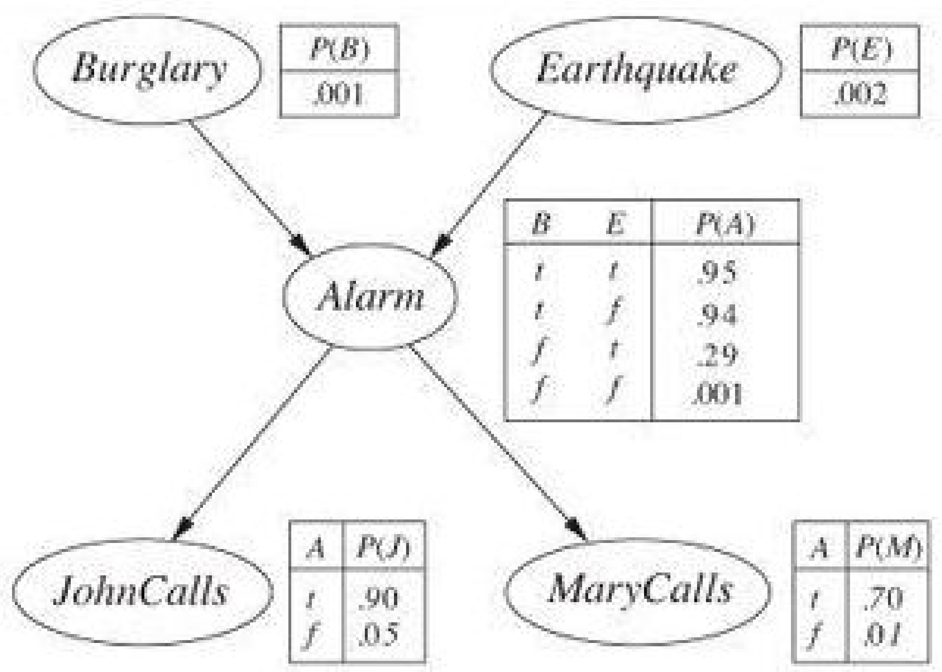
Bayesian Probability

What is a Bayesian Network?

A Bayesian network is a directed graph in which each node is annotated with quantitative probability information. The set of nodes and links specifies the conditional independence relationships.
Example:



Why should you care?

Bayesian Networks should tickle your interest because of their vast potential for use in everyday situations. They can predict the probability for things such as weather on wedding day:
(http://stattrek.com/probability/bayes-theorem.aspx)

Marie is getting married tomorrow, at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year. Unfortunately, the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time. When it doesn't rain, he incorrectly forecasts rain 10% of the time. What is the probability that it will rain on the day of Marie's wedding?

Or finding the source of defective products:
(https://onlinecourses.science.psu.edu/stat414/node/43)

A desk lamp produced by The Luminar Company was found to be defective (D). There are three factories (A, B, C) where such desk lamps are manufactured. A Quality Control Manager (QCM) is responsible for investigating the source of found defects. This is what the QCM knows about the company's desk lamp production and the possible source of defects:

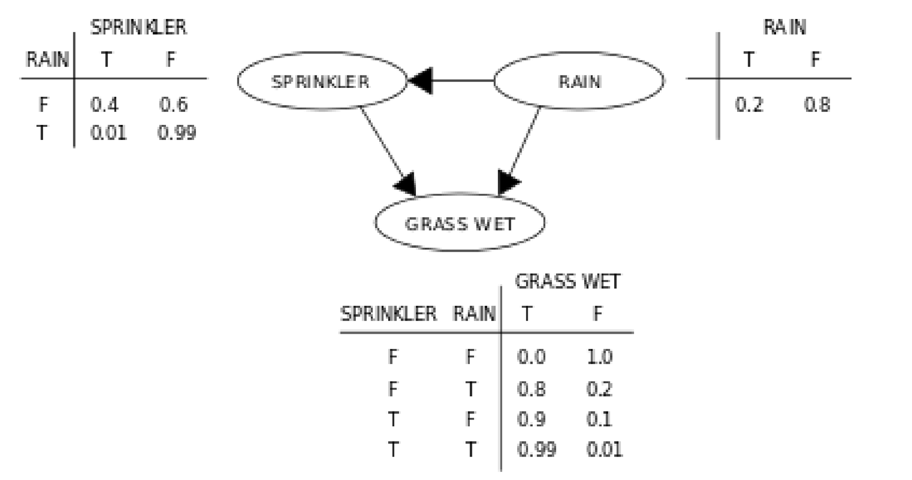
Factory	% of total production	Probability of defective lamps
A	$0.35 = P(A)$	$0.015 = P(D \mid A)$
B	$0.35 = P(B)$	$0.010 = P(D \mid B)$
C	$0.30 = P(C)$	$0.020 = P(D \mid C)$

