CO557 - Ethics in AI
Algorithmic fairness methods in machine learning
Instructor: Dr Viktoriia Sharmanska

Submission: **April 27, 2020**.

Coursework:
Effect of regularisation on accuracy-fairness trade-off.

The standard machine learning (ML) methods such as logistic regression, support vector machines use a regularization parameter to trade-off accuracy and generalization (lambda in the lecture slides).

1. Perform an analysis how varying this hyperparameter improve/worsen/satisfy the fairness metric when optimising for accuracy. Analyse and report both, accuracy and fairness metrics, when varying the hyperparameter.

2. Choose an algorithmic fairness method (e.g. reweighing), and perform an analysis how varying the hyperparameter(s) improve/worsen/satisfy the fairness and accuracy metrics.

3. Based on 1-2, suggest potential strategies for model selection when training a fair and accurate ML model. Perform several (e.g. five) random train/test splits and report the results using mean and standard deviation over the splits.

Perform the empirical evaluations on (at least) two datasets: *Adult Income* and *Compas* (covered in the lab session, can be downloaded from the aif360 library). Analyse at least two fairness metrics along with accuracy.

Details of the report:
You are expected to write a 4-page report. Please use the provided latex or word template. You must also submit your implementation codes. Please make sure it is possible to run your code as is.

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:
1) An analysis of algorithmic fairness methods beyond binary sensitive features. The student should describe how to adapt the fairness metrics and/or methods to a non-binary sensitive feature and report them in empirical evaluations.
2) An analysis of the fairness 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.

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 clear written style and 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 organized 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 organized 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.