Motivation

The knowledge that an intelligent agent possesses about her world is necessarily uncertain and incomplete. Even in situations when she could acquire a complete and certain information, it could be impractical.

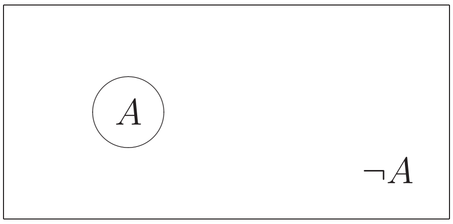
Artificial intelligence has long tried to create models for reasoning and acting under uncertainty, eg. by including certainty factors for the fact statements in the logical reasoning. Examples of such models are: modal logic, three-valued logic, nonmonotonic logic, fuzzy logic, probabilistic logic, and others.

The practical usefulness of such methods turn out to be limited. And only quite recently did artificial intelligence turn to using the probability directly. This turned out to be a good and fruitful approach. The approach to knowledge representation based on probability is one of the most dynamically expanding application area of artificial intelligence. The reasoning mechanism in this approach is the mathematical probability theory.

Prior probability

The **prior probability** designates numerical odds of some chance event occurring, when no other information which may influence the event is known (such as, if it did in fact occur).

A graphical visualization of events and their probability:



$P(A) =$ area of the circle

$P(\neg A) =$ circle complement

area of the rectangle $= 1$

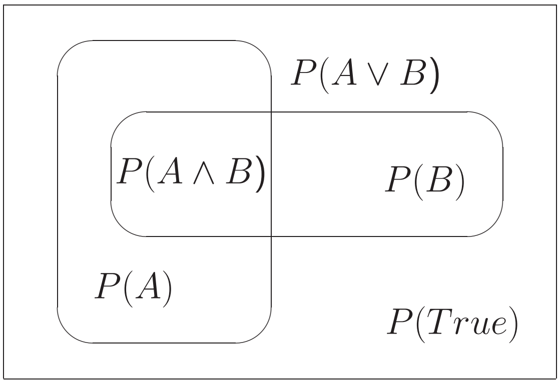
Eg.: the probability, that a patient coming to a doctor's office has the Severe Acute Respiratory Syndrome (SARS)[1] may be $P(\mathit{SARS}) = 0.0001$

However, if the doctor knew, that the patient has just arrived from Hong-Kong and had all the typical symptoms of the acute respiratory syndrome then the probability of her having the SARS virus would have to be determined differently.

[1] Explanation: this example was conceived in 2003 when there was an outbreak of SARS in China. SARS is a coronavirus causing severe respiratory infections, with initial symptoms resembling a flu. There is no known effective therapy, but since 2004 the number of infections dropped to zero, worldwide.

The probability axioms

- $0 \leq P(A) \leq 1$
- $P(\text{True}) = 1$
- $P(\text{False}) = 0$
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$



More on the probability axioms

From the above axioms of probability one can derive many interesting properties:

$$\begin{aligned} P(\neg A) &= 1 - P(A) && (1) \\ P(A) &= P(A \wedge B) + P(A \wedge \neg B) && (2) \end{aligned}$$

(and others).

The probability axioms have a deep significance and a number of important properties follow from them. For example, an agent who adheres to them in her beliefs, and makes some betting decisions accordingly, is guaranteed not to make a mistake.

In other words, if in some probabilistic game an agent utilized probabilities violating the above axioms, and continued to accept bets in accordance to them, then there exists a winning strategy for her opponents.

Random variables

A random variable represents some chance event, which may assume values from some set (the domain of the random variable).

