Maximizing Fairness Under Accuracy Constraints (gamma and fine-gamma) and Handling Conditional Discrimination (LM and LPS)

Group 7

How do the algorithms work?

Maximizing Fairness Under Accuracy Constraints

This method produces a model by solving a constrained optimization problem. The problem maximizes fairness while constraining accuracy.

This method was proposed by Zafar, Valera, Rodriguez, and Gummadi in 2015. It is recommended when the covariance between the class labels and sensitive attributes is high.

$$\begin{aligned} \textbf{minimize} \quad & \left| \frac{1}{N} \sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, d_{\boldsymbol{\theta}}(\mathbf{x}_i) \right| \\ \textbf{subject to} \quad & L(\boldsymbol{\theta}) \leq (1+\gamma) L(\boldsymbol{\theta}^*), \end{aligned}$$

Maximizing Fairness Under Accuracy Constraints

In the optimization problem, the objective function has been derived as an estimate of the covariance between the sensitive attributes and the distance from each data points feature vector to the decision boundary.

The constraint states that the loss must be less than or equal to the optimal loss multiplied by 1 + a parameter, gamma.

$$\begin{aligned} \text{minimize} \quad & \left| \tfrac{1}{N} \textstyle\sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, d_{\boldsymbol{\theta}}(\mathbf{x}_i) \right| \\ \text{subject to} \quad & L(\boldsymbol{\theta}) \leq (1+\gamma) L(\boldsymbol{\theta}^*), \end{aligned}$$

$$\begin{aligned} \operatorname{Cov}(\mathbf{z}, d_{\boldsymbol{\theta}}(\mathbf{x})) &= \mathbb{E}[(\mathbf{z} - \bar{\mathbf{z}}) d_{\boldsymbol{\theta}}(\mathbf{x})] - \mathbb{E}[(\mathbf{z} - \bar{\mathbf{z}})] \bar{d}_{\boldsymbol{\theta}}(\mathbf{x}) \\ &= \mathbb{E}[(\mathbf{z} - \bar{\mathbf{z}}) d_{\boldsymbol{\theta}}(\mathbf{x})] \\ &\approx \frac{1}{N} \sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, d_{\boldsymbol{\theta}}(\mathbf{x}_i), \end{aligned} \tag{2}$$

Maximizing Fairness Under Accuracy Constraints with Fine Gamma

One variant of this method makes it possible to protect certain individuals/groups from being misclassified.

Using fine-grained accuracy constraints, loss constraints can be set on an individual basis like in the optimization problem below. The constraint states that the loss on a particular individual must be less than or equal to the optimal loss on that particular individual multiplied by 1 + a parameter, gamma.

$$\begin{aligned} \text{minimize} \quad & \left| \frac{1}{N} \sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}}) \, \boldsymbol{\theta}^T \mathbf{x}_i \right| \\ \text{subject to} \quad & L_i(\boldsymbol{\theta}) \leq (1 + \gamma_i) L_i(\boldsymbol{\theta}^*) \quad \forall i \in \{1, \ldots, N\}, \end{aligned}$$

Data Processing and Metrics

We processed the data by one-hot-encoding categorical variables, dropping irrelevant columns, replacing missing values with the mode, and separating the sensitive variable (race) from the input data. We then split the input and output data into test and train sets.

The metrics we assessed for model A3 were accuracy (percentage of testing points classified correctly) and covariance between the sensitive feature and distance from the decision boundary, as a measure of fairness.

Baseline Results

For the baseline model, we achieve an accuracy of approximately 99%.

Optimizing Fairness with Accuracy Constraints Results

When optimizing fairness with accuracy constraints, we achieve an accuracy of approximately 45%, a significant reduction in accuracy from the baseline. However, we see more fairness, with equal proportions of each sensitive attribute group classified as positive.

Interestingly, we see that the model predicts positive for every testing point. We feel that this is either a result of choosing to optimize fairness or a bug in the training function we used that was provided by Zafar, Valera, Rodriguez, and Gummadi's paper's GitHub.

Optimizing Fairness with Accuracy Constraints (Fine Gamma) Results

When optimizing fairness with accuracy constraints with the fine gamma variant, we achieve the same accuracy of 45% and also have equal proportions of each sensitive attribute group classified as positive.Again, we see that the model predicts positive for every testing point.

