Is it impossible to be fair?

draft for JFI blog post

last modified: September 25, 2018

1 TL;DR

2 Crates and boxes

2.1 Toy situation

Here's a toy situation:

> Imagine a crate of locked boxes. Each box is labeled with a number. Inside each box is a ball, which is either red or green, and a cube, which is either blue or yellow. You draw a box uniformly at random from the crate, read its label, and predict the color of the ball inside.

[INSERT FIGURE: literal crate, boxes, and person making prediction.]

The toy situation is odd. Why bother thinking about it? Because, surprising as it might seem, the toy situation models important practical situations, so results proved about the toy situation apply to these important practical situations too. In particular, some recent results about the toy situation seem to show that in important practical situations, it's impossible to be fair: no matter how the predictions are made—whether by human or machine, whether based on algorithm, experience or intuition, and however much evidence is gathered—the predictions aren't fair. That's why the toy situation matters. We'll get to the practical situations in Section TBD. For now let's think more about the toy situation.

2.2 Representing a crate

Back to the crate of boxes. We can represent the contents of a crate with frequency tables.

	blue	yellow
red	2	2
green	2	1

label: '1'

	blue	yellow
red	1	2
green	2	2

label: '2'

[TBD Maybe a better running example would have three labels and be less symmetrical?]

The tables show which boxes are in the crate. For example: there are 2 boxes containing a red ball, blue cube and labeled '1'; there is 1 box containing a red ball, blue cube and labeled '2'. And so on. For each color of ball (red or green), each color of cube (blue or yellow) and each label (here, '1' or '2'), the tables have an entry telling us how many boxes containing that color of ball, that color of cube, and with that label are in the crate. From this, we can infer other information. For example: there are $2 + 2 + 2 + 1 + 1 + 2 + 2 + 2 = 14$ boxes in total; there are $2 + 2 + 2 + 1 = 7$ boxes labeled '1'; there are $1 + 2 = 3$ boxes labeled '2' and containing a red ball.

2.3 Working out probabilities

You draw a box uniformly at random from the crate. That means all of the 14 boxes have an equal chance, 1 in 14 or about 7%, of being drawn. It's like drawing names from a hat. That's a helpful feature of the toy situation. It makes life simple. It means that working out the probabilities of drawing boxes of different kinds just comes down to counting.

Examples. The probability of drawing a box containing a red ball, blue cube and labeled '1' is 2 in 14, or about 14%, because there are 2 boxes of that kind and 14 boxes in total. Similarly, the probability of drawing a box containing a red ball, blue cube and labeled '2' is 1 in 14, or about 7%, the probability of drawing a box labeled '1' is 7 in 14, or 50%, and the probability of drawing a box labeled '2' and containing a red ball is 3 in 14, or about 21%.

We can get fancier. The probability of drawing a box which is either labeled '1' or contains a red ball is 10 in 14, or about 71%. We can get fancier still. What's the probability of drawing a box containing a red ball, given that you draw a box labeled '1'? Well, there are 7 boxes labeled '1' and, of these 7 boxes, 4 contain a red ball. So it's 4 in 7, or about 57%. Similarly, the probability of drawing a box containing a yellow cube, given that you draw a box containing a green

ball, is 3 in 7, or about 43%. And the probability of drawing a box containing a green ball and yellow cube, given that you draw a box labeled '2' or containing a blue cube, is 3 in 11, or about 27%. When we have the frequency tables and draw boxes uniformly at random, calculating probabilities is easy. Just count.

2.4 Predictions and strategies

Remember your task in the toy situations: you draw a box uniformly at random, read its label and then guess the color of the ball inside. If the boxes were unlocked, you could just open them and check the color of the ball directly. But life is hard. You have to proceed indirectly, by reading the label and then predicting the color of the ball.

The description of the toy situation doesn't specify what kinds of predictions you can make. It might be, for example, that you have to make categorical predictions: predict red or predict green, like a jury's verdict in court. More sophisticated predictions are possible, such as assigning a probability to the ball's being red or being green. We'll see other options later.

