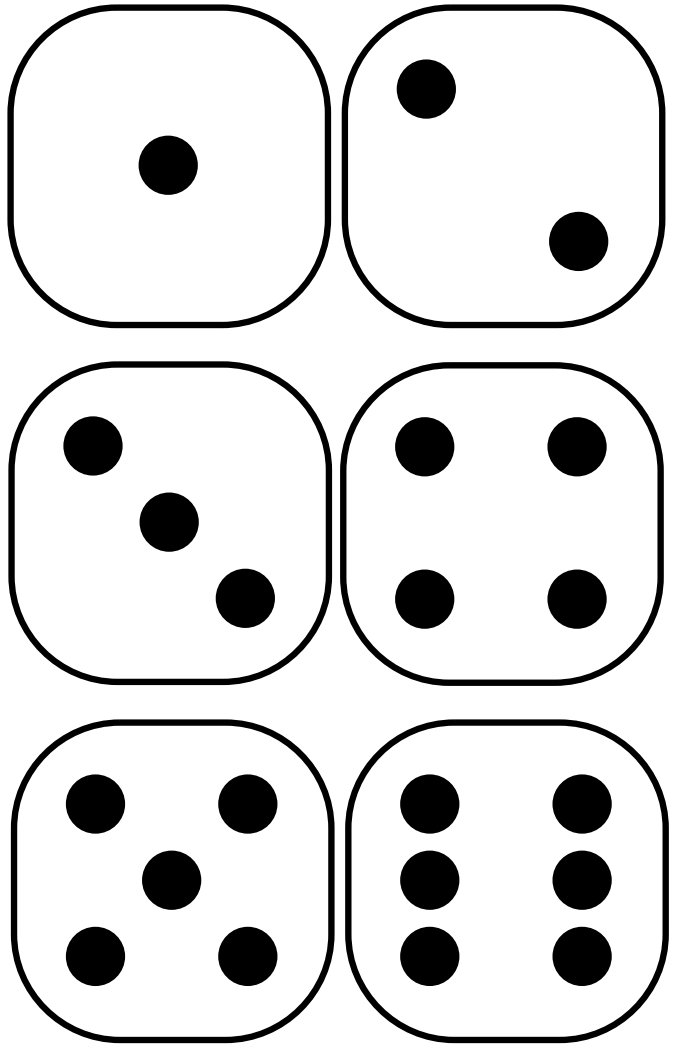
Statistics and Machine Learning 1

Lecture 1B: First Steps in Probability

Mark Muldoon
Department of Mathematics, Alan Turing Building
University of Manchester

Week 1

Prototypical examples



Axioms of probability

We'll consider *events* $A, B, \ldots$, which stand for things like "Rolling a six on a die." We'll write $\mathrm{P}(A)$ for the probability of event A, and the most fundamental rules of probability (or *axioms*) are, in ordinary language:

1. The probability that *some* event happens is 1. E.g.

$$\mathrm{P}(\text{Roll } 1, 2, 3, 4, 5 \text{ or } 6) = 1.$$

2. All probabilities must be positive or zero, i.e. $\mathrm{P}(A) \geq 0$ for any A.
3. If A and B are *disjoint* (they do not involve any of the same outcomes) then $\mathrm{P}(A \text{ or } B) = \mathrm{P}(A) + \mathrm{P}(B)$. For example,

$$\mathrm{P}(\text{Roll } 1 \text{ or Roll } 6) = \mathrm{P}(\text{Roll } 1) + \mathrm{P}(\text{Roll } 6).$$

These rules need to be formulated somewhat more carefully to deal with continuous random variables such as measurements like 'the time difference is 34.5 seconds'.

Rules of probability

- *Independence:* If A and B are *independent* of each other then

$$\mathrm{P}(A \text{ and } B) = \mathrm{P}(A) \times \mathrm{P}(B). \tag{1}$$

  Example: A is your coin toss and B is mine. The probability of us both getting tails is $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$.

- *Conditional probability:* Provided A is possible, then the probability of B *given* A is

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(B \text{ and } A)}{\mathrm{P}(A)}. \tag{2}$$

  Example: The probability of rolling 6 on a die given that the number rolled is even is $\frac{1}{3}$ and the probability of rolling 6 on a die given that the number rolled is odd is 0.

