Machine Learning

11 – Modeling Uncertainty

SS 2018

Gunther Heidemann

Overview

Part I: Uncertainty and probability

- Random variables
- Joint distribution
- Inference
- Independence and conditional independence
- Bayes rule

Part II: Bayes networks

Part I: Uncertainty and Probability

Modeling uncertainty

Environments may be uncertain due to several causes:

- Environment is only partially observable.
- Sensors are unreliable.
- The results of actions are uncertain.
- High complexity.

We deal with uncertainty using probabilities of propositions.

Alternative: Model uncertainty using probabilities of rules. such as

$$\mathit{LawnSprinkler} \;\; |\!\rightarrow_{0.99} \;\; \mathit{WetGras},$$

$$\mathit{WetGras} \;\; |\!\rightarrow_{0.7} \;\; \mathit{Rain}.$$

Modeling uncertainty

Probabilities summarize several factors:

- Missing knowledge,
- Incapability to devise complete models of complex domains,
- Chance.

Random variables

- Modeling uncertainty using random variables.
- Types of random variables:
  - *Boolean* random variable:
    - E.g. *Cavity* (Is there a cavity in my tooth?)
    - Values: <*true, false*>
  - *Discrete* random variable:
    - E.g. *Weather* has one of the values
      <*sun, rain, cloudy, snow*>
    - Values must describe the domain sufficiently and be mutually exclusive.
  - *Continuous* random variable:
    - Values are real numbers.
    - E.g. *Length* $\in [1, 20]$.

Propositions

- A *proposition* is made by assigning a value to a random variable:
  - *Weather* = *sun*
  - *Length* = 2,4
- Complex propositions are made by using logical operators to connect simple propositions:

$$Weather = sun \vee Cavity = false$$

- Notation:
  - Random variables with capital: *Weather*, but values: *sun*.
  - But:

cavity	means	*Cavity*	*= true,*
$\neg$ *cavity*	means	*Cavity*	*= false,*
sun	means	*Weather = sun.*	

Atomic events

- Atomic event:

  A *complete* specification of the state of the domain (the agent may be uncertain about the state).

- Example: Domain is fully described by the boolean variables *Cavity* and *Toothache.*

  Then there are 4 atomic events:

  $Cavity = false \quad \wedge \quad Toothache = false$

  $Cavity = false \quad \wedge \quad Toothache = true$

  $Cavity = true \quad \wedge \quad Toothache = false$

  $Cavity = true \quad \wedge \quad Toothache = true$

- Atomic events are mutually exclusive and describe the domain completely.

Probability distribution

- A-priori or unconditional probabilities of propositions:

  P(*Cavity* = *true*) = 0.1 or P(*Weather* = *sun*) = 0.72 denote the probability of guesses. The probabilities may change when new information becomes available.

- The probability distribution **P** comprises the probabilities of all values:

  P(*Weather*) = <0.72, 0.1, 0.08, 0.1>

  for values <*sun, rain, cloudy, snow*>.

- **P** is *normalized*, i.e., sum = 1.

Probability distribution

- Joint probability distribution for several random variables comprises all atomic states:

  $\mathbf{P}(\mathit{Weather}, \mathit{Cavity})$ is a 4 × 2 matrix:

Weather	=		*sun*	*rain*	*cloudy*	*snow*
Cavity	=	true	0.144	0.02	0.016	0.02
Cavity	=	*false*	0.576	0.08	0.064	0.08

- The joint probability distribution holds the entire knowledge about the domain!

Conditional probability

- Conditional or posterior probability:

  E.g. P(*cavity* | *toothache*) = 0.8

  i.e., the information *toothache* is known (but no more).

- Notation for conditional distributions:

  P(*Cavity* | *Toothache*) = 2-component vector of 2-comp. vectors

- If the additional information *cavity* is known, *Toothache* is irrelevant:

  P(*cavity* | *Toothache, cavity*) = <1,1>.

- *sun* is irrelevant given *toothache*:

  P(*cavity* | *toothache, sun*) = P(*cavity* | *toothache*) = 0.8

  More general:

  P(*Cavity* | *Toothache, sun*) = **P**(*Cavity* | *Toothache*).

- Domain knowledge of this kind facilitates finding the joint probability distribution.