It's helpful to distinguish *predictions* and *strategies*. After you draw a box from the crate, you make a prediction based on its label. Before you draw a box from the crate, you don't know what the label will be. But you can decide for each label what to predict if you draw a box with that label. You're deciding your strategy. A strategy, then, is a complete contingency plan: it lists your prediction for each label. It takes the form: if <label>, then <prediction>. We're interested in properties of the strategies.

Let's look at a simple example. Suppose the crate is as in Figure TBD and you have to make categorical predictions. Given that you have to make categorical predictions, four strategies are available: if '1' predict red and if '2' predict red; if '1' predict red and if '2' predict green; if '1' predict green and if '2' predict red; if '1' predict green and if '2' predict green. For each strategy we can ask, *What's the probability of predicting incorrectly when using that strategy?* That property—let's call it the strategy's *overall error rate*—is one of the properties we'll be interested in.

Take the first strategy: if '1' predict red and if '2' predict red. What's the probability you predict incorrectly when using that strategy? Well, when using that strategy you predict incorrectly just if you draw a box containing a green ball, and the probability of drawing a box containing a green ball is 7 in 14, or 50%. So the strategy's overall error rate is 50%.

Take the second strategy: if '1' predict red and if '2' predict green. What's the probability you predict incorrectly when using that strategy? Well, when using that strategy you predict incorrectly just if you draw a box labeled '1'

containing a green ball or you draw a box labeled '2' containing a red ball, and the probability of that is 6 in 14, or about 43%. So the strategy's overall error rate is about 43%. Similarly, the third strategy's overall error rate is about 57% and the fourth's is 50%.

The strategy which has the lowest overall error rate is the second: if '1' predict red and if '2' predict green. You might have guessed this just by looking at the frequency tables. Among the boxes labeled '1' more are red than green and among the boxes labeled '2' more are green than red. So it makes sense that predicting red for boxes labeled '1' and green for boxes labeled '2' has the lowest overall error rate.

For each strategy we can also ask, *What's the probability you predict incorrectly when using that strategy, given that you draw a box containing a blue cube?* and *What's the probability you predict incorrectly when using that strategy, given that you draw a box containing a yellow cube?* These properties—let's call them the strategy's *error rate among blue cube boxes* and *error rate among yellow cube boxes*—are like the overall error rate, but relative to a subpopulation (blue cube boxes, yellow cube boxes) rather than the whole population (all boxes).

Take the first strategy. What's the probability you predict incorrectly when using that strategy, given that you draw a box containing a blue cube? Well, as we already noted, you predict incorrectly just if you draw a box containing a green ball. There are 7 boxes containing a blue cube and among these 7 boxes 4 contain a green ball. So the strategy's error rate among blue cube boxes is 4 in 7, or about 57%. Similarly, its error rate among yellow cube boxes is 3 in 7, or about 43%. Note that the error rates across blue and yellow cube boxes are different: the probability that you predict incorrectly given that you draw a box containing a blue cube is significantly higher than given that you draw a box containing a yellow cube.

Take the second strategy. What's the probability you predict incorrectly when using that strategy, given that you draw a box containing a blue cube? Well, as we already noted, you predict incorrectly just if you draw a box labeled '1' containing a green ball or you draw a box labeled '2' containing a red ball. There are 7 boxes containing a blue cube and among these 7 boxes 2 are labeled '1' and contain a green ball and 1 is labeled '2' and contains a red ball. So the strategy's error rate among blue cube boxes is 3 in 7, or about 43%. Similarly, its error rate among yellow cube boxes is also 3 in 7, or about 43%. The error rates across blue and yellow cube boxes are the same.

Exercise: Work out the third and fourth strategies' error rates among blue and yellow cube boxes.

This will be the general pattern in what follows: we'll think about properties of strategies—among all boxes, among blue cube boxes, and among yellow cube

boxes.

2.5 Form of results

Most of the results described later are negative: they tell us that *no* strategy has certain combinations of properties. More carefully, the results are of the form:

> No strategy of kind *so-and-so* has *such-and-such* properties (unless the contents of the crate happens to be like *this-and-that*).

The coming sections will fill in the place-holders: the so-and-so, such-and-such and this-and-that.

3 Why care?

Why care about the toy situation? Because the toy situation models important practical situations, so results proved about the toy situation apply to these important practical situations too. Let's look at some examples.

