Redes Bayesianas e Inteligencia Artificial: Aplicaciones en Educación

Inteligencia Artificial y Educación
Programa de Doctorado de Física e Informática
Universidad de La Laguna

Información Básica

- **Profesor: Fernando Pérez Nava**
  – Teléfono: 922845048
  – e-mail: fdoperez@ull.es
  – Despacho: Edif de la ETSII. Segunda Planta
- **Algunas Referencias Bibliográficas:**
  – Español:
    - **S. Russel y P. Norvig,** *Inteligencia Artificial. Un enfoque moderno,* 2002, Prentice-Hall. Cap 14-17
    - **N. J. Nilsson,** *Inteligencia Artificial. Una nueva síntesis,* 2000, McGraw Hill. Cap 19-20
    - **F.J. Díez,** *Introducción al Razonamiento Aproximado.* Dpto. Inteligencia Artificial, UNED, 2001. http://ia-serv.dia.uned.es/~fjdiez/libros/razaprox.zip
    - **E. Castillo, J.M. Gutiérrez y A.S. Hadi,** *Sistemas Expertos y Modelos de Redes Probabilísticos,* Monografías de la Academia Española de Ingeniería, Madrid, 1998. http://personales.unican.es/gutierjm/papers/BookCGH.pdf
  – Inglés:
    - **F.V. Jensen,** *Bayesian Networks and Influence Diagrams,* Aalborg University, 2001 http://www.cs.auc.dk/~fvj/BSS99/book99.ps
    - **H. Bengtsson,** *Bayesian Networks,* Lund Institute of Technology, 1999. http://www.maths.lth.se/matstat/staff/hb/hbbn99.pdf
  – Cursos:
    - K.B. Laskey Computational Models for Probabilistic Inference (George Manson Univ.) http://ite.gmu.edu/~klaskey/CompProb/
    - N. Friedman Probabilistic Methods in AI (Hebrew University) http://www.cs.huji.ac.il/~pmai/index.html

Contenidos

1. **Introducción a las Redes Bayesianas.**
   – Modelado
   – Inferencia
   – Decisión
   – Aprendizaje
2. **Aplicaciones de Redes Bayesianas en Educación**
   – Sistemas Tutoriales Inteligentes

1

Incertidumbre

- **En muchos dominios de interés para la I.A es necesario trabajar con incertidumbre:**
  – "Falta de conocimiento seguro y claro de algo". (Diccionario RAE)
- **Algunas fuentes de incertidumbre**
  – Ignorancia
    - Puede que en un determinado campo el conocimiento sea incompleto. (Medicina)
    - Aunque se pudiera completar el conocimiento, puede ser necesario tomar decisiones con información incompleta.
    - En otros campos la ignorancia es irreducible
      – Presente en modelos físicos
        » ¿Cuál será el resultado del lanzamiento de una moneda?
      – Presente en la vida real
        » ¿Es la otra persona sincera?
  – Vaguedad e Imprecisión
    - Algunos conceptos son vagos o imprecisos.
      – Las personas altas, guapas, felices etc.

1 Razonamiento con Incertidumbre

- **Objetivo:**
  - Ser capaz de razonar sin tener todo el conocimiento relevante en un campo determinado utilizando lo mejor posible el conocimiento que se tiene.
- **Implementación**
  - Es difícil cumplir estos requerimientos utilizando las técnicas clásicas de la IA (lógica).
  - Deben de introducirse modelos para manejar información vaga, incierta, incompleta y contradictoria.
  - Crucial para un sistema funcione en el "mundo real"

1 Actuar con Incertidumbre

- **El propósito último de un sistema inteligente es actuar de forma óptima utilizando el conocimiento del sistema y un conjunto de percepciones.**
- **Para actuar se necesita decidir que hacer.**
- **¿Cuál es la forma correcta de decidir?**
  - La decisión racional:
    - Cuando se tienen distintas opciones un sistema debe decidirse por aquella acción que le proporcione el mejor resultado.
  - Cuando hay incertidumbre para poder decidir racionalmente se requiere:
    - La importancia de los distintos resultados de una acción
    - La certidumbre de alcanzar esos resultados cuando se realiza la acción.

1 Principales Modelos de Representación de la Incertidumbre

- **Modelos Simbólicos**
  - Lógicas por Defecto
  - Lógicas basadas en Modelos Mínimos
    - La asunción del mundo cerrado
    - Terminación de predicados
    - Circunscripción
- **Modelos Numéricos**
  - Probabilidad
    - Redes Bayesianas
  - Teoría de Dempster-Shaffer
  - Lógica difusa

1 Representación Numérica de la Incertidumbre: Probabilidad

- **La Teoría de la Probabilidad (TProb)**
  - Es un área de las Matemáticas que ha sido aplicada a problemas de razonamiento con incertidumbre
  - Es una teoría elegante, bien entendida y con mucha historia (formalizaciones a partir de mediados del siglo XVII)
  - Asigna valores numéricos (llamados probabilidades) a las proposiciones.
  - Nos dice, dadas las probabilidades de ciertas proposiciones, y algunas relaciones entre ellas como asignar probabilidades a las proposiciones relacionadas
  - Relación con la Lógica Proposicional:
    - En la Lógica Proposicional las proposiciones son ciertas o falsas.
    - Con la Tprob las proposiciones son también ciertas o falsas pero se tiene un grado de creencia en la certeza o falsedad.

1 ¿Qué son las Probabilidades?

- **A pesar de su larga historia los valores numéricos que representan las probabilidad no tiene una interpretación única.**
- **Algunas Interpretaciones:**
  - Frecuentista: Es el valor, cuando el número de pruebas tiende a infinito, de la frecuencia de que ocurra algún evento.
  - Subjetiva: Es un grado de creencia acerca de un evento incierto
- **Aún así:**
  - Existe un consenso sobre el modelo matemático que soporta la Teoría

1 Los Valores Numéricos de la Probabilidad

- **Dada una proposición A, denotaremos por P(A) a la probabilidad de dicha proposición.**
  - A="El resultado del lanzamiento de un dado es 2"
  - A="El paciente tiene sarampión"
  - A="Mañana saldrá el sol"
- **Los valores de la Probabilidad satisfacen tres axiomas:**
  - AX 1: $0 \leq P(A) \leq 1$
  - AX 2: P(Proposición Verdadera)=1
  - AX 3: $P(A \vee B)=P(A)+P(B)$
    - Siempre que A y B sean mutuamente exclusivos, es decir $\neg(A \wedge B)$

1 Consecuencias de los axiomas de Probabilidad

- **Ley de Probabilidad Total**
  - $P(A)=P(A \wedge B)+ P(A \wedge \neg B)$
    - Es una consecuencia del tercer axioma:
      - AX 3: $P(A \vee B)=P(A)+P(B)$
        - » Siempre que A y B sean mutuamente exclusivos, es decir $\neg(A \wedge B)$
  - En general, si B_i i=1...n es un conjunto completo y mutuamente excluyente de proposiciones:

    $P(A)=P(A \wedge B_1)+P(A \wedge B_2)+...+P(A \wedge B_n)=\Sigma\, P(A \wedge B_i)$

    A esta operación se la llama "marginalización"
- **Otras consecuencias:**
  - $P(\neg A)=1-P(A)$
  - P(Proposición Falsa)=0
  - $P(A \vee B)=P(A)+P(B)-P(A \wedge B)$

1 Variables Aleatorias

- **Muchas veces tenemos un evento con un conjunto de resultados:**
  - Completo

    Se conocen todos los posibles resultados
  - Mutuamente excluyente

    No se pueden dar dos resultados distintos simultáneamente.

  Ejemplos
  - Si tiramos una moneda, el resultado es cara o cruz
    - Completo: solo puede salir cara o cruz
    - Excluyente: si sale cara no puede salir cruz
  - La temperatura de un paciente puede estar en un conjunto de intervalos: =<36.4, 36.5-37.4, 37.5-38.4, 38.5-39.4, >=39.5
    - Completo: la temperatura está en alguno de los intervalos
    - Excluyente: la temperatura no puede estar en dos intervalos al mismo tiempo

Variables Aleatorias

- **En lugar de tener una proposición por resultado se introduce el concepto de Variable aleatoria**
- **Se permiten proposiciones de la forma Variable = resultado**
  - Por ejemplo, si M="Resultado de tirar una moneda con valores posibles cara y cruz" se permiten las proposiciones:
    - M=cara y M=Cruz y podemos hablar de
    - P(M=cara) y P(M=cruz) que representan la probabilidad de obtener una cara y una cruz respectivamente
- **Abreviaturas**
  - Se suele escribir P(M=cara) como P(cara), cuando el contexto lo permite
  - Si una variable aleatoria como Sarampión toma únicamente los valores verdadero o falso se suele escribir P(Sarampión =verdadero) como P(sarampión) y P(Sarampión =falso) como P(¬ sarampión)

Distribuciones de Probabilidad

- **Dada una Variable Aleatoria nos gustaría conocer la probabilidad para cada valor que pueda tomar**
- **Esta descripción se llama distribución de probabilidad (Dprob) de la variable aleatoria y consiste en listar los valores de probabilidad para cada valor de la variable**
- **Ejemplo:**
  - Distribución de probabilidad de la variable Llueve

