Project 4: Algorithm Implementation and Evaluation

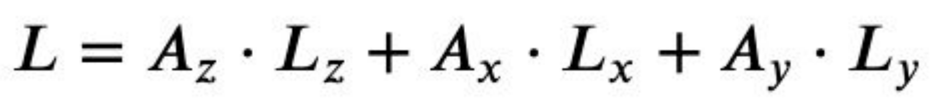
Group 2: Wen Chen(cw3229)
Haoyu He(hh2982)
Chenghao Lu(cl4259)
Ranran Tao(rt2796)
Aubrey Yan(xy2543)
Xinyu Zhu(xz3136)

A1: Learning Fair Representations (LFR)

Basic Idea:

In A1, the authors propose a learning algorithm for fair classification that aims to achieve both group fairness and individual fairness.

This LFR algorithm minimizes an objective function (shown below) with three terms corresponding to the goals of statistical parity, information preservation, and accurate classification.

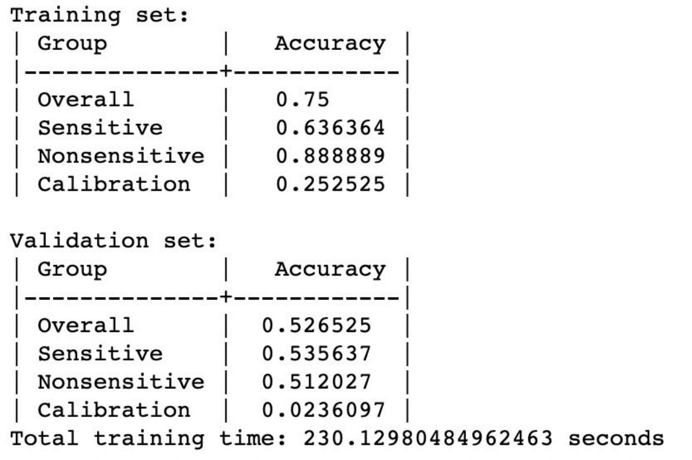


$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$$

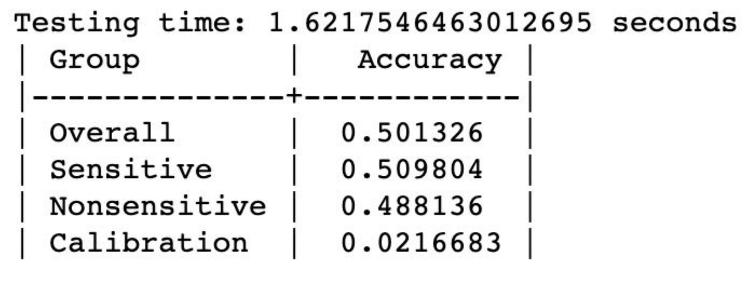
A1: Learning Fair Representations (LFR)

Results:

After implementing LFR, we achieved an overall accuracy of around 50% with a small calibration of ~0.02. This result suggests that the LFR algorithm can effectively balance fairness and accuracy in classification tasks, making it a good approach for promoting fairness.



```
Training set:
| Group          |    Accuracy |
|----------------+-------------|
| Overall        |    0.75     |
| Sensitive      |    0.636364 |
| Nonsensitive   |    0.888889 |
| Calibration    |    0.252525 |

Validation set:
| Group          |    Accuracy |
|----------------+-------------|
| Overall        |  0.526525   |
| Sensitive      |  0.535637   |
| Nonsensitive   |  0.512027   |
| Calibration    |  0.0236097  |
Total training time: 230.12980484962463 seconds
```



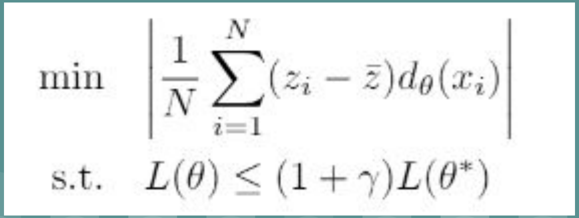
```
Testing time: 1.6217546463012695 seconds
| Group          |    Accuracy |
|----------------+-------------|
| Overall        |  0.501326   |
| Sensitive      |  0.509804   |
| Nonsensitive   |  0.488136   |
| Calibration    |  0.0216683  |
```

A3: Maximizing Fairness Under Accuracy Constraints (Gamma and Fine-gamma)

Basic Idea:

The goal of this algorithm is to ensure compliance with the disparate impact by maximizing the fairness subject to accuracy constraints. Minimizing the corresponding absolute decision boundary covariance over the training set with constraints could help us find the decision boundary parameters.

