HST.582J / 6.555J / 16.456J Biomedical Signal and Image Processing
Spring 2007

Chapter 10 - A PROBABILITY PRIMER

CDFs, PMFs, PDFs, Expectation and all that...

10 Introduction

In many experiments there is some element of randomness that we are unable to explain. Probability and statistics are mathematical tools for reasoning in the face of such uncertainty. Here, we are primarily interested in the use of probability for decision making, estimation, and cost minimization. Probabilistic models allow us to address such issues quantitatively. For example; "Is the signal present or not?" Binary : YES or NO, "How certain am I?" Continuous : Degree of confidence. Here we introduce some very basic notions of probability and random variables and their associated cumulative distribution functions (CDF), probability mass functions (PMF), and probability density functions (PDF). From these we can define the notion of expectation, marginalization, and conditioning. This description is meant to be concise, limiting the discussion to those concepts which can be applied to decision problems we encounter in biomedical signal and image processing[1].

10.1 Sample Spaces and Events

Basic probability can be derived from set theory where one considers a **sample space**, associated **events**, and a **probabilty law**. The **sample space**, denoted Ω, is the exhaustive set of finest grain outcomes of an experiment. An **event** is a subset of the **sample space**. A **probability law**, denoted P, assigns numerical values to **events**. While the sample space may be continuous or discrete, we restrict the following discussion to discrete (countable) sets in order introduce basic probability concepts.

Examples of sample spaces are:

- the set of outcomes of rolling a 6-sided die,
- the set of outcomes of rolling a 6 sided die AND flipping a coin,
- the set of outcomes when drawing 3 (or 4) cards from 47

Examples of events for the above sample spaces are:

- the roll of the die is greater than 4,

[1] more accurately, we consider concepts which can be applied to a limited, but useful class of decision problems.

- the flip of the coin is heads AND the roll of the die is odd,
- the draw of 3 (or 4) cards contains at least one 2 (or one King).

Events A and B are *mutually exclusive* events if observing one precludes observing the other (equivalent to $A \cap B = \emptyset$). The **probability** of an event, denoted $P\{A\}$ or $P\{B\}$ is the number of times we expect to observe the event relative to the number of trials. Letting N be the number of experimental trials and N_A, N_B be the respective number of times events A and B are observed within N trials. The frequentist (empirical, Laplace) approach defines probabilities in terms of a limit over infinite trials (i.e. as $N \to \infty$) while the axiomatic (Kolmogorov, Bayesian) approach defines probabilities in terms of axioms (from which all other probabilistic properties can be derived). While the frequentist approach agrees with our intuition, the axiomatic approach constitutes the modern view of probability.

Empirical Definition	Axiomatic Definition
$P\{A\} = \lim_{N\to\infty}\left(\frac{N_A}{N}\right)$	$0 \le P\{A\}, P\{B\} \le 1$
$P\{B\} = \lim_{N\to\infty}\left(\frac{N_B}{N}\right)$	$P\{\Omega\} = 1$
$P\{A\cup B\} = \lim_{N\to\infty}\left(\frac{N_A+N_B}{N}\right)$	if $A \cap B = \emptyset$ $P\{A\cup B\} = P\{A\} + P\{B\}$

Note that $A \cap B$ and AB are both used to signify the event "A and B", while $A \cup B$ and $A + B$ are both used to signify the event "A or B".

10.1.1 Uniform Sample Spaces (Discrete Case)

If the elements of the sample space are equally likely (which is the case for all of the examples above), then probabilities are computed as follows.

1. $N =$ the number of experimental outcomes.
2. $N_A =$ the number of experimental outcomes which satsify the event of interest A.
3. $P(A) = \frac{N_A}{N}$.

Many probability questions simply come down to your ability to count, although counting is not always a straightforward process as discussed next.

10.2 Counting:Permutations and Sampling

Permutations and sampling schemes are useful for computing probabilities of events which are defined as sets of from the sample space where all elements of the sample space have equal probability.

