Bayesian Statistics Tutorial

Urinx

mkk@hust.edu.cn

July 1, 2017

Basics

conditional probabilities:

$$p(x|y) := \frac{p(x,y)}{p(y)}$$

the joint probalility of x and y:

$$p(x,y) = p(x|y)p(y) = p(y|x)p(x)$$

Theorem: Bayes Rule

Denote by X and Y random variables then the following holds

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

An Example

Question

Assume that a patient would like to have such a test carried out on him. The physician recommends a test which is guaranteed to detect HIV-positive whenever a patient is infected. On the other hand, for healthy patients it has a 1% error rate. That is, with probability 0.01 it diagnoses a patient as HIV-positive even when he is, HIV-negative. Moreover, assume that 0.15% of the population is infected.

Now the patient has the test carried out and the test returns HIV-negative. In this case, logic implies that he is healthy, since the test has 100% detection rate. In the converse case things are not quite as straightforward.

So what's the $p(X = \texttt{HIV+} | T = \texttt{HIV+})$?

An Example

$p(t\|x)$	$X = \texttt{HIV}-$	$X = \texttt{HIV}+$
$T = \texttt{HIV}-$	0.99	0
$T = \texttt{HIV}+$	0.01	1

$$p(X = \texttt{HIV}+) = 0.0015$$

An Example

By Bayes rule we may write

$$p(X = \texttt{HIV+}|T = \texttt{HIV+}) = \frac{p(T = \texttt{HIV+}|X = \texttt{HIV+})p(X = \texttt{HIV+})}{p(T = \texttt{HIV+})}$$

While we know all terms in the numerator, $p(T = \texttt{HIV+})$itself is unknown. That said, it can be computed via

$$\begin{aligned} p(T = \texttt{HIV+}) &= \sum_{x \in \{\texttt{HIV+}, \texttt{HIV-}\}} p(T = \texttt{HIV+}, x) \\ &= \sum_{x \in \{\texttt{HIV+}, \texttt{HIV-}\}} p(T = \texttt{HIV+}|x)p(x) \\ &= 1.0 \cdot 0.0015 + 0.01 \cdot 0.9985 \end{aligned}$$

Substituting back into the conditional expression yields

$$p(X = \texttt{HIV+}|T = \texttt{HIV+}) = \frac{1.0 \cdot 0.0015}{1.0 \cdot 0.0015 + 0.01 \cdot 0.9985} = 0.1306$$

How can we improve the diagnosis

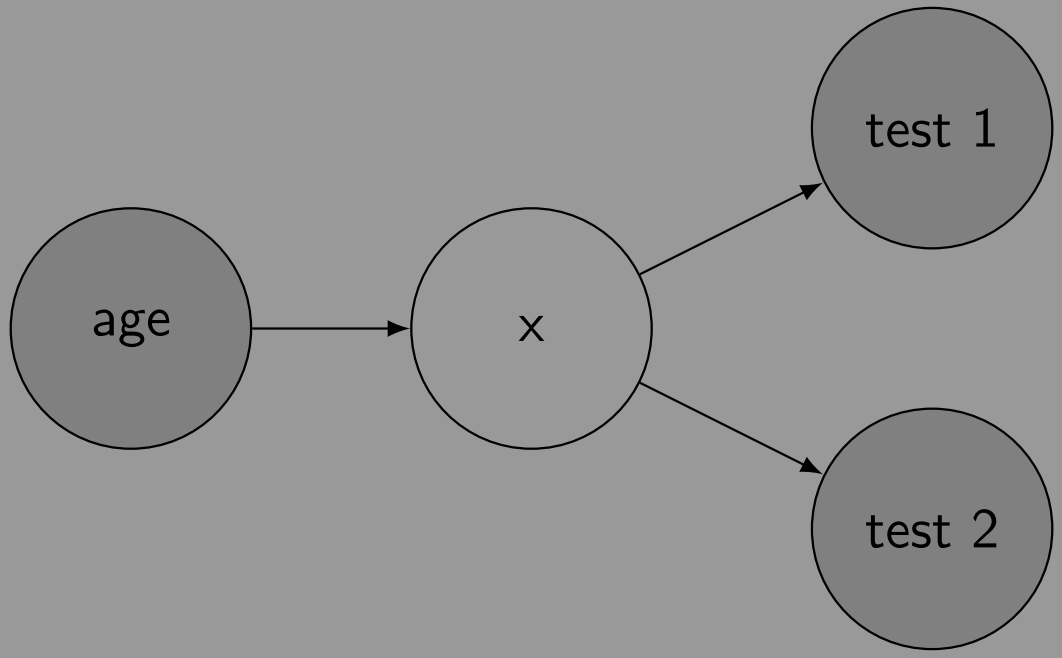

Figure: A graphical description of our HIV testing scenario. Knowing the age of the patient influences our prior on whether the patient is HIV positive (the random variable X). The outcomes of the tests 1 and 2 are independent of each other given the status X. We observe the shaded random variables (age, test 1, test 2) and would like to infer the un-shaded random variable X.