3.1 First example

Think of prisoners coming up for parole. Each prisoner has a case file giving information about their situation: their age, what they were convicted of, their behavior in prison, their prospects if released, and so on. Prisoners come before a judge, who reads their case files and decides whether to grant them parole. Some prisoners, if they were granted parole, would re-offend within a year; others wouldn't. The judge takes that into consideration: she predicts based on the case file whether the prisoner would re-offend if granted parole and her prediction is a factor—one among many—in her decision.

To say that the judge predicts whether the prisoner would re-offend is not to say much. It's not to say, for example, that the judge writes down her prediction in her report. It's merely to say that whether the prisoner would re-offend is one factor in her decision. Let's focus on that factor.

How should the judge predict based on the case file whether the prisoner would re-offend? That's a very hard question. But we can identify one constraint. Prisoners coming up for parole can be divided into groups: say, white and non-white. However the judge makes predictions based on the case file—whatever her strategy—it shouldn't be biased in favor of white prisoners and against non-white prisoners, or vice versa.

3.2 Second example

Think of professors coming up for tenure. Each professor has a tenure file giving information about their situation: their areas of specialization, their publication record, recommendation letters, and so on. Tenure applications come before the dean, who reads the tenure files and decides whether to grant tenure. Some professors, if they were granted tenure, would bring in grant money for the university; others wouldn't. The dean takes that into consideration: she predicts based on the tenure file whether the professor would bring in grant money and her prediction is a factor—one among many—in her decision.

To say that the dean predicts whether the professor would bring in grant money is not to say much. It's not to say, for example, that the dean writes down her prediction in her report. It's merely to say that whether the professor would bring in grant money is one factor in her decision. Let's focus on that factor.

How should the dean predict based on the tenure file whether the professor would bring in grant money? That's a very hard question. But we can identify one constraint. Applicants for tenure can be divided into groups: say, male and female. However the dean makes predictions based on the case file—whatever her strategy—it shouldn't be biased in favor of male professors and against female professors, or vice versa.

3.3 Decision situations

General description of prediction scenario, e.g. handwriting recognition, quality control, to show how general problem is.

COMPAS

3.4 How do the crates and boxes fit in?

Spelling out the analogy:

boxes	prisoners	faculty members
color of ball (red or green)	re-offend (yes or no)	successful career (yes or no)
color of cube (blue or yellow)	race (white or non-white)	sex (male or female)
label	case file	tenure file

Just as we can't directly check the color of the ball, so too we can't directly check whether the prisoner will re-offend or the professor will have a successful career. (Of course, we can wait and see. But we can't directly check *now*.) We

can, however, use the label, case file, or tenure file as a guide.

Two points. (1) Are race and gender, or things correlated with them, included in the case and tenure files? Maybe or maybe not. The results I describe don't assume either way. (2) The judge or the dean may base her prediction on more than the case or tenure file. Perhaps she knows the person under consideration already; perhaps she's told things not written down in the file. That doesn't matter. The case or tenure files should be taken as place-holders: they stand for everything on which the decision-maker can base her decision.

Does the toy situation model these real-life situations? On the one hand, it's easy to see where the probabilities come from for the crate of boxes: you draw boxes uniformly at random. It's not so easy to see where probabilities come from for the prisoners or professors: prisoners and professors don't come up for parole or tenure at random. On the other hand, scientific practice shows that probabilistic models can be useful models of real-life phenomena, even when it's not clear where the probabilities come from in the real-life phenomena. In any case, I won't pursue the issue here.

4 What kinds of predictions can you make?

What kinds of predictions can you make? We'll consider a few options.

Remember that a strategy is a contingency plan: it lists your prediction for each label. It takes the form: if <label>, then <prediction>. So when we change the kinds of predictions you can make, we also change the strategies available.

