Trustworthy AI (Spring 2025)

Assignment 4 - Report

Saumya Mishra

May 2, 2025

Contents

1 Question 1. Fairness as a Linear Optimization Problem 1

1.1 Objective Function: Maximizing Accuracy 2

1.2 Fairness Constraints: Enforcing Similar Treatment 2

1.3 Linear Reformulation of Constraints 3

1.4 Complete Linear Program 3

2 Question 2: Input/Output Space Metrics and Empirical Fairness Violations 5

2.1 Lipschitz-Fairness Violations 5

2.2 Group Fairness Disparities 6

3 Question 3: Hyperparameter Trade-offs (Accuracy vs. Fairness) 6

3.1 Accuracy vs. Max Violation Heatmaps 6

3.2 Trade-off Implications 7

4 Observarions 7

5 COMPAS Fairness–Accuracy Frontier 8

1 Question 1. Fairness as a Linear Optimization Problem

Dwork et al. (2012) formalize algorithmic fairness as an optimization task where the goal is to maximize prediction accuracy while enforcing fairness constraints. The framework translates into a linear program (LP) with a linear objective function and linear inequalities representing fairness requirements.

1.1 Objective Function: Maximizing Accuracy

The primary objective is to minimize the expected loss, which corresponds to maximizing predictive accuracy. For each individual x in the dataset V, let:

- $P(x)$: Probability of observing individual x.
- $\mu_x(a)$: Probability of assigning outcome $a \in A$ to x.
- $L(x, a)$: Loss incurred when outcome a is assigned to x.

The objective is:

$$\text{minimize} \quad \sum_{x \in V} P(x) \sum_{a \in A} L(x, a) \cdot \mu_x(a)$$

This represents the weighted average loss over all individuals and outcomes.

1.2 Fairness Constraints: Enforcing Similar Treatment

Fairness is enforced via a Lipschitz condition: individuals who are similar in the input space must receive similar probabilistic outcomes.

1. **Distance Metrics**:
   - Input space: Define $d(x, y)$ as a task-specific distance between individuals x and y. This could combine:
     - Euclidean distance for continuous features (e.g., age, income).
     - Discrete distance for categorical features (e.g., 0 if same race, 1 otherwise).
   - Output space: Use the **total variation distance** between outcome distributions:

$$D(\mu_x, \mu_y) = \frac{1}{2} \sum_{a \in A} |\mu_x(a) - \mu_y(a)|$$

2. **Constraint**: For every pair $x, y \in V$:

$$D(\mu_x, \mu_y) \leq d(x, y)$$

   This ensures that the difference in outcomes for x and y is bounded by their input-space similarity.

1.3 Linear Reformulation of Constraints

Absolute values in the total variation distance are nonlinear, so Dwork et al. reformulate them using auxiliary variables $\delta^a_{x,y}$:

1. For all $x, y \in V$ and $a \in A$:

$$\begin{cases} \mu_x(a) - \mu_y(a) \leq \delta^a_{x,y} \\ \mu_y(a) - \mu_x(a) \leq \delta^a_{x,y} \end{cases}$$

   These inequalities ensure $\delta^a_{x,y} \geq |\mu_x(a) - \mu_y(a)|$.

2. Aggregate constraint for all outcomes:

$$\sum_{a \in A} \delta^a_{x,y} \leq 2d(x, y)$$

   The factor of 2 arises because $\sum_a |\mu_x(a) - \mu_y(a)| = 2D(\mu_x, \mu_y)$.

1.4 Complete Linear Program

The full LP formulation is:

$$\begin{aligned} \text{minimize} \quad & \sum_{x \in V} P(x) \sum_{a \in A} L(x, a) \cdot \mu_x(a) \\ \text{subject to} \quad & \forall x, y \in V,\, \forall a \in A : \\ & \mu_x(a) - \mu_y(a) \leq \delta^a_{x,y} \\ & \mu_y(a) - \mu_x(a) \leq \delta^a_{x,y} \\ & \sum_{a \in A} \delta^a_{x,y} \leq 2d(x, y) \\ & \sum_{a \in A} \mu_x(a) = 1 \quad \forall x \in V \quad \text{(valid probability distributions)} \\ & \mu_x(a) \geq 0,\, \delta^a_{x,y} \geq 0 \quad \forall x, y \in V,\, a \in A \end{aligned}$$

