Lecture 11: Quantifying Uncertainty

Artificial Intelligence

CS-6364

Acting under uncertainty

- Logical agents assume propositions are
  - True
  - False
  - Unknown → *acting under uncertainty*

Example: diagnosis (for medicine) → dental diagnosis using 1st order logic:

$$\forall p\ Symptom(p, Toothache) \Rightarrow Disease(p, Cavity)$$

Wrong: not all patients *p* with toothaches have cavities! Some have gum disease, an abscess or other problems

$$\forall p\ Symptom(p, Toothache) \Rightarrow$$
$$Disease(p, Cavity) \vee Disease(p, GumDisease) \vee Disease(p, Abscess)$$

Conclusion: To make the rule true, we have to add an almost unlimited list of possible causes

Medical Diagnosis

Trying to use first-order logic to cope with a domain like medical diagnosis fails because

1. Laziness – too much work to list the complete list of rules + too hard to use such rules

   Example of causal rule:

   $$\forall p\ Disease(p, Cavity) \Rightarrow Symptom(p, Toothache)$$

   wrong, not all cavities cause pain → need to augment the antecedent with all conditions that cause toothaches

2. Theoretical ignorance: Medical science has no complete theory for the domain
3. Practical ignorance: Even if we know all the rules, we might be uncertain about a particular patient, because not all necessary tests have been or can be run

Degree of belief

When propositions are not known to be true or false, the agent can at best provide a <u>degree of belief</u> in relevant sentences.

- The main tool for dealing with degrees of belief is the Probability Theory

```
function DT-AGENT(percept) returns an action
  static belief_state, probabilistic beliefs about the current state of the world
         action, the agent's action

  update belief_state based action on and percept
  calculate outcome probabilities for actions,
      given action descriptions and current belief_state
  select action with highest expected utility
      given probabilities of outcomes and utility information
  return action
```

Uncertainty and Rational Decisions

- *To make choices, an agent must have preferences between the possible outcomes of possible plans of action.*
- *An **outcome** is a fully specified state.*
- *To represent and reason with **preferences,** we rely on utility theory.*
- *Preferences, as expressed by utilities, are combined with probabilities in the theory of rational decisions called decision theory:*

Decision Theory = probability theory + utility theory

- *The fundamental idea of decision theory is that an agent is rational if and only if it choses the action that yields the highest expected utility, averaged over all the possible outcomes of the actions. This is called the principle of **maximum expected utility** (MEU)*

Basic Probability Notation

- Possible Worlds – like logical assertions, <u>probabilistic assertions</u> are about possible worlds!
- *How probable the worlds are???* → The set of possible worlds is called **the sample space Ω**.
- The possible worlds are *mutually exclusive* and *exhaustive*.

  2 possible worlds cannot both be the case!

  one possible world must be the case!!

- **Example**: if you roll the dice – there are 36 possible worlds! Each denoted as ω
- A fully specified **probability model** associates a numerical probability $P(\omega)$ with each possible world! Then the basic axioms of probability say:

$0 \leq P(\omega) \leq 1$ for every ω and

$\sum_{\omega \in \Omega} p(\omega) = 1$

What probabilities are about!

- Probabilistic assertions and queries are not about possible worlds, but about sets of them!

Example: we might be interested in the cases where two dice add up to 11! Or cases where doubles are rolled, etc. These sets are called **events**.

All events are described by **propositions** in a formal language (e.g. FOL). For each proposition, the corresponding events correspond to possible worlds where the propositions are TRUE!

The probability associated with a proposition is defined by the sum of the probabilities of the worlds in which it holds:

For every proposition ϕ *we have* $P(\phi) = \sum_{\omega \epsilon \phi} P(\omega)$

Example: when rolling fair dice, we have P(Total=11)=P((5,6))+P((6,5)) = 1/36+1/36=1/18

Axioms of Probability

1. $0 \leq P(\omega) \leq 1$ for every ω
2. $\sum_{\omega \in \Omega} p(\omega) = 1$
3. $P(\mathit{true}) = 1, \quad P(\mathit{false}) = 0$
4. $P(a \lor b)$ 0= $P(a) + P(b) - P(a \land b)$
5. We can deduce $P(\neg a) = 1 - P(a)$ because
   $P(\mathit{true}) = P(a) + P(\neg a) - P(\mathit{false})$

$a \lor \neg a$

$= 0$

$= 1$

$\Rightarrow \quad P(a) + P(\neg a) = 1$

Prior probability

- **Prior probability** (or unconditional probability) associated with a proposition *a* is the degree of belief accorded to it in the absence of any other information:

  Example: p(*cavity* = *true*) = 0.1

Important: P(a) can be used only when there is no other information. As soon as some new information is known, we must reason with the conditional probability of a, given new information!

- *Most of the time we have some additional evidence!*

Example: If I am going to the dentist for a regular checkup, the probability p(*cavity*) = 0.2 might be of interest, but it I go the dentist because I have a toothache, then what matters is the conditional probability: P(*cavity*/*toothache*)=0.6

However, p(*cavity*) = 0.2 is still valid after toothache is observed, it is just not that useful anymore!

Prior probability -2

- ***Interpretation:*** P(*cavity*|*toothache*)=0.6

does NOT mean "whenever toothache is true, conclude that cavity is true with P=0.6", but rather:

"whenever toothache is true, and we have no further information, conclude that cavity is true with P=0.6"

Conditional Probability

- *When new evidence concerning a previously unknown random variable is found, prior probabilities no longer apply.*
- We use conditional probabilities:
  - If *a – a proposition*
  - and *b* – a *proposition*

P(*a*|*b*) denotes "the probability of *a*, given that all we know is *b*"

- Example: P(*cavity*|*toothache*) = 0.8
- How do we compute P(*a*|*b*)?

$$P(a \mid b) = \frac{P(a \wedge b)}{P(b)}$$

- Same rule: P(*a*∧*b*) = P(*a*|*b*)P(*b*) – the product rule!

