EXPLAINABILITY FOR FAIR MACHINE LEARNING

Tom Begley, Tobias Schwedes, Christopher Frye, & Ilya Feige

Presented by:
Yicong Li, Ziwei Gu, Hongwen Song
Fair Machine Learning
Fair Machine Learning - Challenges

- **Defining fairness** is hard:
  - Many competing definitions, often incompatible with each other
    - statistics-based, causal-reasoning-based, etc.
  - focus on group outcomes vs. individual outcomes
  - Requires contextual understanding

| Name                          | Closest relative | Note       |
|-------------------------------|------------------|------------|
| Statistical parity           | Independence    | Equivalent |
| Group fairness                | Independence    | Equivalent |
| Demographic parity            | Independence    | Equivalent |
| Conditional statistical parity| Independence    | Relaxation |
| Equal opportunity             | Separation       | Relaxation |
| Equalized odds                | Separation       | Equivalent |
| Conditional procedure accuracy equality | Separation       | Equivalent |
| Disparate mistreatment        | Separation       | Equivalent |
| Balance for positive class    | Separation       | Relaxation |
| Balance for negative class    | Separation       | Relaxation |
| Predictive equality           | Separation       | Relaxation |
| Conditional use accuracy equality | Suficiency       | Equivalence |
| Predictive parity             | Suficiency       | Relaxation |
| Calibration                   | Suficiency       | Equivalence |
Fairness definitions - Example

- demographic parity: “f(x) be unconditionally independent of sensitive attr. a.”

If 100 female students and 100 male students apply to Harvard University, demographic parity is achieved if the percentage of female students admitted is the same as the percentage of male students admitted, irrespective of whether one group is on average more qualified than the other.
Fairness definitions - Example

• **equalized odds**: “f(x) be independent of sensitive attr. a given y”

If 100 female students and 100 male students apply to Harvard University, equalized odds is achieved if qualified female and male students both have the same chance of being admitted, and unqualified female and male students have the same chance of being rejected.

**Female students**

|       | Qualified | Unqualified |
|-------|-----------|-------------|
| Admitted | 45        | 2           |
| Rejected | 45        | 8           |
| Total   | 90        | 10          |

**Male students**

|       | Qualified | Unqualified |
|-------|-----------|-------------|
| Admitted | 5         | 18          |
| Rejected | 5         | 72          |
| Total   | 10        | 90          |
Fair Machine Learning - Challenges

- Explanation methods can be manipulated
  
  - [You shouldn’t trust me: Learning models which conceal unfairness from multiple explanation methods, Dimanov, ECAI, 2020]

Demonstrates that an explanation attack can easily mask a model’s discriminatory use of a sensitive feature without hurting accuracy [Dimanov, 2020]

Importance ranking histograms for gender as the sensitive feature on the adult test set of the original (left) and modified (right) models.
Fair Machine Learning - Proposed Solution

• A **unified** approach that works for many **group-fairness** criteria
  • demographic parity, equalised odds, conditional demographic parity
  • for each definition, choose Shapley value functions that attribute overall fairness to individual features.

• **Cannot hide unfairness** by manipulating explanations
  • Fairness Shapley values collectively must sum to the chosen fairness metric
Methodology
Explaining Model Accuracy

- If the team earns a total value $v(N)$, the Shapley value $\phi_v (i)$ attributes a portion to player $i$ according to:

$$\phi_v (i) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! (n - |S| - 1)!}{n!} [v(S \cup \{i\}) - v(S)]$$ (1)

- Binary classification problem:

$$f_y (x) = (1 - y)(1 - f(x)) + y f(x)$$ (2)

- Define a value function by marginalising over out-of-coalition features:

$$v_{f_y} (x) (S) = \mathbb{E}_{p(x') [f_y(x_{S \cup x' \setminus N \setminus S})$$ (3)

- Global explanation of the model’s performance:

$$\Phi_f (i) = \mathbb{E}_{p(x,y)} [\phi_{f_y} (x)(i)]$$ (4)

- Aggregating global Shapley values

$$\sum_i \Phi_f (i) = \mathbb{E}_{p(x,y)} [f_y (x)] - \mathbb{E}_{p(x')p(y)} [f_y (x')]$$ (5)
Methodology

• Explainable Fairness:
  • How Shapley value paradigm can be adapted to explain fairness.

• Meta Algorithm:
  • Motivation: axiomatic properties of Shapley values
  • Applying existing training-time fairness interventions, wherein one trains a perturbation to the original model, rather than a new model entirely.
Explaining Model Fairness

• To explain fairness in a model’s decisions, they define a new value function that captures this effect

\[ g_a(x) = f(x) \cdot \frac{(-1)^a}{p(a)} \quad \text{a: sensitive attribute} \quad (6) \]

Demographic parity calls for \( f(x) \) to be unconditionally independent of \( a \)

• The value function on coalitions is defined through marginalisation:

\[ v_{g_a}(x)(S) = \mathbb{E}_{p(x')} [g_a(x_{\mathcal{S}} \sqcup x'_{N\setminus\mathcal{S}})] \quad (7) \]

\[ \Phi_g(i) = \mathbb{E}_{p(x,a)} [\phi_{g_a}(x)(i)] \quad (8) \]

The joint distribution of features and protected attribute from which the data is sampled

\[ \sum_i \Phi_g(i) = \int dx \ p(x|a = 0) f(x) - \int dx \ p(x|a = 1) f(x) \quad (9) \]

Each feature’s marginal contribution to the overall demographic disparity in the model
Learning Corrective Perturbations

• The linearity axiom of the Shapley values guarantees that the fairness Shapley values of a linear ensemble of models are the corresponding linear combination of Shapley values of the underlying models.

