Probability Basics

A Computational Introduction to Robotics

Fall 2020

Learning Objectives

- Define the concept of a probability.
- Learn Bayes' rule and other rules for manipulating probabilities.

1 Motivation

It turns out that the notion of probability provides us with a powerful method to formalize the concept of uncertainty and allows us to reason about many types of uncertainty in a unified way. Formalizing uncertainty using probability theory will enable us to solve robotics challenges in a principled and robust manner.

- make explicit our assumptions about the uncertainties present within our sensor data and the impact of these uncertainties on predictions based on that data.
- allow us to quantify confidence in our predictions

2 Probability

Hopefully the previous section left you feeling excited to learn more about the theory that underlies these big ideas. Next, we'll take our first steps towards learning this theory.

2.1 Intuition

Most of us are used to thinking that events can be probabilistic, that is we can attach some probability to whether or not they occur. Take for example flipping a coin. We could think of the event that the coin comes up heads as having probability 0.5. That is, there is an even chance that it happens versus doesn't happen. Further, we can say that an event is *observable* if we are able to directly observe whether it occurred. For instance, whether a coin comes up heads is an observable event since you can ultimately observe the outcome of the flip. In contrast, some events would be considered unobservable if they are unable to be directly ascertained by human senses. A classic example of this would be whether a scientific theory is true or not. It is impossible to directly observe whether the theory is true, but you might be able to observe events that are consistent or inconsistent with the theory.

Exercise 1 (10 minutes)

Come up with 3ish examples of observable events that are probabilistic in nature. For each event, provide the probability that it occurs or explain what factors would determine the probability. Some potential ideas to get you going: sporting events, elections, weather, etc.

2.2 Formal Definition

Next, we'll define more formally[1] what we mean by a probability. Having this formal definition will give us the ability to determine useful rules for manipulating and reasoning about probabilities. To define the concept of a probability, we'll need to specify two ingredients.

[1] this is not the full definition of a probability space used in modern mathematics. For the purposes of most people that *use* probability theory on actual problems, the full definition is needlessly complex. For instance, as a grad student I (Paul) never saw the full definition in any of my ML courses. Use the following link if you want a more in-depth discussion of the parts of the formal definition that are tricky (our expectation is that you won't want this discussion!).

2.3 Events

An *event* is something that may or may not occur in response to some random process. For instance, we could define the event that a coin comes up heads when it is flipped. We often use capital letters to indicate events. Since we've been using capital letters to also represent matrices, in our materials we'll use a cool mathy-looking calligraphic font to represent events. For instance, we might use the symbol $\mathcal{H}$ to refer to the event that a coin flip comes up heads. It's important to emphasize that a single random process can have many associated events. For instance, for the coin flip example we might also define $\mathcal{T}$ to be the event that the coin comes up tails (or $\mathcal{U}$ to indicate that event that the coin rotated at least 10 times in the air when we flipped it).

Further, events don't necessarily have to be mutually exclusive. For instance, we might define the event $\mathcal{R}_h$ to indicate the event that the Republican party controls the majority in the House of Representatives following the 2020 election and $\mathcal{D}_s$ to indicate the event that the Democratic party controls the majority in the Senate following the 2020 election. Both (or none) of these events could occur.

2.4 Probability Measure Function

The probability measure function assigns a probability to the occurrence of any particular event. We can think of this probability measure function as taking as input an event and outputting a probability. For instance, $p(\mathcal{E})$ provides the probability that event $\mathcal{E}$ occurs according to probability measure p. All probability measure functions must satisfy the following properties.

- $0 \leq p(\mathcal{E}) \leq 1$: the probability of an event ranges from 0 (an impossible event) to 1 (an event that will always occur).
- Given a set of n events $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n$ that are disjoint (i.e., no two can occur simulta-

neously)

$$p(\mathcal{E}_1 \text{ or } \mathcal{E}_2 \text{ or } \ldots \text{ or } \mathcal{E}_n) = \sum_{i=1}^{n} p(\mathcal{E}_i) \ . \tag{1}$$

The equation above specifies what is sometimes called the union rule of probability. It states that the probability of one of these disjoint events occurring must be equal to the sum of the probability of each of the events occurring. You will also sometimes see Equation 1 written as

$$p(\mathcal{E}_1 \cup \mathcal{E}_2 \cup \ldots \cup \mathcal{E}_n) = \sum_{i=1}^{n} p(\mathcal{E}_i) \ . \tag{2}$$

If you're not familiar with the symbol $\cup$ it is the symbol for a union of two sets. The reason you'll sometimes see this notation is that in the full definition of a probability space an event is defined as a set (as stated in the margin note earlier in this section, you need not worry about the most rigorous definition in this course).

