Mitigation Methodology Document

Categories Covered

Calibrated Equalized Odds Post-processing • Equalized Odds Post-processing • Reject Option Based Classification • Pre-processing • Machine Learning (ML) • Fairness • Risk Assessment.

Name: BIAS MITIGATION POST-PROCESSING FOR INDIVIDUAL AND GROUP FAIRNESS

Link: https://arxiv.org/pdf/1812.06135.pdf

Category: Post-Processing | ROC, EOP, and IGD

Summary

This article proposes methods for increasing both individual and group fairness. Group fairness is defined by splitting a population into protected attributes (such as gender) and seeks for some statistical measure to be equal across groups. Individual fairness seeks for similar individuals to be treated similarly. There are three post-processing algorithms used in this study: (1) Individual Group Debiasing (IGD), (2) Equalized Odds Post-processing (EOP), and (3) Reject Option Classification (ROC). Each of these algorithms are compared by three measures: (a) individual bias, (b) disparate impact, and (c) balanced classification accuracy. The AI Fairness 360 toolkit was used in this analysis for both the EOP and ROC algorithms. Both algorithms are used to mitigate bias in predictions. The EOP algorithm modifies the predicted labels using an optimization scheme to make predictions fairer while the ROC algorithm changes predictions from a classifier to make them fairer. Trade-offs are on page 4.

Name: On Fairness and Calibration

Link: https://proceedings.neurips.cc/paper/2017/file/b8b9c74ac526fffbeb2d39ab038d1cd7-Paper.pdf

Category: Post-Processing | EOP

Summary

Equalized Odds (also referred to as Disparate Mistreatment) is a classification algorithm which aims to ensure that no error type (false-positive or false-negative) disproportionately affects any population subgroup. The article introduces a method called Relaxed Equalized Odds with Calibration, which analyses the trade-offs between false-positive and false-negative rates. The algorithm provided achieves the calibrated Equalized Odds relaxation by post-processing existing calibrated classifiers. Implications of the algorithm is provided along with a few objections. The big takeaway from this article is that it is impossible to satisfying multiple equal-cost constraints. The statistics team should take a look at their probabilistic classifier on pages 3, 4, and 7.

Name: Decision Theory for Discrimination-aware Classification

Link: https://mine.kaust.edu.sa/Documents/papers/ICDM_2012.pdf

Category: Post-Processing | ROC and DAE

Summary

This paper resents two solutions for discrimination-aware classification that neither require data modification nor classifier tweaking. There are two methods used: ROC invokes the reject option and labels instances belonging to deprived and favored groups in a manner that reduces discrimination. Our second solution, called Discrimination-Aware Ensemble (DAE), exploits the disagreement region of a classifier ensemble to relabel deprived and favored group instances for reduced discrimination.

Advantages:

1. Solutions are not restricted to a particular classifier.
2. Solutions require neither modification of learning algorithm nor preprocessing of historical data.
3. Solutions give better control and interpretability of discrimination-aware classification to decision makers.

Statistics team should look at page 3 and analyze the original solutions of ROC and DAE. Page 4 and 5 provide comparisons of their results and previous work. They utilize 4 classifiers in their experimentation: naive Bayes (NBS), logistic regression (Logistic), k-nearest neighbor (IBK), and decision tree (J48). They discovered a decrease in discrimination, but a loss in accuracy. Both solutions provide the decision maker with easy control over the resulting discrimination. One thing to note is how they handled their sensitive attributes (page 6).

Name: Compatible API Reference

Link:

https://aif360.readthedocs.io/en/latest/modules/sklearn.html#module-aif360.sklearn.postprocessing

Category: APIs with AI Fairness 360 | Post-processing, In-processing, and Pre-processing

Summary

Simply look through if you have the time.

Name: Reducing-ai Bias with Rejection Option Based Classification

Link:

https://towardsdatascience.com/reducing-ai-bias-with-rejection-option-based-classification-54fefdb53c2e

Category: Post-processing | ROC

Summary:

This article summarizes some of the key findings in essential post-processing publications such as "*On Fairness and Calibration*", "*Equality of Opportunity in Supervised Learning*" and "*Decision Theory for Discrimination-aware Classification*". The focus is on the usage of ROC. We know discrimination can occur in three places; this article proposes that the most discrimination occurs around the decision boundary (classification threshold). The method used here uses the low confidence region of a classifier for discrimination reduction and reject its predictions. Through this process the hope is to reduce the bias in model predictions. One advantage this method has over other methods is that the final predictions can be manipulated easily.

