Lecture 8 - Reasoning under Uncertainty (Part I)

Jesse Hoey
School of Computer Science
University of Waterloo

February 10, 2020

Readings: Poole & Mackworth (2nd ed.)Chapt. 8 up to 8.4

Uncertainty

Why is uncertainty important?

- Agents (and humans) don't know **everything**,
- but need to make decisions anyways!
- Decisions are made in the absence of information,
- or in the presence of **noisy** information (sensor readings)

The best an agent can do:
know how uncertain it is, and act accordingly

Probability: Frequentist vs. Bayesian

Frequentist view:

probability of heads = # of heads / # of flips

probability of heads **this time** = probability of heads (history)

Uncertainty is **ontological**: pertaining to the world

Bayesian view:

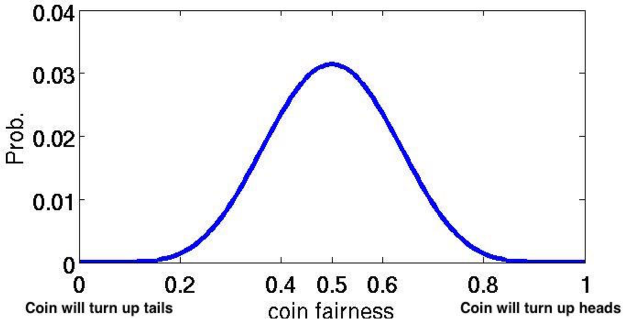
probability of heads **this time** = agent's **belief** about this event

belief of agent A : based on previous experience of agent A

Uncertainty is **epistemological**: pertaining to knowledge

Probability: Bayesian

Bayesian probability
all else being equal (Prior)
before 2 flips

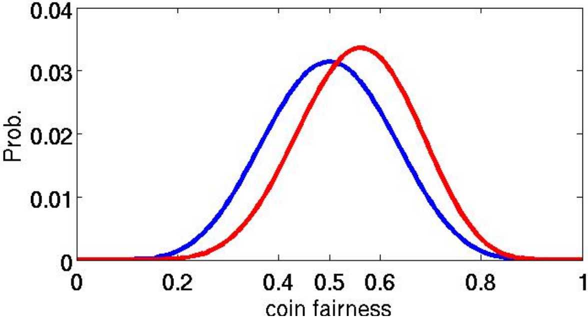


Probability: Bayesian

Bayesian probability

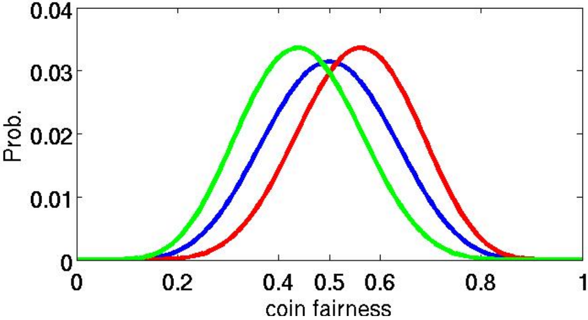
all else being equal (Prior)

after 2 flips heads, heads (Posterior)



Probability: Bayesian

Bayesian probability
all else being equal (Prior)
after 2 flips tails,tails (Posterior)



Probability: Bayesian

Should you wear your seatbelt?
estimate $P(fatality)$ given you do/don't wear it

Probability: Bayesian

Should you wear your seatbelt?
estimate $P(\mathit{fatality})$ given you do/don't wear it

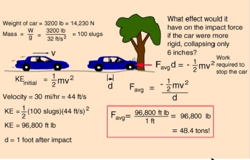
Frequentist:

test	day	result	P(fatality)
-	Sunday (prior to start)	-	?
1	Monday		0.0
2	Tuesday		0.0
3	Tuesday		0.33333
4	Thursday		0.25
5	Friday		0.2
...	...	...	...
N			Number of crashes / N

Probability: Bayesian

Should you wear your seatbelt?
estimate $P(fatality)$ given you do/don't wear it

Bayesian:



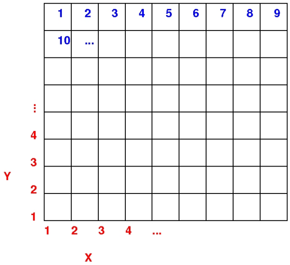
UK government:
"Seatbelts save 2200 lives/year"

car–accidents.com:
"in 63% of fatalities, no seatbelts were worn"

Features

Describe the **world** in terms of a set of **states**: $\{s_1, s_2, ..., s_N\}$

or, as the product of a set of **features**
(also known as **attributes** or **random variables**)



Robot on a grid

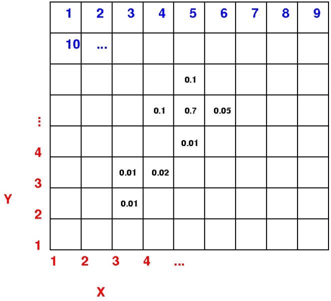
- Number of states = $2^{number\ \ of\ \ binary\ \ features}$
- Features describe the state space in a **factored** form.
- state $\rightarrow$ factorize $\rightarrow$ feature values
- feature values $\rightarrow$ cross product $\rightarrow$ states

Probability Measure

if X is a random variable (feature, attribute),
it can take on values x, where $x \in Domain(X)$
Assume x is discrete

$\mathbf{P}(\mathbf{x})$ is the probability that $X = x$
joint probability $\mathbf{P}(\mathbf{x}, \mathbf{y})$ is the
probability that $X = x$ and $Y = y$ at the same time

Joint probability distribution:



Robot on a grid

Axioms of Probability

Axioms are things we have to assume about probability:

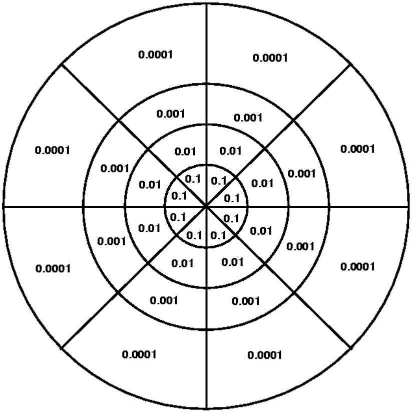
- $P(X) \geq 0$
- $\sum_x P(X = x) = 1.0$
- $P(a \vee b) = P(a) + P(b)$ if a and b are contradictory - can't both be true at the same time e.g. $P(win \vee lose) = P(win) + P(lose) = 1.0$

