3

PROBABILITY

Probability theory is a field devoted to estimating how likely it is that some particular event will happen. A probability of 0% means that the event definitely will not happen, whereas 100% means that it's a sure thing. We can also use probability to express confidence, such as being 80% sure that a piece of fruit is ripe, or that a certain team will win a game.

Probability is one of the pillars on which machine learning is built. Many papers describe their techniques with the language of probability, and lots of documentation follows suit. Library functions can require their input data to have some basic probabilistic properties. Understanding the accuracy and behavior of the systems we build can involve understanding the probabilities of the results they produce.

Probability theory is an enormous subject, with many deep specialties. Since our focus is on using machine learning tools sensibly, we only need command of a few basic terms and topics: different kinds of probability,

how to measure correctness, and a particular way of organizing probabilities called the *confusion matrix*. With a command of these basic ideas, we'll be able to prepare our data to get the best performance out of the tools we'll be using later. Broader and deeper discussions on all the topics we'll cover here, as well as many other topics in this field, may be found in books dedicated to probability (Jaynes 2003; Walpole et al. 2011).

Different Types of Probability

There are many types of probability. We'll discuss a few of them here, beginning with a metaphor.

Dart Throwing

Dart throwing is the classic metaphor for discussing basic probability. The fundamental idea is that we're in a room with a bunch of darts in our hand, facing a wall. Instead of hanging a cork target, we've painted the wall with some blobs of different colors and sizes. We'll throw our darts at the wall, and we'll track which colored region each one lands in (the background counts as a region as well). The idea is illustrated in Figure 3-1.

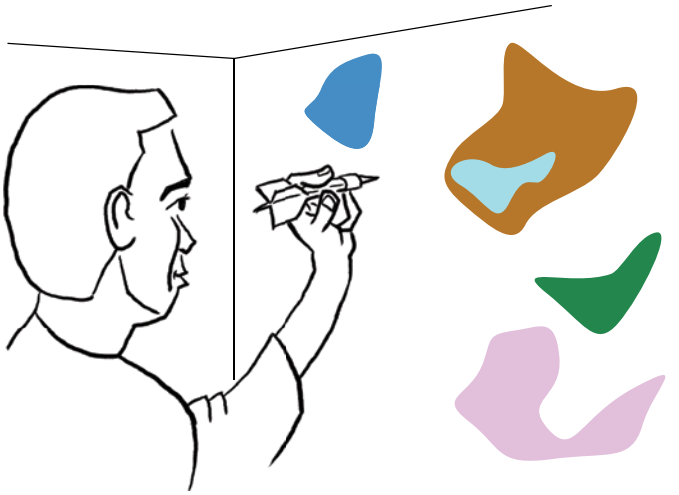


Figure 3-1: Throwing darts at a wall. The wall is covered in blobs of paint of different colors.

We're going to assume from now on that our darts will always strike the wall somewhere (rather than going into the floor or ceiling, for instance). So the probability of each dart striking the wall *somewhere* is 100%. We'll use both floating-point (or real) numbers and percentages for probabilities, so a probability of 1.0 would be a percentage of 100%, a probability of 0.75 would be a percentage of 75%, and so on.

Let's look more closely at our dart-throwing scenario. In the real world, we're more likely to hit the part of the wall that's directly in front us, rather

than, say, something well off to the side. But for the purpose of this discussion, we're going to assume that the probability of our hitting the wall at any point is the same *everywhere*. That is, *every point on the wall has the same chance of being hit by a dart*. Using the language of Chapter 2, we could also say that the probability of striking any given point is given by a uniform distribution.

The heart of the rest of the discussion will be based on comparing the areas of the various regions, and our chances of striking each of those areas. Remember that the background counts as a region (in Figure 3-1, it's the white region).

Here's an example. Figure 3-2 shows a red square on the wall. When we throw a dart, we know it will hit the wall somewhere, with a probability of 1.

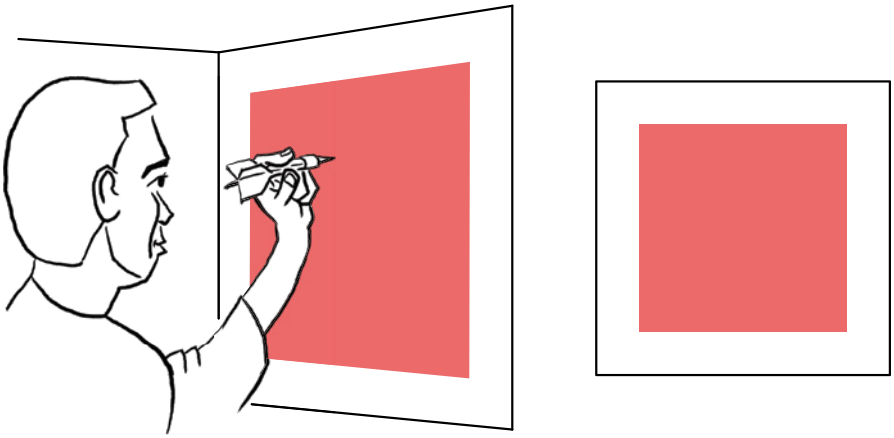


Figure 3-2: We're guaranteed to hit the wall. What's the probability that we'll hit the red square?

What's the probability of hitting the red square? In this figure, the square covers half of the wall's total area. Since our rule is that every point on the wall has an equal likelihood of being hit, when we throw our dart, we have a 50% chance, or a probability of 0.5, of the dart landing in the red square. The probability is just the ratio of the areas. The larger our square, the more points it encloses, and so the more likely it is that we'll land inside of it.

We can illustrate this with a little picture that draws the ratios of the areas. Figure 3-3 shows the ratio for our square with respect to the wall. This kind of diagram, where we draw a "fraction" of one shape above the other, gives us a visual way to track which areas we're talking about and get an intuitive feel for their relative sizes.

Figure 3-3 shows the relative areas accurately, so the area of the red square is really half the area of the white box under it. Using the full-size shapes can make for awkward diagrams when one of the shapes is much larger than the other, so sometimes we'll scale down regions to make the resulting figure fit the page better. That's okay, because the ratio of the areas won't change. Remember that the purpose of these ratios of shapes is to illustrate the relative area of one shape compared to the area of another.

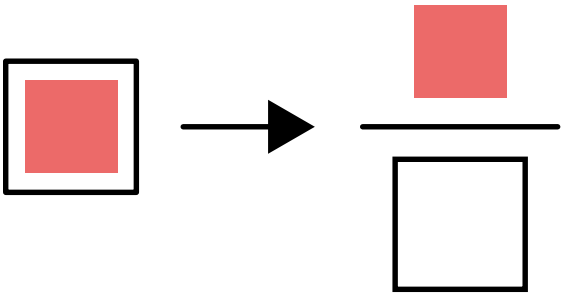


Figure 3-3: The probability of hitting the square in Figure 3-2 is given by the ratio of the area of the square to the area of the wall, here shown as a symbolic fraction.

Simple Probability

When we talk about the probability of something happening, we refer to that something as an *event*. We often refer to events with capital letters, such as A, B, C, and so on. The phrase "the probability of event A happening" simply means the probability that A happens. To save some space, rather than write "the probability of event A happening," or more succinctly "the probability of A," we usually write P(A) (some authors use a lower-case p, writing p(A)).

Let's say A is the event in which we throw a dart and hit the red square from Figure 3-2. We can represent P(A) with a ratio, as we did earlier. Figure 3-4 shows this graphically.

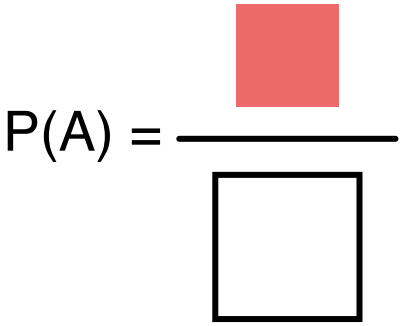


Figure 3-4: We'll say that hitting the square with our dart is event A. The probability of event A occurring is given by the symbolic ratio of areas in Figure 3-3. We write this probability as P(A).

Here, P(A) is the area of the square divided by the area of the wall, so P(A) is ½. This ratio is the probability that, when throwing a dart, we'll hit the square rather than the rest of the wall. We call P(A) a *simple probability*.

Conditional Probability

Let's now talk about probabilities involving two events. Either of these events might happen, or both of them, or neither of them.

For example, we might ask for the probability that a house contains a piano, and the probability that there's a dog inside. There's probably no relationship between these two qualities (or events). We say that two events that are not related to one another in any way are *independent*.

Many types of events are not independent, but have at least some kind of connection. We call these *dependent*. When events are dependent, we might want to find their relationship. That is, we'd like to find the probability of one specific event, when we already know that another specific event has happened (or is happening). For example, suppose we pass a house and hear a dog barking inside. Then we might ask, "What is the probability that there's a dog's chew toy in the house, *given* that we know there's a dog inside?" In other words, we know that one event has happened, and we want to know the probability of the other.

Let's make this a bit more abstract, and discuss two events called A and B. Suppose that we know that B has happened, or equivalently, that B is true. Knowing this, we can ask what's the probability that A is *also* true? We write this probability as P(A|B). The vertical bar represents the word *given*, so we'd say this out loud as "the probability that A is true, given that B is true," or more simply, "the probability of A given B." This is called the *conditional probability* of A given B, since it only applies to the situation, or condition, that B is true. We can also talk about P(B|A), which is the probability that B is true, given that A is true.

We can illustrate this with our picture diagrams. The left diagram in Figure 3-5 shows our wall, with two overlapping blobs labeled A and B. P(A|B) is the probability that our dart landed in blob A, given that we already know it landed in blob B. In the symbolic ratio on the right of Figure 3-5, the top shape is the region that is common to both A and B. That is, it's their overlap, or the area where the dart can land in A, given that we know it landed in B.

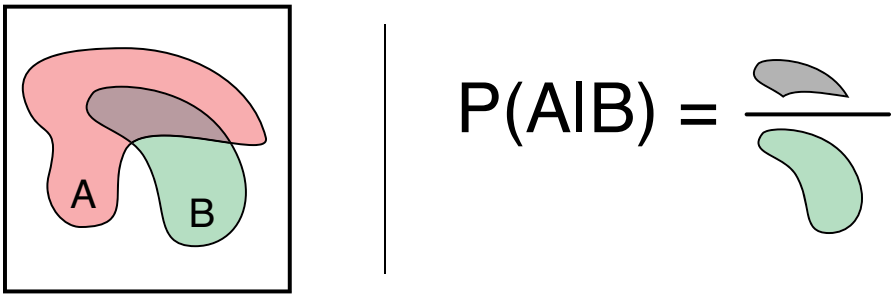


Figure 3-5: Left: The two blobs painted on the wall. Right: The probability of being in A given that the dart is already in B is the ratio of the area of A overlapping B, divided by the area of B.

P(A|B) is a positive number that we can estimate by using our darts. We can estimate P(A|B) by counting all the darts that land in the overlap of A and B, and dividing that number by how many land in any part of B.

Let's see this in action. In Figure 3-6 we've thrown a number of darts at the wall containing the blobs of Figure 3-5. We placed the points to get good

coverage over the whole area, with no two points too close to one another. Dart tips are too hard to see, so we'll show the location of each dart's impact by a black circle, where the center of the circle shows where the dart struck.

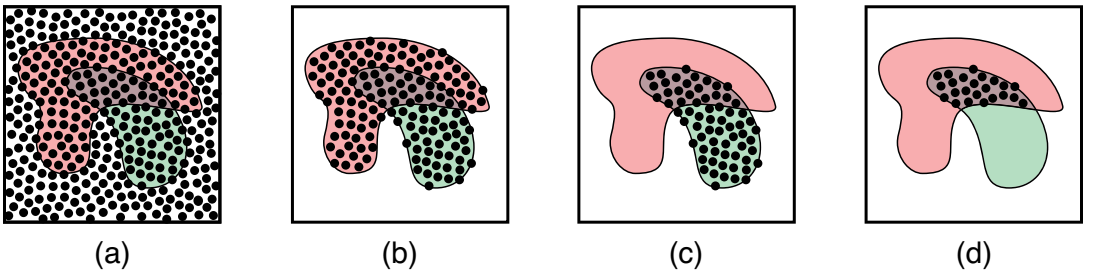


Figure 3-6: Throwing darts at the wall to find P(A|B). (a) Darts striking the wall. (b) All the darts in either A or B. (c) The darts only in B. (d) The darts that are in the overlap of A and B.