4.1 Decision rules

The simplest kind of prediction, which we discussed earlier, is categorical: predict red or predict green. The corresponding strategies take the form: if <label>, predict <color>. Let's call these strategies *decision rules*. Here's a picture of a particular decision rule for the crate in Figure TBD:

[INSERT FIGURE]

4.2 Risk assignments

A more sophisticated kind of prediction is probabilistic. A *probabilistic prediction* assigns a probability to the ball's being red—65%, say. (The probability assigned to the ball's being green is then determined, since the probabilities sum

to 100%.) The corresponding strategies take the form: if <label>, predict red with probability $< X\% >$. Let's call these strategies *risk assignments*. Here's a picture of a particular risk assignment for the crate in Figure TBD:

[INSERT FIGURE]

4.3 Randomization

A decision rule is a deterministic function of the label: same label, same (categorical) prediction. A risk assignment is too: same label, same (probabilistic) prediction. Let's generalize. We won't introduce a new kind of prediction. Instead, we'll introduce a new way of moving from a label to a prediction (whether categorical or probabilistic). The new way is to *randomize*.

Suppose you read a box's label, then flip a coin, and predict red if it lands heads and green if it lands tails. You are implementing a decision rule with randomization. Your prediction is either red or green—categorical predictions—but which prediction you make depends on how the coin lands. Or suppose you read a box's label, then roll a die, and predict red with probability 20% if it lands on 1 or 2, predict red with probability 40% if it lands on 3, 4 or 5, and predict red with probability 90% if it lands 6. You are implementing a risk assignment with randomization. Your prediction is either red with probability 20% or with probability 40%, or with probability 90%—probabilistic predictions—but which prediction you make depends on how the die lands.

In general, a decision rule with randomization takes the form: if <label>, with probability $< p >$ predict *red*. Here's a picture:

[INSERT FIGURE]

In general, a risk assignment with randomization takes the form: if <label>, with probability $< p >$ predict *red with probability* $< X\% >$. Here's a picture:

[INSERT FIGURE]

Risk assignments with randomization use probabilities in two places. Look at the picture. The numbers on the arrows describe the randomization. The numbers in what the arrows point to describe the probabilistic predictions. A probability is part of the content of your prediction—red with probability 60%, say. But also which prediction you make is random—with this probability make this prediction; with that probability, make that prediction.

We've now defined four kinds of strategies: decision rules with and without randomization, and risk assignments with and without randomization. In the next section, we'll define some properties of these strategies. Other strategies

are possible too. These four don't exhaust the options. But they're the strategies we'll focus on.

5 Properties of strategies

5.1 Notation

To save time and ink, let's introduce some notation. For example, instead of writing

> the probability of drawing a box containing a green ball and yellow cube, given that you draw a box labeled '2' or with a blue cube, is 3 in 11,

let's write

> P(Ball = green, Cube = yellow | Label = 2 or Cube = blue) = 3/11.

Or even more briefly,

> P(B=g, C=y | L=2 or C=b) = 3/11.

Similarly, instead of writing

> the probability of incorrectly predicting red is 50%,

we can write

> P(B=g, X=r) = 50%.

I use 'X' instead of 'P' to stand for your prediction, since we already use 'P' to stand for probability.

5.2 Properties of decision rules without randomization

In Section TBD, we looked at a property of decision rules, the overall error rate. Given a crate, a decision rule's overall error rate is the probability of predicting incorrectly when using that decision rule. We also looked at relatives of that property, the error rates among blue and yellow cube boxes. Given a crate, a decision rule's error rate among blue cube boxes is the probability of predicting incorrectly when using that decision rule, given that you draw a box containing a blue cube. And similarly for the error rate among yellow cube boxes.

Let's define some more properties. Here are four key ones:

property	definition in words	definition in symbols
true red probability	probability of correctly predicting red	P(B=r, X=r)
true green probability	probability of correctly predicting green	P(B=g, X=g)
false red probability	probability of incorrectly predicting red	P(B=g, X=r)
false green probability	probability of incorrectly predicting green	P(B=r, X=g)

Several other interesting properties are determined by these. The false green rate is the probability of drawing a box for which you predict green, given that the box contains a red ball. In symbols, P(X=g | B=r). Similarly, the false red rate is the probability of drawing a box for which you predict red, given that the box contains a green ball. In symbols, P(X=r | B=g). The red prediction error is the probability of drawing a box containing a green ball, given that you predict red. In symbols, P(B=g | X=r). Similarly, the green prediction error is the probability of drawing a box containing a red ball, given that you predict green. In symbols, P(B=r | X=g).

We've now defined nine properties of a decision rule: the true/false red/green probabilities, the false red/green rate, the red/green prediction error, and the overall error rate. We could define more if we like, but these ones are key and plenty to be going on with.

