } -1 < CF(h,e_1') \times CF(h,e_2') \leq 0 \\ CF(h,e_1') + CF(h,e_2')(1 + CF(h,e_1')) \\ \quad \text{if } CF(h,e_i') < 0, i = 1,2 \end{cases}$$

Ejemplo

R1: IF A and (B or C) Then H cf 0.8
R2: If D and F Then B cf 0.6
R3: If F or G Then H cf 0.4
R4: If A Then D cf 0.75
R5: If I Then G cf 0.3
Se conoce:
CF(A,Ev) = 1,
CF(C,Ev) = 0.5,
CF(F,Ev) = 0.7,
CF(I,Ev) = -0.4

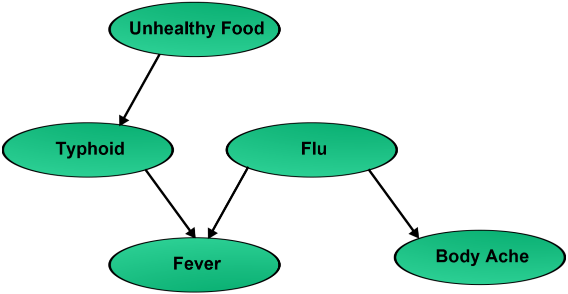
Análisis

- Aunque pretendía apartarse de probabilidad, se ha demostrado [Heckerman 86] que la técnica de MYCIN corresponde a un subconjunto de probabilidad con una serie de suposiciones implícitas:
    - La evidencia es condicionalmente independiente de la hipótesis y su negación.
    - La red de inferencia debe corresponder a un árbol para que los resultados sean coherentes.
    - Las fórmulas para conjunción y disjunción (min y max) sólo son válidas si uno de los términos es subconjunto del otro.
- Estas suposiciones no son válidas en muchas aplicaciones por lo que el método de MYCIN no se puede generalizar.

Redes Bayesianas

- Bayesian networks are directed graphical models that represent the joint distribution of a set of random variables
- In this graphs, the nodes represent random variables and the arcs direct dependencies between variables
- The structure of the graph encodes a set of conditional independence relations between the variables

Example



- *Fever* is independent of *Body ache* given *Flu* (common cause)
- *Fever* is independent of *Unhealthy food* given *Typhoid* (indirect cause)
- *Typhoid* is independent of *Flu* when *Fever* is NOT known (common effect). Knowing Fever makes Typhoid and Flu dependent

Introduction

- In addition to the structure, a Bayesian network considers a set of local parameters, which are the conditional probabilities for each variable given its parents in the graph
- The joint probability of all the variables in the network can be represented based on these local parameters; this usually implies an important saving in the number of required parameters
- Given a Bayesian network we can answer several probabilistic queries. For instance, for the previous example: What is the probability of Fever given Flu? Which is more probable, Typhoid or Flu, given Fever and Unhealthy food?

Bayesian Networks

- A Bayesian network (BN) represents the joint distribution of a set of n (discrete) variables, $X_1, X_2, \ldots, X_n$, as a directed acyclic graph (DAG) and a set of conditional probability tables (CPTs)
- Each node, that corresponds to a variable, has an associated CPT that contains the probability of each state of the variable given its parents in the graph
- The structure of the network implies a set of conditional independence assertions, which give power to this representation

Conditional Independence Assertions

- The conditional independence assertions implied by the structure of a BN should correspond to the conditional independence relations of the joint probability distribution, and vice versa
- If X is conditionally independent of Z given Y:
  - In the probability distribution: $P(X|Y,Z) = P(X|Y)$.
  - In the graph: $I < X \mid Y \mid Z >$.

D-Separation

- Conditional independence assertions can be verified directly from the structure of a BN using a criteria called *D–separation*
- 3 basic BN structures for 3 variables and 2 arcs:
  - Sequential: $X \rightarrow Y \rightarrow Z$.
  - Divergent: $X \leftarrow Y \rightarrow Z$.
  - Convergent: $X \rightarrow Y \leftarrow Z$.
- In the first two cases, X and Z are conditionally independent given Y, however in the third case this is not true

D-Separation

- Given a graph G, a set of variables A is conditionally independent of a set B given a set C, if there is no trajectory in G between A and B such that:
  1. All convergent nodes are or have descendants in C.
  2. All other nodes are outside C.

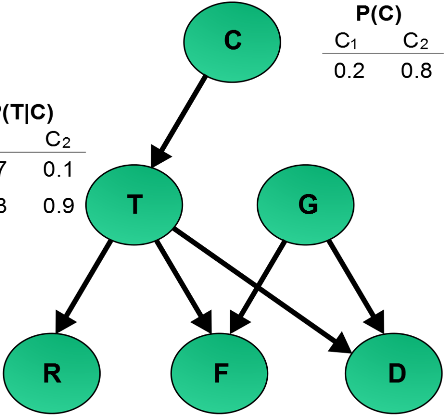
Markov Blanket

- The *Markov Blanket* of a node X, $MB(X)$, is the set of nodes that make it independent of all the other nodes in G, that is $P(X \mid G - X) = P(X \mid MB(X))$
- For a BN, the Markov blanket of X is:
  - the parents of X,
  - the sons of X,
  - and other parents of the sons of X.