Some notes:

- probability between 0-1 is purely convention
- $P(a) = 0$ means a is definitely false
- $P(a) = 1$ means a is definitely true
- $0 < P(a) < 1$ means you have **belief** about the truth of a. It does **not** mean that a is true to some degree, just that you are ignorant of its truth value.
- Probability = measure of ignorance

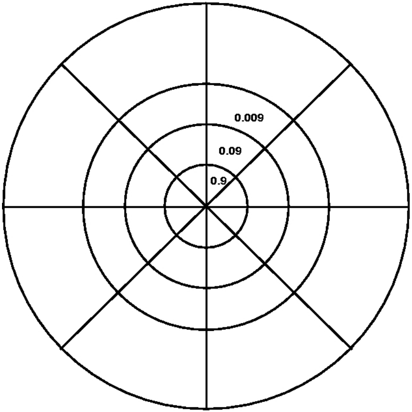
Independence

- describe a system with n features: $2^n - 1$ probabilities
- Use **independence** to reduce number of probabilities
- e.g. radially symmetric dartboard, P(hit a sector)
- $P(sector) = P(r, \theta)$ where $r = 1, \ldots, 4$ and $\theta = 1, \ldots, 8$.
- 32 sectors in total - need to give 31 numbers



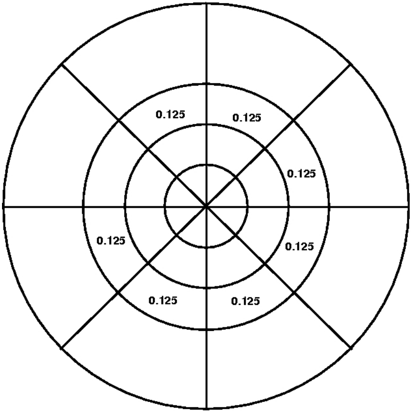
Independence

- describe a system with n features: $2^n - 1$ probabilities
- Use **independence** to reduce number of probabilities
- e.g. radially symmetric dartboard, P(hit a sector)
- assume radial independence: $P(r, \theta) = P(r)P(\theta)$
- only need 7+3=10 numbers



Independence

- describe a system with n features: $2^n - 1$ probabilities
- Use **independence** to reduce number of probabilities
- e.g. radially symmetric dartboard, P(hit a sector)
- assume radial independence: $P(r, \theta) = P(r)P(\theta)$
- only need 7+3=10 numbers



Independence - Example

4 independent Boolean variables, X_1, X_2, X_3, X_4, where x_i means $X_i = \mathit{true}$, and $\overline{x_i}$ means $X_i = \mathit{false}$

$$P(x_1) = 0.4, P(x_2) = 0.2, P(x_3) = 0.5, P(x_4) = 0.8$$

would usually need 16 numbers to specify the joint probability distribution, but if they are all independent, we just need 4:

$$\begin{aligned}
P(x_1, \overline{x_2}, x_3, x_4) &= P(x_1)(1 - P(x_2))P(x_3)P(x_4) \\
&= (0.4)(0.8)(0.5)(0.8) \\
&= 0.128 \\
P(x_1, x_2, x_3 | x_4) &= P(x_1)P(x_2)P(x_3) \\
&= (0.4)(0.2)(0.5) \\
&= 0.04
\end{aligned}$$

Conditional Probability

if X and Y are random variables, then

$P(x|y)$ is the probability that $X = x$ **given** that $Y = y$.

e.g.
$P(\mathit{flies}|\mathit{is\_bird})$ is different than $P(\mathit{flies})$
$P(\mathit{flies}|\mathit{is\_a\_penguin}, \mathit{is\_bird})$ is different again

incorporate independence:
$P(\mathit{flies}|\mathit{is\_bird}, \mathit{has\_feathers}) = P(\mathit{flies}|\mathit{is\_bird})$

Product rule (Chain rule):
$P(\mathit{flies}, \mathit{is\_bird}) = P(\mathit{flies}|\mathit{is\_bird})P(\mathit{is\_bird})$
$P(\mathit{flies}, \mathit{is\_bird}) = P(\mathit{is\_bird}|\mathit{flies})P(\mathit{flies})$
leads to : Bayes' rule
$P(\mathit{is\_bird}|\mathit{flies}) = \frac{P(\mathit{flies}|\mathit{is\_bird})P(\mathit{is\_bird})}{P(\mathit{flies})}$

Sum Rule

We know (an Axiom):

$$\sum_x P(X = x) = 1.0 \;\text{ and therefore that }\; \sum_x P(X = x|Y) = 1.0$$

This means that

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

proof:

$$\begin{aligned}\sum_x P(X = x, Y) &= \sum_x P(X = x|Y)P(Y) \\ &= P(Y)\sum_x P(X = x|Y) \\ &= P(Y)\end{aligned}$$

We call $P(Y)$ the marginal distribution over Y

Conditional Probability

- X and Y are *independent* iff

$$P(X) = P(X|Y)$$

$$P(Y) = P(Y|X)$$

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

so learning Y doesn't influence beliefs about X

- X and Y are *conditionally independent* given Z iff

$$P(X|Z) = P(X|Y, Z)$$

$$P(Y|Z) = P(Y|X, Z)$$

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

so learning Y doesn't influence beliefs about X *if you already know* Z...does **not** mean X and Y are independent

Expected Values

expected value of a function on X, $V(X)$:

$\mathbb{E}(V) = \sum_{x \in Dom(X)} P(x)V(x)$

where $P(x)$ is the probability that $X = x$.

This is useful in decision making, where $V(X)$ is the *utility* of situation X.

