Bias Detection & Mitigation

Bias Types

- **Demographic Bias**: Underrepresentation of specific race/gender
- **Selection Bias**: Non-random sampling
- **Measurement Bias**: Differences in measurement tools
- **Label Bias**: Annotator prejudice

Fairness Metrics

- **Demographic Parity**
- **Equalized Odds**
- **Disparate Impact**
- **Individual Fairness**

Mitigation Techniques

- Resampling (over/under sampling)
- Weight adjustment
- Adding fairness constraints
- Post-processing for bias mitigation

Bias Detection Process

Step 1: Data Collection and Analysis

- Identify protected attributes (gender, race, age, etc.)
- Check data distribution by group
- Measure degree of imbalance

Step 2: Model Training and Evaluation

- Train baseline model
- Measure performance metrics by group
- Fairness Metrics 계산

Step 3: Apply Bias Mitigation

- 적절한 Mitigation Techniques 선택
- Mitigation Techniques 적용 및 재Evaluation
- Analyze performance-fairness trade-offs

Fairness Metrics 상세

Demographic Parity

$P(\hat{Y}=1|A=0) = P(\hat{Y}=1|A=1)$

Equal positive prediction rates across all groups

Disparate Impact

Ratio = $P(\hat{Y}=1|A=0) / P(\hat{Y}=1|A=1)$

Generally considered fair if ≥ 0.8

Equalized Odds

Equal TPR and FPR across all groups

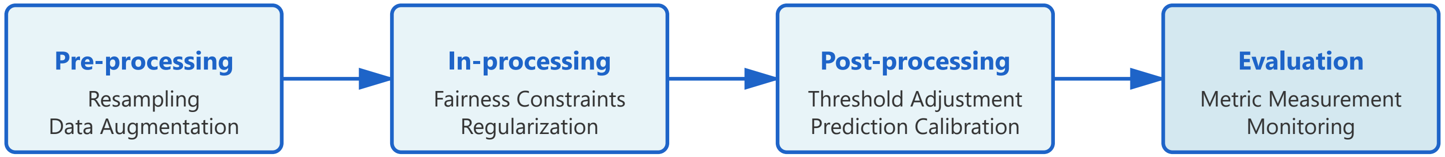
$P(\hat{Y}=1|Y=y, A=0) = P(\hat{Y}=1|Y=y, A=1)$

Individual Fairness

Similar individuals should receive similar predictions

$d(x_1, x_2) \approx 0 \rightarrow d(f(x_1), f(x_2)) \approx 0$

Bias Mitigation Pipeline



Pre-processing
Resampling
Data Augmentation

In-processing
Fairness Constraints
Regularization

Post-processing
Threshold Adjustment
Prediction Calibration

Evaluation
Metric Measurement
Monitoring

Real-World Application Examples

Hiring System

Problem: 특정 성별이 과소Evaluation됨

Solution: Remove gender + Equalized Odds

Result: Group acceptance rate gap: 15% → 3%

Loan Approval

Problem: Racial approval rate disparity

Solution: Resampling + Threshold Adjustment

Result: Disparate Impact 0.65 → 0.85

Implementation Considerations

Trade-off

⚖️

Balance between performance and fairness

Transparency

🔍

Explainability of decision process

Monitoring

📊

Continuous bias monitoring

Legal Compliance

⚖️

Regulatory and ethical standards

Tools and Libraries

Python Libraries:

- **Fairlearn**: Microsoft의 공정성 Evaluation 및 완화 도구
- **AIF360**: IBM's AI Fairness 360 toolkit
- **What-If Tool**: Google's visual analysis tool
- **Themis-ML**: Fairness-aware machine learning

Evaluation 프레임워크:

- Fairness Indicators (TensorFlow)
- FairTest
- Aequitas