

# Inteligencia Artificial Incertidumbre

Primavera 2007

profesor: Luigi Ceccaroni



# Comportamiento bajo incertidumbre

- Casi nunca se pueden hacer las asignaciones epistemológicas de que las proposiciones son ciertas, falsas o desconocidas.
- En la práctica, los programas tienen que saber actuar en situaciones de **incertidumbre**:
  - usando una teoría del mundo simple pero errónea, que no tiene en cuenta la incertidumbre y que funciona la *mayoría* de las veces;
  - manejando el conocimiento incierto y la utilidad de manera racional:
    - Lo correcto a realizar (la **decisión racional**) depende tanto de la importancia relativa de los distintos objetivos como de la verosimilitud y el grado con el cual se conseguirán.

# Manipulación del conocimiento incierto

- Ejemplo de regla para diagnóstico usando lógica de predicados de primer orden:  
$$\forall p \text{ Síntoma}(p, \text{Dolor-de-muelas}) \Rightarrow \text{Enfermedad}(p, \text{Caries})$$
- Esta regla es errónea y, para hacerla cierta, hay que añadir una lista de causas:  
$$\forall p \text{ Síntoma}(p, \text{Dolor-de-muelas}) \Rightarrow \text{Enfermedad}(p, \text{Caries}) \vee \text{Enfermedad}(p, \text{Dolor-de-muelas}) \vee \text{Enfermedad}(p, \text{Absceso}) \dots$$
- Usar la lógica de predicados de primer orden en un dominio como el diagnóstico falla por tres razones principales:
  1. **Pereza:** poner en una lista el conjunto completo de antecedentes y consecuentes que se necesitan para asegurar una regla sin excepciones tiene demasiado trabajo.
  2. **Ignorancia teórica:** la ciencia no tiene una teoría completa para el dominio.
  3. **Ignorancia práctica:** incluso si se conocen todas las reglas, pudiera haber incertidumbre sobre un paciente particular, ya sea porque no se hayan realizado todos los chequeos necesarios o porque no puedan realizarse.

# Manipulación del conocimiento incierto

- En realidad, la conexión entre dolor de muelas y caries no es exactamente una consecuencia lógica en ninguna dirección.
- En dominios sentenciosos, el conocimiento de un agente proporciona sólo un **grado de creencia** en las oraciones.
- La herramienta para tratar con grados de creencia es la **teoría de la probabilidad**, que asigna a cada oración un grado numérico entre 0 y 1.
- La probabilidad proporciona una manera de **resumir** la incertidumbre que se deriva de nuestra pereza e ignorancia.

# Manipulación del conocimiento incierto

- La creencia puede provenir de datos estadísticos o de reglas generales o de una combinación de fuentes de indicios.
- Asignar probabilidad 0 a una oración determinada corresponde a una creencia inequívoca de que la oración es falsa.
- Asignar una probabilidad de 1 corresponde a una creencia rotunda de que la oración es cierta.
- Las probabilidades entre 0 y 1 corresponden a grados intermedios de creencia en la veracidad de la oración.

# Manipulación del conocimiento incierto

- La oración en sí misma es *de hecho* o **verdadera o falsa**.
- El grado de creencia **es diferente del grado de veracidad**.
- Una probabilidad de 0.8 no significa “80% verdadero” sino una expectativa muy fuerte (del 80%) de que algo sea verdadero.
- La teoría de la probabilidad cumple la misma obligación ontológica que la lógica: los hechos del mundo o son verdaderos o no.
- Los grados de veracidad son la materia de la **lógica borrosa**.

# Manipulación del conocimiento incierto

- En lógica, una oración tal como *“El paciente tiene una caries”* es verdadera o falsa.
- En teoría de la probabilidad, la oración *“La probabilidad de que el paciente tiene una caries es 0.8”* hace referencia a creencias de un agente, no directamente al mundo.
- Estas creencias dependen de las percepciones que el agente ha recibido hasta el momento.
- Estas percepciones constituyen la **evidencia** sobre la que se basan las probabilidades.
- Por ejemplo:
  - Un agente saca una carta de un mazo barajado.
  - Antes de mirar la carta, el agente asignaría una probabilidad de  $1/52$  de que se trata del as de picas.
  - Después de mirar la carta, la probabilidad para la misma proposición debería ser 0 o 1.