Variable → (cabecera) · Valores → (filas) · Probabilidades → (columna derecha)

Llueve	P(Llueve)
Verdadero	0.1
Falso	0.9

Proposiciones más Complejas

- **Podemos estar interesados en estudiar varias variables en conjunto.**
  - Por ejemplo
    - P(Sarampión=verdadero ∧ Fiebre=verdadero) que es la probabilidad de que el paciente tenga sarampión y fiebre
  - Generalmente lo escribiremos como:
    - P(sarampión∧ fiebre) o P(sarampión, fiebre)
- **Para ello se necesita asignar probabilidades a cada posible combinación de los valores de las variables.**
- **El listado de todos esos valores se llama la distribución conjunta del conjunto de variables**

Ejemplo de distribución conjunta

- **Distribución conjunta de las variables Sabe_Concepto y Resuelve_Ejercicio P(Sabe_Concepto, Resuelve_Ejercicio):**

Sabe_Concepto (SC)	Resuelve_Ejercicio (RE)	P(Sabe_Concepto, Resuelve_Ejercicio)
Verdadero	Verdadero	0.76
Verdadero	Falso	0.04
Falso	Verdadero	0.18
Falso	Falso	0.02

- **También se puede escribir como:**

Sabe_Concepto (SC)	Resuelve_Ejercicio (RE)	P(Sabe_Concepto, Resuelve_Ejercicio)
sabe_concepto	resuelve_ejercicio	0.76
sabe_concepto	¬resuelve_ejercicio	0.04
¬ sabe_concepto	resuelve_ejercicio	0.18
¬ sabe_concepto	¬resuelve_ejercicio	0.02

- **Recuerda a la tabla de la verdad lógica excepto que:**
  - Describe las probabilidad para cada combinación de valores de las variables
  - Generalmente dichos valores no se pueden calcular a partir de sus componentes

La Importancia de la Distribución Conjunta

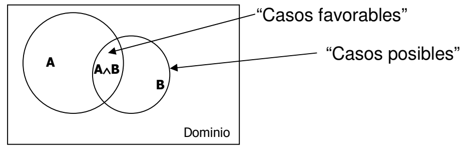
- **La distribución conjunta contiene todo lo que se necesita saber acerca de un conjunto de variables aleatorias.**
- **En particular, la distribución de cada variable individual se puede calcular a partir de la distribución conjunta (y se llama distribución marginal)**
  - Ejemplo: Supongamos las variables aleatorias: Sabe_Concepto y Resuelve_Ejercicio con distribución conjunta P(Sabe_Concepto, Resuelve_Ejercicio)

Sabe_Concepto	Resuelve_Ejercicio	P(Sabe_Concepto, Resuelve_Ejercicio)
sabe_concepto	resuelve_ejercicio	0.76
sabe_concepto	¬resuelve_ejercicio	0.04
¬ sabe_concepto	resuelve_ejercicio	0.18
¬ sabe_concepto	¬resuelve_ejercicio	0.02

- Entonces P(sabe_concepto)= P(sabe_concepto ∧ resuelve_ejercicio)+ P(sabe_concepto ∧ ¬resuelve_ejerciciollueve)=0.76+0.04=0.8.

Probabilidad Condicional

- **Escribiremos P(A|B) para representar la probabilidad de A dado B. Esta probabilidad se llama probabilidad condicional.**
- **Lo podemos interpretar como mi grado de creencia en A cuando todo lo que sé es B.**
  - O de forma alternativa, de los casos en los que se da B, ¿en que proporción se da A?



Probabilidad Condicional Representación gráfica

- **Se define como:**
  - P(A|B)=P(A∧B)/P(B) (Asumiendo P(B)≠0) o equivalentemente
  - P(A∧B)= P(A|B)P(B) (Regla del Producto)

Distribución Condicional

- **Nos permite conocer la probabilidad de que se tomen unos determinados valores por un conjunto de variables aleatorias cuando se saben los valores que han tomado otras.**
  - Ejemplo: P(Resuelve_Ejercicio|Sabe_Concepto)

Resuelve_Ejercicio	P(Resuelve_Ejercicio|Sabe_Concepto)
resuelve_ejercicio	0.95
¬ resuelve_ejercicio	0.05

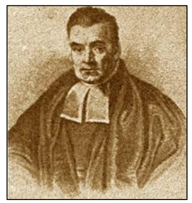
  - Ejemplo: P(Resuelve_Ejercicio| ¬ Sabe_Concepto)

Sabe_Concepto	P(Resuelve_Ejercicio| ¬ Sabe_Concepto)
resuelve_ejercicio	0.10
¬ resuelve_ejercicio	0.90

  - Nótese que Resuelve_Ejercicio|Sabe_Concepto y Resuelve_Ejercicio| ¬ Sabe_Concepto son variables aleatorias

Razonamiento con Probabilidades: La Regla de Bayes

- **Propuesta en 1763 por el Reverendo T. Bayes**
  - P(A|B)= P(B|A) P(A) / P(B)
  - Es una consecuencia de la regla del producto:
    - P(A|B)P(B) = P(A,B) = P(B|A)P(A)



Thomas Bayes

- **De forma intuitiva:**
  - La probabilidad de una hipótesis A dada una evidencia B: P(A|B) es proporcional a probabilidad de la hipótesis P(A) multiplicada por el grado en que la hipótesis predice los datos P(B|A)
- **Aplicabilidad**
  - En muchos problemas dado un conjunto de datos (evidencia) B tenemos que seleccionar la hipótesis A más probable mediante P(A|B)

Regla de Bayes: Forma General

- **Forma general de la Regla de Bayes**
  - Si se tiene un conjunto de proposiciones $\{A_1, A_2, \ldots, A_m\}$ completas y mutuamente excluyente se tiene:

$$P(A_i|B) = \frac{P(B|A_i)\,P(A_i)}{P(B|A_1)\,P(A_1) + \ldots\, P(B|A_n)\,P(A_m)}$$

O lo que es lo mismo, si tiene una variable aleatoria A con valores $a_1, a_2, \ldots, a_m$

$$P(a_i|B) = \frac{P(B|a_i)\,P(a_i)}{P(B|a_1)\,P(a_1) + \ldots\, P(B|a_n)\,P(a_m)}$$

La Regla de Bayes: Ejemplo

- **Intentemos resolver un caso real con probabilidades:**
  - Se pretende determinar si un alumno conoce un concepto en base a la resolución de un ejercicio.
    - En este caso:
      - Hipótesis (SC): Sabe_Concepto (variable aleatoria con dos valores verdadero y falso)
      - Evidencia (RE): Resuelve_Ejercicio (variable aleatoria con dos valores positivo y negativo)
    - Aplicando la Regla de Bayes:

      $P(sc|re) = P(re|sc) \quad P(sc) \quad /(P(re|sc) \quad P(sc) \quad + P(re|\neg sc) \quad P(\neg sc)) = 0.95$

      $P(\neg sc\,|re) = 0.05$
    - Al elegir la hipótesis más probable debemos concluir que si resuelve el ejercicio sabe el concepto

La Regla de Bayes: Ejemplo

- **Continuamos con el ejemplo:**
  - ¿Y si hay varios ejercicios $E_1, \ldots, E_m$?
    - Supondremos que cada ejercicio $RE_1, RE_2, \ldots, RE_m$ es una variable aleatoria que indica si se resuelve con dos valores: verdadero y falso.
  - Entonces si queremos calcular la probabilidad de que el alumno sepa el concepto necesitamos calcular:

    $P(SC|\,E_1, RE_2, \ldots, RE_m) = P(RE_1, \ldots, RE_m|SC)P(SC)/P(RE_1, RE_2, \ldots, RE_m)$
  - Si al alumno se le hace un conjunto de 7 ejercicios:
    - Entonces para almacenar la tabla de probabilidad conjunta $P(RE_1, RE_2, \ldots, RE_m|\ SC)$ se necesitan guardar unos 2^7 números reales (un DVD por alumno).
    - ¿De donde sacamos los números ?
    - ¿Cómo hacemos los cálculos computacionalmente eficientes?

Independencia: ¿Una Solución?