Conditional probability

- Definition of conditional probability:

  $$P(a \mid b) = P(a \wedge b) / P(b) \quad \text{if} \quad P(b) > 0.$$

- Product rule is an alternative formulation:

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

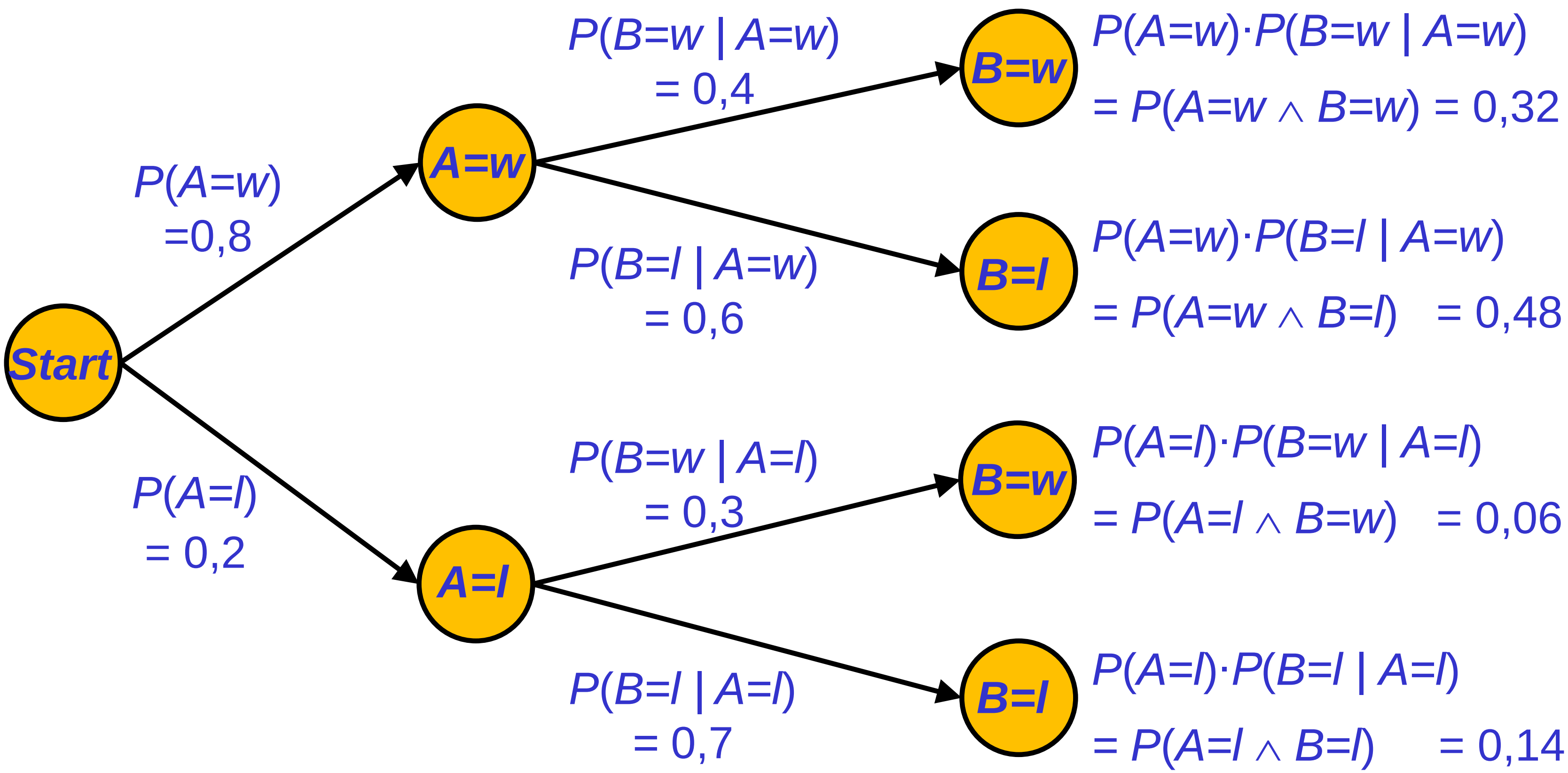
- General version for distributions:

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

  This means 4 × 2 separate equations, not matrix multiplication!

Conditional probability

Example: The german national soccer team plays first against Austria (A), winning chance $P(A{=}w) = 80\%$, then against Brazil (B). Experts say if $A{=}w$, they will be in a good mood and have a 40% chance to defeat Brazil, otherwise only 30%.



