Formal Modeling in Cognitive Science

Joint, Marginal, and Conditional Distributions

Miles Osborne (originally: Frank Keller)

School of Informatics
University of Edinburgh
miles@inf.ed.ac.uk

February 4, 2008

Joint Distributions

Often, we need to consider the relationship between two or more events:

- It is cloudy and raining.
- A cat purring and being groomed.
- Noticing adverts on a page, mouse movements and eye gaze.

Joint distributions allow us to reason about the relationship between multiple events.

Joint Distributions

Previously, we introduced $P(A \cap B)$, the *probability of the intersection* of the two events A and B.

Let these events be described by the random variables X at value x and Y at value y. Then we can write:

$$P(A \cap B) = P(X = x \cap Y = y) = P(X = x, Y = y)$$

This is referred to as the *joint probability* of $X = x$ and $Y = y$.

Note: often the term joint probability and the notation $P(A, B)$ is also used for the probability of the intersection of two events.

Joint Distributions

The notion of the joint probability can be generalised to distributions:

Definition: Joint Probability Distribution

If X and Y are discrete random variables, the function given by $f(x, y) = P(X = x, Y = y)$ for each pair of values (x, y) within the range of X is called the joint probability distribution of X and Y.

Definition: Joint Cumulative Distribution

If X and Y are a discrete random variables, the function given by:

$$F(x, y) = P(X \leq x, Y \leq y) = \sum_{s \leq x} \sum_{t \leq y} f(s, t) \text{ for } -\infty < x, y < \infty$$

where $f(s, t)$ is the value of the joint probability distribution of X and Y at (s, t), is the joint cumulative distribution of X and Y.

Example: Corpus Data

Assume you have a corpus of a 100 words (a corpus is a collection of text; see Informatics 1B). You tabulate the words, their frequencies and probabilities in the corpus:

w	$c(w)$	$P(w)$	x	y
the	30	0.30	3	1
to	18	0.18	2	1
will	16	0.16	4	1
of	10	0.10	2	1
Earth	7	0.07	5	2
on	6	0.06	2	1
probe	4	0.04	5	2
some	3	0.03	4	2
Comet	3	0.03	5	2
BBC	3	0.03	3	0

Example: Corpus Data

We can now define the following random variables:

- X: the length of the word;
- Y: number of vowels in the word.

Examples for probability distributions:

- $f_X(5) = P(\text{Earth}) + P(\text{probe}) + P(\text{Comet}) = 0.14$;
- $f_Y(2) = P(\text{Earth}) + P(\text{probe}) + P(\text{some}) + P(\text{Comet}) = 0.17$.

Examples for cumulative distributions:

- $F_X(3) = f_X(2) + f_X(3) = 0.34 + 0.33 = 0.67$;
- $F_Y(1) = f_X(0) + f_X(1) = 0.03 + 0.80 = 0.83$.

Example: Corpus Data

We can now model the relationship between word length (X) and number of vowels (Y):

- Let $f(x, y) = P(X = x, Y = y)$.
- Examples:
  - $f(2, 1) = P(\text{to}) + P(\text{of}) + P(\text{on}) = 0.18 + 0.10 + 0.06 = 0.34$;
  - $f(3, 0) = P(\text{BBC}) = 0.03$;
  - $f(4, 3) = 0$.

Full distribution:

		x			
		2	3	4	5
	0	0	0.03	0	0
y	1	0.34	0.30	0.16	0
	2	0	0	0.03	0.14

Marginal Distributions

Sometimes, we want to remove the influence of an event:

- Each experiments measures lots of events, but some are less reliable than other events.
- Some events may be irrelevant to some experiment.
- We may only be interested in a subset of the events.

Marginalisation refers to the process of 'removing' the influence of one or more events from a probability.

Marginal Distributions

Definition: Marginal Distribution

If X and Y are discrete random variables and $f(x, y)$ is the value of their joint probability distribution at (x, y), the functions given by:

$$g(x) = \sum_y f(x, y) \quad \text{and} \quad h(y) = \sum_x f(x, y)$$

are the marginal distributions of X and Y, respectively.

Here, we have 'removed' either x or y.

Example: Corpus Data

We had defined the following random variables:

- X: the length of the word;
- Y: number of vowels in the word.