For example: if an agent wished to state probabilities about that kind of weather would be today, she might consider *Weather*$_{\text{today}}$ to be a random variable with values in the set $\{$*Sunny,Cloudy,Rain,Snow*$\}$

The mapping of all the values of a random variable to some probability values is called the **probability distribution** of the variable. The probability distribution for the random variable *Weather*$_{\text{today}}$ can be written as: $\mathbf{P}(\mathit{Weather}_{\text{today}}) = \{0.8, 0.1, 0.09, 0.01\}$

Joint probability distribution

We could consider several random variables describing different chance events. We call an **atomic event** an assignment of values for all the random variables. For example, for two random variables X and Y one can create a table of random events:

	$X = x_1$	$X = x_2$	$\ldots$	$X = x_n$
$Y = y_1$				
$Y = y_2$				
$\ldots$				
$Y = y_k$				

A joint probability distribution (JPD) for a set of random variables is the table of probabilities of all the atomic events. In this table, in row j and column i there is the probability of variable X assuming the value x_i and the variable Y simultaneously assuming the value y_j, or $P(X = x_i \wedge Y = y_j)$. Adding values along rows or columns in this table we obtain the probability of specific values of single variables. The sum of all the probabilities in the whole table is 1.0.

Using the JPD table

Having filled out the JPD table we can compute the probabilities of any events. For example:

• The probability of the event of variable X assuming the value x_1 $P(X = x_1)$ can be computed by summing all the values in the column 1 of the JPD.

• The probability of the event of variable X assuming the value x_1 **or** variable Y assuming the value y_2 can be computed by summing all the values from column 1 and all the values from row 2 of the JPD, counting the slot $(2, 1)$ of the table only once. The result would be exactly the same as that obtained by applying the formula:

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

However, to be able to use probability in this way one has to compute the probabilities of all the atomic events, which for many variables with many values can be expensive.

Computing atomic probabilities

Where do the probability value come from? They can be collected from statistics, or derived from the physical principles and the properties of a phenomenon. We can also associate the probability with an agent, reflecting her point of view of the world.

For example, what is the probability of the event, that the sun will exist tomorrow? One can try to compute it in many different ways, assuming different points of view:

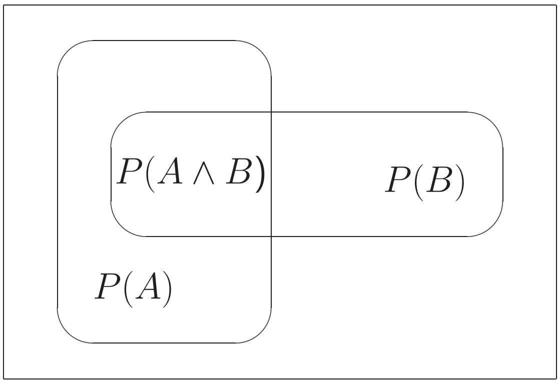
- cannot determine, because the necessary experiments cannot be conducted,
- previous „similar” experiments prove, that the sun „always” existed, so the probability is 1,
- the probability is $1 - \epsilon$ where ϵ is the probability of a star explosion on a given day,
- the probability is $d/(d + 1)$ where d is the number of days the sun existed so far,
- the probability can be determined by building a model of the sun and its evolution, based on similar stars.

Conditional probability

Conditional (*a posteriori*) probability $P(A|B)$ — the probability of an event A, computed in the situations, when B is true. It is connected with the prior probability according the the formula:

$$P(A|B) = \frac{P(A \wedge B)}{P(B)} \tag{3}$$

The above formula can be explained as follows: to compute the probability $P(A|B)$ we must take the fraction of the event $A \wedge B$ occurring in all the cases of the event B.



Another explanation can be given using the inverted formula:

$$P(A \wedge B) = P(A|B)P(B) \tag{4}$$

To compute $P(A \wedge B)$ we must know that B occurred, and knowing that, compute the probability of A. Or the other way around.

An important, often used formula linking the prior probability of an event with its conditional probability, can be obtained from combining formulas (2, slide 4) with (4):

$$P(A) = P(A|B)P(B) + P(A|\neg B)P(\neg B) \tag{5}$$

Let's note, that the conditional probability for some specific condition satisfies all the probability axioms, so has all the properties of the prior probability, for example:

$$P(A|B) + P(\neg A|B) = 1 \tag{6}$$

Why conditional probability?

