A probability primer

Bruno A. Olshausen

March 1, 2004

Abstract

The French mathematician Laplace declared that probability theory is "common sense reduced to calculation." In fact, it is just that - a way of numerically encoding our state of knowledge about variables in the world. Often the variables we care about are some measured data, and we wish to make inferences from this data in the face of uncertainty. But every waking moment, the brain is inundated with "data" in the form of activities impinging upon its sensory epithelium, and it must make inferences about what is "out there" in the world in order to guide appropriate actions. In this sense, probability theory provides a useful quantitative tool for understanding information processing in nervous systems. Here, we shall review the key ideas from probability theory that are commonly encountered in the study of the brain. Much of this material is adapted from the textbook by S. Ross, "A first course in probability theory."

Discrete probability

If we have a variable that takes on a discrete set of outcomes, such as a coin (heads or tails) or a pair of dice (numbers 1-12), then a probability may be assigned to each outcome. The probability assigned may be based on direct empirical data (e.g, by collecting a histogram of the states occupied by the variable over a long length of time), or it may reflect our inferred belief about states that the variable is likely to occupy (analogous to the way a horse-better evaluates the odds on different horses).

Axioms

The probability, P, that a discrete valued variable, X, occupies a specific state, x, is a number between zero and one:

$$0 \leq P(X = x) \leq 1\,. \tag{1}$$

The sum of the probabilities of all outcomes equals one:

$$\sum_x P(X = x) \;=\; 1\,. \tag{2}$$

From here on out we shall use the shorthand $P(x)$ to stand for $P(X = x)$.

Distributions

Uniform

The uniform distribution is the most trivial form of distribution, where all outcomes are equally likely:

$$P(x) = 1/n\,, \quad x = 1, ...n\,. \tag{3}$$

Bernoulli

A Bernoulli random variable is simply a binary variable (i.e., it occupies one of two states) with

$$P(1) = p \tag{4}$$
$$P(0) = 1 - p \tag{5}$$

Binomial

If we draw n independent samples from a Bernoulli variable, the total number of 1's that we get is a *binomial* random variable. The binomial distribution tells us that the probability of getting a total of i one's from n draws is

$$P(i) = \left(\begin{array}{c} n \\ i \end{array} \right) p^i \, (1-p)^{(n-i)}\,. \tag{6}$$

where p is as specified in (4,5). The notation $\left(\begin{array}{c} n \\ i \end{array} \right)$ neans "n choose i." It tells us the number of ways we could get i one's out of n draws:

$$\left(\begin{array}{c} n \\ i \end{array} \right) = \frac{n!}{(n-i)!i!}\,. \tag{7}$$

The probability of any particular outcome with i one's and $n-i$ zero's is $p^i \, (1-p)^{(n-i)}$. Since this can happen $\left(\begin{array}{c} n \\ i \end{array} \right)$ different ways, the total probability of getting i one's is thus given by equation 6.

Poisson

When n is large and p is small, one may approximate the binomial distribution with a Poisson distribution

$$P(i) = e^{-\lambda} \frac{\lambda^i}{i!} \tag{8}$$

where λ may be thought of as a rate parameter that corresponds to np in the binomial distribution, or in general the number of one's occurring in some interval. Among the variables that have been observed to have Poisson distributions are 1) the number of misprints on a page of a book, 2) the number of wrong telephone numbers dialed in a

day, 3) the number of α-particles discharged in a fixed period of time from radioactive material, and 4) the number of spikes discharged from a neuron in a fixed period of time, T, when T is less than 100-200 ms. In the latter case, it should be noted that when natural stimuli are used neurons tend *not* to be Poisson.

Joint distributions

The probability of two or more variables occupying a combined state, $X_1 = x_1$, $X_2 = x_2$,... $X_n = x_n$, is denoted by $P(x_1, x_2, ..., x_n)$ or $P(\mathbf{x})$. Such a distribution obeys the same axioms as above, $0 \leq P(\mathbf{x}) \leq 1$, and $\sum_{\mathbf{x}} P(\mathbf{x}) = 1$.

Conditional probability

We can express the interaction between variables using the conditional probability

$$P(x|y) = \frac{P(x,y)}{P(y)} \tag{9}$$

where the | notation means "given that." Thus, $P(x|y)$ refers to the probability that $X = x$ given that $Y = y$. It should make sense intuitively then that the probability of a joint state x, y, is just $P(x|y)$ multiplied by $P(y)$, which is another way of stating equation 9.