The nine properties are not independent. Far from it: they are intimately linked. This is easy to see in our toy situation, because probabilities all come down to counting. For example, the false green rate is the false green probability divided by the sum of the true red probability and the false green probability. Why? Well, express each term as a ratio of numbers of boxes, which we saw how to do in Section TBD, then simplify.

A confusion table is a helpful way to represent these nine properties of a decision rule and the relations between them:

true red probability, a	false green probability, b	false green rate, $\frac{b}{a+b}$
false red probability, c	true green probability, d	false red rate, $\frac{c}{c+d}$
red prediction error, $\frac{c}{a+c}$	green prediction error, $\frac{b}{b+d}$	overall procedure error, $b+c$

The core of the confusion table is the true/false red/green probabilities, abbreviated as a, b, c, d. The five other entries in the table are determined by these.

For example, take the crate in Figure TBD again. Here is the confusion table for the decision rule: if '1' predict red and if '2' predict green. (Figures are to three decimal places.)

0.286	0.214	0.429
0.214	0.286	0.429
0.429	0.429	0.429

For each of these properties, we can restrict attention to blue cube boxes and yellow cube boxes. Everything is conditionalized on drawing a blue (yellow) cube box.

Confusion table for blue cube boxes:

0.286	0.143	0.333
0.286	0.286	0.500
0.500	0.333	0.429

Confusion table for yellow cube boxes:

0.286	0.286	0.500
0.143	0.286	0.333
0.333	0.500	0.429

Things you might want: equal red prediction errors across groups, equals green prediction errors across groups, equal false green rates across groups, etc.

Why you might want these things.

5.3 Properties of decision rules with randomization

Instead of thinking about decision rules, we could think about decision rules with randomization. Everything still makes sense. Nothing conceptually new. The calculations become only a little harder.

Show how calculations work in general.

Example. Crate as in Figure TBD. Decision rule with randomization: if '1', red with prob .6, green with prob .4; if '2', red with prob .2, green with prob .8. We get overall confusion table (left), for blue cube boxes (middle), for yellow cube boxes (right):

0.214	0.286	0.571	0.200	0.229	0.533	0.229	0.343	0.600
0.186	0.314	0.371	0.229	0.343	0.400	0.143	0.286	0.333
0.464	0.476	0.471	0.533	0.400	0.457	0.385	0.545	0.486

5.4 Properties of risk assignments with or without randomization

Now let's think about risk assignments.

It no longer makes sense to ask, What's the probability of predicting incorrectly?

But other properties make sense. Calibration, Balance$^+$ and Balance$^-$. These properties are to do with accuracy, rather than truth. Accuracy is to probabilistic predictions what truth is to categorical predictions.

Calibration says, roughly speaking, that the probabilities in the predictions match the actual probabilities for each prediction:

More carefully:

$$\text{P(B=r} \mid \text{X=x)} = \text{x}$$

In words: for each probabilistic prediction 'red with probability X%', the probability that the box contains a red ball, given that you make that probabilistic prediction, equals X%.

Relativize across groups, as before.

Balance for the positive class, or Balance$^+$, says, roughly speaking, that probabilistic predictions aren't systematically less accurate for boxes containing red balls in one group than the other. Balance for the negative class, or Balance$^-$ is similar:

Balance$^+$ and Balance$^-$ are ways of lifting the idea of equal prediction errors from decision rules to risk assignments.

6 Nice situations

For some crates, things work out very nicely. There exist strategies with all the properties we could want.

Example 1: Equal base rates.

Example 2: Admitting perfect prediction.

7 Results

Remember the form of the results:

> No strategy of kind *so-and-so* has *such-and-such* properties (unless the contents of the crate happens to be like *this-and-that*).

We've seen various ways to fill in the place-holders. We've seen four different kinds of strategy, a lot of properties of strategies, and some nice crates where things go well.

So we can now state a few results, by filling in different things into the place-holders.

Example 1:

> No decision rule has equal red and green prediction errors across groups, unless the crate happens to have equal base rate or admit perfect prediction.

8 Do the results show that no strategy is fair?

Short answer: don't think so.

Reasons for and against, expanding on handout.

References