Literature review for Fairness Evaluation in Machine Learning

Prepared by-Sumaiya Saima Sultana

Paper 1: Fairness Definitions Explained(Sahil et al 2018)

Contribution of the paper:

- Collected most popular definitions of Fairness & categorized them from three perspectives: A)Statistical, B) Similarity based, C) Casual
- explained their rationale behind the definitions and
- used the definitions to check if a certain dataset exhibits gender related bias

Dataset used: The German Credit dataset 2018 (Link)

Attribute	Attribute
Credit amount	Installment rate
Credit duration	Property
Credit purpose	Residence
Status of existing checking amount	Period of present residency
Number of existing credits	Personal status and gender
Credit history	Age
Installment plans	Foreign worker
Dependents	Other debtors
Employment	Employment length

Target attribute: Credit score (Binary)
Model used: Logistic regression classifier

Keywork: Authors explored whether married/divorced female applicants get unfair treatment comparing with married/divorced male applicants according to various definitions of fairness known from the literature.

(N.B: Dataset does not contain instances from single female)

Paper: Fairness Definitions Explained(Sahil et al 2018)

A) Statistical measures of fairness

Type A: Definitions based on only predicted outcome

1. Group fairness/ statistical parity / equal acceptance rate/ benchmarking
2. Conditional statistical parity

Type B: Definitions based on both predicted and actual outcome

1. Predictive parity/outcome test
2. False positive error rate balance /predictive equality
3. False negative error rate balance/ equal opportunity
4. Equalized odds/ conditional procedure accuracy equality/ disparate mistreatment
5. Conditional use accuracy equality
6. Overall accuracy equality
7. Treatment equality

Type C: Definitions based on predicted probabilities and actual outcome

1. Test fairness/ Calibration/ Matching conditional frequencies
2. Well-calibration
3. Balance for positive class
4. Balance for negative class

B) Similarity based measures of fairness:

1. Casual discrimination
2. Fairness through unawareness
3. Fairness through awareness

C) Casual Reasoning based measures of fairness:

1. Counterfactual fairness
2. No unresolved discrimination
3. No proxy discrimination Fair inference

Statistical based measures of fairness evaluation

Group fairness/Statistical parity: A classifier satisfies this definition if subjects in both protected and unprotected groups have equal probability of being assigned to the positive predicted class.

Conditional statistical parity: The definition is satisfied if subjects in both protected and unprotected groups have equal probability of being assigned to the positive predicted class, controlling for a set of legitimate factors L.

Predictive parity: A classifier satisfies this definition if both protected and unprotected groups have equal PPV – the probability of a subject with positive predictive value to truly belong to the positive class. (/=FDR)

Predictive equality: A classifier satisfies this definition if both protected and unprotected groups have equal FPR – the probability of a subject in the negative class to have a positive predictive value.(/=TNR)

Equal opportunity: A classifier satisfies this definition if both protected and unprotected groups have equal FNR – the probability of a subject in a positive class to have a negative predictive value. (/=TPR)

Equalized odd: A classifier satisfies the definition if protected and unprotected groups have equal TPR and equal FPR.

Conditional use accuracy equality: A classifier satisfies the definition if protected and unprotected groups have equal PPV and equal NPV.

Over all accuracy equality: A classifier satisfies this definition if both protected and unprotected groups have equal prediction accuracy.

Treatment equality: A classifier satisfies this definition if both protected and unprotected groups have an equal ratio of false negatives and false positives.

Test-fairness/Calibration: A classifier satisfies this definition if for any predicted probability score S, subjects in both protected and unprotected groups have equal probability to truly belong to the positive class.

Well-calibration: This definition extends the previous one stating that, for any predicted probability score S, subjects in both protected and unprotected groups should not only have an equal probability to truly belong to the positive class, but this probability should be equal to S.

Balance for positive class: A classifier satisfies this definition if subjects constituting positive class from both protected and unprotected groups have equal average predicted probability score S.

Balance for negative class: A classifier satisfies this definition this definition states if subjects constituting negative class from both protected and unprotected groups also have equal average predicted probability score S.

Paper: Fairness Definitions Explained(Sahil et al 2018)

B) Similarity based measures of fairness:

1. Causal discrimination: A classifier satisfies this definition if it produces the same classification for any two subjects with the exact same set of attributes except protected attribute.

2. Fairness through unawareness: A classifier satisfies this definition if no sensitive attributes are explicitly used in the decision-making process.

3. Fairness through awareness: This definition is a more elaborated and generic version of the previous two: here, fairness is captured by the principle that similar individuals should have similar classification. The similarity of individuals is defined via a distance metric; for fairness to hold, the distance between the distributions of outputs for individuals should be at most the distance between the individuals.

C)Casual Reasoning based measures of fairness:

1. Counterfactual fairness
2. No unresolved discrimination
3. No proxy discrimination Fair inference

Paper 2: Situation testing (Highlights)

- **Definition:** Situation testing is an experimental method aiming to establish discrimination on the spot.[2] In the legal field, situation testing is a systematic research procedure for creating controlled experiments analyzing decision maker's candid responses to applicant's personal characteristics.
- **Goal:** The aim of the method is to reveal and record discriminatory practices whereby a person who possesses a particular characteristic is treated less favorably than a person who does not possess this characteristic in a comparable situation.
- **Methodology:** The approach looks for pairs of people with similar characteristics apart from membership to a protected-by-law group.[1]
    - Given past records of decisions taken in some context, **for each member of the protected group with a negative decision outcome** (someone who may claim to be a victim of discrimination) we look for testers with similar, legally admissible, characteristics, **apart from being or not in the protected group**. If we can observe significantly different decision outcomes between the testers of the protected group and the testers of the unprotected group, we can ascribe the negative decision to a bias against the protected group. Similarity is modelled via a distance function. Testers are searched among the k-nearest neighbors.

