CID - 01911284

1. Introduction

This coursework involved exploring the effect of regularisation on fairness in both standard and fairness-aware machine learning models. For additional exploration, the performance of Florida-trained models on Texas data was analysed.

2. Task 1 – Standard Model

Throughout this coursework, logistic regressors were chosen for their easily variable regularisation/trade-off parameter C. During hyper-parameter tuning, a grid search was employed to observe the effect of varying the 'C' and different 'solver' methods. The search space is described in Table 1.

Table 1: Hyper-parameter Search Space

Parameter	Values
C	10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1
Solver	newton-cg, liblinear, sag, saga, lbfgs

2.1. Validation Set Results and Model Selection

Through grid search, 35 unique models were evaluated on a validation dataset. The metrics presented in Figure 1 represent the mean values obtained from 5 independent trials for each model.

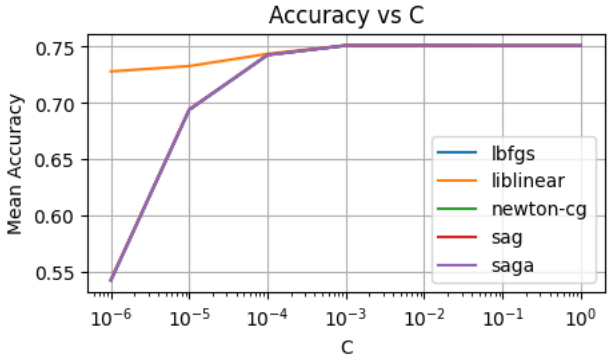

Figure 1: Average accuracy against varying values of C for a standard logistic regressor.

As seen in Figure 1, increased regularization (through decreasing C) reduced model accuracy, over-constraining the model until it was unable to capture underlying patterns. However, accuracy is only reduced once C drops below 10^{-3}. This suggests that C only affects accuracy after a certain threshold is reached. Additionally, it is shown that the solver method has little impact on model performance. This is observed throughout Tasks 1 and 2.

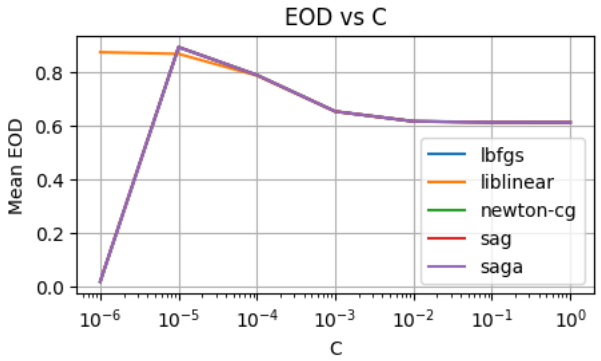

Figure 2: Average EOD against varying values of C for a standard logistic regressor.

Theoretically, regularisation can decrease bias in a model by penalising high dependencies on discriminatory features and reducing model complexity [1]. This is supported by the initial increase in EOD from C = 10^{-6} to C = 10^{-5}. However, EOD subsequently reduces as C increases, eventually plateauing at EOD = 0.6. This suggests that basic regularisation techniques cannot achieve adequate fairness in machine learning models.

For model selection based on the best accuracy and fairness, hyperparameters can be seen in Table 2.

Table 2: Selected Standard Models via Validation Data

Model	Accuracy	EOD	C	Solver
Most Accurate	0.751	0.616	0.01	lbfgs
Most Fair	0.542	0.019	10^{-6}	lbfgs

2.2. Test Set Results

The selected models were then evaluated using the held-out test set. These results can be seen in Table 3:

Table 3: Performance of Selected Standard Models

Model	Accuracy	EOD
Most Accurate	0.752	0.674
Most Fair	0.542	0.019

This consistent performance on the test data suggests the model is acceptable, generalising well to unseen data.

3. Task 2 – Fairness-Aware Model

To create a fairness-aware model, reweighing can be applied to the logistic regressor [2]. Once done, the same grid search and model selection criterion can be employed.

3.1. Validation Set Results and Model Selection

As in Task 1, the C was varied to observe its effect on the fairness-aware model's accuracy and EOD.