# Manipulación del conocimiento incierto

- Una asignación de probabilidad a una proposición es análogo a decir si una oración lógica determinada está producida por la *base de conocimiento*, más que si es o no cierta.
- Todas las oraciones deben así indicar la evidencia con respecto a la cual se está calculando la probabilidad.
- Cuando un agente recibe nuevas percepciones/evidencias, sus valoraciones de probabilidad se actualizan.
- Antes de que la evidencia se obtenga, se habla de probabilidad *a priori* o **incondicional**.
- Después de obtener la evidencia, se habla de probabilidad *a posteriori* o **condicional**.

# Notación básica con probabilidades

- Proposiciones
  - Los grados de creencia se aplican siempre a las **proposiciones**, afirmaciones de que tal o cual es el caso.
  - El elemento básico del lenguaje es la **variable aleatoria**, que puede pensarse como algo que se refiere a una “parte” del mundo cuyo estado es desconocido inicialmente.
  - Por ejemplo, *Caries* podría referirse a si mi muela del juicio inferior izquierda tiene una caries.
  - Cada variable aleatoria tiene un **dominio** de posibles valores que puede tomar.

# Proposiciones

- Como con las variables PSR, las variables aleatorias (VAs) están típicamente divididas en tres clases, dependiendo del tipo de dominio:
  - Las VAs **booleanas**, tal como *Caries*, tienen el dominio *<cierto, falso>*.
  - Las VAs **discretas**, que incluyen las VAs booleanas como un caso especial, toman valores en un dominio contable.
  - Las VAs **continuas** toman sus valores de los números reales.

# Sucesos atómicos

- Un **suceso atómico** es una especificación *completa* del estado del mundo.
- Es la asignación de valores particulares de todas las variables que componen el mundo.
- Ejemplo:
  - Si mi mundo consta sólo de las variables booleanas *Caries* y *Dolor-de-muelas*, entonces hay exactamente cuatro sucesos atómicos.
  - La proposición “*Caries = falso*  $\wedge$  *Dolor-de-muelas = cierto*” es uno de tales sucesos.

# Probabilidad a priori

- La **probabilidad *a priori*** o **incondicional** asociada a una proposición  $a$  es el grado de creencia que se le otorga *en ausencia de cualquier otra información*.
- Se escribe como  $P(a)$ .
- Ejemplo:
  - $P(\text{Caries} = \text{cierto}) = 0.1$  o  $P(\text{caries}) = 0.1$
- Es importante recordar que  $P(a)$  puede usarse sólo cuando no hay otra información.

# Probabilidad a priori

- Para hablar de las probabilidades de todos los valores posibles de una VA:
  - Usaremos una expresión como  $\mathbf{P}(Tiempo)$ , que denota un vector de valores que corresponden a las probabilidades de cada estado individual del tiempo.
  - $\mathbf{P}(Tiempo) = \langle 0.7, 0.2, 0.08, 0.02 \rangle$  (**normalizado**, i.e., suma 1)
  - (El dominio de *Tiempo* es  $\langle soleado, lluvioso, nuboso, nevado \rangle$ )
- Esta expresión define una **distribución de probabilidad a priori** para la VA *Tiempo*.

# Probabilidad a priori

- Expresiones como  $P(\textit{Tiempo}, \textit{Caries})$  se usan para indicar las probabilidades de todas las combinaciones de los valores de un conjunto de VAs.
- En este caso se hablaría de **distribución de probabilidad conjunta** de *Tiempo* y *Caries*.
- Todas las preguntas sobre un dominio se pueden contestar con la distribución conjunta.

# Probabilidad a priori

- La distribución de probabilidad conjunta para un conjunto de VAs proporciona la probabilidad de cada suceso atómico que involucre esas VAs.

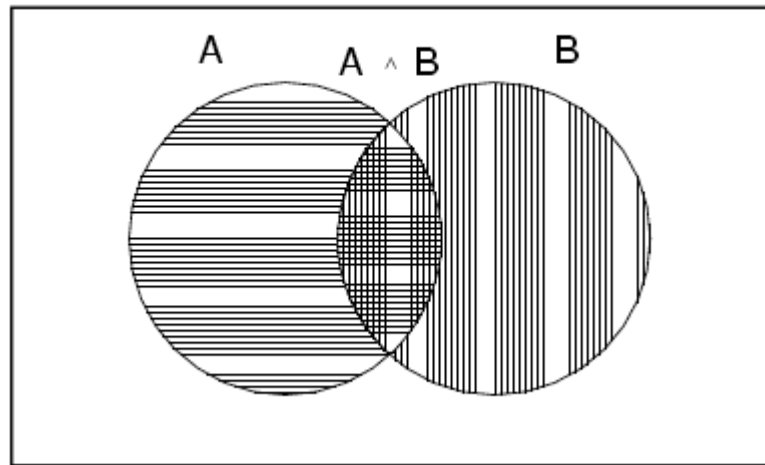
$\mathbf{P}(\text{Tiempo}, \text{Caries})$  = una matriz  $4 \times 2$  de valores de probabilidad:

<i>Tiempo</i> =	soleado	lluvioso	nuboso	nevado
<i>Caries</i> = cierto	0.144	0.02	0.016	0.02
<i>Caries</i> = falso	0.576	0.08	0.064	0.08

# Los axiomas de la probabilidad

- $0 \leq P(a) \leq 1$
- $P(\text{cierto}) = 1$        $P(\text{falso}) = 0$
- $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$

True



# Conditional probability

- Conditional or posterior probabilities:  
e.g.,  $P(\text{cavity} \mid \text{toothache}) = 0.8$   
i.e., given that *toothache* is all I know
- (Notation for conditional distributions:  
 $\mathbf{P}(\text{Cavity} \mid \text{Toothache}) = 2\text{-element vector of } 2\text{-element vectors}$ )
- If we know more, e.g., *cavity* is also given, then we have  
 $P(\text{cavity} \mid \text{toothache}, \text{cavity}) = 1$  (trivial)
- New evidence may be irrelevant, allowing simplification, e.g.,  
 $P(\text{cavity} \mid \text{toothache}, \text{sunny}) = P(\text{cavity} \mid \text{toothache}) = 0.8$
- This kind of inference, sanctioned by domain knowledge, is crucial.

# Conditional probability

- Definition of conditional probability:  
 $P(a | b) = P(a \wedge b) / P(b)$  if  $P(b) > 0$
- **Product rule** gives an alternative formulation:  
 $P(a \wedge b) = P(a | b) P(b) = P(b | a) P(a)$
- A general version holds for whole distributions, e.g.,  
 $P(\textit{Weather}, \textit{Cavity}) = P(\textit{Weather} | \textit{Cavity}) P(\textit{Cavity})$ 
  - (View as a set of  $4 \times 2$  equations, **not** matrix multiplication)
- **Chain rule** is derived by successive application of product rule:  
$$\begin{aligned} P(X_1, \dots, X_n) &= P(X_1, \dots, X_{n-1}) P(X_n | X_1, \dots, X_{n-1}) \\ &= P(X_1, \dots, X_{n-2}) P(X_{n-1} | X_1, \dots, X_{n-2}) P(X_n | X_1, \dots, X_{n-1}) \\ &= \dots \\ &= \prod_{i=1}^n P(X_i | X_1, \dots, X_{i-1}) \end{aligned}$$

# Inference by enumeration

- A simple method for **probabilistic inference** uses observed evidence for computation of posterior probabilities.
- Start with the joint probability distribution:

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	<b>.108</b>	<b>.012</b>	<b>.072</b>	<b>.008</b>
$\neg$ <i>cavity</i>	<b>.016</b>	<b>.064</b>	<b>.144</b>	<b>.576</b>

- For any proposition  $\varphi$ , sum the atomic events where it is true:  $P(\varphi) = \sum_{\omega: \omega \models \varphi} P(\omega)$

# Inference by enumeration

- Start with the joint probability distribution:

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	<b>.108</b>	<b>.012</b>	<b>.072</b>	<b>.008</b>
$\neg$ <i>cavity</i>	<b>.016</b>	<b>.064</b>	<b>.144</b>	<b>.576</b>

- For any proposition  $\varphi$ , sum the atomic events where it is true:  $P(\varphi) = \sum_{\omega:\omega \models \varphi} P(\omega)$
- $P(\textit{toothache}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$

# Inference by enumeration

- Start with the joint probability distribution:

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	<b>.108</b>	<b>.012</b>	<b>.072</b>	<b>.008</b>
$\neg$ <i>cavity</i>	<b>.016</b>	<b>.064</b>	<b>.144</b>	<b>.576</b>

- For any proposition  $\varphi$ , sum the atomic events where it is true:  $P(\varphi) = \sum_{\omega:\omega \models \varphi} P(\omega)$
- $P(\textit{toothache} \vee \textit{cavity}) = 0.108 + 0.012 + 0.016 + 0.064 + 0.072 + 0.008 = 0.28$