How can we improve the diagnosis

Including additional observed random variables

One way is to obtain further information about the patient and to use this in the diagnosis. For instance, information about his age is quite useful. Suppose the patient is 35 years old. In this case we would want to compute $p(X = \texttt{HIV+} | T = \texttt{HIV+}, A = 35)$ where the random variable A denotes the age.

The corresponding expression yields:

$$\frac{p(T = \texttt{HIV+} | X = \texttt{HIV+}, A) p(X = \texttt{HIV+} | A)}{p(T = \texttt{HIV+} | A)}$$

How can we improve the diagnosis

We may assume that the test is independent of the age of the patient, i.e.

$$p(t|x,a) = p(t|x)$$

What remains therefore is $p(X = \mathtt{HIV}+|A)$. Recent US census data pegs this number at approximately 0.9%.

$$p(X = \mathtt{H}+|T = \mathtt{H}+, A) = \frac{p(T = \mathtt{H}+|X = \mathtt{H}+, A)p(X = \mathtt{H}+|A)}{p(T = \mathtt{H}+|A)}$$

$$= \frac{p(T = \mathtt{H}+|X = \mathtt{H}+, A)p(X = \mathtt{H}+|A)}{p(T = \mathtt{H}+|X = \mathtt{H}+, A)p(X = \mathtt{H}+|A) + p(T = \mathtt{H}+|X = \mathtt{H}-, A)p(X = \mathtt{H}+}$$

$$= \frac{p(T = \mathtt{H}+|X = \mathtt{H}+)p(X = \mathtt{H}+|A)}{p(T = \mathtt{H}+|X = \mathtt{H}+)p(X = \mathtt{H}+|A) + p(T = \mathtt{H}+|X = \mathtt{H}-)p(X = \mathtt{H}-|A)}$$

$$= \frac{1 \cdot 0.009}{1 \cdot 0.009 + 0.01 \cdot 0.991} = 0.48$$

How can we improve the diagnosis

Multiple measurements

A second tool in our arsenal is the use of multiple measurements. After the first test the physician is likely to carry out a second test to confirm the diagnosis. We denote by T_1 and T_2 (and t_1,t_2 respectively) the two tests. Obviously, what we want is that T_2 will give us an "independent" second opinion of the situation.

What we want is that the diagnosis of T_2 is independent of that of T_2 given the health status X of the patient. This is expressed as

$$p(t_1, t_2|x) = p(t_1|x)p(t_2|x)$$

which are commonly referred to as conditionally independent.

How can we improve the diagnosis

we assume that the statistics for T_2 are given by

$p(t_2\|x)$	$X = \texttt{HIV}-$	$X = \texttt{HIV}+$
$T_2 = \texttt{HIV}-$	0.95	0.01
$T_2 = \texttt{HIV}+$	0.05	0.99

for $t_1 = t_2 = \texttt{HIV}+$ we have

$$
\begin{aligned}
& p(X = \texttt{H}+|T_1 = \texttt{H}+, T_2 = \texttt{H}+) \\
&= \frac{p(T_1 = \texttt{H}+, T_2 = \texttt{H}+|X = \texttt{H}+)p(X = \texttt{H}+|A)}{p(T_1 = \texttt{H}+, T_2 = \texttt{H}+|A)} \\
&= p(T_1 = \texttt{H}+|X = \texttt{H}+)p(T_2 = \texttt{H}+|X = \texttt{H}+)p(X = \texttt{H}+|A) \,/ \\
& p(T_1 = \texttt{H}+|X = \texttt{H}+)p(T_2 = \texttt{H}+|X = \texttt{H}+)p(X = \texttt{H}+|A) \\
&+ p(T_1 = \texttt{H}+|X = \texttt{H}-)p(T_2 = \texttt{H}+|X = \texttt{H}-)p(X = \texttt{H}-|A) \\
&= \frac{1 \cdot 0.99 \cdot 0.009}{1 \cdot 0.99 \cdot 0.009 + 0.01 \cdot 0.05 \cdot 0.991} = 0.95
\end{aligned}
$$