CPTs

- In the case of a BN, the parameters are the conditional probabilities of each node given its parents in the graph
- If we consider discrete variables:
  - Root nodes: vector of marginal probabilities.
  - Other nodes: conditional probability table (CPT) of the variable given its parents in the graph.

Example



P(C)

C_1	C_2
0.2	0.8

P(T|C)

	C_1	C_2
T_1	0.7	0.1
T_2	0.3	0.9

P(F|T,G)

	T_1,F_1	T_1,F_2	T_2,F_1	T_2,F_2
F_1	0.8	0.6	0.5	0.1
F_2	0.2	0.4	0.5	0.9

Probabilistic inference

- Probabilistic inference consists in *propagating* the effects of certain evidence in a Bayesian network to estimate its effect on the unknown variables
- There are basically two variants of the inference problem in BNs:
  - *Single query inference*: obtaining the posterior probability of a single variable, H, given a subset of known (instantiated) variables, $\mathbf{E}$, that is, $P(H \mid \mathbf{E})$
  - *Conjunctive query inference*: consists in calculating the posterior probability of a set of variables, $\mathbf{H}$ given the evidence, $\mathbf{E}$, that is, $P(\mathbf{H} \mid \mathbf{E})$

Inference algorithms:

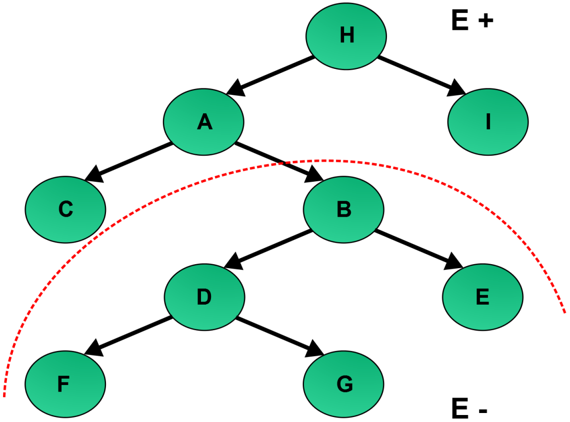
1. Probability propagation (Pearl's algorithm).
2. Variable elimination.
3. Conditioning.
4. Junction tree.
5. Stochastic simulation.

Complexity

- In the worst case the inference problem is *NP-hard* for Bayesian networks
- There are efficient (polynomial) algorithms for certain types of structures (singly connected networks)
- For other structures it depends on the connectivity of the graph.
- In many applications, the graphs are *sparse* and in this case there are inference algorithms which are very efficient

Probability propagation in trees

- Given that the BN has a tree structure, any node divides the network into two independent subtrees



Basic equations

- Given certain evidence, **E** (subset of instantiated variables), the posterior probability for a value *i* of any variable *B*, can be obtained by applying the Bayes rule:

$$P(Bi|\mathbf{E}) = P(Bi)P(\mathbf{E}|Bi)/P(\mathbf{E}) \tag{5}$$

- We can separate the evidence into:

  E-: Evidence in the tree rooted in *B*.

  E+: All other evidence.

- Then:

$$P(Bi|\mathbf{E}) = P(Bi)P(\mathbf{E}-,\mathbf{E}+|Bi)/P(\mathbf{E}) \tag{6}$$

Basic equations

- Given that **E**$+$ and **E**$-$ are independent, by applying the Bayes rule again, we obtain:

$$P(Bi|\mathbf{E}) = \alpha P(Bi|\mathbf{E}+)P(\mathbf{E}-|Bi) \tag{7}$$

  Where α is a normalization constant.

- We define the following terms:

$$\lambda(Bi) = P(\mathbf{E}-|Bi) \tag{8}$$

$$\pi(Bi) = P(Bi|\mathbf{E}+) \tag{9}$$

- Then:

$$P(Bi|\mathbf{E}) = \alpha\pi(Bi)\lambda(Bi) \tag{10}$$

Propagation algorithm

- The computation of the posterior probability of any node B is decomposed into two parts: (i) the evidence coming from the sons of B in the tree (λ), and the evidence coming from the parent of B, (π)
- We can think of each node B in the tree as a simple processor that stores its vectors $\pi(B)$ and $\lambda(B)$, and its conditional probability table, $P(B \mid A)$
- The evidence is propagated via a message passing mechanism, in which each node sends the corresponding messages to its parent and sons in the tree

