Bookmarks

Introduction

1. Probability and Inference

Introduction to Probability (Week 1)
Exercises due Sep 21, 2016 at 21:00 UTC

Probability Spaces and Events (Week 1)
Exercises due Sep 21, 2016 at 21:00 UTC

Random Variables (Week 1)
Exercises due Sep 21, 2016 at 21:00 UTC

Jointly Distributed Random Variables (Week 2)
Exercises due Sep 28, 2016 at 21:00 UTC

Conditioning on Events (Week 2)
Exercises due Sep 28, 2016 at 21:00 UTC

Homework 1 (Week 2)
Homework due Sep 28, 2016 at 21:00 UTC

Inference with Bayes' Theorem for Random Variables (Week 3)
Exercises due Oct 05, 2016 at 21:00 UTC

Independence Structure (Week 3)
Exercises due Oct 05, 2016 at 21:00 UTC

Homework 2 (Week 3)
Homework due Oct 05, 2016 at 21:00 UTC

The Product Rule for Random Variables (Also Called the Chain Rule)

6.008.1x - The Product Rule for Random Variables (Also Called the Ch

0:00 / 0:00 1.25x

Video

Download video file

Transcripts

Download SubRip (.srt) file

Download Text (.txt) file

These notes cover roughly the same content as the video:

THE PRODUCT RULE FOR RANDOM VARIABLES (COURSE NOTES)

In many real world problems, we aren't given what the joint distribution of two random variables is although we might be given other information from which we can compute the joint distribution. Often times, we can compute

Notation Summary (Up Through Week 3)

Mini-project 1: Movie Recommendations (Week 3)
Mini-projects due Oct 12, 2016 at 21:00 UTC

out the joint distribution using what's called the *product rule* (often also called the chain rule). This is precisely the random variable version of the product rule for events.

As we saw from before, we were able to derive Bayes' theorem for events using the product rule for events: $\mathbb{P}(\mathcal{A} \cap \mathcal{B}) = \mathbb{P}(\mathcal{A})\mathbb{P}(\mathcal{B} \mid \mathcal{A})$. The random variable version of the product rule is derived just like the event version of the product rule, by rearranging the equation for the definition of conditional probability. For two random variables X and Y (that take on values in sets $\mathcal{X}$ and $\mathcal{Y}$ respectively), the *product rule* for random variables says that

$$p_{X,Y}(x,y) = p_Y(y)p_{X|Y}(x \mid y) \qquad \textbf{for all } x \in \mathcal{X}, y \in \mathcal{Y} \textbf{ such that } p_Y(y) > 0.$$

Interpretation: If we have the probability table for Y, and separately the probability table for X conditioned on Y, then we can come up with the joint probability table (i.e., the joint distribution) of X and Y.

What happens when $p_Y(y) = 0$? Even though $p_{X|Y}(x \mid y)$ isn't defined in this case, one can readily show that $p_{X,Y}(x,y) = 0$ when $p_Y(y) = 0$.

To see this, think about what is happening computationally: Remember how $p_Y(y)$ is computed from joint probability table $p_{X,Y}$? In particular, we have $p_Y(y) = \sum_x p_{X,Y}(x,y)$, so $p_Y(y)$ is the sum of either a row or a column in the joint probability table (whether it's a row or column just depends on how you write out the table and which random variable is along which axis–along rows or columns). So if $p_Y(y) = 0$, it must mean that the individual elements being summed are 0 (since the numbers we're summing up are nonnegative).

We can formalize this intuition with a proof:

Claim: Suppose that random variables X and Y have joint probability table $p_{X,Y}$ and take on values in sets $\mathcal{X}$ and $\mathcal{Y}$ respectively. Suppose that for a specific choice of $y \in \mathcal{Y}$, we have $p_Y(y) = 0$. Then

$$p_{X,Y}(x,y) = 0 \qquad \textbf{for all } x \in \mathcal{X}.$$

Proof: Let $y \in \mathcal{Y}$ satisfy $p_Y(y) = 0$. Recall that we relate marginal distribution p_Y to joint distribution $p_{X,Y}$ via marginalization:

$$0 = p_Y(y) = \sum_{x \in \mathcal{X}} p_{X,Y}(x, y).$$

Next, we use a crucial mathematical observation: If a sum of nonnegative numbers (such as probabilities) equals 0, then each of the numbers being summed up must also be 0 (otherwise, the sum would be positive!). Hence, it must be that each number being added up in the right-hand side sum is 0, i.e.,

$$p_{X,Y}(x, y) = 0 \qquad \text{for all } x \in \mathcal{X}.$$

This completes the proof. □

Thus, in general:

$$p_{X,Y}(x, y) = \begin{cases} p_Y(y) p_{X|Y}(x \mid y) & \text{if } p_Y(y) > 0, \\ 0 & \text{if } p_Y(y) = 0. \end{cases}$$

Important convention for this course: For notational convenience, throughout this course, we will often just write $p_{X,Y}(x, y) = p_Y(y) p_{X|Y}(x \mid y)$ with the understanding that if $p_Y(y) = 0$, even though $p_{X|Y}(x \mid y)$ is not actually defined, $p_{X,Y}(x, y)$ just evaluates to 0 anyways.

The product rule is symmetric: We can use the definition of conditional probability with X and Y swapped, and rearranging factors, we get:

$$p_{X,Y}(x, y) = p_X(x) p_{Y|X}(y \mid x) \qquad \text{for all } x \in \mathcal{X}, y \in \mathcal{Y} \text{ such that } p_X(x) > 0,$$

and so similarly we could show that

$$p_{X,Y}(x, y) = \begin{cases} p_X(x) p_{Y|X}(y \mid x) & \text{if } p_X(x) > 0, \\ 0 & \text{if } p_X(x) = 0. \end{cases}$$

Again for notational convenience, we'll typically just write $p_{X,Y}(x, y) = p_X(x) p_{Y|X}(y \mid x)$ with the understanding that the expression is 0 when $p_X(x) = 0$.

Interpretation: If we're given the probability table for X and, separately, the probability table for Y conditioned on X, then we can come up with the joint probability table for X and Y.

Importantly, for any two jointly distributed random variables X and Y, the product rule is always true, without making any further assumptions! Also, as a recurring theme that we'll see later on as well, we are decomposing the joint distribution into the product of factors (in this case, the product of two factors).

Many random variables: If we have many random variables, say, X_1, X_2, up to X_N where N is not a random variable but is a fixed constant, then we have

$$\begin{aligned} &p_{X_1,X_2,\ldots,X_N}(x_1,x_2,\ldots,x_N) \\ &= p_{X_1}(x_1)p_{X_2|X_1}(x_2 \mid x_1)p_{X_3|X_1,X_2}(x_3 \mid x_1,x_2) \\ &\quad \cdots p_{X_N|X_1,X_2,\ldots,X_{N-1}}(x_N \mid x_1,x_2,\ldots,x_{N-1}). \end{aligned}$$

Again, we write this to mean that this holds for every possible choice of $x_1, x_2, \ldots, x_N$ for which we never condition on a zero probability event. Note that the above factorization always holds without additional assumptions on the distribution of $X_1, X_2, \ldots, X_N$.

Note that the product rule could be applied in arbitrary orderings. In the above factorization, you could think of it as introducing random variable X_1 first, and then X_2, and then X_3, etc. Each time we introduce another random variable, we have to condition on all the random variables that have already been introduced.

Since there are N random variables, there are $N!$ different orderings in which we can write out the product rule. For example, we can think of introducing the last random variable X_N first and then going backwards until we introduce X_1 at the end. This yields the, also correct, factorization

$$\begin{aligned} &p_{X_1,X_2,\ldots,X_N}(x_1,x_2,\ldots,x_N) \\ &= p_{X_N}(x_N)p_{X_{N-1}|X_N}(x_{N-1} \mid x_N)p_{X_{N-2}|X_{N-1},X_N}(x_{N-2} \mid x_{N-1},x_N) \\ &\quad \cdots p_{X_1|X_2,X_3,\ldots,X_N}(x_1 \mid x_2,\ldots,x_N). \end{aligned}$$