Exploring Impossibility Results for Algorithmic Fairness Using PrSAT

Tina Eliassi-Rad & Branden Fitelson
March 27, 2021

PrSAT — A Decision Procedure for Probability Calculus

PrSAT is a decision procedure for probability calculus that has been implemented in *Mathematica* (it's been tested on versions of *Mathematica* up to v12.3). See Fitelson (2008) for details.

The package is self-contained, and can be downloaded from the following website (which also includes instructions for installation and use).

http://fitelson.org/PrSAT/

We begin by loading the **PrSAT** package (which defines all the *Mathematica* functions we'll be using).

In[1]:=
```
<< PrSAT`
```

Notation, Fairness Measures, and Side Conditions/Auxiliary Assumptions

We will be discussing binary classification. Our binary classifier C can take two values: $C = 1$ or $C = 0$. We will denote these *predicted* values as C and $\neg C$, respectively. The *actual* value of the parameter in question will either take the value $\mathcal{Y} = 1$ or $\mathcal{Y} = 0$, and we will denote these two possibilities as $\mathcal{Y}$ and $\neg\mathcal{Y}$, respectively.

Initially, we will be looking at a single, (binary) protected attribute $\mathcal{A}$, which can take either the value $\mathcal{A} = 1$ or the value $\mathcal{A} = 0$, and we will denote these two possibilities as $\mathcal{A}$ and $\neg\mathcal{A}$, respectively. [In a later section, we will generalize this to the case of *two populations* ($\mathcal{A}$ and $\mathcal{B}$), which need not be mutually exclusive or exhaustive — thus allowing for intersectionality effects, *etc*.]

Assuming these notational conventions, we can define the following four traditional, confusion-matrix-based measures of algorithmic fairness (expressed in pure probability calculus), as follows:

In[2]:=
```
PredictiveParity = Pr[𝒴 | C ∧ 𝒜] == Pr[𝒴 | C ∧ ¬ 𝒜];
TruePositiveParity = Pr[C | 𝒴 ∧ 𝒜] == Pr[C | 𝒴 ∧ ¬ 𝒜];
FalsePositiveParity = Pr[C | ¬ 𝒴 ∧ 𝒜] == Pr[C | ¬ 𝒴 ∧ ¬ 𝒜];
StatisticalParity = Pr[C | 𝒜] == Pr[C | ¬ 𝒜];
```

Terminological notes: Chouldechova (2017) refers to the conjunction of True Positive Parity & False Positive Parity as “Error Base Rate.” This conjunction is also sometimes referred to as “Equalized Odds” or “Positive Rate Parity.” We are keeping these separate, because True Positive Parity is also implicated in several other impossibility results discussed below. True Positive Parity is also known in the literature as “Equal Opportunity” (Mehrabi et al 2019). Predictive Parity is sometimes also referred to as “Positive Predictive Value Parity.”

We will also be working with the following three “side conditions” or “auxiliary assumptions”.

In[6]:=
```
UnequalBaseRates = Pr[𝒴 | 𝒜] ≠ Pr[𝒴 | ¬ 𝒜];
ImperfectClassifier =
  And @@ {Pr[C | ¬ 𝒴 ∧ 𝒜] ≠ 0, Pr[C | ¬ 𝒴 ∧ ¬ 𝒜] ≠ 0, Pr[C | 𝒴 ∧ 𝒜] ≠ 1, Pr[C | 𝒴 ∧ ¬ 𝒜] ≠ 1};
NonZeroPrecision = Pr[𝒴 | C ∧ 𝒜] ≠ 0 ∨ Pr[𝒴 | C ∧ ¬ 𝒜] ≠ 0;
```

Verifying Two Well-Known Impossibility Results with PrSAT

Impossibility #1: Chouldechova (2017)

Here is a simple **PrSAT** verification of Chouldechova's (2017) impossibility theorem, expressed in pure probability calculus using the above notation.