10.2.1 Permutations

A permutation is a rearrangement of k distinct items. The total number of such rearrangements is denoted by $k!$ and computed as

$$k! = \prod_{i=1}^{k} i \qquad \text{where} \qquad 0! \triangleq 1 \tag{1}$$

10.2.2 Sampling without replacement

Sampling without replacement is the process of repeatedly drawing items from a set of distinct items without replacing previous selections. The total number of *different* draws of k distinct items from a set of N items can broken down into two cases: draws where the order of the items does not matter and draws in which it does. If order matters, the sets $\{a, c, d, f, g\}$ and $\{c, a, f, g, d\}$ are considered to be different draws. If order does not matter then they are considered to be the same draw. The total number of possible draws can be computed as:

$$\text{order matters:} \quad \frac{N!}{(N-k)!} \tag{2}$$

$$\text{order does not matter:} \quad \frac{N!}{(N-k)!k!} = \binom{N}{k} \tag{3}$$

Notice that when order does not matter we simply reduce the total number of draws by the number of ways any single draw can be rearranged (i.e. the number of permutations). The notation $\binom{N}{k}$ denotes the function "N choose k" which has the following property:

$$\binom{N}{k} = \binom{N}{N-k}$$

so that the total number of draws of k items from a set of N is equal to the total number of draws of $N - k$ items when order does not matter.

10.2.3 Sampling with replacement

When sampling with replacement, we (not surprisingly) replace our selection back into the set of items before redrawing an item. In this case, the selection of k items may contain repeated elements which slightly complicates cases when order does not matter. This is because we cannot simply

divide by the number of ways k items can be permuted since this would include permutations in which identical items switch positions. The total number of possible draws can be computed as:

$$\text{order matters:} \quad N^k \tag{4}$$

$$\text{order does not matter:} \quad \frac{(N+k-1)!}{(N-1)!k!} = \binom{N+k-1}{k} \tag{5}$$

10.2.4 Conditional Probability

In decision making (and estimation) we are interested in how the observation of one event changes the probability of observing another event. This is denoted $P(A|B)$, the probability of event A given B has occurred or conversely $P(B|A)$, the probability of event B given A has occurred. This probabilistic relationship can be derived from Venn diagrams and has the form:

$$P(A|B) = \frac{P(AB)}{P(B)} \tag{6}$$

$$P(B|A) = \frac{P(AB)}{P(A)} \, . \tag{7}$$

so that it is the probability of jointly observing events "A" and "B" divided by the marginal probability of the conditioning event.

10.2.5 Statistical Independence

Two events A and B are statistically independent if observing one does not alter the probability of observing the other. If both events have nonzero probability, this is equivalent to their joint probability being the product of their marginal probabilities. That is, if A and B are statistically independent events then

$$P(A|B) = P(A) \tag{8}$$

$$P(B|A) = P(B) \tag{9}$$

$$P(AB) = P(A)\,P(B) \quad ; \text{if } P(A) > 0, P(B) > 0 \tag{10}$$

10.2.6 Baye's Theorem

From the definitions of conditional probability we see that the joint probability, that is the probability of A and B can be written as a product of a conditional probability and a marginal probability

$$\begin{aligned} P(AB) &= P(A|B)\,P(B) \\ &= P(B|A)\,P(A) \, . \end{aligned}$$

From this relationship we can easily derive the well known Bayes' rule

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} = \frac{P(AB)}{P(B)} \tag{11}$$

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)} = \frac{P(AB)}{P(A)}\ . \tag{12}$$

Bayes' rule plays a prominent role in many statistical estimation and decision making problems.

10.3 Probability Redux:Venn Diagrams

The previous concepts including sample and event spaces, conditional probabilities, and Bayes' rule can be captured conceptually via Venn diagrams. In particular, conditional probability is easily derived using such diagrams. See the sequence of powerpoint slides on this point.

10.4 Random Variables

So far we have discussed ways in which to compute probabilities of events and how to manipulate joint and marginal probabilities in order to compute conditional probabilities. However, keeping track of different events and their probabilistic relationships is a bit cumbersome. The notion of a random variable simplifies this somewhat. There are two types of scalar random variables; discrete and continuous with which we associate a function enabling us to compute probabilities of specific events. In the case of discrete random variables the argument of the function is an index (or indices in the multi-dimensional case) while in the continuous random variable the argument of the function is a continuous value (or values in the multi-dimensional case). When we discuss multi-dimensional random variables we can define random *vectors* which are comprised of both discrete and continous random variables.

10.4.1 Discrete Random Variables (Probability Mass Functions)

The functions which describes a discrete random variable is a **probability mass function** (PMF). Experimental outcomes are divided into a set of mutually exclusive events. Associated with each event is a probability (a value between 0 and 1). The sum over all probabilities is unity.