This article is concise and easy to read. The statistics team should look the entire article over.

Name: Sample-COMPAS-Risk-Assessment-COMPAS-"CORE"
Link: https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html

Category: Risk-Assessment

Summary:

This is a sample COMPAS Risk Assessment obtained from Wisconsin. The survey consists of 137 questions that asks for information ranging from a defendant's criminal history to his or her social life and thoughts. This survey serves as a visual of what the data collection looks like before we make assessments and apply analytics. All teams can take a look at this risk assessment.

Name: Equality of Opportunity in Supervised Learning

Link: https://arxiv.org/pdf/1610.02413.pdf

Category: Post-processing | EOP

Summary

There are many approaches to mitigate discrimination. The naïve-bayes, demographic parity and more. Disadvantages of these approaches are mentioned on page 2. This paper considers non-discrimination from the perspective of supervised learning. The goal is to predict a true outcome Y from features X based on labeled training data, while ensuring they are "non-discriminatory" with respect to a specified protected attribute A. Ultimately, they want to show how to optimally adjust any learned predictor so as to remove discrimination according to their definitions. They propose an "oblivious" notion, based only on the joint distribution, or joint statistics, of the true target Y, the predictions Y-hat, and the protected attribute A. Our project relates to their notion of oblivious because our risk score is determined by underlying training data that is not public. Similar in their case, the only information about the score is the score itself, which can then be correlated with the target and protected attribute. The Equalized Odds and Equal Opportunity criterion are provided on page 3. Page 4 beings their step process of how they selected an equalized odds or equal opportunity predictor. Core findings are derived from a binary predictor, score function, equalized odds threshold predictor and equal opportunity threshold predictor. The stats team is to look over this paper. The publication consists of many mathematical formulas and statistics throughout its entirety. Assess what you can and bring your findings to our next meeting.

Name: PRIORITY-BASED POST-PROCESSING BIAS MITIGATION FOR INDIVIDUAL AND GROUP FAIRNESS

Link: https://arxiv.org/pdf/2102.00417.pdf

Category: Post-processing

Summary:

This article proposes a priority-based post-processing algorithm to mitigate bias for individual and group fairness. Definitions for individual fairness and group fairness goes as follow: the notion of individual fairness requires that similar individuals should be treated similarly irrespective of socio-economic factors whereas group fairness seeks for some statistical measure to be equal among group defined by protected attributes (such as age, gender, race, and religion). Disparate impact (DI) is a standard measure for group fairness. The advantage of this post-processing model is that the debiasing process can be applied to any black-box model.

Like many other models, there is a trade-off between debiasing and accuracy. There is often a loss in accuracy when mitigating individual bias. As a result, there is a limit to the number of individuals allowed to be debiased. The group fairness (DI) metric is put in place as a threshold to limit the number of individual samples to be debiased. The article introduces a formula known as the *Unfairness Quotient*. The Unfairness Quotient is defined as the difference between the actual model prediction and the prediction after perturbing.

$$b_{xi,di} = abs\ (\hat{c}(x_i,\ d_i') - \hat{c}\ (x_i,\ d_i))$$

The Unfairness Quotient signifies the amount of bias associated with that sample, i.e., more the value, more the injustice and hence higher the priority during debiasing. The statistics team should analyze their algorithm on page 3. This method can be useful in terms of determining a threshold for our debiasing problem.

The protected attribute in their model was gender. One weakness previous post processing algorithms had been that they work poorly with debiasing both group and individual fairness with regression models and datasets with multi-class numerical labels. In their case, they found that the number of samples whose labels need to change to achieve fairness is less in their priority-based algorithm approach. As a result, it runs quicker than the base-line approach and it reduces the bias-accuracy tradeoff.