Here we use one gamma across individuals. This allows a relaxation on the accuracy from the unconstrained case with regard to accuracy so that the algorithm can minimize unfairness. The formula is defined below:



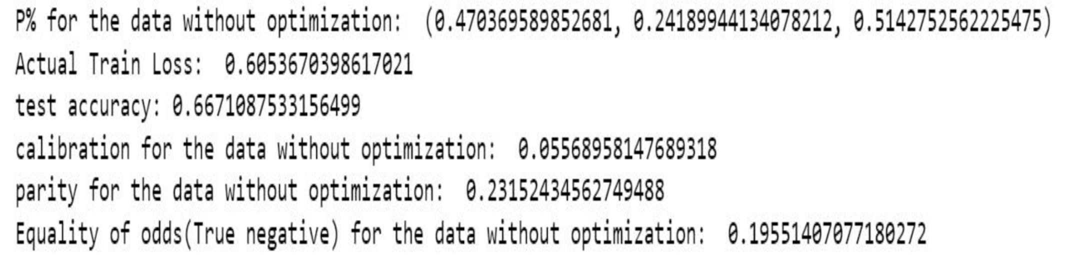
$$\begin{aligned} \text{min} \quad & \left| \frac{1}{N} \sum_{i=1}^{N} (z_i - \bar{z}) d_\theta(x_i) \right| \\ \text{s.t.} \quad & L(\theta) \leq (1 + \gamma) L(\theta^*) \end{aligned}$$

A3: Maximizing Fairness Under Accuracy Constraints (Gamma and Fine-gamma)

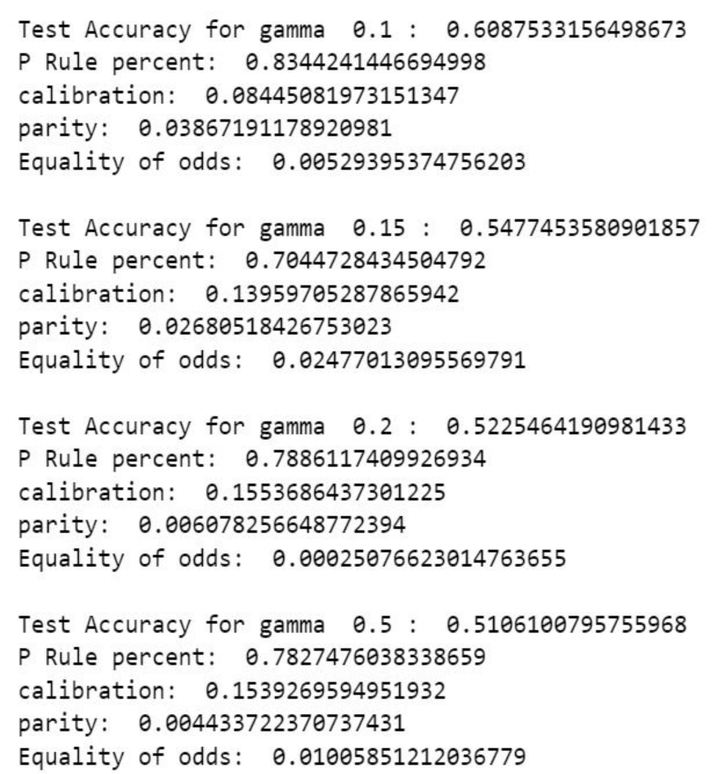
Results:

Improving the p-rule percent can lead to better fairness compared to baseline model. The calibration, parity, equality of odds are metrics used to evaluate the fairness of a binary classifier with respect to two groups. Ideally, we want them to be as close to 0 as possible, indicating that there is no significant difference in the proportion of positive outcomes between the two groups.

A gamma value of 0.1 is relatively good since it leads to increased p-rule value and parity and equality of odds values which are closed to 0, while the accuracy is not decreased much.

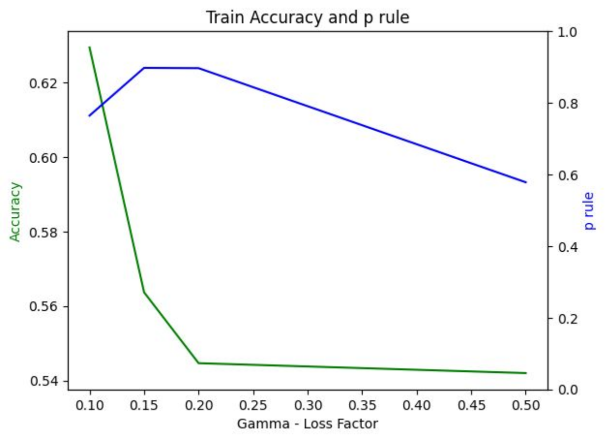


```
P% for the data without optimization:  (0.470369589852681, 0.24189944134078212, 0.5142752562225475)
Actual Train Loss:  0.6053670398617021
test accuracy: 0.6671087533156499
calibration for the data without optimization:  0.05568958147689318
parity for the data without optimization:  0.23152434562749488
Equality of odds(True negative) for the data without optimization:  0.19551407077180272
```