Impossibility #1. *There are no probability models satisfying all four of these fairness constraints*:

(i) Predictive Parity (i.e., Predictive**Parity**)
(ii) True Positive Parity (i.e., **TruePositiveParity**)
(iii) False Positive Parity (i.e., **FalsePositiveParity**)
(iv) Statistical Parity (i.e., **StatisticalParity**)

subject to the following side condition/auxiliary assumption:

(b) there are *unequal base rates* (of $\mathcal{Y}$) in the two populations $\mathcal{A}$ and $\neg\mathcal{A}$ (i.e., **UnequalBaseRates**).

In[9]:=
```
PrSAT[
 {
  PredictiveParity,
  TruePositiveParity,
  FalsePositiveParity,
  StatisticalParity,

  UnequalBaseRates
 },
 BypassSearch → True
]
```

Out[9]= { }

Moreover, *no proper subset of these five conditions is unsatisfiable* (*i.e.*, such that there is no probability assignment Pr(•) which makes all of the claims true), as the following calculations reveal.

In[10]:=
```
NonemptyProperSubsets[S_] := Drop[Drop[Subsets[S], 1], -1];
```

In[11]:=
```
ChouldechovaSubsets = NonemptyProperSubsets[
   {
    PredictiveParity,
    TruePositiveParity,
    FalsePositiveParity,
    StatisticalParity,

    UnequalBaseRates
   }
  ];
```

In[12]:=
```
Select[ChouldechovaSubsets, PrSAT[#, BypassSearch → True] === {} &]
```

Out[12]= { }

So, in this sense, this is the "*minimal*" Chouldechova-style impossibility theorem.

Impossibility #2: Kleinberg et al (2016)

Here is a simple **PrSAT** verification of Kleinberg et al's (2016) impossibility theorem, expressed in pure probability calculus, using the above notation.

Impossibility #2. *There are no probability models satisfying all three of these fairness constraints*:

(i) Predictive Parity (*i.e.*, Predictive**Parity**),
(ii) True Positive Parity (*i.e.*, **TruePositiveParity**), and
(iii) False Positive Parity (*i.e.*, **FalsePositiveParity**),

subject to the following *three* side conditions/auxiliary assumptions:

(a) there are *unequal base rates* (of $\mathcal{Y}$) in the two populations $\mathcal{A}$ and $\neg\mathcal{A}$ (*i.e.*, **UnequalBaseRates**),
(b) our classifier is *imperfect* (*i.e.*, **ImperfectClassifier**), and
(c) *either* $\Pr(\mathcal{Y} \mid C \wedge \mathcal{A}) \neq 0$ *or* $\Pr(\mathcal{Y} \mid C \wedge \neg\mathcal{A}) \neq 0$ (*i.e.*, NonZeroPrecision).

In[13]:=
```
PrSAT[
 {
  PredictiveParity,
  TruePositiveParity,
  FalsePositiveParity,

  UnequalBaseRates,
  ImperfectClassifier,
  NonZeroPrecision
 },
 BypassSearch → True
]
```

Out[13]= { }

To see that the side condition **NonZeroPrecision** is required for this impossibility result, we can use **PrSAT** to find a model of the remaining five conditions.

In[14]:=
```
model = PrSAT[
  {
   PredictiveParity,
   TruePositiveParity,
   FalsePositiveParity,

   UnequalBaseRates,
   ImperfectClassifier
  },
  BypassSearch → True
 ]
```

Out[14]= $\left\{\{\mathcal{A} \to \{a_2, a_5, a_6, a_8\}, C \to \{a_3, a_5, a_7, a_8\}, \mathcal{Y} \to \{a_4, a_6, a_7, a_8\}, \Omega \to \{a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8\}\}, \left\{a_1 \to \frac{17\,275}{94\,869}, a_2 \to \frac{13\,820}{94\,869}, a_3 \to \frac{6910}{31\,623}, a_4 \to \frac{19\,348}{94\,869}, a_5 \to \frac{5528}{31\,623}, a_6 \to \frac{56}{747}, a_7 \to 0, a_8 \to 0\right\}\right\}$

We can use the function **STT** to represent this model as a *stochastic truth-table*, in the sense of Fitelson (2008).

