Lecture Notes: Introduction to Probability

Ted Meeds[1,2]

[1] Informatics Institute, University of Amsterdam
[2] The Centre for Integrative Bioinformatics, Vrije University
tmeeds@gmail.com

Abstract This is a very limited introduction to probability, mostly based on Bishop's first Chapter. There are **many** other better introductions, I recommend `http://people.bath.ac.uk/sw283/core-statistics-nup.pdf` by Simon Wood.

Probability theory provides a consistent framework for the quantification and manipulation of uncertainty and forms one of the central foundations of pattern recognition. Bishop p12

1 Discrete probability

In discrete probability, random variables can take on values from a finite set. The variable names are capitalized, e. g., X, Y etc and the values they take on are lowercase e. g., x, y, x_m, etc. We may use indices to select values from their finite set, e. g., $X \in \{x_i\}$, $i = 1..I$, $Y \in \{y_j\}$, $j = 1..J$.

The probability of a value is equal to the fraction of time we observe that value (from, say, a set of observations/ events / trials). We write either $Pr(X = x)$ or $P(X = x)$ to be the probability of variable X takes on value x. Probabilities are constrained $0 \geq P(X = x) \leq 1$ and $\sum_x P(x) = 1$.

1.1 Example: fruit in boxes

Let B represent the identity of the box (i.e. a random variable); $B \in \{r, b\}$ (i.e. a red or blue box). Let F represent the identity of the fruit; $F \in \{l, o\}$ (i.e. lime or orange – changed from Bishop to match the colours). We can look inside the boxes and count the fruits of each type in each box:

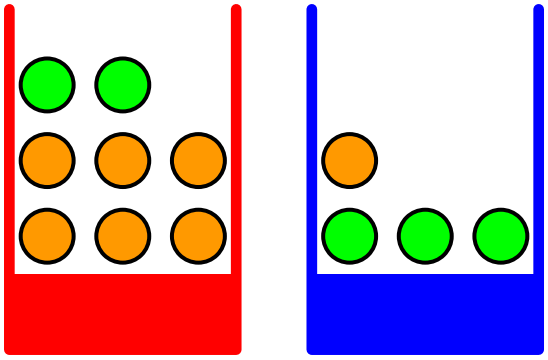

Figure 1. Discrete Probability example (Bishop 1.9). Red/blue box, limes/oranges.

	$F = a$ (limes)	$F = o$ (orange)	
B = r (red box)	2	6	8 in red box
B = b (blue box)	3	1	4 in blue box
	5 limes	7 oranges	12 total

$$\underbrace{P(B = r, F = l)}_{\text{joint probability of “red” and “lime”}} = \frac{2}{12} \tag{1}$$

$$P(F = l) = \frac{5}{12} \tag{2}$$

The quantity $P(F = l)$ can also be computed using the **sum rule**:

$$P(F = l) = \sum_{x \in \{r,b\}} P(F = l, B = x) \tag{3}$$

$$= P(F = l, B = r) + P(F = l, B = b) \tag{4}$$

$$= \frac{2}{12} + \frac{3}{12} = \frac{5}{12} \tag{5}$$

$$P(B = r) = \sum_{x \in \{l,o\}} P(F = x, B = r) \tag{6}$$

$$= P(F = l, B = r) + P(F = o, B = r) \tag{7}$$

$$= \frac{2}{12} + \frac{6}{12} = \frac{8}{12} \tag{8}$$

More generally:

$$P(X = x_i) = \sum_{j=1}^{J} P(X = x_i, Y = y_j) \tag{9}$$

In other words, the sum rule **marginalizes** one (or more) random variable(s) by **summing over** other random variables. This is also called **marginalization**.

The quantity $P(F = l|B = r)$ is read *the probability that the fruit is an apple* **given** *that the box is red*. This is simply the fraction of apples in the red box: $P(F = l|B = r) = 2/8$. This is a **conditional probability** as we are conditioning on the value $B = r$.

