Machine Learning and Bayesian Inference
Some supplementary notes on probability

Sean B. Holden © 2020

1 Introduction

These notes provide a reminder of some simple manipulations that turn up a great deal when dealing with probabilities. The material in this handout—assuming you know it well—should suffice for getting you through most of the AI material on uncertain reasoning. In particular, the boxed results are the really important ones.

Random variables (RVs) are by convention given capital letters. Say we have the RVs $X_1, \ldots, X_n$. Their values are given using lower case. So for example X_1 might be a binary RV taking values `true` and `false`, and X_2 might be the outcome of rolling a die and therefore taking values $\{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$.

The use of probability in AI essentially reduces to representing in some usable way the joint distribution $\Pr(X_1, \ldots, X_n)$ of all the RVs our agent is interested in, because if we can do that then in principle we can compute *any* probability that might be of interest. (This is explained in full below.)

To be clear, the joint distribution is talking about the *conjunction* of the RVs. We'll stick to the convention that a comma-separated list of RVs (or a set of RVs) represents a conjunction. Also, the notation

$$\sum_{x_i \in X_i} (\ldots x_i \ldots)$$

denotes the sum over all *values* of a random variable. So for example if X_1 is binary then

$$\sum_{x_1 \in X_1} \Pr(x_1, X_2) = \Pr(\texttt{true}, X_2) + \Pr(\texttt{false}, X_2)) . \tag{1}$$

This sum can itself take on multiple values, one for each value of X_2. This all extends to summing over *sets* of RVs. Let's define

$$\mathbf{X} = \{X_1, \ldots, X_n\}$$

and the subset

$$\mathbf{X}' = \{X'_1, \ldots, X'_m\} \subseteq \mathbf{X}.$$

Then for any sets $\mathbf{X}$ and $\mathbf{X}' \subseteq \mathbf{X}$ of RVs define $\mathbf{X} \backslash \mathbf{X}'$ to be the set $\mathbf{X}$ with the elements of $\mathbf{X}'$ removed

$$\mathbf{X} \backslash \mathbf{X}' = \{X \in \mathbf{X} | X \notin \mathbf{X}'\}.$$

We'll always be assuming that $\mathbf{X}' \subseteq \mathbf{X}$. Finally

$$\sum_{x' \in \mathbf{X}'} (\ldots, x'_1, \ldots, x'_m, \ldots)$$

means

$$\sum_{x'_1 \in X'_1} \sum_{x'_2 \in X'_2} \cdots \sum_{x'_m \in X'_m} (\ldots, x'_1, \ldots, x'_m, \ldots) ,$$

and is itself a function of the RVs in $\mathbf{X} \backslash \mathbf{X}'$.

2 Standard trick number 1: marginalising

Marginalising is the process of getting rid of RVs that we don't want to have to think about—although in some cases it's used the other way around to introduce RVs. In general, say we want to ignore X_i. Then

$$\Pr\left(\mathbf{X}\backslash\{X_i\}\right) = \sum_{x_i \in X_i} \Pr\left(\mathbf{X}\right).$$

So for example with $\mathbf{X} = \{X_1, X_2\}$, equation 1 is actually telling us that

$$\begin{aligned}\Pr\left(X_2\right) &= \Pr\left(\mathbf{X}\backslash\{X_1\}\right) \\ &= \sum_{x_1 \in X_1} \Pr\left(x_1, X_2\right) \\ &= \Pr\left(\texttt{true}, X_2\right) + \Pr\left(\texttt{false}, X_2\right).\end{aligned}$$

This can obviously be iterated for as many RVs as we like, so if $\mathbf{X}'$ is the set of random variables we're not interested in then

$$\Pr\left(\mathbf{X}\backslash\mathbf{X}'\right) = \sum_{x' \in \mathbf{X}'} \Pr\left(\mathbf{X}\right).$$

These notes assume for the most part that RVs are discrete. Everything still applies when continuous RVs are involved[1], but sums are then replaced by integrals. For example, we can marginalise the two-dimensional Gaussian density

$$p(x_1, x_2) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\left(x_1^2 + x_2^2\right)\right)$$

as follows

$$p(x_1) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(x_1^2 + x_2^2\right)\right) dx_2.$$

(And it turns out that this is itself Gaussian, as we shall see in the lectures.)

3 Standard trick number 2: you can treat a conjunction of RVs as an RV

When we consider events such as $X_1 = \texttt{true}$ and $X_2 = \boxdot$, the *conjunction* of the events is also an event. This goes for any number of events, and any number of RVs as well. Why is that interesting? Well, Bayes' theorem usually looks like this

$$\Pr\left(X|Y\right) = \frac{\Pr\left(Y|X\right)\Pr\left(X\right)}{\Pr\left(Y\right)}.$$

However as a conjunction of RVs can be treated as an RV we can also write things like

$$\Pr\left(X_1, X_5|X_2, X_3, X_{10}\right) = \frac{\Pr\left(X_2, X_3, X_{10}|X_1, X_5\right)\Pr\left(X_1, X_5\right)}{\Pr\left(X_2, X_3, X_{10}\right)}$$

and Bayes' theorem still works.