- **Independencia**
  - Decimos que dos proposiciones A_1 y A_2 son independientes si el conocimiento de una no cambia la probabilidad de la otra
    - Por ejemplo si
      - A_1="Es rubio" , A_2="Tiene la piel clara" ,A_3="Lloverá mañana"
      - A_1 y A_3 son independientes A_1 y A_2 no.
  - Formalmente A_1, A_2 son independientes si $P(A_1|A_2) = P(A_1)$

    o de forma equivalente: $P(A_2|A_1) = P(A_2)$

    o utilizando la regla del producto $P(A_1 \wedge A_2) = P(A_1)\,P(A_2)$
  - Entonces $P(A_1 \wedge A_2 \wedge \ldots \wedge A_n) = P(A_1)\,P(A_2) \ldots P(A_n)$

    Para especificar la distribución conjunta de n variables se necesitan $o(n)$ números en lugar de $o(2^n)$
  - Dos variables aleatorias son independientes si el conocimiento del valor que toma una no cambia la probabilidad de los valores de la otra: $P(A_1{=}c|A_2{=}d) = P(A_1{=}c)$

1 Independencia Condicional

- **Pero...**
  - La condición de independencia es muy restrictiva.
  - Por ejemplo, los resultados de los ejercicios en la enseñanza no suelen ser independientes.
- **Independencia condicional**
  - Se dice que dos proposiciones A_1,A_2 son independientes dada una tercera B si cuando B está presente el conocimiento de una no influye en la probabilidad de la otra: $P(A_1|A_2,B)=P(A_1|B)$
    o de forma equivalente: $P(A_2|A_1,B)=P(A_2|B)$
    o de forma equivalente: $P(A_1 \wedge A_2 |B)= P(A_1|B) P(A_2|B)$
    - Ejemplo:
      - A_1="Tengo congestión nasal" A_2="Tengo fiebre" A_3="Tengo gripe"
      - A_1 y A_2 son dependientes pero son independientes si se conoce A_3.
  - Ahora se tiene: $P(A_1 \wedge A_2 \wedge ... \wedge A_n |B)=P(A_1|B) P(A_2|B) ... P(A_n|B)$
    - Tenemos o(n) números en lugar de o(2^n)

1 Independencia Condicional

- **Finalizamos el ejemplo:**
  - ¿Y si hay varios ejercicios $E_1,E_2,...,E_m$?
  - Como vimos, para calcular la probabilidad de que el alumno sepa el concepto necesitamos calcular:

    $P(SC| E_1,RE_2,...,RE_m)=P(RE_1,,...,RE_m|SC)P(SC)/P(RE_1,RE_2,...,RE_m)$

    Si los resultado de los ejercicios $E_1,E_2,...,E_m$ son independientes dado el concepto (aproximación que suele dar buenos resultados):
    $P(RE_1,,...,RE_m|SC)=P(RE_1|SC) P(RE_2|SC) ... P(RE_m| SC)$
  - El problema a resolver ya es abordable:

1 Representación de la Independencia:Redes Bayesianas

- **La clave hacer factible la inferencia con probabilidades es la introducción explícita de la independencia entre variables**
- **El modelo más extendido de representación de independencias lo constituye las Redes Bayesianas.**
- **En este modelo se representa de forma explícita la dependencia entre variables mediante un grafo**
- **Los nodos del grafo se corresponden con variables y las dependencias se representan mediante arcos entre ellas**

1 Redes Bayesianas:Introducción

- **Eliminan algunos de los problemas asociados al razonamiento probabilístico**
- **Desarrolladas a finales de los 70 (Pearl), se convirtieron durante los 90 en un esquema general de representación de la incertidumbre**
- **Una Red Bayesiana (RB) proporciona una forma compacta y modular de representar la distribución conjunta de varias variables aleatorias**
- **Una RB consta de:**
  - Una parte cualitativa que describe las relaciones entre las distintas variables
  - Una parte *cuantitativa* que describe la fuerza de dichas relaciones mediante probabilidades condicionadas

1 Redes Bayesianas: Inferencia, Decisión y Aprendizaje

- **En una RB, la información proporcionada por una o más variables que se observan (evidencia) se propaga por la red y actualiza nuestra creencia acerca de las variables no observadas. A este proceso se le llama inferencia.**
- **Es posible aprender las probabilidades condicionales que describen las relaciones entre las variables a partir de los datos. Incluso es posible aprender la estructura completa de la red a partir de datos completos o con algunos de sus valores desconocidos.**
- **Las RB pueden utilizarse para tomar decisiones óptimas introduciendo posibles acciones y la utilidad de sus resultados**

1 Redes Bayesianas en la Prensa

- **Cnet.com**

NEWS.COM SPECIAL REPORT

Old school

18th-century theory is new force in computing

By Michael Kanellos
Staff Writer, CNET News.com
February 18, 2003, 4:00 AM PT

Thomas Bayes, one of the leading mathematical lights in computing today, differs from most of his colleagues: He has argued that the existence of God can be derived from equations. His most important paper was published by someone else. And he's been dead for 241 years.

Yet the 18th-century clergyman's theories on probability have become a major part of the mathematical foundations of application development.

Search giant Google and Autonomy, a company that sells information retrieval tools, both employ Bayesian principles to provide likely (but technically never exact) results to data searches. Researchers are also using Bayesian models to determine correlations between specific symptoms and diseases, create personal robots, and develop artificially intelligent devices that "think" by doing what data and experience tell them to do.

Bayes' theorem

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

Despite the esoteric symbols, the idea--roughly speaking--is simple: The likelihood that something will happen can be plausibly estimated by how often it occurred in the past. Researchers are applying the idea to everything from gene studies to filtering e-mail.

A detailed mathematical rundown can be found on the University of Minnesota's Web site. And a Bayes Rule Applet on Gametheory.net lets you answer questions such as "How worried should you be if you test positive for some disease?"

One of the more vocal Bayesian advocates is Microsoft. The company is employing ideas based on probability--or "probabilistic" principles--in its Notification Platform. The technology will be embedded in future Microsoft software and is intended to let computers and cell phones automatically filter messages, schedule meetings without their owners' help and derive strategies for getting in touch with other people.

If successful, the technology will give rise to "context servers"--electronic butlers that will interpret people's daily habits and organize their lives under constantly shifting circumstances.

"Bayesian research is used to make the best gambles on where I should flow with computation and bandwidth," said Eric Horvitz, senior researcher and group manager of the Adaptive Systems & Interaction Group at Microsoft Research. "I personally believe that probability is at the foundation of any intelligence in an uncertain world where you can't know everything."

Probability theorist Thomas Bayes. If you've used Google, you've gone Bayesian.

Developers are using Bayes' centuries-old ideas to take on the Information Age.

They're building programs designed to automatically manage the deluge of data that gets thrown at people each day via e-mail, cell phones, instant messaging programs and the like.

Such a program would collect data over time and create a model of a person's past behavior to determine how best to deal with a newly arrived message.

A simplified example:

1) E-mail arrives
From: The boss
Subject: Read me
Sent: Thu 3:10 PM

Program accesses calendar software
"Thursday, 2 PM: Meeting"

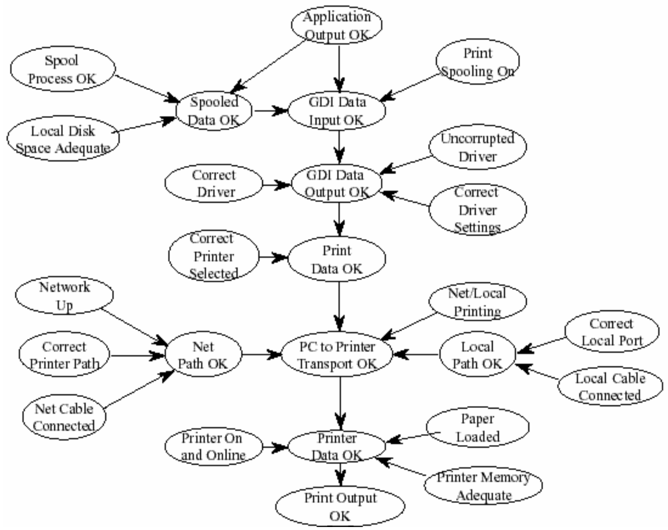
Program considers

http://news.com.com/2009-1001-984695.html

1 Redes Bayesianas: Utilización

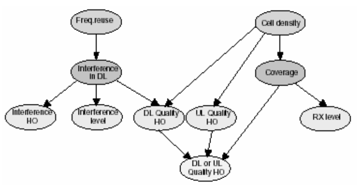
- **Algunas aplicaciones de RB en empresas**
  - Microsoft
    - Answer Wizard (Office)
    - Diagnóstico de problemas de usuario (Aladdin)
    - Home Health en la red de Microsoft (MSN)
  - Intel
    - Diagnóstico de fallos de procesadores
  - HP
    - Diagnóstico de problemas de impresora
  - Nokia
    - Diagnóstico de redes celulares
  - Nasa
    - Sistema de ayuda a la decisión en misiones espaciales

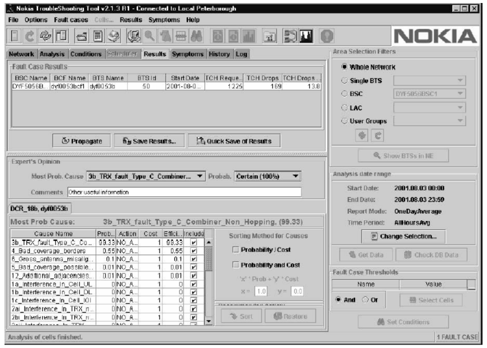
1 Red Bayesiana:Ejemplo



Diagnóstico de Problemas de Impresión (Heckerman)

