PHIL 408Q/PHPE 308D
Fairness

Eric Pacuit, University of Maryland

April 30, 2024

Benjamin Eva (2022). *Algorithmic Fairness and Base Rate Tracking*. Philosophy & Public Affairs, 50(2), pp. 239 - 266.

Base Rate Tracking

If an algorithm assigns one group a higher average risk score than another, that discrepancy has to be justified by a corresponding discrepancy between the base rates of those two groups, and the magnitudes of those discrepancies should be equivalent.

Base Rate Tracking

In slogan form: an algorithm should only treat one groups as much more risky than another if it really is much more risky.

Base Rate Tracking: The difference between the average risk scores assigned to the relevant groups should be equal to the difference between the (expected) base rates of those groups.

Base Rate Tracking...

Base Rate Tracking...

1. ...is motivated by a simple and powerful philosophical intuition about the nature of fairness,

Base Rate Tracking...

1. ...is motivated by a simple and powerful philosophical intuition about the nature of fairness,

2. ...is not undermined by Hedden's coin flipping example or the insurance pricing example, and

Base Rate Tracking...

1. ...is motivated by a simple and powerful philosophical intuition about the nature of fairness,

2. ...is not undermined by Hedden's coin flipping example or the insurance pricing example, and

3. ...significantly expands the diagnostic scope of calibration within groups in some important cases.

Binary Base Rate Tracking

One could naturally try to construct an analogue of base rate tracking for binary classification algorithms:

The difference between the percentage of members of each relevant group that are classed as "positive" should be equal to the (expected) difference between the base rates of those groups.

Binary Base Rate Tracking

To illustrate: binary base rate tracking says that it is unfair for a binary classification algorithm to classify 50 percent of loan applicants from Group 1 as "high risk" while classing only 30 percent of applicants from Group 2 as "high risk" if it is not the case that the (expected) percentage of Group 1 applicants who actually default is not exactly 20 percent greater than the percentage of Group 2 applicants who actually default.

Binary Base Rate Tracking

To illustrate: binary base rate tracking says that it is unfair for a binary classification algorithm to classify 50 percent of loan applicants from Group 1 as "high risk" while classing only 30 percent of applicants from Group 2 as "high risk" if it is not the case that the (expected) percentage of Group 1 applicants who actually default is not exactly 20 percent greater than the percentage of Group 2 applicants who actually default.

Problem: While binary base rate tracking seems to be motivated by the same compelling motivation as standard base rate tracking, it is easy to see that it is actually prone to powerful counterexamples to which the original formulation is immune.

Binary Base Rate Tracking

Suppose that 20 people are split evenly between two rooms, A and B. The A-people are all assigned coins with bias 0.6 and the B-people are assigned coins with bias 0.4.

A binary classification algorithm predicts whether people's coins will land heads when tossed on the basis of their coin's bias. If the bias is 0.6, it predicts that the coin will land heads, and if the bias is 0.4, it predicts that it will land tails.

Then the algorithm will predict that all A-people will toss heads, and that no B-people will toss heads, which seems perfectly fair.

Binary Base Rate Tracking

Suppose that 20 people are split evenly between two rooms, A and B. The A-people are all assigned coins with bias 0.6 and the B-people are assigned coins with bias 0.4.

A binary classification algorithm predicts whether people's coins will land heads when tossed on the basis of their coin's bias. If the bias is 0.6, it predicts that the coin will land heads, and if the bias is 0.4, it predicts that it will land tails.

Then the algorithm will predict that all A-people will toss heads, and that no B-people will toss heads, which seems perfectly fair.

But the difference in the base rates of the two groups is only 20 percent, which is five times less than the 100 percent difference between the percentages of each population that are predicted to toss heads by the algorithm.

Binary Base Rate Tracking

This example illustrates that there is no obvious and plausible analogue of base rate tracking for binary classification algorithms. As it stands, base rate tracking can only be legitimately applied as a necessary condition for the fairness of risk scoring algorithms.

Approximate Fairness

Base Rate Tracking is intended to act as a necessary condition for an algorithm to count as perfectly fair.

Approximate Fairness

Base Rate Tracking is intended to act as a necessary condition for an algorithm to count as perfectly fair.

In practice, few real algorithms will fully satisfy this criterion. However, we can still use the criterion to assess the scale and significance of an algorithm's unfairness by evaluating how far away it is from satisfying base rate tracking.

Approximate Fairness

Base Rate Tracking is intended to act as a necessary condition for an algorithm to count as perfectly fair.

In practice, few real algorithms will fully satisfy this criterion. However, we can still use the criterion to assess the scale and significance of an algorithm's unfairness by evaluating how far away it is from satisfying base rate tracking.

If the difference between the average risk scores is far greater than the difference between the base rates, then the algorithm is very unfair, but if the divergence between those quantities is small, then the unfairness may be slight. As with any evaluative standard, perfection is a rare exception at best, and the fact that the standard is rarely fully satisfied does not undermine its claim to normative significance.