Propositions and probability assertions

- **Random Variables** – are the variables which allow possible worlds to be represented
- *Random variables have domains* → values they may take. Depending on the domain, random variables may be classified as
  - Boolean random variables → have the domain <true, false>

    Example: Cavity = true Cavity = false (¬cavity)
  - Discrete random variables → take values from a countable domain
  - Continuous random variables → values from the real numbers

    Example: the proposition $x=4.02$ asserts that the random variable x has the exact value 4.02. We can also have propositions that use inequalities like $x \leq 4.20$

Prior probability of random variables

- *Sometimes we are interested in the prior probabilities of all possible values of a random variable → use expressions such as P(weather) which denotes <u>a vector of values for the probabilities</u> of each individual state of the weather*

Example: P(*weather* = *sunny*) = 0.7

P(*weather* = *rain*) = 0.1

P(*weather* = *cloudy*) = 0.08

P(*weather* = *snow*) = 0.02

also written as:

P(weather) = <0.7, 0.2, 0.08, 0.02>

*The probability **<u>distribution</u>** of the random variable "weather"*

Probability distributions

Example 1:

Hair_color is a discrete random variable.

It has the domain <blond, brown, red, black, white, none>

Out of a sample of 10000 people, we find that 1872 had blond hair, 4325 had brown hair, 2135 had black hair, 652 had red hair, 321 had white hair and the rest were bald

The *probability distribution* is:

P(*Hair_color*) = <0.1872, 0.4325, 0.0652, 0.2135, 0.0321, 0.0695>

Example 2:

UTD_student is a binary random variable → the domain is True or False

Out of a sample of 1000 youngsters between 19 and 21 in the Dallas area, 321 were students at UTD

P(*UTD_student*) = <0.321, 0.679>

Probability Density Functions

- *For continuous random variables, it is not possible to write out the entire distribution as a table, because there are infinitely many values.* Instead, we define the probability that a random variable takes on some value x as a parameterized function of x.

- Example: Let the random variable X denote tomorrow's maximum rain fall in Dallas. The sentence **P(*X*=*x*) = U[1,3](*x*)** expresses the belief that X is distributed uniformly between 1in and 3in

Conditional Distributions

- If two random variables X and Y define some world, $\mathbf{P}(X|Y)$ gives the values for $P(X=x_i|Y=y_j)$ for each possible i and j.
- Expressed with the product rule, this becomes:

  $P(X=x_1 \wedge Y=y_1) = P(X=x_1|Y=y_1)P(Y=y_1)$

  $P(X=x_1 \wedge Y=y_2) = P(X=x_1|Y=y_2)P(Y=y_2)$

  $\vdots$
- This can be combined in a single equation:

  $\mathbf{P}(X,Y) = \mathbf{P}(X|Y)\mathbf{P}(X)$
- *This denotes a set of equations relating the corresponding individual entries in the tables, not a matrix multiplication of the tables!*

Atomic Events

Example: 2 random variables: Cavity and Toothache

How many atomic events? 4

$$E_1 = (cavity = false) \wedge (toothache = false)$$
$$E_2 = (cavity = false) \wedge (toothache = true)$$
$$E_3 = (cavity = true) \wedge (toothache = false)$$
$$E_4 = (cavity = true) \wedge (toothache = true)$$

Properties:

a) Mutually exclusive: at most one can be true

b) The set of all possible atomic events is exhaustive

What happens if we observe a new random variable: Weather with 4 possible values: {sunny, rain, cloudy, snow}???

We are going to have 16 atomic events!

Full Joint Probability Distribution

Suppose the world consists only of the variables

Cavity	Toothache	Weather		2x2x4=16
↓ 2 values (binary random variable)	↓ 2 values (binary random variable)	↓ 4 values (sunny, rain cloudy, snow)	⇒	entries for the joint distribution

P(Cavity) = <0.23,0.77>
P(Toothache) = <0.35,0.65>
P(Weather) = <0.7,0.2,0.008,0.02>

Represent the joint probability distribution:

	Cavity		**¬Cavity**	
Weather	**Toothache**	**¬Toothache**	**Toothache**	**¬Toothache**
Sunny	?	?	?	?
Rain	?	?	?	?
Cloudy	?	?	?	?
Snow	?	?	?	?

We need to know to compute *conditional probabilities*

Inference Using Full Joint Distribution

<u>Example</u>: The domains consist of three boolean variables: Toothache, Cavity and Catch (the dentist's nasty steel probe catches in my tooth)

	Toothache		**¬Toothache**	
	catch	**¬catch**	**catch**	**¬catch**
Cavity	0.108	0.012	0.072	0.008
¬Cavity	0.016	0.064	0.144	0.576

How may atomic events? 2^3 = 8 (as many entries as in the table!)

e_1	(cavity=false)∧(toothache=false)∧(catch=false)	$P(e_1)=0.576$
e_2	(cavity=false)∧(toothache=false)∧(catch=true)	$P(e_2)=0.144$
e_3	(cavity=false)∧(toothache=true)∧(catch=false)	$P(e_3)=0.064$
e_4	(cavity=false)∧(toothache=true)∧(catch=true)	$P(e_4)=0.016$
e_5	(cavity=true)∧(toothache=false)∧(catch=false)	$P(e_5)=0.008$
e_6	(cavity=true)∧(toothache=false)∧(catch=true)	$P(e_6)=0.072$
e_7	(cavity=true)∧(toothache=true)∧(catch=false)	$P(e_7)=0.012$
e_8	(cavity=true)∧(toothache=true)∧(catch=true)	$P(e_8)=0.108$

Inferring Probabilities

Given any proposition *a*, we can derive its probability as the sum of the probabilities of the atomic events in which it holds

$$P(a) = \sum_{e_i \in e(a)} P(e_i)$$

Example 1: *a* = *cavity* $\vee$ *toothache*

	Toothache		¬Toothache	
	catch	¬catch	catch	¬catch
Cavity	0.108	0.012	0.072	0.008
¬Cavity	0.016	0.064	0.144	0.576

a = cavity + toothache

Six events: e_5, e_6, e_7, e_8, e_3, e_4

P(cavity) = 0.108+0.012+0.072+0.008=0.2

e_1	(cavity=false)$\wedge$(toothache=false)$\wedge$(catch=false)	$P(e_1)=0.576$
e_2	(cavity=false)$\wedge$(toothache=false)$\wedge$(catch=true)	$P(e_2)=0.144$
e_3	(cavity=false)$\wedge$(toothache=true)$\wedge$(catch=false)	$P(e_3)=0.064$
e_4	(cavity=false)$\wedge$(toothache=true)$\wedge$(catch=true)	$P(e_4)=0.016$
e_5	(cavity=true)$\wedge$(toothache=false)$\wedge$(catch=false)	$P(e_5)=0.008$
e_6	(cavity=true)$\wedge$(toothache=false)$\wedge$(catch=true)	$P(e_6)=0.072$
e_7	(cavity=true)$\wedge$(toothache=true)$\wedge$(catch=false)	$P(e_7)=0.012$
e_8	(cavity=true)$\wedge$(toothache=true)$\wedge$(catch=true)	$P(e_8)=0.108$

