02901 Advanced Topics in Machine Learning Summer School: Fairness and Machine Learning

Xiao Hu
Email: xiahaa@space.dtu.dk

Abstract

The short reports presents the mini-project done for the course 02901 Advanced Topics in Machine Learning with regarding to the topic of Fairness and Machine Learning.

Introduction

With the increasing use of machine learning in our society, the concern of fairness of machine learning is rising nowadays(Dwork et al. 2012), especially within the field such as health care, criminal risk assessment, social services. This motivates the central topic of this year's summer school. Throughout five-days lectures, the following topics have been presented:

- Algorithmic fairness: fairness for arbitrary computer algorithms;
- Fairness through awareness: pioneering work on enforcing fairness as a constrained optimization problem;
- Measuring fairness: definitions of fairness;
- Explainability of neural network;
- Causal Bayesian Network and its application to machine ethics and fairness;

This report will mainly cover the second and third points.

Definitions We begin with some definitions (or measures) of the fairness. The four commonly used measures of fairness are (Žliobaitė 2019):

- Equal parity (also known as demographic or statistical parity): equal number selected from each group;
- Proportional parity: equal proportion selected from each group;
- False positive parity: the false positive rates for different groups will be equal;
- False negative parity: the false negative rates for different groups will be equal;

Apparently, there will be conflicts among the four measures. Certain measures are favored in certain cases, e.g. false positive parity for punitive situations whilst false negative parity for preventative situations.

The necessity of awareness There will be inevitably some bias in the data which will make the trained model biased. Simply remove explicit sensible attributes or features cannot guarantee to completely remove bias since implicit correlation may exist in the data, e.g. postal code with race, working hours with gender, etc. Consequently, fairness awareness is required when designing machine learning algorithm to explicitly deal with bias and discrimination.

Approaches Generally speaking, current mainstream approaches for fairness can be divided into three categories:

1. Pre-processing: dedicatedly select the data to remove bias or find representations that do not contain sensitive information.
2. Post-processing: perform corrections on model outputs or predictions.
3. During training: enforce fairness by imposing constraints, regularization, or using an adversary.

From the legal perspective, it is generally considered bad to manipulate decision or inputs but suitable by learning with constraints and re-sampling data.

Training with Fairness Constraints

This section briefly review the training mechanism for fairness classification proposed by (Zafar et al. 2017). The proposed mechanism aims to design fairness-awareness convex-margin-based classifiers (e.g. logistic regression). In section , we show the performance of this mechanism on some real dataset.

Measure for fairness The Measure for fairness in (Zafar et al. 2017) is defined with the p%-rule which is supported by the U.S. Equal Employment Opportunity Commission. Suppose p_a and $p_{\bar{a}}$ being the percentage of subjects having (and not having, resp.) a certain sensitive attribute assigned the positive decision, the p%-rule states that the ratio between p_a and $p_{\bar{a}}$ should be no less than p:100. We further formally define the p%-rule in the binary classification. Denoting $\mathbf{x}$ as the feature vector for binary classification, $y \in \{-1, 1\}$ as the class labels, $f(\mathbf{x})$ as the mapping function from $\mathbf{x}$ to y, for margin-based classifiers, we find $f(\mathbf{x})$ by seeking for a decision boundary $d_{\boldsymbol{\theta}}(\mathbf{x})$ supported by a

set of parameters $\boldsymbol{\theta}$ where $f(\mathbf{x}) = 1$ given $d_{\boldsymbol{\theta}}(\mathbf{x}) \geq 0$ and $f(\mathbf{x}) = -1$ otherwise. The optimum parameters $\boldsymbol{\theta}^*$ can be found by minimizing a loss function $L(\cdot)$ over a training set $\{\mathbf{x}_i, y_i\}_{i=1}^N$, i.e., $\boldsymbol{\theta}^* = argmin_{\boldsymbol{\theta}} L(\mathbf{x}_i, y_i, \theta)$. Considering a sensitive attribute z, the p%-rule requires that:

$$min(\frac{P(d_{\boldsymbol{\theta}}(\mathbf{x}) \geq 0)|z=1}{P(d_{\boldsymbol{\theta}}(\mathbf{x}) \geq 0)|z=0}, \frac{P(d_{\boldsymbol{\theta}}(\mathbf{x}) \geq 0)|z=0}{P(d_{\boldsymbol{\theta}}(\mathbf{x}) \geq 0)|z=1}) \geq \frac{p}{100} \tag{1}$$