However, we also assessed the covariance between the sensitive feature and distance from the decision boundary for A3 with and without fine gamma. While it varies from run to run, we generally see a lower covariance between the sensitive feature and distance from the decision boundary with fine gamma. This implies fine gamma creates a more fair decision boundary on average.

Handling Conditional Discrimination (LM and LPS) A6

In A6, instead of directly aim for an algorithm to deal with the fairness issue, we aim to modify the data before training the model. (Pre-processing)

The idea is to balance the dataset so that the inherent/unexplainable discrimination is avoided.

$$P'(+ \mid e_i, f) = P'(+ \mid e_i, m) = P^*(+ \mid e_i)$$

In detail, we used Local Massaging and Preferential sampling to achieve this goal.

Handling Conditional Discrimination (LM and LPS)

Local Massaging:

Input dataset (X) and output balanced labels Y'.

The idea is to relabel the data points near the decision boundary to the opposite state to counter the unexplainable bias.

for *each partition* $X^{(i)}$ **do**
  learn a ranker $\mathcal{H}_i : X^{(i)} \rightarrow y^{(i)}$;
  rank males using $\mathcal{H}_i$;
  `relabel DELTA` (male) males that are the closest to the decision boundary from + to − (Algorithm 4);
  rank females using $\mathcal{H}_i$;
  `relabel DELTA` (female) females that are the closest to the decision boundary from − to +
end

Handling Conditional Discrimination (LM and LPS)

Local Preferential Sampling:

Input dataset (X) and output resampled dataset X'.

The idea is to remove the samples close to the decision boundary and fill this void with re-sampled data in the neighborhood to weaken the unexplainable bias.

for *each partition* $X^{(i)}$ **do**
  learn a ranker $\mathcal{H}_i : X^{(i)} \rightarrow y^{(i)}$;
  `rank` males using $\mathcal{H}_i$;
  `delete` $\frac{1}{2}$`DELTA` (male) (see Algorithm 4) males + that are the closest to the decision boundary;
  `duplicate` $\frac{1}{2}$`DELTA` (male) males − that are the closest to the decision boundary;
  `rank` females using $\mathcal{H}_i$;
  `delete` $\frac{1}{2}$`DELTA` (female) females − that are the closest to the decision boundary;
  `duplicate` $\frac{1}{2}$`DELTA` (female) females + that are the closest to the decision boundary;
end

Data Processing and Metrics

We did a similar data cleaning process with A3 except that we kept dates in our raw data.

Our metrics is overall model accuracy and the calibration score, which is the difference of the accuracy rate between the two subgroups. I.e, accuracy for predicting African Americans - accuracy for predicting Caucasians.

```
AA_X  =  calib_X[calib_X.race=='African-American']
CA_X  =    calib_X[calib_X.race=='Caucasian']
print(np.abs(accuracy_score(AA_X.label,  AA_X.pred)-accuracy_score(CA_X.label,  CA_X.pred)))
```

Random Forest Results

For comparison purposes, we trained a random forest model using the raw dataset. Here are the results:

From top to bottom, the outputs are the accuracy and the calibration score:

Accuracy
Calibration Score

```
0.9658536585365853
0.006187746733628763
```

```
The rate of Recidivism for African-American is 0.51
The rate of Recidivism for Caucasian is 0.39
The Corrected Recidivism rate should be 0.45
```

Local Massaging Results

After using the local massaging method to balance the dataset, we retrain the random forest model.

From top to bottom, the outputs are the accuracy and the calibration score:

Accuracy	0.9761517615176152
Calibration Score	0.009698079358899103

```
The rate of current Recidivism for African-American is 0.45
The rate of current Recidivism for Caucasian is 0.44
```

Local Preferential Sampling Results

Similarly, after applying the local preferential sampling method, we trained our random forest model again.

From top to bottom, the three outputs are accuracy, confusion matrix and the calibration score

Accuracy	0.9642276422764228
Calibration Score	0.008863928714003433

```
The rate of Recidivism for African-American is 0.51
The rate of Recidivism for Caucasian is 0.39
```

Conclusion

We tried two different approaches to deal with the inherent discrimination within the model. The in-processing method with accuracy constraints (A3) and pre-processing method (A6). In A3 we achieve the goal by posting penalties in the optimization process for unfairness, in A6 we pre-process the data to achieve balance before training the model. A common pattern is, the quality/accuracy of the model is a trade-off with the imposed fairness constraints, and one should think carefully and decides among these two.