Consider a classical puzzle:

> Suppose you're on a game show, and you're given the choice of three doors: behind one door is a car. Having no prior knowledge you pick any door, say No. 1. Then the host, who knows what's behind the doors, opens another door, say No. 3, which is empty. He then offers you a chance to switch your choice by picking door No. 2?
> Is it to your advantage to switch?

The prior probability of winning the car was 1/3. The host somehow "increased" it to 1/2 by opening an empty door. But did he really?

If we unconditionally stick to our original choice then we haven't acquired any new information from opening the other doors, so why should our chance of winning go up? On the other hand, we did obtain new information, so could we use it to increase our odds? But how?

Answer: switching our choice to doors No. 2 increases the chance of winning to 2/3.

We have to use the conditional probability whenever we want to compute the probability of some event in a situation, when we have some knowledge of some other events occurring, which may be related to the event in question.

$P(A)$ is the correct probability of the event A if we don't have any knowledge. If, on the other hand, we know that B occurred, then the correct probability of A is $P(A|B)$, and if we later found out that C is true, then we would have to compute the probability $P(A|B \wedge C)$.

We can consider the prior probability $P(A)$ to be the conditional probability $P(A|)$ in the situation, when we have no knowledge at all.

The conditional probability can be computed from the joint probability distribution table (JPD) using the formula 3.

However, this is normally not done this way.

Bayes' rule

By applying twice the formula 3 we can obtain the following simple formula, called the Bayes' rule, which is the basis for many probabilistic reasoning processes:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{7}$$

Why is this formula so important? Let's go back to the patient with the symptoms of SARS, a very dangerous disease. Suppose a test was done which can detect the virus, and the result came out positive. Should the patient immediately be hospitalized, and a treatment started? Turns out it depends!!

The test may be good, but usually it is not perfect. A good test guarantees giving a high probability of a positive outcome for cases when the virus is present. It turns out to be equally important that it should give a negative outcome for cases when there is no virus.

So the test gives large values of both $P(T^{\oplus}|\mathit{SARS})$ and $P(T^{\ominus}|\neg \mathit{SARS})$. However, what interests a doctor (and of course his patient), are the values of $P(\mathit{SARS}|T^{\oplus})$ and $P(\neg \mathit{SARS}|T^{\ominus})$.

Bayes' rule — an example

To evaluate the probability of having a disease from the results of a blood test, it is necessary to reverse the conditions in the conditional probability, which just means applying the Bayes' rule.

Let us assume the test for SARS gives the positive (correct) result in 95% cases of the presence of the virus. Moreover, in the cases of no virus the same test gives the negative (also correct) result in 90% cases.

$$\begin{aligned} P(\mathit{SARS}) &= 0.0001 \\ P(T^{\oplus}|\mathit{SARS}) &= 0.95 \\ P(T^{\ominus}|\neg \mathit{SARS}) &= 0.90 \end{aligned}$$

Now suppose some patient has been given the test, and it came out positive.
So what is the probability that the patient indeed has the virus?
Let's try to estimate the $P(\mathit{SARS}|T^{\oplus})$?

$$P(\mathit{SARS}|T^{\oplus}) = \frac{P(T^{\oplus}|\mathit{SARS})P(\mathit{SARS})}{P(T^{\oplus})}$$

we are missing the value of $P(T^{\oplus})$, which can be computed as follows:

$$P(T^{\oplus}) = P(T^{\oplus}|\mathit{SARS})P(\mathit{SARS}) + P(T^{\oplus}|\neg\mathit{SARS})P(\neg\mathit{SARS})$$

$$\begin{gathered} P(T^{\oplus}) = 0.95 \times 0.0001 + 0.10 \times 0.9999 \\ P(T^{\oplus}) = 0.000095 + 0.09999 \\ P(T^{\oplus}) = 0.100085 \end{gathered}$$

and finally we compute the interesting value:

$$P(\mathit{SARS}|T^{\oplus}) = \frac{0.95 \times 0.0001}{0.100085}$$

$$P(\mathit{SARS}|T^{\oplus}) = 0.00094919$$

which is below 0.1%!! This is almost ten times more than the average, but perhaps not enough to start a possibly costly and/or harmful therapy.