In Figure 3-6(a) we show all the darts. In Figure 3-6(b) we've isolated just the darts that landed in either A or B (remember it's only the center of each black circle that counts). In Figure 3-6(c) we see the 66 darts that have landed in region B, and in Figure 3-6(d) we see the 23 darts that are in both A and B. The ratio of 23/66 (about 0.35) estimates the probability that a dart landing in B will also land in A. So P(A|B) is about 0.35. That is, if a dart lands in B, then about 35% of the time, it will also be in A.

Note that this process doesn't depend on the absolute area of the colored blobs, such as a number in square inches. It's just the relative size of one area with respect to another, which is the only measure we really care about (if the wall doubled in size and so did the colored regions, the probability of landing in each one wouldn't change).

The bigger the overlap of A and B, the more likely the dart is to land in both. If A surrounds B, as in Figure 3-7, then we *must* have landed in A given that we landed in B. In this case, the overlap of A and B (shown in gray) is the region of B itself. Thus the ratio of the overlap's area to B's area is 100%, or P(A|B) = 1.

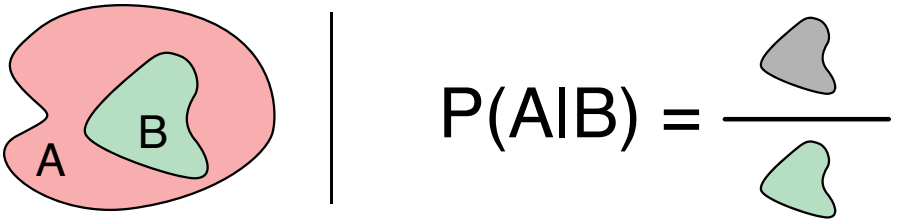


Figure 3-7: Left: Two new blobs on the wall. Right: The probability of landing in A given that we're in B is 1, because A encloses B, and thus their overlap is the same as B.

On the other hand, if A and B don't overlap at all, as in Figure 3-8, then the probability of the dart being in A given that it landed in B is 0%, or P(A|B) = 0.

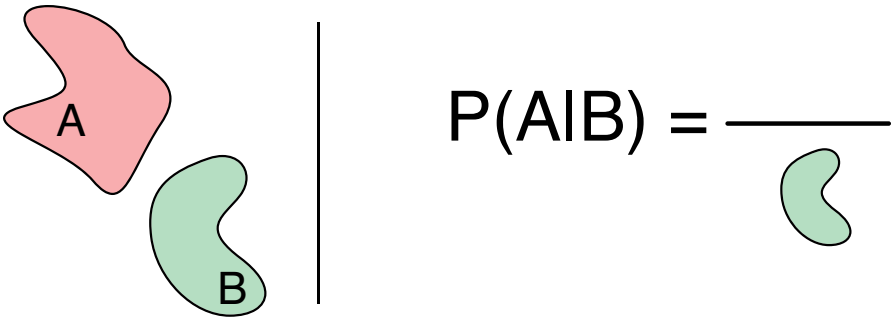


Figure 3-8: Left: Another two new blobs on the wall. Right: The probability of landing in A given that we're in B is 0(or, equivalently, 0%), because there's no overlap between A and B.

The symbolic ratio in Figure 3-8 shows that the area of overlap is 0, and 0 divided by anything is still 0.

For fun, let's flip this around the other way, and ask about P(B|A), or the probability that we're in blob B *given that we're in blob A*. Using the same blobs as in Figure 3-5, the result is shown in Figure 3-9.

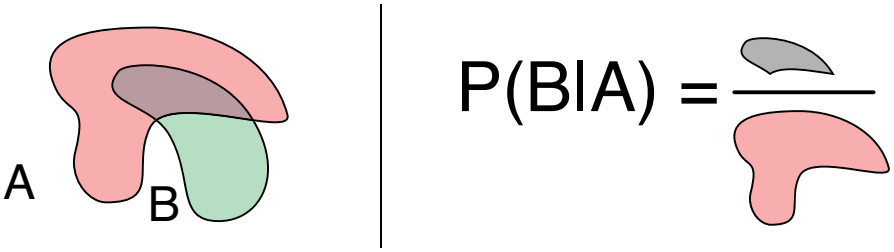


Figure 3-9: The conditional probability P(B|A) is the probability we landed in B, given that we landed in A.

The logic is the same as before. The area of overlap divided by the area of A tells us how much of B appears in A. The more they overlap, the more likely it is that a dart landing in A will also land in B. Let's assign a number to P(B|A). Referring back to Figure 3-6, we see that 104 darts land in A, and 23 in B, so P(B|A) is 23/104 or about 0.22.

Note that the order is important. We can see from Figure 3-5 and Figure 3-9 that P(A|B) does not have the same value as P(B|A). Given the sizes of A, B, and their overlap, the chance of landing in A given that we landed in B is greater than the chance of landing in B given that we landed in A. That is, P(A|B) is about 0.35, but P(B|A) is about 0.22.

Joint Probability

In the last section, we saw a way to express the probability of one event happening, given that another event had already occurred. It would also be helpful to know the probability of both things happening at once. In the

language of our blobs, what's the chance that a dart thrown at the wall will land in *both* blob A and blob B? We write the probability of both A and B happening as P(A,B), where we think of the comma as meaning the word *and*. Thus we read P(A,B) out loud as "the probability of A and B."

We call P(A,B) the *joint probability* of A and B. Using our blobs, we can find this joint probability P(A,B) by comparing the area of the overlap of blobs A and B to the area of the wall. After all, we're asking for the chance that our dart lands in both A and B, meaning inside their overlap, compared to the chance it could land anywhere on the wall. Figure 3-10 shows this idea.

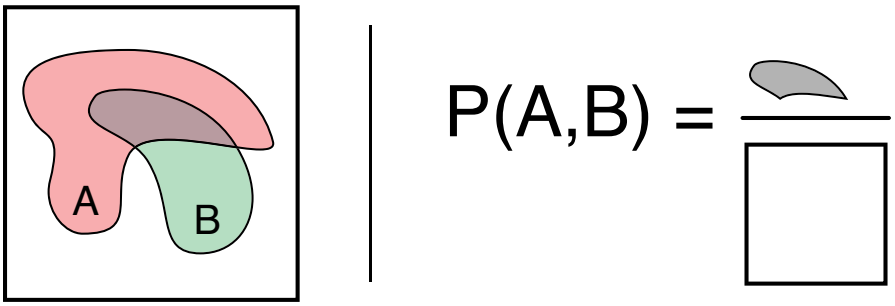


Figure 3-10: The probability that both A and B will occur is called their joint probability, written P(A,B).

There's another way to look at the joint probability that's a little more subtle, but powerful. It's so useful that it will be the heart of Chapter 4. This alternative view of the joint probability combines a simple probability with a conditional probability.

Suppose we know the simple probability of hitting B, or P(B). And suppose we also know the conditional probability P(A|B), or the probability of hitting A, knowing that we hit B. We can combine these into a chain of reasoning: given the probability of hitting B, we'll combine that with the probability of hitting A given that we hit B, to get the probability of hitting both A and B at the same time.

Let's see the chain of reasoning with an example. Suppose that blob B covers half of the wall, so P(B) = ½. Further, suppose that blob A covers a third of blob B, so P(A|B) = ⅓. Then half of our darts thrown at the wall will land in B, and a third of those will fall in A. Since half of the darts fall in B, and a third of those will also fall in A, the total number that land in both B and in A is ½ × ⅓, or ⅙.

This example shows us the general rule: to find P(A,B) we multiply P(A|B) and P(B). This is really quite remarkable: we just found the joint probability P(A,B) using only the conditional probability P(A|B) and the simple probability P(B)! We write this as P(A,B) = P(A|B) × P(B). In practice, we usually leave off the explicit multiplication sign, writing just P(A,B) = P(A|B) P(B).

Figure 3-11 shows what we just did using our little area diagrams.

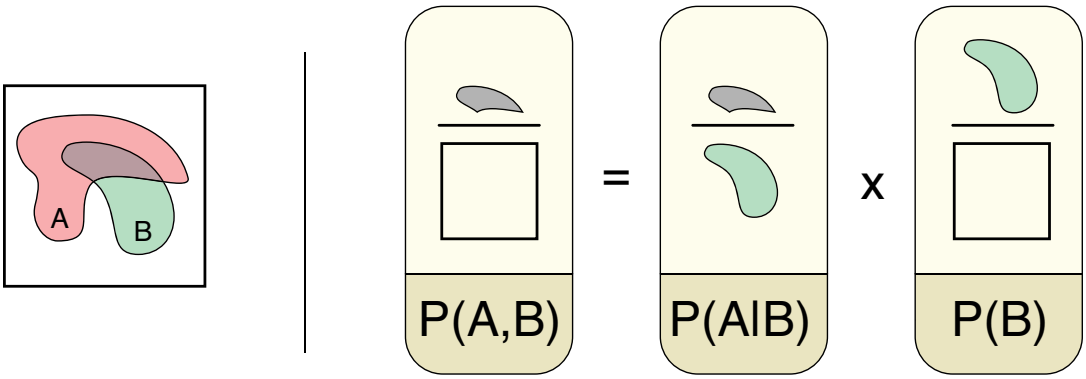


Figure 3-11: Another way to think about the joint probability P(A,B)

Consider the right side of Figure 3-11 and think of the little symbolic ratios as actual fractions. Then the green blobs of area B cancel each other, and we're left with the gray area over the square, showing that the left and right sides of our little equation are, indeed, equal.

We can do this the other way around, too, using event A rather than B. We start with P(B|A) to learn the probability of landing in B given that we landed in A, and then we multiply that by the probability of landing in A, or P(A). The result is P(A,B) = P(B|A) P(A). Graphically, this follows the same pattern as Figure 3-11, only now it's the A blobs that cancel each other.

In symbols, P(B,A) = P(A,B), since both refer to the probability of landing in A *and* B simultaneously. Unlike conditional probability, in joint probability, the order of naming A and B doesn't matter.

These ideas can be a little challenging to get used to, but mastering them will pay off in Chapter 4. It may help to make up a few little scenarios and play with them, imagining different blobs and how they overlap, or even thinking of A and B as actual situations. For instance, imagine an ice cream shop where people can buy different flavors of ice cream, in either a waffle cone or cup. We might say V is true if someone orders vanilla ice cream, and W is true if a person orders their ice cream in a waffle cone. Then P(V) is how likely a random customer will order vanilla, and P(W) is how likely an independently chosen customer will ask for a waffle cone. P(V|W) tells us how likely it is that someone who got a waffle cone ordered vanilla, and P(W|V) tells us how likely it is that someone who ordered vanilla got it in a waffle cone. And P(V,W) tells us how likely it is that a randomly chosen customer got vanilla ice cream in a waffle cone.

Marginal Probability

Another term used for simple probability is *marginal probability*, and understanding where this term comes from will help us understand how we can calculate simple probabilities for multiple events.

Let's start with the word *marginal*, which can seem pretty strange in this context. After all, what does a margin have to do with probability? The legend behind the word *marginal* is that it comes from books that contained tables of precomputed probabilities. The idea is that we (or the printer) would sum up the totals in each row of a table of probabilities, and write those totals in the margin of the page (Glen 2014).

Let's illustrate this idea by returning to our ice cream shop. In Figure 3-12 we show some recent purchases made by our customers. Our shop is brand new and serves only vanilla and chocolate, in either a waffle cone or cup. Based on the purchases of the 150 people who came in yesterday, we can ask the probability of someone buying a cup versus a waffle cone, or vanilla versus chocolate. We find those values by adding up the numbers in each row or column (giving us the number in the margin) and dividing by the total number of customers.