Approximate Fairness

Of course, one might think that the notion of "perfect fairness" is a red herring here, and claim that all we ever have are pragmatically determined standards of what counts as "fair enough."

Approximate Fairness

Of course, one might think that the notion of "perfect fairness" is a red herring here, and claim that all we ever have are pragmatically determined standards of what counts as "fair enough."

When we are dealing with judgments that have life or death outcomes, the standard is much higher than when we are dealing with judgments that, at worst, lead to minor inconveniences for those affected.

Approximate Fairness

Of course, one might think that the notion of "perfect fairness" is a red herring here, and claim that all we ever have are pragmatically determined standards of what counts as "fair enough."

When we are dealing with judgments that have life or death outcomes, the standard is much higher than when we are dealing with judgments that, at worst, lead to minor inconveniences for those affected.

If one prefers to eschew the general ideal of perfect fairness and focus rather on context-dependent notions of sufficient fairness, then one can interpret my arguments as supporting the idea that in order for an algorithm to be "fair enough" in a given context, the divergence between the base rates and the average risk scores should not be "too great," where what counts as "too great," is determined by a range of pragmatic contextual variables.

A Consequence of Base Rate Tracking

Note that as well as requiring that the average risk scores be equal when the base rates are, base rate tracking also requires the converse, i.e., that when the risk scores are equal, the base rates should be too.

A Consequence of Base Rate Tracking

Note that as well as requiring that the average risk scores be equal when the base rates are, base rate tracking also requires the converse, i.e., that when the risk scores are equal, the base rates should be too.

So as well as stipulating that a fair algorithm only treats groups differently when there is a suitable difference in their base rates, base rate tracking also requires that groups should only be treated similarly to the extent that their base rates are similar.

A Consequence of Base Rate Tracking

Note that as well as requiring that the average risk scores be equal when the base rates are, base rate tracking also requires the converse, i.e., that when the risk scores are equal, the base rates should be too.

So as well as stipulating that a fair algorithm only treats groups differently when there is a suitable difference in their base rates, base rate tracking also requires that groups should only be treated similarly to the extent that their base rates are similar.

This is motivated by a natural intuition: that it would be unfair to treat two groups as equally risky if one was in fact more risky than another.

A Consequence of Base Rate Tracking

Note that as well as requiring that the average risk scores be equal when the base rates are, base rate tracking also requires the converse, i.e., that when the risk scores are equal, the base rates should be too.

So as well as stipulating that a fair algorithm only treats groups differently when there is a suitable difference in their base rates, base rate tracking also requires that groups should only be treated similarly to the extent that their base rates are similar.

This is motivated by a natural intuition: that it would be unfair to treat two groups as equally risky if one was in fact more risky than another.

Recalling the correlation between gender and recidivism, an algorithm would seem to be unfair if it assigned males and females similar risk scores even though females had a significantly lower actual rate of recidivism.

A Possible Objection

Going back to Redlining 1, base rate tracking successfully identifies the fact that the algorithm is unfair to black applicants, because the difference between the average risk scores of white and black applicants is far greater than the difference between their base rates. However, base rate tracking still requires that white applicants should be assigned a lower average risk score than black applicants, since black applicants have a higher overall default rate.

A Possible Objection

Going back to Redlining 1, base rate tracking successfully identifies the fact that the algorithm is unfair to black applicants, because the difference between the average risk scores of white and black applicants is far greater than the difference between their base rates. However, base rate tracking still requires that white applicants should be assigned a lower average risk score than black applicants, since black applicants have a higher overall default rate.

And one might plausibly object that this is obviously unfair, since black applicants have the same default rate as white applicants within any given zip code.

This in turn implies that base rate tracking is not a plausible statistical criterion of algorithmic fairness.

Response

If the algorithm was designed to disadvantage black applicants, or if the correlations upon which it relies are the product of unjust historical conditions, then those constitute independent sources of unfairness which need to be appropriately recognized and taken into account in the application of the algorithm.

Response

If the algorithm was designed to disadvantage black applicants, or if the correlations upon which it relies are the product of unjust historical conditions, then those constitute independent sources of unfairness which need to be appropriately recognized and taken into account in the application of the algorithm.

Of course, statistical criteria like base rate tracking are unable to directly diagnose these kinds of unfairness, since they concern the historical origins of the algorithm and the relevant correlations, rather than predictive properties of the algorithm itself.

One can recognize these sources of injustice without thinking that the algorithm and its predictions are themselves intrinsically unfair.

Sources of Unfairness

Base Rate Tracking measures the unfairness that is *intrinsic* to the algorithm, but there are other sources of unfairness of an algorithm:

Sources of Unfairness

Base Rate Tracking measures the unfairness that is *intrinsic* to the algorithm, but there are other sources of unfairness of an algorithm:

- Facts regarding the unjust historical conditions that gave rise to the correlations exploited by the algorithm
- Facts about the unjust intentions of the algorithm's designers.

Concluding Remarks

While statistical criteria like Base Rate Tracking can play an important role in the fight against algorithmic unfairness, the hardest problem will be to develop mechanisms that properly identify and compensate for the way in which algorithms exploit correlations which themselves arise from unfair historical conditions.

