Algorithmic Fairness, Accountability, and Ethics 2022

Intersectional Fairness

Frederik Sloth	frsl@itu.dk	Master in Data Science
Edi Begovic	edbe@itu.dk	Master in Data Science
Anita Justyna Zielinska	anzi@itu.dk	Master in Data Science

Introduction

In modern society, fairness is something we have become more and more aware of. There has been a noticeable tendency for humans to be treated unfairly based on attributes such as their gender, race and age. Avoiding this kind of biased treatment should be a goal in all aspects of modern society - not least in machine learning models. Through the recent years, many scholars have presented approaches to ensuring fairness within machine learning. Their goal has been to ensure that sensitive attributes like race and gender do not have an effect on people's probability of getting positive or negative outcomes in machine learning models.
An aspect that most do not consider, however, is the intersectional fairness of sensitive attributes. Ensuring that all races and sexes are treated fairly is one thing, but another is to ensure the same for the subgroups within these groups. Men and women or asians and caucasians may be treated equally overall, but are white women also treated fairly when compared to asian males? Not ensuring this can lead to fairness gerrymandering, where a model or data seems fair when comparing the fairness of the overall attributes, but turns out to be greatly unfair when considering the subgroups. For example, a university might accept equal numbers of men and women and equal numbers of black and white people, but when you look at the subgroups it becomes clear that they only accept white men and black women. The perceived fairness is therefore a result of fairness gerrymandering, and the problem of mitigating unfairness has not actually been solved. To truly ensure fairness, the intersections must therefore be taken into consideration.
Our goal is to investigate intersectional fairness and explore different approaches for how it is obtained. We do this using the "UCI Census Income" dataset, also known as the "Adult Dataset", and aim to ensure fairness for the intersection between the two sensitive attributed 'sex' and 'race'. We evaluate the fairness using a 'min-max' ratio, which considers the elements of demographic parity, equal opportunity and equalized odds.

Background

Through the years, many papers have gone into the problem of unfairness in machine learning, and presented approaches on how to minimize it. There is however, not a single agreed upon definition for fairness in Machine Learning. (Verma and Rubin, 2018) goes through several of the most prevalent definitions. These include group fairness, conditional statistical parity, equal

opportunity and equalized odds. The concepts of equality of odds and equality of opportunity in supervised learning are introduced by (Hardt et al, 2016) in greater detail as metrics on which to evaluate the fairness.

While the aforementioned approaches present several ways of mitigating unwanted bias with multiple sensitive attributes, they do not consider the specific problem of intersectional bias. As defined by (Foulds et al, 2019), an intersectional definition of fairness is one that considers multiple protected attributes and ensures that all intersections of protected attributes are protected by definition, while still ensuring protection for all protected attributes individually. (Kobayashi and Nakao, 2020) aim to solve this for binary classification by coming up with one possible approach for solving this problem with a method called One-vs.-One mitigation. One-vs.-One mitigation expands upon already existing pre-processing, in-processing and post-processing methods, in order to create a method that is applicable for mitigating intersectional bias for a diverse range of approaches. (Morina et al, 2020) likewise define methods for intersectional fairness for classification problems that are extensions of classical approaches. They present metrics for evaluating intersectional fairness, then present several methods for estimating these metrics, and finally present post processing techniques for binary classifiers in order to achieve intersectional fairness. The paper also uses the 'Adult' dataset, showing it to have ample amounts of data for the intersectional fairness problem. (Ghosh et al, 2021) introduces a worst-case comparison method for intersectional fairness in the form of the 'min-max ratio'. The min-max ratio is then used along with several existing metrics of fairness such as Demographic parity, Conditional statistical parity, Equal opportunity and Group Benefit Equality to encompass intersectionality.

Data

For our project we use the "Census Income Dataset", also known as the "Adult dataset".[1] The dataset was created using the 1994 Census Database. It was made specifically for machine learning tasks, more precisely for making predictions on whether a person makes over 50 thousand dollars in a year. The original Census dataset includes values on a wide range of variables normal for census data, but the Adult Dataset only extracts the ones considered relevant for the machine learning task. The attributes includes of: age, workclass, education, marital status, occupation, relationship, race, sex, hours per week and of course a boolean attribute for whether or not the person earns over 50.000 dollars a year. The dataset is known as 'Adult Data' due to the fact that it only has data on people above the age of 16. This is done as only those above 16 should be considered as having the possibility of working a full time job and thereby be able to make more than 50.000 dollars a year. If children were included it would skew the results, as no child makes 50.000 dollars and therefore age would become a main determinator for predictions.
For considering intersectional fairness we chose three sensitive attributes. The chosen attributes are 'sex' and 'race', both of which are attributes often correlated with unfair treatment for certain groups. The 'sex' attribute is binary, holding either the attribute 'Male' or the attribute 'Female'. The 'race' attribute holds 5 possible values: 'White', 'Black', 'Asian-Pac-Islander', 'Amer-Indian-Eskimo', or 'Other'. Besides that we of course keep the 'over_50k' columns, which

[1] https://archive.ics.uci.edu/ml/datasets/adult

also has two possible values: '<=50K' and '>50K'. This is the target value our classifier should predict on.