[H]

Conditional probability

Chain rule (derived by successive application of product rule):

$$\mathbf{P}(X_1, \ldots, X_n)$$
$$= \mathbf{P}(X_n \mid X_1, \ldots, X_{n-1})\, \mathbf{P}(X_1, \ldots, X_{n-1})$$
$$= \mathbf{P}(X_n \mid X_1, \ldots, X_{n-1})\, \mathbf{P}(X_{n-1} \mid X_1, \ldots, X_{n-2})\, \mathbf{P}(X_1, \ldots, X_{n-2})$$
$$= \ldots$$
$$= \prod_{i=2}^{n} \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1})\, \mathbf{P}(X_1).$$

Inference by enumeration

Joint probability distribution (*catch = dentist has found cavity*):

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

[RN]

- Probability of a proposition φ is the sum of the probabilities of the corresponding atomic events:

$$P(\varphi) = \Sigma_{\omega:\omega \models \varphi} P(\omega).$$

- P(*toothache*) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2
- P(*toothache* ∨ *cavity*) =

  0.108 + 0.012 + 0.016 + 0.064 + 0.072 + 0.008 = 0.28.

Inference by enumeration

Joint probability distribution (*catch = dentist has found cavity*):

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	$\boxed{.108}$	$\boxed{.012}$	.072	.008
¬ *cavity*	$\boxed{.016}$	$\boxed{.064}$	.144	.576

[RN]

Conditional probabilities:

P(¬*cavity* | *toothache*) =

P(¬*cavity* ∧ *toothache*) / P(*toothache*)* =

(0.016+0.064) / (0.108 + 0.012 + 0.016 + 0.064) = 0.4

* Since *toothache* is known, the right side of the table must be normalized to 1.

Inference by enumeration

Joint probability distribution (*catch = dentist has found cavity*):

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	.108	.012	.072	.008
$\neg$ *cavity*	.016	.064	.144	.576

[RN]

Denominator is a normalization constant:

$$\alpha = 1 / P(toothache)$$

$\mathbf{P}(Cavity \mid toothache)$

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

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

$= \alpha\, [\langle 0.108, 0.016 \rangle + \langle 0.012, 0.064 \rangle]$

$= \alpha\, \langle 0.12, 0.08 \rangle = \langle 0.6, 0.4 \rangle$

Inference by enumeration

Idea: Infer distribution of *query variables* (*Cavity*) depending on the *evidence variables* (*toothache*) and sum up over unobserved *hidden variables* (*Catch*).

In general:

For a set $\boldsymbol{X}$ of random variables we want to know

- the joint posterior distributions of the *query variables* $\boldsymbol{Y}$
- for given values $\boldsymbol{e}$ of the *evidence variables* $\boldsymbol{E}$.

The *hidden variables* are $\boldsymbol{H} = \boldsymbol{X} \setminus \{\boldsymbol{Y} \cup \boldsymbol{E}\}$,

and can be removed by “summing out“:

$$\mathbf{P}(\boldsymbol{Y} \mid \boldsymbol{E} = \boldsymbol{e}) = \alpha\, \mathbf{P}(\boldsymbol{Y}, \boldsymbol{E} = \boldsymbol{e}) = \alpha \sum_h \mathbf{P}(\boldsymbol{Y}, \boldsymbol{E}=\boldsymbol{e}, \boldsymbol{H}=\boldsymbol{h})$$

Inference by enumeration

Problem of inference by enumeration:

For n variables with a maximum of d values, the joint probability distribution table comprises $O(d^n)$ values:

- How to find these numbers?
- Memory requirements are $O(d^n)$.
- Time complexity is $O(d^n)$.

Domain knowledge about independence of variables may simplify the joint probability distribution significantly !