It is important that we recognize this problem as distinct from the problem of diagnosing unfairness, that is, intrinsic to the way that a given algorithm makes predictions, since the tools we use to address the latter problem (statistical criteria of algorithmic fairness) are not well suited to addressing the former.

Rush T. Stewart (2022). *Identity and the limits of fair assessment*. Journal of Theoretical Politics 2022, Vol. 34(3), pp. 415 - 442.

Research on algorithmic fairness studies the prospects of unbiased assessment. Bias in error rates is one form of bias, but not the only form and often considered not the most important form. Can bias in error rates and other important forms of bias be simultaneously eliminated?

One lesson that emerges from some of these studies is that eliminating one form of bias can mean that it is impossible to eliminate another. Sometimes, then, we face a conflict between eliminating different forms of bias.

Research on algorithmic fairness studies the prospects of unbiased assessment. Bias in error rates is one form of bias, but not the only form and often considered not the most important form. Can bias in error rates and other important forms of bias be simultaneously eliminated?

One lesson that emerges from some of these studies is that eliminating one form of bias can mean that it is impossible to eliminate another. Sometimes, then, we face a conflict between eliminating different forms of bias.

Here, I argue that, not only do we face a conflict in eliminating different forms of bias, we also face a conflict in eliminating one form of bias across different groupings. Eliminating a certain form of bias across groups for one way of categorizing people in a population can mean that it is impossible to eliminate that form of bias across groups for another way of classifying them.

Consider once again the bias found against black people in the COMPAS data.

In that same Broward County data set, there is a similar amount of bias in error rates against women compared to men, as a companion piece in ProPublica makes clear (Angwin et al., 2016a).

Consider once again the bias found against black people in the COMPAS data.

In that same Broward County data set, there is a similar amount of bias in error rates against women compared to men, as a companion piece in ProPublica makes clear (Angwin et al., 2016a).

Bias against either group is ethically relevant. Satisfying certain central fairness constraints for a race partition does not imply that those constraints are satisfied for a gender partition. Still other partitions could be pertinent.

Consider once again the bias found against black people in the COMPAS data.

In that same Broward County data set, there is a similar amount of bias in error rates against women compared to men, as a companion piece in ProPublica makes clear (Angwin et al., 2016a).

Bias against either group is ethically relevant. Satisfying certain central fairness constraints for a race partition does not imply that those constraints are satisfied for a gender partition. Still other partitions could be pertinent.

The relevant social identities cannot be decided a priori, without appeal to contingent social context and values.

Consider, for example, those who wear a size 8 shoe, or those born between nine and ten in the morning, local time. If size 8 shoes were to become extremely difficult to find then being someone who wears that shoe size may become an important part of one's identity and grounds for solidarity with those similarly unshod.

Consider, for example, those who wear a size 8 shoe, or those born between nine and ten in the morning, local time. If size 8 shoes were to become extremely difficult to find then being someone who wears that shoe size may become an important part of one's identity and grounds for solidarity with those similarly unshod.

Likewise, if an authoritarian ruler were to elect to severely curtail the freedoms of people born between nine and ten in the morning due to some supernatural belief or other, then the hour of one's birth and the persecution it entails for some is, again, likely to become an important aspect of one's identity and grounds for solidarity.

The priority of particular partitions in eliminating bias might reasonably depend not just on past history of discrimination, but also on current deprivation.

What groups suffer discrimination and deprivation is a matter to which we may frequently need to reattend.

Setup

A single property y of interest.

Individuals in N either have property y or lack it: $Y : N \to \{0, 1\}$

Call a function $h : N \to [0, 1]$ an assessor.

For concreteness, interpret $h(i)$ as the assessor's probability that i has property y.

Setup

The quantity $P(Y = 1) = \mu$, for example, is the proportion of people in N that have property y, the prevalence of y in the population.

Call μ the base rate for y in N.

Given a partition $\pi = \{G_1, \ldots, G_m\}$ of N, let $P_k = P(\cdot \mid G_k)$. So, $P_1(Y = 1) = \mu_1$ is the base rate for y in group 1 is μ_1 and $P_2(h = 0.5)$ is the proportion of people to which h assigns 0.5 in G_2, and so on.

Strong Calibration

An assessor is **(strongly) calibrated** if

$$P_k(Y = 1 \mid h = p) = p \text{ for all } p \in [0, 1] \text{ and } k = 1, 2, \ldots, m \text{ such that } P_k(h = p) > 0.$$

Strong Calibration

An assessor is **(strongly) calibrated** if

$$P_k(Y = 1 \mid h = p) = p \text{ for all } p \in [0, 1] \text{ and } k = 1, 2, \ldots, m \text{ such that } P_k(h = p) > 0.$$

E.g., consider weather forecasting. Suppose that each day, a forecaster announces a probability of rain for that day. The forecaster is calibrated if it rains on 10% of the days she announces that it will rain with probability 0.1, and it rains on 85% of the days she predicts rain with probability 0.85, etc.