If $\{A_1, \ldots, A_N\}$ is a set of N mutually exclusive events which enumerates all possible events then the PMF is an indexed table which satisfies the following

$$\begin{aligned} P_i &= P(A_i) \\ P_i &\geq 0 \\ \sum_{i=1}^{N} P_i &= 1 \end{aligned}$$

10.4.2 Continuous Random Variables (Probability Density Functions)

$\frac{\Delta}{T}$ to each. This results in a PMF with equal probability of a sample falling into any interval. In order to consider a continuum of events we could take a limit as Δ grows arbitrarily small to compute the probability that the random variable $X = t$ where $0 \leq t \leq T$.

$$\begin{aligned} P(X = t) &= \lim_{\Delta \to 0} \left(P\left(t - \frac{\Delta}{2} < X \leq t + \frac{\Delta}{2} \right) \right) \\ &= \lim_{\Delta \to 0} \left(\frac{\Delta}{T} \right) \\ &= 0 \end{aligned}$$

As we see, this approach converges to zero probability in the limit. While this result for continuum turns out to be correct (i.e. the probability of observing a given value of a continuous random variable is technically zero), it is not useful for describing random variables on the continuum. Alternatively, we can define a specific type of event, not disimilar to the discretization above, in a way that allows the computation of the probability of **any** event defined on the continuum. Namely, the probability that the random variable X will take a on value less than or equal to x.

$$P_X(x) = \Pr\{X \leq x\} \tag{13}$$

$$1 - P_X(x) = \Pr\{X > x\} \tag{14}$$

$P_X(x)$ is what is known as the cumulative distribution function (CDF) of a random variable and has the following properties

$$\begin{aligned} P_X(-\infty) &= 0 \\ P_X(\infty) &= 1 \\ 0 &\leq P_X(x) \leq 1 \\ P_X(x + \Delta) &\geq P_X(x) \quad ; \quad \Delta \geq 0 \end{aligned}$$

From it we can derive the probability density function (PDF) of a continuous random variable

$$P_X(x) = \int_{-\infty}^{x} p_X(u)\, du \tag{15}$$

$$p_X(x) = \frac{\partial}{\partial x} P_X(x) \tag{16}$$

which has the following properties

$$\begin{aligned} p_X(x) &\geq 0 \\ \int_{-\infty}^{\infty} p_X(u)\, du &= 1 \end{aligned}$$

The PDF and related CDF (one can be derived from the other) allow us to compute the probability of any event that is defined as a sample of the continuous random variable falling in an interval. As an example, if we wish to compute $P(a < X \leq b)$ then we simply integrate the PDF over the appropriate region

$$\begin{aligned} P(a < X \leq b) &= \int_a^b p_X(u)\, du \\ &= P_X(b) - P_X(a) \end{aligned}$$

If the event invovles multiple **disconnected** regions, we simply integrate over each region and add the results

$$P(X \in \{R_1, \cdots, R_N\}) = \sum_{i=1}^{N} \int_{R_i} p_X(u)\, du$$

10.4.3 Expectation

Given a function of a random variable (i.e. $g(X)$) we define it's expected value as:

$$E\{g(X)\} = \begin{cases} \sum_{i=1}^{N} g(x_i) p_X(x_i) & ; X \text{ discrete} \\ \int_{-\infty}^{\infty} g(u) p_X(u)\, du & ; X \text{ continuous} \end{cases} \tag{17}$$

Example functions include

$g(X)$	symbol	statistic
X	μ_x	mean
$(X-\mu_x)^2$	σ_x^2	variance
$(X-\mu_x)^n$	ν_n	nth central moment
$-\log(p_X(X))$	$H(X)$	entropy

Expectation is linear

$$E\{\alpha f(x) + \beta g(x)\} = \alpha E\{f(x)\} + \beta E\{g(x)\} \tag{18}$$

Expectation is with regard to ALL random variables within the arguments. This is important for multi-dimensional and joint random variables.

