February 10, 2025, Lecture Note: Preprocess Interventions

Aadit Shaivalbhai Trivedi, Nihal Niraj Mishra

CS 516: Responsible Data Science and Algorithmic Fairness; Spring 2025
Abolfazl Asudeh; www.cs.uic.edu/∼ asudeh/teaching/archive/cs516spring25/

1 Recap

In the last lecture, we started discussing preprocessing intervention. Preprocessing techniques aim to transform the data so that any underlying discrimination is removed. If the algorithm is allowed to modify the training data, then preprocessing intervention techniques can be used. We discussed sampling and data augmentation.

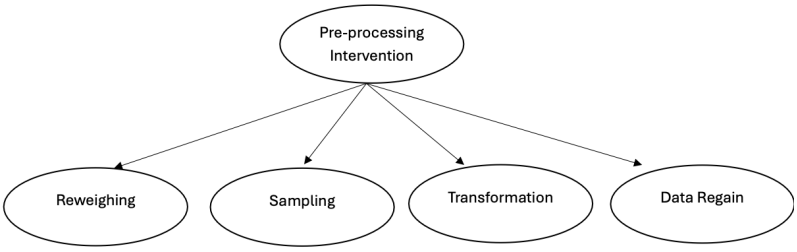

Figure 1: Preprocessing techniques for fairness.

2 Reweighing [1]

Reweighing assigns weights to instances of the training data while leaving the data itself unchanged. It keeps all the data but adjusts the model's learning process to give different importance to different groups, unlike sampling, which modifies the dataset by either duplicating (oversampling) or removing (undersampling) instances. This allows reweighing to correct biases without increasing overfitting risk or discarding valuable information, making it a more efficient and flexible approach to fairness.

The goal is to make sensitive attributes (S) independent of outcome (Y):

$$S \perp\!\!\!\perp Y$$

2.1 Formula for Weight Calculation

For each sample with sensitive attribute s_i and outcome y_j, the weight w_k is calculated as:

$$w_k = \frac{P(s_i)P(y_j)}{P(s_i \text{ and } y_j)}$$

Where:

- $P(s_i)$ is the marginal probability of the sensitive attribute s_i,
- $P(y_j)$ is the marginal probability of the outcome y_j,
- $P(s_i \text{ and } y_j)$ is the joint probability of both s_i and y_j occurring together.

2.2 Example of Weight Calculation

Let's assume the following probabilities for sensitive attribute *gender* and outcome *loan approval*:

- $P(\text{Male}) = 0.5$
- $P(\text{Female}) = 0.5$
- $P(\text{Approved}) = 0.6$
- $P(\text{Denied}) = 0.4$

Joint probabilities for *gender* and *loan approval* are:

- $P(\text{Male and Approved}) = 0.3$
- $P(\text{Male and Denied}) = 0.2$
- $P(\text{Female and Approved}) = 0.3$
- $P(\text{Female and Denied}) = 0.1$

Now, let's calculate the weights for each combination of gender and loan approval:

- **For Male and Approved**:

$$w_k = \frac{P(\text{Male})P(\text{Approved})}{P(\text{Male and Approved})} = \frac{0.5 \times 0.6}{0.3} = 1$$

- **For Male and Denied**:

$$w_k = \frac{P(\text{Male})P(\text{Denied})}{P(\text{Male and Denied})} = \frac{0.5 \times 0.4}{0.2} = 1$$

- **For Female and Approved**:

$$w_k = \frac{P(\text{Female})P(\text{Approved})}{P(\text{Female and Approved})} = \frac{0.5 \times 0.6}{0.3} = 1$$

- **For Female and Denied**:

$$w_k = \frac{P(\text{Female})P(\text{Denied})}{P(\text{Female and Denied})} = \frac{0.5 \times 0.4}{0.1} = 2$$

The weight for *Female and Denied* is 2, meaning the model will give more importance to these samples during training to help address potential bias against this underrepresented group.

2.3 Trade-off

Reweighing improves fairness by adjusting the influence of different groups and reducing biases in the model's predictions.

However, it can reduce predictive accuracy because the model might not learn as effectively from the natural distribution of the data. The model might have to generalize in ways that make it less specific to certain patterns, thus reducing its ability to make precise predictions.

3 Transformation

3.1 Earth Mover's Distance:

The Earth Mover's Distance (EMD) is a method to evaluate dissimilarity between two multi-dimensional distributions in some feature space where a distance measure between single features, known as the ground distance, is given. EMD "lifts" this distance from individual features to full distributions.

In short, given two data distributions, EMD quantifies the minimal effort required to transform one distribution into the other. The minimal change required represents the Earth Mover's Distance.

3.2 Optimized Preprocessing for Discrimination Prevention [2]:

The goal is to transform a biased dataset (x, y, s) into an unbiased dataset (x', y', s')

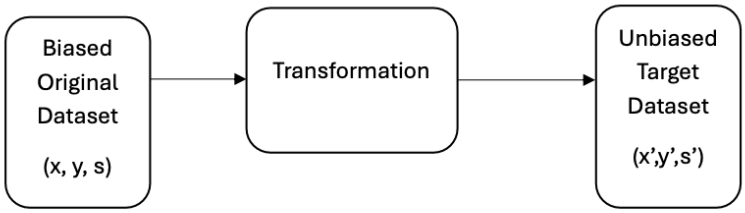

Figure 2: Preprocessing techniques for fairness.