When looking further into the attributes, it can be seen that the data is made up of 67% males and 33% females. In terms of race 'White' by far makes up the largest population with 85.4%, leaving the other races to make up the remaining 14.6% of the entries in the data. Of those 14.6% the 9.6% are 'Black', 3.2% are 'Asian-Pac-Islander', 1% is 'Amer-Indian-Eskimo' and the remaining 0.8% answered 'Other'. For the attribute we predict on, 'over_50k', the distribution is that 76% earn 50k or less yearly, while 24% earn more. The two are therefore not that evenly split.
When working with the data, we encoded it into binary classes. For the 'sex' attribute we simply mapped males 0 and females 1. For race, as 'White' was the by far most prominent value, we split the data into 'White' and 'Non-White', encoded 'White' as 0 and 'Non-White' as 1. For the target attribute 'over_50k' we encoded it so that under 50k was 0 and over 50k was one. Thereby we ended up with 3 attributes of binary values. Figure 1 shows the distribution between the 'sex' attribute, as well as the distribution of the subgroups of 'sex' and 'race'.

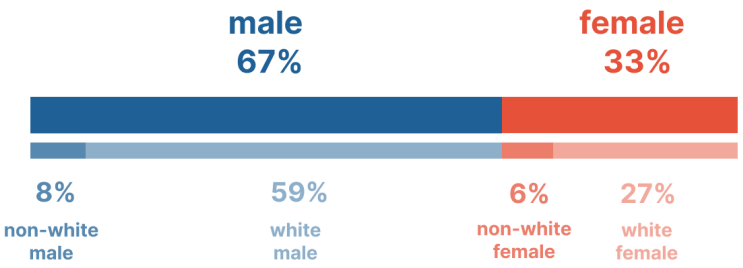


Figure 1 - Class distribution for protected variables

Additionally we kept the 'age' attribute for use in further experiments on intersectional fairness with more than two sensitive attributes.. The age attribute holds integer values in the range of 16 to 90. We looked into the data to find the age at which to split it into two, to get two somewhat even groups. We did this so that when we use it for intersectional fairness we keep the subgroups as large and evenly sized as possible. We found this by using the median age, which was 37, and then encoding everyone younger than 37 as 0 and everyone who is 37 or older as 1.

Approach

While working with a few separate sensitive features, we can achieve a situation, in which a classifier appears to be fair for particular groups, but breaches when considering a convergence of them. In our project, we intend to go towards testing the effectiveness of the geometric de-biasing method set out by (He et al, 2020) for intersecting protected attributes. The method described in the paper focuses on mitigating unfairness on the data level (pre-processing). Using linear algebra, the proposed algorithm removes protected features' impact on the dataset. The general aim of the method is to create a representation for every non-protected feature, that is uncorrelated with protected columns, but highly correlated with the original feature. A big

advantage of such a solution is the possibility of its usage in various models, as it operates on input data only, and has low computational complexity. What distinguishes it from other known approaches is the preservation of the interpretability of the modified features.

Our approach is based on applying two different representations of the sensitive attributes to de-bias the remaining non-sensitive features using the approach developed by (He et al.). Firstly, we simply de-bias on a per-feature basis as the method suggests, such that all features are de-correlated from each sensitive attribute. Secondly, we also encode the intersection of all values of the sensitive attributes as their own binary feature (e.g. *white-female-young*). As the number of intersections grows exponentially relative to the number of sensitive features *and* their values, we limit ourselves to only using binary sensitive attributes. We also only work with up to 3 sensitive attributes (in total 8 intersections), which provide us with a big enough subset of data for all feature combinations. Our aim is to compare how well the de-biasing method handles fairness for the intersection of the sensitive attributes given the two representations.

Fairness measures

To evaluate fairness, we consider three common group-based fairness metrics which are applicable to diverse contexts and are often the basis for more complex measures. These are *demographic parity*, *equalized odds* and *equal opportunity* as defined in Table 1 below.