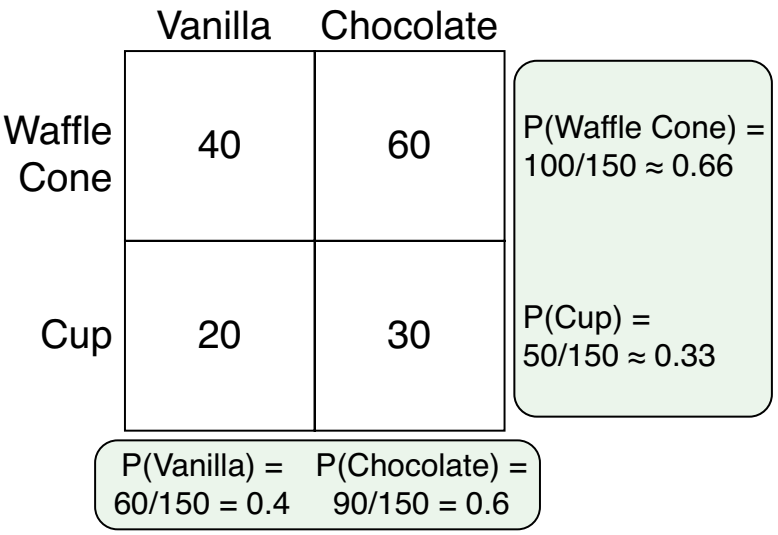


Figure 3-12: Finding marginal probabilities for 150 recent visitors at an ice cream shop. The values in the green boxes (showing the margins of the grid) are the marginal probabilities.

Note that the probabilities of someone buying a cup *or* waffle cone add up to 1, since every customer buys one or the other. Similarly, everyone buys either vanilla or chocolate, so those probabilities also add up to 1. In general, all the probabilities for the various outcomes of any event will always add up to 1, because it's 100% sure that *one* of those choices will occur.

Measuring Correctness

Let's shift gears and look at another important probability concept: given an imperfect algorithm, how likely is it to produce the correct answer? This

is a key question in machine learning, because we will almost always work with systems that fall short of being perfectly accurate. So it's important to understand what kinds of errors they make.

Let's consider a simple classifier with just two classes. We can ask it for the probability that a piece of data is in some specific class (the two classes are then *in category* and *out of category*). For instance, we might ask for the probability that a photograph features a dog, or the probability that a hurricane will hit land, or how likely it is that our high-tech enclosures are strong enough to hold our genetically engineered super-dinosaurs (spoiler: not very).

Naturally, we'd like our classifier to make accurate decisions. The trick is to define what we mean by *accurate*. Just counting the number of incorrect results is the easiest way to measure something that we might call accuracy, but it's not very illuminating. The reason is that there is more than one way to be wrong. If we want to use our mistakes to improve our performance, then we need to identify the different ways our predictions can be wrong and consider how much trouble each kind of error causes us. This kind of analysis applies far beyond just machine learning. The following ideas can help diagnose and solve all kinds of problems where we're making decisions on the basis of labels we've assigned.

Before we dig in, we'll note that some of the terms we'll be using here, such as *precision*, *recall*, and *accuracy*, are used casually in popular and informal writing. But in technical discussions (like in this book), these words have precise definitions and mean different things. Unfortunately, not all authors use the same definitions for these terms, which can cause all kinds of confusion. In this book, we'll stick to the way they're usually used when discussing probability and machine learning, and we'll define them carefully when we come to them later in this chapter. But be aware that these terms appear in lots of places with different meanings or are just left as vague concepts. It's unfortunate when words get overloaded this way, but it happens.

Classifying Samples

Let's narrow our language to the task at hand. We want to know if a given piece of data, or sample, is, or isn't, in a given category. For now, think of this in yes/no question form: Is this sample in the category? There are no "maybe" answers allowed.

If the answer is "yes," we call the sample *positive*. If the answer is "no," we call the sample *negative*. We'll discuss accuracy by comparing the answers we get from our classifier against the real, or correct, labels that we've assigned beforehand. The choice of positive or negative that we've manually assigned to the sample is called its *ground truth* or *actual value*. We'll say that the value that comes back from our classifier is the *predicted value*. In a perfect world, the predicted value would always match the ground truth. In the real world, there are often errors, and our goal here is to characterize those errors.

We'll illustrate our discussion with two-dimensional (2D) data. That is, every sample, or data point, has two values. These might be a person's height and weight, or a weather measurement of humidity and wind speed, or a musical note's frequency and volume. Then we can plot each piece of data on a 2D grid with the X axis corresponding to one measurement and the Y axis to the other.

Our samples will each belong to one of two classes. Let's call them *positive* and *negative*. To identify a sample's correct classification, also called its *ground truth*, we'll use color and shape cues, as in Figure 3-13.

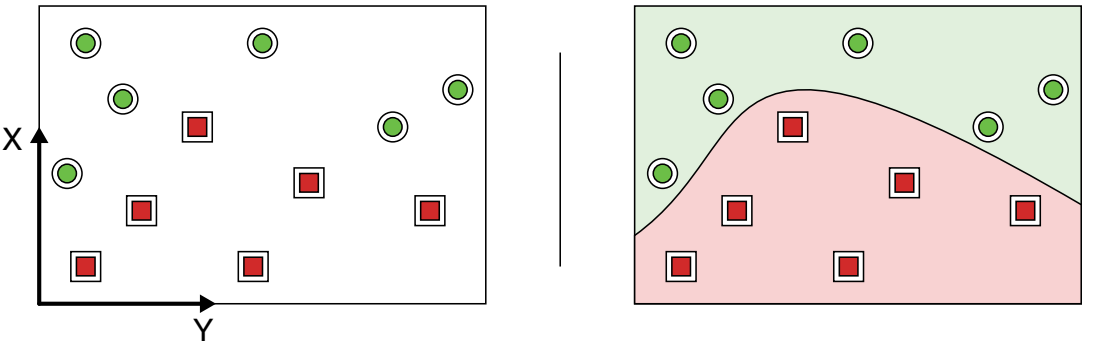


Figure 3-13: Two-dimensional data belonging to two different classes

We'll show the results of our predictions by drawing a *boundary*, or curve, through the collection of points. The boundary may be smooth or twisty. We can think of it as a kind of summary of the classifier's decision-making process. All points in one side of the curve will be predicted to be of one class, while all those on the other side will be predicted to be in the other class. In the right diagram of Figure 3-13, the classifier has done a perfect job of predicting the ground truth of each sample. That's a rare thing.

We sometimes say that the boundary has a positive side and a negative side. This matches up to our class if we think of the classifier as answering the question, "Does this sample belong to the category?" If the answer is positive, then the prediction is "yes," otherwise the prediction is "no." It's often helpful to color in the regions on either side of the boundary, as we've done in Figure 3-13, to make it easy to see which side holds the predictions of positive, and which holds the predictions of negative.

For our dataset, we'll use a set of 20 samples, shown in Figure 3-14. The samples with a ground truth (or manual label) of positive are shown as green circles, while those with a ground truth (or manual label) of negative are drawn as red squares. So the color and shape of each sample corresponds to its ground truth, and not the value assigned by the classifier.

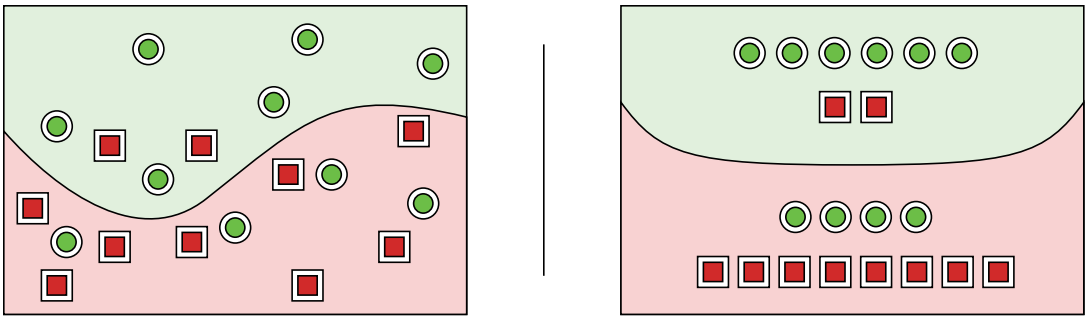


Figure 3-14: Left: The classifier's curve does an okay job of separating the classes, but it makes some mistakes. Right: A schematic version of the same diagram. The curved boundary reminds us that the actual boundary is rarely a straight line.

The job of the classifier is to try to find a boundary so that all the positive samples land on one side, and all the negative samples land on the other. To see how well the classifier's prediction of each sample matches its ground truth, we can just look to see if that sample ended up on the correct side of the classifier's boundary curve. That curve splits the space into two regions. We've used light green to show the positive region, and light red for negative, so every point in the light-green region is predicted, or classified, as positive, and every point in the light-red region is classified as negative.

In a perfect world, all the green circles (the ones with a positive ground truth) would be on the green side of the boundary curve (showing that the classifier predicted them as positive), and all the red squares would be on the red side. But as we can see in the figure, this classifier has made some mistakes. On the left of Figure 3-14 we plotted each piece of data using its two values, along with the boundary curve (and regions) that characterize the classifier's decisions. But we don't really care in this discussion about the specific locations of the points or the shape of the curve. Our interest is in how many points were correctly and incorrectly classified and thus landed on the right and wrong side of the boundary. So in the figure on the right, we've cleaned up the geometry to make it easier to count the samples at a glance.

This diagram represents what typically happens when we run a classifier on a real data set. Some data is classified correctly, and some isn't. If our classifier isn't performing well enough for us, we'll need to take some sort of action—perhaps by modifying the classifier or even throwing it out and making a new one—so it's important to be able to usefully characterize how well it's doing.

Let's find some ways to do that. We'd like to characterize the errors in Figure 3-14 in a way that tells us something about the nature of the classifier's performance, or how well its predictions matched our given labels. It would be nice to know something more than just "right" and "wrong"—we'd like to know the nature of the mistakes, because some mistakes might matter to us a lot, while others might not matter much at all.

The Confusion Matrix

To characterize the classifier's answers, we can make a little table that has two columns, one for each predicted class, and two rows, one for each actual, or ground truth, category. That gives us a 2 by 2 grid, referred to as a *confusion matrix*. The name refers to how the grid, or matrix, shows us the ways in which our classifier was mistaken, or confused, about its predictions. The classifier's output is repeated in Figure 3-15, along with its confusion matrix.

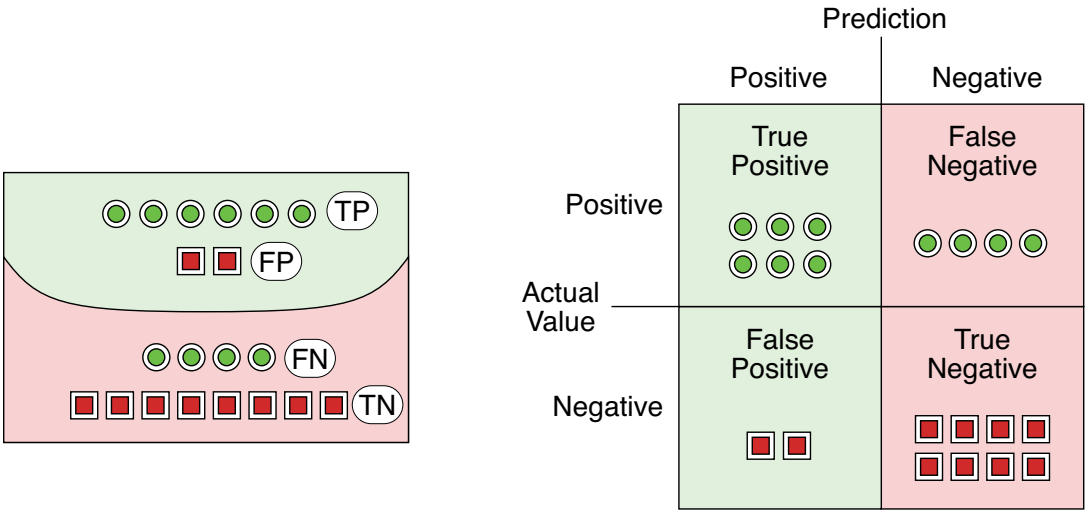


Figure 3-15: We can summarize what went where in Figure 3-14 (repeated here on the left, with labels) into a confusion matrix, which tells us how many samples landed in each of the four classes.