When adding all probabilities in a row we obtain the <u>unconditional</u> or <u>marginal</u> probability

Marginalization and Conditioning Rules

Given any two random variables *Y* and *Z*,

$$P(Y) = \sum_z P(Y, Z)$$ ← marginalization rule

- A variant involves conditional probabilities, instead of joint probabilities:

conditioning rule →

$$P(Y) = \sum_z P(Y \mid Z) P(Z)$$

P(*cavity*) = 0.2   P(¬*cavity*)=0.8

P(*toothache*)=0.108+0.012+0.016+0.06=0.2

P(¬*toothache*)=0.8

P(*catch*)=0.108+0.16+0.072+0.14=0.484

P(¬*catch*)=0.516

	Toothache		**¬Toothache**	
	catch	¬catch	catch	¬catch
Cavity	0.108	0.012	0.072	0.008
¬Cavity	0.016	0.064	0.144	0.576

Marginalization and Conditioning Rules

If we have three random variables *X*, *Y* and *Z*

$$\Rightarrow P(X) = \sum_y \sum_z P(X, y, z) \longleftarrow$$ marginalization rule

$$P(X) = \sum_y \sum_z \underbrace{P(X \mid y, z) P(y \mid z) P(z)}_{\text{Why?}}$$

- From the product rule:

$$P(a \mid bc) = \frac{P(a,b,c)}{P(b,c)} = \frac{P(a,b,c)}{P(b \mid c) P(c)}$$

$$\Rightarrow P(a,b,c) = P(a \mid bc) P(b \mid c) P(c)$$

Conditional Probabilities

$$P(cavity \mid toothache) = \frac{P(cavity \wedge toothache)}{P(toothache)}$$

$$= \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} = 0.6$$

	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

$$\text{Also}: P(\neg cavity \mid toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)}$$

$$= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

- Notice: P(*toothache*) remains the same in both calculations $\Rightarrow$ it acts like a normalization constant for the distribution P(*Cavity*|*toothache*)

Normalization Constants

normalization constant

$$\mathbf{P}(Cavity \mid toothache) = \alpha \mathbf{P}(Cavity, toothache)$$

$$= \alpha[\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)]$$

$$= \alpha[<0.108, 0.016> + <0.012, 0.064>] = \alpha[<0.12, 0.08>]$$

	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

$$= <0.6, 0.4>$$

$$\alpha(0.12 + 0.08) = 1$$

$$\Rightarrow 0.2 = 1/\alpha \Rightarrow \alpha = 1/0.2 = 5$$

$$\Rightarrow \alpha \times 0.12 = 5 \times 0.12 = 0.6$$

$$\alpha \times 0.08 = 5 \times 0.08 = 0.4$$

Normalization Constants are useful shortcuts in many probability computations!

General inference Procedure

A notation:

- X is a query variable (Cavity in the example)
- E is the set of evidence variables (Toothache in the example)
- e are the observed values for the evidence

- The query: P(X|e)
- Evaluated as

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

Independence

	toothache		*¬toothache*	
	catch	*¬catch*	*catch*	*¬catch*
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

Let us add a fourth variable → *Weather*

⇒ The full distribution becomes *P(Toothache, Catch, Cavity, Weather*) which has 32 entries (8 before ×4 values for *Weather*)

This table contains four "editions" of the full table, one for each kind of weather.

- *What relations do these editions have to each other and to the original 3-variable table?*
- How are *P(toothache, catch, cavity, Weather=cloudy*) and *P(toothache, catch, cavity*) related?

- Use the product rule: *P(toothache, catch, cavity, Weather=cloudy*) =

= *P(Weather=cloudy|toothache,catch,cavity*) *P(toothache,catch,cavity*)

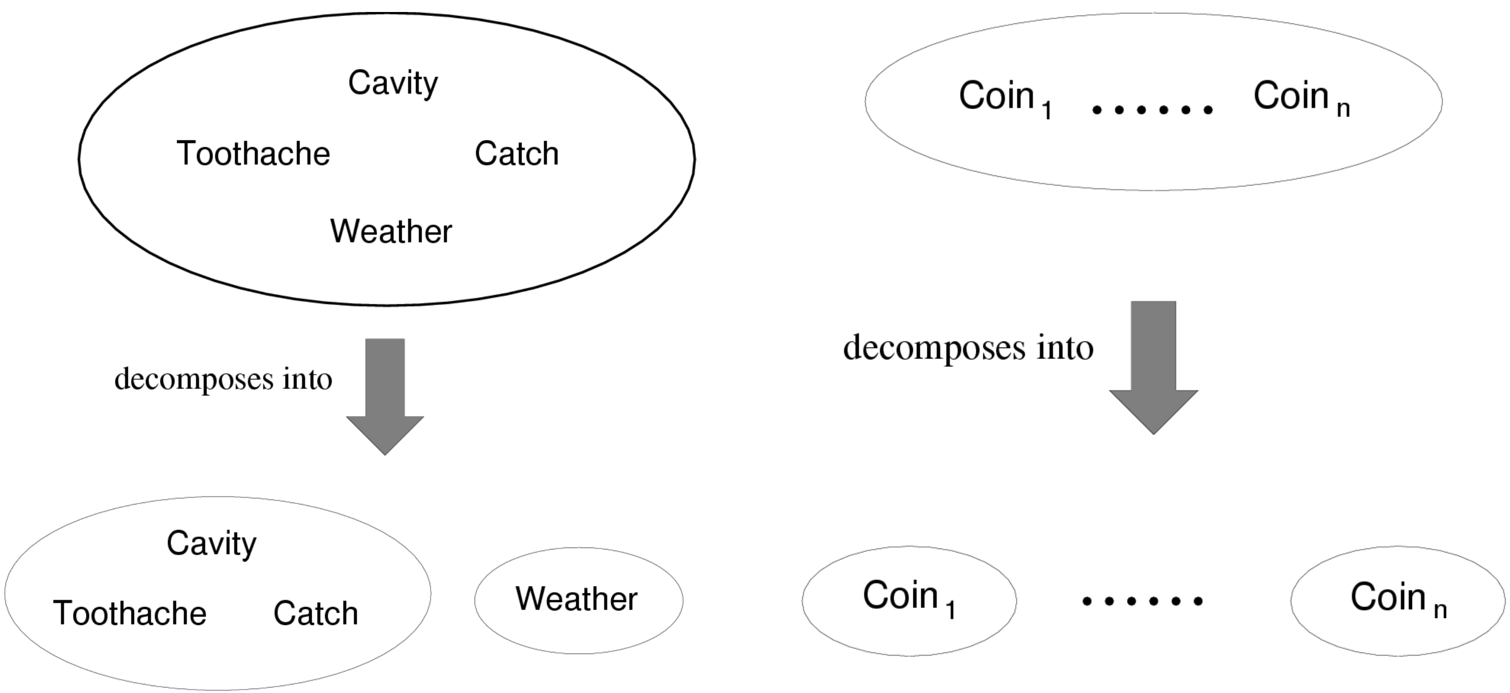
- But *P(Weather=cloudy|toothache,catch,cavity*)= *P(Weather=cloudy)*