Demographic parity	Equal opportunity	Equalized odds
Defined as the difference in *positive rate* between two groups. Each segment of a protected group should receive positive outcomes at equal rates.	Defined as the difference in *true positive rate* for two groups. The same proportion of *qualified* members should receive a positive outcome for both segments.	Defined by the combined difference in *true positive rate* and *true negative rate*. The proportion of *unqualified* members who receive a negative outcome should *also* match.

Table 1 - Fairness metrics

The above fairness criteria are satisfied when their rates match for all protected groups. To extend these metrics for the intersectional case, we use *worst-case disparity* as presented in (Ghosh et al.). A subgroup is defined as a set containing a permutation of members in the groups of the sensitive attributes. One subgroup will therefore exist for every possible combination of values within the sensitive attributes. Values for a given fairness metric as defined above are calculated for every subgroup and the measure is summarized by the ratio of the minimum and maximum value. Considering the worst-case thus puts a bound on the impact on the applied bias mitigation. In (Morina et al) it is stated that given a set of sensitive attributes, if a fairness metric is satisfied for all of its intersectional subsets, then it is also satisfied with respect to any individual sensitive attribute.

Implementation

To assess the effectiveness of the two representations, we implement a simple logistic regression model for classification which we mostly treat as a black box. We use default parameters apart from the hyperparameter 'C' for regularization strength, which is fine-tuned during the training process and also using balanced class weights. Two instances of the model are trained on the same dataset, only de-biased using the two different representations. We use 5-fold cross-validation during training with further 5 split runs for different permutations of the data. For each inner run we calculate all relevant metrics and at last return the aggregate mean value.

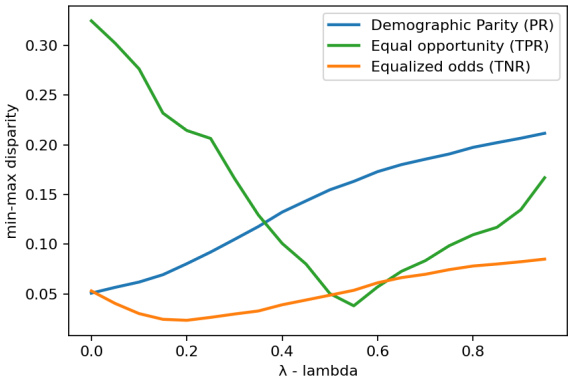
The de-biasing method also allows us to specify the strength of the de-biasing projection with the lambda parameter (*λ=0: full de-correlation, λ=1: original data)*. We'll refer to the strength of this parameter as the applied *fairness level*.

Results

We perform de-biasing on multiple representations of the sensitive attributes to assess the effect on intersectional bias, starting with sex and race as sensitive attributes. We further train each representation with varying levels of fairness during the de-biasing step in 0.05 intervals. All results are based on the baseline classifier (logistic regression) with identical parameters.

In Figure 2 we show disparity for three metrics over varying fairness levels (*λ*) for both de-biasing on individual sensitive attributes (a) and the one-hot encoded representation on each subgroup (b).

(a) de-biased on individual sensitive attributes



(b) de-biased on one-hot encoded subgroups of sensitive attributes

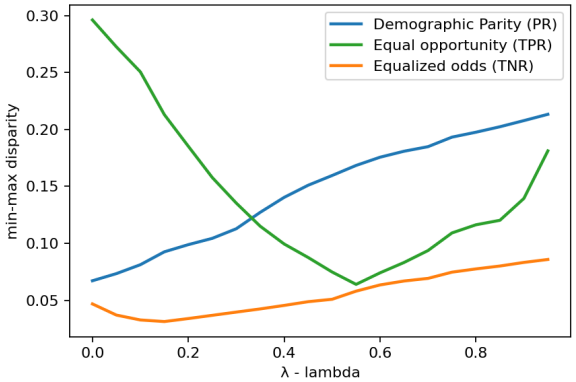


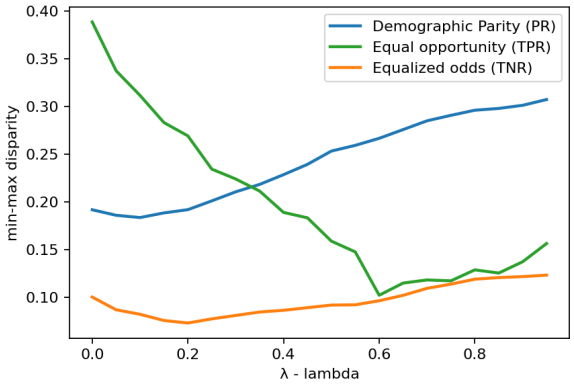
Figure 2 - min-max disparities for differing levels of λ (fairness level) for gender and race as sensitive attributes

We observe little relative change between using the two different representations. Both manage to decrease disparity equally well on *demographic parity* and *equalized odds* (simply represented as the *true negative rate*), while *equal opportunity* sees a steep increase in disparity at the highest fairness level (*λ = 0*). Thus, *equalized odds* as defined in Table 1, minimizing disparity in both

TPR and TNR, is better obtained at a more relaxed choice of fairness level ($\lambda = 0.55$). Model performance also doesn't differ between representations. Overall accuracy (balanced) drops from 75.0% ($\lambda = 1$) to 72,2%($\lambda = 0$) for both cases.