Independence

A and *B* are independent if and only if

$$\mathbf{P}(A \mid B) = \mathbf{P}(A) \quad \text{or} \quad \mathbf{P}(B \mid A) = \mathbf{P}(B).$$

With this, the product rule leads to

$$\mathbf{P}(A, B) = \mathbf{P}(A \mid B)\, \mathbf{P}(B) = \mathbf{P}(A)\, \mathbf{P}(B).$$

Independence

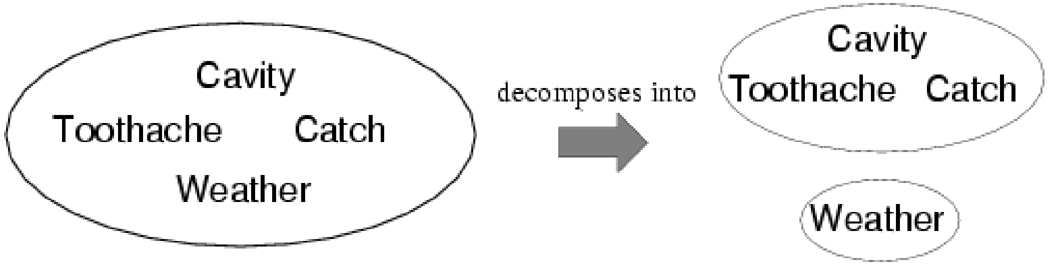
Example:

P(*Toothache, Cavity, Catch, Weather*)

= **P**(*Weather* | *Toothache,Cavity,Catch*) **P**(*Toothache,Cavity,Catch*)

= **P**(*Weather*) **P**(*Toothache,Cavity,Catch*),

since *Weather* is independent of *Toothache, Cavity, Catch*.



[RN]

So the 2 x 2 x 2 x 4 – 1 = 31 independent (sum = 1 !) numbers of the joint probability distribution are reduced to 4 + 2 x 2 x 2 – 1 = 11.

But: Perfect independence is rare (toothache might be influenced by the weather).

Conditional independence

Consider **P**(*Toothache, Cavity, Catch*), which has $2^3 - 1 = 7$ independent probabilities.

Mind *Catch* is not independent of *Toothache*:

In general,

$$\mathbf{P}(Catch \mid Toothache) \neq \mathbf{P}(Catch)\ !$$

Rather, **P**(*Catch*) does depend on the value of *Toothache* - it´s far more likely finding a cavity provided there is *toothache*.

But: This holds only as long as the value of *Cavity* is unknown.

Conditional independence

Assume that

1. for *Cavity = true* the probability that the dentist finds the cavity does not depend on whether there is toothache or not

$$\mathbf{P}(Catch \mid Toothache, \ cavity) = \mathbf{P}(Catch \mid cavity);$$

2. for *Cavity = false* the probability for a catch is independent of *Toothache* likewise:

$$\mathbf{P}(Catch \mid Toothache, \neg cavity) = \mathbf{P}(Catch \mid \neg cavity).$$

Summarizing 1 and 2, *Catch* is ***conditionally independent*** of *Toothache* **given the value of** *Cavity*:

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

Likewise, *Toothache* is conditionally independent of *Catch* given *Cavity*:

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

(e.g., the tooth hurts whether the dentist finds the cavity or not).

Conditional independence

From the conditional independence of *Catch* and *Toothache* for given *Cavity*

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

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

we get the factorization

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

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

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

Conditional independence

Full joint probability distribution derived using chain rule:

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

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

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

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

i.e., 2 + 2 + 1 = 5 independent numbers.

- In many cases, the memory required to represent a joint probability distribution of n variables can be reduced from "exponential in n" to "linear in n" using knowledge about conditional independence.
- Thus, conditional independence is an important and simple method of representing knowledge.

Bayes rule

- Product rule: $P(A, B) = P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A)$

  $\Rightarrow$ Bayes rule: $P(A \mid B) = P(B \mid A) \cdot P(A) / P(B)$

- The same for distributions:

  $$\mathbf{P}(Y \mid X) = \mathbf{P}(X \mid Y)\ \mathbf{P}(Y) / \mathbf{P}(X) = \alpha\, \mathbf{P}(X \mid Y)\ \mathbf{P}(Y).$$

- Useful for assessing *diagnostic* probability from *causal* probability :
  - $P(Cause \mid Effect) =$

    $$P(Effect \mid Cause)\ P(Cause) / P(Effect)$$

  - Example: Let M be meningitis, S stiff neck

    $$P(m \mid s) = P(s \mid m)\ P(m) / P(s) = 0.8 \times 0.0001 / 0.1 = 0.0008$$

    The posterior probability of meningitis is small even for a stiff neck, because the a-priori probability for meningitis is small while the a-priori probability for a stiff neck is much larger.