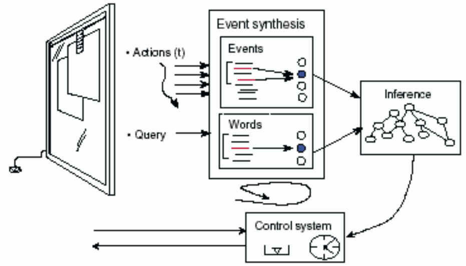
1 Red Bayesiana:Ejemplo



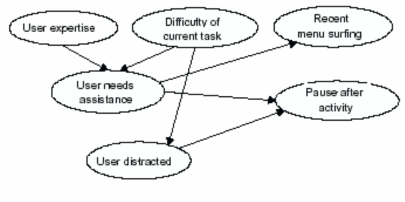


Diagnóstico de Problemas en redes celulares para Nokia (Barco y otros)

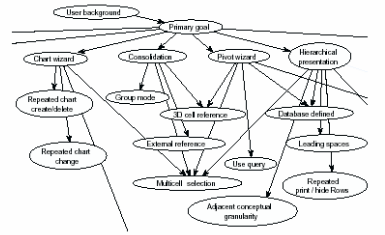
1 Red Bayesiana:Ejemplo



Visión de alto nivel del sistema de RB para Excel



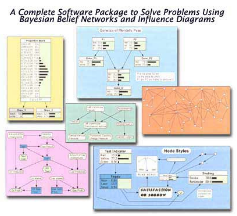
Vista parcial de la red para inferir si el usuario tiene dificultades



Vista parcial de la red para inferir si el usuario tiene dificultades con Excel (Heckerman)

1 Redes Bayesianas: Algunas Herramientas

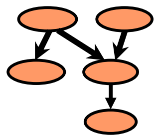
- **Norsys**
  - Programa: Netica
  - Descarga de: http://www.norsys.com/netica.html



- **Microsoft**
  - MSBNx
  - Descarga de: http://research.microsoft.com/adapt/MSBNx/

MSBNx
Bayesian Network Editor and Tool Kit

1 ¿Qué es un Red Bayesiana (RB)?

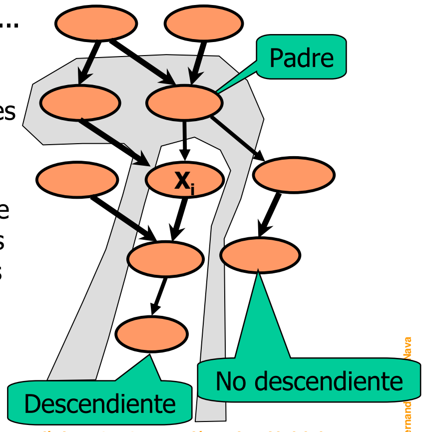


- **Una RB es un grafo dirigido en el que cada nodo contiene información probabilística.**
- **Para determinar una RB hace falta:**
  - Un conjunto de variables aleatorias que forman los nodos de la red. Las variables pueden ser discretas o continuas
  - Un conjunto de enlaces dirigidos (arcos) que conectan parejas de nodos. Si hay un arco del nodo X al Y se dice que X es el padre de Y.
    - El significado intuitivo de un arco desde el nodo X al Y es que X tiene una influencia directa en Y
  - Cada nodo X_i tiene una distribución de probabilidad condicional: $P(X_i|Padres(X_i))$ que mide el efecto de los padres de ese nodo.
  - El grafo no tiene ciclos dirigidos (y por tanto es un grafo dirigido acíclico o DAG)

Red Bayesiana: Significado

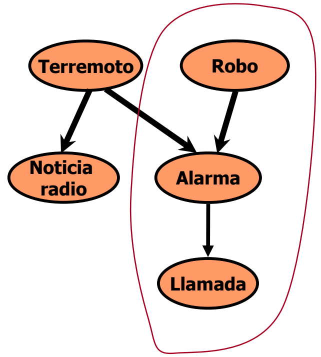
- **Los arcos en una RB proporciona una forma de codificar relaciones de independencia**
- **Estas relaciones se pueden especificar como:**
  - Dada una RB con nodos X_1, X_2,... X_n. Si Padres(X_i) son los padres de X_i y NoDescendientes(X_i) los nodos que no son descendientes de X_i.
  - Entonces para cada variable X_i se tiene que X_i es independiente de sus No Descendientes dados sus Padres. Esto lo expresamos como

    Ind(X_i; NoDescendientes(X_i) | Pa(X_i))



Ejemplos de Independencias

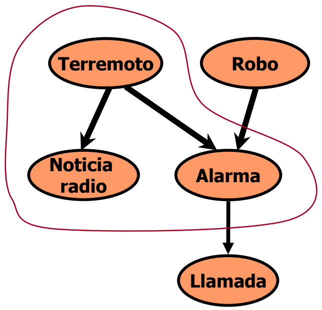
- **Para la RB del ejemplo:**
  - R y L son dependientes:
    - Si hay un robo es más probable suene la alarma, lo que hace más probable que que reciba una llamada.
    - Si recibo una llamada se incrementa la probabilidad de que haya sonado la alarma y por tanto de que me hayan robado.
  - R y L son independientes si se conoce A
    - Si hay un robo ya no es más probable que suene la alarma (ya se sabe si suena o no)
    - Si recibo una llamada ya no se incrementa la probabilidad de que suene la alarma (ya se sabe si suena o no)



L es independiente de sus no-descendientes T,R,N, dados sus padres A

Ejemplos de Independencias

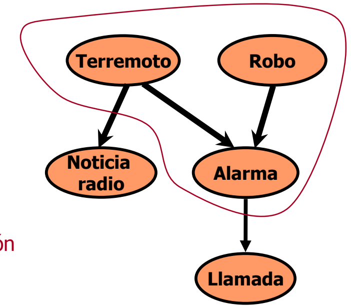
- **Para la RB del ejemplo:**
  - N y A son dependientes:
    - Si oigo en la radio que ha habido un terremoto es más probable que éste haya ocurrido, lo que hace más probable que que suene la alarma.
    - Si suena la alarma se incrementa la probabilidad de que haya ocurrido un terremoto y por tanto de que oiga la noticia en la radio.
  - N y A son independientes si se conoce T
    - Si oigo en la radio que ha habido un terremoto ya no es más probable que éste haya ocurrido. (ya se sabe si ha ocurrido o no).
    - Si suena la alarma ya no se incrementa la probabilidad de que haya ocurrido un terremoto (ya se sabe si ocurrió)



N es independiente de sus no-descendientes R,A,L dados sus padres T

Ejemplos de Independencias

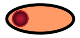
- **Para la RB del ejemplo:**
  - T y R son dependientes si se conoce A
    - Si suena la alarma y ocurre una de las causas (terremoto) me creo menos la otra (alarma)
    - Si suena la alarma y ocurre una de las causas (alarma) me creo menos la otra (terremoto)
    - A este efecto se le llama "eliminación de explicaciones"
  - T y R son independientes:
    - Si desconozco si suena la alarma y ocurre una de las causas (terremoto) no hay razón para creer menos la otra (alarma)
    - Si desconozco si suena la alarma y ocurre una de las causas (alarma) no hay razón para creer menos la otra (terremoto)

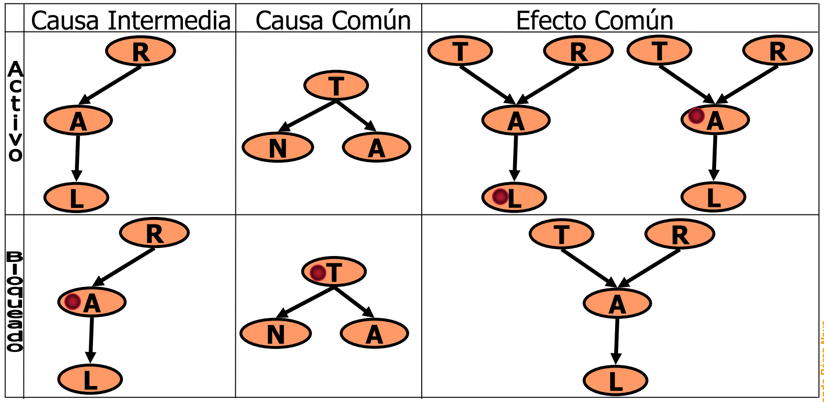


T es independiente de sus no-descendientes R dados sus padres (ninguno).

