Solutions to Machine Learning: A Probabilistic Perspective

By:

Dustin Smith

Contents

Part I.

Foundations

1. Probabilistic Inference

1. Bayes rule for medical diagnosis

   After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease, and the test is 99% accurate (i.e., the probability of testing positive given that you have the disease is 0.99, as is the probability of testing negative given that you dont have the disease). The good news is that this rare disease, striking only one in 10,000 people. What are the chances that you actually have the disease?

   The probability matrix will be given by table 1.1. The probability of the prevalence of the disease is $p(H = 1) = 1/10000$ for the population. The probability that we are infected and tested positive is

$$\begin{aligned} p(H = 1 \mid Y = 1) &= \frac{p(Y = 1 \mid H = 1)p(H = 1)}{p(Y = 1 \mid H = 1)p(H = 1) + p(Y = 1 \mid H = 0)p(H = 0)} \\ &= \frac{0.99 \cdot 0.0001}{0.99 \cdot 0.0001 + 0.01 \cdot (1 - 0.0001)} \\ &= 0.0098 \text{ or } 0.98\% \end{aligned}$$

2. Legal reasoning

   Suppose a crime has been committed. Blood is found at the scene for which there is no innocent explanation. It is of a type which is present in 1% of the population.

   2.1. The prosecutor claims: "There is a 1% chance that the defendant would have the crime blood type if he were innocent. Thus there is a 99% chance he is guilty." This is known as the prosecutor's fallacy. What is wrong with this argument?

   We are given that the probability of prevalence is $P(H = 1) = 0.01$ so $P(H = 0) = 0.99$. Moreover, the the number of people in the community with this rare blood type is 8,000 and there is no innocent explanation for it. That is, the true negative rate, or the probability that person doesn't have this blood and is not guilty is 1. From the probability matrix in table 1.2, we have that true negative rate is 100% and the true positive rate is 1/8000 for the given blood type; therefore, the probability that we are guilty with the rate blood type is

$$\begin{aligned} p(H = 1 \mid Y = 1) &= \frac{p(Y = 1 \mid H = 1)p(H = 1)}{p(Y = 1 \mid H = 1)p(H = 1) + p(Y = 1 \mid H = 0)p(H = 0)} \\ &= \frac{1/8000 \cdot 1/100}{1/8000 \cdot 1/100 + 7999/8000 \cdot 99/100} \\ &\approx 0.000126\% \end{aligned}$$

   2.2. The defender claims: "The crime occurred in a city of 800,000 people. The blood type would found in approximately 8,000 people. The evidence has provided a probability of just 1 in 8,000 that the defendant is guilty, and thus has no relevance." This is known as the defender's fallacy. What is wrong with this argument?

	Y = 0	Y = 1
H = 0	0.99	0.01
H = 1	0.01	0.99

Table 1.1.: The probability matrix for the given rare disease.

	Y = 0	Y = 1
H = 0	1	0
H = 1	7999 / 8000	1 / 8000

Table 1.2.: The probability matrix for guilty and rare blood and vice versa.

As for the defenders argument, the probability the defendant has the rare blood and is not guilty is

$$\begin{aligned} p(H = 1 \mid Y = 0) &= \frac{p(Y = 0 \mid H = 1)p(H = 1)}{p(Y = 0 \mid H = 1)p(H = 1) + p(Y = 0 \mid H = 0)p(H = 0)} \\ &= \frac{7999/8000 \cdot 1/100}{7999/8000 \cdot 1/100 + 1 \cdot 99/100} \\ &\approx 1\% \end{aligned}$$

That is, $p(H = 1 \mid Y = 0) \approx 1\% > 1/8000 = 0.0125\%$.

3. Probabilities are sensitive to the form of the question that was used to generate the answer

My neighbor has two children. Assuming that the gender of a child is like a coin flip, it is most likely, *a priori*, that my neighbor has one boy and one girl, with probability 1/2. The other possibilities–two boys or two girls–have probabilities 1/4, respectively.

3.1. Suppose I ask him whether he has any boys, and he says yes. What is the probability that one child is a girl?

3.2. Suppose instead that I happen to see one of his children run by, and it is a boy. What is the probability that the other child is a girl?

4. Deriving the posterior predictive density for the healthy levels game

W will first consider one-dimensional "rectangles" (i.e., lines); since the dimensions are independent, we can easily generalize to 2d.

For convenience, we will follow the notation of Josh Tenenbaum's Ph.D. thesis. In particular, let $h = \theta$ be the unknown hypothesis or parameter vector, $\mathcal{H}$ be the set of possible hypotheses (rectangles), $\mathcal{H}_y$ be the set of hypotheses consistent with observation y (so the rectangles have to be big enough to capture y), and $\mathcal{H}_{\mathcal{D},y}$ be the set of hypotheses consistent with all the examples in $\mathcal{D}$ as well as with y.

The posterior predictive is given by $p(y \mid \mathcal{D}) = p(y \mid \mathcal{D})\ p(\mathcal{D})$, where

$$p(\mathcal{D}) = \int_{h \in \mathcal{H}} p(h)p(\mathcal{D} \mid h)dh \tag{1.1}$$

$$= \int_{h \in \mathcal{H}_{\mathcal{D}}} \frac{p(h)}{|h|^N} dh \tag{1.2}$$

where we used the fact that $p(\mathcal{D} \mid h) = 1/|h|^N$ if $h \in \mathcal{H}_{\mathcal{D}}$ and is 0 otherwise. Similarly, $p(y, \mathcal{D}) = \int_{h \in \mathcal{H}_{\mathcal{D},y}} p(h)/|h|^N dh$.

To derive the integral in equation (1.2), let us assume the maximum observed value is 0 (we can pick any maximum and recenter the data, since we assume a translation invariant prior). Then the right edge of the rectangle must lie past the data, so $\ell \geqslant 0$. Also, if r is the range spanned by the examples, then the left most data point is $-r$, so the left side of the rectangle must satisfy $\ell - s \leqslant -r$, where s is the size of the rectangle.

4.1. Using these assumptions, show that

$$p(\mathcal{D}) = \frac{1}{N(N-1)r^{N-1}}$$

Hint: use integration by parts

$$I = \int_a^b f(x)g'(x)dx = f(x)g(x)|_a^b - \int_a^b f'(x)g(x)dx$$

4.2. To compute $p(y, \mathcal{D})$, we just need to extend the range from r to $r + d$, where d is the distance from y to the closest observed example. Hence show that

$$p(y \mid \mathcal{D}) = \frac{1}{(1 + d/r)^{N-1}}$$