Absolute Independence

- *Weather* is independent of one's dental problems.

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

The 32-element table can be constructed from one 8-element table and one 4-element table.



Independence in Equations

If propositions *a* and *b* are independent

$$P(a \wedge b) = P(a)\,P(b)$$

$$P(a|b) = P(a)$$

$$\left(P(a\,|\,b) = \frac{P(a)P(b)}{P(b)} \right)$$

- *Independence between variables X and Y is written:*

$$P(X, Y) = P(X)\,P(Y)$$

$$P(X \mid Y) = P(X)$$

$$P(Y \mid X) = P(Y)$$

Bayes Rule

From the product rule: P($a \wedge b$) = P(a/b) P(b)

= P(b/a) P(a)

- Then: P(a/b) P(b) = P(b/a) P(a)

$$P(b \mid a) = \frac{P(a \mid b) P(b)}{P(a)}$$ ← **Bayes' Rule**

- For multi-valued variables, Bayes' Rule is:

$$P(Y \mid X) = \frac{P(X \mid Y) P(Y)}{P(X)}$$

a set of equations, each dealing with specific values of the variables.

Examples

$$P(Cavity \mid Toothache) = \frac{P(Toothache \mid Cavity)P(Cavity)}{P(Toothache)}$$

$$P(cavity \mid toothache) = \frac{P(toothache \mid cavity)P(cavity)}{P(toothache)}$$

$$P(\neg cavity \mid toothache) = \frac{P(toothache \mid \neg cavity)P(\neg cavity)}{P(toothache)}$$

$$P(cavity \mid \neg toothache) = \frac{P(\neg toothache \mid cavity)P(cavity)}{P(\neg toothache)}$$

$$P(\neg cavity \mid \neg toothache) = \frac{P(\neg toothache \mid \neg cavity)P(\neg cavity)}{P(\neg toothache)}$$

Applying Bayes' Rule: The Simple Case

- **Example**: Medical diagnosis

A doctor knows that meningitis causes a stiff neck 50% of the time: P(*s*|*m*)=0.5

The doctor also knows some unconditional facts:

- the prior probability that the patient has meningitis is 1/50,000 $\Rightarrow$ P(*m*)=1/50,000
- the prior probability that any patient has a stiff neck is 1/20 $\Rightarrow$ P(*s*)=1/20

$$P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)} = \frac{0.5 \times 1/50{,}000}{1/20} = 0.0002$$

Only 1 in 5,000 patients with a stiff neck is expected to have meningitis

Another way

$$P(m|s) = 0.0002 = \frac{P(s|m)P(m)}{P(s)} = \frac{0.5 \times 1/50{,}000}{1/20}$$

small because
P(*m*)=1/50,000 <<
P(*s*)=1/20

- We can still compute P(*m*|*s*) without knowing P(*s*)
  - instead compute the posterior probability for each value of the query variable (here *m* and ¬*m*) and normalizing the results:

$$P(s)=P(s|m)P(m) + P(s|\neg m)P(\neg m)$$

Then:

$$P(m|s) = \frac{P(s|m)P(m)}{P(s|m)P(m) + P(s|\neg m)P(\neg m)} = \frac{1}{1 + \frac{P(s|\neg m)P(\neg m)}{P(s|m)P(m)}}$$

Similarly:

$$P(\neg m|s) = \frac{1}{1 + \frac{P(s|m)P(m)}{P(s|\neg m)P(\neg m)}}$$

$$P(m|s) = \alpha < P(s|m)P(m), P(s|\neg m)P(\neg m)>$$

this can be obtained also from applying Bayes' Rule with normalization

Bayes' Rule with Normalization

$$P(m|s) = \alpha < P(s|m)P(m), (P(s|\neg m)P(\neg m)>$$

this can be obtained also from applying Bayes' Rule with normalization

$$P(Y|X) = \alpha\, P(X|Y)P(Y)$$

α is a normalization constant needed to make the entries in $P(Y|X)$ sum to 1

Example of Bayes' Rule with Normalization

We have two discrete random variables:

- X describing weather conditions, with the domain: X={*sunny, rain, cloudy, snow*}
- Y describing clothes: Y={*t-shirt, long-sleeves, coat*}

The distributions of X and Y are:

- X = <0.5, 0.2, 0.29, 0.01>
- Y = <0.5, 0.3, 0.2>

We also have values of joint probabilities:

P(*t-shirt,sunny*)=0.32
P(*t-shirt,rain*)=0.08
P(*t-shirt,cloudy*)=0.09
P(*t-shirt,snow*)=0.01

