Basic Probability Notes

Sravana Reddy

Revised January 2016

1 Introduction

A probability model consists of a **sample space** (which in turn defines a set of **events**), and a **probability function**, also known as a **probability distribution**.

1.1 Sample Space

The sample space Ω is a set of basic outcomes. For example,

- Sample space for a single coin toss: {H, T}
- Sample space for tossing two coins: {HH, TT, HT, TH}
- Sample space for tossing three coins: {HHH, HHT, HTH, HTT, THH, THT, TTH, TTH }
- Sample space for picking a word: set of all words in vocabulary.
- Sample space for picking a sequence of 2 words: set of all pairs of words {(the, the), (the, and), (and, the), (and, and), ...}.

1.2 Event

An **atomic event** is a single member of the sample space. (The event of getting HHH on a 3-coin toss.)

An **event**, generally, is a collection of atomic events, i.e., a subset of the sample space. (The event of getting at least two heads, which is {HHH, HHT, HTH, THH}.)

The event space is the set of all possible events. By this definition, it is the power set of the sample space, and has size $2^{|\Omega|}$. ($|\Omega|$ denotes the size of the sample space Ω.)

1.3 Probability function

A probability function is defined on the sample space.

$$P : \Omega \rightarrow [0, 1] \tag{1}$$

which says that the domain of the function is Ω (the sample space) and the range is all real numbers from 0 to 1. The function must also obey the constraint that the sum of the probabilities of all the atomic events should be 1

$$\sum_{\omega \in \Omega} P(\omega) = 1 \tag{2}$$

1.3.1 Continuous and discrete distributions

Probability functions can be continuous or discrete. A continuous function is one where the sample space is continuous. For example, if you have to build a probability distribution of the amount of rain on a given day, you will need to define probabilities for every real number value for inches of rainfall. In language, we generally do not deal with continuous distributions, since our sample space consists of discrete linguistic units. However, they come up in other related domains like speech processing or computer vision.

1.3.2 Uniform and non-uniform distributions

A uniform distribution assigns equal probability to every atomic event. For example, in a 3-coin toss, all 8 possibilities are equally likely, so $P(\omega) = 1/8$ for all ω.

With uniform distributions, we can derive the probability of a non-atomic event A simply by counting. That is,

$$P(A) = \frac{|A|}{|\Omega|} \tag{3}$$

The probability of getting at least two heads is

$$P(\{HHH, HHT, HTH, THH\}) = 4/8 = 0.5 \tag{4}$$

When the distribution is *not* uniform, we need to invoke the principles of how probabilities combine arithmetically for different combinations of events.

2 Conditional Probabilities and Independent Events

What is the probability that it snows on any given day of the year? Somewhat low. Let's denote this by $P(s)$.

What is the probability that it snows given that it's a winter day (in Boston)? That's quite a bit higher. We denote this probability by $P(s|w)$, read "probability of s given w."

On the other hand, the probability that you aced a test given that it's a winter day in Boston, denoted by $P(a|w)$, is likely no different from the probability that you aced a test at any time, denoted by $P(a)$. That is, $P(a|w) = P(a)$, which means that it being winter does not affect how you do on a test. When this happens, we say the events a and w (acing a test and it being a winter day) are **independent**.

3 Computing Complex Event Probabilities

A good rule of thumb is that the intersection of events ('and') implies multiplication of the individual event probabilities, and the union ('or') implies addition – with some important caveats.

3.1 And = Multiplication under Independent Events

The probability of $P(A \cap B \cap C)$ is the probability of the *intersection* of multiple events. As a shorthand, we often write it as $P(A, B, C)$, replacing the intersection symbol $\cap$ by a comma. This is also known as the **joint** probability of the events A, B, and C.

When A, B, and C are independent, the probability is given by

$$P(A \cap B \cap C) = P(A)P(B)P(C) \tag{5}$$

With a biased coin where $P(H) = 0.7$:

$$\begin{aligned} P(HH) \text{ (probability of a 2-coin toss coming up both heads)} \\ = P(H)P(H) = 0.7^2 = 0.490 \\ P(HHH) = P(H)P(H)P(H) = 0.7^3 = 0.343 \\ P(TTT) = P(T)P(T)P(T) = 0.3^3 = 0.027 \\ P(HTT) = P(H)P(T)P(T) = 0.7 * 0.3^2 = 0.063 \end{aligned}$$

I reiterate that this rule of $P(AB) = P(A)P(B)$ that we have used in the coin examples is only valid because the coin tosses are independent of one another. For most real world events, the independence assumption may not be true.