10.5 Multi-Dimensional Random Vectors

We can define joint probability density and cumulative distribution functions over multiple random variables in a similar fashion as we did for a single random variable. A collection of multiple random variables is known as a random vector. For a two dimensional random vector we define the probability of the event $\{X_1 \leq x_1 \text{ AND } X_2 \leq x_2\}$ as a function of x_1 and x_2. The joint density

is the function we integrate to compute the probability of this event.

$$P_{X_1X_2}(x_1, x_2) = \Pr\{X_1 \leq x_1 \text{ AND } X_2 \leq x_2\} \quad (19)$$

$$= \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} p_{X_1X_2}(u_1, u_2)\, du_2 du_1 \quad (20)$$

$$p_{X_1X_2}(x_1, x_2) = \frac{\partial^2}{\partial x_1 \partial x_2} P_{X_1X_2}(x_1, x_2) \quad (21)$$

Similarly, we can define a N-dimensional joint density and cumulative distribution function over N variables.

$$P(x_1, \ldots, x_N) = \Pr\{X_1 \leq x_1 \text{ AND } X_2 \leq x_2 \ \ldots \ \text{AND } X_N \leq x_N\} \quad (22)$$

$$= \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_N} p(u_1, \ldots, u_N)\, du_N \ldots dx_2 du_1 \quad (23)$$

$$p(x_1, \ldots, x_N) = \frac{\partial^N}{\partial x_1 \ldots \partial x_N} P(x_1, \ldots, x_N) \quad (24)$$

10.5.1 Marginal Densities and Mass Functions

Given 2 random variables X_1 and X_2 and associated CDF $P(x_1, x_2)$ and PDF $p(x_1, x_2)$ (or PMF), the PDF (or PMF) over *just one* of the variables is referred to as the **marginal** distribution or density over that variable. The marginal PDF of X_1 is computed by integrating the joint PDF of X_1 and X_2 over x_2

$$p(x_1) = \int_{-\infty}^{\infty} p(x_1, x_2) dx_2 \quad (25)$$

Note that if X_2 were a discrete random variable, then we would sum over all the possible values of X_2 rather than integrate. The marginal density of X_2 is computing similarly by integrating (or summing) over x_1. A similar form is used for the multidimensional case. For example, given a joint density over four variables $p(x_1, x_2, x_3, x_4)$ if we wish to compute the marginal density $p(x_2, x_4)$ we integrate the joint density over the *remaining* variables as:

$$p(x_2, x_4) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x_1, x_2, x_3, x_4) dx_1 dx_3$$

We can derive the formula above by recalling that the PDF of a random variable is defined as the derivative of the CDF which in turn defines the probability of a specific event over that variable. The equivalent definition for joint random variables (where we have reinstated subscripts for

clarity) is derived as

$$\begin{aligned}
p_{X_1}(x_1) &= \frac{\partial}{\partial x_1} P_{X_1}(x_1) \\
&= \frac{\partial}{\partial x_1} \left(\Pr\{X_1 \leq x_1\}\right) \\
&= \frac{\partial}{\partial x_1} \left(\Pr\{X_1 \leq x_1 \text{ AND } X_2 \text{ equal to anything}\}\right) \\
&= \frac{\partial}{\partial x_1} \left(\Pr\{X_1 \leq x_1 \text{ AND } -\infty \leq X_2 \leq \infty\}\right) \\
&= \frac{\partial}{\partial x_1} \left(\int_{-\infty}^{x_1} \int_{-\infty}^{\infty} p_{X_1,X_2}(u_1, x_2) dx_2 du_1\right) \\
&= \int_{-\infty}^{\infty} p_{X_1,X_2}(x_1, x_2) dx_2
\end{aligned}$$

By similar reasoning, one can derive the marginal density of a subset of variables of an N-dimensional random vector. Again, for component of the random vector

10.5.2 Conditional Densities

In the case of two random variables we are interested in the resulting density when we condition one variable on another. This is especially useful for inference and estimation problems. For the two dimensional case this density has the form

$$p(x_1|x_2) = \frac{p(x_1, x_2)}{p(x_2)} \tag{26}$$

$$= \frac{p(x_1, x_2)}{\int_{-\infty}^{\infty} p(x_1, x_2)\, dx_1} \tag{27}$$

More generally in the multi-dimensional case the conditional density is the ratio of the joint density to the marginal density of the conditioning variables. For example, given a joint density over four variables $p(x_1, x_2, x_3, x_4)$ if we wish to compute the conditional density $p(x_2, x_4|x_1, x_3)$ we form the ratio of the joint density to the marginal of the *conditioning* variables as:

$$\begin{aligned}
p(x_2, x_4|x_1, x_3) &= \frac{p(x_1, x_2, x_3, x_4)}{p(x_1, x_3)} \\
&= \frac{p(x_1, x_2, x_3, x_4)}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x_1, x_2, x_3, x_4)\, dx_2 dx_4}
\end{aligned}$$