[1] A word of caution here. If one wishes to be fully rigorous in dealing with probabilities then care is required in referring to a probability *distribution* or *density*, whether certain items are *measurable* and so on. In most machine learning material, such things tend to be taken for granted without incident.

4 Standard trick number 3: conditional distributions are still distributions

This is perhaps the point I want to make that's most often missed: *a conditional probability distribution is still a probability distribution.* Consequently the first two tricks extend to them without any extra work—you simply apply them while leaving the conditioning RVs (the ones on the right hand side of the | in $\Pr(\ldots|\ldots)$) alone. So, for instance, we can write

$$\Pr(X_1|X_3) = \sum_{x_2 \in X_2} \Pr(X_1, X_2|X_3)$$

or in general for sets of RVs

$$\Pr(\mathbf{X}|\mathbf{Z}) = \sum_{\mathbf{y} \in \mathbf{Y}} \Pr(\mathbf{X}, \mathbf{Y}|\mathbf{Z}).$$

Quite often this trick is used to *introduce* extra RVs in $\mathbf{Y}$ rather than eliminate them. The reason for this is that you can then try to re-arrange the contents of the sum to get something useful. In particular you can often use the following further tricks.

Just as marginalisation still works for conditional distributions, so do Bayes' theorem and related ideas. For example, the definition of a conditional distribution looks like this

$$\Pr(X|Y) = \frac{\Pr(X, Y)}{\Pr(Y)} \tag{2}$$

so

$$\Pr(X, Y) = \Pr(X|Y)\Pr(Y).$$

As the left-hand side of this equation is a joint probability distribution, and conjunctions of RVs act like RVs, we can extend this to arbitrary numbers of RVs to get, for example

$$\begin{aligned}\Pr(X_1, X_2, X_3) &= \Pr(X_1|X_2, X_3)\Pr(X_2, X_3) \\ &= \Pr(X_1|X_2, X_3)\Pr(X_2|X_3)\Pr(X_3).\end{aligned}$$

What's more useful however is to note that Bayes' theorem is obtained from equation 2 and its twin

$$\Pr(Y|X) = \frac{\Pr(X, Y)}{\Pr(X)}$$

by a simple re-arrangement. How might this work if we have conjunctions of random variables? Consider

$$\Pr(X|Y, Z) = \frac{\Pr(X, Y, Z)}{\Pr(Y, Z)}$$

and its twin

$$\Pr(Y|X, Z) = \frac{\Pr(X, Y, Z)}{\Pr(X, Z)}$$

both of which follow from the definition of conditional probability. Re-arranging to eliminate the $\Pr(X, Y, Z)$ gives

$$\Pr(X|Y, Z) = \frac{\Pr(Y|X, Z)\Pr(X, Z)}{\Pr(Y, Z)}.$$

We now have two smaller joint distributions $\Pr(Y, Z)$ and $\Pr(X, Z)$ which we can split to give

$$\begin{aligned}\Pr(X|Y, Z) &= \frac{\Pr(Y|X, Z)\Pr(X|Z)\Pr(Z)}{\Pr(Y|Z)\Pr(Z)} \\ &= \frac{\Pr(Y|X, Z)\Pr(X|Z)}{\Pr(Y|Z)}\end{aligned}$$

or in general, with sets of RVs

$$\Pr(\mathbf{X}|\mathbf{Y}, \mathbf{Z}) = \frac{\Pr(\mathbf{Y}|\mathbf{X}, \mathbf{Z})\Pr(\mathbf{X}|\mathbf{Z})}{\Pr(\mathbf{Y}|\mathbf{Z})}. \tag{3}$$

A word of warning. As conditional distributions are still distributions, it must always be the case that

$$\sum_{\mathbf{x}\in\mathbf{X}} \Pr(\mathbf{X}|\mathbf{Y}) = 1$$

regardless of the value of $\mathbf{Y}$. It is *not* however necessarily the case that

$$\sum_{\mathbf{y}\in\mathbf{Y}} \Pr(\mathbf{X}|\mathbf{Y}) = 1.$$

Do not get this the wrong way around!

5 How to (in principle) compute absolutely anything

Say you want to compute a conditional probability $\Pr(\mathbf{X}|\mathbf{Z})$. By definition

$$\Pr(\mathbf{X}|\mathbf{Z}) = \frac{\Pr(\mathbf{X}, \mathbf{Z})}{\Pr(\mathbf{Z})}$$

and if the complete collection of all the RVs our agent is interested in is $\{\mathbf{X}, \mathbf{Y}, \mathbf{Z}\}$ then both the numerator and the denominator can be computed by marginalising the joint distribution $\Pr(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$. In fact as the denominator serves essentially just to make the left hand side sum to 1 (when we sum over $\mathbf{X}$) so that it's a proper probability distribution, we often treat it just as a constant and write

$$\Pr(\mathbf{X}|\mathbf{Z}) = \frac{1}{Z}\sum_{\mathbf{y}\in\mathbf{Y}} \Pr(\mathbf{X}, \mathbf{Y}, \mathbf{Z}).$$

The quantity Z is called the *partition function* if you're a physicist or *evidence* if you're a computer scientist, for reasons that will become clear during the lectures. Clearly

$$Z = \sum_{\mathbf{x}\in\mathbf{X},\mathbf{y}\in\mathbf{Y}} \Pr(\mathbf{X}, \mathbf{Y}, \mathbf{Z}).$$

6 Further tricks

We now look at some further simple manipulations that are needed to understand the application of Bayes' theorem to supervised learning. Once again, random variables are assumed to be discrete, but all the following results still hold for continuous random variables, with sums replaced by integrals where necessary.

