Annotated Bibliography

Reference

Mirko Bagaric, Dan Hunter, and Nigel Stobbs, Erasing the Bias Against Using Artificial Intelligence to Predict Future Criminality: Algorithms are Color Blind and Never Tire, 88 U. Cin. L. Rev. 1037 (2020) Available at: https://scholarship.law.uc.edu/uclr/vol88/iss4/3

Category: Fairness

Summary

Authors of this article define recidivism as the act of taking measures that will reduce the length of prison terms for some offenders and consequently lower the number of inmates in federal prisons. This article focuses on three components in the criminal justice system where they believe the evaluation of future offending is relevant. These three components are sentencing, bail, and parole. The authors explain the purpose of these three systems and their roles in the criminal justice system.

Suggested Solutions

This article highlights the importance of monitoring the factors used in risk assessments. Each factor must be defined, expressed and identified. Each piece of evidence must be relevant to the sentencing process. Irrelevant factors should not be used in the calculation of sentencing decisions. All algorithms should be properly coded and transparent. People have shown to trust the use of an algorithm when they have seen how it works and how well it determines correct outcomes. One way algorithms could be transparent is by auditing and looking to the external inputs and outputs of the process of the decision. The article concludes by suggesting workings of risk assessments being

---

Reference

Picard, Sarah, et al. *Beyond the Algorithm: Pretrial Reform, Risk Assessment*. Formal Report . New York: Center for Court Innovation, 2019. Report. Link: https://www.courtinnovation.org/sites/default/files/media/documents/2019-06/beyond_the_algorithm.pdf

Category: Fairness

Summary

Beyond the Algorithm addresses risk assessments and their usefulness to society, while also addressing concerns over racial fairness in the criminal justice system. Their attempt isn't to argue for removing risk assessments, but to target the **"kinds or errors" made** by the algorithm. This article uses ProPublica's analysis of errors made by COMPAS as an

example of how risk algorithms could contain racial biases towards minority groups. ProPublica conducted their own study for research purposes to highlight the consequences of implementing risk assessment tools in the criminal justice system. Their assessment was based on nine risk factors:

- **Criminal History**
  - 1. Prior convictions
  - 2. Prior jail or prison sentence
  - 3. Prior failure to appear in court
  - 4. Probation status
- **Current Case Characteristics**
  - 5. Charge type
  - 6. Charge severity
  - 7. Concurrent open cases
- **Demographic Characteristics**
  - 8. Age
  - 9. Gender

Each feature is assigned a weight. The weights represent the number of risk points associated with each risk factor. Scoring placed the defendant in a risk category: minimal, low, moderate, moderate-high, or high-risk. Area under the curve (AUC) statistics revealed the predictive accuracy for Raw Risk score to be 0.745 and Risk Categories scores 0.733. These scores show successfulness by being over 70% accurate. Additionally, their scores were slightly better than COMPAS scores. However, many of its findings were similar to COMPAS in that there were higher false positive rates for African American and Hispanic defendants opposed to White defendants.

Suggested Solutions

To combat the racial disparities in algorithmic systems, they've found the most success in their *Risk-Based Approach* (RBA). This approach reduced the overall detention rate by nine percentage points when compared to another model known as *Business as Usual* (BAU). In the BAU approach, false positive rates were relatively high for all groups. The RBA approach reduced the overall false positive rate to less than 10 percent by improving the accuracy of pretrial detention decisions. Another successful approach was their *Hybrid Charge - and Risk-Based Approach*. This approach reduced the overall detention rates and lessened the racial disparities. Ultimately, they found that if pretrial detention was restricted only to defendants who are charged with violent crimes and who fall into higher-risk categories, such a policy may both reduce incarceration overall and alleviate racial disparities

Website: https://www.courtinnovation.org/about

Instructions on how to find article from website

- Areas of Focus
- Improving Decision-Making
- Risk Assessments

- Beyond the Algorithm: Pretrial Reform, Risk Assessment, and Racial Fairness (Publication)

Reference

Skeem, J, and C Lowenkamp. "Using Algorithms to Address Trade-Offs Inherent in Predicting Recidivism." *Behav Sci Law*, vol. 38, 2020, pp. 259–278., doi:https://doi.org/10.1002/bsl.2465.