As we can see, having the **causal** knowledge about the mechanism of the disease enables us to compute the **diagnostic** probabilities. But why must we compute them each time; why does the manufacturer of the virus test provide the $P(T^{\oplus}|\mathit{SARS})$ and $P(T^{\ominus}|\neg\mathit{SARS})$ figures, instead of delivering the more useful $P(\mathit{SARS}|T^{\oplus})$ value?

This comes from an easier availability of causal data than diagnostic data, which may be harder to determine. For example, if there was a sudden epidemic of *SARS* (Epi), then the value of $P(\mathit{SARS})$ would go up, and, consequently the $P(\mathit{SARS}|T^{\oplus})$. However, the value of $P(T^{\oplus}|\mathit{SARS})$ should remain constant, since it only represents the physiology of the disease and the effect of the test. So the above computation remains valid after considering the higher value of $P(\mathit{SARS})$. [2]

[2] In fact, the $P(T^{\oplus})$ will also change in this case, computed as $P(T^{\oplus}|Epi)$, but we can compute it:

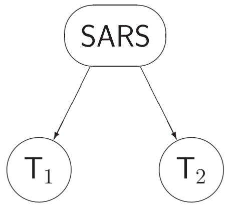
$$P(T^{\oplus}|Epi) = P(T^{\oplus}|\mathit{SARS}, Epi)P(\mathit{SARS}|Epi) + P(T^{\oplus}|\neg\mathit{SARS}, Epi)(1 - P(\mathit{SARS}|Epi))$$

Bayes' rule — condition independence

Let's turn back to the patient with the positive SARS test result. Perhaps the probability of her having the virus is not high enough to place her in a hospital and start a treatment. Suppose another test exists, with different characteristics, and different probability distributions.

We can treat this second test as a third random variable, and after obtaining its result we must compute the probability of SARS as conditional by both test results. In the general case the formula for: $P(\textit{SARS}|\text{Test}_1^{\oplus} \wedge \text{Test}_2^{\oplus})$ will consider the dependency between both test results. This means it is necessary to compute, in the presence of many random variables, a large number of probabilities, which, in theory, wipes out the advantages of using the conditional probability instead of the full JPD.

It is important to notice, however, that the results from both tests depend on the presence of the virus, but not from one another directly. Using this observation we may simplify the formulas, as it is necessary only to compute the conditional probability of the results of both tests.



$$\begin{aligned}
P(\textit{SARS}|T_1^{\oplus}, T_2^{\oplus}) &= \frac{P(\textit{SARS} \cap T_1^{\oplus} \cap T_2^{\oplus})}{P(T_1^{\oplus} \cap T_2^{\oplus})} \\
&= \frac{P(T_1^{\oplus} \cap T_2^{\oplus}|\textit{SARS})P(\textit{SARS})}{P(T_1^{\oplus} \cap T_2^{\oplus})} \\
&= \frac{P(T_1^{\oplus}|\textit{SARS})P(T_2^{\oplus}|\textit{SARS})P(\textit{SARS})}{P(T_1^{\oplus})P(T_2^{\oplus})}
\end{aligned}$$

If both tests had identical characteristics as in the previous example, then a positive result obtained from both test would indicate the probability of the disease of 0.009025, which is now almost a hundred times higher than in case of no information.

Belief networks

Joint probability distributions help us finding answers concerning the problem domain, but it is not easy to use them when there are many variables. Additionally, determining the probabilities for all the atomic events may be cumbersome unless we have performed comprehensive statistical testing.

