Lecture 12: RANDOM VARIABLES [*]

10-606 MATHEMATICAL FOUNDATIONS FOR MACHINE LEARNING

1 Random variables

A *random variable* is a function that takes atomic events as inputs. For example, if we roll three dice, the total number of spots is a random variable: our universe is the set of 3-tuples of numbers in $1:6$, and the total number of spots is a function that maps each of these 6^3 tuples to an integer in $3:18$, e.g., it maps $(4,2,3)$ to 9.

We can have random variables of any type: integer, real, boolean, complex, etc... The type of a random variable is the type of the function's output, e.g., a boolean random variable is a function in $U \to \{T, F\}$.

Events are the simplest random variables: we can think of an event as a function that returns T or 1 if the event happens, and F or 0 otherwise. Other examples include:

- If our atomic events are people, then the height of a person is a random variable. So are height[2] and eye color.
- If our atomic events are complex numbers, then the real part of the next complex number is a random variable.
- If our atomic events are cards that we draw from a deck, then the number of pips on the next card is a random variable; so are the suit and whether the card is a face card.

1.1 Random variables and events

An important use of random variables is to define new events. For example, we could make an event for whether someone's height is greater than 150cm, or whether the next five cards drawn form a straight flush.

Sometimes we go so far as to forget about the underlying probability space, and talk only about random variables. For example, we might talk about probabilities related to height or eye color without mentioning that the underlying atomic events are people. It's important to remember that there still is an underlying probability space, and that all of our statements about probabilities implicitly refer to this space.

1.2 Functions of a random variable

If X is a random variable and f is a function, then $f(X)$ is a random variable too. This is just composition of functions: remember that X is a function of the underlying atomic event $\omega \in U$, so $f(X) = f(X(\omega)) = (f \circ X)(\omega)$.

For example, we could ask for the square of the number of spots on our die. Here our underlying events are the different possible die rolls: $\omega \in U$ where U is the set $\{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$. The function X reads off the number of spots, e.g., $X(\boxdot) = 3$. Then, the function f squares it, so $f(X(\boxdot)) = 9$.

This works for multiple random variables too: if X and Y are random variables and f is a two-argument function, then $f(X, Y)$ is also a random variable. We can define a new function $Z(\omega)$ to represent this

[*]Compiled on Wednesday 21st June, 2023 at 03:47

random variable: $Z = f(X, Y)$ means $Z(\omega) = f(X(\omega), Y(\omega))$.

Often we won't give a separate name to the function of a random variable, we just write something like $X^2 + \cos(3Y)$. The value of this expression is a random variable, since it depends on which atomic event happens: we can write it as $X(\omega)^2 + \cos(3Y(\omega))$.

2 Joint distribution

It is common to define several different random variables for the same probability space, e.g., the height, eye color, and favorite food of each person.

We can describe this situation with a table. The rows correspond to atomic events. There is one column for the probability of the event, and optionally one column for its name. Then there is one column for each random variable, listing its value under each atomic event. Here's an example with two random variables:

p	eye color	height
0.01	blue	150cm
0.03	brown	155cm
0.02	green	152cm
⋮	⋮	⋮

Even if nobody tells us what probability space is being used to define the random variables, we can still *construct* an appropriate one. To do so, we take our atomic events to be all possible joint settings of the random variables, e.g., we might get something like "height = 150cm AND eye color = blue AND favorite food = zucchini".

When we fill in the probability of each of these joint outcomes, the resulting probability space is called the *joint distribution* of our random variables.

2.1 Probability tables

The above table is one way to represent a joint distribution. Often, though, we reshape the table to give a different presentation of the same information. Given the table above, we can reshape it to look like this:

	brown	blue	green	⋯
⋮	⋮	⋮	⋮	⋮
150cm	0.02	0.01	0.005	⋯
151cm	0.025	0.015	0.005	⋯
152cm	0.01	0.03	0.002	⋯
⋮	⋮	⋮	⋮	⋮

In this new version, rows correspond to values of one random variable, height and columns correspond to values of the other random variable, eye color. Each entry of the table tells us the probability of the corresponding atomic event. For example, 3% of our people have height 152cm and blue eyes.

We call each of these variants a probability table, even though the formatting is different. It's clear that the second table contains the same information as the first: each row of the first table corresponds to one cell of the second. But the second table can be useful because it shows relationships among atomic events.

In the above example there are only two random variables, so we get a two-dimensional table: one random

variable on the rows, and the other on the columns. If there are three random variables we get a three-dimensional table, with rows, columns, and layers. If we have four random variables, we have to add a fourth dimension, and so on. As we get to higher and higher dimensional tables, it becomes more difficult to write them on a printed page, but we can still usefully store them as tensors in a computer.