P(*long-sleeves,sunny*)=0.01
P(*long-sleeves,rain*)=0.15
P(*long-sleeves,cloudy*)=0.05
P(*long-sleeves,snow*)=0.09

P(*coat,sunny*)=0.001
P(*coat,rain*)=0.03
P(*coat,cloudy*)=0.019
P(*coat,snow*)=0.15

Example of Bayes' Rule with Normalization

- *From these calculate conditional probabilities P(Y|X), that is probability of clothing given weather.*

P(t-shirt|sunny)=P(t-shirt, sunny)/P(sunny)=0.32/ 0.331 = .967

P(long-sleeves|sunny)= P(long-sleeves, sunny)/P(sunny)= 0.01/0.331= 0.0302

P(coat|sunny)= P(coat, sunny)/P(sunny)= 0.001/0.331= 0.003

P(t-shirt|rain)= P(t-shirt, rain)/P(rain)= 0.08/0.26=0.307

P(long-sleeves|rain)= P(long-sleeves, rain)/P(rain)= 0.15/0.26=0.577

P(coat|rain)= P(coat, rain)/P(rain)= 0.03/0.26=0.1154

P(t-shirt|cloudy)= P(t-shirt, cloudy)/P(cloudy)= 0.09/0.159=0.566

P(long-sleeves|cloudy)= P(long-sleeves, cloudy)/P(cloudy)= 0.05/0.159=0.314

P(coat|cloudy)=P(coat, cloudy)/P(cloudy) = 0.019/0.159=0.1195

P(t-shirt|snow)= P(t-shirt, snow)/P(snow)= 0.01/0.25=0.04

P(long-sleeves|snow)= P(long-sleeves, snow)/P(snow)= 0.09/0.25=0.36

P(coat|snow)= P(coat, snow)/P(snow)= 0.15/0.25=0.6

Computing the Normalization Constant

Bayes's Rule: → trying to guess the weather from the clothes people wear: $P(X|Y) = \alpha\, P(Y|X)P(X) = \alpha\, P(X,Y)$

P(X|Y) add up to 1 for each value of Y

1. $P(sunny|t\text{-}shirt)=\alpha_1 P(t\text{-}shirt, sunny)=\alpha_1\times 0.32= 2\times 0.32=0.64$
2. $P(rain|t\text{-}shirt)=\alpha_1 P(t\text{-}shirt, rain)=\alpha_1\times 0.08=2\times 0.08=0.16$
3. $P(cloudy|t\text{-}shirt)=\alpha_1 P(t\text{-}shirt, cloudy)=\alpha_1\times 0.09=2\times 0.09=0.18$
4. $P(snow|t\text{-}shirt)=\alpha_1 P(t\text{-}shirt, snow)=\alpha_1\times 0.01=2\times 0.01=0.02$

$\alpha_1(0.32+0.08+0.09+0.01) = 1$ $\alpha_1=1/0.5 = 2$

1. $P(sunny|long\text{-}sleeves)=\alpha_2 P(long\text{-}sleeves, sunny)=\alpha_2\times 0.01=3.33\times 0.01=0.033$
2. $P(rain|long\text{-}sleeves)=\alpha_2 P(long\text{-}sleeves, rain)=\alpha_2\times 0.15=3.33\times 0.15=0.5$
3. $P(cloudy|long\text{-}sleeves)=\alpha_2 P(long\text{-}sleeves, cloudy)=\alpha_2\times 0.05=3.33\times 0.05=0.167$
4. $P(snow|long\text{-}sleeves)=\alpha_2 P(long\text{-}sleeves, snow)P=\alpha_2\times 0.09=3.33\times 0.09=0.3$

$\alpha_2(0.01+0.15+0.05+0.09) = 1$ $\alpha_2=1/0.3 = 3{,}333$

1. $P(sunny|coat)=\alpha_3 P(coat, sunny)=\alpha_3\times 0.001=5\times 0.001=0.005$
2. $P(rain|coat)=\alpha_3 P(coat, rain)=\alpha_3\times 0.03=5\times 0.03=0.15$
3. $P(cloudy|coat)=\alpha_3 P(coat, cloudy)=\alpha_3\times 0.019=5\times 0.019=0.095$
4. $P(snow|coat)=\alpha_3 P(coat, snow)=\alpha_3\times 0.15=5\times 0.15=0.75$

$\alpha_3(0.001+0.03+0.019+0.15) = 1$ $\alpha_3=1/0.2 = 5$

Example - Summary

X

Y	sunny	rain	cloudy	snow	
t-shirt	.32	.08	.09	.01	.5
long-sleeve	.01	.15	.05	.09	.3
coat	.001	.03	.019	.15	.2
	.331	.26	.159	.25	

Joint Probabilities

P (X, Y)

weather X

clothes Y

Y

X	t-shirt	long-sleeve	coat	
sunny	.967	.0302	.003	1
rain	.307	.577	.1154	1
cloudy	.566	.314	.1195	1
snow	.04	.36	.6	1

Conditional Probabilities

P (Y | X)

clothes Y

weather X

X

Y	sunny	rain	cloudy	snow	
t-shirt	.64	.16	.18	.02	1
long-sleeve	.033	.5	.167	.3	1
coat	.005	.15	.095	.75	1

Conditional Probabilities

P (X | Y)

More on Bayes' rule

- What happens if one has the conditional probability in one direction but not the other?
  - **Example**: meningitis domain

    The doctor knows that a stiff neck implies meningitis in 1 of 5000 cases → the doctor has quantitative information in the diagnostic direction from symptoms to causes. Lucky case → the doctor has no need to use Bayes's rule

Causal Knowledge

Note: Unfortunately, diagnostic knowledge is often more fragile than causal knowledge.

Why? → If there is a sudden epidemic of meningitis, the prior probability of meningitis P(m) will go up. Because of this, the doctor who designed the diagnostic probability P(m|S) directly from statistical information, will not know how to update P(m|s).

But $P(m|s) = \frac{P(s|m)P(m)}{P(s)}$ → P(m|s) should go up proportionally with P(m)

Important: P(s|m) is <u>unaffected</u> by epidemic because it reflects how meningitis works.