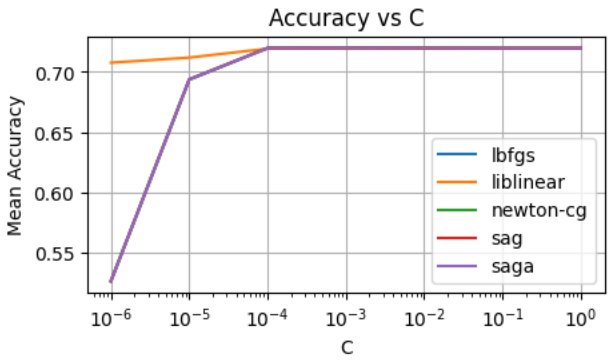

Figure 3: Average accuracy against varying values of C for a logistic regressor with reweighting.

The trend shown in Figure 3 is very similar to Figure 1 as accuracy increases with C (reduced regularisation). However, small differences include a slight reduction in maximum accuracy (72% in comparison to 75%) and the decreased effective range of C. In the fairness-aware model, C only impacts accuracy when below C = 10^{-4}, not C = 10^{-3}.

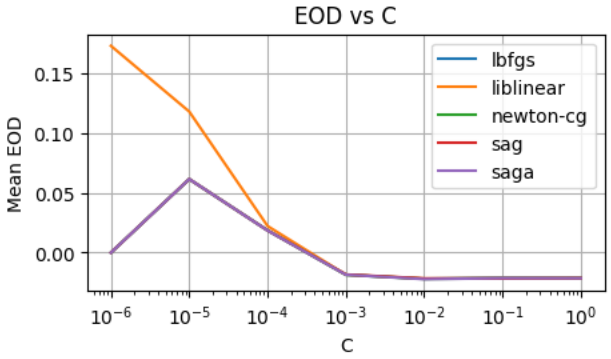

Figure 4: Average EOD against varying values of C for a logistic regressor with reweighting.

Like Figure 2, Figure 4 shows an initial spike in EOD as C increases to 10^{-5} before decreasing and eventually plateauing. However, EOD values in the fairness-aware model are greatly reduced, being close to 0 and reflecting high fairness. Notably, when C is equal to or greater than 10^{-3}, EOD values become negative, slightly favouring the underprivileged class instead.

As seen in Figures 2 and 3, this suggests that reweighting is highly effective in training fair machine learning models with minimal sacrifice in performance.

Hyperparameters for the best accuracy and fairness models can be seen in Table 4.

Table 4: Selected Fairness-Aware Models via Validation Data

Model	Accuracy	EOD	C	Solver
Most Accurate	0.720	-0.018	0.001	saga
Most Fair	0.526	0.000	10^{-6}	lbfgs

3.2. Test Set Results

As in Task 1, the selected models were evaluated on the held-out test set. These results can be seen in Table 5:

Table 5: Performance of Selected Fairness-Aware Models

Model	Accuracy	EOD
Most Accurate	0.720	0.005
Most Fair	0.526	0.000

4. Task 3 – Model Selection Strategy

4.1. Proposed Criterion and Justification

To select a model that is both accurate and fair, a proposed criterion (utility function 'U') should be maximised using the formula below.

$$U = \max(0, 2A - 1) - EOD + abs(A - 1 - abs(EOD))$$

where A is accuracy and EOD is equality of opportunity difference. While seemingly complicated, each term represents a desired aspect of the final model.

First, max(0, 2A – 1) is the model's reward function for accuracy. As a binary classification problem, arguably the worst performance possible is that of a random coin toss, having 50% accuracy. Hence, the max function produces 0 utility for model accuracy below 50% and linearly increasing utility until 100% is reached (U = 1).

Second, –EOD is the model's penalty function for fairness. As lower EOD values are desirable, EOD should be minimised and therefore, EOD is multiplied by -1. Beyond this, no other functions are applied to keep the contribution's magnitude between 0 and 1.

Lastly, abs(A – 1 – abs(EOD)) is the penalty function representing the accuracy-fairness trade-off, penalising models that greatly favour one metric over the other. If EOD represents unfairness (ranging from 0 to 1), then 1 – EOD represents fairness (also ranging from 0 to 1). As such, fairness can be described as 1 – abs(EOD). By taking the absolute difference between fairness and accuracy, a metric for accuracy-fairness difference is derived, resulting in abs(A – 1 – abs(EOD)).

Each term has been engineered to fall within [0, 1], ensuring that each term's contribution is uniformly valued.

4.2. Model Selection with Proposed Criterion

Using "U", the standard and fairness-aware models were re-selected, having utility (U) values of 0.629 and 1.160 respectively. Performance and hyperparameters are shown in Table 6.

*Note the most accurate fairness-aware model from Task 2 was selected again due to having the largest "U" value.