Factorial distribution

When a set of variables are statistically independent—i.e, the outcome of one variable has no effect on the others—then $P(x|y)$ will be the same as $P(x)$. In this case, the joint distribution is said to be *factorial*, meaning that $P(x, y)$ is given by the product of the distributions for each variable alone:

$$P(x,y) = P(x)P(y) \tag{10}$$

or more generally

$$\begin{aligned} P(x_1, x_2, ..., x_n) &= P(x_1) \times P(x_2) \times ... \times P(x_n) &(11)\\ &= \Pi_i P(x_i) &(12) \end{aligned}$$

Continuous variables

A variable that takes on a value along a continuum, such as voltage or light intensity, is assigned a *probability density function*, or p.d.f., which measures the amount of probability per unit of the variable. For example, the p.d.f. for the voltage on a car battery, $p(V)$, might be a bell-shaped function that is peaked at 12 volts with some spread on either side. The value of the function at a given point does not denote the probability of being exactly at that voltage (since the continuum of voltage is infinitely divisible), but rather the "probability per volt." If you want to know the probability

of the voltage being found between 11.9 and 12.1, then you would integrate $p(V)$ over that interval:

$$P(11.9 \leq V \leq 12.1) = \int_{11.9}^{12.1} p(V)dV \,. \tag{13}$$

Thus, if you want to speak of the probability of a continuous variable being at a certain value, you must necessarily specify a level of precision.

Axioms

We denote the p.d.f. of a continous random variable x as $p(x)$. An important distinction between a p.d.f. and the probability of a discrete variable is that a p.d.f. is bounded only below by zero, but is not bounded above

$$p(x) \geq 0 \,. \tag{14}$$

It must in any case integrate to one

$$\int_{-\infty}^{\infty} p(x)dx = 1 \,. \tag{15}$$

Distributions

Uniform

The most trivial form of p.d.f. is the uniform distribution, in which the variable has non-zero probability over a finite interval from a to b. The p.d.f. over this interval is then

$$p(x) = \frac{1}{b-a} \quad a \leq x \leq b \tag{16}$$

Normal (Gaussian)

Perhaps the most ubiquitous distribution of all is the normal distribution which forms the classic bell-shaped curve:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{17}$$

Although the normal distribution was introduced by the French mathematician de Moivre as an approximation to the binomial distribution when n is large, Gauss somehow managed to stamp his name on it. So a continuous variable distributed as in (17) is commonly referred to as a "Gaussian distributed" variable. The parameter μ sets the center or mean of the distribution, while the parameter σ sets its spread or variance[1]

The reason the normal distribution is so commonly used to describe natural phenomena is due to the *Central Limit Theorem*, which states that the sum of N random variables will tend to be normally distributed as $N \to \infty$.

[1] See the 10 Deutsche Mark note for illustration.

Exponential (Laplacian)

The exponential distribution

$$p(x) = \lambda e^{-\lambda x} \;, \quad x \geq 0 \tag{18}$$

is often used to describe the amount of time one must wait before an event occurs (such as an earthquake). In its two-sided form,

$$p(x) = \frac{\lambda}{2} e^{-\lambda |x|} \;, \quad -\infty \leq x \leq \infty \tag{19}$$

it is known as the Laplacian and is often used to model natural image statistics.

Function of a random variable

Let's say we have a random variable x with distribution $p_x(x)$. Now if another variable y is a deterministic function of x

$$y = f(x) \,, \tag{20}$$

what is the corresponding distribution of $p_y(y)$? The way to figure it out is shown in figure 1. The basic idea is that the area under the distribution must be preserved

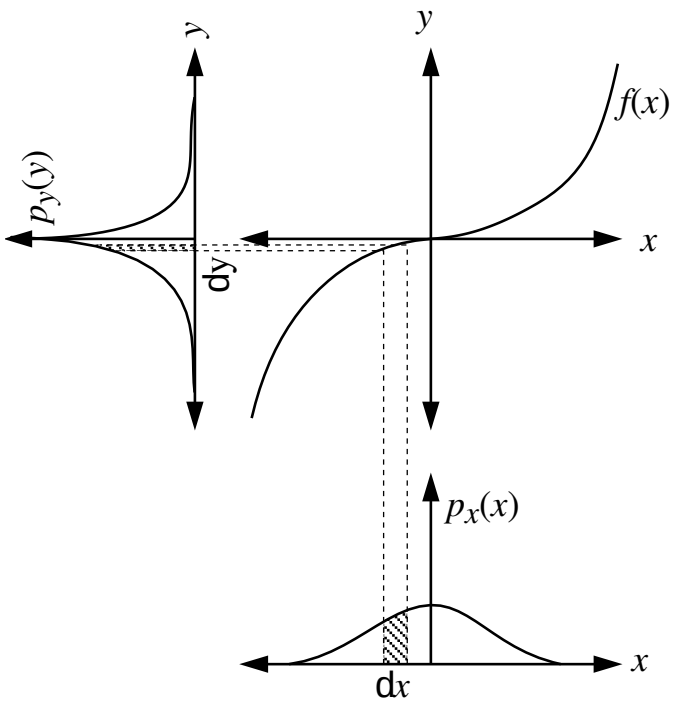

Figure 1: Distributions $p_x(x)$ and $p_y(y)$ must have equal area for corresponding intervals.