6.1 Some (slightly) unconventional notation

In the machine learning literature there is a common notation intended to make it easy to keep track of which random variables and which distributions are relevant in an expression. While this notation is common within the field, it's rarely if ever seen elsewhere; it is however very useful.

A statistician would define the *expected value* of the random variable X as

$$\mathbb{E}[X] = \sum_{x \in X} x \Pr(x)$$

or when we're interested in the expected value of a function of a random variable

$$\mathbb{E}[f(X)] = \sum_{x \in X} f(x) \Pr(x)$$

where f is some function defined on X. Here, it is implicit that the RV X has some probability distribution, which we will denote by $P(X)$. With complex expressions involving combinations of functions defined on random variables with multiple underlying distributions it can be more tricky to keep track of which distributions are relevant. Thus the notation

$$\mathbb{E}_{x \sim P(X)}[f(X)]$$

is intended to indicate explicitly that the distribution of X is $P(X)$, in situations where we don't write out the full definition

$$\mathbb{E}_{x \sim P(X)}[f(X)] = \sum_{x \in X} f(x) \Pr(x)$$

to make it clear. The same notation is also often applied to statements about probabilities rather than expected values.

6.2 Expected value and conditional expected value

The standard definition of the expected value of a function f of a random variable X is

$$\mathbb{E}_{x \sim P(X)}[f(X)] = \sum_{x \in X} f(x) \Pr(x)$$

as already noted. We can also define the *conditional expected value* of $f(X)$ given Y as

$$\mathbb{E}_{x\sim P(X|Y)}\left[f(X)|Y\right] = \sum_{x\in X} f(x)\Pr\left(x|Y\right).$$

Now here's an important point: *the value of this expression depends on the value of* Y. Thus, the conditional expected value is itself a function of the random variable Y. What is its expected value? Well

$$\begin{aligned}
\mathbb{E}_{y\sim P(Y)}\left[\mathbb{E}_{x\sim P(X|Y)}\left[f(X)|Y\right]\right] &= \sum_{y\in Y}\mathbb{E}_{x\sim P(X|Y)}\left[f(X)|Y\right]\Pr\left(y\right)\\
&= \sum_{y\in Y}\sum_{x\in X} f(x)\Pr\left(x|y\right)\Pr\left(y\right)\\
&= \sum_{y\in Y}\sum_{x\in X} f(x)\Pr\left(x,y\right)\\
&= \sum_{x\in X} f(x)\sum_{y\in Y}\Pr\left(x,y\right)\\
&= \sum_{x\in X} f(x)\Pr\left(x\right)\\
&= \mathbb{E}_{x\sim P(X)}\left[f(X)\right]
\end{aligned}$$

or in the more usual notation

$$\mathbb{E}\left[\mathbb{E}\left[f(X)|Y\right]\right] = \mathbb{E}\left[f(X)\right].$$

6.3 Expected value of the indicator function

For any $b \in \{\texttt{true}, \texttt{false}\}$ the *indicator function* $\mathbb{I}$ is defined as

$$\mathbb{I}\left[b\right] = \begin{cases} 1 & \text{if } b = \texttt{true} \\ 0 & \text{if } b = \texttt{false} \end{cases}.$$

Let f be a Boolean-valued function on a random variable X. Then

$$\begin{aligned}
\mathbb{E}_{x\sim P(X)}\left[\mathbb{I}\left[f(x)\right]\right] &= \sum_{x\in X}\mathbb{I}\left[f(x)\right]\Pr\left(x\right)\\
&= \sum_{x\in X, f(x)\text{ is }\texttt{true}}\mathbb{I}\left[f(x)\right]\Pr\left(x\right) + \sum_{x\in X, f(x)\text{ is }\texttt{false}}\mathbb{I}\left[f(x)\right]\Pr\left(x\right)\\
&= \sum_{x\in X, f(x)\text{ is }\texttt{true}}\Pr\left(x\right)\\
&= P_{x\sim P(x)}\left[f(x) = \texttt{true}\right]
\end{aligned}.$$

In other words, *the probability of an event is equal to the expected value of its indicator function.* This provides a standard method for calculating probabilities by evaluating expected values. So for example if we roll a fair die and consider $f(X)$ to be true if and only if the outcome is even then

$$\Pr\left(\text{outcome is even}\right) = \mathbb{E}\left[\mathbb{I}\left[f(X)\right]\right] = 1/6 + 1/6 + 1/6 = 1/2$$

as expected.