In[15]:=
```
STT[model]
```

Out[15]=

$\mathcal{A}$	$\mathcal{C}$	$\mathcal{Y}$	var	Pr
T	T	T	a_8	0
T	T	F	a_5	$\frac{5528}{31623}$
T	F	T	a_6	$\frac{56}{747}$
T	F	F	a_2	$\frac{13820}{94869}$
F	T	T	a_7	0
F	T	F	a_3	$\frac{6910}{31623}$
F	F	T	a_4	$\frac{19348}{94869}$
F	F	F	a_1	$\frac{17275}{94869}$

And, we can verify the correctness of this **model** using the function **EvaluateProbability**.

In[16]:=
```
EvaluateProbability[
 {
  PredictiveParity,
  TruePositiveParity,
  FalsePositiveParity,

  UnequalBaseRates,
  ImperfectClassifier
 },
 model]
```

Out[16]= {True, True, True, True, True}

Moreover, no proper subset of these six conditions is unsatisfiable (as the following calculations reveal).

In[17]:=
```
KleinbergSubsets = NonemptyProperSubsets[
   {
    PredictiveParity,
    TruePositiveParity,
    FalsePositiveParity,

    UnequalBaseRates,
    ImperfectClassifier,
    NonZeroPrecision
   }
  ];
```

In[18]:=
```
Select[KleinbergSubsets, PrSAT[#, BypassSearch → True] === {} &]
```

Out[18]= { }

So, in this sense, this is the "*minimal*" Kleinberg et al-style impossibility theorem.

A Simplification of Kleinberg et al (2016) Based on *Regularity*

If we assume that the prior probability function Pr(•) is *regular* (*i.e.*, that it only assigns extremal probability to non-contingent propositions), then we can obtain a simplification of the impossibility result of Kleinberg et al, involving the three fairness constraints.

(i) Predictive Parity (*i.e.*, Predictive`Parity`),
(ii) True Positive Parity (*i.e.*, `TruePositiveParity`), and
(iii) False Positive Parity (*i.e.*, `FalsePositiveParity`),

and the following *two* side conditions/auxiliary assumptions:

(a) there are *unequal base rates* (of $\mathcal{Y}$) in the two populations $\mathcal{A}$ and $\neg\mathcal{A}$ (*i.e.*, `UnequalBaseRates`), and
(b) Pr(•) is *regular* (note: this auxiliary assumption is a built-in option of the `PrSAT` function).

Note: regularity is slightly stronger than Kleinberg et al's two side-conditions. So, this is a slightly weaker impossibility result than Kleinberg et al. But, regularity is much simpler and easier to work with.

In[19]:=
```
PrSAT[
 {
  PredictiveParity,
  TruePositiveParity,
  FalsePositiveParity,

  UnequalBaseRates
 },
 BypassSearch → True, Probabilities → Regular
]
```

Out[19]= { }

Moreover, *assuming regularity, no proper subset of these four conditions is unsatisfiable* (as the following calculations reveal).

In[20]:=
```
KleinbergSubsets2 = NonemptyProperSubsets[
    {
     PredictiveParity,
     TruePositiveParity,
     FalsePositiveParity,

     UnequalBaseRates
    }
   ];
```

In[21]:=
```
Parallelize[Select[KleinbergSubsets2, PrSAT[#, Probabilities → Regular] === {} &]]
```

Out[21]=
```
{ }
```

So, in this sense, this is the "*minimal*" Kleinberg et al-style (regularity-based) impossibility theorem.

Four Other (Confusion-Matrix-Based) Fairness Measures

Here are four other (confusion-matrix-based) fairness measures that have been discussed in the literature.