The **product rule** is used to compute the joint probability from a marginal and conditional probability:

$$P(B = r, F = l) = P(B = r)P(F = l|B = r) \tag{10}$$

$$= \frac{8}{12} \cdot \frac{2}{8} \tag{11}$$

$$= \frac{2}{12} \tag{12}$$

The sum and product rules are fundamental operators in probability theory. In machine learning they are used to **reverse** the conditioning relationships to derive the **posterior parameter distribution**, that is, the distribution of parameters **given** the data. The application of the sum and product rules for this purpose is called **Bayes theorem** (or **Bayes rule**):

$$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)} \tag{13}$$

where $P(X) = \sum_Y P(X|Y)P(Y)$ (product rule for constructing the joint then the sum rule for marginalization). Notice how we used the $X|Y$ relationship to the $Y|X$ relationship.

Example: fruit in boxes.

$$P(B = r|F = l) = \frac{P(F = l|B = r)P(B = r)}{P(F = l)} \tag{14}$$

where $P(F = a) = \sum_B P(F = l|B)P(B)$.

2 Continuous Probability Distributions

If random variables can now take on a continuous value (rather than values from a discrete set), we use probability density functions (pdf) to describe their distributions. For example, $p(X = x_i)$ is the density of variable X taking on value x_i. Note that the density is not between 0 and 1, but 0 and ∞.

The sum rule extends naturally to continuous distributions by replacing the summation with an integral. The product rule remains the same.

The sum rule:

$$P(B = r) = \int_{-\infty}^{\infty} p(F = x, B = r)dx \tag{15}$$

Just like for discrete distributions, the integral over all possible values sums to 1.

$$\int_{-\infty}^{\infty} p(x)dx = 1 \tag{16}$$

One can think of dx as the infinitesimal change in volume. The total volume is 1.

It is important to note that distributions nearly always composed of a function that depends on the variable, say $f(x)$ and a constant C that does not depend of the variable but which makes the integral equal to 1; we call these **normalizing constants**. Since the normalizing constant does not depend on x, we can pull it out of the integral.

$$\int_{-\infty}^{\infty} p(x)dx = 1 \tag{17}$$

$$\int_{-\infty}^{\infty} C \cdot f(x)dx = 1 \tag{18}$$

$$C \int_{-\infty}^{\infty} f(x)dx = 1 \tag{19}$$

$$\int_{-\infty}^{\infty} f(x)dx = 1/C \tag{20}$$

It is important to understand this when working through the linear algebra required for finding analytic solutions to distributions using the product and sum rules.

3 Probability Models of Data

The sum and product rules are used everywhere in Machine Learning when dealing with probabilistic models. We often drop the variable name and use the value directly, i. e., $p(x)$ or $P(x)$.

Recall that in a data set there are N data vectors x_n, $n = 1, 2, \ldots, N$, called $\mathcal{D}$. Given a model class, we have model parameters $\boldsymbol{\theta}$ and a likelihood function $p(x_n|\boldsymbol{\theta})$ (e. g., a Gaussian distribution centered at $\boldsymbol{\theta}$). If the data a iid, we can use the product rule to write the full likelihood $p(\mathcal{D}|\boldsymbol{\theta}) = \prod_n p(x_n|\boldsymbol{\theta})$. If we add a **prior** distribution over $\boldsymbol{\theta}$, we can use the product rule again to get the joint distribution $p(\mathcal{D}, \boldsymbol{\theta}) =$

$p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. The marginal distribution is found using the sum rule: $p(\mathcal{D}) = \int p(\mathcal{D}, \boldsymbol{\theta}) d\boldsymbol{\theta}$. Then finally, we can use Bayes rule to get the posterior: $p(\theta|\mathcal{D}) = p(\mathcal{D}, \boldsymbol{\theta})/p(\mathcal{D})$. Summarized:

For probability models we have:

$$posterior = \frac{likelihood \times prior}{evidence} \tag{21}$$

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})}{p(\mathcal{D})} \tag{22}$$

where $p(\mathcal{D}) = \int p(\mathcal{D}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta}) d\boldsymbol{\theta}$. In other words the **evidence** sums over (for continuous $\boldsymbol{\theta}$) all values of $\boldsymbol{\theta}$.

The sum and product rules apply equally to probability density functions over continuous variables. The sum is replaced with an integral. Please read Bishop chapter 1 for these rules and Bishop chapter 2 the parts on the Gaussian distribution.

References