Table 6: Selected Models based on Proposed Criterion "U"

Model	Accuracy	EOD	C	Solver
Standard	0.543	0.020	10^{-6}	sag
Fairness-Aware	0.720	-0.018	0.001	saga

While the grid search resulted in models of ranging performance, model variance was still limited, having multiple models that perform similarly. This led to the selection of models that closely resemble those that were previously selected. For the standard model, the proposed criterion selects a model that is very similar to the "most fair" model. For the fairness-aware model, the selected model is also the "most accurate" model.

4.3. Test Set Results

As in Task 1 and 2, the newly selected models were evaluated on the held-out test set. These results can be seen in Table 6:

Table 7: Performance of Selected Fairness-Aware Models

Model	Accuracy	EOD	Utility
Standard	0.545	0.024	0.636
Fairness-Aware	0.720	0.005	1.159

The proposed criterion was semi-effective, resulting in the selection of a good fairness-aware model, but an arguably poor standard model.

While the standard model is very fair, its 54% accuracy is extremely poor. Alternatively, other models with acceptable accuracy scores (approximately 70+%) could have been chosen despite their higher EOD scores (~60%) during validation. For model selection, it's arguably more advantageous to use a biased but high-performing model over one that's fair but ineffective, highlighting the complex trade-off between accuracy and fairness.

Future work includes improving the criterion with weighted accuracy/fairness contributions to better reflect their perceived utility. Furthermore, non-linear functions can be employed, reflecting the utility of accuracy and EOD combinations at different levels. For example, logarithmic functions can be used to reflect the decreasing additional utility that incremental improvements in accuracy and fairness bring.

5. Additional Exploration – Florida-trained Model Performance on Texas Data

By testing models outside the environments within which they were trained, model robustness and generalisation ability can be analysed. Key points for observation include the impact on model accuracy and EOD and model selection using the proposed criterion. For this exploration, the previously selected 6 six models are used again to compare performance.

5.1. Test Results on Texas Data

Model performances on Texas data are shown in Table 8.

Table 8: Performance of Previous Models on Texas Dataset

Model		Accuracy	EOD	Utility
Standard	Most Accurate	0.720	0.665	0.391
	Most Fair	0.596	0.029	0.789
	Highest Utility	0.600	0.040	0.800
Fairness-Aware	Most Accurate /Highest Utility	0.688	0.003	1.064
	Most Fair	0.579	0.000	0.738

Derived from Table 8, Table 9 shows the difference in performance from Florida test data to Texas test data.

Table 9: Performance Difference between Florida and Texas data

Model		Accuracy	EOD	Utility
Standard	Most Accurate	-0.032	-0.009	-0.014
	Most Fair	0.054	0.01	0.164
	Highest Utility	0.055	0.016	0.164
Fairness-Aware	Most Accurate /Highest Utility	-0.032	-0.002	-0.095
	Most Fair	0.053	0	0.16

As seen in Table 9, the model performance on Texas data both improved and worsened compared to Florida tests.

Expectedly, the "most accurate' models worsened in all metrics, having slight reductions in accuracy and minuscule decrements in EOD and utility.

Conversely, the "most fair" models (as well as the "highest utility" standard model) saw improvements in accuracy, EOD, and utility. Notably, these models increased their accuracies by 5+%, which is a small but significant improvement from their previously abysmal accuracy scores of ~55%.

5.2. Evaluation and Conclusion284

The worsened performance of the 'most accurate' models suggests that proficient models will experience a decline in performance when tested outside their training environment [3]. This is supported by the fact that patterns are not guaranteed to persist from environment to environment, rendering the model unable to leverage these learned trends. Fortunately, performance only decreased slightly, suggesting that Texas has similar hiring patterns to Florida, making it nearly just as predictable.

The increased performance by other models can be explained by their previously poor performance, having only achieved high fairness at the expense of terrible accuracy, akin to that of a random model. As patterns for prediction were hardly learned in the training environment (Florida), it is likely to observe improved performances when testing in other environments (like Texas).

6. References

[1] Kamishima, T., Akaho, S., & Sakuma, J. (2011). Fairness-aware Learning through Regularization Approach. https://doi.org/10.1109/ICDMW.2011.83

[2] Lee, J. G., Roh, Y., Song, H., & Whang, S. E. (2021). Machine Learning Robustness, Fairness, and their Convergence. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 21, 4046–4047. https://doi.org/10.1145/3447548.3470799

[3] Adragna, R., Creager, E., Madras, D., & Zemel, R. (2020). Fairness and Robustness in Invariant Learning: A Case Study in Toxicity Classification. https://github.com/adragnar/irm-toxicity-classification