YOU ARE NOT ALLOWED TO SHARE THE CONTENT WITH OTHERS OR DISSEMINATE THE CONTENT

NUS CS-CS5562: Trustworthy Machine Learning October 25, 2023

Assignment 5

Lecturer: Reza Shokri

Introduction

The objective of this assignment is to understand common ways of measuring and mitigating bias in machine learning as well as long term effects of (fair) machine learning. This assignment contains the following 5 parts:

1. **On the Neutrality of Data** in which we implement common fairness metrics and use them to inspect some common misconceptions about the inherent neutrality of data driven decision making.
2. **Separation vs. Calibration** in which we show that in general separation and calibration cannot be satisfied simultaneously.
3. **Post-processing algorithms** in which we learn how to develop a post-processing algorithm to debias an existing classifier.
4. **Negative Feedback Loops** in which we explore how feedback loops can affect existing bias.
5. **Delayed Impact of Fair Machine Learning** in which we learn about the impact that enforcing fairness metrics can have.

Tasks 1, 3, 4, and 5 come with a jupyter notebook, you will need to implement the code in each notebook and write a report about the tasks. Details about the exact items to be reported are given in the corresponding task descriptions in this document.

The notebooks were tested using GoogleColab and we recommend using GoogleColab to solve them to avoid errors.

- You need to use the LaTeX template we provided in the `report/report.tex`.
- **The report should ONLY contain EIGHT pages at most. Anything that exceeds eight pages will be ignored.** See the report template for further details on what to include.
- You are required to submit the completed notebooks, your LaTeX file(s) and the compiled PDF file for the report (name it `report.pdf` and keep it in the `report` folder).

1 On the Neutrality of Data

A general conception among many parts of society and many data scientists is that data-driven decision-making is inherently neutral. Many see machine learning algorithms as a way to overcome human biases in decision-making. This section explores this notion.

Intended Learning outcomes

After finishing this task you should be able to:

1. implement common observational fairness metrics and apply them to detect bias in a machine learning model,
2. argue whether data driven decision making is inherently neutral,
3. discuss the concept of "fairness through unawareness", and
4. inspect a classifier to detect variables that encode bias.

1.1 Fairness definitions

In this task we consider *demographic parity*, *equalized odds* and *predictive parity* as three examples of demographic fairness notions. For a broader overview of these notions see Chapter 2 of Barocas et al. [2019] or this tutorial.

For simplicity, we consider only binary classification and the case of two groups. Also, note that usually, fairness is a to-be-satisfied constraint. Yet, for this assignment, we are interested in measuring the degree of violation of each criterion. Hence, our fairness notions return numbers in $\mathbb{R}$, where 0 indicates that the classifier is fair. A positive number shows that Group A has an advantage and a negative number that Group B is advantaged.

Demographic parity, also known as independence, or statistical parity, requires that the outcome of a classifier is independent of the group membership.

Definition 1 (Demographic disparity) *Given two disjoint groups $A, B \subseteq \mathbb{R}^n$ and a distribution of points $\mathcal{D}$ over $\mathbb{R}^n$ we define the* demographic disparity *of a binary classifier $f\colon R^n \rightarrow \{0,1\}$ as:*

$$DP(f, \mathcal{D}) := \mathbb{P}_{x \sim \mathcal{D}}[f(x) = 1 | x \in A] - \mathbb{P}_{x \sim \mathcal{D}}[f(x) = 1 | x \in B]$$

Equalized odds Hardt et al. [2016] allows for a dependence of the predicted label on the sensitive attribute, but only as far as the true label justifies.

Definition 2 (Equalized odds violation) *Given two groups $A, B \subseteq \mathbb{R}^n$ and a distribution of labeled points $\mathcal{D}$ over $\mathbb{R}^n \times \{0,1\}$ we measure the amount a binary classifier $f\colon R^n \rightarrow \{0,1\}$ violates* equalized odds *as:*

$$EO(f, \mathcal{D}) := max_{\hat{y} \in \{0,1\}} \left| \left[\mathbb{P}_{(x,y) \sim \mathcal{D}}[f(x) = 1 | x \in A, y = \hat{y}] - \mathbb{P}_{(x,y) \sim \mathcal{D}}[f(x) = 1 | x \in B, y = \hat{y}] \right] \right|$$

The final notion, predictive parity, or sufficiency, notes that given the predicted outcome, the group membership is independent of the actual outcome.