We extended our experiment to also incorporate age as a sensitive attribute. Given that the feature is continuous, we make it binary by binning at the median value (age $\geq$ 37). With three binary sensitive attributes, the total number of subgroups is 8, ranging between 1,015 and 10,548 instances for the smallest and largest subgroup, respectively. Further, for the smallest subgroup only 43 instances (4%) have a positive label (high income), making some subgroups very susceptible to uncertainty.

(a) de-biased on individual sensitive attributes



(b) de-biased on one-hot encoded subgroups of sensitive attributes

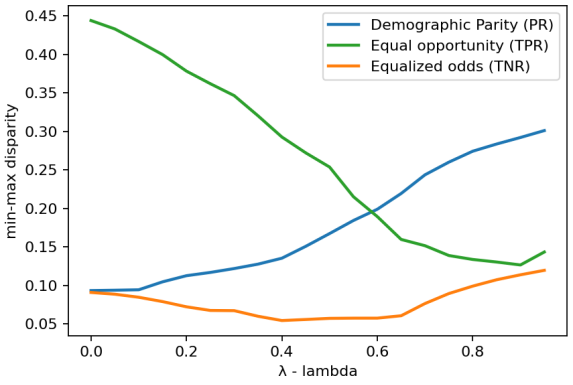


Figure 3 - min-max disparities for differing levels of λ (fairness level) for gender, race and age as sensitive attributes

As when using only two sensitive attributes, we see little improvement in disparity for *equal opportunity* and *equalized odds*. Only for *positive rates* do we see a clear decrease in disparity as fairness level increases. We do observe a noticeable decrease in disparity for *positive rates* with the subgroup representation going from 0.19 to 0.10 at the highest fairness level (Figure 3b). Even with the age attribute removed and de-correlated for, overall model accuracy (balanced) only decreased from 74.9% to 71.6%.

Discussion

The results showed little to no improvement in terms of disparity for both *equal opportunity* and *equalized odds*. This is the case both when de-biasing on individual sensitive attributes and when using one-hot encoded attributes. It therefore seems that the geometric de-biasing method does not have much of an effect on these aspects of fairness. For demographic disparity however, we clearly see an improvement. This goes for both the individual fairness and the subgroup based fairness as both perform equally well. When adding additional sensitive attributes, in this case 'age', the cases with one-hot encoded subgroups start performing better than the individual disparity. The results therefore seem to imply that for when the data has more sensitive attributes we want to avoid discriminating on, doing intersectional encoding should be preferred. More experiments should be run however to confirm this, which would be a goal in future work.

Overall, however, enforcing intersectional fairness appears to give just as good or better results as only enforcing individual fairness. As ensuring intersectional fairness additionally to individual fairness helps avoid fairness gerrymandering, and it achieves comparable results to the individual debiasing with marginal differences in performance, intersectional fairness can only be recommended.

The de-biasing method presented in (He et al.) works by eliminating linear correlation between the features used for model training and the sensitive attributes. Their results only reflect on positive rates (*demographic parity*) and Peasrson correlation between features as a measure of fairness. We therefore didn't have prior expectations on the method's effectiveness on reducing disparity for other metrics based on error-rates. Further, a significant proportion of the non-sensitive attributes from the Adult dataset are categorical, represented through binary dummy variables. This may have reduced the effectiveness of the method. Future work includes replicating the experiment with the same premises using other de-biasing techniques based on pre-processing data. The inherent problem of using one-hot encodings for sensitive categorical attributes is covered by (Mougan et al.). They show that concatenated one-hot encodings for sensitive attributes consistently discriminate more (on *equal opportunity*) than using target encodings.