Bayesian decision making is then

$\mathbb{E}(V(\texttt{decision})) = \sum_{outcome} P(outcome|decision)V(outcome)$

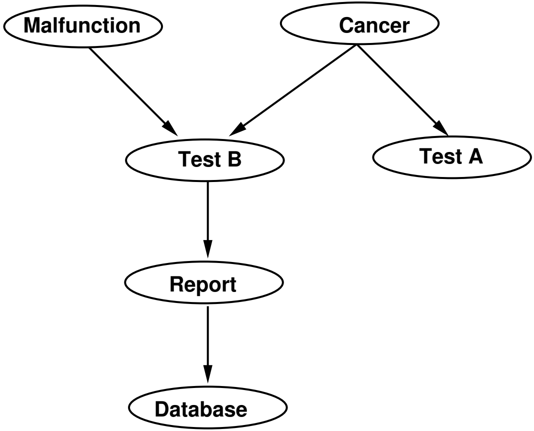
Value of Independence

- complete independence reduces both *representation* and *inference* from $O(2^n)$ to $O(n)$
- Unfortunately, complete mutual independence is rare
- Fortunately, most domains do exhibit a fair amount of *conditional independence*
- **Bayesian Networks** or **Belief Networks** (BNs) encode this information

Belief Networks

Bayesian network or **belief network**

- Directed Acyclic graph
- Encodes independencies in a graphical format
- Edges give $P(X_i|parents(X_i))$



Correlation and Causality

- Directed links in Bayes' net $\approx$ causal
- However, not always the case: chocolate $\rightarrow$ Nobel or Nobel $\rightarrow$ chocolate?
- In a Bayes net, it doesn't matter!
- But, some structures will be easier to specify

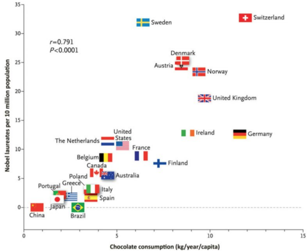


Figure 1. Taken from Messerli (2012). Correlation between countries' annual per capita chocolate consumption and the number of Nobel laureates per 10 million population. Copyright "The New England Journal of Medicine".

In this example, its probably
chocolate $\leftarrow$ *"Switzerland – ness"* $\rightarrow$ *Nobel*

Bayesian networks - example

If Jesse's alarm doesn't go off (A), Jesse probably won't get coffee (C); if Jesse doesn't get coffee, he's likely grumpy (G). If he is grumpy then it's possible that the lecture won't go smoothly L. If the lecture does not go smoothly then the students will likely be sad S.

$$A \longrightarrow C \longrightarrow G \longrightarrow L \longrightarrow S$$

A=Jesse's alarm doesn't go off
C=Jesse doesn't get coffee
G=Jesse is grumpy
L=lecture doesn't go smoothly
S=students are sad

Conditional Independence

- If you learned any of A, C, G, or L, would your assessment of $P(S)$ change?
  - If any of these are seen to be true, you would increase $P(s)$and decrease $P(\bar{s})$.
  - So S is not independent of A, C, G, L.
- If you knew the value of L (true or false), would learning the value of A, C, or G influence $P(S)$?
  - Influence that these factors have on S is mediated by their influence on L.
  - Students aren't sad because Jesse was grumpy, they are sad because of the lecture.
  - So S is independent of A, C, and G, given L

Conditional Independence

- So S is independent of A, C, and G, given L
- Similarly:
  - S is independent of A and C, given G
  - G is independent of A, given C
  - ...
- This means that:
  - $P(S|L, G, C, A) = P(S|L)$
  - $P(L|G, C, A) = P(L|G)$
  - $P(G|C, A) = P(G|C)$
  - $P(C|A)$ and $P(A)$ don't "simplify"

Conditional Independence

Chain rule (product rule):

$$P(S, L, G, C, A) = \\ P(S|L, G, C, A)P(L|G, C, A)P(G|C, A)P(C|A)P(A)$$

Independence:

$$P(S, L, G, C, A) = P(S|L)P(L|G)P(G|C)P(C|A)P(A)$$

So we can specify the full joint probability using the five local conditional probabilities

$$P(S|L), P(L|G), P(G|C), P(C|A), P(A)$$

Example Quantification

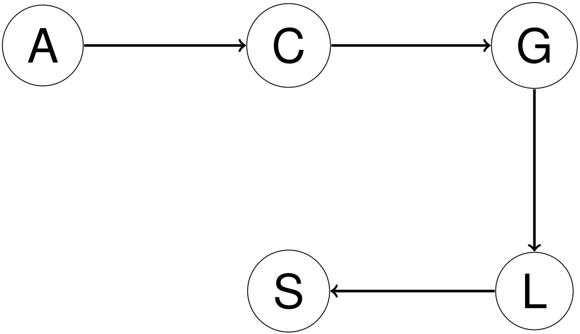
$\mathbf{P}(\mathbf{A} = \mathbf{true}) = 0.3$

$\mathbf{P}(\mathbf{C} = \mathbf{true}|\mathbf{A}) =$

A	$P(C = true\|A)$
t	0.8
f	0.15

$\mathbf{P}(\mathbf{G} = \mathbf{true}|\mathbf{C}) =$

C	$P(G = true\|C)$
t	1.0
f	0.2



$\mathbf{P}(\mathbf{S} = \mathbf{true}|\mathbf{L}) =$

L	$P(S = true\|L)$
t	0.9
f	0.3

$\mathbf{P}(\mathbf{L} = \mathbf{true}|\mathbf{G}) =$

G	$P(L = true\|G)$
t	0.7
f	0.2

Inference is Easy

$$A \longrightarrow C \longrightarrow G \longrightarrow L \longrightarrow S$$

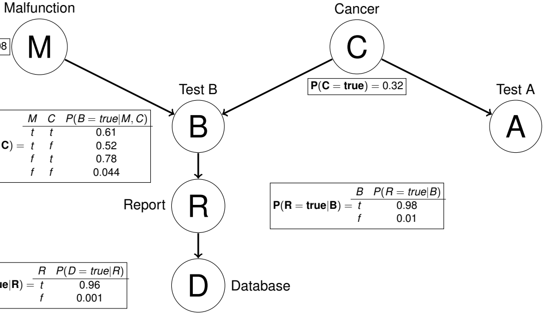
Want to know $P(g)$? - Use sum rule to sum over parents of g

$$\begin{aligned}
P(g) &= \sum_{c_i \in Dom(C)} P(g, c_i) \\
&= \sum_{c_i \in Dom(C)} P(g|c_i)P(c_i) \\
&= \sum_{c_i \in Dom(C)} P(g|c_i) \sum_{a_j \in Dom(A)} P(c_i|a_j)P(a_j) \\
&= P(g|c)[P(c|a)P(a) + P(c|\overline{a})P(\overline{a})] \\
&\quad + P(g|\overline{c})[P(\overline{c}|a)P(a) + P(\overline{c}|\overline{a})P(\overline{a})] \\
&= 1.0 * [0.8 * 0.3 + 0.15 * 0.7] + 0.2 * [0.2 * 0.3 + 0.85 * 0.7 \\
&= 0.4760
\end{aligned}$$

Bayesian Networks

A Bayesian Network (Belief Network, Probabilistic Network) or BN over variables $\{X_1, X_2, \ldots, X_N\}$ consists of:

- a DAG whose nodes are the variables
- a set of Conditional Probability tables (CPTs) giving $P(X_i | Parents(X_i))$ for each X_i



$\mathbf{P(M = true)} = 0.08$

$\mathbf{P(C = true)} = 0.32$

$\mathbf{P(B = true|M, C)} =$

M	C	$P(B = true \mid M, C)$
t	t	0.61
t	f	0.52
f	t	0.78
f	f	0.044

$\mathbf{P(A = true|C)} =$

C	$P(A = true \mid C)$
t	0.80
f	0.15

$\mathbf{P(R = true|B)} =$

B	$P(R = true \mid B)$
t	0.98
f	0.01

$\mathbf{P(D = true|R)} =$

R	$P(D = true \mid R)$
t	0.96
f	0.001

Semantics of a Bayes' Net

The structure of the BN means that :

every X_i is conditionally independent of all its nondescendants given its parents:

$$P(X_i|S, Parents(X_i)) = P(X_i|Parents(X_i))$$

for any subset $S \subseteq NonDescendants(X_i)$

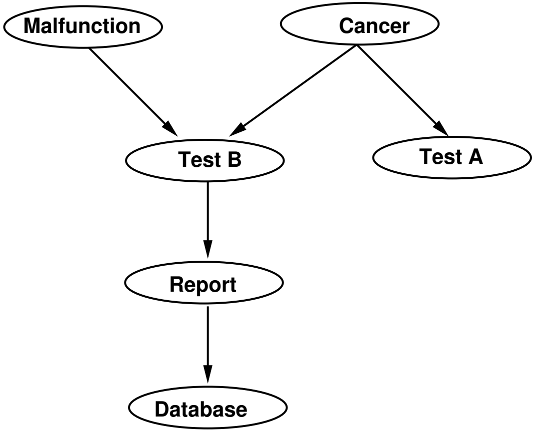
The BN defines a **factorization** of the joint probability distribution. The joint distribution is formed by multiplying the conditional probability tables together.

Constructing belief networks

To represent a domain in a belief network, you need to consider:

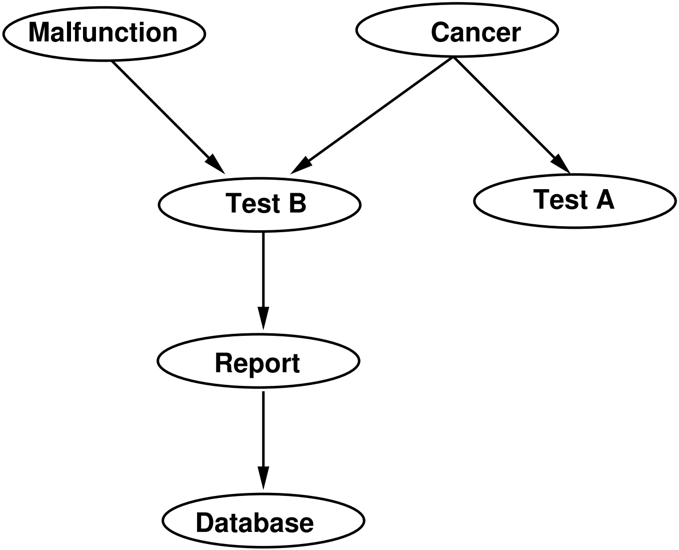
- What are the relevant variables?
  - What will you observe? - this is the evidence
  - What would you like to find out? - this is the query
  - What other features make the model simpler? - these are the other variables
- What values should these variables take?
- What is the relationship between them? This should be expressed in terms of local influence.
- How does the value of each variable depend on its parents? This is expressed in terms of the conditional probabilities.

Bayesian Networks - Independence assumptions

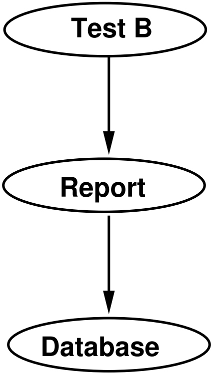


- **Test B** depends on **Cancer** and **Malfunction**
- **Test A** depends only on **Cancer**
- **Report** depends only on **Test B**
- **Database** depends only on **Report**

Three Basic Bayesian Networks

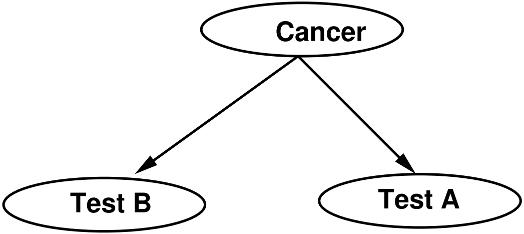


Three Basic Bayesian Networks



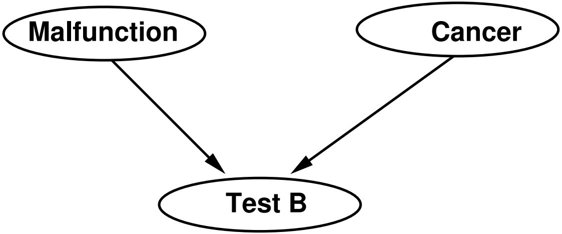
Database and Test B independent if Report is observed

Three Basic Bayesian Networks



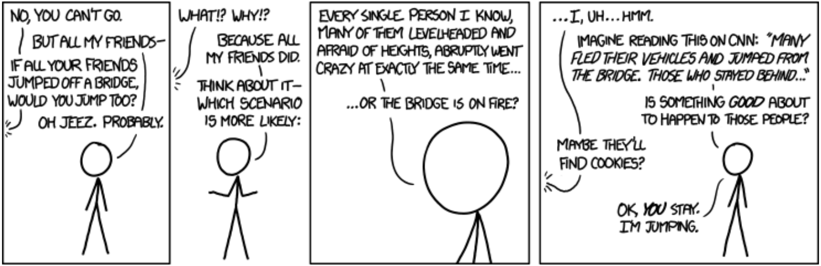
Test B and Test A are independent if Cancer is observed

Three Basic Bayesian Networks

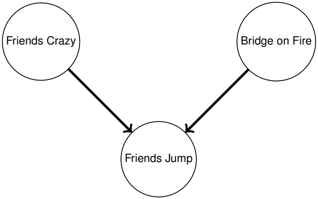


Malfunction and Cancer are independent if Test B is not observed

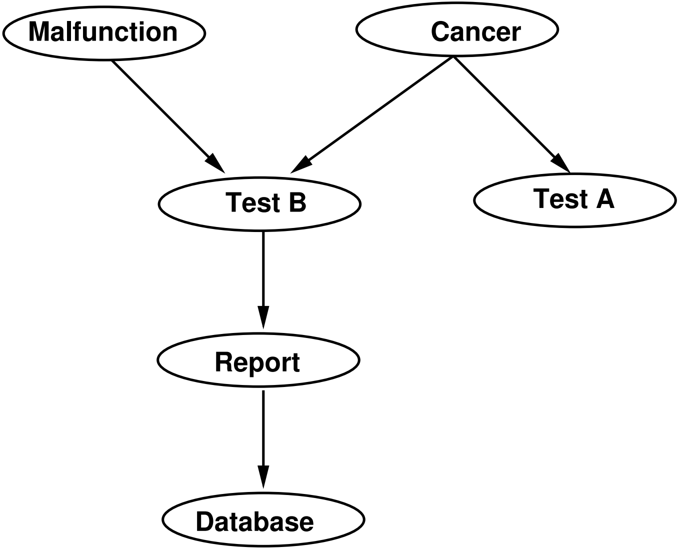
Three Basic Bayesian Networks



http://imgs.xkcd.com/comics/bridge.png



Three Basic Bayesian Networks



Testing Independence

Given a BN, how do we determine if two variables X,Y are independent (given evidence E)?

D-separation : A set of variables *E* d-separates X and Y if it blocks every undirected path in the BN between X and Y

But what does “block” mean?

Blocked Paths

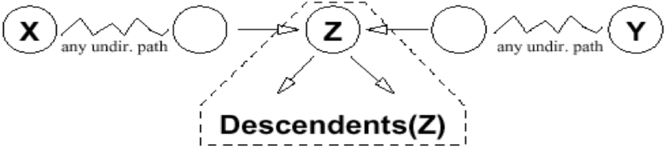
(1)

If Z in evidence, the path between X and Y blocked

(2)

If Z in evidence, the path between X and Y blocked

(3)



If Z is ***not*** in evidence and ***no*** descendent of Z is in evidence, then the path between X and Y is blocked

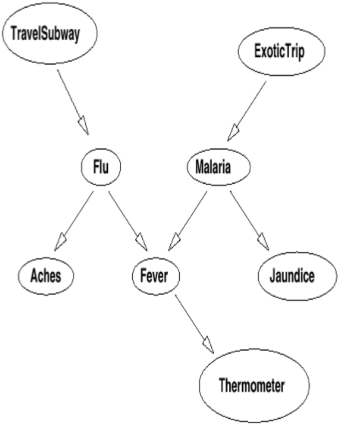
Markov Blanket

The **Markov Blanket** of a node *v* is:

- the parents, children, and the (other) parents of children
- the minimal set of nodes that d-separates *v* from all other variables

The joint distribution over the **Markov Blanket** allows for the computation of the distribution of *v*.

D-Separations: Example



- TravelSubway and Thermometer (given no evidence)?
- TravelSubway and Thermometer (given Flu or Fever)?
- TravelSubway and Malaria (given Fever)?
- TravelSubway and Exotic Trip (given Jaundice)?
- TravelSubway and Exotic Trip (given Jaundice and Thermometer)?
- TravelSubway and Exotic Trip (given Malaria and Thermometer)?

Updating belief: Bayes' Rule

Agent has a **prior belief** in a **hypothesis**, h, $P(h)$,

Agent observes some **evidence** e
that has a **likelihood** given the hypothesis: $P(e|h)$.

The agent's **posterior belief** about h after observing e, $P(h|e)$,

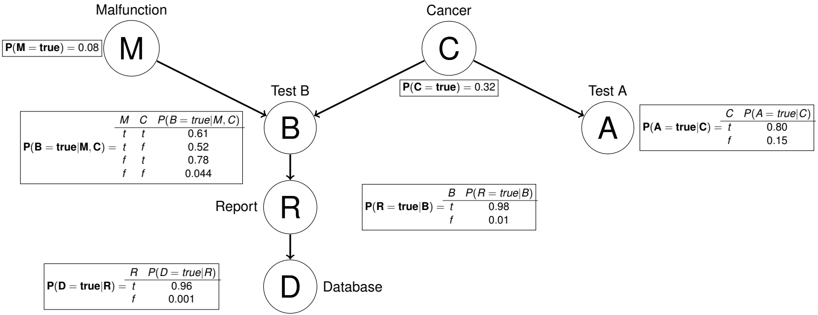
is given by **Bayes' Rule:**

$$P(h|e) = \frac{P(e|h)P(h)}{P(e)} = \frac{P(e|h)P(h)}{\sum_h P(e|h)P(h)}$$

Why is Bayes' theorem interesting?

- Often you have causal knowledge:
  $P(\textit{symptom} \mid \textit{disease})$
  $P(\textit{light is off} \mid \textit{status of switches and switch positions})$
  $P(\textit{alarm} \mid \textit{fire})$
  $P($*image looks like*  $\mid$ *a tree is in front of a car*$)$
- and want to do evidential reasoning:
  $P(\textit{disease} \mid \textit{symptom})$
  $P(\textit{status of switches} \mid \textit{light is off and switch positions})$
  $P(\textit{fire} \mid \textit{alarm})$.
  $P($*a tree is in front of a car* $\mid$ *image looks like* $)$