Name: Certifying and removing disparate impact

Link: https://arxiv.org/pdf/1412.3756.pdf

Category: Fairness

Summary:

This article focuses on key definitions regarding fairness, which follows: How to measure that fairness? What protected attributes to use when testing for fairness? What methods should be used to empirically show the effectiveness of those test? We learn the notion of disparate impact, which occurs when a selection process has different outcomes for different groups, even if the initial intent was meant to be neutral.

The article provides a threshold known as the 80% rule. If the conditional probability of positive YES without the protected attribute X, divided by the conditional probability of positive YES given protected attribute X, is less than or equal to 0.8.

$$\frac{\Pr(C = YES|X = 0)}{\Pr(C = YES|X = 1)} \leq \tau = 0.8$$

This equation is useful because the threshold monitors the quality of the classifier. This processes also involves a regression algorithm which will be used to minimize the balanced error rate (BER). There were three different classifiers used for measuring discrimination and to test the accuracy of a classification after the repair algorithm: Logistic Regression (LR), Support Vector Machine (SVM), and Gaussian Naive Bayes (GNB). For their experiment, they analyzed Adult Income and German Credit data sets. In their results, they discovered a decay in utility as fairness increased. The statistics team should focus their attention on the usage of (DI) and their minimizing balanced error rate (BER) on page 10.

Name: Data Mining for Discrimination Discovery

Link: tkdd.pdf (unipi.it)

Category: Fairness

Summary:

This article uses the civil rights definition of discrimination where it refers to unfair or unequal treatment of people based on membership to a category or a minority, without regard to individual merit. Discrimination often occurs in in situation involving credit, mortgage, insurance, labor market, and education. Algorithms often have trouble detecting discrimination because other attributes such as personal data, economic and cultural indicators often act as proxies for indirect discrimination. For example, redlining with zip codes. The goal of this article was to uncover discrimination in historical decision records by means of data mining techniques.

Two notions are addressed in this article: potentially discriminatory (direct discrimination) and potentially non-discriminatory (indirect discrimination). Here is the model they followed:

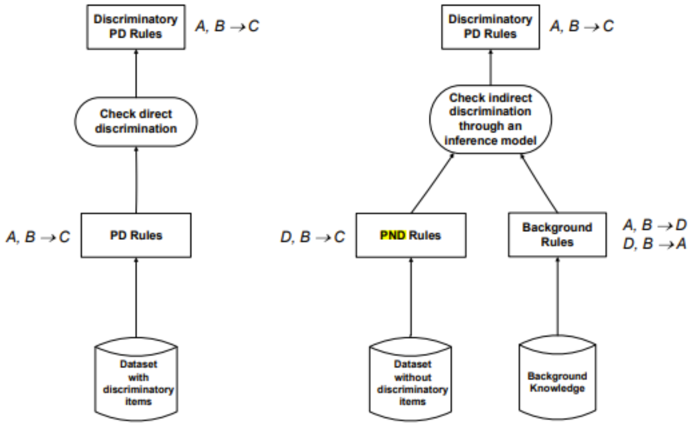

Discriminatory PD Rules
A, B → C
Check direct discrimination
A, B → C
PD Rules
Dataset with discriminatory items
Discriminatory PD Rules
A, B → C
Check indirect discrimination through an inference model
D, B → C
PND Rules
Background Rules
A, B → D
D, B → A
Dataset without discriminatory items
Background Knowledge

Fig. 1. Modelling the process of direct (left) and indirect (right) discrimination analysis.

Their PND model can help us uncover indirect discrimination in our model. In their case, they were able to identify discrimination by potentially discriminatory rules through some deduction starting from potentially non-discriminatory rules and background knowledge. In our case, we could use priors pulled from our allegation's dataset.