This can be generalize to multiple sensible attributes $\mathbf{z} \in \mathbb{R}^m$. Since it is challenging to incorporate the p%-rule directly in training (non-convexity), the decision boundary covariance is proposed to measure fairness and define classifiers satisfying the p%-rule approximately. The decision boundary covariance is defined with $\mathbf{z}$ and $d_{\boldsymbol{\theta}}(\mathbf{x})$ as follows:

$$\begin{aligned} Cov(\mathbf{z}, d_{\boldsymbol{\theta}}(\mathbf{x})) &= \mathbb{E}[(\mathbf{z} - \bar{\mathbf{z}}) d_{\boldsymbol{\theta}}(\mathbf{x})] - \mathbb{E}[(\mathbf{z} - \bar{\mathbf{z}})] \bar{d}_{\boldsymbol{\theta}}(\mathbf{x}) \\ &\approx \frac{1}{N} \sum_{i=1}^{N} (\mathbf{z} - \bar{\mathbf{z}}) d_{\boldsymbol{\theta}}(\mathbf{x}) \end{aligned} \tag{2}$$

It is clear that the decision boundary covariance (2) is a convex function with respect to the decision boundary parameters $\boldsymbol{\theta}$ for convex margin-based classifiers like logistic regression or SVM since $d_{\boldsymbol{\theta}}(\mathbf{x})$ is convex to $\boldsymbol{\theta}$. This facilitates adding fairness without increasing the complexity of training.

Optimization framework Based on the decision boundary covariance (2), two optimization frameworks can be derived. The first optimization aims to maximize accuracy under fairness constraints, which can be written as:

$$\begin{aligned} &\text{minimize: } L(\boldsymbol{\theta}) \\ &\text{subject to: } \frac{1}{N} \sum_{i=1}^{N} (\mathbf{z} - \bar{\mathbf{z}}) d_{\boldsymbol{\theta}}(\mathbf{x}) \leq \mathbf{c}, \\ &\qquad \frac{1}{N} \sum_{i=1}^{N} (\mathbf{z} - \bar{\mathbf{z}}) d_{\boldsymbol{\theta}}(\mathbf{x}) \geq -\mathbf{c} \end{aligned} \tag{3}$$

where $\mathbf{c}$ is the covariance threshold that trades of fairness and accuracy (the smaller $\mathbf{c}$ is, the tighter the p% rule and larger loss in accuracy would be).

The second optimization treats fairness in a soft way by maximizing fairness under accuracy constraints. By treating accuracy as hard constraints, this optimization framework can be used to train a classifier which causes least possible disparate impacts while subjecting the given accuracy constraints:

$$\begin{aligned} &\text{minimize: } |\frac{1}{N} \sum_{i=1}^{N} (\mathbf{z} - \bar{\mathbf{z}}) d_{\boldsymbol{\theta}}(\mathbf{x})| \\ &\text{subject to: } L(\boldsymbol{\theta}) \leq (1 + \gamma) L(\boldsymbol{\theta}^*) \end{aligned} \tag{4}$$

where $L(\boldsymbol{\theta}^*)$ is the optimal loss over the training set provided by the unconstrained classifier. $\gamma \geq 0$ specifies the maximum tolerant loss increase with respect to $L(\boldsymbol{\theta}^*)$. In case that the loss function is additive over the training set, i.e. $L(\boldsymbol{\theta}) = \sum_{i=1}^{N} L_i(\boldsymbol{\theta})$, the aforementioned loss constraint can be customized for each individual separately in the training set by setting different γ_i.

Logistic Regression logistic regression classifier maps $\mathbf{x}$ to y by means of a probability distribution:

$$p(y = 1|\mathbf{x}, \boldsymbol{\theta}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \mathbf{x}}}$$

The corresponding loss function can be written as $L(\theta) = -\sum_{i=1}^{N} log(p(y_i|\mathbf{x}_i, \boldsymbol{\theta}))$. The decision boundary is simply the hyperplane defined by $d_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta}^T \mathbf{x} = 0$. Thus, taking the first optimization as an example, (3) will adopt the form as:

$$\begin{aligned} &\text{minimize: } -\sum_{i=1}^{N} log(p(y_i|\mathbf{x}_i, \boldsymbol{\theta})) \\ &\text{subject to: } \frac{1}{N} \sum_{i=1}^{N} (\mathbf{z} - \bar{\mathbf{z}}) \boldsymbol{\theta}^T \mathbf{x} \leq \mathbf{c}, \\ &\qquad -\frac{1}{N} \sum_{i=1}^{N} (\mathbf{z} - \bar{\mathbf{z}}) \boldsymbol{\theta}^T \mathbf{x} \leq \mathbf{c} \end{aligned} \tag{5}$$