A problem fundamental with ensuring intersectional fairness is the number of intersections. Ensuring individual fairness can already be a large task, especially in cases with many sensitive attributes that we want to avoid discrimination on. Yet when looking at intersectional fairness, this becomes an exponentially larger task. If we consistently have sensitive attributes with binary values, as we did here, the number of subgroups would increase at the power of two for every single new attribute we wanted to include in the intersectional fairness. If the attributes were to have more than two values, the number of subgroups would increase even faster. It would therefore quickly become infeasible to ensure fairness for all of the intersections. This can be argued as being one of the reasons behind many approaches only considering individual fairness and not intersectional fairness, as intersectional fairness can often end up being a much more expensive task. At the same time, for every new sensitive attribute that is added, all subgroups would be split into two. This way, subgroups consistently become smaller and smaller - in best cases decreasing in size by 50% and in the worst cases being divided unevenly such that some drop significantly in size. As a result of this, subgroups quickly become really small if the dataset is not really large and fairly distributed. (Mougan et al.) also discusses the importance of *induced reducible bias* which encompasses unfairness that is introduced by the large variance found in the small subgroups. This is an aspect which we haven't considered in our data selection.

Ethics

In data science, one of the priorities should always be to provide a faithful representation of the world and constantly search for the true essence of certain correlations, rather than thoughtlessly allowing hurtful patterns to be amplified. In our case, it seems fundamental that a person's earnings should depend only on obviously related features, like education completed, area of occupation, or number of hours worked per week and not on superficial attributes like race and

sex. The conviction that these assumptions are self-evident stems from treating human nature as one of the fundamentals of the justification system. The natural consequence of this is that human rights (also in terms of income) are equal. The data we worked with only mimicked ingrained stereotypes, so the model without any intervention would duplicate them harming certain groups. In this case, the utilitarian approach seems to coincide with our motivation for de-biasing data - utilitarianism considers the interests of all humans equally.

Modern AI models give us a variety of possible solutions that can be selected and adapted to specific requirements, while GOFAI doesn't offer this flexibility, mainly due to restrictions on internal data representation. What is more, symbolic AI can't deal appropriately with uncertainty, because it is simply a representation of human logic. Putting all the trust in abstract internal representations in modern AI simply moves the uncertainty towards the human interpretable metrics they represent, for instance, fairness. After all, our fairness metrics can only be defined by our human intuitions, drawing arbitrary lines between properties in the data. Our understanding of fairness might change over time as societal values and our perception of human attributes evolve. Symbolic AI is dependent on the logic we imbue into it, paying no regard to societal changes, but modern AI can at least update its intuition over time if supplied with new data.
Using modern AI, it is easier to correct human errors and unfairness by having alternative underlying representations that relate to our target, increasing fairness.

Conclusion

In this paper/project we examine the impact of fairness on intersectional subgroups of sensitive attributes using a linear de-biasing technique introduced by (He et al.). We focus on the impact of encodings for categorical sensitive attributes and their intersections and assess their importance through experiments. We implement a logistic regression model for classification, we examine the performance and fairness of the two representations. Results show that de-biasing has little impact on model accuracy while improving fairness on demographic parity. In our testing, using few binary sensitive attributes, one-hot encoded intersectional subgroups and individual sensitive attributes only differ marginally in relation to min-max disparity for various fairness metrics. Our results further show little improvement in other error-based fairness metrics apart from *demographic parity*. We therefore, as part of future work, want to further evaluate the findings using alternative techniques for de-biasing datasets. Overall, our results show no substantial performance penalty when encoding subgroups as separate features, while improving intersectional fairness.

References

Foulds, James R., et al. “An Intersectional Definition of Fairness.” 2019, https://arxiv.org/abs/1807.08362.

Ghosh, Avijit, et al. “Characterizing Intersectional Group Fairness with Worst-Case Comparisons.” 2021, https://arxiv.org/abs/2101.01673.

Hardt, Moritz, et al. “Equality of Opportunity in Supervised Learning.” 2016, https://arxiv.org/pdf/1610.02413.pdf.

He, Yuzi, et al. “A Geometric Solution to Fair Representations.” 2020, https://dl.acm.org/doi/pdf/10.1145/3375627.3375864.

Kobayashi, Kenji, and Yuri Nakao. “One-vs.-One Mitigation of Intersectional Bias: A General Method to Extend Fairness-Aware Binary Classification.” 2020, https://arxiv.org/pdf/2010.13494.pdf.

Morina, Giulio, et al. “Auditing and achieving intersectional fairness in classification problems.” 2020, https://arxiv.org/pdf/1911.01468.pdf.

Mougan, Carlos, et al. “Fairness implications of encoding protected categorical attributes.” 2022, https://arxiv.org/abs/2201.11358.

Verma, Sahil, and Julia Rubin. “Fairness Definitions Explained.” 2018, https://fairware.cs.umass.edu/papers/Verma.pdf.