Category: Fairness

Summary

Risk assessment algorithms evaluate the likelihood of recidivism by weighing risk factors such as criminal history and drug use. There are concerns that any benefits of reducing incarceration rates are offset by racial justice costs as these algorithms have displayed racial disparities in incarceration. This study compares how alternative strategies for "debiasing" risk assessment algorithms affect fairness trade-offs inherent in predicting violent recidivism when the rate of recidivism is unequal across races. Fairness in this study is viewed as balance in error rates for outcomes and balance in accuracy rates for predictions. ProPublica's investigation demonstrates that COMPAS, a risk assessment instrument, misclassified Black defendants twice as much as Whites as medium or high risk to reoffend which subsequently could result in harsher treatment of Black defendants. The algorithm displays bias where fairness is focused on a balance in error rates for outcomes.

Reference

Zavrsnik, Ales. "Algorithmic Justice: Algorithms and Big Data in Criminal Justice Settings." *European Journal of Criminology*, 2019, pp. 1–20., doi:10.1177/1477370819876762.

Summary

This article highlights the societal trend of 'algorithmic governance' reflected in the rising use of big data, algorithms, and machine learning in the criminal justice sector. The author provides a brief review of algorithmic governance and its growing use in policing and the courts (i.e., predictive policing and algorithmic justice). Through the use of examples and past research, the purposes of this paper are 1) to examine the problematic shift and heavy reliance on big data, algorithms, and machine learning in the criminal justice systems for understanding crime and how to address it and 2) demonstrate how algorithmic justice can conflict with civil liberties, legal doctrines, and criminal procedure laws.

In addressing the first goal of the paper, Zavrsnik discusses how the use of big data, algorithms, and machine learning are producing new knowledge, thresholds, and justifications to address crime that may impinge upon civil liberties, contradict concepts in criminal procedure rules, and allow abuses of power. Citing examples of the use of algorithms and machine learning to identify "terrorist sleepers" in Germany, track "persons of interest" through sentiment analysis (predicting future personal states/behavior) using a Twitter analysis tool in Slovenia during the Occu Movement, and track consumers in order to influence their spending habits, the author points out how people's civil rights and liberties are abused and blurring boundaries in the areas of security and criminal procedures.

In addressing the second purpose of the paper, the author addresses the risks and downfalls of algorithmic justice. First, the existence of false positives in predictive policing is discussed, Second, there are problems in creating reliable and valid databases and predictive algorithms if data used is unreliable and invalid and unintentional problems can arise even with the creator's best intentions. Third, creating good algorithms requires examining the purpose for its creation and whether it betters society. Fourth, there is a need to address runaway algorithms that result in unintended consequences such as racial discrimination in prison sentencing. Fifth, blurring probability causality and certainty when using predictive algorithms in judicial decisions. Sixth, Zavrsnik discusses the issues with machine bias and the argument for more empathic human judgment in the justice system and how de-biasing may not be desirable. Seventh, there are limitations algorithms that conflict with changes in judicial cases. Eighth, he discusses the pitfalls of algorithms as judicial cases change whereby algorithms may hinder judicial evolution. Ninth, relying on algorithms (even ones that produce errors) without human judgment is discussed. Lastly, human rights implications due to over policing and under policing as a result of algorithms and the rights of convicted persons to question the algorithms is explored.

AI Fairness 360 Demo

The AI Fairness 360 Demo provides users with a sample COMPAS dataset which allows us to explore areas where biases are located. Four mitigation strategies are provided. The successfulness of each strategy is measured based on five bias metrics.

1. **Sex and race are the focused attributes.**
    - Protected Attribute: Sex

      Privileged Group: ***Female***, Unprivileged Group: ***Male***

    - Protected Attribute: Race

      Privileged Group: ***Caucasian***, Unprivileged Group: ***Not Caucasian***

1. **Bias mitigation algorithms**

"The choice of which to use depends on whether you want to fix the data (pre-process), the classifier (in-process), or the predictions (post-process)"

Definitions

- **Reweighing** - Weights the examples in each (group, label) combination differently to ensure fairness before classification. (pre-process)

- **Optimized Pre-Processing** - Learns a probabilistic transformation that can modify the features and the labels in the training data. (pre-process)

- **Adversarial Debiasing** - Learns a classifier that maximizes prediction accuracy and simultaneously reduces an adversary's ability to determine the protected attribute from the predictions. This approach leads to a fair classifier as the predictions cannot carry any group discrimination information that the adversary can exploit. (in-process)

- **Reject Option Based Classification -** Changes predictions from a classifier to make them fairer. Provides favorable outcomes to unprivileged groups and unfavorable outcomes to privileged groups in a confidence band around the decision boundary with the highest uncertainty. (post-process).