Summary

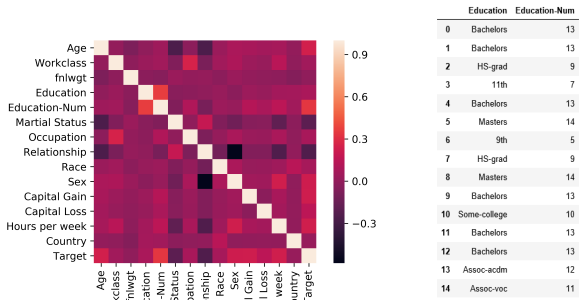
- By approximating the computationally intractable p% rule with the decision boundary covariance, fairness can be model as a convex function for margin-based classifier, which is computationally tractable and optimization-feasible.
- With the decision boundary covariance, fairness can be explicitly treated in training either as the hard constraint (first optimization) or as the optimization objective function (soft fashion, second optimization).
- Apparently, the way of dealing with fairness belongs to the third category, as mentioned in the introduction section.
- If the boundary decision is not a convex function on $\boldsymbol{\theta}$, then the fairness constraint will make the original unconstrained optimization problem non-convex. This may make the training more complex.
- Only work for margin-based classifier.

Experiments

Here we evaluate the performance on real dataset. The Adult dataset (Dua and Graff 2017) is used here. The Adult dataset contains 45222 subjects, each with 14 features and a binary label indicating whether a subject's incomes is above (positive) or below (negative) 50K USD. 30000 samples are randomly sampled from the whole dataset (70% for training). The implementation is based on the repository[1]. However, there are some bugs that needs to be fixed. The notebook is available[2].

[1] github.com/mbilalzafar/fair-classification.

[2] github.com/xiahaa/machine_learning_course/adv_summer_school



	Education	Education-Num
0	Bachelors	13
1	Bachelors	13
2	HS-grad	9
3	11th	7
4	Bachelors	13
5	Masters	14
6	9th	5
7	HS-grad	9
8	Masters	14
9	Bachelors	13
10	Some-college	10
11	Bachelors	13
12	Bachelors	13
13	Assoc-acdm	12
14	Assoc-voc	11

Figure 1: Correlation of Adult dataset.

Figure 2: Correlation of Education and Education Num.

Sensitive Attributes As sensitive attributes, we considered "sex" (female and male).

Preliminary Analysis A brief review of the Adult data-set could be seen from Fig 1. It is clear that there is a strong correlation (see Fig 2) between "Education" and "Education-Num". Assuming "Education" as the sensible attribute, simply removing "Education" will not solve the fairness problem because "Education-Num" correlates with "Education". Although this is a pretty simple example, it verifies the aforementioned statement that make sensible attributes blind will not solve the bias and fairness due to the correlation of data.

First Optimization: Maximizing accuracy under fairness constraint Here we present the experimental result by training logistic regression using the (2) with different values of covariance threshold **c**. From the left figure, we can see apparently that by decreasing **c** (tightening fairness constraint), the p% rule will increase (indicating more fair with respect to the sensible attribute). The right figure shows the trade-off between accuracy and fairness. The relative loss is

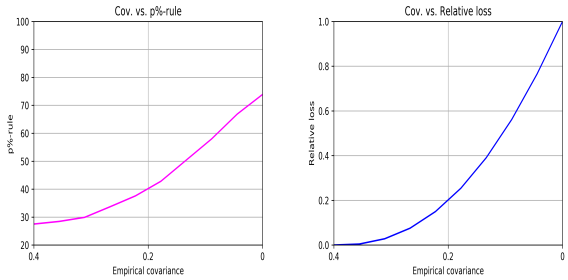


Figure 3: Maximizing accuracy under fairness constraints: single sensible attribute ("sex"). Left figure shows the the p% rule variation with respect to the empirical covariance threshold **c**. Right figure shows the relative loss variation with respect to the empirical covariance threshold **c**.