As Figure 3-15 shows, each of the four cells in the table has a conventional name, which describes a specific combination of the predicted and actual values. The six positive green circles were correctly predicted as positive, so they go into the *true positive* category. In other words, they were predicted to be positive, and they actually were positive, so the prediction of positive was correct, or true. The four green circles that were incorrectly classified as negative go into the *false negative* category, because they were incorrectly, or falsely, labeled as negative. The eight red negative squares were correctly classified as negative, so they all go into the *true negative* category. Finally, the two red squares that were incorrectly predicted to be positive go into *false positive*, because they were incorrectly, or falsely, predicted to be positive.

We can write this more concisely using two-letter abbreviations for the four classes and a number describing how many samples fell into each category. Figure 3-16 shows the form that the confusion matrix is usually shown in.

Unfortunately, there is no universal agreement on where the various labels go in confusion matrix diagrams. Some authors put predictions on the left and actual values on top, and some place positive and negative

in the opposite locations than shown here. When we encounter a confusion matrix, it's important to look at the labels and make sure we know what each box represents.

		Prediction: Positive	Prediction: Negative
Actual Value	Positive	TP 6	FN 4
	Negative	FP 2	TN 8

Figure 3-16: The confusion matrix of Figure 3-15 in its conventional form

Characterizing Incorrect Predictions

We mentioned earlier that some errors might matter more to us than others. Let's see why that might be.

Suppose that we work for a company that makes toy figurines in the likeness of popular TV characters. Our toys are a hit right now, so our production line is running at full capacity. Our job is to take the manufactured figurines, box them, and ship them off to retail stores.

Suddenly, one day we're told that our company has lost the rights to sell a particular character named Glasses McGlassface. If we accidentally ship any of those figurines, we'll get sued, so it's important to make sure that none of them leave our factory. Unfortunately, the machines are still cranking them out, and if we stop the production line to update the machines, we'll fall way behind on our orders. We decide the better approach is to keep making the forbidden figurines, but spot them after they've been made and throw them into a bin for recycling. So our goal is to identify each Glasses McGlassface and throw it in the bin, making sure none of them get out the door.

Figure 3-17 shows the situation.

We need to work fast, so we might make some mistakes. In Figure 3-17 we see one figurine that we incorrectly recycled. That is, when answering the question, "Is this Glasses McGlassface?" we incorrectly said, "yes." Using our language from the last section, this doll is a false positive. How big a problem is that?

In this case, it's not a big deal (as long as we don't do it too often). Our goal is to make sure that every Glasses McGlassface is correctly identified and removed. Missing even one would cost us a lot. But a false positive costs us only a little, since we'll melt down the plastic and reuse it. So in this situation, false positives, while not desirable, are tolerable.

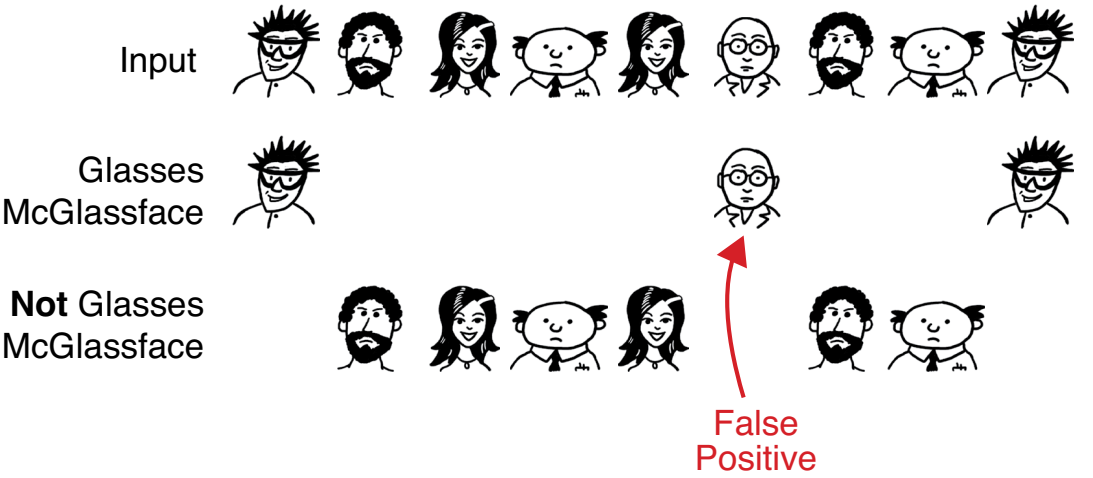


Figure 3-17: Glasses McGlassface is the first character on the top row. We want to remove any doll that could be that character. Our selections are in the middle row.

Suppose we've later noticed that some figurines are not having their eyes painted on properly. Giving a child a toy without eyes could be traumatic, so we definitely want to catch them all. As before, we'll look at every toy, this time asking, "Are the eyes present?" If not, we throw the figurine into a bin for recycling. Figure 3-18 shows the idea.

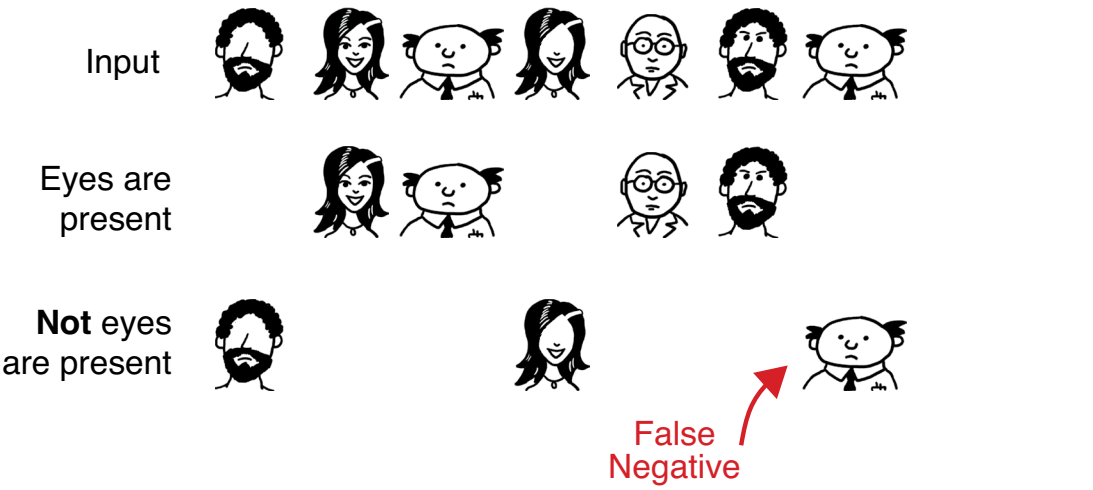


Figure 3-18: A new group of toys. Now we're looking for any with mispainted eyes. Our selections are in the bottom row.

Here we have a false negative: the doll has its eyes painted in, but we said it didn't. In this situation, a few false negatives aren't so bad. As long as we're sure to remove every doll that is missing its eyes, it's okay if we remove a few with their eyes present.

To sum up, true positives and true negatives are the easy cases to understand. How we should respond to false positives and false negatives is dependent on our situation and our goals. It's important to know what our question is, and what our policy is, so we can work out how we want to respond to these different types of errors.

Measuring Correct and Incorrect

Let's return to our overview of true and false positives and negatives, as summarized in a confusion matrix. Looking at a confusion matrix can be, well, confusing, so people have created a variety of terms to help us talk about how well our classifier is performing.

We'll illustrate these terms using a medical diagnosis scenario, where *positive* means someone has a particular condition, and *negative* means they're healthy. Suppose that we're public health workers who have come to a town that's experiencing an outbreak of a terrible but completely imaginary disease called *morbus pollicus (MP)*. Anyone who has MP needs to have their thumbs surgically removed right away, or the disease will kill them within hours. It's therefore critical that we correctly diagnose everyone with MP. But we definitely don't want to make any incorrect diagnoses that lead to removing anyone's thumbs if their life is not in danger—thumbs are important!

Let's imagine that we have a laboratory test for detecting MP. The lab test is flawless, so it always gives us the correct answer: a positive diagnosis means the person has MP, and a negative diagnosis means they do not. Using this test, we've checked every person in town, and we now know whether or not they have MP. But our lab test is slow, and expensive. We're worried about future outbreaks, so based on what we've just learned, we develop a fast, cheap, and portable field test that will predict immediately if someone does or does not have MP.

Unfortunately, our field test is not perfectly reliable, and sometimes makes incorrect diagnoses. Although we know our field test is flawed, when we're in the middle of an outbreak it may be the only tool we have. So we want to characterize how often the field test is correct and how often it's wrong, and when it's wrong, we want to characterize the ways it's wrong.

To work this out, we need data. We've just heard of another town where a few people have reported MP. We'll check every person in town with both tests: our perfect (but slow and expensive) lab test, and our imperfect (but quick and cheap) field test. In other words, the lab test gives us the ground truth for each person, and the field test gives us a prediction. The lab test is too expensive to always run both tests on every person, but we can afford it this once.

By comparing the field test predictions with the lab test label, we'll know all four quadrants of the confusion matrix for our field test:

True Positive: the person has MP, and our field test correctly says that they have it.

True Negative: the person does *not* have MP, and our field test agrees.

False Positive: the person does *not* have MP, but our field test says that they do.

False Negative: the person has MP, but our field test says they don't.

Both true positive and true negative are correct answers, while false negative and false positive are incorrect. A false positive means we'd operate without cause, and a false negative would leave someone at risk of dying.

If we build a confusion matrix for our field test by attaching numbers to each of the four cells, we can use those values to determine how well our field test is performing. We will be able to characterize its performance with a few well-known statistics. The *accuracy* will tell us how often the field test gives us a correct answer, the *precision* will tell us something about false positives, and the *recall* will tell us about false negatives. These values are the standard way that people talk about the quality of a test like this, so let's look at those values now. Then we'll come back to our confusion matrix for the field test, compute these values, and see how they help us interpret the test's predictions.

Accuracy

Each of the terms we'll discuss in this section is built from the four values in the confusion matrix. To make things a bit easier to discuss, we'll use the common abbreviations: TP for true positive, FP for false positive, TN for true negative, and FN for false negative.

Our first term to characterize the quality of a classifier is *accuracy*. The accuracy of the predictions made for any collection of samples is a number from 0 to 1. It's a measure of the percentage of samples that were assigned to the correct category. So it's just the sum of the two "correct" values, TP and TN, divided by the total number of samples measured. Figure 3-19 shows the idea graphically. In this figure, as in the ones to come, the samples we're counting for any given computation will be shown, and the samples that don't contribute to that value will be omitted.

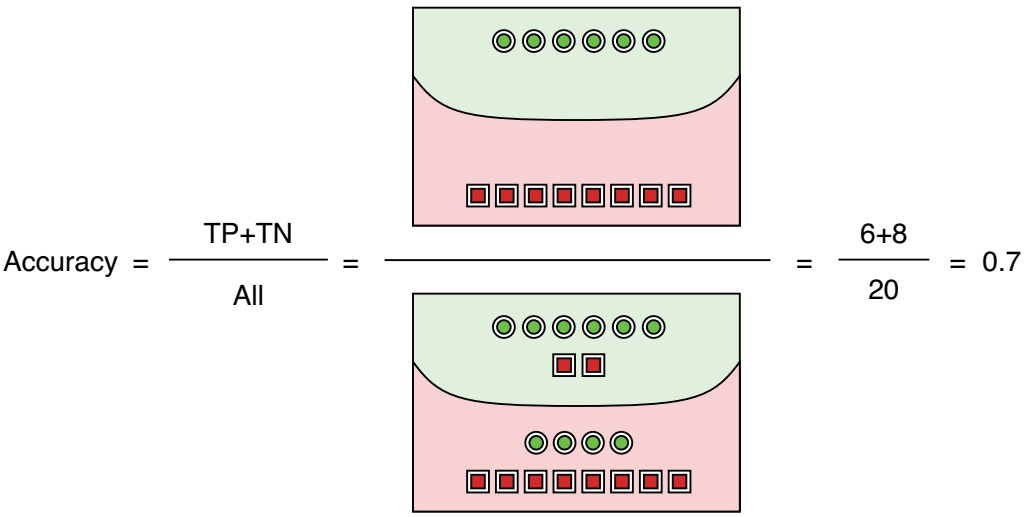


Figure 3-19: Accuracy is a number from 0 to 1 that tells us how often our prediction is correct.