Messages

- A message sent from node B to its parent A:

$$\lambda_B(Ai) = \sum_j P(B_j \mid A_i)\lambda(B_j) \tag{11}$$

- A message sent from node B to its son S_k:

$$\pi_k(Bi) = \alpha\pi(B_j)\prod_{l \neq k}\lambda_l(B_j) \tag{12}$$

where l refers to each one of the sons of B

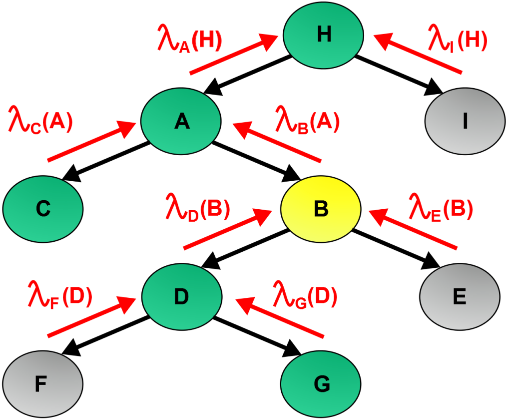
Combination and Propagation

- Each node can receive several λ messages, which are combined via a term by term multiplication for the λ messages received from each son:

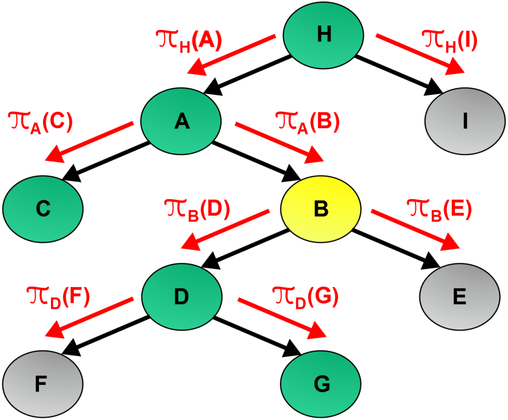
$$\lambda(Ai) = \prod_{j=1}^{m} \lambda_{Sj}(Ai) \tag{13}$$

- The propagation algorithm starts by assigning the evidence to the known variables, and then propagating it through the message passing mechanism until the root of the tree is reached for the λ messages, and the leaves are reached for the π messages

Bottom-up propagation



Top-down propagation



Initial Conditions

Leaf nodes: If not known, $\lambda = [1, 1, ..., 1]$ (a uniform distribution). If known, $\lambda = [0, 0, ..., 1, ..., 0]$ (one for the assigned value and zero for all other values).

Root node: If not known, $\pi = P(A)$ (prior marginal probability vector). If known, $\pi = [0, 0, ..., 1, ..., 0]$ (one for the assigned value and zero for all other values).

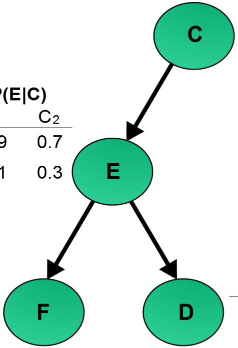
Propagation - example

P(C)

C_1	C_2
0.8	0.2

P(E|C)

	C_1	C_2
E_1	0.9	0.7
E_2	0.1	0.3



P(F|E)

	E_1	E_2
F_1	0.9	0.5
F_2	0.1	0.5

P(D|E)

	E_1	E_2
D_1	0.7	0.4
D_2	0.3	0.6

- Consider that the only evidence is $F = false$ - initial conditions for the leaf nodes are: $\lambda_F = [1, 0]$ and $\lambda_D = [1, 1]$ (no evidence)

Example - λ propagation

- Multiplying the λ vectors by the corresponding CPTs:

$$\lambda_F(E) = [1,0][\begin{array}{c} 0{,}9, 0{,}5 \\ 0{,}1, 0{,}5 \end{array}] = [0{,}9, 0{,}5]$$

$$\lambda_D(E) = [1,1][\begin{array}{c} 0{,}7, 0{,}4 \\ 0{,}3, 0{,}6 \end{array}] = [1,1]$$

- Then, $\lambda(E)$ is obtained by combining the messages from its two sons:

$$\lambda(E) = [0{,}9, 0{,}5] \times [1,1] = [0{,}9, 0{,}5]$$

- Propagation to its parent, C:

$$\lambda_E(C) = [0{,}9, 0{,}5][\begin{array}{c} 0{,}9, 0{,}7 \\ 0{,}1, 0{,}3 \end{array}] = [0{,}86, 0{,}78]$$

Example - π propagation

- Given that C is not instantiated, $\pi(C) = [0{,}8, 0{,}2]$
- Propagate to its son, E, which also corresponds to multiplying the π vector by the corresponding CPT:

$$\pi(E) = [0{,}8, 0{,}2] \begin{bmatrix} 0{,}9, 0{,}7 \\ 0{,}1, 0{,}3 \end{bmatrix} = [0{,}86, 0{,}14]$$