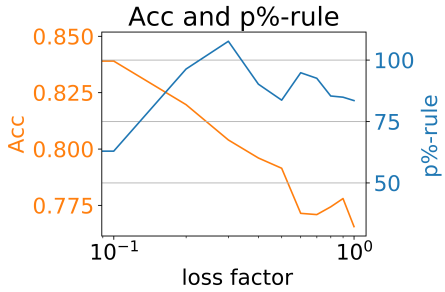


Figure 4: Maximizing fairness under accuracy constraint: single sensible attribute ("sex"). Loss factor denotes the γ in (3).

computed as $L_{rel} = \frac{L_{cons} - L_{uncons}}{max\{L_{cons} - L_{uncons}\}}$. L_{cons} is the loss of constrained optimization and L_{uncons} is the loss of unconstrained optimization.

Second Optimization: Maximizing fairness under accuracy constraint In this experiment, we train the model with the second optimization. γ is ranged from 0 to 1 with a step of 0.1. We compute the accuracy as well as the corresponding p% rule values. Intuitively speaking, we would assume by increasing γ, accuracy will drop while the p% rule will rise, which partially matches with the trend shown in Fig 4. However, continuously increasing γ didn't further improve the p% rule but degraded the accuracy, which should be investigated deeply (unfortunately, due to the time issue, I haven't found a explanation to this issue).

Conclusion & Discussion

Fairness is an important issue for data analysis and machine learning. Although fairness is relatively a very young research field, it is very fortunate that public concern has been raised. From the definition of fairness in different perspectives, it is apparent that no uniform fairness can be defined and attained. We need to match fairness according to the purpose of the task. Increasing number of approaches have been proposed to deal with this issue. However, it is still preliminary since many methods only works for traditional machine learning algorithms rather than deep neural network. Nonetheless, be aware of fairness will definitely help us build better and more fair machine learning algorithms.

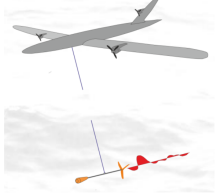
References

Dua, D., and Graff, C. 2017. UCI machine learning repository.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226. ACM.

Žliobaitė, I. 2019. Measuring algorithmic fairness.

Zafar, M. B.; Valera, I.; Rogriguez, M. G.; and Gummadi, K. P. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, 962–970.



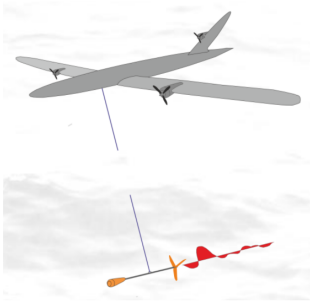
Poster session for 02901 Advanced Topics in Machine Learning

Xiao Hu
xiahaa@space.dtu.dk

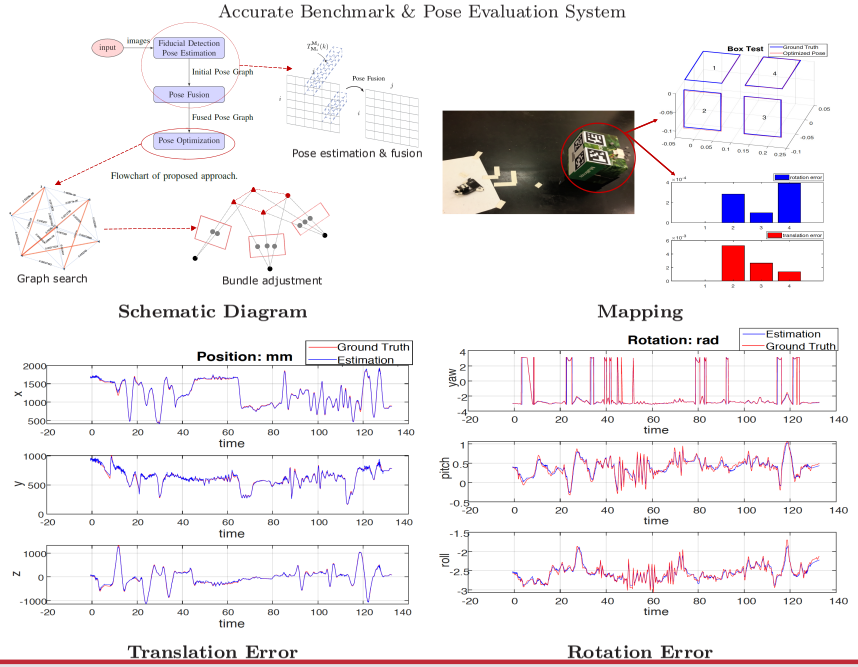
Project: UAV-QMS

The UAV-QMS project aims to develop a cost-efficient and long range Unmanned Aerial Vehicle for high-Quality Magnetic Surveying. Due to the sensitivity of the surveying sensors, direct positioning the magnetic bird via traditional sensors like GPS is infeasible. Relative pose estimation using computer vision will be investigated in the project. The focus of the PhD thesis is to investigate:

- Pose estimation.
- Navigation.