Rules of probability

- *Total Probability:* If we have a collection of events B_n (with $n = 1, 2, \ldots$) such that:
  - $\sum_n \mathrm{P}(B_n) = 1$ and
  - only one of the B_n can happen at a time, so that if, say, B_1 happens, then B_2 and B_3 can't,

  then

$$\mathrm{P}(A) = \sum_n \mathrm{P}(A|B_n)\mathrm{P}(B_n). \tag{3}$$

Example: total probability

A machine will shut itself off with probability 0.5 if the temperature T falls below 10 degrees, which happens with probability 0.01. It will also shut itself off with probability 0.3 if the temperature rises above 20 degrees, which has probability 0.001. At intermediate temperatures there is no probability of shutdown. The total probability of a shutdown is then:

$$\begin{aligned}\mathrm{P}(\text{Shutdown}) &= \mathrm{P}(\text{Shutdown} \mid T < 10) \times \mathrm{P}(T < 10) \\ &\quad + \mathrm{P}(\text{Shutdown} \mid 10 \leq T \leq 20) \times \mathrm{P}(10 \leq T \leq 20) \\ &\quad + \mathrm{P}(\text{Shutdown} \mid T > 20) \times \mathrm{P}(T > 20) \\ &= 0.5 \times 0.01 + 0 \times 0.989 + 0.3 \times 0.001 \\ &= 0.053\end{aligned}$$

where we've used

$$\mathrm{P}(T < 10) + \mathrm{P}(10 \leq T \leq 20) + \mathrm{P}(T > 20) = 1$$

so

$$\mathrm{P}(10 \leq T \leq 20) \;=\; 1 - \mathrm{P}(T < 10) + \mathrm{P}(T > 20) \;=\; 1 - 0.01 + 0.001.$$

Rules of probability

- *Bayes' Theorem:* This follows from other rules, and is used a lot in modern data science:

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\mathrm{P}(A)}{\mathrm{P}(B)}. \tag{4}$$

- We will consider probabilistic reasoning with data in great detail in Lecture 7.

Example: health screening

Many diseases (glaucoma, many cancers . . .) are assessed via *screening tests*, which are typically quick, inexpensive and somewhat imprecise ways to determine whether a patient should be referred for more expensive, decisive examination.

		Patient is actually:	
		Ill	*Healthy*
Test says:	*Ill*	True Positive (TP)	False Positive (FP)
	Healthy	False Negative (FN)	True Negative (TN)

Standard measures of the quality of a screening test include:

Sensitivity (sometimes also called *recall* or *true positive rate*) is the ratio $TP/(TP+FN)$. It is the probability with which the genuinely ill are detected correctly.

Specificity (sometimes also called the *true negative rate*) is the ratio $TN/(TN+FP)$. It is the probability with which genuinely healthy are detected correctly.

Health screening, Bayes rule & false alarms

Suppose 1% of all people are ill with a disease. Everyone is screened with a test that is 99% sensitive and specific (i.e. only 1% false positives and negatives). The probability of being ill given a positive test is:

$$\mathrm{P}(\mathrm{ill}|+) = \frac{\mathrm{P}(+|\mathrm{ill})\mathrm{P}(\mathrm{ill})}{\mathrm{P}(+|\mathrm{ill})\mathrm{P}(\mathrm{ill}) + \mathrm{P}(+|\mathrm{healthy})\mathrm{P}(\mathrm{healthy})}$$

$$= \frac{0.99 \times 0.01}{0.99 \times 0.01 + 0.01 \times 0.99} = \frac{1}{2}.$$

The key thing to remember about this example is:

Rare conditions are very hard to detect, even with excellent tests: true positives can be overwhelmed by false negatives.

Random variables and distribution functions

- A random variable X takes values, and we attribute probabilities to ranges of these values.
- The properties of random variables can be encoded in functions.
- The (cumulative) distribution function is the most fundamental such function, and its relationship to probabilities is:

$$\boxed{F(x) = \mathrm{P}(X \leq x)} \tag{5}$$

- Note that here we are assuming that all values are numbers and can be ordered. This is not actually a very restrictive assumption, provided we can encode complex data like images, sounds etc. numerically.

