Assignment: Effect of regularisation on accuracy-fairness trade-off

Deadline: refer to Cate

Summary: You have to submit a **4-page report** using the provided **template** and an **implementation code**. For submission, zip your report and code into a single file called fairness.zip.

The main task of the assignment is to study model selection that takes into account accuracy and fairness metrics when training the machine learning models. The task is to compare standard machine learning models versus fairness-based machine learning models with the following model selection criteria: most accurate, most fair, and most accurate+fair.

Task 1. The standard machine learning methods such as logistic regression, support vector machines, multi-layer perceptron use a trade-off hyperparameter (lambda, learning rate) to trade-off accuracy and generalisation (Lecture 2). The first task regularisation is to analyse whether or not better generalisation could correspond to fairer models.

Specifically, for a given machine learning model of your choice, use **training data** to perform 5-fold cross validation. By varying the trade-off hyperparameter, select (1) the model with the highest accuracy, and (2) the model with the best fairness metric across 5 folds. After this model selection step, compute and report final accuracy and fairness for both models (1) and (2) **on the test set**.

Task 2. Now choose an algorithmic fairness method, e.g. reweighing (Lecture 2), and perform the same analysis, i.e. how varying the hyperparameter(s) impacts accuracy and fairness metrics during model selection.

Specifically, using the same 5-fold cross validation as in Task 1, perform validation by varying the trade-off hyperparameter and select (3) the model with the highest accuracy, and (4) the model with the best fairness metric across 5 folds. After model selection, compute and report final accuracy and fairness for both models (3) and (4) **on the test set**.

Task 3. Based on your observations, suggest a model selection strategy (criterion) that accounts for both, accuracy and fairness. Compare the standard machine learning model versus the fairness-based machine learning model using the proposed criterion - what is the effect (if any)?

Specifically, using the same 5-fold cross validation as before, perform validation by varying the trade-off hyperparameter. Select (5) the best standard model, and (6) the best fairness-based model using the proposed criterion. After model selection, compute and report final accuracy and fairness for both models (5) and (6) **on the test set**.

Metrics. As metrics, evaluate and report accuracy and fairness metric of equality of opportunity, i.e. true positive rate (TPR) difference between the sensitive groups (Lecture 1). The sensitive groups are dataset specific, use your best judgement, or interest.

Reporting final results. In total we have six models (1)-(6) with accuracy and fairness results on the test set to be reported and analysed. You may report these results as a table. You may include the plots of cross validation results (ox: hyperparameters, oy: performance measure – eg. accuracy, fairness, both) to demonstrate how model selection has been performed.

Dataset Perform the empirical evaluations on at least two datasets: the **AdultIncome dataset** available from the aif360 library and a dataset of your choice (e.g. from the same library).

The train/test split should be set as 0.7/0.3.

Toolboxes You can use any of your favourite classifiers. Some of the classifiers that we have discussed in the class are: logistic regression, neural networks (multi-layer perceptron), support vector machines (Lecture 2). Feel free to use some of the machine learning toolboxes such as Weka [1] (in Java), scikit-learn [2] (in Python), shogun [3] (in C++), or stats [4] (in Matlab). Feel free to use PyTorch [5], or any other deep learning frameworks (JAX, TensorFlow) [6].

Feel free to use the fairness-based toolbox such as aif360 [7]. See Lecture 2 and the Tutorial session on how to use this toolbox.

Details of the report

You are expected to write a report with up to 4-pages. Please use the provided latex or word template. Make sure to describe the key results, observations and conclusions in the main text. You may include supplementary plots in the appendix. References may go on an additional fifth page. You must also submit your implementation codes. Please make sure it is possible to run your code as is.

[1] http://www.cs.waikato.ac.nz/ml/weka/

[2] https://scikit-learn.org/stable/

[3] http://www.shogun-toolbox.org

[4] https://uk.mathworks.com/help/stats/index.html

[5] https://pytorch.org/

[6] Colab allows you to enable a GPU accelerator.

[7] https://aif360.readthedocs.io/en/latest/index.html

Marking Criteria

70% – 100% **Excellent**

Shows very good understanding supported by evidence that the student has extrapolated from what was taught, through extra study or creative. Work at the top end of this range is of exceptional quality. Report will be excellently structured, with proper references and discussion of existing relevant work. The report will be neatly presented, interesting and clear with a disinterested critique of what is good and bad about approach taken and thoughts about where to go next with such work. Possible options how to extrapolate from what was taught:

e.g. reweighing

- Perform an analysis of algorithmic fairness methods beyond binary sensitive features, for example, *Race or Age*. The student should describe how to adapt the fairness metrics and/or methods to a non-binary sensitive feature and report them in the empirical evaluations.

- Perform an analysis of the machine learning methods beyond lecture materials. The report should describe the approach in sufficient details, its advantages and disadvantages comparing to other (taught) methods. Methodological and empirical evidence of the approach tested should be presented in the report.

- Compare two scenarios when sensitive attribute is a feature of the main input X and when we exclude it from the features of X. Analyse how this change the performance of the models 1-6 using at least one of the datasets.

Important: The report should clearly indicate the extra content if any, e.g. by having a specific section.

60% – 69% **Good**

The work will be very competent in all respects. Work will evidence substantially correct and complete knowledge, though will not go beyond what was taught. Report should be well-structured and presented with proper referencing and some discussion/critical evaluation. Presentation will generally be of a high standard, with some discussion of related work.

55% – 59% **Satisfactory**

Will be competent in most respects. There may be minor gaps in knowledge, but the work will show a reasonable understanding of fundamental concepts. Report will be generally well-structured and presented with references, though may lack depth, appropriate critical discussion or discussion of further developments, etc.

50% – 54% **Borderline**

The work will have some significant gaps in knowledge but will show some understanding of fundamental concepts. Report should cover the fundamentals but may not cover some aspects of the work in sufficient detail. The work may not be organised in the most logical way and presentation may not be always be appropriate. There will be little or no critical evaluation or discussion. References may be missing, etc.

30% – 49% **Poor**

The work will show inadequate knowledge of the subject. The work is seriously flawed, displaying major lack of understanding, irrelevance or incoherence. Report badly organised and incomplete, possibly containing irrelevant material. May have missing sections, no discussion, etc.

Below 30% **unacceptable (or not submitted)**

Work is either not submitted or, if submitted, so seriously flawed that it does not constitute a bona-fide report/script.