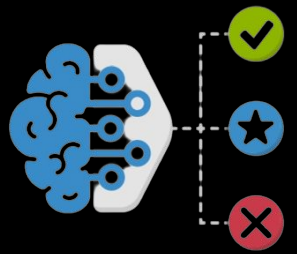
Learning Fairness Representations
VS
Fairness-aware Classifier with Prejudice Remover Regularizer

Group 4



Problem Statement

Goal: To implement and compare Algorithm 1 & Algorithm 5 and determine which model does a better job in **fairly** predicting two-year recidivism.

Algo 1: Learning Fair Representations
Algo 5: Fairness-aware Classifier with Prejudice Remover Regularizer

Metric of Performance: Equality of Odds using Precision

P(predicted to recidivate| Yes Caucasian) = P(predicted to recidivate| Yes African American)

Dataset

1. We are using the COMPAS dataset that contains the criminal history, prison time, demographics and COMPAS risk scores for defendants from Broward County from 2013 and 2014

2. The dataset can be found at https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis

Data Preprocessing

Input X: age, sex (1 - M, 0 - F), juv_fel_count, juv_misd_count, juv_other_count, priors_count, c_charge_degree (1- M, 0 - F), days_b_screening_arrest, is_recid (1 - Yes, 0 - No) and is_violent_recid (1 - Yes, 0 - No)

Output Y: two_year_recid (1 - Yes, 0 - No) . Indicates whether or not an individual has recidivated within two years

Sensitive feature S: race (1 - Caucasian, 0 - African-American)

Data Split: 80% train, 20% test

Learning Fair Representations

(Zemel et al., 2013)

The idea is to take X and transform it into a different form (Z) that achieves the following:

- Eliminates any details that would reveal whether someone belongs to a protected group (race, binary variable S)
- When X is transformed into Z, the probability of a person from a protected group ($S = 1$) being mapped to Z is the same as for someone from an unprotected group ($S = 0$).
- Most of the relevant information from X is retained when it is transformed into Z.
- Z can be used to map to Y (predicted recidivism).

LFR: Building the Model

Finding Z

- Input X and Y for both $S = 0$ and $S = 1$ into the Encoder model (a neural network)
- Model outputs a low-dimensional representation of the input, which is the latent variable Z.

The Final Model (predicting Y)

- A logistic regression model is built using the latent variable Z as input features.
- To train the model, the loss function is defined as a weighted sum of Lx, Ly, and Lz, where
  - Lx: the difference between the original input features and the features transformed by the model
  - Ly: the difference between the predicted labels and the actual labels
  - Lz: how different the representations of the protected and unprotected groups are
- The model is trained by looping over various hyperparameters and finding the combination of hyperparameters that minimizes the overall loss function.

Fairness-aware Classifier with Prejudice Remover Regularizer

(Kamishima et al., 2012)

The goal of a fairness-aware classifier is to ensure that the decisions made by the base algorithm are not influenced by factors such as race, or other protected characteristics. It is designed to reduce bias and discrimination in the decision making process by achieving the following:

- Identifies the protected attributes (race, binary variable S)
- Incorporates a regularizer into the objective function of the learning algorithm
- The regularizer uses a prejudice index to penalize the base model for using features that would reveal whether someone belongs to a protected group (S)

PR: Building the Model

Main Components

1. Sigmoid function
2. Prejudice Index

The Final Model (predicting Y)

- A logistic regression model is built using X as input features.
- To train the model, the sigmoid activation function is used to calculate the loss function.
- The prejudice remover regularizer penalizes the model using a prejudice index, by comparing the model's predictions to the actual distribution of the sensitive feature (race).
- If the predicted distributions are different from the actual distribution, the penalty increases, leading the model to adjust its objective function.
- The model is trained by looping over various hyperparameters and finding the combination of hyperparameters that minimizes the overall loss function.

Results (LFR vs PR)

Learning Fair Representations

	Precision	Accuracy
Caucasian	95.54%	97.77%
African American	95.41%	97.71%

Prejudice Remover

	Precision	Accuracy
Caucasian	96.37%	92.52%
African American	93.79%	94.33%

To maximise fairness (equality of odds) between protected and unprotected groups, we would recommend the LFR model to predict two-year recidivism in the Compas dataset. The LFR model has about equal rates of true positives (precision), whereas the PR model has lower precision (higher false positive) rate in African-Americans, in line with biases against the group.

References

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23 (pp. 35-50). Springer Berlin Heidelberg.

Larson, J., Angwin, J., Kirchner, L., & Mattu, S. (2016, May 23). How we analyzed the compas recidivism algorithm. ProPublica. Retrieved April 12, 2023, from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013, May). Learning fair representations. In International conference on machine learning (pp. 325-333). PMLR.

THANK YOU