Fairness

Jeff Edmonds *
York University
jeff@cse.yorku.ca

Karan Singh
York University
singkara@cse.yorku.ca

Ruth Urner
York University
ruth@cse.yorku.ca

November 9, 2020

1 High School

Consider some person. Let t indicate their inner math ability. Let this be randomly chosen from some fixed distribution with $Pr[t] = T(t)$. Let $i = 0$ indicate that they are disadvantaged and $i = 1$ advantaged. Let $G = G(t, i)$ give their expected high school grade. This can be any function that is strictly increasing in both t and i. In particular, $\forall t\ G(t, 0) < G(t, 1)$. However, this grading process has some error/noise in approximating the ability t. We choose to model this be having the error added to t, namely let $t' = t + Error$ and $G = Grade(t', i)$.

Here $Error$ is our random error variable. Let $p = e(u)$ give that $Pr[Error = u] = p$. We require $e(u)$ to be strictly "sub-exponential" in a range $[a, b]$ and zero outside this range. Wolfram alpha shows that it includes Gaussians, quadratics, ?polynomials?, uniform, ??? It might as well have mean zero.

Lets suppose that the resulting grade is known to be $G = g$. This gives us that the measured ability is $t_i = t' = Grade^{-1}(g, i)$. Note that for a disadvantaged person ($i = 0$) to get the same grade g as an advantaged person ($i = 1$), they need to have a higher math abilities, i.e. $t_0 > t_1$. Because $t_i = t' = t + Error$, we have that $t = t_i - Error \in [t_i - a, t_i - b]$. Denote this range as $[a_i, b_i]$.

Claim: The employer wants to know i, but because he would rather hire someone with $i = 0$.

We will simplify the notation by having $[g\&0]$ be short hand for the condition that $[G = g\ \&\ i = 0]$.
Let c denote the threshold for the employer to hire the person.

Theorem 1

- *If the distribution $T(t)$ is uniform within the conditioned range $[a_1, b_0]$, then*
  - $Exp[t|g\&0] = t_0 > t_1 = Exp[t|g\&1]$.

*Thanks to Frances for her encouragement.

- *If $T(t)$ graceful and the error range $[a_1, b_0]$ is not too big, then it is reasonable to approximate $T(t)$ to be linear within $[a_1, b_0]$. In this case,*
  - *????* $Exp[t|g\&0] - Exp[t|g\&1] = t_0 - t_1$.
  - *????* $\forall c$, $Pr[t \geq c|g\&0] = Pr[t \geq c - t_0 + t_1|g\&1]$.
- *For worst case $T[t]$,*
  - $Exp[t|g\&0] > Exp[t|g\&1]$.
  - $\forall c$, $Pr[t \geq c|g\&0] > Pr[t \geq c|g\&1]$, *or both are zero, or both are one. This is tight for a specially chosen distribution on t and (mostly) uniform $Error$.*

Proof of Theorem 1: ($T(t) = wt + v$) By definition, $Pr[t \geq c|g\&i] = \frac{Pr[t \geq c\&g\&i]}{Pr[g\&i]}$. Fix some error value $u \in [a, b]$. It has probability $E(u)\Delta u$. Because $t_i = t' = t + Error$, we have that $t = t_i - u$. The probability of the person having this ability is $T(t_i - u)\delta t$. This gives that $Pr[g\&i] = \int_{u \in [a,b]} T(t_i - u)E(u)\Delta u \delta t$ and that
$Pr[t \geq c\&g\&i] = \int_{u \in [c',b]} T(t_i - u)E(u)\Delta u \delta t$
$= \int_{u \in [c',b]} [w(t_i - u) + v]E(u)\Delta u \delta t$
$= (wt_i + v) \times \int_{u \in [c',b]} E(u)\Delta u \delta t + w \times \int_{u \in [c',b]} uE(u)\Delta u \delta t$
$= (wt_i + v) \times m + w \times n = w't_i + v'$.
$Pr[t \geq c\&g\&i] = (wt_i + v) \times 1 + w \times n' = wt_i + v''$.
$Pr[t \geq c|g\&i] = \frac{w't_i + v'}{wt_i + v''}$
$= \frac{w'}{w} + \frac{w''}{wt_i + v''}$. ■