Transmisión de información en la red

- **Un camino del grafo puede estar:**
  - Activo si pasa información por el.
  - Bloqueado: si no pasa

Nodo con evidencia (observado) 

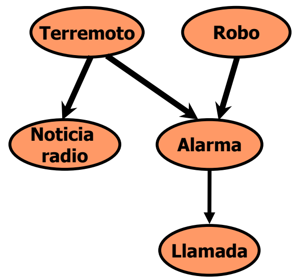


Teorema de Factorización

- **Dada la codificación de independencias de una RB**

$$P(X_1,\ldots,X_n) = \prod_i P(X_i \mid Pa(X_i))$$

- **Ejemplo**
  - Teorema de Factorización:
    P(L,A,N,T,R) =
    P(R) P(T) P(N|T) P(A|R,T) P(L|A)



Factorización:Consecuencias

- **Se puede describir P utilizando probabilidades condicionales "locales"**
- **Si G es un grafo disperso, es decir el número de padres de cada variable está acotado: $|Pa(X_i)| \leq k$ con k un número "pequeño" se obtiene:**
  - Representación compacta

    El número de parámetros para describir la función de distribución conjunta es lineal en el número *n* de variables aleatorias o(*n*)

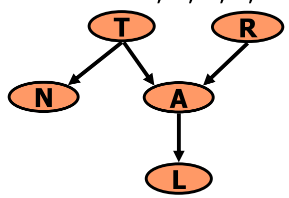
    Nótese que el número de parámetros requerido en general es de orden o(2^n)
  - Representación modular

    Añadir una nueva variable no obliga a actualizar todos los parámetros de la representación

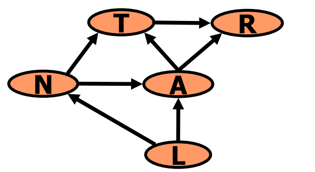
Construcción de RB

- **Un algoritmo de construcción de RB**
  - Elegir un grupo de variables $X_1,\ldots,X_n$ que describen un dominio
  - Fijar un orden en $X_1,\ldots,X_n$ (por ejemplo de las causas a los efectos)
  - Mientras haya variables
    - Elegir la siguiente variable X_i y añadir un nodo para ella
      Selecionar Padres(X_i) como el conjunto mínimo de $\{X_1,\ldots,X_{i-1}\}$, de forma que Ind(X_i ; $\{X_1,\ldots,X_{i-1}\}$ - Pa_i | Pa_i)
- **La red resultante depende del orden:**

Orden:T, R, A, L, N



Orden: L, N, A, T, R



1

La elección de la ordenación y la causalidad

- **La elección de la ordenación puede tener un impacto drástico en la complejidad de la Red Bayesiana.**

- **Heurística para construir la RB:**
  - Construir la RB utilizando la ordenación causal entre las variables

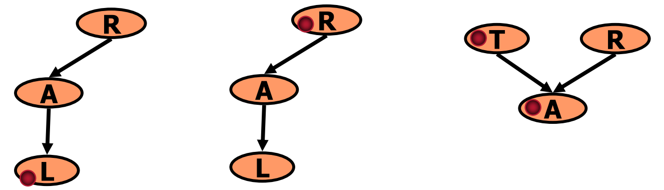
- **Justificación**
  - Generalmente se puede asumir que los grafos generados a partir de relaciones causales cumplen las condiciones de independencia

1

Inferencia en Redes Bayesianas

- **Inferencia:**
  - Se pretende hallar la distribución de probabilidad de determinadas variables de interés dados los valores de otras variables que se observan.

- **Principales tipos de Inferencia**
  - Diagnóstico — Predicción — Intercausal



P(robo|llama) — P(llama|robo) — P(robo|alarma,¬terremoto)

1

Inferencia en Redes Bayesianas

- **De manera formal**
  - Supondremos que:
    - La red bayesiana está formada por las variables: $\{ X_1, \ldots, X_n \}$
    - Las variables de interés son $\mathbf{X}_I = \{ X_1, \ldots, X_i \}$
    - Las variables observadas (con evidencia) son: $\mathbf{X}_O = \{ X_{i+1}, \ldots, X_j \}$
      - Los valores que toman dichas variables (evidencia) son $\mathbf{e} = \{ e_{i+1}, \ldots, e_j \}$
    - El resto de variables son $\mathbf{X}_R = \{ X_{j+1}, \ldots, X_n \}$
  - El problema a resolver es:
    Calcular:

$$P(\mathbf{X}_I \mid \mathbf{X}_O = \mathbf{e}) = \frac{P(\mathbf{X}_I, \mathbf{X}_O = \mathbf{e})}{P(\mathbf{X}_O = \mathbf{e})} =$$

$$\frac{P(X_1, X_2, \ldots, X_i, X_{i+1} = e_{i+1}, X_{i+2} = e_{i+2}, \ldots, X_j = e_j)}{P(X_{i+1} = e_{i+1}, X_{i+2} = e_{i+2}, \ldots, X_j = e_j)}$$

1

Algoritmos de inferencia

- **Los diversos algoritmos propuestos se pueden dividir en:**
  - Algoritmos exactos
    - Calcular de forma exacta la inferencia solicitada.
    - La complejidad de resolver de forma exacta el problema general de inferencia en Redes Bayesianas es NP-duro.
    - Tipos
      - Para redes específicas:
        - » Árboles (Pearl), (Complejidad lineal)
        - » Poliárboles (Kim, Pearl), (Complejidad lineal)
      - Para redes generales
        - » Eliminación de Variables, Árbol de uniones (Lauritzen y Spiegelhalter)
  - Algoritmos aproximados
    - Se basan calcular de forma aproximada la inferencia solicitada simulando la distribución de la red bayesiana.
    - Aproximar una distribución con una tolerancia dada es también NP-duro.
    - Algunos algoritmos
      - Muestreo lógico (Henrion)
      - Ponderación de la verosimilitud (Fung y Chang)

1 Algoritmo de Inferencia por Eliminación de Variables

- **El problema es calcular:**

$$P(\mathbf{X}_I \mid \mathbf{X}_O = \mathbf{e}) = \frac{P(\mathbf{X}_I, \mathbf{X}_O = \mathbf{e})}{P(\mathbf{X}_O = \mathbf{e})}$$

- **El numerador es igual a:**

$$P(\mathbf{X}_I, \mathbf{X}_O = \mathbf{e}) = \sum_{\mathbf{X}_R} P(\mathbf{X}_I, \mathbf{X}_O = \mathbf{e}, \mathbf{X}_R) =$$

$$\sum_{X_{j+1}, X_{j+2}, \ldots, X_n} P(X_1, X_2, \ldots, X_i, X_{i+1} = e_{i+1}, X_{i+2} = e_{i+2}, \ldots, X_j = e_j, X_{j+1}, X_{j+2}, \ldots, X_n)$$

- **El denominador es igual a:**

$$P(\mathbf{X}_O = \mathbf{e}) = \sum_{\mathbf{X}_I} P(\mathbf{X}_I, \mathbf{X}_O = \mathbf{e}) =$$

$$\sum_{X_1, X_2, \ldots, X_i} P(X_1, X_2, \ldots, X_i, X_{i+1} = e_{i+1}, X_{i+2} = e_{i+2}, \ldots, X_j = e_j)$$

  – Por tanto, si se tiene el numerador el denominador se puede calcular a partir de éste.

Fernando Pérez Nava

1 Eliminación de Variables: Cálculo del numerador

- **El problema ahora es calcular:**

$$P(\mathbf{X}_I, \mathbf{X}_O = \mathbf{e}) = \sum_{\mathbf{X}_R} P(\mathbf{X}_I, \mathbf{X}_O = \mathbf{e}, \mathbf{X}_R) =$$

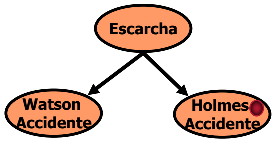
$$\sum_{X_{j+1}, X_{j+2}, \ldots, X_n} P(X_1, X_2, \ldots, X_i, X_{i+1} = e_{i+1}, X_{i+2} = e_{i+2}, \ldots, X_j = e_j, X_{j+1}, X_{j+2}, \ldots, X_n)$$

- **Idea General:**
  – Para realizar de forma eficiente la suma anterior:
    - Paso 1
      Usar el teorema de factorización para factorizar la distribución conjunta:
      $$P(X_1, X_2, \ldots, X_i, X_{i+1}, X_{i+2}, \ldots, X_j, X_{j+1}, X_{j+2}, \ldots, X_n) = \prod_i P(X_i, pa(X_i))$$
    - Paso 2
      Fijar las variables observadas a sus valores de evidencia
    - Paso 3
      Eliminar de forma iterativa del sumatorio las variables que no son de interés ni de evidencia

Fernando Pérez Nava

1 Eliminación de variables: Ejemplo con evidencia

- **La red de la Escarcha**



P(E)

E	
e	0.7
¬e	0.3

P(W|E)

E \ W	w	¬w
e	0.8	0.2
¬e	0.1	0.9

P(H|E)

E \ H	h	¬h
e	0.8	0.2
¬e	0.1	0.9

- **Calculemos P(w|h)**
  – Factorización:

$\boxed{P(E)P(W|E)P(H|E)}$

Sustitución de la evidencia

$\boxed{P(E)P(W|E)P(h|E)}$

Eliminamos E

$$g_E(W) = \sum_E P(E)P(W \mid E)P(h \mid E)$$

$g_E(W)$

w	0.7 x 0.8 x 0.8+0.3 x 0.1 x 0.1 = 0.451
¬w	0.7 x 0.2 x 0.8+0.3 x 0.9 x 0.1 = 0.139

$\boxed{P(W,h)}$

Normalizamos

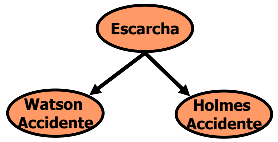
w	≈ 0.77
¬w	≈ 0.23

$\boxed{P(W|h)}$

Fernando Pérez Nava

1 Eliminación de variables: Ejemplo sin evidencia

- **La red de la Escarcha**



P(E)

E	
e	0.7
¬e	0.3

P(W|E)

E \ W	w	¬w
e	0.8	0.2
¬e	0.1	0.9

P(H|E)