Definition 3 (Predictive disparity) *Given two groups $A, B \subseteq \mathbb{R}^n$ and a distribution of labeled points $\mathcal{D}$ over $\mathbb{R}^n \times \{0,1\}$ we measure the* predictive disparity *of a binary classifier $f\colon R^n \rightarrow \{0,1\}$ as:*

$$PredPar(f, \mathcal{D}) := max_{\hat{y} \in \{0,1\}} \left| \left[\mathbb{P}_{(x,y) \sim \mathcal{D}}[y = 1 | x \in A, f(x) = \hat{y}] - \mathbb{P}_{(x,y) \sim \mathcal{D}}[y = 1 | x \in B, f(x) = \hat{y}] \right] \right|$$

Note on absolute values: In Definitions 2 and 3 , we use absolute values to make the expressions more succinct. However, you should implement them such that positive values indicate an advantage for Group A and negative values indicate an advantage for Group B.

Note on the names: By now, you will have noticed we use many different names for each fairness notion we mention. Unfortunately, this is a reflection of the fairness literature. We use (and keep repeating) these names to prepare you for reading the different research papers. If you ever come across a new-sounding definition, one good resource to check is the *Dictionary of criteria* at the end of Chapter 3 in Barocas et al. [2019] which provides an overview of all these names.

1.2 Warm up

Use the provided in `Assignment05_Task01.ipynb` to download the adult dataset and train a decision tree. Follow the instructions given on preprocessing and parameter choice in the code. Report the overall accuracy on the test set.

1.3 Measuring fairness

Implement the functions to measure the *demographic (dis)parity*, *equalized odds* and *predictive parity* of a binary classifier (based on the templates provided in the notebook). Report the metrics for the attribute 'sex' on the test set of the classifier trained in the previous task.

1.4 Inherent neutrality of data

Given the results from above and accepting demographic disparity as an appropriate metric, discuss whether the statement "Data-driven decision making is inherently fair" is correct. Give reasons why it might or might not hold. Give some potential reasons for the (un)fair behavior of the classifier.

1.5 Fairness through unawareness

Another common assumption is that just removing the sensitive attribute will lead to a fair classifier. Let's explore this assumption as well. Remove the 'sex' attribute from the training data of the adult dataset and repeat the training of the classifier from Task 1.2. Report the resulting accuracy and fairness metrics from Task 1.3.

1.6 Removing an additional attribute

Inspect the classifier you created in the previous task and all the attributes in the training set. Understand which variables are used by the classifier and what they mean. Find the variable that is the most likely culprit for the still existing bias. Remove this variable from training and create a third classifier. Report the resulting accuracy and fairness metrics from Task 1.3. Succinctly describe your approach and your observations. Describe some potential pitfalls of the fairness through unawareness approach.

2 Separation vs Sufficiency

Given the different fairness notions and their trade-offs, a natural solution would be to satisfy multiple fairness conditions at the same time. So, you would gain the benefits of all and avoid their shortcomings. Unfortunately, this is not possible. Except for some corner cases the three fairness notions we introduced in Section 1.1 are pairwise incompatible Barocas et al. [2019]. In this exercise, you'll show one of these incompatibilities.

Intended learning outcome

After finishing this task you should be

1. Able to use standard results of probability theory to show incompatibility of statistical fairness criteria

Prove the following theorem:

Theorem 4 *Let X, Y, A be random variables, , if there exists a function $R = r(x)$ such that $R \perp A|Y$ and $Y \perp A|R$ then*

$$A \perp Y.$$

3 Post-processing algorithms

Intended Learning outcomes

After finishing this task you should be able to:

1. Develop and implement the optimization problem to enforce fairness constraints on an existing classifier.
2. Study the trade-off between accuracy and strictly enforcing a fairness criteria.
3. Discuss some inherent limitations in post-processing algorithms.