We want the accuracy to be 1.0, but usually it will be less than that. In Figure 3-19, we have an accuracy of 0.7, or 70%, which isn't great. The accuracy doesn't tell us in what way the predictions are wrong, but it does give us a broad feeling for how much of the time we get the right result. Accuracy is a rough measurement.

Let's now look at two other measures that provide more specific characterizations of our predictions.

Precision

Precision (also called *positive predictive value [PPV]*) tells us the percentage of our samples that were properly labeled positive, relative to all the samples we labeled as positive. Numerically, it's the value of TP relative to TP + FP. In other words, precision tells us what percentage of our positive predictions were correct.

If the precision is 1.0, then every sample we labeled as positive was correctly predicted as positive. As the percentage falls, it carries with it our confidence in these predictions. For example, if the precision is 0.8, then we can only be 80% sure that any given sample that's labeled positive has the correct label. Figure 3-20 shows the idea visually.

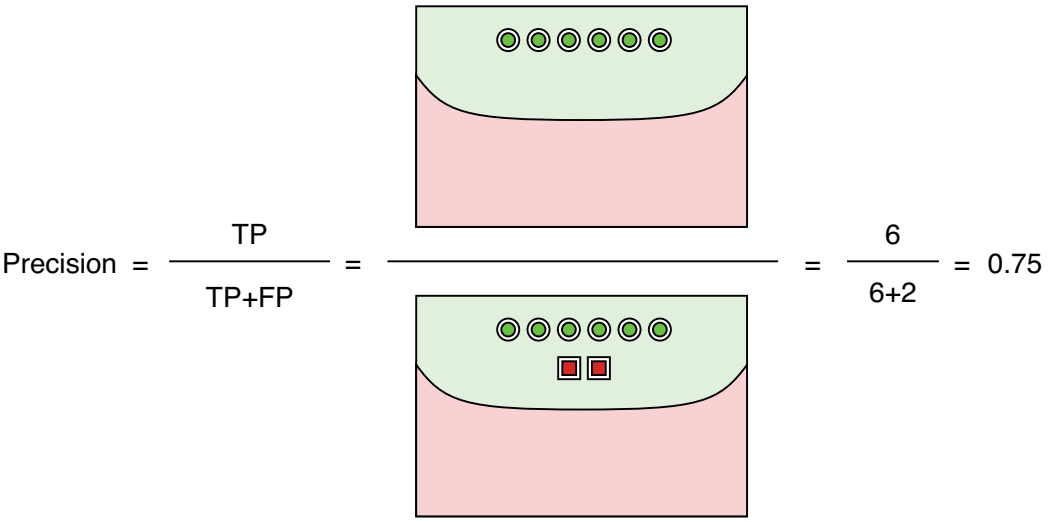


Figure 3-20: The value of precision is the total number of positive samples that really are positive, divided by the total number of samples that we labeled as positive.

When the precision is less than 1.0, it means we labeled some samples as positive when we shouldn't have. In our healthcare example from before with our imaginary disease, a precision value of less than 1.0 means that we'd perform some unnecessary operations. An important quality of precision is that it doesn't tell us if we actually found all the positive objects: that is, all the people who had MP. Precision ignores all samples except those labeled as positive.

Recall

Our third measure is *recall* (also called *sensitivity*, *hit rate*, or *true positive rate*). This tells us the percentage of the samples we correctly predicted to be positive, relative to all the samples that really were positive. That is, it tells us the percentage of positive samples that we correctly predicted.

When recall is 1.0, then we correctly predicted every positive event. The more that recall drops below that number, the more positive events we missed. Figure 3-21 shows this idea visually.

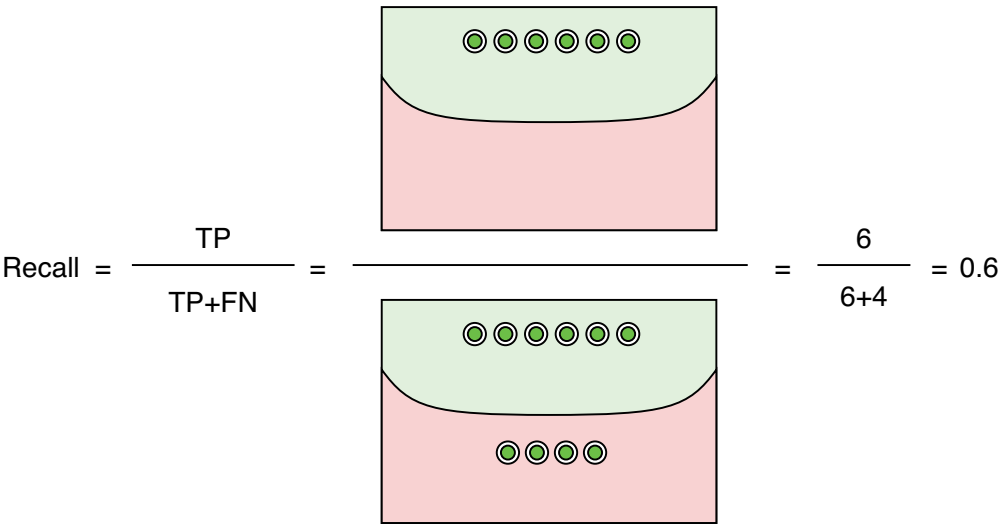


Figure 3-21: The value of recall is the total number of correctly-labeled positive samples, divided by the total number of samples that should have been labeled as positive.

When recall is less than 1.0, it means that we missed some positive answers. In our healthcare example, it means we would misdiagnose some people with MP as not having the disease. The result is that we wouldn't operate on those people, even though they're infected and in danger.

Precision-Recall Tradeoff

When we're categorizing data into two classes, and we can't eliminate false positives and false negatives, there's a tradeoff between precision and recall: as one goes up, the other goes down. That's because as we reduce the number of false positives (and therefore increase precision), we necessarily also increase the number of false negatives (and therefore reduce recall). Let's see how this comes about.

Figure 3-22 shows 20 pieces of data. They start out as negatives (red squares) at the far left, and gradually become positives (green circles) as we move right. We'll draw a boundary line vertically somewhere, predicting everything to its left as negative, and everything to its right as positive.

We want all the red squares to be predicted as negative, and all the green circles to be positive. Because they're mixed up, there's no boundary that separates the two groups perfectly.

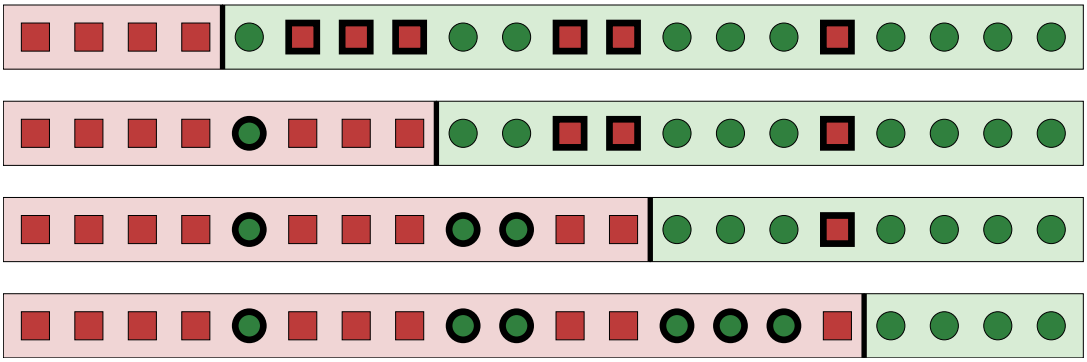


Figure 3-22: As we move the boundary line to the right, from top to bottom, we decrease the number of false positives (red squares with a heavy border), but increase the number of false negatives (green circles with a heavy border).

In the top row of Figure 3-22, the boundary is near the left end. All the green circles are correctly marked positive, but many of the red squares are false positives (shown with a thick outline). As we move the boundary to the right in lower rows, we reduce the number of false positives, but we increase the number of false negatives, because now we're predicting more green circles to be negative.

Let's increase the dataset size to 5000 elements. The data will be like Figure 3-22, so each entry will be positive with a probability given by its distance from the left end. The leftmost graph of Figure 3-23 shows the number of true positives and true negatives as we move the decision boundary from the far left to the far right. The middle graph shows the number of false positives and false negatives, and the rightmost graph shows the resulting accuracy.

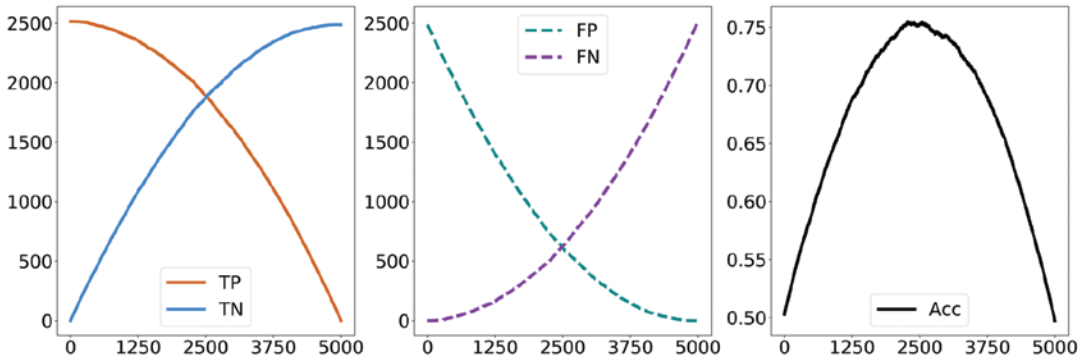


Figure 3-23: Left: The number of true positives and true negatives as we move the boundary. Middle: The number of false positives and false negatives. Right: The accuracy.

To find precision and recall, we'll gather together TP and FP in the left graph of Figure 3-24, and TP and FN in the middle. At the right, we show the result of combining these pairs with TP, following the earlier definitions to compute the precision and recall for each position of the boundary.

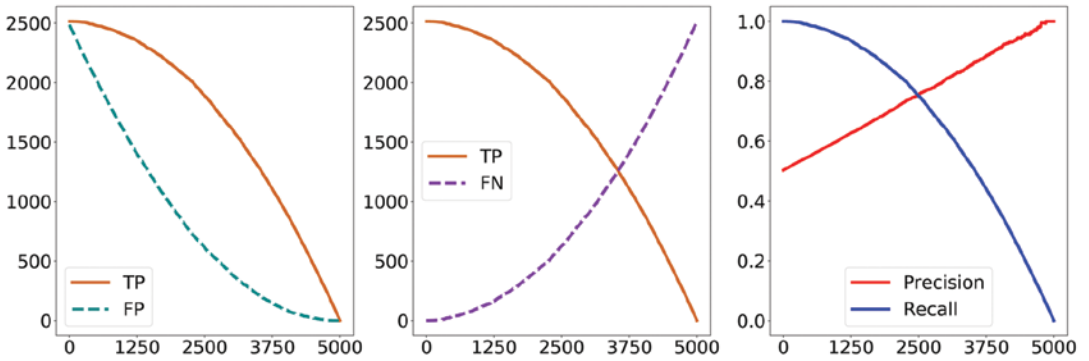


Figure 3-24: TP and FP, TP and FN, and precision and recall as we move the boundary from far left to right

Notice that as we increase precision, we decrease recall, and vice-versa. That's the precision-recall tradeoff.