1. There are 5 bias metrics: Statistical Parity Difference, Equal Opportunity Difference, Average Odds Difference, Disparate Impact, and Theil Index
1. Each Bias mitigation algorithm is compared to the original bias metrics found by COMPAS.

Relates to project

I recommend the statistics team take a close look at steps (3) and (4) to further see which algorithm provides a more balanced approach toward fairness. At this point in the project, we are still unsure which direction we want to head in. Mitigation in one area could also lead to unfairness in another. Becoming attentive to tradeoffs early on could help us as we move forward.

Accessing Demo

- Click link: https://aif360.mybluemix.net/
- "Try a Web Demo"
- Choose Compas (ProPublica recidivism) dataset
- Click "Next"

Reference

Pranay K. Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R. Varshney, Ruchir Puri. *BIAS MITIGATION POST-PROCESSING FOR INDIVIDUAL AND GROUP FAIRNESS*. Brighton, UK: Institute of Electrical and Electronics Engineers, 2018. Report. Link: https://arxiv.org/pdf/1812.06135.pdf

Category: Post-Processing

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.

Relates to project

We know that post-processing is one approach we can look upon as we think about how we can mitigate bias. This article provides empirical results and explains in detail the trade-offs between each of these algorithms. Majority of the trade-offs are mentioned on page 4. The statistics team should take a look at these.

Reference

Flavio P. Calmon, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. "Optimized-Data Pre-Processing for Discrimination Prevention." 2017. Report. Link: https://arxiv.org/pdf/1704.03354.pdf

Category: Pre-processing

Summary

Pre-processing is one of the three areas of focus when considering mitigating bias in algorithms. Many publications discuss several approaches (such as *disparate impact* in relation to group fairness, individual fairness, and equal error rates) towards making algorithmic systems fairer. This article proposes an optimization formula comprised of three goals: controlling discrimination (trade-offs), limiting distortion in individual data samples, and preserving utility. The distortion constraint included in this model distinguishes it from the previous approaches. Their goal is to determine a randomized

mapping that transforms the given dataset into a new dataset, which may be used to train a model, and similarly transforms data to which the model is applied, i.e., test data. The optimizations in this study utilizes a probabilistic framework for discrimination-prevention pre-processing in supervised learning (pages 2-9). Additionally, ProPublica COMPAS recidivism data is used in this dataset and compared to their results (page 9-12).

Relates to project

This study has shown decreases in recidivism for certain groups. Looking forward, if we decide to take the pre-processing approach towards the project, we could possibly draw ideas from here. The mathematical steps are included in this article. The Domain and Statistics team can continue to read over these steps and make sense of it. It could help the Development team to view this article as well to see if the methods used here are applicable for us.

Reference

Hardt, Moritz, Eric Price and Nathan Srebro. "Equality of Opportunity in Supervised Learning." 11 October 2016. Pages 1 - 22. Formal Report. <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 begins 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.

Reference

Lohia, Pranay . "PRIORITY-BASED POST-PROCESSING BIAS MITIGATION FOR INDIVIDUAL AND GROUP." AI, India: IBM Research , 31 January 2021. Report. 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 follows: 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 groups 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 unbiased. 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 was 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.

Reference:

Feldman, Michael, et al. "Certifying and Removing Disparate Impact." *Certifying and Removing Disparate Impact*. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 16 July 2015. Pages 259–268. Formal Report. <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 tests? 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 process 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.

Reference:

Ruggieri, Salvatore, Dino Pedreschi and Franco Turini. "Data Mining for Discrimination Discovery." *Data Mining for Discrimination Discovery*. ACM Transactions on Knowledge Discovery from Data, May 2010. Pages 1 - 49. Formal Report.

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:

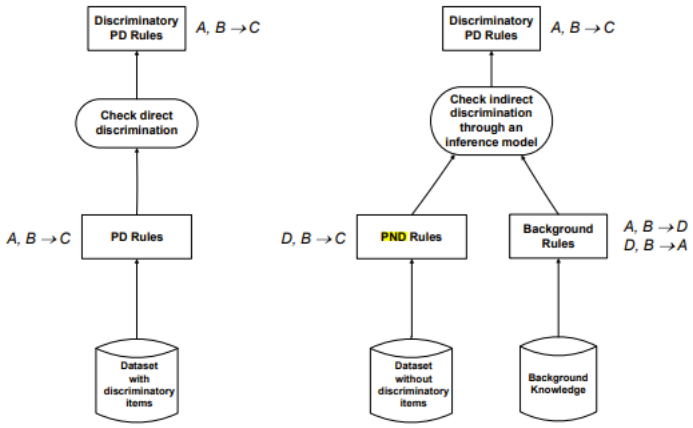


Fig. 1. Modelling the process of direct (left) and indirect (right) discrimination analysis.

