INDS 4997 Capstone in Data Science Course

Mitigation Methodology Document

Types of Post-processing Mitigation Algorithms

Calibrated Equalized Odds Post-processing

Use to mitigate bias in predictions. Optimizes over calibrated classifier score outputs that lead to fair output labels.

Equalized Odds Post-processing

Use to mitigate bias in predictions. Modifies the predicted labels using an optimization scheme to make predictions fairer.

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).

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: The interaction between classification and reject performance for distance-based reject-option classifiers

Link: https://www.sciencedirect.com/science/article/pii/S0167865505003089

Category: Post-processing | ROC

Summary:

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

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

Category: Post-processing

Summary:

Name: Certifying and removing disparate impact

Link: [1412.3756] Certifying and removing disparate impact (arxiv.org)

Category:

Summary:

Name: Data Mining for Discrimination Discovery

Link: tkdd.pdf (unipi.it)

Category:

Summary:

Name: Data preprocessing techniques for classification without discrimination

Link: [PDF] Data preprocessing techniques for classification without discrimination | Semantic Scholar

Category:

Summary:

Name: Handling Discriminatory Biases in Data for Machine Learning

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

Category:

Summary: