Lecture 2 - Introduction to Probability Theory

> Probability theory is nothing but common sense reduced to calculation. P. Laplace (1812)

Objectives

- To use probability theory to represent states of knowledge.
- To use probability theory to extend Aristotelian logic to reason under uncertainty.
- To learn about the **pruduct rule** of probability theory.
- To learn about the **sum rule** of probability theory.
- What is a **random variable**?
- What is a **discrete random variable**?
- When are two random variable **independent**?
- What is a **continuous random variable**?
- What is the **cumulative distribution function**?
- What is the **probability density function**?

Readings

Before coming to class, please read the following:

- Chapter 1 of Probabilistic Programming and Bayesian Methods for Hackers (http://nbviewer.ipython.org/github/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/blob/master/Chapter1_Introduction/Chapter1.ipynb)
- Chapter 1 (http://home.fnal.gov/~paterno/images/jaynesbook/cc01p.pdf) of (Jaynes, 2003).
- Chapter 2 (http://home.fnal.gov/~paterno/images/jaynesbook/cc02p.pdf) of (Jaynes, 2003) (skim through).

The basic desiderata of probability theory

It is actually possible to derive the rules of probability based on a system of common sense requirements. Paraphrasing Chapter 1 (http://home.fnal.gov/~paterno/images/jaynesbook/cc01p.pdf) of \cite{jaynes2003}), we would like our system to satisfy the following desiderata:

1) *Degrees of plausibility are represented by real numbers.*

2) *The system should have a qualitative correspondance with common sense.*

3) *The system should be consistent in the sense that:*

- *If a conclusion can be reasoned out in more than one way, then every possible way must lead to the same result.*
- *All the evidence relevant to a question should be taken into account.*
- *Equivalent states of knowledge must be represented by equivalent plausibility assignments.*

How to speak about probabilities?

Let

- A be a logical sentence,
- B be another logical sentence, and
- and I be all other information we know.

There is no restriction on what A and B may be as soon as none of them is a contradiction. We write as a shortcut:

$$\text{not A} \equiv \neg,$$

$$A \text{ and } B \equiv A, B \equiv AB,$$

$$A \text{ or } B \equiv A + B.$$

We **write**:

$$p(A \mid BI),$$

and we **read**:

> the probability of A being true given that we know that B and I is true

or (assuming knowledge I is implied)

> the probability of A being true given that we know that B is true

or (making it even shorter)

> the probability of A given B.

$p(\text{something} \mid \text{everything known}) = \text{probability samething is true conditioned on what is known.}$

$p(A \mid B, I)$ is just a number between 0 and 1 that corresponds to the degree of plaussibility of A conditioned on B and I. 0 and 1 are special.

- If

$$p(A \mid BI) = 0,$$

  we say that we are certain that A is false if B is true.
- If

$$p(A \mid BI) = 1,$$

  we say that we are certain that A is false if B is false.
- If

$$p(A \mid BI) \in (0, 1),$$

  we say that we are uncertain about A given that B is false. Depending on whether $p(A \mid B, I)$ is closer to 0 or 1 we beleive more on one possibiliy or another. Complete ignorance corresponds to a probability of 0.5.

The rules of probability theory

According to Chapter 2 (http://home.fnal.gov/~paterno/images/jaynesbook/cc02m.pdf) of \cite{jaynes2003} the desiderata are enough to derive the rules of probability. These rules are:

- The **obvious rule** (in lack of a better name):

$$p(A \mid I) + p(\neg A \mid I) = 1.$$

- The **product rule** (also known as the Bayes rule or Bayes theorem):

$$p(AB \mid I) = p(A \mid BI)p(B \mid I).$$

  or

$$p(AB \mid I) = p(B \mid AI)p(A \mid I).$$

These two rules are enough to compute any probability we want. Let us demonstrate this by a very simple example.

Example: Drawing balls from a box without replacement

Consider the following example of prior information I:

> We are given a box with 10 balls 6 of which are red and 4 of which are blue. The box is sufficiently mixed so that so that when we get a ball from it, we don't know which one we pick. When we take a ball out of the box, we do not put it back.