In this example, the precision follows a straight line, whereas the recall is a curve. To get a feeling for why, consider that the sum TP + FP of the curves shown at the left of Figure 3-24 is a diagonal line from northwest to southeast, whereas the sum TP + FN of the curves in the middle of the figure is a horizontal line. Dividing the TP curve by these two differently oriented lines gives us the different shapes of the precision and recall curves.

For other kinds of data sets, all of these curves would look different, but the precision-recall tradeoff would remain: the better the precision, the worse the recall, and vice-versa.

Misleading Measures

Accuracy is a common measure, but in machine learning precision and recall appear more frequently because they're useful for characterizing the performance of a classifier and comparing it against others. But both precision and recall can be misleading if taken all by themselves, because extreme conditions can give us a great value for either measure, whereas overall performance is lousy.

These misleading results can come from many sources. Perhaps the most common, and difficult to catch, is when we're not careful enough about what we ask the computer to do for us. For example, our organization might want us to produce a classifier that delivers extremely high precision or recall. That may sound desirable, but let's see why it could be a mistake.

To see the problem, consider what might happen if we ask for one of the two extremes of *perfect precision* and *perfect recall*. We'll invent lousy boundary curves to demonstrate the issues, but keep in mind that these can come out naturally from an algorithm tasked to produce perfect precision or recall.

One way to create a boundary curve with perfect precision is to look through all of the samples and find the one we are most certain is really true. Then we draw the curve so that the point we selected is the only positive sample, and everything else is negative. Figure 3-25 shows the idea.

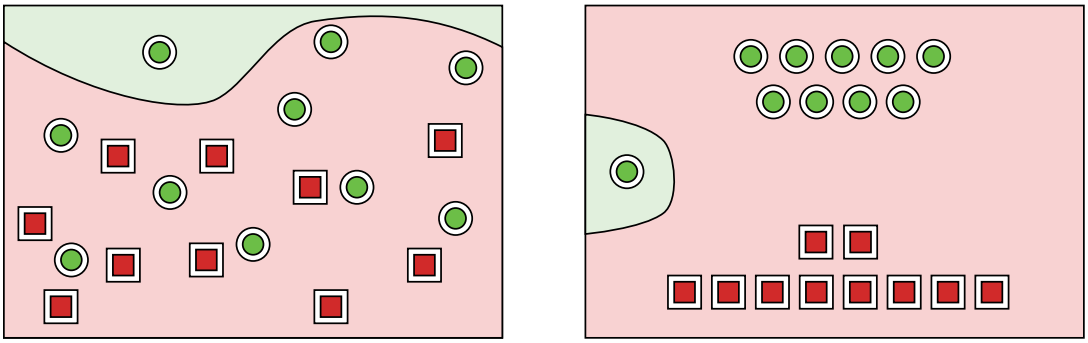


Figure 3-25: Left: This boundary curve gives us a perfect score for precision. Right: A schematic version of the figure on the left.

How does this give us perfect precision? Remember that precision is the number of true positives (here only 1) divided by the total number of points labeled positive (again, just 1). So we get the fraction 1/1, or 1, which is a perfect score. But the accuracy and recall are both pretty awful because we've also created lots of false negatives, as shown in Figure 3-26.

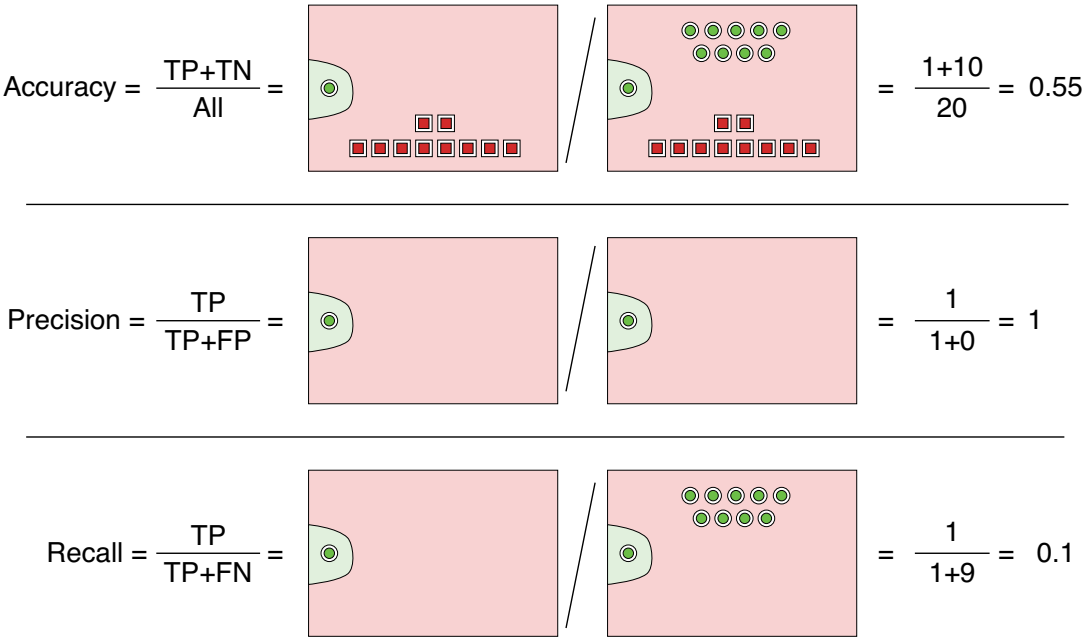


Figure 3-26: These figures all share the same boundary curve, which has labeled exactly one green circle as positive, and all the others as negative.

Let's do a similar trick with recall. To create a boundary curve with perfect recall is even easier. All we have to do is label everything as positive. Figure 3-27 shows the idea.

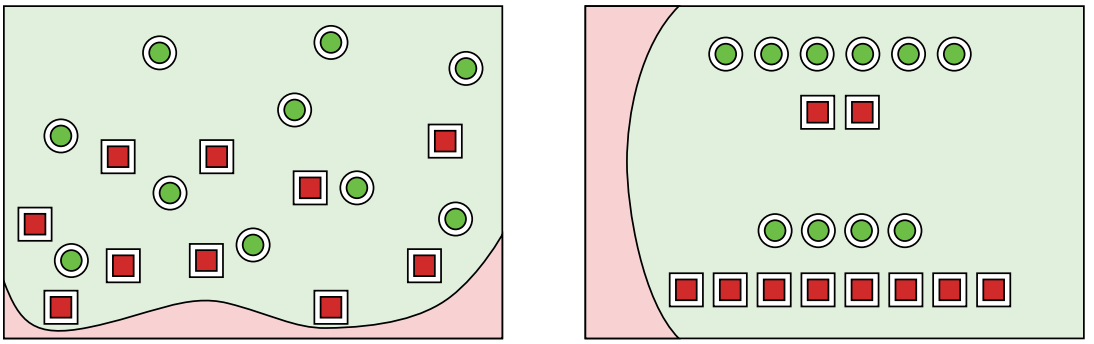


Figure 3-27: Left: This boundary curve gives us a perfect recall score. Right: A schematic version of the figure on the left.

We get perfect recall from this because recall is the number of correctly labeled true points (here, all 10 of them) divided by the total number of true points (again, 10). So 10/10 is 1, or a perfect score for recall. But of course, accuracy and precision are both poor, because every negative sample is now a false positive, as shown in Figure 3-28.

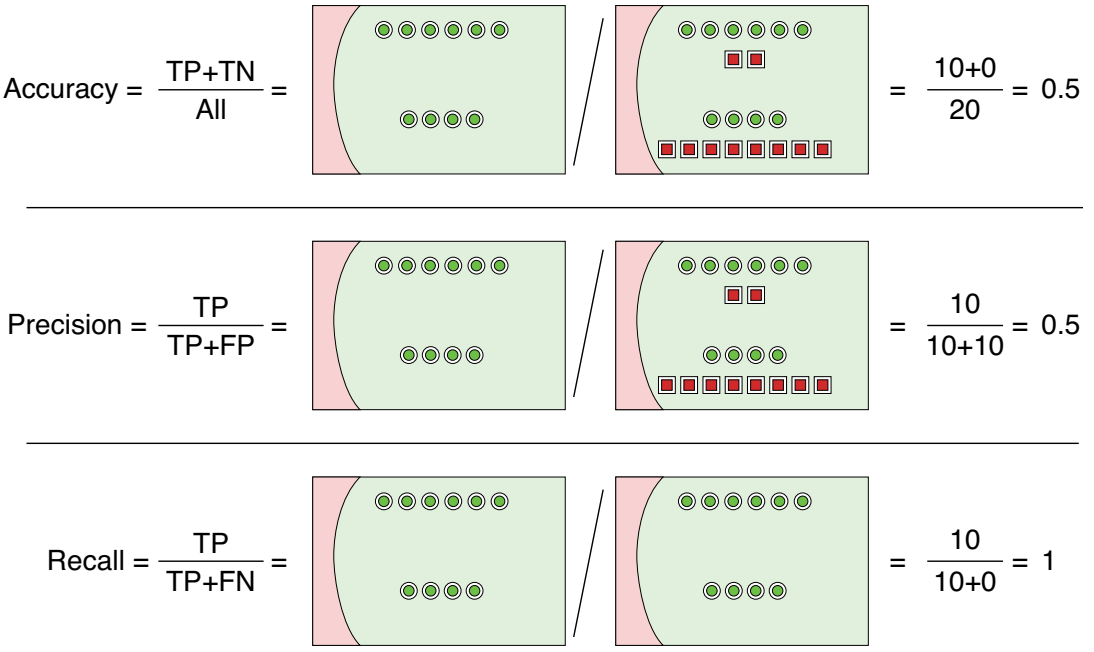


*Figure 3-28: **Perfect recall.** All of these figures share the same boundary curve. With this curve, every point is predicted to be positive. We get a perfect recall, because every positive point is correctly labeled. Unfortunately, accuracy and precision both have very low scores.*

The moral of Figures 3-26 and 3-28 is that asking for perfect precision or perfect recall is unlikely to give us what we really want, which is perfect correctness. We want accuracy, precision, and recall to all be near 1, but if we're not careful, we can get a great score for just one of these measures by picking an extreme solution that performs poorly when we look at the results in just about any other way.

f1 Score

Looking at both precision and recall is informative, but they can be combined with a bit of mathematics into a single measure called the *f1 score*. This is a special type of "average" called a *harmonic mean*. It lets us look at a single number that combines both precision and recall (the formula appears later on in the last lines of Figure 3-30 and Figure 3-32).

Figure 3-29 shows the f1 score visually.

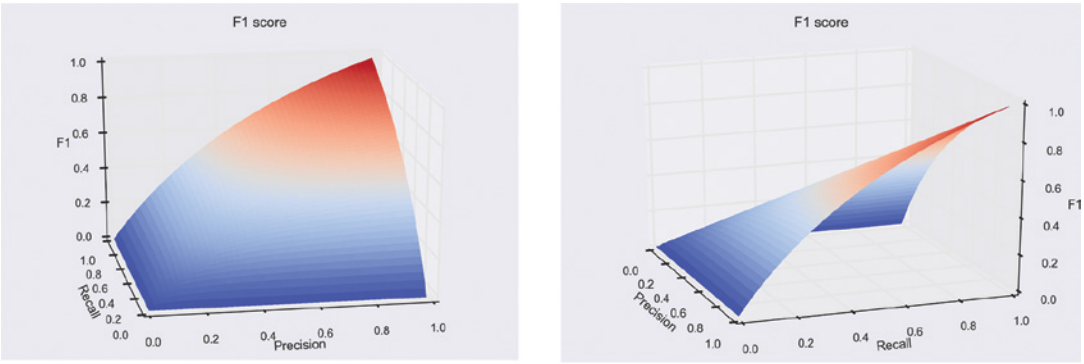


Figure 3-29: The f1 score is 0 when either precision or recall is also 0, and 1 when both are 1. In between it slowly rises as both measures increase.

