Bayesian Networks

1. Bayesian Statistics

Bayesian statistics is the foundation for Bayesian networks. The difference between Bayesian statistics and Frequentist statistics is that Bayesian statistics would have a "prior" belief of the distribution of the data (or we could say Frequentist statistics always use uniform prior), and update that belief based on the data seen.

An example of using Bayesian statistics to solve probability problem: if there exists a type of test that doctors use to determine if a person have a disease. The probability that if a person has the disease and the test result is positive is 95%, that is: $P(Test = +ve \mid Disease = true) = 0.95$. And we also have the probability of the person have the disease and the test is negative (the test doesn't think that the person has the disease), which is $P(Test = -ve \mid Disease = true)$, is 5%. We also have the probability of the person doesn't the disease and the test come up positive (a false alarm), $P(Test = +ve \mid Disease = false) = 0.05$. Also, let's suppose that the disease is a really rare one, only 1 percent of overall population have this disease. So, our prior is $P(Disease = true) = 0.01$ (1%).

So, let's say if a person goes to the hospital and take the test, and the test result shows up positive, according to Bayesian statistics, what is the probability of the person actually have the disease?

To approach this problem, we will need Bayes' rule:

$$P(Y = y \mid X = x) = \frac{P(X = x \mid Y = y) * P(Y = y)}{\sum_{yi \in Y} P(X = x \mid Y = yi) * P(Y = yi)}$$

We can represent the denominator as:

$$P(X = x) = \sum_{yi \in Y} P(X = x \mid Y = yi) * P(Y = yi)$$

Which is called marginal likelihood. It is used to normalize the probability so that it sums up to 1 for different x.

The nominator is the joint probability of the independent variable $X = x$ and $Y = y$. In other words, it's the probability of the events $X = x$ and $Y = y$ happen at the same time:

$$P(X = x, Y = y) = P(X = x \mid Y = y) * P(Y = y) = P(Y = y \mid X = x) * P(X = x)$$

Now that we have all the math tools, we can start analyzing this problem. The probability we want to find is:

$$P(Disease = True \mid Test = Positive)$$

Which reads as: the probability of the person actually has the disease when the test result is positive.

Therefore, according to Bayes' rule, we need to put the corresponding number in the following equation:

$$P(Disease = T \mid Test = +)$$
$$= \frac{P(Test = + \mid Diease = T) * P(Diease = T)}{P(Test = + \mid Diease = T) * P(Diease = T) + P(Test = + \mid Diease = F) * P(Diease = F)}$$
$$= \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.05 \times 0.99} = 0.16101694915254236$$

The result is quite interesting. It tells us: if a person is tested positive for this disease, the probability of that person actually having the disease is 16.1%.

Now you may say that people might not take the test seriously because the result above. Well another way to look at it is: if a person is tested positive by this test, then he is 16 times more likely than general public to have the disease!

Now let's look at another interesting fact, from the equation above:

$$\frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.05 \times 0.99} \approx \frac{1\%}{6\%}$$

Notice the denominator is 6% and the nominator is 1%, and the difference is 5%. So what does that 5% mean?

Remember in the beginning we talked about a probability $P(Test = +ve \mid Disease = false) = 0.05$. That is, in general, we expected about 5% all general public to get tested positive without actually having the disease. Plus the 1% that actually have the disease, we 6% in total who got tested positive.

So, what is the difference between Frequentist statistics and Bayesian statistics for this problem? A Frequentist would assume uniform prior, that is, the probability of a person drawn from general public would have 50% to have the disease (that would assume 50% of general public have the disease). That is an unreasonable assumption, but that's what we can only do without the knowledge of any prior.

$$P(Disease = T \mid Test = +)$$
$$= \frac{P(Test = + \mid Diease = T) * P(Diease = T)}{P(Test = + \mid Diease = T) * P(Diease = T) + P(Test = + \mid Diease = F) * P(Diease = F)}$$
$$= \frac{0.95 \times 0.5}{0.95 \times 0.5 + 0.05 \times 0.5} = 0.95$$

Therefore, a Frequentist statistician would believe that if a person is tested positive, then that person has a 95% probability to have the disease.

So let's dig deeper in to the problem, let's consider this: what is the probability of the person having the disease when he was tested positive on two independent tests? (note the independence assumption is really important)

Therefore, the probability we want is:

$$P(Disease = T \mid Test1 = +, Test2 = +)$$

According to Bayes' rule, the above probability equals:

$$\frac{P(Test1 = + \mid Disease = True) * P(Test2 = + \mid Disease = True) * P(Disease = True)}{P(Test1 = +, Test2 = +)}$$

The denominator equals:

$$P(Test1 = + \mid Disease = True) * P(Test2 = + \mid Disease = True) * P(Disease = True) \\ + P(Test1 = + \mid Disease = True) * P(Test2 = + \mid Disease = False) * P(Disease = False)$$

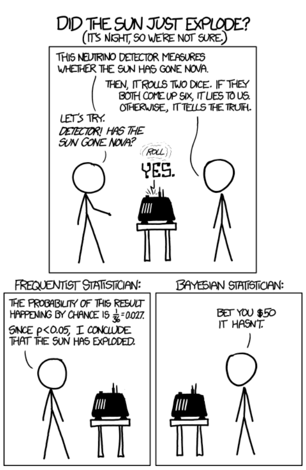
For the second part of the denominator, yes, it is possible for someone to get tested positive twice even that person does not have the disease (although that possibility is quite low).

Plugging all the numbers, this time we have:

$$P(Disease = T \mid Test1 = +, Test2 = +) = 0.7826$$

This tells us: if a person gets tested positive for both two of the independent tests, the probability of that person actually having the disease is 0.7826. As we can see, the posterior increased a lot.

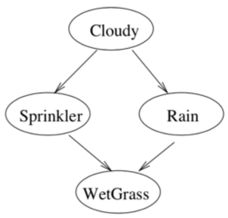
A piece of cartoon to show the difference between Frequentist thinking and Bayesian thinking:



2. Bayesian Networks

Bayesian network is a combination of graphical models and Bayes' rule and probability. It uses graph with directed edges, each node represents a random variable, and the directed edges captures the relationship between random variables.

P(C=F)	P(C=T)
0.5	0.5



C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

An advantage of Bayes' networks is each random variable can be inferred through its direct parents. For example, the probability of "the sprinkler is on when it's cloudy" can be calculated:

$$P(S = T \mid Cloudy = T) = \frac{P(Cloudy = T \mid S = T) * P(S = T)}{P(Cloudy = T \mid S = T) * P(S = T) + P(Cloudy = T \mid S = F) * P(S = F)}$$

And the probability of "the grass is wet given sprinkler is on and it's not raining:

$$P(W = T \mid S = T, R = F) = \frac{P(S = T, R = F \mid W = T) * P(W = T)}{P(S = T, R = F)}$$

The conditional probabilities can be obtained from the training set with either Laplace smoothing or M-estimate.