Importantly, a probability table is more than just a matrix or tensor: the labels on the rows, columns, etc..., are meaningful as well. So for example, it doesn't matter whether we put the values of random variable X on rows and the values of random variable Y on columns or vice versa; we'll still consider it the same table. And when we manipulate two probability tables, we will match their dimensions based on whether they correspond to the same random variable.

Probability tables give us the probabilities for atomic events. To get the probability of a compound event, we can use the sum rule: we add up the appropriate table entries. For example, to get the probability that a person's height is 151cm, we can sum across the 151cm row of the above table.

2.2 Marginal distribution

Given a joint distribution over several random variables (like height, eye color, and favorite food in the example above), we can ask for the distribution over just some subset of the variables (say just height and eye color). The smaller distribution is a *marginal* of the larger one, and forming it is called *marginalizing*.

We can compute the marginal probability table using the sum rule: we just sum out each dimension that we want to marginalize. For example, given the following table (where we've simplified our joint distribution by pretending there are just two heights and just three eye colors)

	brown	blue	green
short	0.1	0.2	0.1
tall	0.1	0.3	0.2

we can marginalize out height to get a distribution over just eye color:

brown	blue	green
0.2	0.5	0.3

Each entry above is the sum of one column of the joint table, e.g., the middle entry is

$$P(\text{blue}) = P(\text{blue} \wedge \text{short}) + P(\text{blue} \wedge \text{tall}) = 0.2 + 0.3 = 0.5$$

We can interpret marginalization as deciding to ignore irrelevant information: we have a distribution over many random variables, but for our current purpose we only want to consider some subset of them.

Knowledge check 1

1. **Numerical answer**: In the example above, what is the marginal distribution over height i.e., the distribution after we marginalize out eye color?
   - **Answer**:

tall	short
0.4	0.6

   This is computed by summing over the rows in the simplified joint distribution probability above.

2.3 Conditional distribution

Given a joint distribution, we can also ask what happens if we fix the values of some of the variables. This is called *conditioning on* or *observing* these values, and the result is a *conditional distribution*.

For example, in the table above, we could ask, suppose that we know that the person is tall; then what is the distribution over eye color? We write this as $P(\text{eye color} \mid \text{tall})$, and say "probability of eye color given tall".

To find the answer, we throw away all parts of the joint table that don't match our observations. Then we renormalize: divide by the sum of the remaining entries so that the table sums to 1. Here's the result for the conditional distribution of eye color given tall:

brown	blue	green
$1/6$	$1/2$	$1/3$

For example,

$$P(\text{blue} \mid \text{tall}) = \frac{P(\text{blue} \wedge \text{tall})}{P(\text{brown} \wedge \text{tall}) + P(\text{blue} \wedge \text{tall}) + P(\text{green} \wedge \text{tall})} = \frac{0.3}{0.6}$$

If we look back at the definition of marginalization above, we can see that the denominator is the same as $P(\text{tall})$, the marginal probability of our observation. So we can also write

$$P(\text{blue} \mid \text{tall}) = \frac{P(\text{blue} \wedge \text{tall})}{P(\text{tall})} = \frac{0.3}{0.6}$$

2.3.1 Notation for conditional distributions

If we write $P(X = T, Y = F \mid Z = 3, W = 1)$, showing specific values for all the variables, this expression evaluates to a single number: the probability that X is true and Y is false, given that Z is 3 and W is 1. If the ordering of the variables is clear, we can omit the variable names, and just write the values: $P(T, F \mid 3, 1)$. Each expression like $X = T$ represents an event, which happens exactly when a particular random variable takes a particular value.

On the other hand, if we omit a specific value for some variable, we mean a table where we look at all possible specific values. For example, $P(X, Y \mid Z = 3, W = 1)$ represents a table showing all possible values of X and Y, and listing the probabilities of the corresponding events. This particular table represents a conditional probability distribution: it will sum to 1 over X and Y.

Another example table might be $P(X = T, Y = F \mid Z, W)$, showing the probability that X is true and Y is false for all possible observations Z, W. This table is *not* a distribution: it is not required to (and generally won't) sum to 1.

The general rule is that if we sum over **everything** on the **left** of the conditioning bar, we get 1. Any other sum can be arbitrary: if we sum over only some of the variables on the left of the bar, or if we sum over anything on the right of the bar, we won't typically get 1.

Knowledge check 1

1. **True or False**: The table corresponding to $P(X, Y = F \mid Z = 3, W = 1)$ must sum to 1 i.e., $P(X, Y = F \mid Z = 3, W = 1)$ is a valid conditional probability distribution.
   - **Answer**: False, here we are only summing over the variable X as Y is fixed to some value, so this table may not *necessarily* sum to 1. Note that it is still possible for this table to sum to 1, e.g., if Y is always false when $Z = 3$ and $W = 1$ (think about why this would the table would sum to 1 in this setting), but it is not guaranteed to.