Name: Data preprocessing techniques for classification without discrimination

Link: https://link.springer.com/content/pdf/10.1007%2Fs10115-011-0463-8.pdf

Category: pre-processing

Summary:

This article introduces four key methods for preprocessing and learning a classifier. Those methods are suppression, massaging the data, weighing, and sampling. These methods are defined on page 3. Their experiment focused on gender discrimination in terms of hiring/employment. The favored group was male, and the unfavorable group was female. If they could find a statistically significant difference in the hiring proportions, this would indicate discrimination. One method they used was the standard statistical one-sided null hypothesis (h0 : m2 ≥ m1) approach. If the hypothesis gets rejected, the probability is high that there is discrimination. One result they discovered was that there is a linear trade-off between lowering the discrimination and lowering the accuracy.

$$acc(C) = \frac{tp + tn}{d} = \frac{tp_b + tn_b + tp_w + tn_w}{d}$$

$$disc(C) = \frac{tp_w + fp_w}{d_w} - \frac{tp_b + fp_b}{d_b}$$

An area the statistics team should focus on is their methods for determining accuracy and discrimination. They utilize the true positive, true negative, and false negative values to analyze trade-offs. Their goal is to minimize *disc(C).* The optimal equation is provided on page 10. We can utilize this approach in our project to strengthen our argument. Our proof of concept would be strengthened if we can use statistical analysis to show there is significant difference is risk-scores based off gender, age, or race.

Name: Handling Discriminatory Biases in Data for Machine Learning

Link: https://towardsdatascience.com/machine-learning-and-discrimination-2ed1a8b01038

Category: Machine Learning

Summary:

This article provides an overview of the COMPAS algorithm and summarizes some of the analysis found by ProPublica. It reveals stories of those effected by the biases in the COMPAS algorithm. The distinguishment between disparate impact and disparate treatment is provided as well. In our project we plan to use race, gender, and age as protected attributes. This article provides several additional protected attributes to draw from such as religion, disability, or national origin. There are many ways to optimize accuracy in algorithms. The author goes into detail about how to optimize for fairness. Those are: formalizing a non-discrimination criterion (1), demographic parity (2), equalized odds (3), and well-calibrated systems (4). Further explanation of these methods is provided at the center of the article. The statistics team should review each of these methods and compare the trade-offs.

Name: Learning Fair Representations

Link: Learning Fair Representations (toronto.edu)

Category: Fairness

Summary:

This article proposes a learning algorithm for fair classification that accommodates for both individual fairness and group fairness. Group fairness is defined as people of protected variable have similar proportion to total population. Individual fairness is defined by people of similar qualifications will be rated similarly. Key methods that were used in their project were statistical parity.

Equation: $$P(Z = k|\mathbf{x}^{+} \in X^{+}) = P(Z = k|\mathbf{x}^{-} \in X^{-}), \forall k$$

X^{+} and **X^{-}** - represents sub-groups.

Z - represents a random variable.

K – represents a set of prototypes.

$$\hat{y}_n = \sum_{k=1}^{K} M_{n,k} w_k$$

$\hat{y}_n$ - is the prediction for y_n based on marginalizing over each prototype's prediction for y.

We can utilize statistical parity in our project to promote group and individual fairness for race, gender, and age.

Name: A Confidence-Based Approach for Balancing Fairness and Accuracy

Link: 1601.05764.pdf (arxiv.org)

Category: Fairness

Summary:

The objective of this article is to provide statistical methods that maintain the high accuracy of these learning algorithms while reducing the degree to which they discriminate against individuals because of their membership in a protected group. In our case, if the protected attribute is race or gender – the classifier should not correlate someone's race or gender to the likelihood of them getting a higher risk score due to their membership of a particular group.

There are three key focal points this article addresses which could be useful to our application. Those are: The Shifted Decision Boundary (SDB), Statistical parity, and K-nearest-neighbor. SDB is a method based on the theory of margins and help optimize trade-offs in relation to boosting, support vector machines, and logistic regression. Statistical parity is defined as the probability of someone in protected group being approved or the probability of anyone being approved. Many of its metrics are similar to disparate impact (group fairness). K-nearest-neighbor (kNN) classifies similar individuals similarly (individual fairness). In our project, we currently have a regression model that could be a piece of an SDB mentioned in the article. Additionally, statistical parity/k-nearest-neighbor could be different ways to measure bias in the current algorithm and our future algorithm.