In[22]:=
```
NegativePredictiveValueParity = Pr[𝒴 | ¬ C ∧ 𝒜] == Pr[𝒴 | ¬ C ∧ ¬ 𝒜];
OverallAccuracyEquality =
  Pr[C | 𝒴 ∧ 𝒜] + Pr[¬ C | ¬ 𝒴 ∧ 𝒜] == Pr[C | 𝒴 ∧ ¬ 𝒜] + Pr[¬ C | ¬ 𝒴 ∧ ¬ 𝒜];
EqualizingDisincentives = Pr[C | 𝒴 ∧ 𝒜] - Pr[C | ¬ 𝒴 ∧ 𝒜] ==
   Pr[C | 𝒴 ∧ ¬ 𝒜] - Pr[C | ¬ 𝒴 ∧ ¬ 𝒜];
TreatmentEquality = Pr[C | ¬ 𝒴 ∧ 𝒜] / Pr[¬ C | 𝒴 ∧ 𝒜] == Pr[C | ¬ 𝒴 ∧ ¬ 𝒜] / Pr[¬ C | 𝒴 ∧ ¬ 𝒜];
```

Terminological Notes: In the literature on fairness measures, the conjunction of Predictive Parity & Negative Predictive Parity is known as "Conditional Use Accuracy." See Caton & Haas (2020, §3.2.2) for a survey of confusion-matrix-based fairness measures (we adopt their terminology for our last three measures).

Automating the Search for Impossibility Results — And Impossibilities #3 and #4

We can use **PrSAT** to search for (e.g.) all 3-element subsets of our total set of 8 fairness constraints above which are jointly unsatisfiable (assuming Unequal Base Rates & Regularity). In this way, we can fully automate the search for impossibility results (in the general style of Kleinberg et al).

Here's how we can use **PrSAT** to exhaustively search for all 3-element sets of constraints (from our list of 8 fairness constraints, above) that are jointly unsatisfiable, in the presence of Unequal Base Rates and Regularity.

In[26]:=
```
ElementQ[o_, s_] := Or @@ (# === o & /@ s);
```

In[27]:=
```
ThreeElementSubsets = Subsets[
   {
    PredictiveParity,
    TruePositiveParity,
    FalsePositiveParity,
    StatisticalParity,
    NegativePredictiveValueParity,
    OverallAccuracyEquality,
    EqualizingDisincentives,
    TreatmentEquality
   }, {3}
  ];
```

There are 56 3-element subsets. Instead of exhaustively searching this entire set, we will examine only those 3-element sets containing `TruePositiveParity` (and we'll only search for 20 seconds per subset). There are 21 of these.

In[28]:=
```
SearchSet = Select[ThreeElementSubsets, ElementQ[TruePositiveParity, #] &];
```

Our short and incomplete search produces 11 impossibilities out of these 21 3-element subsets of constraints.

In[29]:=
```
Impossibilities =
  Quiet[Parallelize[Select[SearchSet, Quiet[TimeConstrained[PrSAT[Union[#,
            {UnequalBaseRates}], Probabilities → Regular], 20]] === {} &]]];
```

In[30]:=
```
Length[Impossibilities]
```

Out[30]=
```
11
```

That is, we now know that there are (at least) 11 jointly unsatisfiable 3-element subsets (containing `TruePositiveParity`) of our set of 8 fairness measures. Two of these 11 correspond to the impossibility result of Kleinberg et al. (*i.e.*, they are both notational variants of their result).

But, some of the impossibilities among these 11 appear to be novel.

For instance, as far as we know, the following two impossibility results are novel (at least, we haven't seem them in the literature).

Impossibility #3. Statistical Parity, True Positive Parity, and Negative Predictive Parity are jointly unsatisfiable (assuming unequal base rates & regularity).

In[31]:=
```
PrSAT[
 {
  StatisticalParity,
  TruePositiveParity,
  NegativePredictiveValueParity,

  UnequalBaseRates
 }, BypassSearch → True, Probabilities → Regular
]
```

Out[31]= { }

And, there are no proper subsets of this set that are unsatisfiable.