E \ H	h	¬h
e	0.8	0.2
¬e	0.1	0.9

- **Calculemos P(W)**
  – Factorización:

$\boxed{P(E)P(W|E)P(H|E)}$

Eliminamos H

$$g_H(E) = \sum_H P(H \mid E)$$

$g_H(E)$

e	0.8 + 0.2 = 1.0
¬e	0.1 + 0.9 = 1.0

$\boxed{P(E)P(W|E)g_H(E)}$

Eliminamos E

$$g_E(W) = \sum_E g_H(E)P(E)P(W \mid E)$$

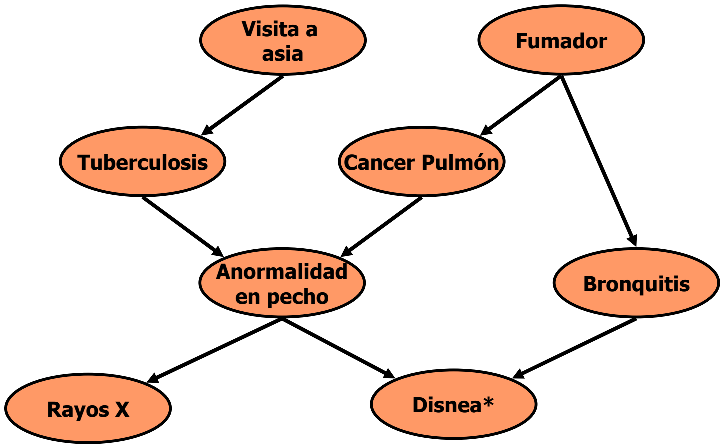
w	1.0 x 0.7 x 0.8+1.0 x 0.3 x 0.1 = 0.59
¬w	1.0 x 0.7 x 0.2+1.0 x 0.3 x 0.9 = 0.41

$\boxed{P(W)}$

Fernando Pérez Nava

1 Un Ejemplo más Complejo

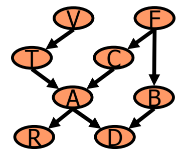
- **La Red Asia**



*Disnea=Dificultad para respirar

Fernando Pérez Nava

1 Eliminación de variables: Ejemplo con evidencia



- **Ejemplo: calcular P(C|v,f,d)**
  - Como siempre calculamos P(C,v,f,d) y normalizamos
  - Escribimos la factorización:

    $P(V)P(F)P(T \mid V)P(C \mid F)P(B \mid F)P(A \mid T,C)P(R \mid A)P(D \mid A,B)$

    y sustituimos la evidencia V=cierto,F=cierto,D=cierto

    $P(v)P(f)P(T \mid v)P(C \mid f)P(B \mid f)P(A \mid T,C)P(R \mid A)P(d \mid A,B)$

    tenemos por tanto que eliminar:R,T,A,B (puesto que V,F,D están fijos)
  - Proceso de eliminación:

$$\boxed{P(v)P(f)P(T \mid v)P(C \mid f)P(B \mid f)P(A \mid T,C)P(R \mid A)P(d \mid A,B)}$$

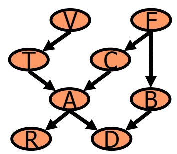
Eliminamos R

$$\boxed{h_R(A,C) = \sum_R P(R \mid A)}$$

$$\boxed{P(v)P(f)h_R(A)P(T \mid v)P(C \mid f)P(B \mid f)P(A \mid T,C)P(d \mid A,B)}$$

Fernando Pérez Nava

1 Eliminación de variables: Ejemplo con evidencia



- **Proceso de eliminación: continuación**
  - Eliminando: T,A,B

$$\boxed{P(v)P(f)P(T \mid v)P(C \mid f)P(B \mid f)P(A \mid T,C)P(R \mid A)P(d \mid A,B)}$$

$$\boxed{P(v)P(f)h_R(A)P(T \mid v)P(C \mid f)P(B \mid f)P(A \mid T,C)P(d \mid A,B)}$$

Eliminamos T

$$\boxed{h_T(A,C) = \sum_T P(T \mid v)P(A \mid T,C)}$$

$$\boxed{P(v)P(f)h_R(A)h_T(A,C)P(C \mid f)P(B \mid f)P(d \mid A,B)}$$

Eliminamos A

$$\boxed{h_A(C,B) = \sum_A h_r(A)h_t(A,C)P(d \mid A,B)}$$

$$\boxed{P(v)P(f)h_A(C,B)P(C \mid f)P(B \mid f)}$$

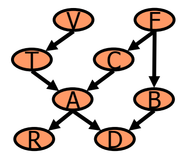
Eliminamos B

$$\boxed{h_b(C) = \sum_B h_B(C,B)P(B \mid f)}$$

$$\boxed{P(v)P(f)P(C \mid f)h_B(C)} \xrightarrow{\text{Normalización}} \boxed{P(C \mid v,f,d)}$$

Fernando Pérez Nava

1 Eliminación de variables: Ejemplo sin evidencia



- **Ejemplo: Calcular P(D)**
  - Eliminaremos: V,F,R,T,C,A,B
  - Aplicando el teorema de factorización:

$$\boxed{P(V)P(F)P(T \mid V)P(C \mid F)P(B \mid F)P(A \mid T,C)P(R \mid A)P(D \mid A,B)}$$

Eliminamos V

$$\boxed{h_V(T) = \sum_V P(T \mid V)P(V)}$$

$$\boxed{h_V(T)P(F)P(C \mid F)P(B \mid F)P(A \mid T,C)P(R \mid A)P(D \mid A,B)}$$

Eliminamos F

$$\boxed{h_F(C,B) = \sum_F P(C \mid F)P(B \mid F)P(F)}$$

$$\boxed{h_V(T)h_F(C,B)P(A \mid T,C)P(R \mid A)P(D \mid A,B)}$$

Eliminamos R

$$\boxed{h_R(A) = \sum_A P(R \mid A)}$$

$$\boxed{h_V(T)h_F(C,B)h_R(A)P(A \mid T,C)P(D \mid A,B)}$$

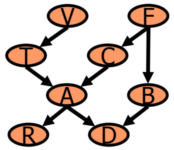
Eliminamos T

$$\boxed{h_T(A,C) = \sum_T h_v(T)P(A \mid T,C)}$$

$$\boxed{h_T(A,C)h_F(C,B)h_R(A)P(D \mid A,B)}$$

Fernando Pérez Nava

Eliminación de variables: Ejemplo con evidencia



- **Ejemplo: Calcular P(D)**
  - Eliminaremos V,F,R,T,C,A,B
  - Proceso de eliminación (continuación)

$\boxed{P(V)P(F)P(T \mid V)P(C \mid F)P(B \mid F)P(A \mid T,C)P(R \mid A)P(D \mid A,B)}$

$\boxed{h_V(T)P(F)P(C \mid F)P(B \mid F)P(A \mid T,C)P(R \mid A)P(D \mid A,B)}$

$\boxed{h_V(T)h_F(C,B)P(A \mid T,C)P(R \mid A)P(D \mid A,B)}$

$\boxed{h_V(T)h_F(C,B)h_R(A)P(A \mid T,C)P(D \mid A,B)}$

$\boxed{h_T(A,C)h_F(C,B)h_R(A)P(D \mid A,B)}$

Eliminamos C

$$\boxed{h_C(A,B) = \sum_C h_T(A,C)h_F(C,B)}$$

$\boxed{h_C(A,B)h_R(A)P(D \mid A,B)}$

Eliminamos A

$$\boxed{h_A(D,B) = \sum_A h_C(A,B)h_R(A)P(D \mid A,B)}$$

$\boxed{h_A(D,B)}$

Eliminamos B

$$\boxed{P(D) = \sum_B h_A(B,D)}$$

$\boxed{P(D)}$

Complejidad del Algoritmo de Eliminación

- **Estudiemos la complejidad:**
  - En cada paso se calcula:

$$h_X(Y_1,\dots Y_k) = \sum_X h'_X(X,Y_1,\dots Y_k), \qquad h'_X(X,Y_1,\dots Y_k) = \prod_{i=1}^{m} f_i(X,Y_{i1},\dots Y_{il_i})$$

  - Entonces:
    - Para cada valor de Y_1, Y_2,..Y_k hacemos |Val(X)| sumas, por tanto el número total de sumas es $|Val(X)|\times|Val(Y_1)|\times\dots\times|Val(Y_k)|$
    - Para cada valor de X,Y_1, Y_2,..Y_k hacemos m multiplicaciones, por tanto el número total de de multiplicaciones es de m $|Val(X)|\times|Val(Y_1)|\times\dots\times|Val(Y_k)|$
  - Por tanto: la complejidad es exponencial en el número de variables de los factores intermedios
  - Es necesario buscar "buenos" ordenes de eliminación de variables para reducir el tamaño de los factores intermedios.
  - Sin embargo, el problema de buscar la ordenación óptima tiene también complejidad exponencial, por lo que se emplean diversas heurísticas.

Decisión con Redes Bayesianas

- **Decisión:**
  - En el tema anterior se presentó como razonar en presencia de incertidumbre. Veremos ahora como tomar decisiones (actuar) bajo incertidumbre
- **Ejemplo:**
  - Quiero para hacer una fiesta en el jardín de mi casa pero puede que llueva ¿debería hacer la fiesta dentro de casa?

