FAIR-TAN: Fairness-Aware Income Prediction via Task-Audit Networks

A Dual-Objective Neural Network Approach for Ethical ML

Overview

What is FAIR-TAN?

• FAIR-TAN is a machine learning model designed to predict income levels based on demographic and socio-economic data.

• Unlike traditional ML models, FAIR-TAN audits its own predictions for fairness, ensuring that sensitive attributes like sex and race do not lead to biased outcomes.

• It combines two objectives:

  - TaskNet for income prediction.

  - AuditNet for fairness auditing and mitigation of bias.

• The goal is to balance predictive accuracy with fairness in the decision-making process.

Key Features

• Dual-Objective Architecture: Simultaneously optimizes for task prediction accuracy and fairness.

• Fairness Auditing: Ensures that model predictions are not disproportionately influenced by sensitive attributes.

• Real-World Application: Useful in areas where fairness is critical, like hiring, lending, or public policy.

Core Components

TaskNet:

• Function: TaskNet is responsible for predicting whether an individual's income is greater than or less than $50K.

• Input: Takes demographic and socio-economic features (e.g., age, education, hours worked).

• Output: A predicted class label (either <=50K or >50K).

• Loss Function: The model is trained using cross-entropy loss to minimize prediction error.

AuditNet:

• Function: AuditNet audits TaskNet's predictions for fairness.

• Input: Receives the output of TaskNet (predicted income class).

• Output: A fairness score that indicates whether the prediction is biased with respect to sensitive attributes.

• Loss Function: AuditNet is trained using binary cross-entropy loss to detect fairness violations.

Mathematical Formulation (TaskNet Loss)

TaskNet Loss Function (Cross-Entropy):

• Cross-entropy loss is used to evaluate how well TaskNet predicts the income class.

Formula:

$$L_{task} = -\sum y_i \log(\hat{y}_i)$$

Where:

• y_i is the true income label (0 for <=50K, 1 for >50K).

• $\hat{y}_i$ is the predicted probability from TaskNet.

Why Cross-Entropy?

• Cross-entropy loss is commonly used for classification tasks because it penalizes incorrect predictions, and the penalty increases as the predicted probability diverges from the true label.

Mathematical Formulation (AuditNet Loss)

AuditNet Loss Function (Binary Cross-Entropy):

• Binary cross-entropy is used to evaluate how well AuditNet detects fairness violations in TaskNet's predictions.

Formula:

$L_{audit} = -\sum s_i \log(\hat{s}_i) + (1 - s_i) \log(1 - \hat{s}_i)$

Where:

• s_i is the sensitive attribute (e.g., sex or race) for the i-th instance.

• $\hat{s}_i$ is the fairness score predicted by AuditNet.

Why Binary Cross-Entropy?

• Binary cross-entropy is used because AuditNet is essentially classifying whether the model's prediction is biased (binary outcome: biased vs. unbiased).

Total Loss Function

Combined Objective (TaskNet + AuditNet):

• The model is trained to minimize both task prediction error and fairness violations.

Total Loss Formula:

$L_{total} = L_{task} + \lambda L_{audit}$

Where:

• λ is a hyperparameter that controls the trade-off between accuracy and fairness.

• When λ is large, the model prioritizes fairness more heavily, even if it sacrifices some predictive accuracy.

Why Combine Losses?

• Combining the losses enables joint optimization, where the model simultaneously improves its predictions while minimizing fairness violations. This way, we address both model performance and ethical considerations in a single unified training process.

Architecture Overview

FAIR-TAN Architecture:

• The architecture consists of two primary networks:

- TaskNet: A standard feedforward neural network that takes demographic and socio-economic features as input and outputs income class predictions.

- AuditNet: A second neural network that takes TaskNet's predictions as input and evaluates fairness based on sensitive attributes like sex or race.

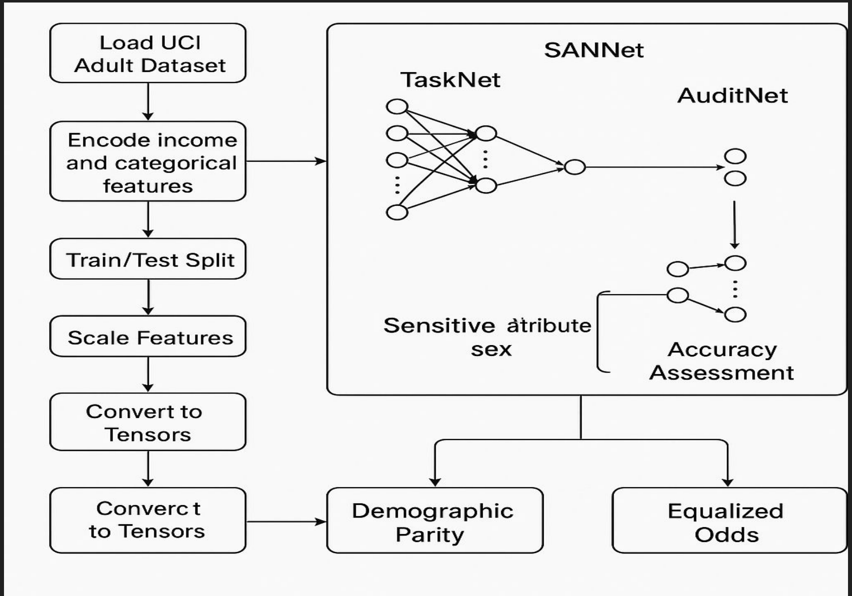
Network Flow:

• Input Features: Demographic and socio-economic data (age, education, hours worked, etc.).

• TaskNet: Produces a predicted income classification (<=50K or >50K).

• AuditNet: Audits the predictions for fairness, generating a fairness score that helps reduce bias.

Visual View



Load UCI Adult Dataset
Encode income and categorical features
Train/Test Split
Scale Features
Convert to Tensors
Converc t to Tensors
SANNet
TaskNet
AuditNet
Sensitive atribute sex
Accuracy Assessment
Demographic Parity
Equalized Odds

Results and Evaluation

Evaluation Metrics:

• Accuracy: Measures the proportion of correct income predictions.

Fairness Compliance:

• Demographic Parity: Ensures that different demographic groups receive positive predictions at equal rates.

• Equalized Odds: Ensures that both the False Positive Rate (FPR) and True Positive Rate (TPR) are similar across groups.

• Audit Accuracy: Measures how well AuditNet identifies fairness violations.

Impact:

• FAIR-TAN ensures that machine learning models do not unfairly favor or disadvantage specific demographic groups.

• This framework can be applied to sensitive areas such as hiring, lending, and law enforcement, where fairness is crucial to avoid discrimination.