Binary Classification

Kevin Sun (nusnivek)

Consider a standard True/False (T/F) quiz: a student is presented a sequence of statements and must identify whether each statement is true or false. Their final score is the total number of questions answered correctly, divided by the total number of questions.

A True/False quiz is an example of a general problem called *binary classification*. It sounds fancy, but it's not that bad. "Classifying" something means assigning it to a particular group, also known as a *class*. The "binary" part simply means that there are only two possible classes. (In the case of a T/F quiz, the two classes are "True" and "False".)

There are many other examples of binary classification tasks, and they appear in a wide variety of contexts. Here are just a few:

- An email service (e.g., Gmail) must classify an email as "Spam" or "Not Spam".
- An HIV test must classify a patient's blood as "Has HIV" or "No HIV".
- An interviewer must classify an applicant as "Accept" or "Reject".
- A jury must classify a defendant as "Guilty" or "Not Guilty".

What if there are multiple classes?

A generalization of binary classification is known as *multiclass* (or *multinomial*) classification, and one familiar example is the multiple choice quiz. In this case, the classes are usually "A", "B", "C", and "D". Of course, there are also many examples of this in the real world. Post offices must classify handwritten numbers as "0", "1", "2", etc. Libraries must classify books as "Horror", "Science Fiction", "Fantasy", etc. And self-driving cars must classify objects as "Person", "Sign", "Car", etc. But in this lesson, we'll focus on the case where there are only two classes.

1 Terminology

Now that we've seen how True/False quizzes are just another binary classification problem, we shall introduce some terminology. In a generic binary classification problem, the names of the two classes are "Positive" and "Negative". This follows the language of medical tests: a blood test result is *positive* if it has a particular disease, and *negative* otherwise. The *samples* are the things that have to be classified, and the *test* is the thing doing the classifying. The table below illustrates this terminology being applied to the examples of binary classification tasks given above.

Test	Sample	"Positive"	"Negative"
student	a statement on T/F quiz	"True"	"False"
email service	an email	"Spam"	"Not Spam"
HIV test	a patient's blood	"Has HIV"	"No HIV"
interviewer	an applicant	"Hire"	"Do Not Hire"
jury	a defendant	"Guilty"	"Not Guilty"

In binary classification, we assume that every sample is either in the "Positive" class or the "Negative" class. The test must predict, or guess, the class of each sample based on how it looks (without knowing its actual class, of course). Thus, for each sample, there are four possibilities, depending on the *predicted* class and the *actual* class (as shown below).

	Sample is Positive (P)	Sample is Negative (N)
Test guesses "Positive"	**True Positive** (TP)	**False Positive** (FP) Type I error
Test guesses "Negative"	**False Negative** (FN) Type II error	**True Negative** (TN)

Notice that the name of each possibility has two parts: True/False (whether or not the test was correct), and Positive/Negative (whether the test output "Positive" or "Negative"). The errors are the ones that start with "False", so there are two types of errors: False Positives (also known as Type I), and False Negatives (Type II).

Naming conventions

Don't confuse the "True/False" here with "True/False" quizzes! Here, "True" is an adjective meaning "correctly classified" whereas previously, "True/False" were the names of the two classes (which we've renamed to be "Positive/Negative").

It is also valid to rename the "False" class as "Positive", but generally speaking, "Positive" refers to the *presence* of something and "Negative" refers to the *absence* of something. For example, a student must determine whether or not a statement on a T/F quiz *has* validity, and an interviewer must determine whether or not an applicant *has* the qualifications to work at the company. Also, generally speaking, the "Negative" class corresponds to the larger "default" class (e.g., "Not Guilty").

Every sample is either positive or negative, so the total number of samples is P + N. Also, since every positive sample is either correctly classified (TP) or incorrectly classified (FN), the total number of positive samples is equal to the sum of True Positives and False Negatives, i.e., P = TP + FN. Similarly, N = TN + FP.

2 Scoring methods

So what can we do with TP, FP, TN, and FN? Consider the following situation: you're trying to decide if you should use Email Service A (ESA) or Email Service B (ESB), so you want to compare their spam filters. You currently have a batch of 100 emails, and you've manually identified 20 of them as "Spam" (i.e., in the "Positive" class). Naturally, you run both services on your batch of 100 and see how they perform:

	TP	TN	FP	FN
Email Service A	19	70	10	1
Email Service B	15	78	2	5

(Notice that in each row, the four numbers add up to 100, TP+FN = 20, and TN+FP = 80.) Out of the 20 spam emails, ESA got 19 correct while ESB only got 15. However, out of the 80 non-spam emails, ESA only got 70 correct while ESB got 78. So which one is "better"?