- Given a set of (not necessarily disjoint) events $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n$ where at least one of these n events must occur

$$p(\mathcal{E}_1 \text{ or } \mathcal{E}_2 \text{ or } \ldots \text{ or } \mathcal{E}_n) = 1 \ . \tag{3}$$

This rule just states that if we have an exhaustive set of events (that cover all possible cases), at least one of them *must* occur.

2.5 Complement Rule for Probability

Given the definition of probability detailed above, it follows that if the probability of an event happening is $p(\mathcal{E})$ then the probability of the event NOT happening is $1 - p(\mathcal{E})$. The following are common ways of to express this relationship (we'll use Equation 4 in this class). These all say the same thing (the only difference is notation).

$$\begin{aligned} p(\neg\mathcal{E}) &= 1 - p(\mathcal{E}) \\ p(\text{not } \mathcal{E}) &= 1 - p(\mathcal{E}) \\ p(\overline{\mathcal{E}}) &= 1 - p(\mathcal{E}) \\ p(\mathcal{E}') &= 1 - p(\mathcal{E}) \\ p(\mathcal{E}^c) &= 1 - p(\mathcal{E}) \end{aligned} \tag{4}$$

We point out these alternate notations not to confuse you (we'd never do that!) but to help you interpret various external resources you might find on these topics.

Exercise 2 (10 minutes)

Here are some diagnostic questions to make sure that you got the basic ideas.

(a) Suppose $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ are disjoint events. Further, suppose that one of these events must occur. Which of the following functions are valid probability measure functions?

$$p_1(\mathcal{E}_1) = \frac{1}{10}, p_1(\mathcal{E}_2) = \frac{1}{5}, p_1(\mathcal{E}_3) = \frac{7}{10}$$

$$p_2(\mathcal{E}_1) = \frac{11}{10}, p_2(\mathcal{E}_2) = \frac{-1}{10}, p_2(\mathcal{E}_3) = 0$$

$$p_3(\mathcal{E}_1) = \frac{1}{10}, p_3(\mathcal{E}_2) = \frac{1}{5}, p_3(\mathcal{E}_3) = \frac{1}{2}$$

$$p_4(\mathcal{E}_1) = 1, p_4(\mathcal{E}_2) = 0, p_4(\mathcal{E}_3) = 0$$

(b) The Birthday Problem is a well-known probability problem often used in discrete math courses. According to the Wikipedia article on the Birthday Problem, the probability that at least two students among the 70 students in machine learning this semester share the same birthday is 0.999. What is the probability that no two students share the same birthday?

3 Bayes' Rule

External Resource(s) (60 minutes)

Learning Objectives

Note that these learning objectives have been written to be very specific (based on feedback from the course survey). When you first read them, you probably won't know what they mean in detail. As you go through the readings, hopefully the more precise statement of these learning objectives will be useful for assessing your understanding of the provided resources.

- When Bayes' rule is useful (i.e., when $p(A|B)$ is easier to work with than $p(B|A)$).
- The idea of a conjoint probability $p(\mathcal{A}, \mathcal{B})$ (note: alternate notations include $p(\mathcal{A}$ and $\mathcal{B})$ and $p(\mathcal{A} \cap \mathcal{B})$).
- The definition of a conditional probability $p(\mathcal{A}|\mathcal{B}) = \frac{p(\mathcal{A},\mathcal{B})}{p(\mathcal{B})}$.
- The equation for the product rule $p(\mathcal{A}, \mathcal{B}) = p(\mathcal{B})p(\mathcal{A}|\mathcal{B}) = p(\mathcal{A})p(\mathcal{B}|\mathcal{A})$ (Allen calls this the probability of a conjunction).
- The equation for Bayes' rule $p(\mathcal{A}|\mathcal{B}) = \frac{p(\mathcal{B}|\mathcal{A})p(\mathcal{A})}{p(\mathcal{B})}$.