Generally speaking, the f1 score will be low when either precision or recall is low and will approach 1 when both measures also approach 1.

When a system is working well, sometimes people just cite the f1 score as a shorthand way of showing that both precision and recall are high.

About These Terms

The terms accuracy, precision, and recall may not seem obviously connected to what they measure. Let's make those connections, which can help us remember what these terms mean.

Accuracy tells us what percentage of the samples we predicted correctly. If we predicted every label perfectly, accuracy would be 1. As the percentage of mistakes increases, accuracy drops toward 0. To characterize our mistakes, we want to know our rate of false positives and false negatives. This is what precision and recall are for.

Precision reveals our percentage of false positives, or how many samples we incorrectly predicted to be positive. So this measures the specificity, or precision, of our positive prediction. The larger the value of precision, the

more confidence we have that a positive prediction is accurate. In terms of our medical example, if our test has high precision, then it's likely that a positive diagnosis means that person really has MP. But precision doesn't tell us how many infected people we improperly declared to be disease-free.

Recall reveals our percentage of false negatives. If we think of our system as finding, or recalling, just the positives from a set of data, this tells us how well we've done. The better our recall, the more confidence we have that we correctly retrieved all the positive samples. In our medical example, if our test has high recall, then we can feel confident that we've identified everyone with MP. But recall doesn't tell us how many healthy people we incorrectly identified as having MP.

Other Measures

We've seen the measures of accuracy, recall, precision, and f1. There are lots of other terms that are sometimes used in discussions of probability and machine learning (Wikipedia 2016). We won't encounter most of these terms in this book, but we'll summarize them here to provide a one-stop reference that gathers all the definitions in one place.

Figure 3-30 provides this summary. Don't bother memorizing any unfamiliar terms and their meanings. The purpose of this table is to offer a convenient place to look these things up when needed.

Common Name	Other Names	Abbreviation	Definition	Interpretation
True Positive	Hit	TP	True sample labeled True	Correctly labeled True sample
True Negative	Rejection	TN	False sample labeled False	Correctly labeled False sample
False Positive	False Alarm, Type I Error	FP	False sample labeled True	Incorrectly labeled False sample
False Negative	Miss, Type II Error	FN	True sample labeled False	Incorrectly labeled True sample
Recall	Sensitivity, True Positive Rate	TPR	TP/(TP+FN)	% of True samples correctly labeled
Specificity	True Negative Rate	SPC, TNR	TN/(TN+FP)	% of False samples correctly labeled
Precision	Positive Predictive Value	PPV	TP/(TP+FP)	% of samples labeled True that really are True
Negative Predictive Value		NPV	TN/(TN+FN)	% of samples labeled False that really are False

False Negative Rate		FNR	FN/(TP+FN)=1-TPR	% of True samples incorrectly labeled
False Positive Rate	Fall-out	FPR	FP/(FP+TN)=1-SPC	% of False samples incorrectly labeled
False Discovery Rate		FDR	FP/(TP+FP)=1-PPV	% of samples labeled True that are really False
True Discovery Rate		TDR	FN/(TN+FN)=1-NPV	% of samples labeled False that are really True
Accuracy		ACC	(TP+TN)/(TP+TN+FP+FN)	Percent of samples correctly labeled
f1 score		f1	(2*TP)/((2*TP)+FP+FN)	Approaches 1 as errors decline

Figure 3-30: Common confidence terms derived from the confusion matrix

This table is a lot to take in. We provide an alternative that presents the terms graphically, using our distribution of samples from Figure 3-14, repeated here as Figure 3-31.

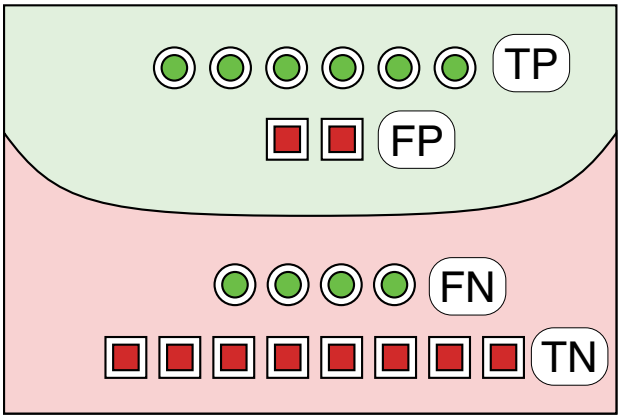


Figure 3-31: The data from Figure 3-14, with the labels from Figure 3-15

Reading from top to bottom, we have six positive points correctly labeled (TP=6), two negative points incorrectly labeled (FP=2), four positive points incorrectly labeled (FN=4), and eight negative points correctly labeled (TN=8).

With these points, we can illustrate the measures of Figure 3-30 by combining these four numbers, or their pictures, in different ways. Figure 3-32 shows how we'd compute the measures using just the relevant pieces of the data.

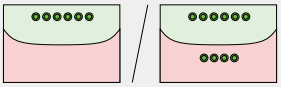
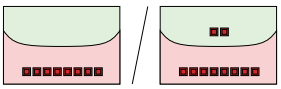
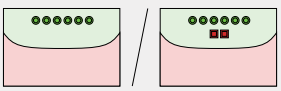
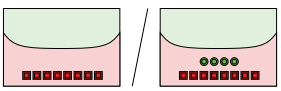
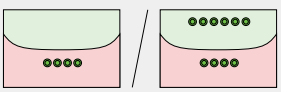
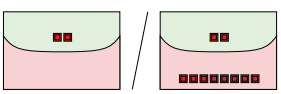
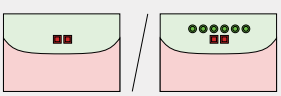
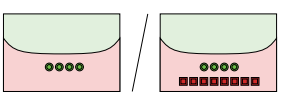
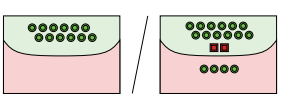
Recall	TPR	$\frac{TP}{TP+FN}$	 $= \frac{6}{6+4} = 6/10 = 0.6$
Specificity	TNR	$\frac{TN}{TN+FP}$	 $= \frac{8}{8+2} = 8/10 = 0.8$
Precision	PPV	$\frac{TP}{TP+FP}$	 $= \frac{6}{6+2} = 6/8 = 0.75$
Negative Predictive Value	NPV	$\frac{TN}{TN+FN}$	 $= \frac{8}{8+4} = 8/12 \approx 0.66$
False Negative Rate	FNR	$\frac{FN}{FN+TP}$	 $= \frac{4}{4+6} = 4/10 = 0.4$
False Positive Rate	FPR	$\frac{FP}{FP+TN}$	 $= \frac{2}{2+8} = 2/10 = 0.2$
False Discovery Rate	FDR	$\frac{FP}{TP+FP}$	 $= \frac{2}{2+6} = 2/8 = 0.25$
True Discovery Rate	TDR	$\frac{FN}{TN+FN}$	 $= \frac{4}{4+8} = 4/12 \approx 0.33$
Accuracy	ACC	$\frac{TP+TN}{TP+FP+TN+FN}$	 $= \frac{6+8}{6+2+4+8} = 14/20 = 0.7$
F1 Score	F1	$\frac{2\,TP}{2\,TP+FP+FN}$	 $= \frac{2*6}{(2*6)+2+4} = 12/18 \approx 0.66$

Figure 3-32: Our statistical measures of Figure 3-29 in visual form using the data of Figure 3-30

Constructing a Confusion Matrix Correctly

Understanding a test (or classifier) from its statistical measures can be difficult. There's a lot to take in, and keeping everything straight and organized can be a challenge. It's important to rise to this challenge, because most real-world tests (in every field) are imperfect, as are most machine-learning systems. In general, they need to be understood in terms of their statistical performances.

The confusion matrix is a simple but powerful way to simplify and summarize our understanding. But we have to build and interpret it carefully, or we can too easily come to the wrong conclusions. To wrap up this chapter, let's look more closely at how to properly build and interpret a confusion matrix.

The plan will be to return to our imaginary disease of MP, attach some numbers to our confusion matrix, and ask some questions about the quality of our fast, but inaccurate, field test. Recall that we earlier said that we'd measure everyone in a town with our slow and expensive, but perfectly accurate, lab test (giving us the *ground truth*), as well as our faster, cheaper, and imperfect field test (giving us *predictions*).

Let's suppose that these measurements show that the field test has a high true positive rate: we found that 99% of the time, someone with MP is correctly diagnosed. Since the TP rate is 0.99, the false negative (FN) rate, which contains all of the people with MP who we did *not* correctly diagnose, is 1 – 0.99 = 0.01.

The test does a bit worse for people who *don't* have MP. We'll suppose that the true negative (TN) rate is 0.98, so 98 times out of 100 when we predict that someone is not infected, they really aren't. But this means that the false positive (FP) rate is 1 – 0.98 = 0.02, so 2 people in 100 who don't have MP will get an incorrect positive diagnosis.

Let's suppose that we've just heard of a suspected outbreak of MP in a new town of 10,000 people. From experience, given the amount of time that has passed, we expect that 1% of the population is already infected. *This is essential information.* We're not testing people blindly. We *already know* that there's only a 1 in 100 chance that someone has MP. It will be essential for us to include this information to correctly understand the results from our field test.

So we pack up our gear and head into town at top speed.

There's no time to send our results to the big and slow lab, so we get everyone to come down to city hall to get tested with our field test. Suppose someone comes up positive. What should they do? How likely is it that they have MP? Suppose instead the test is negative. What should those people do? How likely is it that they *don't* have it?

We can answer these questions by building a confusion matrix. If we jump into it, we might build a confusion matrix just by popping the values above into their corresponding boxes, as in Figure 3-33. But this is *not* the way to go! This matrix is incomplete and will lead us to the wrong answers to our questions.

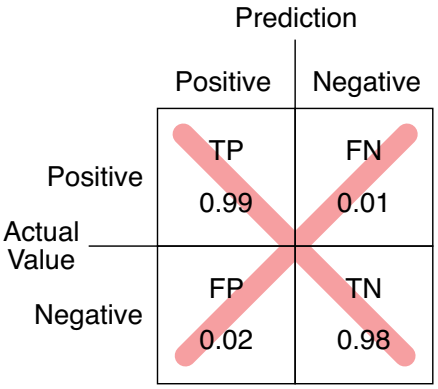


Figure 3-33: This is not the confusion matrix we're looking for.

The problem is that we're ignoring a critical piece of information: only 1% of the people in town will have MP right now. The chart in Figure 3-33 doesn't include that knowledge and therefore isn't telling us what we need to know.

In Figure 3-34, we work out the proper matrix by considering the 10,000 people in town and analyzing what we expect from the test by using our knowledge of the infection rate and the test's measured performance.