Proof of Theorem 1: (Worst case $T(t)$) Because $t_1 < t_0$, there are two cases. In the first case, this t interval is disjoint for $i = 0$ and $i = 1$, namely $a_1 \leq b_1 \leq a_0 \leq b_0$. It is trivial to prove the result for this case. Hence, lets assume the second case, namely $a_1 \leq a_0 \leq b_1 \leq b_0$. Note that abilities outside the range $t \in [a_1, b_0]$ are simply not possible given $G = g$ and hence will be ignored. Abilities in the range $t \in [a_1, a_0]$ only help prove our result because they are low and are possible for $i = 1$ but not for $i = 0$. The worst case for the theorem is when the distribution on t simply does not allow such t to arise. Similarly for $t \in [b_1, b_0]$. Hence, without loss of generality, lets assume that the distribution on t has range $t \in [a_0, b_1]$. If our threshold c is outside this range, then both probabilities in the theorem will either be zero or one. Hence, lets assume $c \in (a_0, b_1)$.

The theorem states that the result is tight. We know that $Pr[t|0] = Pr[t|1]$ for $t \in (a_0, b_1)$. If uniform $Error$ in the range $[a, b]$, then $Pr[t \geq c|g\&0] = Pr[t \geq c|t \in [a_0, b_1]] = Pr[t \geq c|g\&1]$

By definition, $Pr[t \geq c|g\&i] = \frac{Pr[t \geq c\&g\&i]}{Pr[g\&i]}$. Fix some abilities value t. Let $T(t)\delta t$ denote the probability of the person having ability t. Because $t_i = t' = t + Error$, we have that $Error = t_i - t$. Define $E_i(t) = E(t_i - t)$. This gives $Pr[Error = t_i - t] = E_i(t)\delta t$. This gives that $Pr[g\&i] = \int_t T(t)E_i(t)\delta t \delta t$ and that $Pr[t \geq c\&g\&i] = \int_{t \geq c} T(t)E_i(t)\delta t \delta t$. Given a function $r(t)$, define $F(r) = \frac{\int_{t \geq c} r(t)\delta t \delta t}{\int_t r(t)\delta t \delta t}$ to be the fraction of the area under the curve $r(t)$ that is to the right of the $t = c$. This gives that $Pr[t \geq c|g\&i] = F(TE_i)$.

Let $h(t) = \frac{E_0(t)}{E_1(t)}$, let c' be the constant $h(c)$, and let $h'(t) = h(t)/c'$. Lemma 1 proves $h(t)$ is strictly increasing. And hence, $\forall t \in [a_0, c)$ we have $h'(t) < 1$ and $\forall t \in (c, b_1]$ we have

$h'(t) > 1$. Hence, multiplying $r(t)$ by $h'(t)$ decreases $r(t)$ for those t before c and increase those after. It follows that the fraction of the area under the curve $r(t)$ that is right of the $t{=}c$ increases, i.e. $F(r) < F(h'r)$. Similarly, $F(r) = F(c'r)$.

The result follows $Pr[t \geq c|g\&1] = F(TE_1) = F(Tc'E_1) \leq F(Tc'E_1h') = F(TE_0) = Pr[t{\geq}c|g\&0]$. ■

Lemma 1 *Suppose that $E(u) > 0$ and is "sub exponential" in a range $[a, b]$ and zero outside this range. Let $f(t) = E(t_0 - t)$ and $g(t) = E(t_1 - t)$ for $t_1 < t_0$. Let $h(t) = \frac{f(t)}{g(t)}$. Then $h(t)$ is strictly increasing.*

Proof of Lemma 1: $h'(t) = \frac{f'(t)g(t) - f(t)g'(t)}{g^2(t)}$. Hence, to prove that $h'(t) > 0$, it is sufficient to prove that $\frac{g(t)}{g'(t)} > \frac{f(t)}{f'(t)}$. By the definitions of f and g, it is sufficient to prove that $\frac{E(t_1-t)}{E'(t_1-t)} > \frac{E(t_0-t)}{E'(t_0-t)}$. Because $t_1 < t_0$, it is sufficient to prove that $\frac{E(u)}{E'(u)}$ is strictly decreasing.

This can be visualized as follows. In graph of $E(u)$, draw the line though the point $\langle u, E(u)\rangle$ tangent to the curve. Let u_0 be the value of u at where the line crosses the u-axis and $\Delta u(u) = u - u_0$. Note that $E'(u) = \frac{E(u)}{\Delta u(u)}$ and hence $\frac{E(u)}{E'(u)} = \Delta u(u)$. Hence, it is sufficient to prove that $\Delta u(u)$ is strictly decreasing. Now visualize watching $\Delta u(u)$ change as u increases. Note how for concave functions it seem (though not totally clear) that Δu decreases.