In[32]:=
```
Subsets3 = NonemptyProperSubsets[
   {
    StatisticalParity,
    TruePositiveParity,
    NegativePredictiveValueParity,

    UnequalBaseRates
   }
  ];
```

In[33]:=
```
Select[Subsets3, PrSAT[#, BypassSearch → True, Probabilities → Regular] === {} &]
```

Out[33]= { }

So, this appears to be a new “minimal” inconsistent set of three fairness measures (assuming unequal base rates & regularity).

Impossibility #4. Predictive Parity, True Positive Parity, and Treatment Equality are jointly unsatisfiable (assuming unequal base rates & regularity).

In[34]:=
```
PrSAT[
 {
  PredictiveParity,
  TruePositiveParity,
  TreatmentEquality,

  UnequalBaseRates
 }, BypassSearch → True, Probabilities → Regular
]
```

Out[34]= { }

And, there are no proper subsets of this set that are unsatisfiable.

In[35]:=
```
Subsets4 = NonemptyProperSubsets[
    {
     PredictiveParity,
     TruePositiveParity,
     TreatmentEquality,

     UnequalBaseRates
    }
   ];
```

In[36]:=
```
Select[Subsets4, PrSAT[#, BypassSearch → True, Probabilities → Regular] === {} &]
```

Out[36]= { }

So, this appears to be another new “minimal” inconsistent set of three fairness measures (assuming unequal base rates & regularity).

It is worth noting that there are no 2-element subsets of our set of 8 fairness measures that are jointly unsatisfiable (assuming unequal base rates and regularity).

In[37]:=
```
TwoElementSubsets = Subsets[
    {
     PredictiveParity,
     TruePositiveParity,
     FalsePositiveParity,
     StatisticalParity,
     NegativePredictiveValueParity,
     OverallAccuracyEquality,
     EqualizingDisincentives,
     TreatmentEquality
    }, {2}
   ];
```

In[38]:=
```
Select[TwoElementSubsets,
 PrSAT[Union[#, {UnequalBaseRates, Pr[𝒜] == 1/2, Pr[𝒴 | 𝒜] == 2/3, Pr[𝒴 | ¬ 𝒜] == 1/3}],
    BypassSearch → True, Probabilities → Regular] === {} &]
```

Out[38]= { }

Note: in this last search, we added a few equational constraints to speed-up the search for satisfying probability models. Since models *were* found for each 2-element subset (even with these additional constraints added), this results in no loss of generality for present purposes.

`PrSAT` is a very useful tool for discovering, verifying, and generalizing impossibility results (in this notebook, we're just scratching the surface).

Generalizing The Impossibilities — Beyond $\mathcal{A}$ and $\neg\mathcal{A}$

In our analysis above, we assumed that we had one (binary) protected attribute $\mathcal{A}$. This is tantamount to assuming we're dealing with two populations (protected and non-protected) which are mutually exclusive and exhaustive. All of the results above can be generalized by supposing we are dealing with two populations $\mathcal{A}$ and $\mathcal{B}$ (instead of $\mathcal{A}$ and $\neg\mathcal{A}$), which need not be mutually exclusive or exhaustive (thus allowing for the presence of intersectionality effects, *etc.*). In fact, we need not assume anything about the relationship between populations $\mathcal{A}$ and $\mathcal{B}$ (except what is already encoded in the fairness measures/side conditions). Here are the corresponding (eight) generalized fairness conditions.