Let A be the sentence:

> The first ball we draw is blue.

Intuitively, we would set the probability of A equal to:

$$p(A \mid I) = \frac{4}{10}.$$

This choice can actually be justified, but we will come to this later in this course. From the "obvious rule", we get that the probability of not drawing a blue ball, i.e., the probability of drawing a red ball in the first draw is:

$$p(\neg A \mid I) = 1 - p(A \mid I) = 1 - \frac{4}{10} = \frac{6}{10}.$$

Now, let B be the sentence:

> The second ball we draw is red.

What is the probability that we draw a red ball in the second draw given that we drew a blue ball in the first draw? Just before our second draw, there remain 9 bals in the box, 3 of which are blue and 6 of which are red. Therefore:

$$p(B \mid AI) = \frac{6}{9}.$$

We have not used the product rule just yet. What if we wanted to find the probability that we draw a blue during the first draw and a red during the second draw? Then,

$$p(AB \mid I) = p(A \mid I)p(B \mid AI) = \frac{4}{10}\frac{6}{9} = \frac{24}{90}.$$

What about the probability o a red followed by a blue? Then,

$$p(\neg AB \mid I) = p(\neg A \mid I)p(B \mid AI) = [1 - p(A \mid I)]p(B \mid \neg AI) = \frac{6}{10}\frac{5}{9} = \frac{30}{90}.$$

Other rules of probability theory

All the other rules of probability theory can be derived from these two rules. To demonstrate this, let's prove that:

$$p(A + B \,|\, I) = p(A \,|\, I) + p(B \,|\, I) - p(AB \,|\, I).$$

Here we go:

$$\begin{aligned}
p(A + B \,|\, I) &= 1 - p(\neg A \neg B \,|\, I) \text{ (obvious rule)} \\
&= 1 - p(\neg A \,|\, \neg BI) p(\neg B \,|\, I) \text{ (product rule)} \\
&= 1 - [1 - p(A \,|\, \neg BI)] p(\neg B \,|\, I) \text{ (obvious rule)} \\
&= 1 - p(\neg B \,|\, I) + p(A \,|\, \neg BI) p(\neg B \,|\, I) \\
&= 1 - [1 - p(B \,|\, I)] + p(A \,|\, \neg BI) p(\neg B \,|\, I) \\
&= p(B \,|\, I) + p(A \,|\, \neg BI) p(\neg B \,|\, I) \\
&= p(B \,|\, I) + p(A \neg B \,|\, I) \\
&= p(B \,|\, I) + p(\neg B \,|\, AI) p(A \,|\, I) \\
&= p(B \,|\, I) + [1 - p(B \,|\, AI)] p(A \,|\, I) \\
&= p(B \,|\, I) + p(A \,|\, I) - p(B \,|\, AI) p(A \,|\, I) \\
&= p(A \,|\, I) + p(B \,|\, I) - p(AB \,|\, I).
\end{aligned}$$

The sum rule

Now consider a finite set of logical sentences, $B_1, \ldots, B_n$ such that:

1. One of them is definitely true:

$$p(B_1 + \cdots + B_n \,|\, I) = 1.$$

2. They are mutually exclusive:

$$p(B_i B_j \,|\, I) = 0, \text{ if } i \neq j.$$

The **sum rule** states that:

$$P(A \,|\, I) = \sum_i p(AB_i \,|\, I) = \sum_i p(A \,|\, B_i I) p(B_i \,|\, I).$$

