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

Deadline: 31 Jan 2024

Summary: You have to submit a **3-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 a model selection procedure 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-aware machine learning models with the following model selection criteria: most accurate, most fair, and most accurate & fair.

Task 1. Standard machine learning models such as logistic regression, support vector machines, multi-layer perceptron use free parameters, called hyperparameters, such as regularisation coefficient, number of hidden layers, learning rate, etc. to trade-off complexity and generalisation (Lecture Part 2). The first task is to analyse whether or not better generalisation could correspond to a fairer model.

Task 1 a) Specifically, split the data into train/test sets (***the train/test split should be set as 0.7/0.3***). Further split the train set into two sets: train-train/train-val sets (***the train-train/train-val split should be set as 0.8/0.2***). Repeat the procedure of splitting train set into train-train/train-val sets 4 more times (make sure you shuffle the train data before you repeat this procedure). In the end, you should have ***5 sets of train-train/train-val data splits***.

The above procedure is a standard method for estimating the performance of a machine learning algorithm. A single run of train/val split may result in a noisy estimate of model performance. Different splits of the data may result in very different results. *Repeated experiments* provides a way to improve the estimated performance of a machine learning model. This involves simply partitioning train data into train-train/train-val multiple times and reporting the mean result across all train-train/train-val sets. This *mean result* is expected to be a more accurate estimate of the true unknown underlying mean performance of the model on the dataset.

Task 1 b) Perform model selection by varying the trade-off hyperparameter and select the model ***with the highest accuracy*** across 5 sets of train-train/train-val data splits. After model selection, compute and report final accuracy and fairness for this standard model **on the test set**.

Task 1 c) Perform model selection by varying the trade-off hyperparameter and select the model ***with the best fairness metric*** across 5 sets of train-train/train-val data splits. After model selection, compute and report final accuracy and fairness for this standard model **on the test set**.

Task 2. Now choose a fairness-aware method, e.g. reweighing of Kamiran and Calders 2012, learning fair representations of Zemel et al. 2013 (Lecture Part 2), and perform the same analysis, i.e. how varying the hyperparameter(s) impacts accuracy and fairness metrics during model selection.

Task 2 a) Specifically, using the same 5 sets of train-train/train-val data splits as in Task 1, vary the trade-off hyperparameter and select the model ***with the highest accuracy*** across 5 sets of train-train/train-val data splits. After model selection, compute and report final accuracy and fairness for this fairness-aware model **on the test set**.

Task 2 b) Using the same 5 sets of train-train/train-val data splits as in Task 1, vary the trade-off hyperparameter and select the model ***with the best fairness metric*** across 5 sets of train-train/train-val data splits. After model selection, compute and report final accuracy and fairness for this fairness-aware model **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-aware machine learning model using the proposed criterion - what is the effect (if any)?

Task 3 a) Specifically, using the same 5 sets of train-train/train-val data splits as before, vary the trade-off hyperparameter and select ***the best standard model*** using the proposed criterion. After model selection, compute and report final accuracy and fairness for this model **on the test set**.

Task 3 b) Specifically, using the same 5 sets of train-train/train-val data splits as before, vary the trade-off hyperparameter and select ***the best fairness-aware model*** using the proposed criterion. After model selection, compute and report final accuracy and fairness this model **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 Part 1). The main task is binary classification with a binary sensitive feature to be analysed.

Reporting final results. In total we have **six models (from tasks 1b, 1c, 2a, 2b, 3a, 3b) 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 model selection results (ox: hyperparameters, oy: performance measure – eg. accuracy, fairness, both) to demonstrate how hyperparameters selection has been performed.

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 Part 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

[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

free to use PyTorch [5], or any other deep learning frameworks (JAX, TensorFlow) [6].

For the fairness-aware methods, you are encouraged to use AI Fairness 360 (AIF360) [7]. Consult the Tutorial session on how to use this toolbox.

Dataset Perform the empirical evaluations on the **ACS American Community Survey dataset** available from the folktables. This ACS dataset is the more recent, more comprehensive version of the Adult Income (which was based on 1994 Census database) dataset that you have used in the Tutorial, for more information please read: Ding, Frances and Hardt, Moritz and Miller, John and Schmidt, Ludwig, Retiring Adult: New Datasets for Fair Machine Learning, Advances in Neural Information Processing Systems, 2021.

Please use the 2018 yearly ACS, data for Florida state, and the ACSEmployment as a prediction task (predict whether an individual is employed). This is done as follows (first, you need to install folktables: `pip install folktables`):

```
import folktables
from folktables import ACSDataSource
import numpy as np

#(Age) must be greater than 16 and less than 90, and (Person weight) must be
    greater than or equal to 1
def employment_filter(data):
    """
    Filters for the employment prediction task
    """
    df = data
    df = df[df['AGEP'] > 16]
    df = df[df['AGEP'] < 90]
    df = df[df['PWGTP'] >= 1]
    return df

ACSEmployment = folktables.BasicProblem(
    features=[
        'AGEP', #age; for range of values of features please check Appendix B.4 of
            Retiring Adult: New Datasets for Fair Machine Learning NeurIPS 2021 paper
        'SCHL', #educational attainment
        'MAR', #marital status
        'RELP', #relationship
        'DIS', #disability recode
        'ESP', #employment status of parents
        'CIT', #citizenship status
        'MIG', #mobility status (lived here 1 year ago)
        'MIL', #military service
        'ANC', #ancestry recode
        'NATIVITY', #nativity
        'DEAR', #hearing difficulty
        'DEYE', #vision difficulty
```

[5] https://pytorch.org/

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

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

```
        'DREM', #cognitive difficulty
        'SEX', #sex
        'RAC1P', #recoded detailed race code
        'GCL', #grandparents living with grandchildren
    ],
    target='ESR', #employment status recode
    target_transform=lambda x: x == 1,
    group='DIS',
    preprocess=employment_filter,
    postprocess=lambda x: np.nan_to_num(x, -1),
)

data_source = ACSDataSource(survey_year='2018', horizon='1-Year', survey='person')
acs_data = data_source.get_data(states=["FL"], download=True) #data for Florida
    state
features, label, group = ACSEmployment.df_to_numpy(acs_data)
```

The above example use disability (binary) as the sensitive attribute.

You need to pre-process the data so that it is AIF360-ready, for example, you can do as follows:

```
from aif360.datasets import StandardDataset
import pandas as pd

data = pd.DataFrame(features, columns = ACSEmployment.features)
data['label'] = label

favorable_classes = [True]
protected_attribute_names = [ACSEmployment.group]
privileged_classes = np.array([[1]])
data_for_aif = StandardDataset(data, 'label', favorable_classes = favorable_classes,
                      protected_attribute_names = protected_attribute_names,
                     privileged_classes = privileged_classes)
privileged_groups = [{'DIS': 1}]
unprivileged_groups = [{'DIS': 2}]
```

Details of the report

You are expected to write a report with up to 3-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 fourth page. You must also submit your source code. 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:

- Propose a well-researched, well-founded solution for Task 3.
- Perform an analysis of algorithmic fairness methods beyond binary sensitive features, for example, *Race* (RAC1P feature in the ACS dataset) or *Age* (AGEP feature in the ACS dataset). 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.
- 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.
- Perform model selection on data for Florida state (states=[”FL”]), while test on data for Texas (states=[”TX”]). Analyse how this change the performance of the models 1-6.

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.