Types of random variable

- Most random variables are either continuous or discrete, although combinations of the two sometimes occur (can you think of an example?).
- We have to treat these cases somewhat differently.

Discrete Example

Let $X = 0$ if we get a tail from a coin toss and $X = 1$ if we get a head. We write

$$\mathrm{P}(X = 0) = \frac{1}{2}, \tag{6}$$

$$\mathrm{P}(X = 1) = \frac{1}{2}. \tag{7}$$

Continuous Example

Let X be the distance (in, say, cm.) a dart lands from the bullseye. We cannot attribute a non-zero probability to any *particular* value of X, but can make statements like

$$\mathrm{P}(1 < X < 2) = 0.233. \tag{8}$$

Probability (density) functions

Discrete Random Variables

We let

$$\mathrm{P}(X = x) = f(x). \tag{9}$$

Here, f is the probability function. Where the possible values of X are integers (we will assume this from now on) you will often see

$$\mathrm{P}(X = n) = p_n, \tag{10}$$

where p_n is called the *probability mass function* (pmf).

Continuous Random Variables

We let

$$\mathrm{P}(a \leq X \leq b) = \int_{x=a}^{x=b} f(x)\mathrm{d}x. \tag{11}$$

Here, f is the *probability density function* (pdf).
NB: $f(x) \geq 0$, but $f(x)$ does *not* need to be less than or equal to 1.

Relating distribution functions

Recall that the (cumulative) distribution function

$$F(x) = \mathrm{P}(X \leq x)$$

is defined the same way for both types of random variable. This is related to each type's probability function as follows:

Integer Random Variables

For any x,

$$F(x) = \sum_{n \leq x} p_n. \qquad (12)$$

Note that this will 'jump' at the integers.

Continuous Random Variables

Note that

$$F(b) - F(a) = \int_{x=a}^{x=b} f(x)\mathrm{d}x. \quad (13)$$

That is, f is the derivative of F.

Expectation

- The expectation of a function g of a random variable is written

$$\mathbb{E}[g(X)] \tag{14}$$

- It is calculated as follows:

Integer Random Variables

$$\mathbb{E}[g(X)] = \sum_n g(n)p_n. \tag{15}$$

Continuous Random Variables

$$\mathbb{E}[g(X)] = \int_{x=-\infty}^{x=\infty} g(x)f(x)\mathrm{d}x. \tag{16}$$

Examples:

- X is a die roll, $g(x) = 1$ if x is odd, $g(x) = 0$ if x is even: $\mathbb{E}[X] = 0.5$.
- X is the height of a English adult male picked blindfold from the phone book: $\mathbb{E}[X] = 175\,\mathrm{cm}$.

Summary statistics

- *Moments:* The kth moment is $\mathbb{E}[X^k]$.
- *Mean:* The mean is

$$\text{Mean}(x) = \mathbb{E}[X]. \tag{17}$$

- *Variance:* The variance is

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2. \tag{18}$$

- *Mode:* A mode is a value of x for which the probability function $f(x)$ is a local maximum—mathematically this means an x^* such that

$$\left.\frac{\mathrm{d}f}{\mathrm{d}x}\right|_{x^*} = 0, \qquad \left.\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\right|_{x^*} < 0.$$

- *Median:* The median is the value of x for which $F(x) = 0.5$.
- *Percentile:* The pth percentile is the value of x for which $F(x) = p/100$ (so the median is the 50th percentile).

Arithmetic with random variables

Suppose we have two random variables, X_1 and X_2 that are independent and have means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively.

- A random variable W defined by $W = X_1 + X_2$ has mean and variance given by

$$\mu_W = \mu_1 + \mu_2 \qquad \text{and} \qquad \sigma_W^2 = \sigma_1^2 + \sigma_2^2.$$

- A random variable Y defined by $Y = aX_1 + b$ has mean and variance given by

$$\mu_Y = a\mu_1 + b \qquad \text{and} \qquad \sigma_Y^2 = a^2\sigma_1^2.$$