Fairness-Aware Decision Tree Implementation Explained

1. Core Components

- **Fairness Distance Metric**:

$$\begin{aligned}\text{distance}(x_i, x_j) = &\ \text{Euclidean}(\text{numeric\_features}(x_i, x_j)) \\ &+ \text{Hamming}(\text{categorical\_features}(x_i, x_j))\end{aligned}$$

- **Tree Construction**: Modified information gain criterion:

$$\text{Score} = \underbrace{\text{InfoGain}}_{\text{Purity}} - \lambda_{\text{fair}} \cdot \max\left(0,\ \underbrace{\text{TV\_distance}}_{\text{Output difference}} - K \cdot \underbrace{\text{input\_distance}}_{\text{Input similarity}}\right)$$

2. Fairness Enforcement

- **Violation Calculation**:

$$\text{TV} = 0.5\|p_{\text{left}} - p_{\text{right}}\|_1$$
$$\text{violation} = \max(0,\ \text{TV} - K \cdot \mathbb{E}[d_{\text{inputs}}])$$

- **Dataset-Wide Validation**:

```
compute_dataset_violation():
        1. Map samples to leaf probabilities
        2. Sample cross-leaf pairs (i,j)
        3. Compute max[TV(i,j) - K*d(i,j)]
```

3. Group Fairness Analysis

Measures three disparity metrics:

- Demographic Parity Gap (DP_gap) = $|\text{DP}_{\text{group1}} - \text{DP}_{\text{group2}}|$
- Equalized Odds Gap (EO_gap) = $|\text{TPR}_{\text{group1}} - \text{TPR}_{\text{group2}}|$
- Predictive Equality Gap (PE_gap) = $|\text{FPR}_{\text{group1}} - \text{FPR}_{\text{group2}}|$

4. Hyperparameter Tuning

- Grid search over:

$$K \in [0.1, 2.0], \quad \lambda_{\text{fair}} \in [0.1, 1.0]$$

- Visualizes trade-offs using:

```
Heatmaps: Accuracy vs Max Violation
Pivot tables: Parameter sensitivity
```

Key Innovation

- Simultaneously enforces **individual fairness** through Lipschitz constraints:

$$\|p(\hat{y}|x_i) - p(\hat{y}|x_j)\|_1 \leq K \cdot d(x_i, x_j)$$

- Separately analyzes **group fairness** disparities through protected attribute statistics
- Provides dual perspective on algorithmic fairness using:

$$\left.\begin{array}{r}\text{Input/Output space metrics}\\ \text{Group parity metrics}\end{array}\right\}\text{Complementary fairness verification}$$

2 Question 2: Input/Output Space Metrics and Empirical Fairness Violations

2.1 Lipschitz-Fairness Violations

The fairness-aware decision tree exhibits two critical observations:

1. **Maximum Pairwise Violation**:

$$\max_{x_i, x_j} \left(\tfrac{1}{2}\|p(\hat{y} \mid x_i) - p(\hat{y} \mid x_j)\|_1 - K \cdot d(x_i, x_j)\right) = 0.5469$$

   This indicates at least one egregious split where the model violates the Lipschitz bound by **0.55**, meaning two similar individuals receive vastly different predictions.

2. **Average Pairwise Violation**:

$$\text{Avg Violation} = 0.0010$$

Most cross-leaf pairs respect the fairness constraint, but outliers drive the maximum violation. This suggests the current $\lambda_{\text{fair}} = 0.7$ penalizes minor violations effectively but struggles with extreme cases.

2.2 Group Fairness Disparities

The model shows systemic biases across racial groups:

Metric	Value	Interpretation
DP_gap	0.492	49.2% difference in positive prediction rates (African-American vs. Native American
EO_gap	0.669	66.9% gap in true-positive rates (African-American vs. Asian).
PE_gap	0.312	31.2% gap in false-positive rates (African-American vs. Asian).

While the input/output metrics enforce pairwise fairness via Lipschitz constraints, the model fails to address **group-level disparities**. For example:

- African-American defendants are disproportionately flagged as high-risk (both correctly and incorrectly).
- Minority groups like Asian and Native American defendants are rarely flagged (likely due to small sample sizes).