Allen Downey (ever heard of him?) wrote an excellent book called Think Bayes that introduces Bayesian analysis. The first chapter (which you should read) starts with a less formal definition of probability than the one we gave earlier. The chapter then gives intuitions around conjoint probability (the probability that multiple events occur simultaneously), conditional probability (the probability that some event occurs conditioned on another event having occurred), and finally to Bayes' rule (a surprisingly easy theorem to derive that allows you to write one conditional probability distribution in terms of another). The Monty Hall problem in section 1.7 is probably okay to skim (see Allen's note at the end of that section for why this is the case).

Allen's treatment of the material is, of course, not the only one out there (we like it for its focus on building intuition and focusing on the key ideas). Here are some other resources you might consider checking out (they are optional).

- Khan Academy Video on Bayes' Theorem shows some simple applications of Bayes' rule and explains why it is a convenient way to reason about the probability of hypothesis given data).
- Veritasium Episode on Bayes' Theorem has a bit more history and philosophy of Bayes' Theorem along with some nice visualizations. It also includes the presenter walking on a very scenic mountain (for some reason), so there's that if nothing else.
- Julia Galef's video A Visual Guide to Bayesian Thinking

- I (Paul) ran across this example of applying Bayes' rule to a real world problem. It was created by a grad school friend of mine and is hilarious (lots of Cat Memes). I did notice that there is a mistake in the math at the 8:12 mark in the video (he states that $p(\text{alarm}|\text{no theft}) = 1 - p(\text{alarm}|\text{theft})$, which is not necessarily the case). It's still a good video though.

Exercise 3 (20 minutes)

⚠ Notice

If you are having trouble dealing with the denominator that you get when you apply Bayes' rule, you may want to skip ahead to the section on the *Marginalization Rule for Probabilities* and then return to these problems.

(a) You are given three coins. Two are of type 1, we'll call this type $\mathcal{C}_1$, and one is of type 2, we'll call this type $\mathcal{C}_2$. If you flip a coin of type 1, the coin will come up heads with probability $\frac{4}{5}$ (i.e., $p(\mathcal{H}|\mathcal{C}_1) = \frac{4}{5}$). If you flip a coin of type 2, the coin will come up heads with probability $\frac{1}{2}$ (i.e., $p(\mathcal{H}|\mathcal{C}_2) = \frac{1}{2}$). Suppose you choose one of the three coins (there is no way for you to tell them apart), flip it once, and it comes up heads. What is the probability that you flipped a coin of type 2 (i.e., what is $p(\mathcal{C}_2|\mathcal{H})$)? After you compute your answer, compare it with the probability that a randomly selected coin was of type 2 (before you flipped it and observed heads). Does the relationship between this prior probability and the posterior probability make sense?

(b) You train a neural network to identify whether an image contains a picture of a Chihuahua or a Blueberry Muffin (you know you want to click the link!). Let's further assume that there are no images that contain both a muffin and a Chihuahua (of course as we all know Chihuahuas love muffins). Based on your project report for module 1, you know that if the image contains a Chihuahua, your model will identify it as such with probability 0.9. Also, you know that if the model contains a muffin, your model will identify it as such with probability 0.8. You now decide to test your model on a dataset which contains 80% muffins. Assuming that your model predicted that an image contained a muffin, what is the probability that it actually contains a muffin? (you can assume that the performance of your model doesn't change when run on this new dataset).

4 Marginalization Rule for Probabilities

The application of Bayes' rule often proceeded according to the following outline. First, we would define an event we want to reason about. For instance, we might define $\mathcal{D}$ as the event that a person has a disease and $\mathcal{S}$ as the event that a particular symptom is

observed. If we want to know $p(\mathcal{D}|\mathcal{S})$ we apply Bayes' rule like so.

$$p(\mathcal{D}|\mathcal{S}) = \frac{p(\mathcal{S}|\mathcal{D})p(\mathcal{D})}{p(\mathcal{S})} \tag{5}$$