Pose Estimation



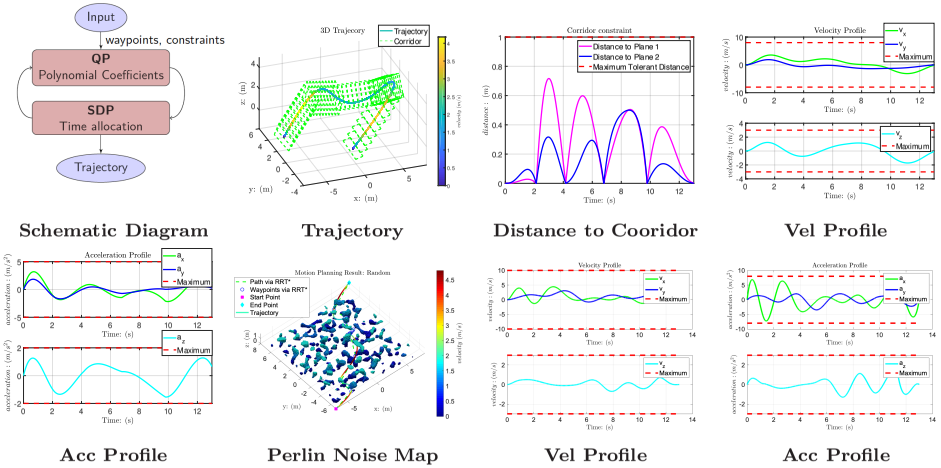
References

[1] Hu, X., et al. A Novel Robust Approach for Correspondence-Free Extrinsic Calibration. 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2019).

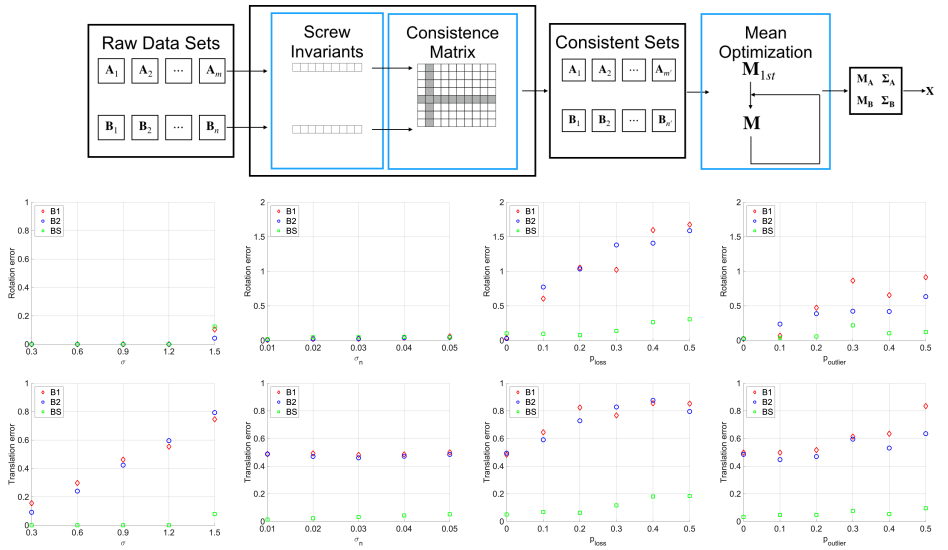
[2] Hu, X., Olesen, D., & Knudsen, P. (2019). Trajectory Generation Using Semidefinite Programming For Multi-Rotors. In Proceedings of the European Control Conference (ECC 2019) (pp. 2577-2582). IEEE.

[3] Hu, X., Jakobsen, J., Knudsen, P., & Wei , J. (2018). Accurate Fiducial Mapping For Pose Estimation Using Manifold Optimization. In 2018 International Conference on Indoor Positioning and Indoor Navigation (IPIN) (pp. 206-212).

Navigation



Calibration&Synchronization



ML&Fairness

Machine learning will be applied in this project for object detection, tracking, pose estimation. Learning based pose estimation is still very controversial, especially compared with traditional geometry based solutions where the underlying mathematics are known explicitly. In this case, bias existed in the data may result in biased pose estimation. **Fairness** could be helpful to design unbiased pose estimation netwrok.