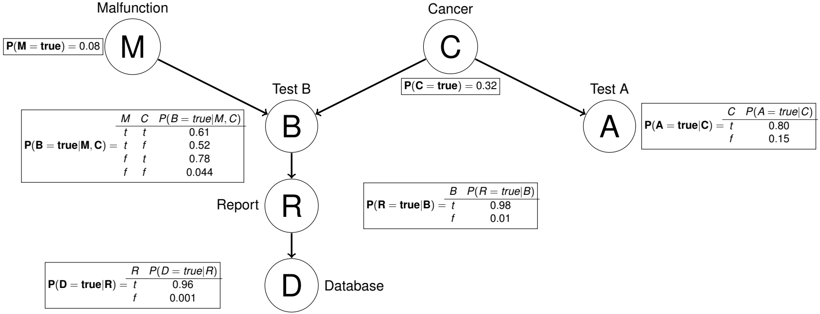
Probabilistic Inference



Before you get any information

- $P(Cancer) = 0.32$
- $P(Malfunction) = 0.08$

Probabilistic Inference



$\mathbf{P(M = true)} = 0.08$

$\mathbf{P(C = true)} = 0.32$

$\mathbf{P(B = true | M, C)} =$

M	C	$P(B = true \mid M, C)$
t	t	0.61
t	f	0.52
f	t	0.78
f	f	0.044

$\mathbf{P(A = true | C)} =$

C	$P(A = true \mid C)$
t	0.80
f	0.15

$\mathbf{P(R = true | B)} =$

B	$P(R = true \mid B)$
t	0.98
f	0.01

$\mathbf{P(D = true | R)} =$

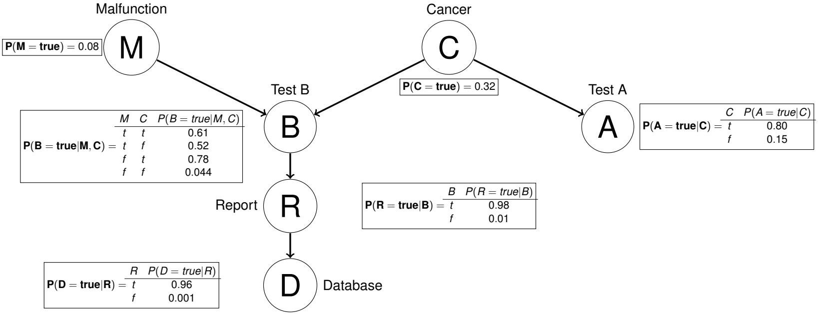
R	$P(D = true \mid R)$
t	0.96
f	0.001

Suppose the doctor reads a positive Test B in the Database so the evidence gives Database=true (not directly Test B= true) we want to know $P(Cancer = true | Database = true)$

- $P(Cancer = true | Database = true) = 0.80$
- $P(Malfunction = true | Database = true) = 0.14$

(we will see how to get these numbers later)

Probabilistic Inference



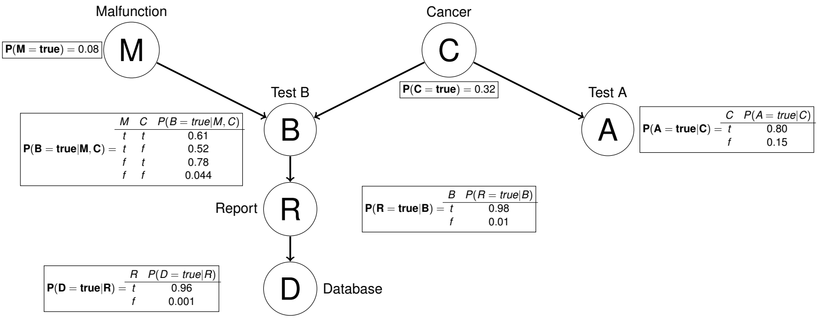
Suppose Test A is positive as well

we want $P(Cancer = true | Database = true \wedge TestA = true)$

- $P(Cancer = true | Database = true \wedge TestA = true) = 0.95$
- $P(M = true | Database = true \wedge TestA = true) = 0.08$

(we will see how to get these numbers later)

Probabilistic Inference



Suppose Test A is negative, though!

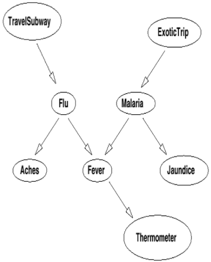
we want $P(Cancer = true|Database = true \wedge TestA = false)$

- $P(Cancer = true|Database = true \wedge TestA = false) = 0.48$
- $P(M = true|Database = true \wedge TestA = false) = 0.27$

(we will see how to get these numbers later)

Simple Forward Inference (Chain)

Computing marginal requires simple forward propagation of probabilities

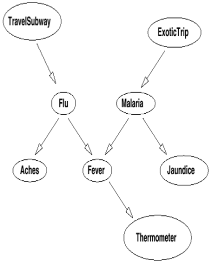


- $P(J) = \sum_{M,ET} P(J, M, ET)$ (marginalisation - sum rule)
- $P(J) = \sum_{M,ET} P(J|M, ET)P(M|ET)P(ET)$ (chain rule)
- $P(J) = \sum_{M,ET} P(J|M)P(M|ET)P(ET)$ (conditional indep).
- $P(J) = \sum_{M} P(J|M) \sum_{ET} P(M|ET)P(ET)$ (distribution of sum)

Note: all terms on the last line are CPTs in the BN
Note: only ancestors of J considered. Why?

Simple Forward Inference (Chain)

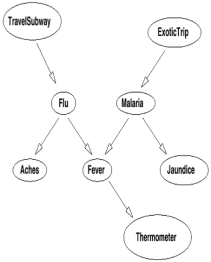
Same idea when evidence "upstream"



- $P(J|et) = \sum_M P(J, M|et)$ (marginalisation)
- $P(J|et) = \sum_M P(J|M, et)P(M|et)$ (chain rule)
- $P(J|et) = \sum_M P(J|M)P(M|et)$ (conditional indep).

Simple Forward Inference

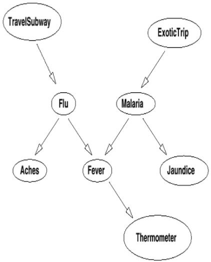
With multiple parents the evidence is "pooled"



$$P(Fev) = \sum_{Flu,M,TS,ET} P(Fev, Flu, M, TS, ET)$$

$$= \sum_{Flu,M} P(Fev|M, Flu)[\sum_{TS} P(Flu|TS)P(TS)][\sum_{ET} P(M|ET)P(ET)]$$

also works with “upstream” evidence



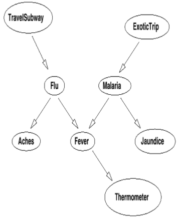
$$P(Fev|ts,\overline{m}) = \sum_{Flu} P(Fev, Flu|\overline{m}, ts)$$

$$= \sum_{Flu} P(Fev|Flu, ts, \overline{m}) P(Flu|ts, \overline{m})$$

$$= \sum_{Flu} P(Fev|Flu, \overline{m}) P(Flu|ts)$$

Simple Backward Inference

When evidence is downstream of query, then we must reason "backwards". This requires Bayes' rule