It's tempting to judge a test based on its *accuracy* — the fraction of all samples that it classified correctly. But as we'll see, this can lead to misleading conclusions. To reduce the chances of this happening, we should also consider two other methods of scoring: *sensitivity* and *specificity*. All three scoring methods are defined below.

1. **Accuracy:** Out of **all of the samples**, determine what fraction the test correctly answered. Mathematically, this is $(\mathrm{TP}+\mathrm{TN})/(\mathrm{P}+\mathrm{N})$. Note that this is the way that teachers typically grade True/False quizzes.

2. **Sensitivity:** Out of **all of the *positive* samples**, determine what fraction the test correctly answered (i.e., guessed "Positive"). Mathematically, this is TP/P.

3. **Specificity:** Out of **all of the *negative* samples**, determine what fraction the test correctly answered (i.e., guessed "Negative"). Mathematically, this is TN/N.

For all three measurements, the minimum score is 0% and the maximum score is 100%. So a perfect test has 100% accuracy, 100% sensitivity, and 100% specificity, while a test that misclassifies *every* sample would get 0% across the board. For the ESA/ESB data given above, we have the following values.

	Accuracy	Sensitivity	Specificity
Email Service A	89/100 = 89%	19/20 = 95%	70/80 = 87.5%
Email Service B	93/100 = 93%	15/20 = 75%	78/80 = 97.5%

Okay, but which method is the *best*? If we could only use a single scoring method, then accuracy would be the best choice.[1] To see this, suppose a student's grade on a T/F quiz was completely determined by sensitivity (instead of accuracy). In other words, the teacher only looks at the true statements, and counts the fraction of these that the student answered "True". In this case, the student could "hack" the quiz and guarantee themselves a sensitivity of 100% by simply classifying every statement as "True"! Similarly, if the teacher only cared about specificity, then the student could guarantee themselves 100% by classifying every statement as "False". Notice that accuracy can't be similarly "hacked" because *every* question is graded, so every mistake gets caught.

[1] This is assuming that false positives and false negatives have the same "cost," as is the case on a standard True/False quiz. As we'll see later, this assumption generally does not hold.

So why don't we just use accuracy? The problem with *just* using accuracy is that it can lead to misleading conclusions. For example, suppose there are two HIV tests: Test A and Test B. There are 100 patients, and each patient gets their blood tested twice: once by Test A, and once by Test B. We discover that the accuracy of Test A is 95% while the accuracy of Test B is only 90%. Thus, if we go with accuracy, we would conclude that Test A is better than Test B.

But then we learn the following: *Test A always outputs "No HIV"*. Its accuracy was 95% because out of the 100 patients, only 5 of them have HIV (so P = 5, N = 95). Moreover, we find out that *Test B correctly classified all 5 positive samples.* Using this additional information, we can deduce the sensitivity and specificity of both tests (shown below).

	Accuracy	Sensitivity	Specificity
Test A	95/100 = 95%	0/5 = 0%	95/95 = 100%
Test B	90/100 = 90%	5/5 = 100%	85/95 ≈ 89.5%

This example illustrates why accuracy is not enough when evaluating solutions to binary classification problems in the real world. A bogus test (e.g., Test A) could, depending on the samples being tested, achieve higher accuracy than a genuine test (Test B). Of course, we hope that there aren't any bogus tests in real life. But still, if we only look at the table above — is Test B *actually* better? Maybe the increased sensitivity isn't "worth" the decreased in specificity. How can we tell? This is the topic of the final section.

Remembering "sensitivity" and "specificity"

Recall the typical definition of sensitive: *quick to detect slight changes.* For example, a sensitive tooth detects more pain than a non-sensitive tooth. Increasing the sensitivity of your mouse causes the cursor to move more quickly. In our context, an HIV test that is highly sensitive would quickly detect the presence of HIV in someone's blood.

On the other hand, the opinions of a picky eater can be highly *specific.* Perhaps they only eat chicken nuggets at a particular temperature from a particular restaurant: everything else gets rejected. If a teacher gives highly specific instructions for an assignment, then they're likely to reject many submissions for not meeting their standards. Similarly, an HIV test that is highly specific would reject any blood sample that doesn't contain "high enough" levels of HIV.