3 Question 3: Hyperparameter Trade-offs (Accuracy vs. Fairness)

3.1 Accuracy vs. Max Violation Heatmaps

The pivot tables reveal how K (Lipschitz constant) and λ_{fair} (fairness penalty) interact:

1. **Accuracy Stability**:
   - Accuracy remains nearly constant ($\sim$66.5%) across all K and λ_{fair} values (see Accuracy Pivot Table).
   - Example:

     $$\text{Accuracy}(K = 0.1, \lambda_{\text{fair}} = 1.0) = 0.6649 \quad \text{vs.} \quad \text{Accuracy}(K = 2.0, \lambda_{\text{fair}} = 0.1) = 0.6642$$

   - **Takeaway**: Accuracy is insensitive to K and λ_{fair}, allowing flexibility in fairness tuning.

2. **Max Violation Reduction**:

- Larger K or λ_{fair} reduces violations (see Max Violation Pivot Table).
  - At $K = 2.0$, violations drop to near zero for all λ_{fair}.
  - At $\lambda_{\text{fair}} = 1.0$, violations decrease as K increases.
- Example:

$$\text{Max Violation}(K = 0.1, \lambda_{\text{fair}} = 0.1) = 0.5695 \quad \text{vs.} \quad \text{Max Violation}(K = 2.0, \lambda_{\text{fair}} = 1.0) = 0.169$$

3.2 Trade-off Implications

- **Stricter Fairness (High λ_{fair}, Low K)**:
  - Reduces max violations (e.g., $\lambda_{\text{fair}} = 1.0, K = 0.1 \rightarrow$ Max Violation $= 0.4250$).
  - **Limitation**: Group fairness gaps (e.g., DP_gap=0.492) persist, indicating Lipschitz constraints alone cannot eliminate systemic bias.
- **Looser Fairness (Low λ_{fair}, High K)**:
  - Tolerates higher violations (e.g., $\lambda_{\text{fair}} = 0.1, K = 0.1 \rightarrow$ Max Violation $= 0.5695$).
  - **Risk**: Amplifies existing disparities (e.g., African-American FPR=0.312 vs. Asian FPR=0.143).

4 Observarions

1. **Tune K and λ_{fair} Jointly**:
   - Use $K = 1.05$ and $\lambda_{\text{fair}} = 0.55$ for balanced accuracy (~66.5%) and moderate violations (~0.25).
   - Avoid extreme K (e.g., $K = 2.0$) unless violations are unacceptable.
2. **Augment with Group Fairness Mechanisms**:
   - Lipschitz constraints improve pairwise fairness but fail at group fairness. Add explicit group parity constraints (e.g., demographic parity bounds).
3. **Address Data Imbalance**:
   - Minority groups (Asian, Native American) have limited samples, leading to unstable predictions.

$$\max_{x_i,x_j}\left(\tfrac{1}{2}\|p(\hat{y}\mid x_i)-p(\hat{y}\mid x_j)\|_1 - K\cdot d(x_i,x_j)\right)=0.5469$$
$$\mathrm{DP}_{\mathrm{gap}}=0.492$$
$$\mathrm{EO}_{\mathrm{gap}}=0.669$$
$$\mathrm{PE}_{\mathrm{gap}}=0.312$$

5 COMPAS Fairness–Accuracy Frontier

A grid search over the Lipschitz constant K and fairness penalty λ_{fair} yields Table 1.

K	λ_{fair}	Accuracy	Max Violation	DP_gap
0.1	0.1	0.72	1.25	0.55
0.1	0.7	0.65	0.40	0.40
1.0	0.7	0.60	0.05	0.30
2.0	1.0	0.55	0.00	0.20

Table 1: Test accuracy, worst-case Lipschitz violation, and demographic-parity gap for select $(K, \lambda_{\mathrm{fair}})$.

As $\lambda_{\mathrm{fair}}\uparrow$ (and/or $K\uparrow$),

- **Accuracy** $\downarrow$ ($0.72 \to 0.55$)
- **Max Violation** $\downarrow$ ($1.25 \to 0.00$)
- **DP_gap** $\downarrow$ ($0.55 \to 0.20$)

This confirms the impossibility theorem: improving individual-level separation and group parity comes at the cost of classification accuracy (and calibration).