# Inference by enumeration

- Start with the joint probability distribution:

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	<b>.108</b>	<b>.012</b>	<b>.072</b>	<b>.008</b>
$\neg$ <i>cavity</i>	<b>.016</b>	<b>.064</b>	<b>.144</b>	<b>.576</b>

- Conditional probabilities:

$$\begin{aligned} P(\neg \text{cavity} \mid \text{toothache}) &= \frac{P(\neg \text{cavity} \wedge \text{toothache})}{P(\text{toothache})} \\ &= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} \\ &= 0.4 \end{aligned}$$

# Normalization

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	<b>.108</b>	<b>.012</b>	<b>.072</b>	<b>.008</b>
$\neg$ <i>cavity</i>	<b>.016</b>	<b>.064</b>	<b>.144</b>	<b>.576</b>

- Denominator can be viewed as a **normalization constant**  $\alpha$

$$\begin{aligned} \mathbf{P}(\text{Cavity} \mid \text{toothache}) &= \alpha, \mathbf{P}(\text{Cavity}, \text{toothache}) \\ &= \alpha, [\mathbf{P}(\text{Cavity}, \text{toothache}, \text{catch}) + \mathbf{P}(\text{Cavity}, \text{toothache}, \neg \text{catch})] \\ &= \alpha, [<0.108, 0.016> + <0.012, 0.064>] \\ &= \alpha, <0.12, 0.08> = <0.6, 0.4> \end{aligned}$$

General idea: compute distribution on query variable by fixing **evidence variables** and summing over **hidden variables**

# Inference by enumeration

Typically, we are interested in:

the posterior joint distribution of the **query variables**  $\mathbf{X}$  given specific values  $\mathbf{e}$  for the **evidence variables**  $\mathbf{E}$ .

Let the **hidden variables** be  $\mathbf{Y}$

Then the required summation of joint entries is done by summing out the hidden variables:

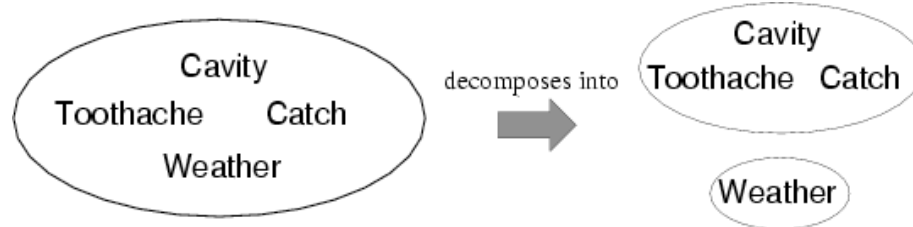
$$\mathbf{P}(\mathbf{X} \mid \mathbf{E} = \mathbf{e}) = \alpha \mathbf{P}(\mathbf{X}, \mathbf{E} = \mathbf{e}) = \alpha \sum_{\mathbf{y}} \mathbf{P}(\mathbf{X}, \mathbf{E} = \mathbf{e}, \mathbf{Y} = \mathbf{y})$$

- $\mathbf{X}$ ,  $\mathbf{E}$  and  $\mathbf{Y}$  together exhaust the set of random variables.
- Obvious problems:
  - Worst-case time complexity  $O(d^n)$  where  $d$  is the largest arity and  $n$  is the number of variables
  - Space complexity  $O(d^n)$  to store the joint distribution
  - How to define the probabilities for  $O(d^n)$  entries, when variables can be millions?

# Independence

- $A$  and  $B$  are independent iff

$$\mathbf{P}(A|B) = \mathbf{P}(A) \quad \text{or} \quad \mathbf{P}(B|A) = \mathbf{P}(B) \quad \text{or} \quad \mathbf{P}(A, B) = \mathbf{P}(A) \mathbf{P}(B)$$



$$\mathbf{P}(\textit{Toothache}, \textit{Catch}, \textit{Cavity}, \textit{Weather}) \\ = \mathbf{P}(\textit{Toothache}, \textit{Catch}, \textit{Cavity}) \mathbf{P}(\textit{Weather})$$

- 32 entries reduced to 12; for  $n$  independent biased coins,  $O(2^n) \rightarrow O(n)$
- Absolute independence powerful but rare
- Dentistry is a large field with hundreds of variables, none of which are independent. What to do?