    En el problema aparecen:
    - Dos posibles acciones: preparar la fiesta dentro o prepararla fuera.
    - Cuatro posibles resultados: la fiesta se hace dentro o la fiesta se hace fuera con lluvia o sin lluvia
    - Una distribución de probabilidad sobre los resultados (no sé si lloverá o no)
    - Un conjunto de preferencias sobre los resultados: prefiero hacer la fiesta fuera sin lluvia

		Lugar	
		dentro	fuera
Tiempo	lluvioso	Aliviado	Deprimido
	¬lluvioso	Arrepentido	Contento

Preferencias sobre los resultados

  - ¿Cómo tomar entonces la decisión óptima?

Teoría de la Decisión

- **La teoría de la decisión bajo incertidumbre:**
  - Es la combinación de dos elementos básicos: la teoría de la probabilidad y la teoría de la utilidad.
  - La teoría de la utilidad
    - Establece una estructura de preferencias racional sobre los resultados de nuestras acciones basada en un conjunto de axiomas (Savage)
    - Afirma que bajo esos axiomas es posible encontrar una función de utilidad U que asigna un valor numérico a la utilidad de cada resultado que cumple:
      - Si el resultado A se prefiere al B entonces U(A)≥U(B)
      - Si A y B son indiferentes entonces U(A)=U(B)
      - Si se dan varios posibles resultados A_1, A_2,..., A_n con probabilidades p_1, p_2,..., p_n la utilidad del resultado conjunto es $U([p_1, A_1; p_2, A_2;\dots p_n, A_n])=p_1U(A_1)+\dots p_nU(A_n)$
  - Ejemplo de función de utilidad

U(Tiempo, Lugar)

		Lugar	
		dentro	fuera
Tiempo	lluvioso	Aliviado (60)	Deprimido (0)
	¬lluvioso	Arrepentido (50)	Contento (100)

El Principio de la Máxima Utilidad Esperada (MUE)

- **El principio de la MUE es un principio fundamental en Inteligencia Artificial.**
- **Establece que:**
  - La forma racional de decidir para un agente es elegir aquella acción cuyo resultado le sea sea más útil (más preferido).
  - Entonces dada una acción A con:
    - Posibles resultados: {$Resultado_1(A)$,$Resultado_2(A)$,... $Resultado_n(A)$}
    - Probabilidades de obtener dichos resultados si se tiene la evidencia E: $P(Resultado_1(A)|Hacer(A),E)$,..., $P(Resultado_n(A)|Hacer(A),E)$

    será óptima si hace máxima la utilidad esperada:

$$U(A \mid E) = \sum_{i=1}^{n} P(Resultado_i(A) \mid Hacer(A), E)\, U(Resultado_i(A))$$

    - Esto es, es la suma de la utilidad de cada resultado multiplicado por la probabilidad de obtenerlo.

El Principio de la Máxima Utilidad Esperada (MUE): Ejemplo

- **Ejemplo:**
  - En el caso de la fiesta las acciones son:
    - A_1="preparar fiesta dentro", A_2="preparar fiesta fuera"
  - Los resultados son:
    - $Res_1(A_1)$= (lluvioso,dentro), $Res_2(A_1)$= (¬lluvioso,dentro)
    - $Res_1(A_2)$= (lluvioso,fuera), $Res_2(A_2)$= (¬lluvioso,fuera)
  - La función de utilidad es:

U(Tiempo,Lugar)

Tiempo \ Lugar	dentro	fuera
lluvioso	Aliviado (60)	Deprimido (0)
¬lluvioso	Arrepentido (50)	Contento (100)

  - Las probabilidades son:
    - $P(Res_1(A_1)|Hacer(A_1))$=P(lluvioso,dentro)/P(dentro)=P(lluvioso)
    - $P(Res_2(A_1)|Hacer(A_1))$=P(¬lluvioso)
    - $P(Res_1(A_2)|Hacer(A_2))$=P(lluvioso)
    - $P(Res_2(A_2)|Hacer(A_2))$=P(¬lluvioso)

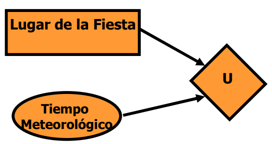
P(T)

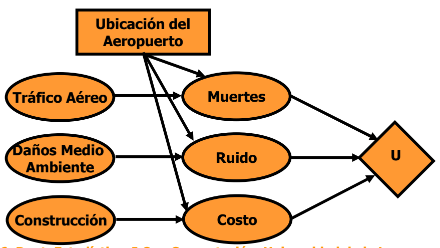
T	
lluvioso	0.4
¬lluvioso	0.6

  - Las utilidades son:
    - U(dentro)=U(lluvioso,dentro)P(lluvioso,dentro|dentro)+U(¬lluvioso,dentro)P(¬lluvioso,dentro|dentro)= 60*0.4+50*0.6=54
    - U(fuera)=U(lluvioso,fuera)P(lluvioso,fuera|fuera)+U(¬lluvioso,fuera)P(¬lluvioso,fuera|fuera)= 0*0.4+100*0.6=60 (decisión óptima)

Redes de Decisión (Diagramas de influencia)

- **Una red de decisión define un escenario con una sucesión de observaciones y decisiones**
- **Una red de decisión está compuesta de:**
  - Nodos aleatorios (óvalos): Representan variables aleatorias de la misma forma que las redes de creencia.
  - Nodos de decisión (rectángulos): Representan puntos para los cuales puede decidirse que acción emprender
  - Nodos de utilidad (rombos): Representan la función de utilidad.
- **Ejemplos**





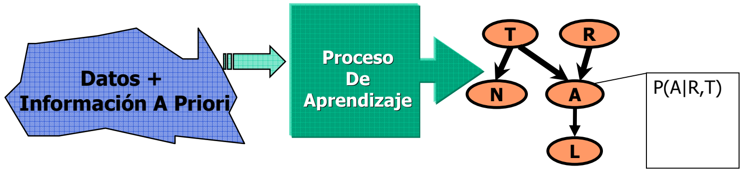
Aprendizaje

- **Aprendizaje:**
  - "Cualquier cambio en un sistema que le permita obtener un mejor rendimiento la segunda vez que realiza la misma tarea u otra tarea similar" (Simon)
- **¿Por qué realizar aprendizaje en sistemas basados en conocimiento?**
  - El proceso de adquisición del conocimiento es muy caro
  - Frecuentemente no se tienen expertos disponibles
  - Por el contrario generalmente es posible disponer de grandes cantidades de datos.
- **El Aprendizaje nos permite diseñar sistemas basados en datos**
  - Además estos datos pueden combinarse con las opiniones de distintos expertos

1

Aprendizaje en Redes Bayesianas

- **Proceso de Aprendizaje General**
  - Inferir la estructura y tablas de probabilidad condicional a partir de datos e información a priori



- **El Problema del Aprendizaje en Redes Bayesianas**

	Estructura Conocida	**Estructura Desconocida**
Datos completos	Estimación Paramétrica	Optimización sobre estructuras
Datos Incompletos	Optimización Paramétrica	Técnicas Combinadas

1

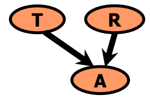
Aprendizaje en RB: Datos Completos y Estructura Conocida (DC/EC)

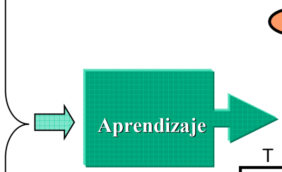
- **En este caso:**
  - La estructura de la red es conocida.
  - El Proceso de Aprendizaje nos proporciona los parámetros que describen las tablas de probabilidad condicional

Casos

T, R, A
< t, ¬r, ¬a>
.
.
.
<¬t, r, ¬a>

T	R	P(A|R,T)	
t	r	?	?
t	¬r	?	?
¬t	r	?	?
¬t	¬r	?	?