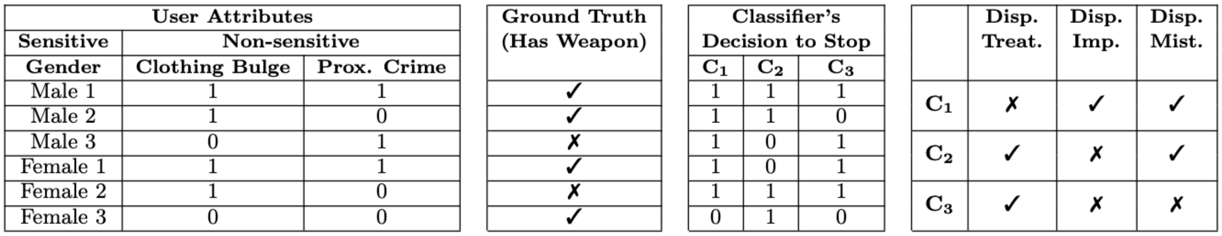
Name: Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment

Link: [1610.08452.pdf (arxiv.org)]

Category: Fairness

Summary:

Unfairness can be classified by many definitions. Disparate mistreatment measures the 'false positives' of various protected groups are different (i.e., stop-and-frisk, loan approval, etc.). Disparate treatment measures different outputs for people with similar non-sensitive attributes. Lastly, disparate impact measures different sensitive groups get different output. This article provides an experiment that measures stop-and-frisk bias. Here, it shows their sensitivity analysis based on attributes along with their decision rules.



User Attributes		
Sensitive	Non-sensitive	
Gender	Clothing Bulge	Prox. Crime
Male 1	1	1
Male 2	1	0
Male 3	0	1
Female 1	1	1
Female 2	1	0
Female 3	0	0

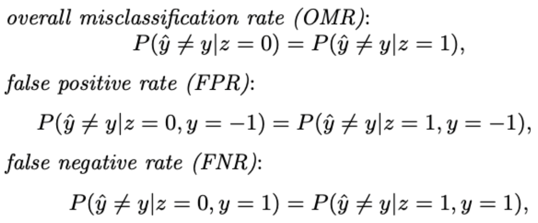
Ground Truth (Has Weapon)
✓
✓
✗
✓
✗
✓

Classifier's Decision to Stop		
C_1	C_2	C_3
1	1	1
1	1	0
1	0	1
1	0	1
1	1	1
0	1	0

	Disp. Treat.	Disp. Imp.	Disp. Mist.
C_1	✗	✓	✓
C_2	✓	✗	✓
C_3	✓	✗	✗

This gives us an idea of another way we can measure the difference in discrimination. One question we could address is, "How many African Americans receive false 'High' recidivism predictors vs Caucasians?". Similar to their analysis, we can calculate the overall misclassification rate given certain parameters. This would allow us to maximize accuracy and fairness.

Method:



overall misclassification rate (OMR):

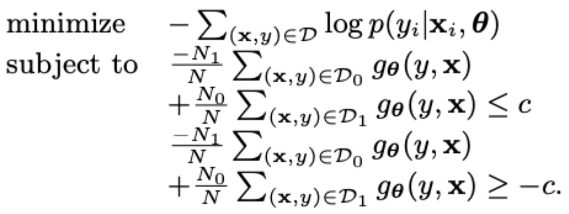
$$P(\hat{y} \neq y | z = 0) = P(\hat{y} \neq y | z = 1),$$

false positive rate (FPR):

$$P(\hat{y} \neq y | z = 0, y = -1) = P(\hat{y} \neq y | z = 1, y = -1),$$

false negative rate (FNR):

$$P(\hat{y} \neq y | z = 0, y = 1) = P(\hat{y} \neq y | z = 1, y = 1),$$



$$\begin{array}{ll} \text{minimize} & -\sum_{(\mathbf{x},y)\in\mathcal{D}} \log p(y_i|\mathbf{x}_i, \boldsymbol{\theta}) \\ \text{subject to} & \frac{-N_1}{N}\sum_{(\mathbf{x},y)\in\mathcal{D}_0} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \\ & +\frac{N_0}{N}\sum_{(\mathbf{x},y)\in\mathcal{D}_1} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \leq c \\ & \frac{-N_1}{N}\sum_{(\mathbf{x},y)\in\mathcal{D}_0} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \\ & +\frac{N_0}{N}\sum_{(\mathbf{x},y)\in\mathcal{D}_1} g_{\boldsymbol{\theta}}(y,\mathbf{x}) \geq -c. \end{array}$$

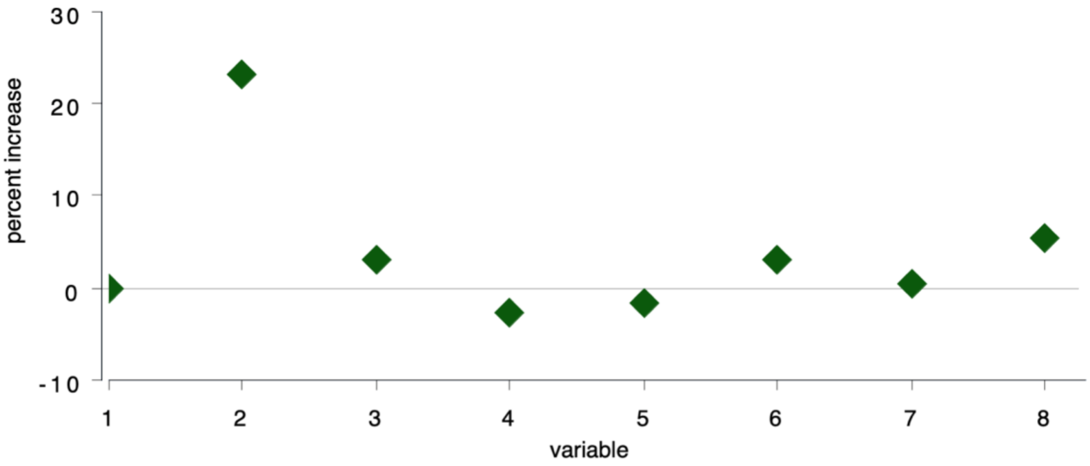
Name: RANDOM FORESTS

Link: randomforest2001.pdf (berkeley.edu)

Category: Fairness

Summary:

This article uses a method referred to as random trees which is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. To do this we need to find the accuracy in order to debias our dataset. We need to start by assigning each small subset of the data an “out-of-bag” classification then we would need to create the classification tree so that it corresponds to each variable. We would then run the subset through the classification tree and then compare the output to the class and return the change in misclassification. We can use this in our project to find which attributes have the most weight, then cross examine with a method to find which attributes carry bias to see where the majority of the algorithm's bias comes from. Below is an example given to us and it shows a single variable can carry more importance than any others regarding its accuracy.



Name: A peek into the black box: exploring classifiers by randomization

Link: A peek into the black box: exploring classifiers by randomization | SpringerLink

Category: Fairness

Summary:

The empirical investigation shows that the novel algorithm is indeed able to find groupings of interacting attributes exploited by the different classifiers. These groupings allow for finding similarities among classifiers for a single dataset as well as for determining the extent to which different classifiers exploit such interactions in general. As our focus is on reducing bias of course, we will need to also make sure we do not have unacceptable reductions in accuracy. This algorithm will allow us to make sure that any attributes we remove from our algorithm does not cause too much loss in accuracy.

Name: Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems

Link: datta-sen-zick-oakland16.pdf (cmu.edu)

Category: ML

Summary:

Algorithmic transparency is an emerging research area aimed at explaining decisions made by algorithmic systems. The call for algorithmic transparency has grown in intensity as public and private sector organizations increasingly use large volumes of personal information and complex data analytics systems for decision-making. As we go through our project, we try to solve questions so that we can query our data which will allow for data-driven questions to be answered. We can find out where the issues mainly lay and can uncover them behind the semantics within the data.

Name: Auditing Black-box Models for Indirect Influence

Link: auditing_icdm_2016.pdf (friedler.net)

Category: ML

Summary:

In this paper, we present a technique for auditing black-box models, which lets us study the extent to which existing models take advantage of particular features in the dataset, without knowing how the models work. Our work focuses on the problem of indirect influence: how some features might indirectly influence outcomes via other, related features. As a result, we can find attribute influences even in cases where, upon further direct examination of the model, the attribute is not referred to by the model at all. The issue of indirect influence is basically the core issue of algorithmic fairness in the criminal justice system. When we remove race as an attribute that does not mean we do not see its indirect influence throughout many other important attributes. When we use these strategies, we may be able to see which variables serve as proxies for problematic attributes such as race or gender.