Probability I: Reference

James Geddes

August 31, 2021

Ensembles

An *ensemble*, X, is a pair, $X = (\mathcal{A}_X, p_X)$, consisting of

1. A set, $\mathcal{A}_X$, called the *sample space*; and
2. A map, $p_x : \mathcal{A}_X \to [0, 1]$, called the *probability mass function*, such that

$$\sum_{a \in \mathcal{A}_X} p_X(a) = 1.$$

An element of the sample space is called an *outcome*. We sometimes say 'discrete probability distribution,' 'probability distribution,' or just 'distribution,' when we mean 'ensemble'.[1]

An *event* is a subset $E \subset \mathcal{A}_X$ of the sample space. It is events that have probabilities. In the general theory—not described here—it is not necessary that every subset of the sample space be an event; although in ensembles they generally are. The probability of an event, denoted $P_X(E)$, is given by

$$P_X(E) = \sum_{a \in E} p_X(a).$$

When the event contains just a single outcome, $E = \{a\}$, you may see the probability written as $P(X = a)$ (or $p(x = a)$) or sometimes simply $P(a)$.[2]

[1] The notion of an ensemble, along with much of my understanding, comes from MacKay (2003).

[2] The notation is pretty confusing. Perhaps an expression like $P(X = a)$ is intended to be read as 'the probability that a random variable, X, takes on the value a.' However, this begs the question of what a random variable is. I could not find a formal definition in either Bishop or Barber, although both texts use the term without discussion when introducing probability. Furthermore, most definitions say that a random variable is a (measurable) function on a probability space (and typically real-valued at that); whereas there is no function here, just the set of outcomes.

Even MacKay (from whom I took the idea of an ensemble in the first place) defines an ensemble as "a triple $(x, \mathcal{A}_X, \mathcal{P}_X)$ where the *outcome* x is the value of a random variable." (Emphasis as in the original.) I do not understand this definition: x cannot be the *value* of a random variable—which one would it be? It could *be* a random variable—but again, that is begging the question of what a random variable is and why we need the $\mathcal{P}_X$ as well. (On the other hand, it would not be the first time that I have realised that MacKay was saying something cleverer than I thought.)

Alternatively, perhaps the notation is a tacit acknowledgement that probabilities are defined not on outcomes but on events. For example, Gelman writes things like '$\Pr(\theta > 1)$' to mean 'the probability that θ is greater than 1.' Under this view, $P(X = a)$ would really mean '$P_X(\{x \in \mathcal{A}_X \mid x = a\})$,' only that seems too cumbersome to write.

Perhaps it is clearest to say that the argument of P really is an event (that is, a set of outcomes) and that any other notation is a shorthand for this.

Joint ensembles

Suppose $\mathcal{A}_X$ and $\mathcal{A}_Y$ are sets. A *joint ensemble* $(\mathcal{A}_X \times \mathcal{A}_Y, p_{X\times Y})$ is an ensemble whose sample space is $\mathcal{A}_X \times \mathcal{A}_Y$.

Let X and Y be two independent ensembles. An example of a joint ensemble is the *product ensemble*, which consists of:

1. The sample space $\mathcal{A}_X \times \mathcal{A}_Y$; and
2. The probability mass function $p_{X\times Y}(x, y) = p_X(x)p_Y(y)$.

Let $(\mathcal{A}_X \times \mathcal{A}_Y, p_{X\times Y})$ be a joint ensemble.

The *marginal distribution* on $\mathcal{A}_X$ is that given by the probability mass function

$$p(x) = \sum_y p_{X\times Y}(y),$$

and similarly for the marginal distribution $p(y)$ on $\mathcal{A}_Y$.

The *conditional distribution* is, for each $y \in \mathcal{A}_Y$, an ensemble on $\mathcal{A}_X$ given by the probability mass function

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

and similarly for the conditional $p(y \mid x)$ on $\mathcal{A}_Y$.

Facts about joint ensembles

The *chain rule* (or the 'product rule'):

$$P(x, y) = P(x \mid y)P(y).$$

Bayes' theorem (or 'Bayes' rule'):

$$P(x \mid y) = \frac{P(y \mid x)P(x)}{P(y)}.$$

Information

Let $(\mathcal{A}_X, p_X)$ be an ensemble.

The *information content* of an event $E \subset \mathcal{A}_X$ is

$$h(E) \equiv \log_2 \frac{1}{P_X(E)}.$$

The units of $h(E)$ are bits.[3] We usually consider only the information content, $h(a)$, of events, a, that are single outcomes:

$$h(a) = \log_2 \frac{1}{p_X(a)}.$$

[3] A bit is also called a 'Shannon.' If the logarithm is taken to base e, the units are 'nats;' if it is taken to base 10, the units are 'bans'.

The *Shannon entropy* of the ensemble is the expected information content of an outcome:

$$H(X) = \langle h \rangle_X = \sum_{a \in \mathcal{A}_X} p_X(a) \log_2 \frac{1}{p_X(a)}.$$

The entropy of a product ensemble, $(\mathcal{A}_X \times \mathcal{A}_Y, p_X \times p_Y)$ is

$$H(X \times Y) = H(X) + H(Y).$$

We denote the entropy of a joint ensemble as $H(X, Y)$. The conditional distribution, $p(x \mid y)$, is a distribution on $\mathcal{A}_X$, so we can compute its entropy—it is the *conditional entropy* of X given y (note the capitalisation):

$$H(X \mid y) = \sum_{x \in \mathcal{A}_X} p(x \mid y) \frac{1}{p(x \mid y)}.$$

The average of this over Y is the conditional entropy of X given Y:

$$\begin{aligned} H(X \mid Y) &= \sum_{y \in \mathcal{A}_y} p(y) \sum_{x \in \mathcal{A}_X} p(x \mid y) \frac{1}{p(x \mid y)} \\ &= \sum_{(x,y) \in \mathcal{A}_X \times \mathcal{A}_Y} p(x, y) \log_2 \frac{1}{p(x \mid y)}. \end{aligned}$$

The *chain rule for conditional entropy* follows from the chain rule:

$$H(X, Y) = H(Y) + H(X \mid Y).$$

Let $P = (\mathcal{A}, p)$ and $Q = (\mathcal{A}, q)$ be two ensembles over the same sample space. The *Kullback–Leibler divergence* of P from Q, is

$$\begin{aligned} D_{KL}(P \parallel Q) &\equiv \sum_{a \in \mathcal{A}} p(a) \Big(\log_2 \frac{1}{q(a)} - \log_2 \frac{1}{p(a)} \Big) \\ &= \sum_{a \in \mathcal{A}} p(a) \log_2 \frac{p(a)}{q(a)}. \end{aligned}$$

The Kullback–Leibler divergence is 'the average information content of a sample drawn from P, if we *thought* it came from Q,' less the actual entropy of P.

Here is a suggestive result. Consider a joint ensemble $X \times \Lambda$, where the joint distribution is written as

$$p(x, \lambda) = p(x \mid \lambda) p(\lambda),$$

and we think of λ as the model and x as the data in a Bayesian analysis. Write the posterior as $\Lambda \mid X$. Then:

$$D_{KL}((\Lambda \mid X) \parallel \Lambda) = H(X) - H(X \mid \Lambda).$$

My translation of this is 'the change from the prior to the posterior is the information given by the data, less the information about the data that was already in the prior.' (Note this is all on average, not for a particular observation.)

References

MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms Cambridge Univ.* Cambridge: Cambridge University Press. URL: http://www.inference.org.uk/mackay/itila/.