# Conditional independence

- $\mathbf{P}(\textit{Toothache}, \textit{Cavity}, \textit{Catch})$  has  $2^3 - 1$  (because the numbers must sum to 1) = 7 independent entries
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:  
$$\mathbf{P}(\textit{catch} \mid \textit{toothache}, \textit{cavity}) = \mathbf{P}(\textit{catch} \mid \textit{cavity})$$
- The same independence holds if I haven't got a cavity:  
$$\mathbf{P}(\textit{catch} \mid \textit{toothache}, \neg \textit{cavity}) = \mathbf{P}(\textit{catch} \mid \neg \textit{cavity})$$
- *Catch* is **conditionally independent** of *Toothache* given *Cavity*:  
$$\mathbf{P}(\textit{Catch} \mid \textit{Toothache}, \textit{Cavity}) = \mathbf{P}(\textit{Catch} \mid \textit{Cavity})$$
- Equivalent statements:  
$$\mathbf{P}(\textit{Toothache} \mid \textit{Catch}, \textit{Cavity}) = \mathbf{P}(\textit{Toothache} \mid \textit{Cavity})$$
  
$$\mathbf{P}(\textit{Toothache}, \textit{Catch} \mid \textit{Cavity}) = \mathbf{P}(\textit{Toothache} \mid \textit{Cavity}) \mathbf{P}(\textit{Catch} \mid \textit{Cavity})$$

# Conditional independence

- Full joint distribution using chain rule:

$$\mathbf{P}(\textit{Toothache}, \textit{Catch}, \textit{Cavity})$$

$$= \mathbf{P}(\textit{Toothache} \mid \textit{Catch}, \textit{Cavity}) \mathbf{P}(\textit{Catch}, \textit{Cavity})$$

$$= \mathbf{P}(\textit{Toothache} \mid \textit{Catch}, \textit{Cavity}) \mathbf{P}(\textit{Catch} \mid \textit{Cavity}) \mathbf{P}(\textit{Cavity})$$

$$= \mathbf{P}(\textit{Toothache} \mid \textit{Cavity}) \mathbf{P}(\textit{Catch} \mid \textit{Cavity}) \mathbf{P}(\textit{Cavity})$$

I.e.,  $2 + 2 + 1 = 5$  independent numbers

- In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in  $n$  to linear in  $n$ .
- Conditional independence is our most basic and robust form of knowledge about uncertain environments.

# Bayes' Rule

- Product rule  $P(a \wedge b) = P(a | b) P(b) = P(b | a) P(a)$   
 $\Rightarrow$  **Bayes' rule:**  $P(a | b) = P(b | a) P(a) / P(b)$
- or in distribution form  
$$\mathbf{P(Y|X) = P(X|Y) P(Y) / P(X) = \alpha P(X|Y) P(Y)}$$
- Useful for assessing **diagnostic** probability from **causal** probability:
  - $P(\text{Cause}|\text{Effect}) = P(\text{Effect}|\text{Cause}) P(\text{Cause}) / P(\text{Effect})$
  - E.g., let  $M$  be meningitis,  $S$  be stiff neck:  
 $P(m|s) = P(s|m) P(m) / P(s) = 0.8 \times 0.0001 / 0.1 = 0.0008$
  - Note: posterior probability of meningitis still very small!

# Bayes' Rule and conditional independence

$$\begin{aligned} P(\text{Cavity} \mid \text{toothache} \wedge \text{catch}) \\ &= \alpha P(\text{toothache} \wedge \text{catch} \mid \text{Cavity}) P(\text{Cavity}) \\ &= \alpha P(\text{toothache} \mid \text{Cavity}) P(\text{catch} \mid \text{Cavity}) P(\text{Cavity}) \end{aligned}$$

- This is an example of a **naïve Bayes** model:

$$P(\text{Cause}, \text{Effect}_1, \dots, \text{Effect}_n) = P(\text{Cause}) \prod_i P(\text{Effect}_i \mid \text{Cause})$$



- Total number of parameters (the size of the representation) is **linear** in  $n$ .

# Summary

- Probability is a rigorous formalism for uncertain knowledge.
- **Joint probability distribution** specifies probability of every **atomic event**.
- Queries can be answered by summing over atomic events.
- For nontrivial domains, we must find a way to reduce the joint size.
- **Independence** and **conditional independence** provide the tools.