The transformation process should satisfy the following constraints:

1. **Utility Constraint:**
$$P(x, y) \approx P(x', y')$$
The joint distribution of features and labels should not change significantly, ensuring that the model's predictive performance is preserved.

2. **Individual Distortion Constraint:**
$$(x_i, y_i) \approx (x'_i, y'_i)$$
Each individual's transformed data point (x'_i, y'_i) should be as close as possible to the original (x_i, y_i), preventing large distortions.

3. **Independence Constraint:**
$$S \perp\!\!\!\perp Y$$
The sensitive attribute S should be statistically independent of the label Y, ensuring fairness in predictions.

4 Certifying and Removing Disparate Impact [3]

Introduction:

To prevent disparate impact, it is crucial to ensure that the input features X do not allow for the prediction of the sensitive attribute S. This fundamental approach underpins efforts to remove biases and ensure fairness in automated decision-making systems.

Disparate Impact Rule:

$$t \leq \frac{P(f(x) = 1|S = 0)}{P(f(x) = 1|S = 1)} \leq \frac{1}{t}$$

Where $f(x)$ represents a binary classifier, and t often is set around 0.8, reflecting the 80% rule used in disparate impact assessments to ensure fairness across different demographic groups.

Task Independence:

Achieving fairness and eliminating disparate impact is independent of the specific tasks or models used. If the features X cannot predict the sensitive attribute S, no classifier trained on X can manifest disparate impact, emphasizing the importance of data preparation and feature selection in building fair models.

Certifying and removing disparate impacts are critical to developing fair machine learning systems. These efforts are supported by rigorous mathematical formulations and practical algorithmic interventions. The transition to discussing in-process interventions highlights the continuity in applying fairness throughout the machine learning pipeline—from data preparation to algorithmic decision-making.

4.1 Disparate Impact Certification

Disparate impact certification involves verifying that data or algorithmic decisions do not lead to unintended biases against protected groups. This is achieved by assessing the predictability of sensitive attributes from other data features. Certification is accomplished if these attributes cannot be predicted beyond chance.

4.2 Balanced Error Rate (BER)

Balanced Error Rate (BER) is used to assess the fairness of the algorithm by measuring how inaccurately the sensitive attribute S is predicted:

$$BER(g, X, S) = \frac{1}{2}\left(Pr[g(X) = 0|S = 1] + Pr[g(X) = 1|S = 0]\right)$$

A higher BER indicates that the sensitive attribute S is less predictable from the features X, suggesting better fairness.

4.2.1 Non-predictability and Epsilon Predictability

There exists no function g that can accurately predict S from X. This non-predictability is crucial for certifying that the model or data handling practices are free of bias concerning the sensitive attribute S.

If the Balanced Error Rate (BER) is bounded by ϵ, then S is considered ϵ-predictable. Specifically, $BER(g) \geq \epsilon$ implies that S does not influence the outcome significantly, maintaining fairness.

4.2.2 Linking BER to Disparate Impact

Suppose there is a classifier $f(X)$ with $BER(f)$ meeting specific criteria. If ϵ satisfies the condition:

$$\epsilon \geq \frac{1}{2} + \left(1 - \frac{1}{0.8}\right) \times P(f = 1|S = 0)$$

then the disparate impact is ensured to be at least 0.8. This condition shows that the probability of a favorable outcome for the unprotected group $S = 0$ is not disproportionately high, which aligns with fairness standards.

4.2.3 Proof Sketch for BER Analysis

$$\begin{aligned} BER(g) &= \frac{P(g = 1|S = 0) + P(g = 0|S = 1)}{2} \\ &= \frac{P(g = 1|S = 0) + (1 - P(g = 1|S = 1))}{2} \end{aligned}$$

Disparate impact requirement:

$$\frac{P(g = 1|S = 0)}{P(g = 1|S = 1)} \geq t$$

This leads to an inequality for BER(g):

$$BER(g) \leq \frac{P(g = 1|S = 0) + 1 - \frac{P(g=1|S=0)}{t}}{2}$$

Which simplifies to:

$$\frac{1}{2} + \frac{P(g = 1|S = 0)(1 - \frac{1}{t})}{2} = \varepsilon$$

5 Introduction to In-Process Interventions

In-process interventions involve modifying the learning algorithm itself to embed fairness directly into the model training process.

5.1 Objective

Adjust the training process to minimize loss while respecting fairness constraints, ensuring that the final model decisions do not favor one group over another unjustly.

5.2 Approach

Include terms in the loss function that penalize deviations from desired fairness metrics and introduce constraints into the optimization problem that ensure outputs meet specific fairness criteria, such as equal false positive rates across different groups.

6 References

1. Kamiran, Faisal, and Toon Calders. "Data preprocessing techniques for classification without discrimination." Knowledge and Information Systems 33.1 (2012): 1-33. DOI: 10.1007/s10115-011-0463-8

2. Calmon, Flavio, et al. "Optimized pre-processing for discrimination prevention." Advances in Neural Information Processing Systems 30 (2017). NeurIPS 2017 Paper

3. Feldman, Michael, et al. "Certifying and removing disparate impact." Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015. DOI: 10.1145/2783258.2783311