Bayes rule and conditional independence

Naive Bayes model:

Let denote C the cause, E_i its effects.

$$\mathbf{P}(C, E_1, \ldots E_n) = \mathbf{P}(E_1 \mid C, E_2 \ldots E_n)\, \mathbf{P}(C, E_2 \ldots E_n)$$

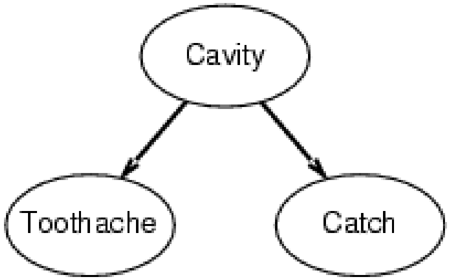
$$= \mathbf{P}(E_1 \mid C, E_2 \ldots E_n)\, \mathbf{P}(E_2 \mid C, E_3 \ldots E_n)\, \mathbf{P}(C, E_3 \ldots E_n)$$

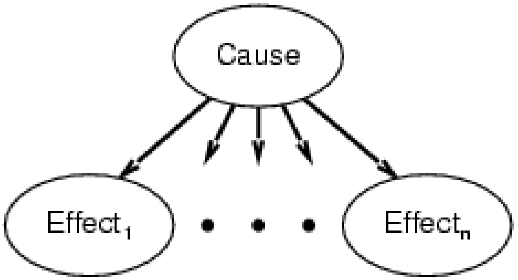
etc. $$= \prod_{i=1}^{n-1} \mathbf{P}(E_i \mid C, E_{i+1} \ldots E_n)\, \mathbf{P}(E_n \mid C)\, \mathbf{P}(C)$$

$$= \prod_{i=1}^{n} \mathbf{P}(E_i \mid C)\, \mathbf{P}(C)$$

since effect E_i is conditionally independent of the other effects E_j, $j \neq i$, for a given value of cause C.

The number of parameters is linear in n.





[RN]

Part II: Bayes networks

Bayes networks

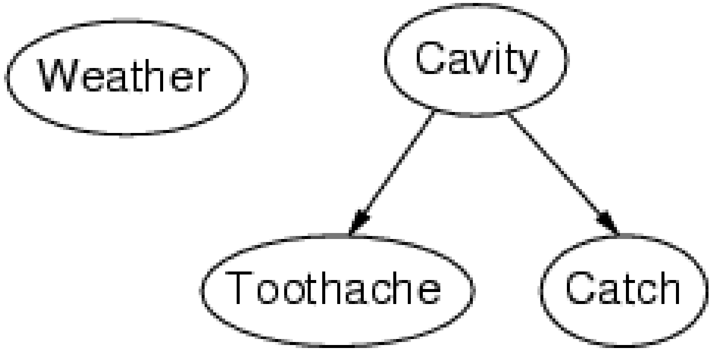
- Bayes networks are a form of graphical notation for propositions on conditional probabilities and thus a suitable way to express joint probability distributions.
- Syntax:
  - A set of nodes, one for each random variable.
  - A directed acyclic graph
    - Edge from *A* to *B* means: „*A* influences *B*“.
    - *A* is a parent node of *B*.
  - A conditional probability distribution for each node depending on its parent nodes:

$$\mathbf{P}\,(X_i \mid Parents(X_i))$$

- In the simplest case, the conditional probability distribution is represented as a conditional probability table (CPT), which specifies the distribution over X_i for each combination of parent values.

Example

The topology of the net encodes conditional independence assertions:

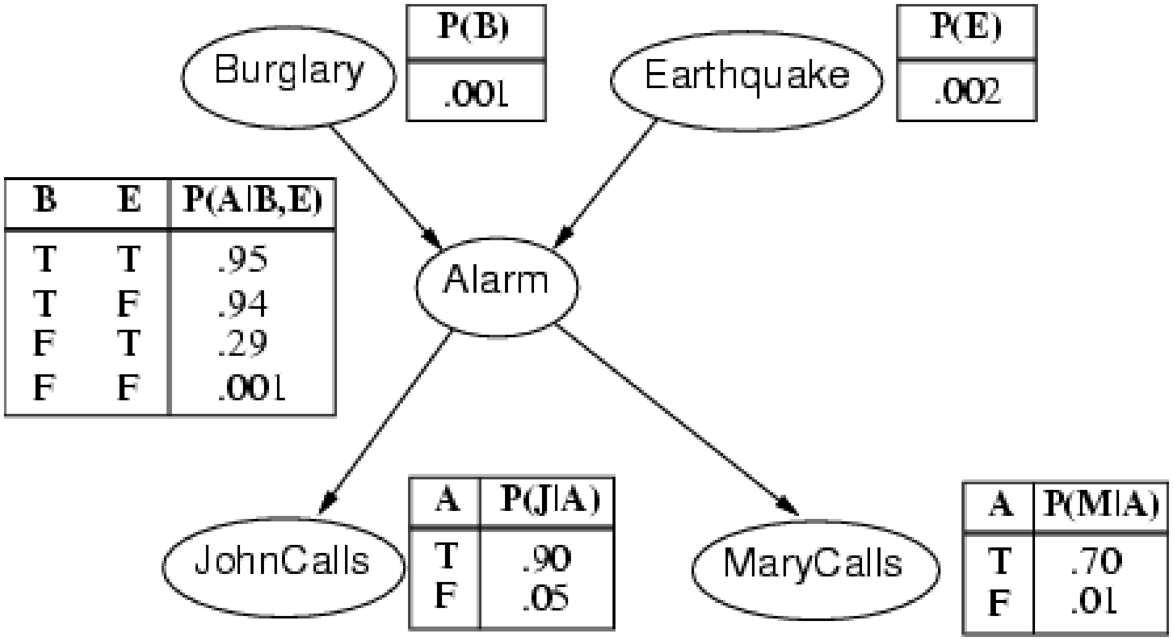


[RN]

- *Weather* is independent of the other variables.
- *Toothache* und *Catch* are conditionally independent given the value of *Cavity*.

Example

- I´m not at home. My neighbor John calls because my alarm is ringing but neighbor Mary does not call. The alarm should indicate burglars, but sometimes it is set off by minor earthquakes. Is there a burglary?
- Variables: *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*
- The joint probability distribution of the five varibles comprises $2^5-1 = 31$ independent numbers (without further knowledge about independencies).
- Network topology reflects causal knowledge:
  - A burglary can set off the alarm.
  - An earthquake can set off the alarm.
  - The alarm can cause Mary to call.
  - The alarm can cause John to call.



[RN]

Independencies: E.g., *MaryCalls* is not independent on *JohnCalls* (because the alarm makes Mary´s calling likely, but the same applies to John).

But *MaryCalls* is conditionally independent on *JohnCalls* given *Alarm*.

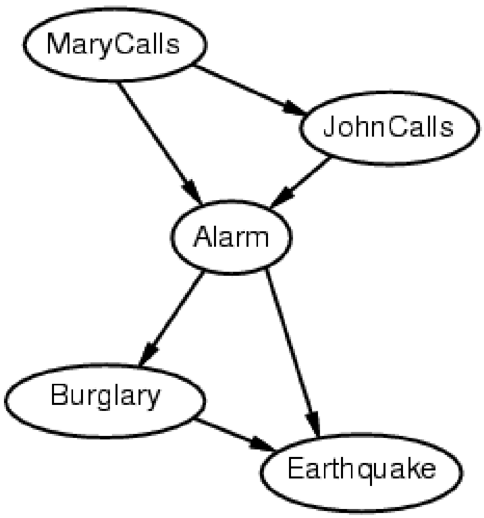
Knowledge about conditional independencies reduces joint probability distribution to 1 + 1 + 4 + 2 + 2 = 10 independent numbers.

The joint probability distribution is the product of the local conditional probability distributions:

$$\mathbf{P}\,(X_1, \ldots ,X_n) \;=\; \Pi_{i\,=\,1..n}\; \mathbf{P}\,(X_i \,|\, Parents(X_i)).$$