If $r(u) = \frac{E(u)}{E'(u)}$, then $r'(u) = \frac{(E'(u))^2 - E(u)E''(u)}{(E'(u))^2} = 1 - \frac{E(u)E''(u)}{(E'(u))^2}$ To prove that $r'(t) < 0$, it is sufficient to prove that $E''(u) < \frac{(E'(u))^2}{E(u)}$. Suppose this was tight. Then if you tell me $E(0)$ and $E'(0)$, then the whole function is determined. By breaking the domain u into ϵ sized pieces, u effectively becomes an integer and hence we can do induction on it. By way of induction, assume that we have determined $E(u)$ and $E'(u)$. From this we determine $E''(u) = \frac{(E'(u))^2}{E(u)}$, $E(u{+}\epsilon) = E(u) + \epsilon E'(u)$, and $E'(u{+}\epsilon) = E'(u) + \epsilon E''(u)$. Wolfram alpha gives that $E(u) = c_1 e^{c_2 u}$. Certainly, $h(t) = \frac{f(t)}{g(t)} = \frac{c_1 e^{c_2(t_0-t)}}{c_1 e^{c_2(t_1-t)}} = e^{c_2(t_0-t_1)}$ is not strictly increasing but constant as we could expect from things being made tight.

Here is a leap, if the solution to $E''(u) = \frac{(E'(u))^2}{E(u)}$ is $E(u) = c_1 e^{c_2 u}$ and for $E(u) = c_1 e^{c_2 u}$ we have that $E''(u) = c'E(u)$, then does it follow that $E''(u) < \frac{(E'(u))^2}{E(u)}$ can be simplified to $E''(u) < c'E(u)$?

Any way, my intuition is that this poses very few restriction on $E(u)$ because it should include every function that ????grows slower than exponential???? ■

Lemma 2 *If $\forall c$, $Pr[t{\geq}c|g\&0] > Pr[t{\geq}c|g\&1]$, then $Exp[t|g\&0] > Exp[t|g\&1]$.*

Proof of Lemma 2: $Exp[t|g\&0] = \int_{t\geq 0} Pr[t|g\&0]t\delta t = \int_{t\geq 0}\int_{c\in[0,t]} Pr[t|g\&0]\delta c\delta t = \int_{c\geq 0}\int_{t\geq c} Pr[t|g\&0]\delta t\delta c = \int_{c\geq 0} Pr[t \geq c|g\&0]\delta c > \int_{c\geq 0} Pr[t \geq c|g\&1]\delta c = Exp[t|g\&1]$. ■

2 New Problem

Consider some person.
Let $i = 0$ indicate that they are disadvantaged and $i = 1$ advantaged.
Let t indicate their inner abilities. Let this be randomly chosen from some distribution $\mathcal{D}$. Note that it does not depend on i.
Let $G_{highschool}(t, i)$ give their expected high school grade. This can be any function that is strictly increasing in both t and i. In particular, $\forall t,\ G_{highschool}(t, 0) < G_{highschool}(t, 1)$.
Let $G_{university}(t)$ give their expected university grade, again any increasing function of t. Note that it does not depend on i.
Let $Error_{highschool}$ and $Error_{university}$ be any independent error random variables. They might as well have mean zero.
What is the minimum requirement for these error functions for the proof to go through?
Let $g_{highschool} = G_{highschool}(t, i) + Error_{highschool}$ be their actual high school grade.
Let $g_{university} = G_{university}(t) + Error_{university}$ be their actual university grade.
Let $\mu_{highschool}$ be the threshold of $g_{highschool}$ for accepting the person into university.
Let $c_{employer}$ be the threshold of t that is acceptable for the employer.
The proof will be easier if you assume each possible value of t, G, and $Error$ is rounded down to the nearest integer multiple of ϵ because then $Pr[g = g']$ will be non-zero and you wont have to stick integrations everywhere.
Claim: The employer wants to know $g_{highschool}$ and i, but because he would rather hire someone with $i = 0$.
$\forall g'_{highschool} \geq \mu_{highschool}$, $\forall g'_{university}$,

$$
\begin{aligned}
&Pr[t \geq c_{employer} \mid \\
&\qquad g_{highschool} = g'_{highschool} \\
&\qquad g_{university} = g'_{university} \\
&\qquad \text{and } i = 0] \\
&> Pr[t \geq c_{employer} \mid \\
&\qquad g_{highschool} = g'_{highschool} \\
&\qquad g_{university} = g'_{university} \\
&\qquad \text{and } i = \text{unknown }] \\
&> Pr[t \geq c_{employer} \mid \\
&\qquad g_{highschool} = g'_{highschool} \\
&\qquad g_{university} = g'_{university} \\
&\qquad \text{and } i = 1]
\end{aligned}
$$