In order to calculate $p(\mathcal{S})$, some of the resources simply gave a number (e.g., in the Khan Academy video the premise was that you Googled to find this value), used a convenient trick to get it (as in Allen's M&M example), or used the following calculation (as in the Veritasium and the Car Alarm videos).

$$p(\mathcal{S}) = p(\mathcal{D})p(\mathcal{S}|\mathcal{D}) + p(\neg\mathcal{D})p(\mathcal{S}|\neg\mathcal{D}) \tag{6}$$

We wanted to revisit this calculation as it is hiding away some pretty powerful and interesting stuff. This calculation can be derived using the technique of marginalizing a probability measure function. The basic motivation for this technique is that sometimes you'd like to compute the probability of some event, $\mathcal{A}$, but it is difficult to do so directly. Instead you can introduce another event, $\mathcal{B}$, and write $p(\mathcal{A})$ as:

$$p(\mathcal{A}) = p(\mathcal{A}, \mathcal{B}) + p(\mathcal{A}, \neg\mathcal{B}) \tag{7}$$

In the equation above we sometimes say that we are *marginalizing out* $\mathcal{B}$ (by summing over the two possibilities: that $\mathcal{B}$ occurred and that $\mathcal{B}$ did not occur).

Exercise 4 (15 minutes)

Using Equation 7 and the product rule of probability (also called the conjunction rule), show that Equation 6 is true. Remember that the product rule states that $p(A, B) = p(A)p(B|A)$ or, equivalently, $p(A, B) = p(B)p(A|B)$.

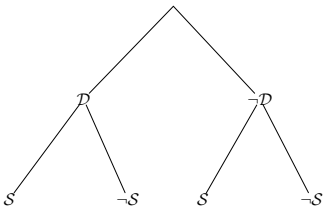
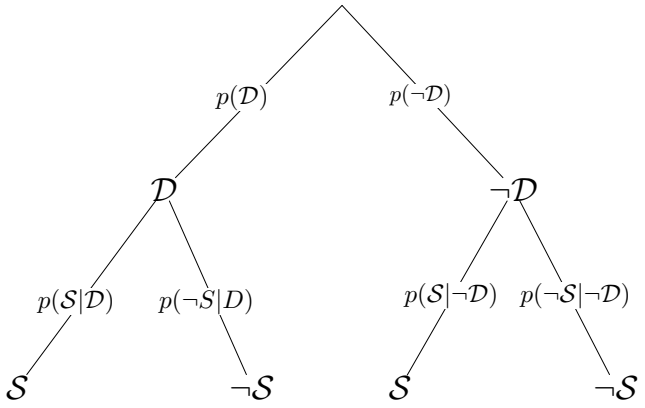


Figure 1: A tree diagram of the events $\mathcal{D}$ (has a disease) and $\mathcal{S}$ (has a symptom).

Another way to think about marginalization is to draw a tree where you have the event, which you are marginalizing out ($\mathcal{D}$ in the previous exercise) at the first level of the tree and the variable you want to know the probability of ($\mathcal{S}$ in the previous exercise) at the next junction in the tree (see Figure 1).

Further, we annotate the arrows with the conditional probability of the event conditioned on the things further up in the tree (note that for $\mathcal{D}$ there is nothing further up the tree, so we just write $p(\mathcal{D})$ or $p(\neg\mathcal{D})$).



If you want to find a joint probability (e.g., $p(\mathcal{D}, \neg\mathcal{S})$), follow the corresponding path, multiplying probabilities as you go. For example, from examining the graph above we get

$$p(\mathcal{D}, \neg\mathcal{S}) = p(\mathcal{D})p(\neg\mathcal{S}|\mathcal{D}) \ . \tag{8}$$

A marginal probability for an event (e.g., $p(\mathcal{S})$) can be found by summing over all paths that arrive at the event. For example, from examining the graph above we see that there are two paths to $\mathcal{S}$, which allows us to compute $p(\mathcal{S})$ as

$$p(\mathcal{S}) = p(\mathcal{D})p(\mathcal{S}|\mathcal{D}) + p(\neg\mathcal{D})p(\mathcal{S}|\neg\mathcal{D}) \ . \tag{9}$$

Exercise 5 (20 minutes)

Do problem 2 from this assignment. They use $\mathcal{E}'$ to refer to the event $\neg\mathcal{E}$.