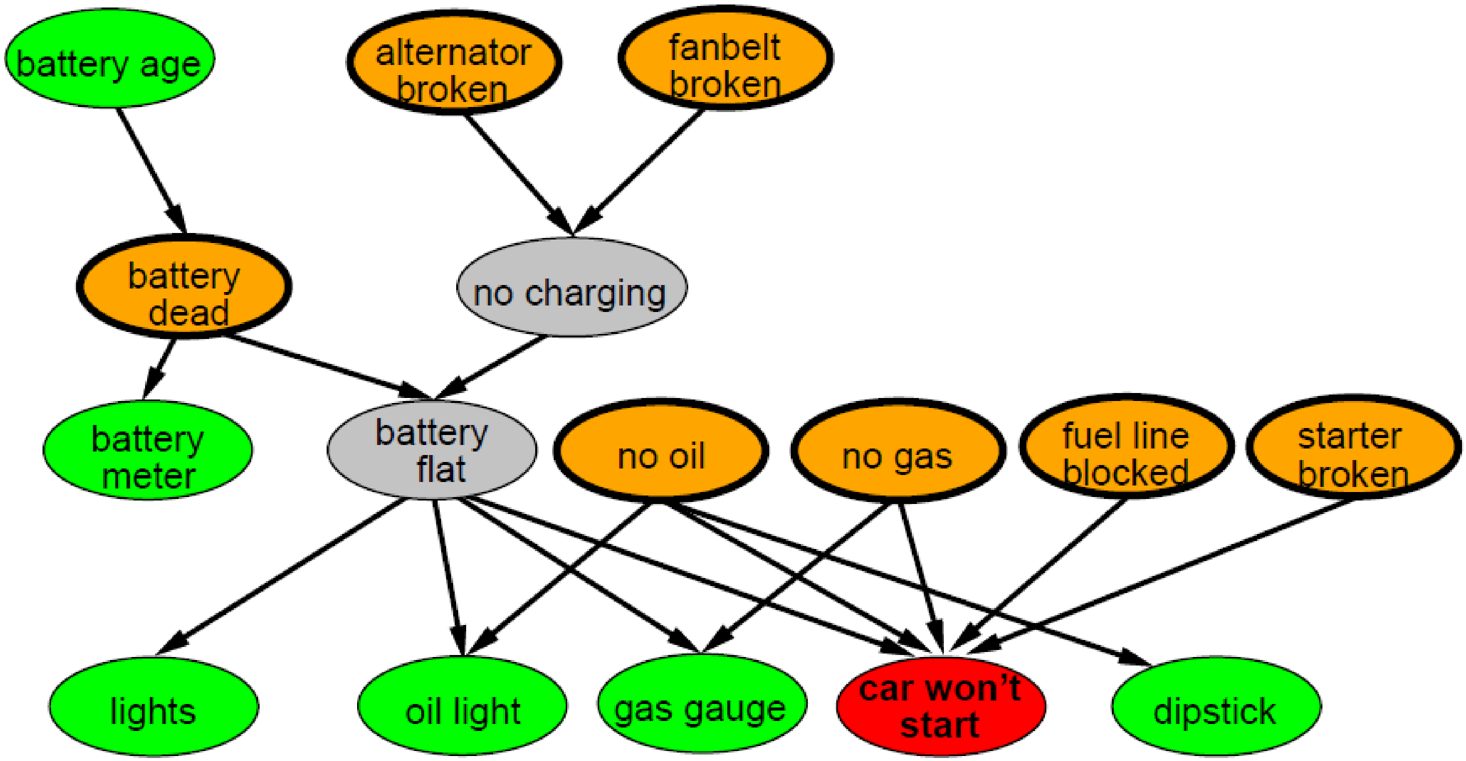
Note:

- It is convenient to construct the edges in the causal direction, but the opposite (diagnotic direction) is possible as well.
- However, dependencies change and so may
- the number of independent parameters, and
- the network becomes more difficult to interpret.



[RN]

Example


battery age
alternator broken
fanbelt broken
battery dead
no charging
battery meter
battery flat
no oil
no gas
fuel line blocked
starter broken
lights
oil light
gas gauge
car won't start
dipstick

[RN]

Summary

- Uncertainties can be represented by assigning probabilities to propositions (but that´s not the only way).
- The joint probability distribution assigns a probability to each atomic event and thus holds the complete domain knowledge.
- Inference is possible by summing up probabilities of atomic events.
- Independence reduces the complexity of the joint probability distribution but is rarely perfect in reality.
- Conditional independence is more feasible.
- Bayes nets are a convenient representation for the dependence / conditional independence of random variables.
- Causal direction of edges is easier to interpret.
- Bayes nets are easier to design for domain experts than conditional probabilities.

Image sources

[M] Online material available at www.cs.cmu.edu/~tom/mlbook.html
for the textbook: Tom M. Mitchell: *Machine Learning*, McGraw-Hill

[RN] Stuart Russell, Peter Norvig: *Artificial Intelligence*, Pearson

[H] Gunther Heidemann, 2012.