We can prove this by induction, but let's just prove it for $n = 2$:

$$\begin{aligned}
p(A \,|\, I) &= p[A(B_1 + B_2 \,|\, I] \\
&= p(AB_1 + AB_2 \,|\, I) \\
&= p(AB_1 \,|\, I) + p(AB_2 \,|\, I) - p(AB_1 B_2 \,|\, I) \\
&= p(AB_1 \,|\, I) + p(AB_2 \,|\, I),
\end{aligned}$$

since

$$p(AB_1 B_2 \,|\, I) = p(A \,|\, B_1 B_2 \,|\, I) p(B_1 B_2 \,|\, I) = 0.$$

Let's go back to our example. We can use the sum rule to compute the probability of getting a red ball on the second draw independently of what we drew first. This is how it goes:

$$\begin{aligned} p(B\,|\,I) &= p(AB\,|\,I) + p(\neg AB\,|\,I) \\ &= p(B\,|\,AI)p(A\,|\,I) + p(B\,|\,\neg AI)p(\neg A\,|\,I) \\ &= \frac{6}{9}\frac{4}{10} + \frac{5}{9}\frac{6}{10} \\ &= \dots \end{aligned}$$

Example: Medical Diagnosis

This example is a modified version of the one found in Lecture 1 (http://www.zabaras.com/Courses/BayesianComputing/IntroToProbabilityAndStatistics.pd of the Bayesian Scientific Computing course offered during Spring 2013 by Prof. N. Zabaras at Cornell University.

We are going to examine the usefullness of a new tuberculosis test. Let the prior information, I, be:

> The percentage of the population infected by tuberculosis is 0.4%. We have run several experiments and determined that:
>
> - If a tested patient has the disease, then 80% of the time the test comes out positive.
> - If a tested patient does not have the disease, then 90% of the time, the test comes out negative.

Suppose now that you administer this test to a patient and that the result is positive. How confident are you that the patient does indeed have the disease?

Let's use probability theory to answer this question. Let A be the event:

> The patient's test is positive.

Let B be the event:

> The patient has tuberculosis.

According to the prior information, we have:

$$p(B\,|\,I) = p(\text{has tuberculosis}\,|\,I) = 0.004,$$

and

$$p(A\,|\,B,I) = p(\text{test is positive}\,|\,\text{has tuberculosis},I) = 0.8.$$

Similarly,

$$p(A \mid \neg B, I) = p(\text{test is positive} \mid \text{does not have tuberculosis}, I) = 0.1.$$

We are looking for:

$$\begin{aligned} p(\text{has tuberculosis} \mid \text{test is positive}, I) &= P(B \mid A, I) \\ &= \frac{p(AB \mid I)}{p(A \mid I)} \\ &= \frac{p(A \mid B, I)p(B \mid I)}{p(A \mid B, I)p(B \mid I) + p(A \mid \neg B, I)p(\neg B \mid I)} \\ &= \frac{0.8 \times 0.004}{0.8 \times 0.004 + 0.1 \times 0.996} \\ &\approx 0.031. \end{aligned}$$

How much would you pay for such a test?

Conditional Independence

We say that A and B are **independent** (conditional on I), and write,

$$A \perp B \mid I,$$

if knowledge of one does not yield any information about the other. Mathematically, by $A \perp B \mid I$, we mean that:

$$p(A \mid B, I) = p(A \mid I).$$

Using the product rule, we can easily show that:

$$A \perp B \mid I \iff p(AB \mid I) = p(A \mid I)p(B \mid I).$$

Question

- Give an example of I, A and B so that $A \perp B \mid I$.

Now, let C be another event. We say that A and B are **independent** conditional on C (and I), and write:

$$A \perp B \mid C, I,$$

if knowlege of C makes information about A irrelevant to B (and vice versa). Mathematically, we mean that:

$$p(A \mid B, C, I) = p(A \mid C, I).$$

Question

- Give an example of I, A, B, C so that $A \perp B \mid C, I$.

Random Variables

The formal mathematical definition of a random variable involves measure theory and is well beyond the scope of this course. Fortunately, we do not have to go through that route to get a theory that is useful in applications. For us, a **random variable** X will just be a variable of our problem whose value is unknown to us. Note that, you should not take the word "random" too literally. If we could, we would change the name to **uncertain** or **unknown** variable. A random variable could correspond to something fixed but unknown, e.g., the number of balls in a box, or it could correspond to something truely random, e.g., the number of particles that hit a Geiger counter (https://en.wikipedia.org/wiki/Geiger_counter) in a specific time interval.