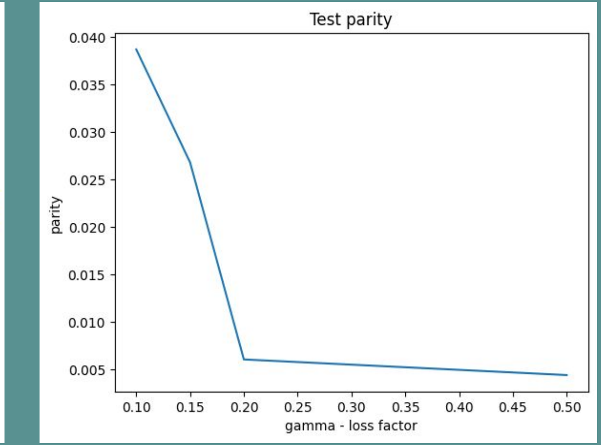


```
Test Accuracy for gamma  0.1 :  0.6087533156498673
P Rule percent:  0.8344241446694998
calibration:  0.08445081973151347
parity:  0.03867191178920981
Equality of odds:  0.00529395374756203

Test Accuracy for gamma  0.15 :  0.5477453580901857
P Rule percent:  0.7044728434504792
calibration:  0.13959705287865942
parity:  0.02680518426753023
Equality of odds:  0.02477013095569791

Test Accuracy for gamma  0.2 :  0.5225464190981433
P Rule percent:  0.7886117409926934
calibration:  0.1553686437301225
parity:  0.006078256648772394
Equality of odds:  0.00025076623014763655

Test Accuracy for gamma  0.5 :  0.5106100795755968
P Rule percent:  0.7827476038338659
calibration:  0.1539269594951932
parity:  0.004433722370737431
Equality of odds:  0.01005851212036779
```

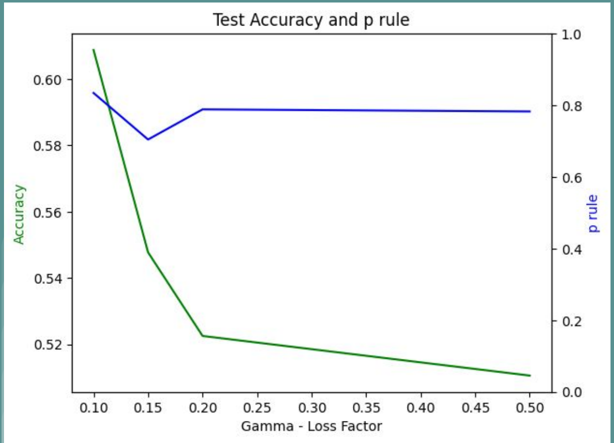
A3: Maximizing Fairness Under Accuracy Constraints



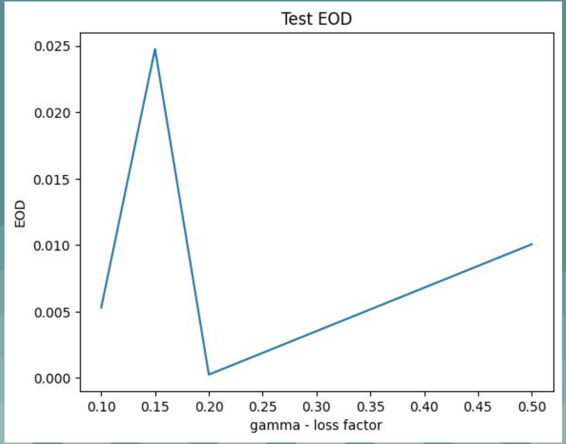
Train Accuracy and p rule



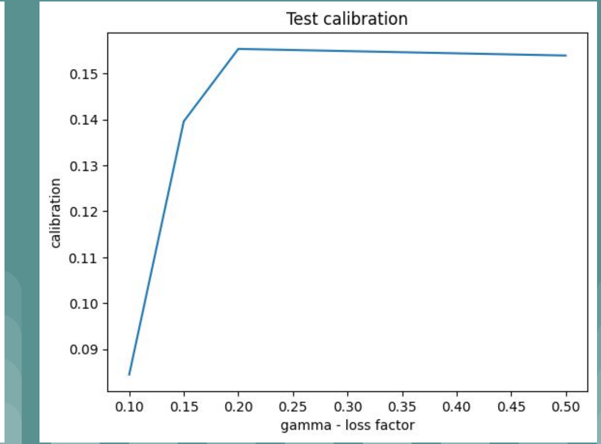
Test parity



Test Accuracy and p rule



Test EOD



Test calibration

Conclusion

In conclusion, it is possible to balance fairness and accuracy in classification tasks through our A1 and A3 algorithms. Both of them have shown good results in promoting fairness in binary classifiers. Further exploration of these algorithms could potentially lead to the development of fair and accurate machine learning models for real-world applications.

Thanks for listening!