• Motivated by this they consider the problem of learning an additive perturbation to an existing model in order to impose fairness.

\[ f_\theta = f + \delta_\theta \]

\[ \delta_\theta(f(x), x, a) = \sigma\left(\sigma^{-1}(f(x)) + \tilde{\delta}_\theta(f(x), x, a)\right) - f(x). \]

Any training-time fairness algorithm used to learn the auxiliary perturbation e.g. Agarwal et al. (2018) and Zhang et al. (2018)

The key idea is to reduce fair classification to a sequence of cost-sensitive classification problems, whose solutions yield a randomized classifier with the lowest (empirical) error subject to the desired constraints.
Experiments & Results
Datasets

1. Adult dataset - UCI Machine Learning Repository (Dua & Graff, 2017)

   *Task: predict whether an individual earns more than $50K per year based on their demographics*

1. COMPAS recidivism dataset (Larson et al., 2016)

   *Task: predict recidivism risk based on demographics*
Explainability

- Marital Status
- Sex
- Relationship
Robustness of Fairness Explanation

- Suppressing the importance of the sex feature.

- Demographic Parity Difference: 0.193 -> 0.184
Learnt Perturbations - Performance

- No significant reduction under the fairness definition of demographic parity.

| Method                        | Accuracy [%] at demographic parity difference |
|-------------------------------|-----------------------------------------------|
|                               | 0.1 | 0.08 | 0.06 | 0.04 | 0.02 | 0.01 | 0.005 |
| **Adult**                     |     |      |      |      |      |      |       |
| Agarwal et al.                | 84.71 | 84.32 | 83.94 | 83.82 | 83.29 | 83.29 |      |
| Agarwal et al. - perturbed    | 84.69 | 84.43 | 83.82 | 83.82 | 83.35 | 83.23 |      |
| Zhang et al.                  | 84.65 | 84.18 | 84.06 | 83.58 | 83.18 | 83.15 | 83.15 |
| Zhang et al. - perturbed      | 84.74 | 84.48 | 83.78 | 83.61 | 83.14 | 82.99 | 82.96 |
| Feldman et al. (post)         | 84.69 | 84.35 | 84.12 | 83.67 | 83.32 | 83.30 | 83.01 |
| **COMPAS**                    |     |      |      |      |      |      |       |
| Agarwal et al.                | 74.05 | 74.05 | 73.77 | 73.67 | 73.11 | 73.11 | 73.01 |
| Agarwal et al. - perturbed    | 74.24 | 74.24 | 73.86 | 73.86 | 73.20 | 72.73 | 72.73 |
| Zhang et al.                  | 75.19 | 75.19 | 75.19 | 74.62 | 74.15 | 74.15 | 74.15 |
| Zhang et al. - perturbed      | 74.24 | 74.24 | 74.24 | 73.30 | 73.30 | 73.20 | 72.73 |
| Feldman et al. (post)         | 74.81 | 74.81 | 74.81 | 74.24 | 74.24 | 73.20 | 72.35 |
Learnt Perturbations - Performance

No significant reduction under the fairness definition of equalised odds.

Table 2: Accuracy associated with decreasing equalised odds thresholds.

| Method                | Adult Accuracy [\%] at equalised odds difference |
|-----------------------|-------------------------------------------------|
|                       | 0.1  | 0.08 | 0.06 | 0.04 | 0.02 | 0.01 | 0.005 |
| Agarwal et al.        | 85.32 | 85.32 | 85.13 | 84.30 | 84.18 | -    | -     |
| Zhang et al.          | 85.13 | 85.04 | 85.04 | 84.86 | 84.33 | 75.43 | 75.43 |
| Zhang et al. - perturbed | 85.26 | 85.11 | 85.06 | 84.97 | 84.23 | 83.53 | -     |
| Hardt et al.          | 82.77 | 82.77 | 82.77 | 82.77 | 82.77 | 82.77 | 82.77 |
| COMPAS                |       |       |       |       |       |       |       |
| Agarwal et al.        | 75.19 | 75.19 | 74.05 | 74.05 | 73.86 | 73.39 | 73.39 |
| Zhang et al.          | 74.62 | 74.62 | 74.62 | 74.62 | 74.62 | 74.62 | 73.48 |
| Zhang et al. - perturbed | 74.34 | 74.34 | 74.34 | 74.34 | 73.48 | 72.44 | 72.44 |
| Hardt et al.          | 71.31 | 71.31 | 71.31 | 70.45 | 68.75 | -     | -     |
Learnt Perturbations - Flexibility

- Fully **model-agnostic** with respect to the original model, as any model structure or access requirements **apply only to the perturbation**, and not the original model.

- If the original model is complex, we have the option of **training a lightweight perturbation** to the complex model, and may not need to rerun an expensive training procedure.
Learnt Perturbations - Stability

- The proposed perturbative approach has less variance and higher mean accuracy.

Figure 3: Accuracy violin plots of experimental outcomes binned by achieved level of fairness.
Limitations & Discussion question

● What do you think are the advantages and disadvantages of the perturbation method proposed in the paper against other training-time fairness algorithm?

● From the fairness shapley value, if we observe that a certain feature contributes a lot to the unfairness (e.g., marital status in demographic parity difference), is it always correct to remove that particular feature from the model?

● After reading this paper, what’s your opinion about using interpretability methods to validate the fairness of machine learning models?
Thanks & Questions