In[39]:=
```
PredictiveParityG = Pr[𝒴 | C ∧ 𝒜] == Pr[𝒴 | C ∧ ℬ];
TruePositiveParityG = Pr[C | 𝒴 ∧ 𝒜] == Pr[C | 𝒴 ∧ ℬ];
FalsePositiveParityG = Pr[C | ¬ 𝒴 ∧ 𝒜] == Pr[C | ¬ 𝒴 ∧ ℬ];
StatisticalParityG = Pr[C | 𝒜] == Pr[C | ℬ];
NegativePredictiveValueParityG = Pr[𝒴 | ¬ C ∧ 𝒜] == Pr[𝒴 | ¬ C ∧ ℬ];
OverallAccuracyEqualityG =
  Pr[C | 𝒴 ∧ 𝒜] + Pr[¬ C | ¬ 𝒴 ∧ 𝒜] == Pr[C | 𝒴 ∧ ℬ] + Pr[¬ C | ¬ 𝒴 ∧ ℬ];
EqualizingDisincentivesG = Pr[C | 𝒴 ∧ 𝒜] - Pr[C | ¬ 𝒴 ∧ 𝒜] == Pr[C | 𝒴 ∧ ℬ] - Pr[C | ¬ 𝒴 ∧ ℬ];
TreatmentEqualityG = Pr[C | ¬ 𝒴 ∧ 𝒜] / Pr[¬ C | 𝒴 ∧ 𝒜] == Pr[C | ¬ 𝒴 ∧ ℬ] / Pr[¬ C | 𝒴 ∧ ℬ];
```

And, here is the generalized assumption of Unequal Base Rates.

In[47]:=
```
UnequalBaseRatesG = Pr[𝒴 | 𝒜] ≠ Pr[𝒴 | ℬ];
```

Here is a verification of the generalized Chouldechova (2017) impossibility theorem.

In[48]:=
```
PrSAT[
 {
  PredictiveParityG,
  TruePositiveParityG,
  FalsePositiveParityG,
  StatisticalParityG,

  UnequalBaseRatesG
 },
 BypassSearch → True
]
```

Out[48]= { }

Here is a verification of the generalized Kleinberg et al (2016) impossibility theorem.

In[49]:=
```
ImperfectClassifierG =
  And @@ {Pr[C | ¬ 𝒴 ∧ 𝒜] ≠ 0, Pr[C | ¬ 𝒴 ∧ ℬ] ≠ 0, Pr[C | 𝒴 ∧ 𝒜] ≠ 1, Pr[C | 𝒴 ∧ ℬ] ≠ 1};
NonZeroPrecisionG = Pr[𝒴 | C ∧ 𝒜] ≠ 0 ∨ Pr[𝒴 | C ∧ ℬ] ≠ 0;
```

In[51]:=
```
PrSAT[
 {
  PredictiveParityG,
  TruePositiveParityG,
  FalsePositiveParityG,

  UnequalBaseRatesG,
  ImperfectClassifierG,
  NonZeroPrecisionG
 },
 BypassSearch → True
]
```

Out[51]= { }

The other two impossibilities discussed above also generalize in this fashion.

In[52]:=
```
PrSAT[
 {
  StatisticalParityG,
  TruePositiveParityG,
  NegativePredictiveValueParityG,

  UnequalBaseRatesG
 }, BypassSearch → True, Probabilities → Regular
]
```

Out[52]= { }

In[53]:=
```
PrSAT[
 {
  PredictiveParityG,
  TruePositiveParityG,
  TreatmentEqualityG,

  UnequalBaseRatesG
 }, BypassSearch → True, Probabilities → Regular
]
```

Out[53]= { }

Downloading this *Mathematica* Notebook

If you are reading a PDF version of this *Mathematica* notebook and you'd like to download the notebook itself, it is available at

http://fitelson.org/exploring_impossibility.nb

References

Chouldechova, A. “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments.” *Big data* 5.2 (2017): 153-163.

Fitelson, B. “A Decision procedure for Probability Calculus with Applications.” *Review of Symbolic Logic* 1.1 (2008): 111-125.

Kleinberg, J, S. Mullainathan, and M. Raghavan. “Inherent trade-offs in the fair determination of risk scores.” *arXiv preprint* arXiv:1609.05807 (2016).

Mehrabi, N., et al. “A survey on bias and fairness in machine learning.” *arXiv preprint* arXiv:1908.09635 (2019).

Caton, S., and C. Haas. “Fairness in Machine Learning: A Survey.” *arXiv preprint* arXiv:2010.04053 (2020).