- We now propagate to its son D; however, given that E has another son, F, we also need to consider the λ message from this other son, thus:

$$\pi(D) = [0{,}86, 0{,}14] \times [0{,}9, 0{,}5] \begin{bmatrix} 0{,}7, 0{,}4 \\ 0{,}3, 0{,}6 \end{bmatrix} = [0{,}57, 0{,}27]$$

Example - posterior probabilities

- Given the λ and π vectors for each unknown variable, we just multiply them term by term and then normalize to obtain the posterior probabilities:

$$P(C) = [0{,}86, 0{,}2] \times [0{,}86, 0{,}78] = \alpha[0{,}69, 0{,}16]$$

$$= [0{,}815, 0{,}185]$$

$$P(E) = [0{,}86, 0{,}14] \times [0{,}9, 0{,}5] = \alpha[0{,}77, 0{,}07]$$

$$= [0{,}917, 0{,}083]$$

$$P(D) = [0{,}57, 0{,}27] \times [1, 1] = \alpha[0{,}57, 0{,}27] = [0{,}67, 0{,}33]$$

Introduction

- Learning a Bayesian network includes two aspects:
  - Structure Learning. There are two main types of methods: (i) global methods based on search and score, and (ii) local methods that use conditional independence tests
  - Parameter Learning. When the structure is known, parameter learning consists in estimating the conditional probability tables (CPTs) from data.

Parameter Learning

- If we have *sufficient* and complete data for all the variables, and we assume the topology of the BN is known, parameter learning is straight forward
- The CPT for each variable can be estimated from the data based on the frequency of each value obtaining a *maximum likelihood* (ML) estimator
- For example, to estimate the CPT of B given it has two parents, A, C:

$$P(B_i \mid A_j, C_k) \sim NB_iA_jC_k/NA_jC_k \tag{14}$$

Structure learning

- Structure learning consists in obtaining the topology of the BN from the data
- This is a complex problem because: (i) the number of possible structures is *huge* even with a few variables (it is super-exponential on the number of variables); (ii) a very large database is required to obtain good estimates of the statistical measures
- There are several techniques depending on the type of structure – trees, polytrees, general DAG

General case

- For the general case several methods have been proposed, which can be divided into two main classes:
    1. Global methods: these perform a heuristic search over the space of network structures, starting from some initial structure, and generating a variation of the structure at each step. The *best* structure is selected based on a score that measures how well the model represents the data. Common scores are BIC and MDL
    2. Local methods: these are based on evaluating the (in)dependence relations between subsets of variables given the data, to sequentially obtain the structure of the network. The most well known variant of this approach is the PC algorithm

Conclusiones

- Las redes bayesianas permiten modelar problemas con incertidumbre mediante la reperesentación de las variables de interés y sus relaciones de dependencia / independencia condicional.
- Una vez modelado un problema mediante redes bayesianas, existen algoritmos eficientes para estimar las probabilidades posteriores de alguna variables dadas otras variables (evidencia).
- Es posible aprender la estructura y parámetros de una red bayesiana a partir de datos, conocimiento experto o combinación de ambos.
- Las redes bayesianas forman parte de un conjunto de técnicas conocidas como Modelos Gráficos Probabilistas que incluyen a los clasificadores bayesianos, modelos ocultos de Markov, campos de Markov, diagramas de influencia, etc.

Conclusiones

- Existen, básicamente, otras dos alternativas bien fundamentadas para el manejo de incertidumbre: teoría de Dempster-Shafer y Lógica Difusa.
- La teoría de Dempster-Shafer permite diferenciar entre ignorancia e iguales probabilidades, aunque para problemas grandes se vuelve demasiado compleja
- La lógica difusa tiene la ventaja de que permite representar más fácilmente ciertos conceptos no bien definidos y, en particular las reglas difusas se asemejan a los sistemas basados en reglas. Sin embargo, tiene la dificultad de la falta de una semántica clara.

Referencias

- David Heckerman, Probabilistic Interpretations for MYCIN's Certainty Factors, Proceedings of the First Conference on Uncertainty in Artificial Intelligence (UAI1985)
- Russel and Norvig, Cap. 13 y 14
- Sucar, L. E, *Probabilistic Graphical Models*, Springer 2015
- Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Francisco (1988)
- Neapolitan, R.E.: Learning Bayesian Networks. Prentice Hall, New Jersey (2004)

Tarea

- Repite el ejemplo de MYCIN asumiendo que todas las evidencias tienen un CF igual a 1
- Para la red bayesiana de la lámina 50: (i) determina las relaciones de independencia condicional entre las variables; (ii) Si se sabe que el valor de $C = c_1$, calcula la probabilidad posterior de las otras variables.