For example, looking at the event that it is sunny on a given day, the probability that it is sunny tomorrow (S_1) is not independent of the event that it is sunny today (S_0). Let $P(S_0, S_1)$ denote the probability that it is sunny today and tomorrow. Then,

$$P(S_0, S_1) = P(S_0)P(S_1|S_0) \tag{6}$$

where $P(S_1|S_0)$ is the probability that it is sunny tomorrow *given* that it is sunny today. Keep in mind that the following equation also holds true.

$$P(S_0, S_1) = P(S_1)P(S_0|S_1) \tag{7}$$

Which way that you factor a joint probability into a product of conditional probabilities depends only on how much information you are given. In the cases of chronological events, or reading left to right in a text, it just happens to be that the probability of something happening in the future will be conditioned on the past, rather than the other way around.

If S_1 is independent of S_0, it follows by definition that $P(S_1|S_0) = P(S_1)$. Eq. 6 becomes

$$P(S_0, S_1) = P(S_0)P(S_1) \tag{8}$$

which is the same rule for the joint probability of independent events that you saw earlier.

3.2 Or = Addition under Disjoint Events

The probability of $A \cup B \cup C$ – a *union* of multiple events – can be derived by

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) \tag{9}$$

when the events are **disjoint** from one another: that is, they share no common outcomes. For example, under the same 0.7/0.3 biased coin:

$$P(H \cup T) \text{ (probability of a coin toss coming up heads or tails)}$$
$$= P(H) + P(T) = 0.7 + 0.3 = 1.0$$
$$P(HHH \cup TTT) \text{ (probability of a 3-coin toss coming up all H or all T)}$$
$$= P(HHH) + P(TTT) = 0.7^3 + 0.3^3 = 0.370$$
$$P(\text{first coin is T}) = P(THH \cup THT \cup TTH \cup TTT)$$
$$= 0.3 \cdot 0.7^2 + 0.3 \cdot 0.7 \cdot 0.3 + 0.3^2 \cdot 0.7 + 0.3^3 = 0.300$$
$$P(\text{exactly two H}) = P(HHT \cup HTH \cup THH)$$
$$= 0.7^2 \cdot 0.3 + 0.7 \cdot 0.3 \cdot 0.7 + 0.3 \cdot 0.7^2 = 0.450$$
$$P(\text{exactly two T}) = P(TTH \cup THT \cup HTT)$$
$$= 0.3^2 \cdot 0.7 + 0.3 \cdot 0.7 \cdot 0.3 + 0.7 \cdot 0.3^2 = 0.063$$
$$P(\text{exactly two H or exactly two T}) = P(\text{exactly two H}) + P(\text{exactly two T}) = 0.513$$

However, if we are looking for the probability of 'exactly two H or the first coin is T', we cannot simply sum the probability of 'exactly two H' and the probability of 'first coin is T', since the events share a common outcome: namely, `THH`.

Sample space	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
Exactly 2 H		x	x		x			
First coin is T					x	x	x	x

Notice that summing $P(\text{exactly two H}) + P(\text{first coin is T})$ amounts to double-counting the probability at the intersection. So,

$$P(\text{exactly two H or first coin is T})$$
$$= P(\text{exactly two H}) + P(\text{first coin is T}) - P(\text{exactly two H and first coin is T})$$
$$= P(\text{exactly two H}) + P(\text{first coin is T}) - P(THH)$$
$$= 0.450 + 0.300 - 0.147 = 0.603$$

3.3 Marginalization

Consider the situation when you are given the joint probabilities of me eating ice-cream and it being the winter, spring, fall, and summer respectively: $P(i, w)$, $P(i, f)$, $P(i, sp)$, and $P(i, su)$. What is the probability of me eating ice-cream on any given day, regardless of

the weather? We often call this the problem of **marginalizing** over events – in this case, the different seasons. It involves taking the union of the joint probabilities for each of the different seasons, assuming that they're disjoint. That is,

$$P(i) = P(i, w) + P(i, f) + P(i, sp) + P(i, su) \tag{10}$$

4 Bayes Rule

Bayes Rule follows from the factorization of joint probability.

$$P(A \cap B) = P(A)P(B|A)$$
$$P(A \cap B) = P(B)P(A|B)$$

Therefore,

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

4.1 Application to Learning

Bayes Rule comes up pretty much everywhere! A notable appearance is when we want to estimate the probability of a model θ given observed data X (for example, the probability of having a disease given a positive test, or the probability of a certain grammar given a set of English sentences).

$$P(\theta|X) = \frac{P(\theta)P(X|\theta)}{P(X)} \tag{12}$$

The factor $P(\theta)$ is known as the **prior**: i.e., our prior estimate of the model that we may have from world knowledge – we may know that a disease is rare and $P(\theta) = 0.01$. When we have no world knowledge, we set the prior to be uniform: $P(\theta)$ equally likely for all θ. For example, if we do not know whether a disease is rare, we will set $P(\text{disease}) = P(\text{no disease}) = 0.5$.

$P(X|\theta)$ is known as the **likelihood** of the model. This is the probability that θ assigns to the observed data X.

The quantity on the left-hand side, $P(\theta|X)$, is the **posterior** estimate of the model given the observed data.