SMART DUBAI

DUBAI AI ETHICS: SUPPLEMENTARY DOCUMENT

Feedback Draft, 24th September 2018

دبـــي الذكية
SMART DUBAI

Pointers to key resources for technical experts

Feedback Draft, 24th September 2018

This document aims to help technical experts (e.g. data scientists, machine learning engineers, AI engineers) investigate how to apply ethical AI.

Note that this is not a comprehensive document. Ethical AI is an evolving field, with further advances expected.

A list of resources is provided, for reference only, rather than a formal bibliography. Relevant entities should evaluate on their own the suitability and risks of each method.

Contents

Solutions to build and measure fairness in AI systems

Methods proposed for providing fair AI can be divided into two categories: (1) Discrimination discovery: detecting and evaluating discrimination that exist in data; (2) Discrimination prevention: building models that tend not to make discriminatory decisions even if trained from a biased dataset.

1. **Detect bias and evaluate fairness** [1][2]
    a. **Evaluate 'Fairness' as equal error rates / accuracy across different groups:**
        e.g. Equal True Positive Rates, Equal False Negative Rates, Overall Accuracy Equality, etc.
    b. **Evaluate 'Fairness' as equal mean values, or equal positive rates across different groups:**
        e.g. Mean Difference, Normalised Difference, Balance in score for positive/negative class, Calibration, AUC Parity, etc.
    c. **Evaluate 'Fairness' as conditional independence over certain features**
        e.g. Given certain characteristics of an individual, the decisions on them should be independent of their demographic group (e.g. gender)

    Online articles:

    - *"Mirror Mirror" - Reflections on Quantitative Fairness*
      https://speak-statistics-to-power.github.io/fairness/
    - *Fairness Measures*
      http://fairness-measures.org/

    Journals, Proceedings, or Work in progress:

    - *Group Fairness Under Composition, C. Dwork and C.Ilvento, 2018*

    d. **Evaluation of Individual Fairness**
        i.e. two individuals who are similar for a specific task should be classified/rated similarly

    Journals, Proceedings, or Work in progress:

    - *Learning fair representations, Zemel et al., 2013*
    - *Fairness through awareness, Dwork et al., 2011*
    - *Individual Fairness Under Composition, C. Dwork and C.Ilvento, 2018*

[1] *Evaluation metrics partially adapted from Mirror Mirror- Reflections on Quantitative Fairness (see https://speak-statistics-to-power.github.io/fairness/). Licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.*

[2] *Evaluation metrics partially adapted from Fairness Measures (see http://fairness-measures.org/). Licensed under a CC-BY-4.0 license.*

e. Machine Learning methods for detection and evaluation

Journals, Proceedings, or Work in progress:

- *k-NN as an Implementation of Situation Testing for Discrimination Discovery and Prevention, Luong et al., 2011*
- *Combating discrimination using bayesian networks, K. Mancuhan and C. Clifton, 2014*
- *Data mining for discrimination discovery, Ruggieri et al., 2014*

f. Statistical methods for detection and evaluation

Journals, Proceedings, or Work in progress:

- *Measuring racial discrimination, Panel on Methods for Assessing Discrimination, Blank et al., 2004.*
- *Statistical evidence of discrimination, D. Kaye, 1982*
- *The uses and misuses of statistical proof in age discrimination claims, T. Tinkham., 2010*

2. Remove unfairness

a. Remove unfairness by preprocessing training datasets

Journals, Proceedings, or Work in progress:

- *Certifying and removing disparate impact, Feldman et al., 2015*
- *Quantifying explainable discrimination and removing illegal discrimination in automated decision making, Kamiran et al., 2013*
- *Combating discrimination using bayesian networks, K. Mancuhan and C. Clifton, 2014*

b. Remove unfairness in models by adding a regulariser

Journals, Proceedings, or Work in progress:

- *Fairness-aware classifier with prejudice remover regularizer, Kamishima et al., 2012*
- *Learning fair representations, Zemel et al., 2013*
- *Controlling attribute effect in linear regression, Calders et al., 2013*
- *Discrimination aware decision tree learning, Kamiran et al., 2010*
- *Learning fair representations, Dwork et al., 2013*

c. Remove unfairness by post-processing trained models

Journals, Proceedings, or Work in progress:

- *A methodology for direct and indirect discrimination prevention in data mining, S. Hajian and J. Domingo-Ferrer, 2013*

d. Remove unfairness by post-processing model outcomes

Journals, Proceedings, or Work in progress:

- *Discrimination aware decison tree learning, F. Kamiran, T. Calders, and M. Pechenizkiy, 2010*

Other resources on AI Fairness:

Course Materials:

- *CS294: Fairness in Machine Learing (Lectures), UC Berkeley*
  https://fairmlclass.github.io/

Tools:

- *Fairness Measures - Code Repository*
  https://github.com/megantosh/fairness_measures_code/tree/master
- *FairML: Auditing Black-Box Predictive Models*
  https://github.com/adebayoj/FairML

Solutions to AI systems documentation

3 possible methods to document and explain data have been proposed:

1. **Datasheets for datasets**
    - A method to document how and why a dataset was created, what information it contains, what tasks it should and should not be used for, and whether it might raise any ethical or legal concerns
2. **Data statement schema**
    - Built for NLP systems
    - Seek to address ethics, exclusion, and bias issues in NLP systems
3. **Dataset nutrition label**

Journals, Proceedings, or Work in progress:

- *Datasheets for datasets, Timnit Gebru et al., 2018*
- *Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science, Emily M. Bender and Batya Friedman,*
- *The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards, Sarah Holland et al., 2018*

Methods for general documentation:

1. **Supplier's Declarations of Conformity**
    - A method that gives a comprehensive list of items and Q&As about AI systems to be documented
2. **Social Impact Statement for Algorithms**
    - A documentation method proposed to ensure Ethical AI

Journals, Proceedings, or Work in progress:

- *Increasing Trust in AI Services through Supplier's Declarations of Conformity, Hind et al., 2018*

Online articles:

- *Principles for Accountable Algorithms and a Social Impact Statement for Algorithms, FAT/ML* http://www.fatml.org/resources/principles-for-accountable-algorithms

Solutions to provide explainable AI

In general there are 3 types of methods to provide explainable AI[3]:

1. **Use directly interpretable models**
    - e.g. Linear Model, Logistic Regression, Decision Tree, Decision Rules (If-Then rules), RuleFit, Naive Bayesian Methods, K-Means Clustering, Random Forest / Boosting, Hidden Markov Models, etc.
    - Explanations of the model should be provided by e.g. listing features selected and their weights

    Books:

    - *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable, Molnar. C. , 2018*
      https://christophm.github.io/interpretable-ml-book/

    Reports:

    - *Explainable AI - Driving business value through greater understanding, PricewaterhouseCoopers LLP. , 2018*
      https://www.pwc.co.uk/audit-assurance/assets/explainable-ai.pdf

2. **Use model-agnostic interpretation methods**
    - Methods that do not specify the underlying ML models, could in general be applicable to black-box models
    - e.g. Partial Dependence Plot, Individual Conditional Expectation, Feature Interaction, Feature Importance, Global Surrogate Model, Accumulated Local Effects plot, Local Surrogate Model, Shapley Value Explanations, etc.

    Books:

    - *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable, Molnar. C. , 2018*
      https://christophm.github.io/interpretable-ml-book/

    Journals, Proceedings, or Work in progress:

    - *Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models, Apley, D. W., 2016*
    - *"Why Should I Trust You?" Explaining the Predictions of Any Classifier, Ribeiro et al., 2016*

[3] *Types of methods adapted from Interpretable Machine Learning - A Guide for Making Black Box Models Explainable, Molnar. C. , 2018 (see https://christophm.github.io/interpretable-ml-book/). Licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.*

Tools:

- *LIME Python package & R package*
  https://github.com/marcotcr/lime
  https://cran.r-project.org/web/packages/lime/index.html

3. **Use example-cased interpretation methods**
   - e.g. Counterfactual Explanations, Prototype and Criticisms, Influential Instances, and etc.

Books:

- *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable, Molnar. C. , 2018*
  https://christophm.github.io/interpretable-ml-book/

Journals, Proceedings, or Work in progress:

- *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR, S. Wachter, B. Mittelstadt and C. Russell, 2018*
- *Inverse Classification for Comparison-based Interpretability in Machine Learning, Laugel et al., 2017*
- *"Examples are not Enough, Learn to Criticize! Criticism for Interpretability", B. Kim, R. Khanna, and O. Koyejo, 2016*

Methods have also also been proposed according to 3 different stages of AI system development:[4]

1. **Pre-building**
   e.g. Visualisation, Exploratory data analysis
2. **Building**
   e.g. Rule-based, per-feature-based, Case-based, Sparsity method, Monotonicity method
3. **Post-buidling**
   e.g. Sensitivity analysis, gradient-based methods, Mimic/Surrogate models, Investigation of hidden layers

Tutorials:

Interpretable Machine Learning: The fuss, the concrete and the questions, B.Kim and F. Doshi-Velez, 2017

[4] *Adapted from Tutorials: Interpretable Machine Learning: The fuss, the concrete and the questions, B.Kim and F. Doshi-Velez, 2017 (see https://people.csail.mit.edu/beenkim/papers/BeenK_FinaleDV_ICML2017_tutorial.pdf).*

https://people.csail.mit.edu/beenkim/papers/BeenK_FinaleDV_ICML2017_tutorial.pdf

Other resources on AI Explainability:

Journals, Proceedings, or Work in progress:

- *Towards A Rigorous Science of Interpretable Machine Learning, F. Doshi-Velez and B. Kim, 2017*

Other resources

Organisations and Conferences

ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*)

https://fatconference.org/

ACM SIGKDD

http://www.kdd.org/

NIPS Conference

ICML Conference

Reading lists on ethical AI

Toward ethical, transparent and fair AI/ML: a critical reading list

https://medium.com/@eirinimalliaraki/toward-ethical-transparent-and-fair-ai-ml-a-critical-reading-list-d950e70a70ea