Given a classifier $f : R^n \rightarrow \{0, 1\}$ that violates an observational fairness criterion, we can construct a new classifier f' using a post-processing algorithm such that f' satisfies the fairness criterion. For this, following the approach outlined in Hardt et al. [2016], f is treated as a black-box and f' has only access to the predicted label and the sensitive attribute. Let $p_{y,a} = \mathbb{P}[f' = 1|f = y, A = a]$ denote the probability* that f' predicts label 1 conditioned on f predicting $y \in \{0, 1\}$ and the point having the sensitive attribute $a \in \{0, 1\}$. f''s behavior can be completely captured in terms of f's behavior and the four variables: $p_{1,0}$, $p_{0,0}$, $p_{1,1}$ and $p_{0,1}$.

1. Write down the constraints f' needs to satisfy in order to satisfy exact equalized odds. These constraints should be linear in $p_{0,0}$, $p_{0,1}$, $p_{1,0}$ and $p_{1,1}$ and may contain probabilities over f, Y, and A.
2. Express the accuracy of f' linearly in terms of $p_{0,0}$, $p_{0,1}$, $p_{1,0}$ and $p_{1,1}$. You may also use probabilities over f, Y, and A.
3. Use the results of the previous two steps to implement `Postprocessing_equalized_odds` a class given in `Assignment05_Task03.ipynb`.

*In this exercise we use the training data to calculate these values.

4. Train a simple classifier using the code in `Assignment05_Task03.ipynb`. Then use your post-processing algorithm on this classifier to create a classifier that satisfies equalized odds.

5. It is not always necessary to perfectly satisfy a fairness criterion, and in some settings it might come at a considerable cost to accuracy. Given $\lambda \in [0, 1]$, *relaxed equalized odds* is satisfied if:

   $$|\mathbb{P}[f' = 1|Y = y, A = 0] - \mathbb{P}[f' = 1|Y = y, A = 1]| \leq \lambda \qquad y \in \{0, 1\}$$

   Develop an algorithm `Postprocessing_relaxed_equalized_odds` in which the user can decide to which extent the fairness criterion needs to be satisfied. Report the accuracy for $\lambda \in [0, 0.01, 0.02, \ldots, 0.1]$.

6. Discuss potential limitations of post-processing algorithms.

4 Negative feedback loops

Intended learning outcomes

After finishing this task you should be

1. Aware that feedback loops can lead to biased behavior
2. Able to simulate a feedback loop based on a simple urn model
3. Able to discuss the influence of prior beliefs in feedback loops
4. Design an intervention to avoid outsized influence of the prior

So far, we have considered classifiers that are trained once and then deployed. In practice, a classifier is often updated based on the data obtained while the classifier is in deployment. These updates can lead to problematic feedback loops Ensign et al. [2018].

Consider the following toy scenario. We have a police precinct overseeing two neighborhoods with one police car. Each day the police can decide in which of the neighborhoods to send the car. Given the true crime rates λ_A, λ_B the probability to observe a crime is as follows:

	Crime observed in A	Crime observed in B
Car in A	$w_d\lambda_A + w_r\lambda_A$	$w_r\lambda_B$
Car in B	$w_r\lambda_A$	$w_d\lambda_B + w_r\lambda_B$,

where w_d, w_r are the amounts of crime discovered by the police and reported by the community respectively ($w_d + w_r = 1$). Every day the police decide to go to one of the two neighborhoods. The probability to go to a neighborhood is proportional to the number of crimes observed in the neighborhood so far. An observed crime is added to the historical records. We are interested in the following scenarios:

1. Both neighborhoods have the same crime rate ($\lambda = 0.1$), none of the crimes are reported to the police ($w_r = 0$) the historic records show 10 crimes in neighborhood A and 10 crimes in neighborhood B.

2. Both neighborhoods have the same crime rate ($\lambda = 0.1$), none of the crimes are reported to the police ($w_r = 0$) the historic records show 15 crimes in neighborhood A and 5 crimes in neighborhood B.