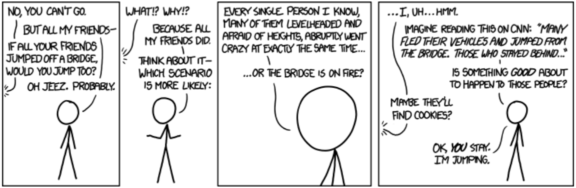


$$
\begin{aligned}
P(ET|j) &\propto P(j, ET) \\
&= P(j|ET)P(ET) \quad \text{(Bayes' rule)} \\
&= \sum_M P(j, M|ET)P(ET) \\
&= \sum_M P(j|M, ET)P(M|ET)P(ET) \\
&= \sum_M P(j|M)P(M|ET)P(ET)
\end{aligned}
$$

normalising constant is $\frac{1}{P(j)}$, but this can be computed as

$$P(j) = \sum_{ET} P(ET, j)$$

Backward Inference

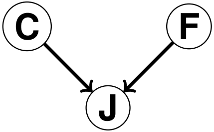


`http://imgs.xkcd.com/comics/bridge.png`

$\mathbf{P}(\mathbf{C} = \mathbf{true}) = 0.0001$

$\mathbf{P}(\mathbf{F} = \mathbf{true}) = 0.1$

F: Bridge on Fire
C: All friends Crazy
J: All friends Jump
What is $P(F|J = true)$?



$\mathbf{P}(\mathbf{J} = \mathbf{true}|\mathbf{F}, \mathbf{C}) =$

F	C	$P(J = true\|F, C)$
t	t	0.95
t	f	0.99
f	t	0.99
f	f	0.01

Variable Elimination

- intuitions above : polytree algorithm
- works for simple networks without loops
- more general algorithm: Variable Elimination
- applies sum-out rule repeatedly
- distributes sums

Factors

A factor is a representation of a function from a tuple of random variables into a number.
We will write factor f on variables $X_1, \ldots, X_j$ as $f(X_1, \ldots, X_j)$.
We can assign some or all of the variables of a factor
$\rightarrow$(this is "restricting" a factor):

- $f(X_1 = v_1, X_2, \ldots, X_j)$, where $v_1 \in dom(X_1)$, is a factor on $X_2, \ldots, X_j$.
- $f(X_1 = v_1, X_2 = v_2, \ldots, X_j = v_j)$ is a number that is the value of f when each X_i has value v_i.

The former is also written as $f(X_1, X_2, \ldots, X_j)_{X_1 = v_1}$, etc.

Example factors - Restricting a factor

$r(X, Y, Z)$:

X	Y	Z	val
t	t	t	0.1
t	t	f	0.9
t	f	t	0.2
t	f	f	0.8
f	t	t	0.4
f	t	f	0.6
f	f	t	0.3
f	f	f	0.7

$r(X{=}t, Y, Z)$:

Y	Z	val
t	t	0.1
t	f	0.9
f	t	0.2
f	f	0.8

$r(X{=}t, Y, Z{=}f)$:

Y	val
t	0.9
f	0.8

$$r(X{=}t, Y{=}f, Z{=}f) = 0.8$$

Multiplying factors

The **product** of factor $f_1(X, Y)$ and $f_2(Y, Z)$, where Y are the variables in common, is the factor $(f_1 \times f_2)(X, Y, Z)$ defined by:

$$(f_1 \times f_2)(X, Y, Z) \quad = \quad f_1(X, Y) f_2(Y, Z).$$

Multiplying factors example

f_1:

A	B	val
t	t	0.1
t	f	0.9
f	t	0.2
f	f	0.8

f_2:

B	C	val
t	t	0.3
t	f	0.7
f	t	0.6
f	f	0.4

$f_1 \times f_2$:

A	B	C	val
t	t	t	0.03
t	t	f	0.07
t	f	t	0.54
t	f	f	0.36
f	t	t	0.06
f	t	f	0.14
f	f	t	0.48
f	f	f	0.32

Summing out variables

We can sum out a variable, say X_1 with domain $\{v_1, \ldots, v_k\}$, from factor $f(X_1, \ldots, X_j)$, resulting in a factor on $X_2, \ldots, X_j$ defined by:

$$
\begin{aligned}
&(\sum_{X_1} f)(X_2, \ldots, X_j) \\
&= \quad f(X_1 = v_1, \ldots, X_j) + \cdots + f(X_1 = v_k, \ldots, X_j)
\end{aligned}
$$

Summing out a variable example

f_3:

A	B	C	val
t	t	t	0.03
t	t	f	0.07
t	f	t	0.54
t	f	f	0.36
f	t	t	0.06
f	t	f	0.14
f	f	t	0.48
f	f	f	0.32

$\sum_B f_3$:

A	C	val
t	t	0.57
t	f	0.43
f	t	0.54
f	f	0.46

Evidence

If we want to compute the posterior probability of Z given evidence $Y_1 = v_1 \wedge \ldots \wedge Y_j = v_j$:

$$
\begin{aligned}
& P(Z | Y_1 = v_1, \ldots, Y_j = v_j) \\
= & \frac{P(Z, Y_1 = v_1, \ldots, Y_j = v_j)}{P(Y_1 = v_1, \ldots, Y_j = v_j)} \\
= & \frac{P(Z, Y_1 = v_1, \ldots, Y_j = v_j)}{\sum_Z P(Z, Y_1 = v_1, \ldots, Y_j = v_j).}
\end{aligned}
$$

So the computation reduces to the probability of $P(Z, Y_1 = v_1, \ldots, Y_j = v_j)$.
We normalize at the end.

Probability of a conjunction

Suppose the variables of the belief network are $X_1, \ldots, X_n$.
To compute $P(Z, Y_1 = v_1, \ldots, Y_j = v_j)$, we sum out the other variables, $Z_1, \ldots, Z_k = \{X_1, \ldots, X_n\} - \{Z\} - \{Y_1, \ldots, Y_j\}$.
We order the Z_i into an elimination ordering.

$$
\begin{aligned}
& P(Z, Y_1 = v_1, \ldots, Y_j = v_j) \\
&\quad = \sum_{Z_k} \cdots \sum_{Z_1} P(X_1, \ldots, X_n)_{Y_1 = v_1, \ldots, Y_j = v_j}. \\
&\quad = \sum_{Z_k} \cdots \sum_{Z_1} \prod_{i=1}^{n} P(X_i | parents(X_i))_{Y_1 = v_1, \ldots, Y_j = v_j}.
\end{aligned}
$$

Computing sums of products

Computation in belief networks reduces to computing the sums of products.

- How can we compute $ab + ac$ efficiently?

Computing sums of products

Computation in belief networks reduces to computing the sums of products.

- How can we compute $ab + ac$ efficiently?
- Distribute out the a giving $a(b + c)$

Computing sums of products

Computation in belief networks reduces to computing the sums of products.

- How can we compute $ab + ac$ efficiently?
- Distribute out the a giving $a(b + c)$
- How can we compute $\sum_{Z_1} \prod_{i=1}^{n} P(X_i | parents(X_i))$ efficiently?

Computing sums of products

Computation in belief networks reduces to computing the sums of products.

- How can we compute $ab + ac$ efficiently?
- Distribute out the a giving $a(b + c)$
- How can we compute $\sum_{Z_1} \prod_{i=1}^{n} P(X_i | parents(X_i))$ efficiently?
- Distribute out those factors that don't involve Z_1.

Variable elimination algorithm

To compute $P(Z|Y_1 = v_1 \wedge \ldots \wedge Y_j = v_j)$:

- Construct a factor for each conditional probability.
- Set the observed variables to their observed values.
- Sum out each of the other variables (the $\{Z_1, \ldots, Z_k\}$) according to some elimination ordering.
- Multiply the remaining factors. Normalize by dividing the resulting factor $f(Z)$ by $\sum_Z f(Z)$.

Summing out a variable

To sum out a variable Z_j from a product $f_1, \ldots, f_k$ of factors:

- Partition the factors into
  - those that don't contain Z_j, say $f_1, \ldots, f_i$,
  - those that contain Z_j, say $f_{i+1}, \ldots, f_k$

We know:

$$\sum_{Z_j} f_1 \times \cdots \times f_k = f_1 \times \cdots \times f_i \times \left(\sum_{Z_j} f_{i+1} \times \cdots \times f_k \right).$$

- Explicitly construct a representation of the rightmost factor. Replace the factors $f_{i+1}, \ldots, f_k$ by the new factor.

Example I

$\mathbf{P}(\mathbf{A} = \mathbf{true}) = 0.3$

$\mathbf{P}(\mathbf{C} = \mathbf{true}|\mathbf{A}) =$

A	$P(C = true\|A)$
t	0.8
f	0.15

$\mathbf{P}(\mathbf{G} = \mathbf{true}|\mathbf{C}) =$

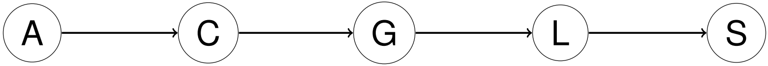
C	$P(G = true\|C)$
t	1.0
f	0.2

$\mathbf{P}(\mathbf{L} = \mathbf{true}|\mathbf{G}) =$

G	$P(L = true\|G)$
t	0.7
f	0.2

$\mathbf{P}(\mathbf{S} = \mathbf{true}|\mathbf{L}) =$

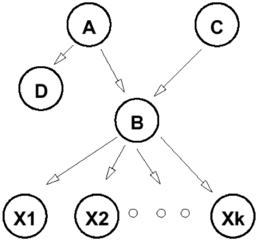
L	$P(S = true\|L)$
t	0.9
f	0.3



Notes on VE

- Complexity is linear in number of variables, and exponential in the size of the largest factor
- When we create new factors: sometimes this blows up
- Depends on the elimination ordering
- For **polytrees**: work outside in
- For general BNs this can be hard
- simply **finding** the optimal elimination ordering is NP-hard for general BNs
- inference in general is NP-hard

Variable Ordering: Polytrees



- eliminate "singly-connected" nodes ($D, A, C, X_1, \ldots, X_k$) first
- Then no factor is ever larger than original CPTs
- If you eliminate B first, a factor is created that includes $A, C, X_1, \ldots, X_k$

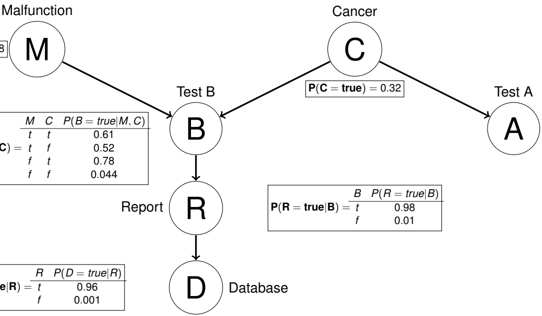
Variable Ordering: Relevance

- Certain variables have no impact
- In ABC network above, computing $P(A)$ does not require summing over B and C

$$\begin{aligned} P(A) &= \sum_{B,C} P(C|B)P(B|A)P(A) \\ &= P(A)\sum_{B} P(B|A)\sum_{C} P(C|B) = P(A) * 1.0 * 1.0 \end{aligned}$$

- Can restrict attention to relevant variables:
- Given query Q and evidence $\mathbf{E}$, complete approximation is:
  - Q is relevant
  - if any node is relevant, its parents are relevant
  - if $E \in \mathbf{E}$ is a descendent of a relevant variable, then E is relevant
- irrelevant variable: a node that is not an ancestor of a query or evidence variable

Example II



$\mathbf{P}(\mathbf{M} = \mathbf{true}) = 0.08$

$\mathbf{P}(\mathbf{C} = \mathbf{true}) = 0.32$

$\mathbf{P}(\mathbf{B} = \mathbf{true}|\mathbf{M}, \mathbf{C}) =$

M	C	$P(B = true\|M, C)$
t	t	0.61
t	f	0.52
f	t	0.78
f	f	0.044

$\mathbf{P}(\mathbf{A} = \mathbf{true}|\mathbf{C}) =$

C	$P(A = true\|C)$
t	0.80
f	0.15

$\mathbf{P}(\mathbf{R} = \mathbf{true}|\mathbf{B}) =$

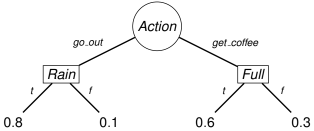
B	$P(R = true\|B)$
t	0.98
f	0.01

$\mathbf{P}(\mathbf{D} = \mathbf{true}|\mathbf{R}) =$

R	$P(D = true\|R)$
t	0.96
f	0.001

Other Representations for Probability distributions

- Decision Tree:



- Noisy Or: $P(x|Y_1, \ldots, Y_k)$ where X is caused by parents Y
- Logistic Regression where parents Y caused by X.

$$P(x|Y_1, \ldots, Y_k) = sigmoid(\sum_i w_i Y_i)$$

Next:

- Reasoning under Uncertainty Part II (Poole & Mackworth (2nd ed.)Chapter 8.5-8.9)