Conclusions: Using Causal or model-based knowledge provides robustness → feasible probabilistic reasoning

Combining Evidence

- Until now, we considered probabilistic information available in the form $P(effect|cause)$

single evidence

- What happens when there are multiple pieces of evidence?

Example: dentist domain

$P(Cavity \mid toothache \wedge catch) = \alpha\langle 0.108, 0.016\rangle \cong \langle 0.871, 0.129\rangle$

This will not scale up to larger numbers of variables

$P(Cavity \mid toothache \wedge catch) = \alpha P(toothache \wedge catch \mid Cavity)P(Cavity)$

→ we need to know the values of the conditional probabilities of the conjunction $toothache \wedge catch$ for all values of Cavity

If we have *n* possible evidence variables (X rays, diet, oral hygiene,…) these are 2^n possible combinations of observed values, and for each we need to know the conditional probabilities.

Solution

Consider the notion of independence

- Three variables: Cavity, Toothache, Catch →which are independent?
  - (Cavity, Catch)? → if the probe catches in the tooth, it probably has a cavity and that probably causes a toothache
  - (Toothache, Catch)? → if there is a cavity, there will be a toothache, regardless of the probe catching the tooth
  - (Toothache,Cavity)? → if there is a cavity, it might cause a toothache, but toothaches are not only caused by cavities

Conditional Independence

$$P(toothache \wedge catch|Cavity) = P(toothache|Cavity) \times P(catch|Cavity)$$

Conditional independence of toothache and catch given Cavity

We also know:

$$P(Cavity|toothache \wedge catch) = \alpha P(toothache \wedge catch|Cavity) \times P(Cavity)$$

then interpret:

$$P(Cavity|toothache \wedge catch) = \alpha P(toothache|Cavity) \times P(catch|Cavity) \times P(cavity)$$

$$P(\underbrace{Cavity}_{\text{cause}}|\underbrace{toothache}_{\text{effect1}} \wedge \underbrace{catch}_{\text{effect2}}) =$$

$$\alpha \underbrace{P(toothache|Cavity)}_{\text{effect1 given cause}} \times \underbrace{P(catch|Cavity)}_{\text{effect2 given cause}} \times \underbrace{P(Cavity)}_{\text{cause}}$$

Computational Complexity

General definition of conditional independence of two variables X and Y given a third variable Z

$$P(X,Y\mid Z) = P(X\mid Z)P(Y\mid Z)$$

Example:

$$P(Toothache, Catch\mid Cavity) = P(Toothache\mid Cavity)P(Catch\mid Cavity)$$

We can now derive the decomposition:

$$P(Toothache, Catch, Cavity) =$$

$$= P(Toothache, Catch\mid Cavity)P(Cavity)P(Catch\mid Cavity) \text{ (product rule)}$$

$$= \underbrace{P(Toothache\mid Cavity)}_{\text{Table 1}}\underbrace{P(Catch\mid Cavity)}_{\text{Table 2}}\underbrace{P(Cavity)}_{\text{Table 3}}$$

Initial table-size = 2^3-1 = 7 values (2^3 all, but since they sum to 1, we do not need the last one)

Table1-size = 2 x (2^2-1) – 1 = 5 values

Table2-size = Table1-size

Table3-size = 2^1-1 = 1 → the size of the representation grows as O(n) instead of O(2^n)

Separation

$$P(Cause, Effect_1, \ldots, Effect_n) = P(Cause)\prod_i P(Effect_i \mid Cause)$$

$$\begin{aligned} &P(Toothache, Catch, Cavity) = \\ &\quad = P(Toothache|Cavity)P(Catch|Cavity)P(Cavity) \end{aligned}$$

here Cavity separates Toothache and Catch because it is a Cause to both of them!

Naïve Bayes model

It is called Naïve because it works surprisingly well even when the effect variables are not conditionally independent

The Wumpus World Revisited

After finding a breeze in both [1,2] and [2,1], the agent is stuck because there is no safe place to explore

1,4	2,4	3,4	4,4
1,3	2,3	3,3	4,3
1,2 B OK	2,2	3,2	4,2
1,1 OK	2,1 B OK	3,1	4,1

Goal: compute the probability that the 3 neighboring squares contain a pit

$$P(Pit[1,3]) \times P(Pit[2,2]) \times P(Pit[3,1])$$

Information:

- a pit causes a breeze in all neighboring squares
- each square other than [1,1] contains a pit with probability 0.2

Step 1: Identify random variables

$P_{i,j}$=1 if square [i,j] contains a pit

[i,j] = only for observed squares, e.g [1,1],[1,2],[2,1]

B_{ij} = 1 if square [i,j] is breezy

P_{ij}, B_{ij} are boolean random variables

Probabilistic Reasoning for the Wumpus

Next step: specify the full distribution:

$$P(P_{11}, P_{12}, P_{13}, P_{14}, P_{21}, P_{22}, P_{23}, P_{24}, P_{31}, P_{32}, P_{33}, P_{34},$$
$$P_{41}, P_{42}, P_{43}, P_{44}, B_{11}, B_{12}, B_{21}) =$$
$$= P(B_{11}, B_{12}, B_{21} \mid P_{11}, \ldots, P_{44}) P(P_{11}, \ldots, P_{44})$$

The prior probability of a pit configuration:

$$P(P_{11}, \ldots, P_{44}) = \prod_{i,j=1,1}^{4,4} P(P_{ij})$$

If there are n pits,

$$P(P_{11}, \ldots, P_{44}) = (0.2)^n \times (0.8)^{16-n}$$

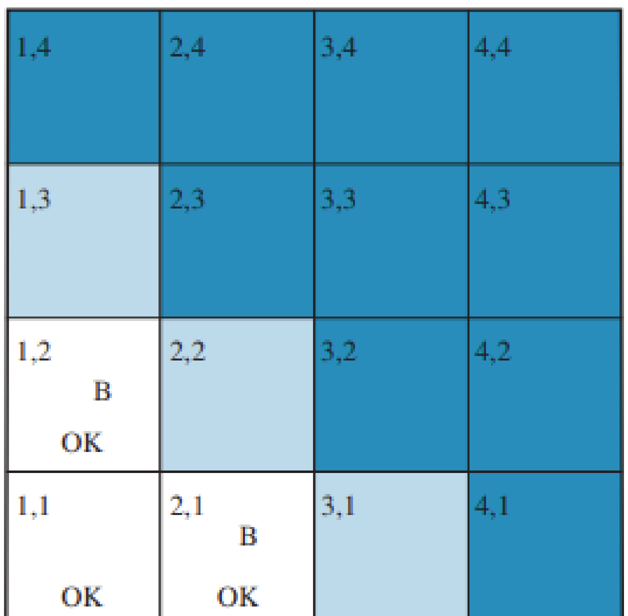
1,4	2,4	3,4	4,4
1,3	2,3	3,3	4,3
1,2 B OK	2,2	3,2	4,2
1,1 OK	2,1 B OK	3,1	4,1