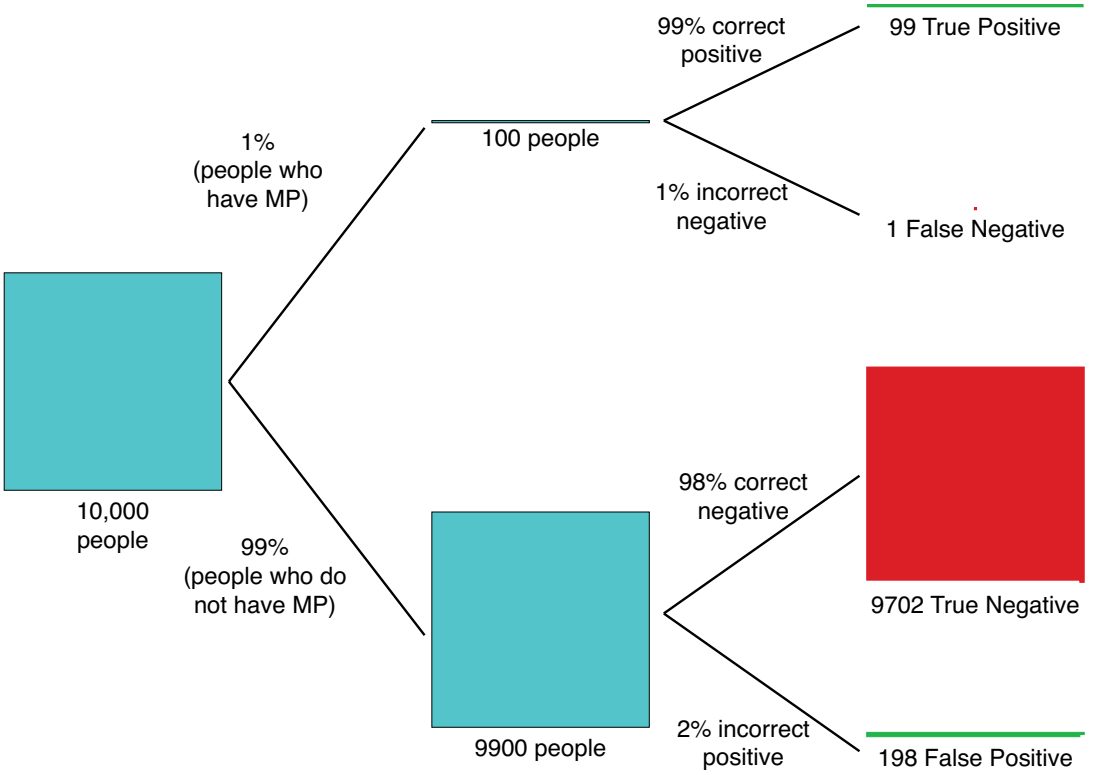


Figure 3-34: Working out the populations we expect from our infection rate and our test

Figure 3-34 forms the heart of the correct process, so let's walk through it. We start at the left with 10,000 people in town. Our essential starting information is that we already know from prior experience that 1 person out of 100, or 1% of the population, will be infected with MP. That's shown in the upper path, where we show the 1% of 10,000, or 100, people who have MP. Our test will correctly come up positive for 99 of them, and negative for only 1. Returning to our starting population, on the lower path we follow the 99%, or 9900, people who are not infected. Our test will correctly identify 98% of them, or 9,702 people, as being negative. 2% of those 9900, or 198 people, will get an incorrect positive result.

Figure 3-34 tells us the values that we *should* use to populate our confusion matrix, because they incorporate our knowledge of the 1% infection rate. From our 10,000 tests, we'll expect (on average) 99 true positives, 1 false negative, 9,702 true negatives, and 198 false positives. These values give us the proper confusion matrix in Figure 3-35.

		Prediction: Positive	Prediction: Negative
Actual Value	Positive	TP 99	FN 1
	Negative	FP 198	TN 9702

Figure 3-35: The proper confusion matrix for our MP test, incorporating our knowledge of the 1% infection rate

Comparing this to the matrix in Figure 3-33, the TN rate has changed by a lot! Instead of 98 (what we'd get by multiplying 0.98 by 100), we have 9702. The value for FP has also gone through a huge change, from 2 to 198. These improved values will make a big difference in our interpretations of the test results.

Now that we have the right matrix, we're ready to answer our questions. Suppose someone gets a positive test result. What's the chance that they really do have MP? In statistical terms, what's the conditional probability that someone has MP, given that the test says they do? More simply, what percentage of the positive results we get back are true positives? That's just what precision measures. In this case, the precision is 99 / (99 + 198), or 0.33, or 33%.

Wait a second. What does this mean? Our test has a 99% probability of correctly diagnosing MP, yet 2/3 of the times when it gives us a positive result, that person does *not* have the disease. More than half of our positive results are wrong!

That definitely seems weird.

And that's why we're going through this example. Understanding probabilities can be tricky. Here we have a test with a 99% true positive rate, which sounds pretty great. Yet the majority of our positive diagnoses are wrong.

This surprising result comes about because even though the chance of missing an infected person is very small, there's a huge number of healthy people being tested. So we get a whole lot of those rare incorrect positive diagnoses, and they add up fast. The result is that if someone gets a positive result, we should *not* operate right away. We should instead interpret this result as a signal to do the more expensive and accurate test.

Let's look at these numbers using our region diagrams. We'll have to distort the sizes of the areas in Figure 3-36 in order to make them big enough to see.

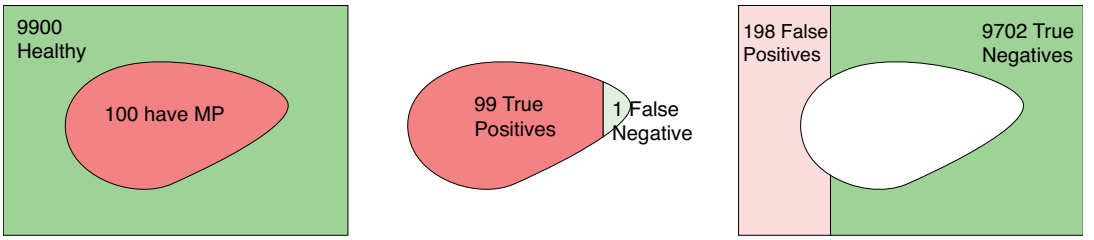


Figure 3-36: Left: The population contains 100 people with MP, and 9900 without. Middle and Right: The results of our test. The sizes of the shapes are not to scale.

We saw earlier that the precision tells us the chance that someone who is diagnosed as positive really does have MP. This is illustrated at the far left of Figure 3-37. We can see that the field test incorrectly labels people without MP as positive, giving us a precision of 0.33. That tells us to be suspicious of positive results, because $1 - 0.33 \approx 0.66$, or 66%, of those results will be wrong.

What if someone gets a negative result? Are they really clear? That's the ratio of true negatives to the total number of negatives, or TN / (TN + FN), which Figure 3-29 gives the name of *negative predictive value*. In this case it's 9702 / (9702 + 1). That's well over 0.999, or 99.9%. So if someone gets back a negative result, there's only about 1 chance in 10,000 that the test was wrong and they do have MP. We can tell them that, and let them decide if they want the slower, more expensive test.

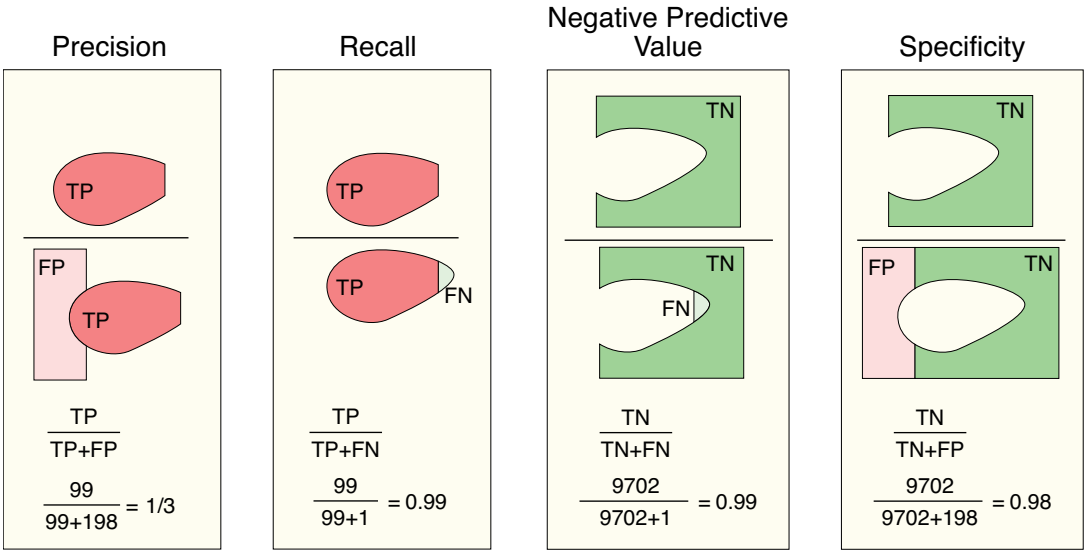


Figure 3-37: Four statistics describing our test for MP based on the results of Figure 3-36. Precision: What percentage of our positives are accurate? Recall: What is our percentage of finding all the positives? Negative Predictive Value: What percentage of our negatives are accurate? Specificity: What is our percentage of finding all the negatives?

To summarize, the chance that a positive result means that someone actually does have MP is only about 33%. On the other hand, a negative result is 99.9% sure to be really negative.

Figure 3-37 shows a couple of other measurements. The recall tells us the percentage of people that are properly diagnosed as positive. Since we only missed one person out of 100, that value is 99%. The specificity tells us the percentage of people that are properly diagnosed as negative. Since we gave 198 incorrect negative diagnoses, that result is a little less than 1.

To summarize, out of 10,000 people in this town with a 1% infection rate, out test will only miss 1 case of MP. But we'll get nearly 200 incorrect positive diagnoses (that is, false positives), which can unduly scare and worry people. Some might even have the surgery right away, rather than wait for the slower test. Since we wanted to be very sure of correctly finding every person with MP, our test is overly zealous in telling people they're infected.

As we saw earlier, if we wanted to make a test that would never miss any person with MP, we could simply label everyone as positive, but that's not useful. The goal in real situations with imperfect systems is to balance the false negatives and false positives in a way that serves our purposes, while keeping those errors in mind.

Our example of MP was imaginary, but the real world is full of situations where people are making decisions based on incorrect confusion matrices or bad questions. And some of those decisions are related to real and very serious health issues.

For example, many women have had needless mastectomies because their surgeons misunderstood the probabilities from a breast exam and gave their patients bad counseling (Levitin 2016). Recommending someone undergo an unnecessary surgery is a dangerous mistake. Men were also operated on without cause, because many were given bad advice based on their doctors misunderstanding the statistics of using elevated PSA levels as evidence for prostate cancer (Kirby 2011).

Probability and statistics can be subtle. It's essential that we go slow, think things through, and make sure that we're interpreting our data correctly.

Now we know that we shouldn't be fooled by hearing that some test is "99% accurate," or even that it "correctly identifies 99% of the positive cases." In our town where only 1% of the people are infected, and a test with an impressive 99% true positive rate, anyone with a positive diagnosis is more than likely to *not* really have the disease.

The moral is that statistical claims in any situation, from advertising to science, need to be looked at closely and placed into context. Often, terms like "precision" and "accuracy" are used colloquially or casually, which, at best, makes them difficult to interpret. Even when these terms are used in their technical sense, bare claims of accuracy and related measures can easily be misleading and can lead to poor decisions.

When it comes to probability, don't trust your gut. There are surprises and counterintuitive results that lie in wait all over the place. Go slow and think it through.

Summary

We've seen a lot in this chapter! We covered some of the most important ideas in probability. We saw a term for how likely it is for some event A to happen, or for some event A to happen given that some other event B already happened, or for events A and B to happen together.

We then looked at a few statistical measures that let us characterize how well a test is able to properly identify the positive and negative samples in a dataset. We saw that we can use these measures to help us interpret the results of any decision-making process. And we organized those terms into a confusion matrix, which helps us make sense of all that information.

And we saw that statistics can be misleading. If we're not careful, we can create tests (or classifiers) that seem to do a great job according to one set of measurements, but are lousy in other ways. It's important to go slow, think things through, and use language carefully when working with probability.

In Chapter 4, we'll apply some of these ideas to a method of reasoning about probabilities that is widely used in machine learning. This will give us another tool to help us later design learning algorithms that will learn, and are able to usefully perform the tasks we want of them.

References

Glen, S. 2014. "Marginal Distribution." Statisticshowto.com. *http://www.statisticshowto.com/marginal-distribution/.*

Jaynes, E.T. 2003. *Probability Theory: The Logic of Science.* Cambridge, UK: Cambridge University Press.

Kirby, Roger. 2011. *Small Gland, Big Problem.* London, UK: Health Press.

Levitin, Daniel J. 2016. *A Field Guide to Lies: Critical Thinking in the Information Age.* New York: Viking Press, 2016.

Walpole, Ronald E., Raymond H. Myers, Sharon L. Myers, and Keying E. Ye. 2011.*Probability and Statistics for Engineers and Scientists (9th Edition).* New York: Pearson.

Wikipedia. 2016. "Sensitivity and Specificity." *https:// en.wikipedia.org/wiki/Sensitivity_and_specificity.*