Joint distribution of X and Y:

		x				
		2	3	4	5	
	0	0	0.03	0	0	
y	1	0.34	0.30	0.16	0	
	2	0	0	0.03	0.14	

Example: Corpus Data

We had defined the following random variables:

- X: the length of the word;
- Y: number of vowels in the word.

Joint distribution of X and Y:

		x				
		2	3	4	5	$\sum_x f(x, y)$
	0	0	0.03	0	0	*0.03*
y	1	0.34	0.30	0.16	0	*0.80*
	2	0	0	0.03	0.14	*0.17*

Marginal distribution of Y.

Example: Corpus Data

We had defined the following random variables:

- X: the length of the word;
- Y: number of vowels in the word.

Joint distribution of X and Y:

		x				
		2	3	4	5	$\sum_x f(x,y)$
	0	0	0.03	0	0	*0.03*
y	1	0.34	0.30	0.16	0	*0.80*
	2	0	0	0.03	0.14	*0.17*
$\sum_y f(x,y)$		*0.34*	*0.33*	*0.19*	*0.14*	

Marginal distribution of Y. Marginal distribution of X.

Conditional Distributions

Sometimes, we know an event has happened already and we want to model what will happen next:

- Yahoo's share price is low and Microsoft will buy it.
- Yahoo's share price is low and Google will buy it.
- It is cloudy and it might rain.

Conditional probabilities allow us to reason about causality.

Conditional Distributions

Previously, we defined the *conditional probability* of two events A and B as follows:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Let these events be described by the random variable $X = x$ and $Y = y$. Then we can write:

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f(x,y)}{h(y)}$$

where $f(x,y)$ is the joint probability distribution of X and Y and $h(y)$ is the marginal marginal distribution of y.

Conditional Distributions

Definition: Conditional Distribution

If $f(x,y)$ is the value of the joint probability distribution of the discrete random variables X and Y at (x,y) and $h(y)$ is the value of the marginal distributions of Y at y, and $g(x)$ is the value of the marginal distributions of X at x, then:

$$f(x|y) = \frac{f(x,y)}{h(y)} \quad \text{and} \quad w(y|x) = \frac{f(x,y)}{g(x)}$$

are the conditional distributions of X given $Y = y$, and of Y given $X = x$, respectively (for $h(y) \neq 0$ and $g(x) \neq 0$).

Example: Corpus Data

Based on the joint distribution $f(x, y)$ and the marginal distributions $h(y)$ and $g(x)$ from the previous example, we can compute the conditional distributions of X given $Y = 1$:

		x			
		2	3	4	5
y	1	$\frac{f(2,1)}{h(1)} = \frac{0.34}{0.80} =$ 0.43	$\frac{f(3,1)}{h(1)} = \frac{0.30}{0.80} =$ 0.38	$\frac{f(4,1)}{h(1)} = \frac{0.16}{0.80} =$ 0.20	$\frac{f(5,1)}{h(1)} = \frac{0}{0.80} =$ 0

Independence

A key idea is the notion of *independence*:

- Is a cat purring caused by grooming?
- Does object recognition depend upon shoe size?

Deciding whether two events are independent of each other is central for understanding phenomena.

Independence

The notion of *independence* of events can be generalised to probability distributions:

Definition: Independence

If $f(x, y)$ is the value of the joint probability distribution of the discrete random variables X and Y at (x, y), and $g(x)$ and $h(y)$ are the values of the marginal distributions of X at x and Y at y, respectively, then X and Y are independent iff:

$$f(x, y) = g(x)h(y)$$

for all (x, y) within their range.

Example: Corpus Data

Marginal distributions from the previous example:

		x				
		2	3	4	5	$h(y)$
	0	0	0.03	0	0	0.03
y	1	0.34	0.30	0.16	0	0.80
	2	0	0	0.03	0.14	0.17
$g(x)$		0.34	0.33	0.19	0.14	

Now compute $g(x)h(y)$ for each cell in the table:

		x			
		2	3	4	5
	0	0.01	0.01	0.01	0.00
y	1	0.27	0.26	0.15	0.12
	2	0.06	0.06	0.03	0.02

X and Y are *not independent.*

Summary

- A joint probability distribution models the relationship between two or more events.
- marginalisations allow us to remove events from distributions.
- with conditional distributions, we can relate events to each other.
- two distributions are independent if the joint distribution is the same as the product of the two marginal distributions.