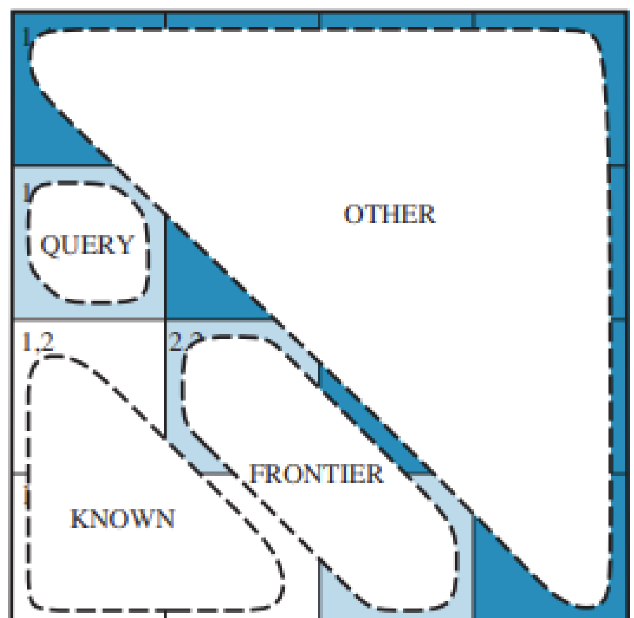
Combining Evidence

The evidence: - observed breeze in each square visited
+ each square visited contains no pit

$$b = \neg b_{11} \wedge b_{12} \wedge b_{21}$$
$$known = \neg p_{11} \wedge \neg p_{12} \wedge \neg p_{21}$$
$$Query: \quad P(P_{13} \mid known, b)$$

How likely is that [1,3] contains a pit, given the observation so far





Answering the Query

- To answer $P(P_{13}|known,b)$ we sum over the entries from the full distribution.
- Let *unknown* be a composite variable consisting of the P_{ij} variables for squares other than *known* and the query square [1,3]

➔ $$P(P_{13} \mid known, b) = \sum_{unknown} P(P_{13}, unknown, known, b)$$

How many squares?

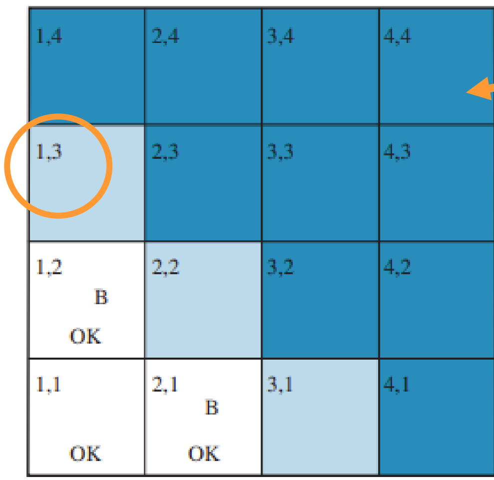
16-3-1=12 → the summation contains

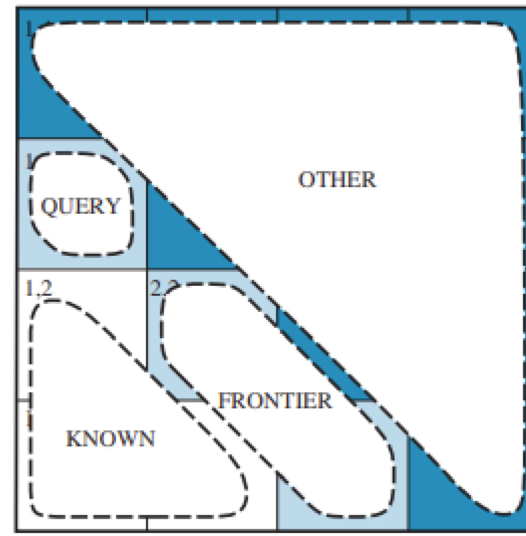
2^{12} = 4096 terms (too many!)

known query

Careful Computation-1

Why would the contents of [4,4] affect whether [1,3] has a pit?





- Let us consider the set Frontier made of variables that are adjacent to the visited squares. Frontier={[2,2],[3,1]}
- Let us also consider the set Other containing variables of the unknown squares (10 of them)
- The observed breezes are conditionally independent!

Careful Computation-2

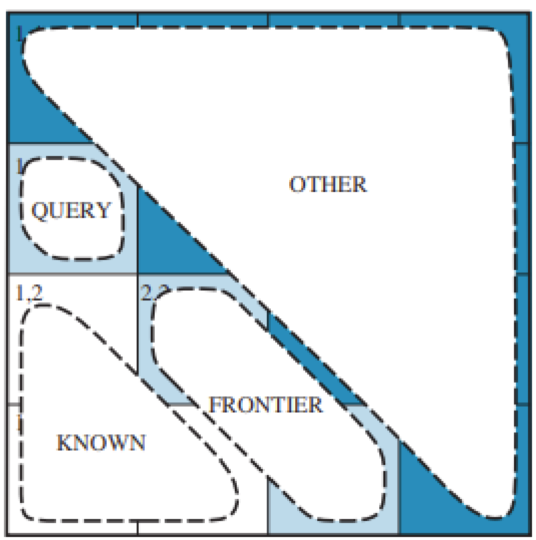
- The observed breezes are conditionally independent

$$P(P_{13}|known, b) = \alpha \sum_{unknown} P(b|P_{13}, unknown, known) \times P(P_{13}\,, known, unknown) =$$

$$= \alpha \sum_{frontier} \sum_{other} P(b|known, P_{13}, frontier, other) \times P(P_{13}, known, frontier, other) =$$