3. Neighborhood B has a slightly higher crime rate ($\lambda_A = 0.1, \lambda_B = 0.11$), none of the crimes are reported to the police ($w_r = 0$) the historic records show 10 crimes in neighborhood A and 10 crimes in neighborhood B.

4. Neighborhood B has a slightly higher crime rate ($\lambda_A = 0.1, \lambda_B = 0.11$), none of the crimes are reported to the police ($w_r = 0$) the historic records show 15 crimes in neighborhood A and 5 crimes in neighborhood B.

5. Neighborhood B has a slightly higher crime rate ($\lambda_A = 0.1, \lambda_B = 0.11$), most of the crimes are reported to the police ($w_r = 0.9$) the historic records show 15 crimes in neighborhood A and 5 crimes in neighborhood B.

Implement a simulator with the code given in `Assignment05_Task04.ipynb` and

1. estimate for any of the above scenarios the probability that after 10 years (i.e., 3650 days), the police's belief about the ratio of crime rates is within ±0.1 of the true ratio ($\frac{\lambda_A}{\lambda_A+\lambda_B}$),

2. discuss the influence the initial belief has on the final belief of the police,

3. design an intervention that improves the chance that the polices belief is correct. Note, for this intervention, the police can only change its own behavior (i.e., how it decides which neighborhood to visit or what crimes to record.) It cannot change historic records, underlying crime rates, or community reporting

behavior. Describe the approach and report the new probability for the second scenario.

For simplicity you can assume that 10000 runs are enough to get a valid estimate of the probabilities (this might take a couple of minutes to run).

5 Delayed impact of fair machine learning

Intended learning outcomes

After finishing this task you should be

1. Able to measure the impact of different classification policies on the long term welfare of groups in a simple population model
2. Aware that enforcing fairness metric can have negative long term impacts
3. Able to prove the potential of fairness metrics to cause harm in the given population model

We consider two *groups* A and B, which comprise a g_{A} and $g_{\mathsf{B}} = 1 - g_{\mathsf{A}}$ fraction of the total population, and a *bank* which makes a binary decision to approve a loan for each individual in each group. Individuals in each group are assigned *scores* in $\mathcal{X} := \{1, 2, \ldots, C\}$, and the scores for group $j \in \{\mathsf{A}, \mathsf{B}\}$ are distributed according $\pi_j \in \text{Simplex}^{C-1}$. Let $\rho \colon \mathcal{X} \to [0, 1]$ be a function that determines the probability that an individual with score x succeeds to pay the loan back. The bank selects a *policy* $\tau := (\tau_{\mathsf{A}}, \tau_{\mathsf{B}}) \in [C] \times [C]$, where τ_j corresponds to the threshold above which people in a group are accepted for a loan. One should think of a score as an abstract quantity which summarizes how well an individual is suited to being selected.[†]

The bank is utility-maximizing (potentially under fairness constraints). The expected utility to the bank is given by the expected return from a loan, which we model

[†] This is a slightly simplified from the setting first discussed in Liu et al. [2018]

as an affine function of $\rho(X): u(x) = u_+\rho(x) + u_-(1-\rho(x))$, where u_+ denotes the profit when a loan is repaid and u_- the loss when someone defaults on a loan (i.e., $u_+ > 0 > u_-$). The bank's expected utility for a policy τ is given by

$$\mathcal{U}(\tau) = \sum_{j\in\{\mathsf{A},\mathsf{B}\}} g_j \sum_{x\in\mathcal{X},\tau_j\leq x} \pi_j(x)u(x).$$

This exercise focuses on the effect the selection policy τ has on the groups A and B. Individual outcomes of being granted a loan are based on the probability of repaying the loan. Similar to the utility of the bank we can define the change in well-being of the individual as $\Delta(x) = c_+\rho(x) + c_-(1-\rho(x))$. The constant c_+ denotes the gain in credit score if the loans are repaid and c_- the score penalty in the case of default (i.e., $c_+ > 0 > c_-$). The average change of the mean score μ_j for group j is defined as

$$\Delta\mu_j(\tau) = \sum_{x\in\mathcal{X},\tau_j\leq x} \pi_j(x)\Delta(x).$$

The file `transrisk_performance_by_race_ssa.csv` contains the transaction risk (i.e., the risk that an individual with this score defaults on a loan) for each score (0-100) by race. Assume that $c_- = -150$ and $c_+ = 75$ and that the population proportions between "Black" and "Non-Hispanic white" is 18% to 82%. We ignore the remaining groups.