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.

Reference:

Kamiran, Faisal and Toon Calders. "Data preprocessing techniques for classification." *Data preprocessing techniques for classification*. Springer, 3 December 2011. Paper. 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 in risk-scores based on gender, age, or race.

Reference:

Stewart, Matthew. "Handling Discriminatory Biases in Data for Machine Learning." Towards Data Science, 30 March 2019. Post. <https://towardsdatascience.com/machine-learning-and-discrimination-2ed1a8b01038>.

Category: ML

Summary:

This article provides an overview of the COMPAS algorithm and summarizes some of the analysis found by ProPublica. It reveals stories of those affected 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.

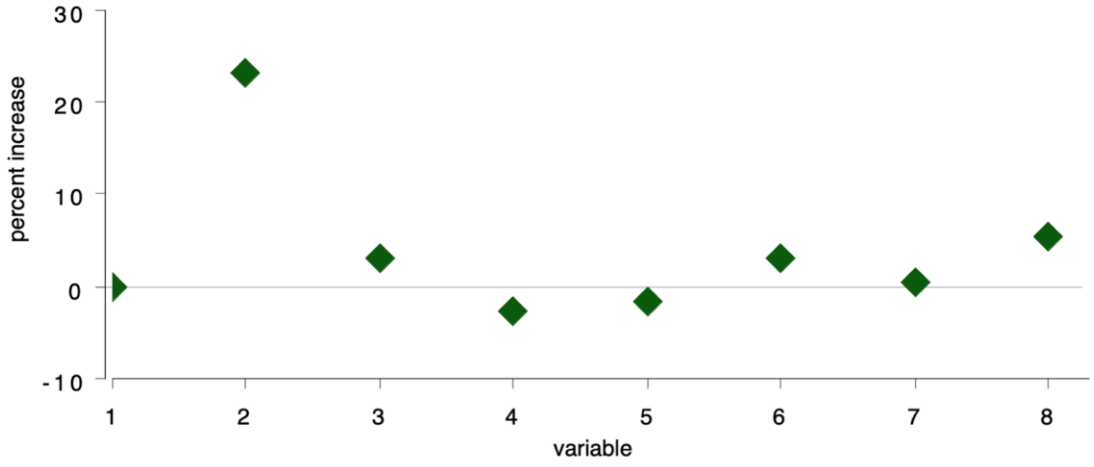
Reference:

Breiman, Leo. “Breiman's Idea for Testing Classifiers.” *Random Forests,* Jan. 2001, www.stat.berkeley.edu/~breiman/randomforest2001.pdf.

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.



Reference:

Henelius, A., Puolamäki, K., Boström, H. *et al.* A peek into the black box: exploring classifiers by randomization. *Data Min Knowl Disc* 28, 1503–1529 (2014). https://doi.org/10.1007/s10618-014-0368-8.

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.

Reference:

Datta, Anupam, et al. "Algorithmic Transparency via Quantitative Input Influence." *Theory and Experiments with Learning Systems*, Carnegie Mellon University, Pittsburgh, USA, www.andrew.cmu.edu/user/danupam/datta-sen-zick-oakland16.pdf.

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.

Reference:

Adler, Philip, and Casey Falk. "Auditing Black-Box Models for Indirect Influence." *Auditing Black-Box Models*, sorelle.friedler.net/papers/auditing_icdm_2016.pdf.

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.

Reference:

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, Cynthia Dwork; Proceedings of the 30th International Conference on Machine Learning, PMLR 28(3):325-333, 2013.

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

$\mathbf{X}^+$ and $\mathbf{X}^-$ - represents sub-groups.

Z - represents a random variable.

K – represents a set of prototypes.

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

$\hat{\mathbf{y}}_\mathbf{n}$ - 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.

Reference:

Fish, Benjamin et al. "A Confidence-Based Approach for Balancing Fairness and Accuracy." *SDM* (2016).

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.

Reference:

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In Proceedings of the 26th International Conference on World Wide Web (WWW '17). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1171–1180. DOI:https://doi.org/10.1145/3038912.3052660

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			Ground Truth (Has Weapon)
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	✓

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):

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

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