$$
\begin{aligned}
&Pr[t \geq c_{employer} \mid \\
&\qquad g_{highschool} \geq \mu_{highschool} \\
&\qquad g_{university} = g'_{university} \\
&\qquad \text{and } i = 0] \\
&> Pr[t \geq c_{employer} \mid \\
&\qquad g_{highschool} \geq \mu_{highschool} \\
&\qquad g_{university} = g'_{university}
\end{aligned}
$$

$$
\begin{aligned}
&\text{and } i = \text{unknown }] \\
> Pr[t \geq c_{employer} \mid & \\
& g_{highschool} \geq \mu_{highschool} \\
& g_{university} = g'_{university} \\
& \text{and } i = 1]
\end{aligned}
$$

3 Rewarding Improvement

Fairness in machine learning is topic that is growing in importance. Affirmative action is a way of giving a step up to a disadvantaged group in order to improve the group as a whole's situation in the long run. In contrast, we propose rewarding individuals from disadvantaged pasts who have managed to "make it" in their present despite these disadvantages, **because** we feel that this is an good indication that they may also "make it" in their future.

X = **Past and Group Membership:** Let X represent the data we know about an individual about their past, eg their race, gender, neighborhoods, handicaps, parents, ...
Note that though we call X a "group", in practice X could have many fields to such an extent that each individual is effectively in their own group.

Y = **Present Accomplishments:** Let Y denote how well they are doing now, eg high school or SAT scores, jobs, clubs, or letters of reference about behavior from school/jail.

Z = **Future Accomplishments:** Let Z denote what we are trying to predict about the person so that we can make some decision about them. For example, we may be deciding whether to accept them to our University, let them out of jail, or give them a mortgage, while Z denotes how well they will succeed if we say yes.

Unfair: The first thing to try is to use machine learning to learn a predictor $\widehat{Z}$ that takes $\langle X, Y \rangle$ as input and outputs $\widehat{Z} = \widehat{Z}(X, Y)$ as a prediction of Z. The concern is that this is potentially unfair as the predictor might make $\widehat{Z}(X, Y)$ lower when X is disadvantaged.

Human Interaction: One way to fix this is for humans to identify which X are disadvantaged and to work into the machine learning policies to benefit these people.

Information Hiding: One way to fix this is to not tell the machine learner this group information X, making the prediction $\widehat{Z}(Y)$. This may still be unfair if the learner can somehow infer X from Y and then use this to be prejudiced against the disadvantaged group X when deciding Z.

Insult to Injury: We argue that simply dropping the group information X when learning $\widehat{Z}(Y)$ is also unfair because there really are ways in which being in such a disadvantaged group X "causes" one to do worse in the present situation Y. Then holding it against them that they did poorly in Y by giving them a low prediction $\widehat{Z}(Y)$ is adding insult to injury

Rewarding Improvement: We argue that if someone managed to do ok in Y even though they were disadvantaged in X, then we should reward them by giving them a better $\widehat{Z}(X, Y)$. One might formalize this is by stating that if $X_1 << X_2$ and $Y_1 \approx Y_2$ then $\widehat{Z}(X_1, Y_1)$ should "fairly" be bigger than $\widehat{Z}(X_2, Y_2)$.

Predicting Y: We don't want to rely on human bias to decide which groups X are disadvantaged. Lets use our best machine learning practices to predict one's present Y

from one's past X using some predictor $\widehat{Y}(X)$. Then if a group X has a particularly small predicted present $\widehat{Y}(X)$, then we will use this as the definition of X being a *disadvantaged* group X. Note that $\widehat{Y}(X)$ speaks about the group X and not about individuals in the group. Though we did say that X could be unique to an individual, in which case, this is saying that this individual has features of their past which are considered to be disadvantaged.

Improvement: If an individual is from a disadvantaged group, then we predict their present to be bad, i.e. $\widehat{Y}(X)$ is small. But if despite this, the individual is presently doing well, i.e. Y is big, then this individual is particularly bright, resilient, and hard working. This should be rewarded. Define $Improvement(X, Y)$ to be $Y - \widehat{Y}(X)$. Note that this is positive if the individual is presently better than average in his group X and is negative if worse. Note that here we are assuming Y is a real value. This will have to be fudged if Y is a vector.

Formally Rewarding Improvement: We recommend defining the individual's *adjusted present* to be $Y' = Y + \alpha \cdot Improvement(X, Y) = Y + \alpha(Y - \widehat{Y}(X))$.

Predicting Future Z: The second step is to then forget the group X as recommended before for sake of fairness, but now to use the adjusted present Y' as a prediction $\widehat{Z}(Y')$ of Z instead of the actual present Y. Our hypothesis that we want to examine in this paper is that this will be a better predictor, i.e. $\widehat{Z}(Y')$ is closer to Z than $\widehat{Z}(Y)$. The intuition goes as follows. If $Y >> \widehat{Y}(X)$, then this individual is sufficiently bright, resilient, and hard working to over come their disadvantage X. These same qualities will help this person do well in their future Z.

Oh, if Y is a vector but Z is a real number, maybe it is better to define $\widehat{Z}(X, Y)$ to be $\widehat{Z}(Y) + \alpha \cdot Improvement(X, Y) = \widehat{Z}(Y) + \alpha(Y - \widehat{Y}(X))$.

Individual Fairness: The claim is that this is NOT *affirmative action* geared to help individuals from the disadvantaged group X, but is actually strongly motivated by wanting to choose the best candidates by accurately predicting Z.

Reverse Discrimination: Of course if $Improvement(X, Y) = Y - \widehat{Y}(X)$ is negative, then this individual is doing worse than expected given their group X. This new method would be a disservice to this person. For example, an upper class white male with poor marks will be disadvantaged.

Group Fairness: Group fairness is a method of being kinder to disadvantaged groups X when predicting Z. Inadvertently, we are doing this. If the group X is disadvantaged then $\widehat{Y}(X)$ by definition will be small. Then because of the negative sign, $Y' = Y + \alpha(Y - \widehat{Y}(X))$ will increase as the advantage of X decreases. As such our prediction $\widehat{Z}(Y')$ will also increase.

Contradiction: We must confess to the obvious contradiction. Note that $\widehat{Z}(Y')$ is a function of Y' and Y' is a function of $\langle X, Y \rangle$. Hence, if our very first predictor $\widehat{Z}(X, Y)$ was trained well on this data then it should do at least as good of a job as $\widehat{Z}(Y')$ in

predicting Z. Ok, maybe our hypothesis is wishful thinking. However, I have a button that says "Having abandoned my search for truth, I am looking for a good fantasy."

```
X=T+E
Y= (X-Exp(E))  = (T+E-Exp(E))
(Y | T=t)  is norm with mean t and variance var(E)

================
If var(EA)=var(EB)

For every Threshold_A, exits Threshold_B
   BY = 1  iff  Y > Threshold
      for all  t,   Pr( Y | T=t and A) = Pr( Y | T=t and B)

If my t is small then I want my variance to be big.
so that I increase prob of getting in when I should not.
If my t is big then I want my variance to be small.
so that I decrease prob of not getting in when I should.

If my variance is big., then I am happy when my t is small
because variiane increases prob of getting in when I should not.
So I am less happy when you raise the threshold.

If my variance is big., then I am unhappy when my t is big
because variiane increases prob of not getting in when I should.
So I am more happy when you lower the threshold.


====================
[E-Exp(E)]/SD(E)  is   Norm(0,1)
Y= (X-Exp(E))/SD(E)  = (T+E-Exp(E))/SD(E)
(Y | T=t)  is norm with mean t/SD(E) and variance 1

BY = 1  iff  Y > Threshold
  iff c Y > c Threshold

====================
Assume Var(EA) = Var(EB) + dV

If B, then change X to X' = X + Norm(0,dV)


[E-Exp(E)]/SD(E)  is   Norm(0,1)
Y= (X-Exp(E))/SD(E)  = (T+E-Exp(E))/SD(E)
(Y | T=t)  is norm with mean t/SD(E) and variance 1
```

```
BY = 1  iff  Y > Threshold
  iff c Y > c Threshold

================
X=T+E
Y= (X-Exp(E))  = (T+E-Exp(E))

Equal Opportunity
For every Threshold,
   Threshold_i =  ??(Threshold, var(Ei), Ti
   BY = 1  iff  Y > Threshold_i
Pr( Y | T>Threshold and A) = Pr( Y | T>Threshold and B)

Demographic Parity
         Exp( Y | A) = Exp( Y | B)
Does not care how what your T is .
```