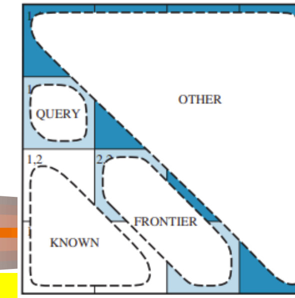
$$= \alpha \sum_{froniter} \sum_{other} P(b|known, P_{13}, frontier) \times P(P_{13}, frontier, other)$$

conditional independent



$$b = \neg b_{11} \wedge b_{12} \wedge b_{21}$$
$$known = \neg p_{11} \wedge \neg p_{12} \wedge \neg p_{21}$$

Continue Computation



$$\alpha \sum_{frontier} \sum_{other} \underbrace{P(b|known, P_{13}, frontier)} \times P(P_{13}, frontier, other)$$

the first term in this expression does not depend on the other variables → we can move the summation inwards

$$= \alpha \sum_{frontier} P(b|known, P_{13}, frontier) \sum_{other} \underbrace{P(P_{13}, frontier, other)}$$

The prior term can be factored

$$P(P_{13}|known, b)$$

$$= \alpha \sum_{frontier} P(b|known, P_{13}, fringe) \times \sum_{other} P(P_{13}) \times P(known) \times P(frontier \times P(other)$$

$$= \alpha P(known) P(P_{13}) \sum_{frontier} P(b|known, P_{13}, frontier) \times P(frontier) \times \sum_{other} P(other)$$

$$= \alpha' P(P_{13}) \sum_{frontier} P(b|known, P_{13}, frontier) \times P(frontier)$$

Finishing

$$P(P_{13}|known, b) = \alpha' P(P_{13}) \sum_{frontier} P(b|known, P_{13}, frontier) \times P(frontier)$$

$$\boldsymbol{frontier = \{[2,2],[3,1]\}}$$

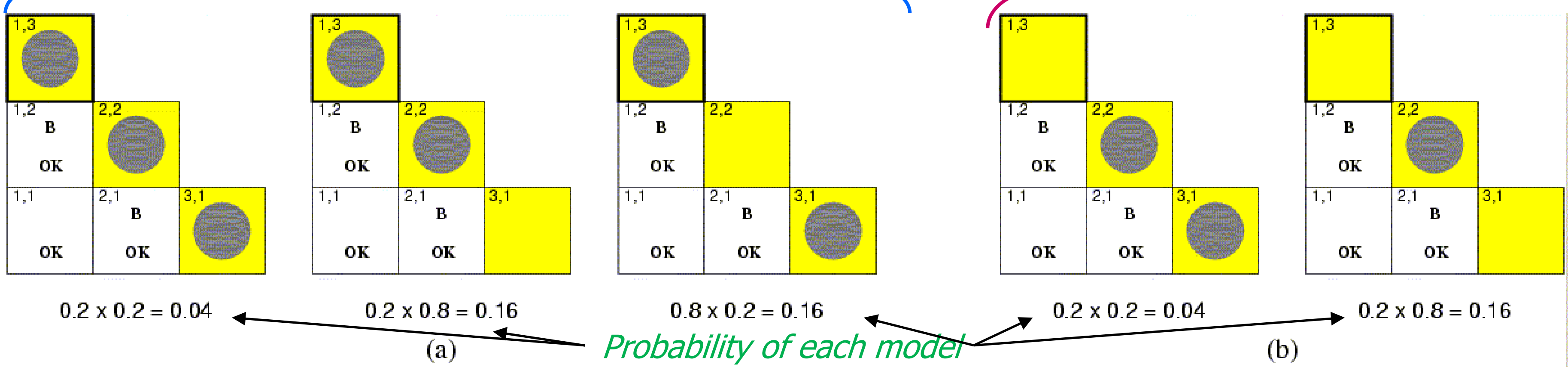
How do we build models for the Frontier?

- Since B_{12} and B_{21} we may have a pit in P_{13},P_{22} or P_{31}

	P_{13}	P_{22}	P_{31}	
(a)	1	1	1	**Model1**
	1	1	0	**Model2**
	1	0	1	**Model3**
(b)	0	1	1	**Model4**
	0	1	0	**Model5**

Three models with $P_{1,3}$=True

Two models with $P_{1,3}$=False



Probability of each model

Likelihood of Pit at [1,3]

$$P(P_{13}|known, b) = \alpha' P(P_{13}) \sum_{frontier} P(b|known, P_{13}, frontier) \times P(frontier)$$

Use *Models 1,2 & 3* *for* P_{13}*=True*

Model 4,5 *for* P_{13}*=False*

Models

P_{13}	P_{22}	P_{31}	
1	1	1	**Model1**
1	1	0	**Model2**
1	0	1	**Model3**
0	1	1	**Model4**
0	1	0	**Model5**

$\rightarrow$ $P(P_{13}|known,b) = \alpha' <0.2 \times (0.04 + 0.16 + 0.16), 0.8 \times (0.04+0.16)>$

$P(P_{13}=1)$ P(Model1) P(Model2) P(Model3) P(Model4) P(Model5)

$= \alpha' <0.2 \times 0.36, 0.8 \times 0.2>$

$= \alpha' <0.072, 0.016> \approx <0.3103, 0.6897>$

Because $\alpha' \times 0.072 + \alpha' \times 0.16 = 1 \rightarrow \alpha' = 1/0.232 = 4.3103448$

Interpretation

From $P(P_{13}|known,b) = <0.3103, 0.6897>$ *we know that* [1,3] (and [3,1] by symmetry) *contains a pit roughly 31% probability*.

- *Similarly, $P(P_{22}|known,b)$ contains a pit roughly 86% probability*
- → The wumpus should avoid [2,2]!

<u>Lessons</u>:

- seemingly complicated problems can be formulated precisely in probability theory and solved using simple algorithms
- Efficient solutions are obtained when independence and conditional independence relationships are used to simplify the summations
- Independence corresponds to our natural understanding of how the problem should be decomposed