From the two examples above we see that the joint can always be expressed as product of a conditional density and the marginal of the conditioning variables.

$$p(x_1, x_2) = p(x_1)\, p(x_2|x_1) \tag{28}$$

$$= p(x_2)\, p(x_1|x_2) \tag{29}$$

$$p(x_1, x_2, x_3, x_4) = p(x_1, x_3)\, p(x_2, x_4|x_1, x_3)$$

10.5.3 Bayes' Rule

Examining the relationship between joint and conditional densities results in Bayes' rule as applied to densities. Noting that

$$p(x_1)\,p(x_2|x_1) \;=\; p(x_2)\,p(x_1|x_2)$$

results in what is also known as Bayes' rule

$$p(x_2|x_1) \;=\; \frac{p(x_2)\,p(x_1|x_2)}{p(x_1)} \tag{30}$$

$$p(x_1|x_2) \;=\; \frac{p(x_1)\,p(x_2|x_1)}{p(x_2)} \tag{31}$$

10.5.4 Independent Random Variables

Two random variables X_1 and X_2 are statistically independent if their joint density can be expressed as a product of their marginal densities

$$p(x_1, x_2) \;=\; p(x_1)p(x_2)$$

More generally for the elements of a random vector are statistically independent if they can be expressed in the form

$$p(x_1, \ldots, x_N) \;=\; \prod_{i=1}^{N} p(x_i) \tag{32}$$

This is a stronger statement of statistical independence as it implies that the random variables are statistically independent over *all* events.

10.5.5 Expectation

Expectation over random vectors is defined in the same way as for the scalar case. That is

$$E\left\{g(x_1, \cdots, x_N)\right\} \;=\; \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \cdots, x_N)\,p(x_1, \cdots, x_N)\,dx_1 \cdots dx_N \tag{33}$$

It is important to remember that unless otherwise specified, expectation is taken with regard to every random variable within the brackets. As in the scalar case, expectation over random vectors is linear in the following sense

$$E\left\{\alpha f(x_1) + \beta g(x_2) + \gamma h(x_1, x_2)\right\} \;= \alpha E\left\{f(x_1)\right\} + \beta E\left\{g(x_2)\right\} + \gamma E\left\{h(x_1, x_2)\right\} \tag{34}$$

The derivation follows

$$
\begin{aligned}
E\left\{\alpha f\left(x_1\right)+\beta g\left(x_1\right)+\gamma h\left(x_1, x_2\right)\right\} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\big(\alpha f\left(x_1\right)+\beta g\left(x_1\right) \\
&\qquad +\gamma h\left(x_1, x_2\right)\big) p_{X_1 X_2}\left(x_1, x_2\right) dx_1 dx_2 \\
&= \int_{-\infty}^{\infty} \alpha f\left(x_1\right) p_{X_1}\left(x_1\right) \underbrace{\int_{-\infty}^{\infty} p_{X_2|X_1}\left(x_2|x_1\right) dx_2}_{1} dx_1 + \\
&\quad \int_{-\infty}^{\infty} \beta g\left(x_2\right) p_{X_2}\left(x_2\right) \underbrace{\int_{-\infty}^{\infty} p_{X_1|X_2}\left(x_1|x_2\right) dx_1}_{1} dx_2 + \\
&\quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \gamma h\left(x_1, x_2\right) p_{X_1 X_2}\left(x_1, x_2\right) dx_1 dx_2 \\
&= \alpha \int_{-\infty}^{\infty} f\left(x_1\right) p_{X_1}\left(x_1\right) dx_1 + \beta \int_{-\infty}^{\infty} g\left(x_2\right) p_{X_2}\left(x_2\right) dx_2 + \\
&\quad \gamma \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h\left(x_1, x_2\right) p_{X_1 X_2}\left(x_1, x_2\right) dx_1 dx_2 \\
&= \alpha E\left\{f\left(x_1\right)\right\}+\beta E\left\{g\left(x_1\right)\right\}+\gamma E\left\{h\left(x_1, x_2\right)\right\}
\end{aligned}
$$