for corresponding intervals in x and y. In other words, if we integrate $p_x(x)$ over the interval $x_0 - \frac{\delta x}{2} \leq x \leq x_0 + \frac{\delta x}{2}$, we should get the same answer as when we integrate $p_y(y)$ over the interval $f(x_0 - \frac{\delta x}{2}) \leq y \leq f(x_0 + \frac{\delta x}{2})$:

$$\int_{x_0-\frac{\delta x}{2}}^{x_0+\frac{\delta x}{2}} p_x(x)dx = \int_{f(x_0-\frac{\delta x}{2})}^{f(x_0+\frac{\delta x}{2})} p_y(y)dy \tag{21}$$

or in the limit as $\delta x \to 0$ we have

$$p_x(x_0)\delta x = p_y(f(x_0))\delta f(x_0)\,. \tag{22}$$

Thus,

$$p_y(y) = p_x(x)\left[\frac{dy}{dx}\right]^{-1}\,. \tag{23}$$

In other words, you get the p.d.f. for y by simply taking the p.d.f. for x and weighting it by the inverse derivative of f. Note that this holds only if f is monotonic.

Generating random variables

Equation 23 suggests a method for drawing samples from an arbitrary distribution. Most computers are equipped with a random number generator that produces numbers uniformly distributed between 0 and 1. So, if we let $p_y(y)$ be the uniform distribution between 0 and 1 and $p_x(x)$ is the desired distribution, then we have

$$1 = p_x(x)[f']^{-1}\,. \tag{24}$$

Thus,

$$f(x) = \int_{-\infty}^{x} p_x(u)du\,. \tag{25}$$

In other words, if we take random numbers generated from the uniform [0,1] distribution, pass them through the inverse of the cumulative distribution for $p_x(x)$, what comes out are numbers that are distributed as though they came from $p_x(x)$!

Moments

A *moment* provides a way of characterizing the distribution of a random variable with a single number that is obtained by taking the expected value of a function of the variable. A few popular moments are described here.

Mean (first moment)

The mean, μ, of a distribution attempts to characterize the average value of a random variable drawn from the distribution:

$$\mu = E[x] \tag{26}$$

$$= \int p(x)\,x\,dx \tag{27}$$

where $E[]$ denotes "expected value." Note that for some distributions—e.g., a bimodal distribution—the mean does not in any sense characterize the *typical* value of a variable drawn from the distribution.

Variance (second moment)

The variance, σ^2, of a distribution attempts to characterize its spread:

$$\begin{aligned} \sigma^2 &= E[(x-\mu)^2] && (28) \\ &= \int p(x)\,(x-\mu)^2\,dx && (29) \end{aligned}$$

For a Gaussian distribution, the variance is simply given by the parameter σ^2 (by definition). A Poisson distribution has its variance equal to the mean, which is given by the parameter λ. Thus, a test that is typically applied in order to tell whether a random variable is consistent with a Poisson process is to calculate the ratio of variance to mean to see if it is one.

Skew (third moment)

The skew of a distribution attempts to characterize its lopsidedness:

$$\begin{aligned} \text{skew} &= \frac{1}{\sigma^3} E[(x-\mu)^3] && (30) \\ &= \frac{1}{\sigma^3} \int p(x)\,(x-\mu)^3\,dx\,. && (31) \end{aligned}$$

If the distribution is perfectly symmetric then the skew will be zero. But simply observing that the skew is zero does not necessarily imply the distribution is symmetric.

Kurtosis (proportional to fourth moment)

The kurtosis attempts to measure the peakedness of a distribution:

$$\begin{aligned} \kappa &= \frac{1}{\sigma^4} E[(x-\mu)^4] - 3 && (32) \\ &= \frac{1}{\sigma^4} \int p(x)\,(x-\mu)^4\,dx - 3\,. && (33) \end{aligned}$$

The reason for subtracting three is so that $\kappa = 0$ for a Gaussian distribution. A distribution with positive kurtosis may be more peaked or have heavier tails than a Gaussian, while a distribution with negative kurtosis may look more like a loaf of bread.

Principle of maximum entropy

More generally, we may take the expected value of any arbitrary function $\phi(x)$ in an attempt to characterize a distribution:

$$\begin{aligned} \alpha &= E[\phi(x)] && (34) \\ &= \int p(x)\,\phi(x)\,dx\,. && (35) \end{aligned}$$

What can we say about the distribution from the value obtained from the expected value of such an arbitrary function? The *principle of maximum entropy* states that if we have to guess a particular distribution, we should choose the one with maximum entropy that satisfies the constraints. If the constraints are of the form (35), then the distribution we choose should be of the form

$$p(x) = \frac{1}{Z_\lambda} e^{-\lambda\phi(x)} \tag{36}$$

where λ is choosen to satisfy the constraint (35), and Z_λ is a normalizing constant. For example, the Gaussian distribution is the maximum entropy distribution for a fixed variance.