T	R	P(A|R,T)	
t	r	0.99	0.01
t	¬r	0.2	0.8
¬t	r	0.8	0.2
¬t	¬r	0.01	0.99

1

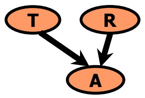
Aprendizaje de Redes Bayesianas en Netica

- **Sólo resuelve los casos de Estructura Conocida**
- **Estimación de los parámetros (Datos Completos)**
  - Dado un caso que proporciona valores para un nodo y sus padres, la nueva probabilidad condicional p' y el nuevo número de casos de los padres e' se actualizan para esos valores en función de los anteriores p y e como:

    e'=e+1 p'=(p×e+1)/(e+1) (si el estado del nodo coincide con el valor del caso para ese nodo)

    p'=(p×e)/(e+1) (si no coincide con el valor)

- **Ejemplo:**



T	R	P(A|R,T)	
t	r	0.99	0.01
t	¬r	0.2	0.8
¬t	r	0.8	0.2
¬t	¬r	0.01	0.99

T	R	E
t	r	1
t	¬r	3
¬t	r	20
¬t	¬r	980

Casos

T, R, A
< t, ¬r, ¬a>
.
< ¬t, r, a>
.
<¬t, r, ¬a>

T	R	P'(A|R,T)	
t	r	0.99	0.01
t	¬r	0.2	0.8
¬t	r	0.81	0.19
¬t	¬r	0.01	0.99

T	R	E'
t	r	1
t	¬r	3
¬t	r	21
¬t	¬r	980

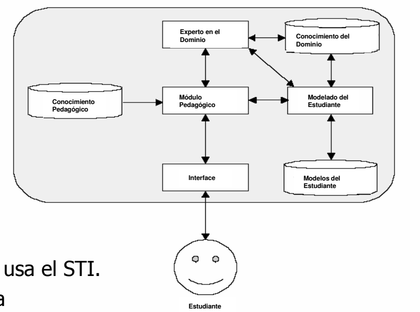
2

Aplicaciones de Redes Bayesianas en Educación

- **Sistemas Tutoriales Inteligentes (STI)**
  - Son sistemas informáticos para la enseñanza de estudiantes.
- **¿Por qué construirlos?**
  - Idealmente permiten un "profesor" por alumno y por tanto se tiene un profesor que se adapta:
    - A las características personales del alumno
    - A su ritmo de aprendizaje
    - A sus horarios
  - Sin embargo...
    - La capacidad actual de procesamiento del lenguaje natural no permite conversar de forma normal con un STI
    - Siempre habrá estudiantes para los que la enseñanza por ordenador no sea adecuada
- **Objetivo actual:**
  - Construir sistemas informáticos para ayudar al profesor, en clase, en el lugar de trabajo o en casa.

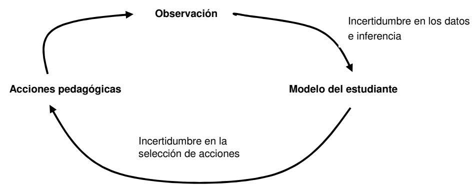
2 Arquitectura de un STI

- **Elementos de la arquitectura**
  - Experto en el dominio
    - Representación del Conocimiento del dominio a enseñar
  - Modelado del estudiante
    - Representación del estudiante que usa el STI.
    - Es el módulo de mayor importancia
  - Modelos del estudiante
    - Almacenamiento de los distintos modelos de estudiantes que usan el sistema
  - Módulo Pedagógico
    - Subsistema que toma las decisiones acerca de cómo enseñar el dominio basado en el Conocimiento pedagógico.
  - Interface
    - Módulo de comunicación del STI con el estudiante



2 Incertidumbre en los STI

- **Incertidumbre en el modelado del estudiante**
  - En los datos
    - El STI debe construir el modelo del estudiante a partir de un conjunto de datos muy limitado. (Generalmente limitados a respuestas del teclado y ratón)
  - En la inferencia
    - Las reglas para la construcción del modelo del estudiante a partir de datos suelen ser heurísticas (y por tanto subóptimas).
  - En la selección de acciones
    - La incertidumbre en el modelo del estudiante se traslada a la selección de la acción pedagógica más adecuada.



2 STI con Redes Bayesianas

- **Modelado del estudiante con Redes Bayesianas**
  - Las propuestas se pueden dividir en tres grupos:
    - Centradas en expertos
      - Basadas en utilizar expertos que especifican de manera general y de forma directa o indirecta la estructura completa y las tablas de probabilidad condicional del modelo del estudiante
      - Ejemplos:
        - » ANDES (Gertner & Van Lehn 2000) http://www.andes.pitt.edu/
        - » HYDRIVE (Miselvy & Gitomer, 1996)
        - » DT- Tutor (Murry & VanLenh, 2000)
        - » ADELE (Ganeshan y otros 2000)
      - Ventajas
        - » La utilización de expertos proporciona modelos de gran calidad
      - Principal inconveniente:
        - » Los modelos resultantes de las propuestas de los expertos incluyen tantas variables que puede ser infactible trabajar con la red bayesiana en tiempo real.

2 STI con Redes Bayesianas

- Centradas en la eficiencia
  - Basadas en la idea de restringir los tipos de modelos permitidos y "ajustar" el conocimiento del dominio a dichos modelos. Estas restricciones se eligen generalmente de forma que se optimice algún aspecto de la eficiencia como por ejemplo el tiempo de realizar inferencias sobre la red.
  - Ejemplos:
    - » (Reye, 1998)
    - » (Murray, 1998)
    - » (Collins y otros, 1996)
    - » (Mayo y Mitrovic, 2000)
  - Ventajas
    - » Eficiencia
    - » Los modelos utilizados permiten modelar la adquisición del conocimiento por parte del alumno a través del tiempo.
  - Inconvenientes
    - » La búsqueda de la eficiencia puede introducir simplificaciones incorrectas acerca del dominio.

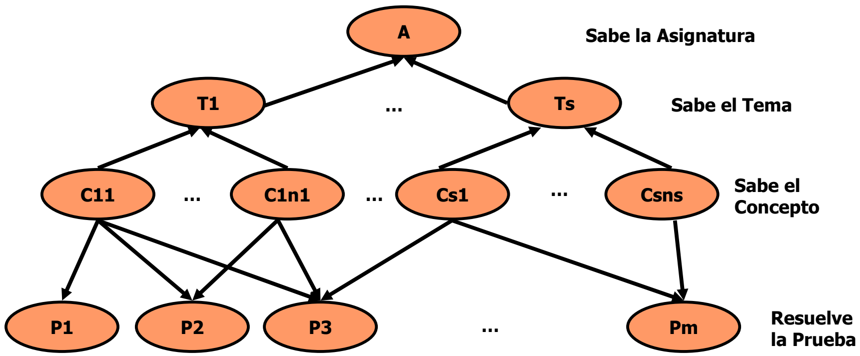
STI con Redes Bayesianas

- Centradas en los datos
  - Basadas en la idea de aprender tanto la estructura como las probabilidades condicionales de la red del trabajo en tiempo real del tutor.
  - Ejemplos:
    - » MANIC (Stern y otros, 1999)
    - » CAPIT (Mayo y Mitrovic, 2001)
  - Ventajas
    - » Tienden a ser más simples al estar basados en variables observadas
    - » Permiten evaluar la calidad del modelo
    - » Los modelos utilizados permiten modelar la adquisición del conocimiento por parte del alumno a través del tiempo.
  - Inconvenientes
    - » Requieren grandes cantidades de datos

Selección de acciones pedagógicas

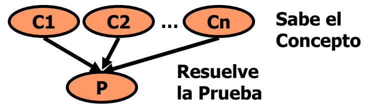
- **Una vez obtenido el modelo del estudiante, éste debe usarse para elegir la acción pedagógica óptima**
  - Tipos de estrategias
    - Heurísticas
      - Utilizan la salida del proceso de inferencia como entrada a una regla de selección heurística
      - Ejemplos
        - » ANDES, ADELE
    - Diagnóstico
      - Seleccionan la acción que maximizan la certidumbre de que el estudiante ha adquirido los conceptos del dominio
      - Ejemplos
        - » (Collins y otros, 1996)
    - Teoría de la decisión
      - Seleccionan la acción que maximiza su utilidad esperada
      - Ejemplos
        - » CAPIT, DT-Tutor

Ejemplo del Modelo del Alumno: Parte Cualitativa

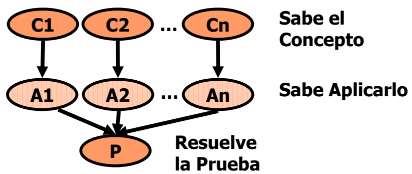


Ejemplo del Modelo del Alumno: Parte Cuantitativa (1)

- **Tablas de Probabilidad Conceptos-Ejercicios**



  - Cuando un ejercicio depende de varios conceptos la tabla de probabilidad condicional puede ser muy grande.
  - Generalmente los conceptos no son independientes, pero se puede asumir que la capacidad de aplicarlos cada concepto correctamente cuando se sabe si es independiente. Entonces se introduce la red:



2 Ejemplo del Modelo del Alumno: Parte Cuantitativa (2)

- **Llamaremos:**
  - $P(A_i=0|C_i=1)=d_i$ a la probabilidad de "descuido", el alumno sabe el concepto, pero se equivocó al aplicarlo.
  - $P(A_i=1|C_i=0)=s_i$ a la probabilidad de "suerte", el alumno no sabe el concepto, pero acertó al aplicarlo.
- **Entonces:**
  - Cuando para resolver un ejercicio es necesario conocer todos los conceptos aparece el modelo de probabilidades condicionales "Noisy AND".
  - Cuando para resolver un ejercicio es necesario conocer algún concepto aparece el modelo de probabilidades condicionales "Noisy OR".

2 Ejemplo del Modelo del Alumno: Parte Cuantitativa (3)

- **"Noisy Or" en el programa Netica:**
  - Parámetro $p_i=1-s_i/(1-g_i)$
  - Parámetro $leak=\prod(1-g_i)$
- **"Noisy And" en el programa Netica:**
  - Parámetro $p_i=1-g_i/(1-s_i)$
  - Parámetro $Inh=1-\prod(1-s_i)$
- **Las relaciones entre Conceptos, Temas y Asignaturas se modelan de forma similar**