1. In this question, we consider another fairness definition: **equal opportunity**. This is a relaxation of equalized odds that only considers the true positive rates across protected groups. To satisfy equal opportunity, we require:

   $$\mathbb{P}_{(x,y)\sim\mathcal{D}}[f(x)=1|x\in A, y=1] = \mathbb{P}_{(x,y)\sim\mathcal{D}}[f(x)=1|x\in B, y=1]$$

   Write down the constraints for a policy to satisfy *demographic parity* and *equal opportunity* in this setting.

2. For a loss profit ratio of $\frac{u_-}{u_+} = -10$ use the code provided in `Assignment05_Task04.ipynb` to compare the average change of the mean score of the two groups under the following conditions:

(a) The bank uses a policy that optimizes their utility (`maxUtil`).

(b) The bank uses a policy that optimizes their utility while satisfying *Demographic parity*

(c) The bank uses a policy that optimizes their utility while satisfying *Equal opportunity*

Note: For the theoretical considerations, we define ρ as independent from the group (i.e., once you know the score of an individual, its success probability is independent of its group membership). However, the dataset contains group-specific risk scores. Hence, you should use group-specific risk scores for the experiments.

Now we will study some theoretical properties of the maxUtil policy and the effect of fairness constraints. For this we make some simplifying assumptions:

(i) $\mathcal{X} = [0, C]$ and all other definitions are updated accordingly to a continious space (e.g., $\sum_{x \in \mathcal{X}} \to \int_{x \in \mathcal{X}}$).

(ii) π_j is continuous and non-zero on $\mathcal{X}$.

(iii) $\forall x \in \mathcal{X}\colon \rho_A(x) = \rho_B(x)$

(iv) $\rho(x)$ is non-decreasing

For a selection rate $\beta \in [0, 1]$ let $r_j^{-1}(\beta)$ be the threshold value (or policy) that implies the selection rate.

3. Let A be the underprivileged group, $\tau^*_{\mathsf{A}}, \beta^*_{\mathsf{A}}$ be the optimal policy and selection rate for group A (according to $\Delta\mu_{\mathsf{A}}$). Assume that $u(x) \geq 0 \Rightarrow \Delta(x) \geq 0$. Show that $0 \leq \Delta\mu_{\mathsf{A}}(\tau^{\text{maxUtil}}_{\mathsf{A}}) \leq \Delta\mu_{\mathsf{A}}(\tau^*_{\mathsf{A}})$. That means that the policy cannot cause active harm.

4. Let $\beta \in [0, 1]$ with $\beta^{\texttt{maxUtil}}_{\mathsf{B}} > \beta > \beta^{\texttt{maxUtil}}_{\mathsf{A}}$ be fixed, show that there exists a population proportion g_0 such that, for all $g_{\mathsf{A}} \in [0, g_0], \beta^{\texttt{DP}}_{A} > \beta$. In particular

if $\Delta\mu_{\mathsf{A}}(r^{-1}_{\pi_{\mathsf{A}}}(\beta)) = 0$ demographic parity causes active harm (i.e. reducing the mean score of group A.).

Note: You can assume without proof that the banks overall utility in the constraint problem is strongly concave in the selection rate and

$$\frac{\partial r^{-1}_{\pi_j}(\beta)}{\partial \beta} = \frac{-1}{\pi_j(r^{-1}_{\pi_j}(\beta))}.$$

5. Use the code provided in `Assignment05_Task05.ipynb` to find a hypothetical g_0 such that demographic parity causes active harm to group A (i.e., $\Delta\mu_{\mathsf{A}} < 0$).

References

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. In *Conference on Fairness, Accountability and Transparency*, pages 160–171. PMLR, 2018.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.