Discrete Random Variables

We say that a random variable X is discrete, if the possible values it can take are discrete (possibly countably infinite). We write:

$$p(X = x \,|\, I)$$

and we read "the probability of X being x". If it does not cause any ambiguity, sometimes we will simplify the notation to:

$$p(x) \equiv p(X = x \,|\, I).$$

Note that $p(X = x)$ is actually a discrete function of x which depends on our beliefs about X. The function $p(x) = p(X = x \,|\, I)$ is known as the probability density function of X.

Now let Y be another random variable. The **sum rule** becomes:

$$p(X = x \,|\, I) = \sum_y p(X = x,\, Y = y \,|\, I) = \sum_y p(X = x \,|\, Y = y,\, I) p(Y = y \,|\, I),$$

or in simpler notation:

$$p(x) = \sum_y p(x, y) = \sum_y p(x \,|\, y) p(y).$$

The function $p(X = x,\, Y = y \,|\, I) \equiv p(x, y)$ is known as the joint *probability mass function* of X and Y.

The **product rule** becomes:

$$p(X = x,\, Y = y \,|\, I) = p(X = x \,|\, Y = y,\, I) p(Y = y \,|\, I),$$

or in simpler notation:

$$p(x, y) = p(x \,|\, y) p(y).$$

We say that X and Y are **independent** and write:

$$X \perp Y \,|\, I,$$

if knowledge of one does not yield any information about the other. Mathematically, Y gives no information about X if:

$$p(x\,|\,y) = p(x).$$

From the product rule, however, we get that:

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

from which we see that the joint distribution of X and Y must factorize as:

$$p(x,y) = p(x)p(y).$$

It is trivial to show that if this factorization holds, then

$$p(y\,|\,x) = p(y),$$

and thus X yields no information about Y either.

Continuous Random Variables

A random variable X is continuous if the possible values it can take are continuous. The probability of a continuous variable getting a specific value is always zero. Therefore, we cannot work directly with probability mass functions as we did for discrete random variables. We would have to introduce the concepts of the **cumulative distribution function** and the **probability density function**. Fortunately, with the right choice of mathematical symbols, the theory will look exactly the same.

Let us start with a real continuous random variable X, i.e., a random variable taking values in the real line R. Let $x \in \mathrm{R}$ and consider the probability of X being less than or equal to x:

$$F(x) := p(X \leq x\,|\,I).$$

$F(x)$ is known as the **cumulative distribution function** (CDF). Here are some properties of the CDF whose proof is left as an excersise:

- The CDF starts at zero and goes up to one:

$$F(-\infty) = 0 \text{ and } F(+\infty) = 1.$$

- $F(x)$ is an increasing function of x, i.e.,

$$x_1 \leq x_2 \implies F(x_1) \leq F(x_2).$$

- The probability of X being in the interval $[x_1, x_2]$ is:

$$p(x_1 \leq X \leq x_2\,|\,I) = F(x_2) - F(x_1).$$

Now, assume that the derivative of $F(x)$ with respect to x exists. Let us call it $f(x)$:

$$f(x) = \frac{dF(x)}{dx}.$$

Using the fundamental theorem of calculus, it is trivial to show Eq. (???) implies:

$$p(x_1 \leq X \leq x_2 \,|\, I) = \int_{x_1}^{x_2} f(x) dx.$$

$f(x)$ is known as the **probability density function** (PDF) and it is measured in probability per unit of X. To see this note that:

$$p(x \leq X \leq x + \delta x \,|\, I) = \int_{x}^{x+\delta x} f(x') dx' \approx f(x) \delta x,$$

so that:

$$f(x) \approx \frac{p(x \leq X \leq x + \delta x \,|\, I)}{\delta x}.$$

The PDF should satisfy the following properties:

- It should be positive

$$f(x) \geq 0,$$

- It should integrate to one:

$$\int_{-\infty}^{\infty} f(x) dx = 1.$$

Notation about the PDF of continuous random variables

In order to make all the formulas of probability theory the same, we define for a continuous random variable X:

$$p(x) := f(x) = \frac{dF(x)}{dx} = \frac{d}{dx} p(X \leq x \,|\, I).$$

But keep in mind, that if X is continuous $p(x)$ is not a probability but a probability density. That is, it needs a dx to become a probability.

Let the PDF $p(x)$ of X and the PDF $p(y)$ of Y (Y is another continuous random variable). We can find the PDF of the random variable X conditioned on Y, i.e., the PDF of X if Y is directly observed. This is the **product rule** for continuous random variables:

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

where $p(x, y)$ is the **joint PDF** of X and Y. The **sum rule** for continous random variables is:

$$p(x) = \int p(x, y) dy = \int p(x \,|\, y) p(y) dy.$$

The similarity between these rules and the discrete ones is obvious. We have prepared a table to help you remember it.

Concept	Discrete Random Variables	Continuous Random Variables
$p(x)$	in units of robability	in units of probability per unit of X
sum rule	$\sum_y p(x, y) = \sum_y p(x \mid y)p(y)$	$\int_y p(x, y)dy = \int_y p(x \mid y)p(y)$
product rule	$p(x, y) = p(x \mid y)p(y)$	$p(x, y) = p(x \mid y)p(y)$

Expectations

Let X be a random variable. The expectation of X is defined to be:

$$\mathrm{E}[X] := \mathrm{E}[X \mid I] = \int x p(x) dx.$$

Now let $g(x)$ be any function. The expectation of $g(X)$, i.e., the random variable defined after passing X through $g(\cdot)$, is:

$$\mathrm{E}[g(X)] := \mathrm{E}[g(X) \mid I] = \int g(x) p(x) dx.$$

As usual, calling $\mathrm{E}[\cdot]$ is not a very good name. You may think of $\mathrm{E}[g(X)]$ as the expected value of $g(X)$, but do not take it too far. Can you think of an example in which the expected value is never actually observed?

Conditional Expectation

Let X and Y be two random variables. The conditional expectation of X given $Y = y$ is defined to be:

$$\mathrm{E}[X \mid Y = y] := \mathrm{E}[X \mid Y = y, I] = \int x p(x \mid y) dx.$$

Properties of Expectations

The following properties of expectations of random variables are extremely helpful. In what follows, X and Y are random variables and c is a constant:

- Sum of random variable with a constant:

$$\mathrm{E}[X + c] = \mathrm{E}[X] + c.$$

- Sum of two random variables:

$$\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y].$$

- Product of random variable with constant:

$$\mathrm{E}[cX] = c\mathrm{E}[X].$$

- If $X \perp Y$, then:

$$\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y].$$

NOTE: This property does not hold if X and Y are not independent!

- If $f(\cdot)$ is a convex function, then:

$$f(\mathrm{E}[X]) \leq \mathrm{E}[f(X)].$$

NOTE: The equality holds only if $f(\cdot)$ is linear!

Variance of a Random Variable

The variance of X is defined to be:

$$\mathrm{V}[X] = \mathrm{E}\Big[X - \mathrm{E}[X])^2\Big].$$

It is easy to prove (and a very useful formulat to remember), that:

$$\mathrm{V}[X] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2.$$

Covariance of Two Random Variables

Let X and Y be two random variables. The covariance between X and Y is defined to be:

$$\mathrm{C}[X, Y] = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]$$

Properties of the Variance

Let X and Y be random variables and c be a constant. Then:

- Sum of random variable with a constant:

$$\mathrm{V}[X + c] = \mathrm{V}[X].$$

- Product of random variable with a constant:

$$\mathrm{V}[cX] = c^2\mathrm{V}[X].$$

- Sum of two random variables:

$$\mathrm{V}[X + Y] = \mathrm{V}[X] + \mathrm{V}[Y] + 2\mathrm{C}(X, Y).$$

- Sum of two independent random variables:

$$\mathrm{V}[X + Y] = \mathrm{V}[X] + \mathrm{V}[Y].$$

References

(Jaynes, 2003) E T Jaynes, ``*Probability Theory: The Logic of Science*'', 2003. online (http://bayes.wustl.edu/etj/prob/book.pdf)