The QCM would like to answer the following question: If a randomly selected lamp is defective, what is the probability that the lamp was manufactured in factory C?

These are two simple examples to show the depth of potential for this inference based probability network. Let's dive into how this future-telling magic actually works.

Let's get technical. (source: wikipedia)

Suppose that there are two events which could cause grass to be wet: either the sprinkler is on or it's raining. Also, suppose that the rain has a direct effect on the use of the sprinkler (namely that when it rains, the sprinkler is usually not turned on). Then the situation can be modeled with a Bayesian network (shown to the right). All three variables have two possible values, T (for true) and F (for false).



The joint probability function is:

$$\Pr(G,S,R) = \Pr(G|S,R)\Pr(S|R)\Pr(R)$$

where the names of the variables have been abbreviated to G = Grass wet (yes/no), S = Sprinkler turned on (yes/no), and R = Raining (yes/no).

The model can answer questions like "What is the probability that it is raining, given the grass is wet?" by using the conditional probability formula and summing over all nuisance variables:

$$\Pr(R=T|G=T) = \frac{\Pr(G=T,R=T)}{\Pr(G=T)} = \frac{\sum_{S\in\{T,F\}}\Pr(G=T,S,R=T)}{\sum_{S,R\in\{T,F\}}\Pr(G=T,S,R)}$$

Using the expansion for the joint probability function {\displaystyle \Pr(G,S,R)} {\displaystyle \Pr(G,S,R)} and the conditional probabilities from the conditional probability tables (CPTs) stated in the diagram, one can evaluate each term in the sums in the numerator and denominator. For example,

$$\begin{aligned}\Pr(G=T,S=T,R=T) &= \Pr(G=T|S=T,R=T)\Pr(S=T|R=T)\Pr(R=T)\\ &= 0.99\times 0.01\times 0.2\\ &= 0.00198.\end{aligned}$$

Then the numerical results (subscripted by the associated variable values) are

$$\Pr(R=T|G=T) = \frac{0.00198_{TTT} + 0.1584_{TFT}}{0.00198_{TTT} + 0.288_{TTF} + 0.1584_{TFT} + 0.0_{TFF}} = \frac{891}{2491} \approx 35.77\%.$$

Concluding that the probability that it is raining when the grass is wet is 35.77%

What we have done.

Using publically available CORGIS Datasets (https://think.cs.vt.edu/corgis/), we have sets of data (displayed on the next panel) categorized by state all surrounding a particular year. From these datasets, we have created a set of targets from conditionally independent relationships. One example of this is displayed on the following panel and showcases how education, income, and poverty correlate to murder rate. You can see that the murder rate is higher for people with a high school education or higher, houses with slighly lower income, and states with a lower poverty rate. This calculation was correct on 44/50 counts.

Make your own Bayesian combination!

Here are the data sets we have available, categorized by state:

Crime Data
- Assault Rate
- Murder Rate
- Rape Rate
- Robbery Rate
- Burglary Rate
- Larceny Rate
- Motor Rate

Education
- StateFunding
- Attendance Rate
- Bachelor's Degree or Higher
- High School or Higher

Ethnicity
- American Indian and Alaska Native Alone
- Asian Alone
- Black Alone
- Hispanic or Latino
- Native Hawaiian and Other Pacific Islander Alone
- Two or More Races
- White Alone

Wealth
- Median Household Income
- Persons Below Poverty Level

Example: How education, income, and poverty affect murder rate