3 Finding the tradeoff

For most binary classification tasks, there's a tradeoff between minimizing false positives (FP) and minimizing false negatives (FN). A test that errs on the side of "Positive" has a higher risk of FPs but lower risk of FNs, and vice versa for a test that errs on the side of "Negative". So in general, how do we decide which test is better? **The answer depends on the *cost* of the errors.**

In different contexts, the *cost*[2] of an FP and the *cost* of an FN can be very different. For example, the FP cost from an HIV test might be that the patient takes medications that are expensive and unnecessary. On the other hand, the FN cost might be that the patient's health deteriorates due to the untreated HIV. It's often unclear which error (FP or FN) incurs a higher cost because they depend on many factors (e.g., the severity of the virus, the patient's financial situation), some of which could be fuzzy or unknown.

A toy situation. Let's consider the fictitious email services ESA and ESB again, whose results on 100 emails (20 spam, 80 non-spam) are shown below.

	TP	TN	FP	FN
Email Service A	19	70	10	1
Email Service B	15	78	2	5

Recall that ESA has an accuracy of 89% while ESB has an accuracy of 93%, so ESB has higher accuracy. In this situation, the cost of an FP is having a spam email in your inbox; let's call this 1 unit of "sadness." The cost of an FN is potentially missing an important email; let's estimate this as 3 units of sadness. Then ESA has an overall cost of 13 while ESB has an overall cost of 17. So even though ESB had a higher accuracy than ESA, its overall cost was higher due to the large number of false negatives. Put another way, ESB should have guessed "Positive" more times, so its sensitivity was too low.

3.1 A more realistic situation

Let's end with a more realistic (and difficult) situation: the jury/defendant problem. (Recall that "Positive" corresponds to the defendant being guilty.) The FP cost is that a non-criminal gets punished (e.g., goes to jail); the FN cost is that a criminal gets away. Obviously it's impossible to quantify these costs, but our general sentiment is that an FP costs more than an FN. This is why many societies hold the legal principle of "innocent until proven guilty," so juries should err on the side of "Not Guilty".

A thought experiment. Suppose there are $n + 1$ defendants: one of them is a lovely person who has never committed any crimes. The remaining n people have all committed heinous crimes, and they plan to commit more heinous crimes. Unfortunately, there is not enough evidence to separate the n criminals from the one innocent person. So your task is the following: **you must *simultaneously* classify *all* $n + 1$ defendants as "Not Guilty" or "Guilty"**.

If you choose "Not Guilty", all $n + 1$ defendants are released. If you choose "Guilty", all $n + 1$ defendants get jailed for life. When $n = 0$, the choice is obvious. But how large does n need to be for you to cross the line from "Not Guilty" to "Guilty"? In other words, when does the "cost" of n free criminals surpass the "cost" of a jailed non-criminal?

[2] In this context, "cost" doesn't necessarily refer to a financial value (though it could). Instead, it refers to the total loss incurred by all parties (one of which might be"society"), which could include things such as money, time spent, emotional well-being, and environmental impact.

	Accuracy	Jailed non-criminals (FPs)	Free criminals (FNs)
"Not Guilty"	$1/(n+1) \to 0\%$	0	n
"Guilty"	$n/(n+1) \to 100\%$	1	0

Of course, criminal trials don't actually work this way, but the idea is the same. Suppose there are ten defendants, and the jury decides "Guilty" for each one with 90% certainty. Then the probability that they're all guilty is only $(0.9)^{10} \approx 35\%$, which means there's a good chance (at least 65%) that at least one innocent person got jailed. In other words, by convicting at 90% certainty, this jury crosses the line from "Not Guilty" to "Guilty" when n reaches 9. Whether you think this is high or low depends, at least partially, on whether or not you think 90% qualifies as "beyond a reasonable doubt" (a common legal term).

The main takeaway is this: even when there are only two choices, making the "right" choice can be very difficult (and is often fraught with subjectivity). But understanding that there's a tradeoff between FPs and FNs, and quantifying that tradeoff as clearly as possible, is a fundamental aspect of making an informed decision.

4 Summary

- A binary classification problem involves a *test* determining whether a *sample* is in the "Positive" class or "Negative" class. For example, an email service must determine whether an email is "Spam" or "Not Spam".
- A test can make two kinds of errors: a *false positive* occurs when the test guess that a negative sample is positive, and a *false negative* is the opposite error.
- The most familiar way to evaluate the results of a test is looking at its *accuracy*, but we should also consider its *sensitivity* and *specificity*. A test with 95% accuracy might not actually be "better" than a test with 90% accuracy.
- There is a tradeoff between minimizing false positives and minimizing false negatives. Finding the right balance involves calculating the overall cost of each error, which is often abstract and difficult to fully determine.
- **Further reading.** Binary classification is just one problem that people study in a field known as *machine learning* (ML), sometimes described as the intersection of computer science and statistics. There are tons of huge ML applications; a few notable examples include medical diagnosis, online advertising, and self-driving cars.