As with the SARS virus example, we can build a graph of the actual random variable dependencies, and after determining their conditional probabilities we can efficiently compute probability of any event. More precisely, we will call a **belief network** (Bayesian network, probabilistic network) the following graph:

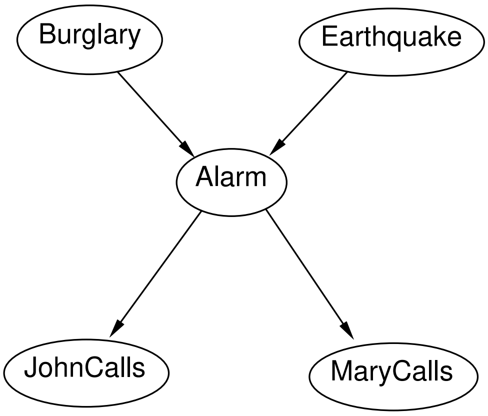
- the nodes of the network are random variables,
- the arcs of the network are directed, and an arc $X \longrightarrow Y$ has the intuitive meaning: "variable X has a direct influence on Y",
- each node X has an associated conditional probability table expressing the influence effected on X by its parents (predecessors in the graph),
- the network may not have directed cycles.

The network is constructed by first determining its topology, and the conditional probabilities for those nodes which have direct dependencies (incoming arcs).

The advantage of the belief networks consists in the relative ease, with which we can determine these direct dependencies. The probabilities of other events can be computed from the network.

Belief networks — an example

Let us consider an alarm system in a house: it reacts to break-ins, as well as, unfortunately, to minor earthquakes. The neighbors: John and Mary, are retired and stay at home. They have agreed to cooperate, and call us when they hear the alarm going off. John is very good at hearing the alarm, but is somewhat overreactive, and sometimes takes other events (such as telephone calls) for an alarm, and also calls then. Mary, on the other hand, recognizes the alarm correctly, but often listens to loud music, and sometimes fails to report.
What we are interested in is determining the probability of a burglary in progress at any given time, given who does, and who does not, call us.



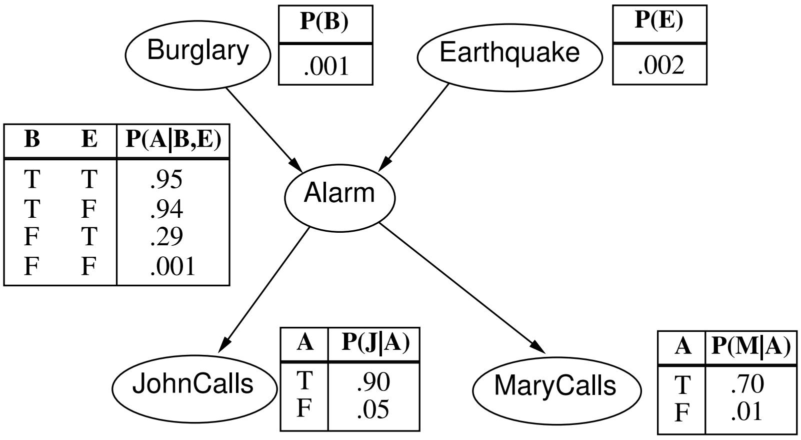
This network ignores some important details, such as whether Mary does listen to music at the specific moment of time, or whether there is an earthquake. Such details are cumbersome to determine precisely, but are accounted for — along with all other uncertainty — in the conditional probabilities of all the random variables.

In general, we must determine the probability distribution for all the random variables, conditional on other variables represented on our network. Specifically, we must provide the probability for each value of each random variable X for all the combinations of values of all random variables, on which X is conditional.

Burglary	*Earthquake*	$\mathbf{P}(\mathit{Alarm}\mid\mathit{Burglary},\mathit{Earthquake})$	
		True	False
True	True	0.950	0.050
True	False	0.940	0.060
False	True	0.290	0.710
False	False	0.001	0.999

The set of all such probabilities make up the **conditional probability table** (CPT). For all the variables that are not conditional we must determine the prior probability. In these cases the CPT has only one row with probabilities of different values of the random variable (adding up to 1.0).

The complete belief network for the example:



P(B)
.001

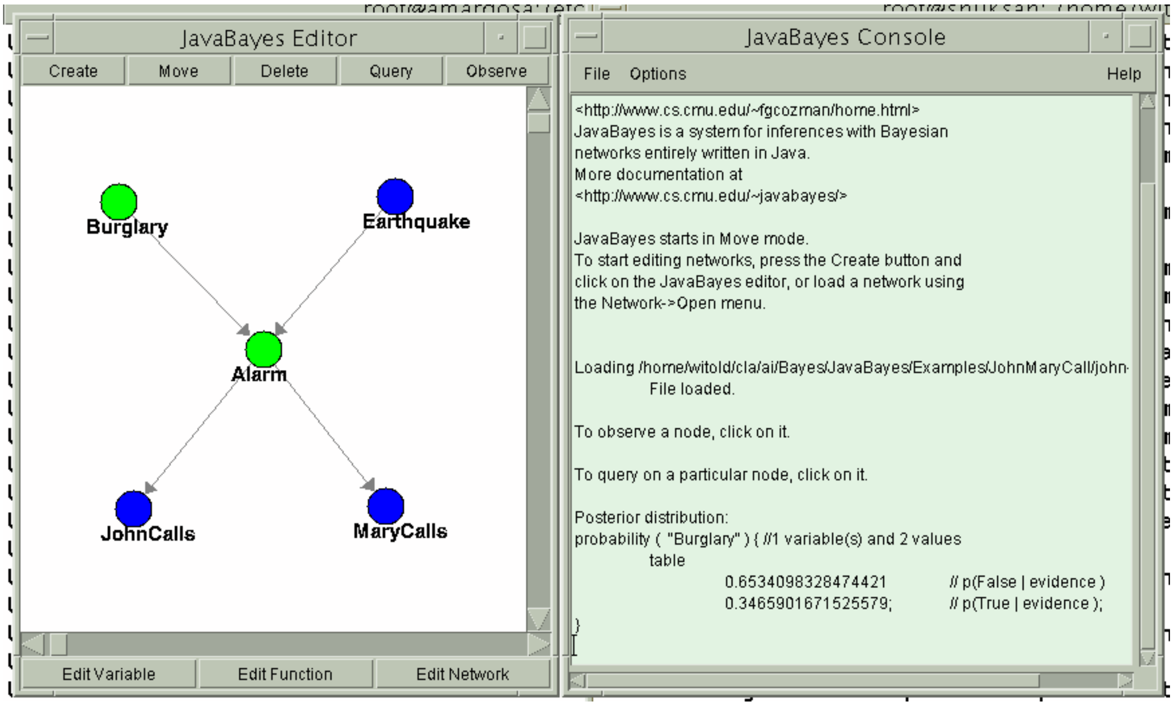
P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

Example network in JavaBayes



```
<http://www.cs.cmu.edu/~fgcozman/home.html>
JavaBayes is a system for inferences with Bayesian
networks entirely written in Java.
More documentation at
<http://www.cs.cmu.edu/~javabayes/>

JavaBayes starts in Move mode.
To start editing networks, press the Create button and
click on the JavaBayes editor, or load a network using
the Network->Open menu.


Loading /home/witold/cla/ai/Bayes/JavaBayes/Examples/JohnMaryCall/john-
            File loaded.

To observe a node, click on it.

To query on a particular node, click on it.

Posterior distribution:
probability ( "Burglary" ) { //1 variable(s) and 2 values
            table
                        0.6534098328474421          // p(False | evidence )
                        0.3465901671525579;         // p(True | evidence );
}
```