Situation testing: Pseudo-algorithm

- **Step 1:** Select data samples for which Sensitive attribute(Gender) value = Protected group (Female) & target attribute value = not favorable outcome (default payment in this case)
- **Step 2:** Define distance function for different data types and set threshold for final distance function
  - Interval scale: Standardize using z-score and calculate absolute difference between z scores
  - Nominal scale: Distance is binary function testing equality
  - Ordinal scale: Mapped to interval-scaled values and absolute differences were measures
- **Step 3:** Pre-processing of data to adopt the distance functions
- **Step 4:** Calculate distance for each selected sample with non-protected group data samples
- **Step 5:** For each of the non-protected group data sample with distances lower than threshold:
  - If target attribute value = not favorable outcome
    - No unfairness found for the problem case
  - If target attribute value = favorable outcome
    - Unfair case detected!!

Paper 3: Fairness testing (Highlights)

- This paper defines software fairness and discrimination and develops a testing-based method for measuring discrimination. **This paper's main contributions are**:
  1. Formal definitions of **software fairness** and **discrimination**, including **a causality-based improvement** on the state-of-the-art definition of algorithmic fairness.
  2. **Themis**, a technique and open-source implementation— https://github.com/LASER-UMASS/Themis—formeasuring discrimination in software.
  3. **A formal analysis of the theoretical foundation of Themis**, including proofs of monotonicity of discrimination that led to provably sound **two-to-three orders of magnitude improvements in test suite size**, a proof of the relationship between fairness definitions, and a proof that Themis is more efficient on systems that exhibit more discrimination.
  4. An **evaluation of the fairness** of 20 real-world software instances (based on 8 software systems), 12 of which were designed with fairness in mind, demonstrating that (i) even when fairness is a design goal, developers can easily introduce discrimination in software, and (ii) Themis is an effective fairness testing tool.

- Two simplifying assumptions of this paper:
  1. **Software** under test has been defined as **black box**, that takes input characteristics and gives output; the authors claim internal complexity of the software has been avoided without losing generality.
  2. All the input and output characteristics are assumed to be **categorical variables**. This assumption simplifies measure of causality.

  Note: The definitons introduced here cannot be directly applied for non-categorical variables but can be altered using binning techniques.

  Authors identify themselves that this work does not apply to broader class of data types but will be extended in future.

Paper: Fairness testing(Highlights)

The fairness definitions mentioned in the paper are given below:

Casual discrimination: "We define software to be causally fair with respect to **input characteristic χ** if for all inputs, **varying the value of χ does not alter the output**. For example, a sentence recommendation system is fair with respect to race if there are no two individuals who differ only in race but for whom the system's sentence recommendations differs.

In addition to capturing causality, this definition **requires no oracle**—the equivalence of the output for the two inputs is itself the oracle—which helps fully automate test generation. Causal discrimination score seeks out causality in software and identifies **changing which characteristics directly affects the output.**

Group discrimination: Group discrimination says that to be fair with respect to an input characteristic, distribution of outputs for each group should be similar.

The Calders-Verwer (CV) score [19] measures the strength of group discrimination as the difference between the largest and the smallest outcome fractions; if 30% of people <40 get the loan, and 40% of people >40 get the loan, then loan is 40% − 30% = 10% group discriminating. (See paper for limitations of group discrimination)

Paper: Themis (Highlights)

- Themis is an automated test suite generator , an implementation of previously discussed paper, to measure two types of discrimination: Group discrimination & casual discrimination
  - **Group discrimination** is the maximum difference in the fractions of software outputs for each sensitive input group. For example, loan’s group discrimination with respect to race compares the fractions of green and purple applicants who get loans. If 35% of green and 20% of purple applicants get loans, then loan’s group discrimination with respect to race is 35% − 20% = 15%. With more than two races, the measure would be the difference between the largest and smallest fractions.
  - **Causal discrimination** is the frequency with which equivalences classes of inputs (recall Figure 1) contain at least two inputs on which the software under test produces different outputs. For causal discrimination, each equivalence class contains inputs with identical non-sensitive attribute values but varied sensitive attribute values. For example, loan’s causal discrimination with respect to age and race is the fraction of equivalence classes that contain a pair of individuals with identical name, income, savings, employment status, and requested loan amount, but different race or age, for which loan approves a loan for one but not the other.

Literature for robustness improvement in terms of fairness

1. Blake Lemoine, Brian Zhang, and M Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. (2018)

Methodology: For creating an unbiased model, this paper proposes a model in which they are trying to maximize accuracy of the predictor on target attribute, and at the same time minimize the ability of the adversary to predict the protected or sensitive attribute.

This adversarial approach can be incorporated with 3 different definitons of fairness measurement: Demographic Parity, Equality of Odds, Equality of Opportunity; making it generalized.

The authors also claim the adversarial approach described can be applied regardless of how simple or complex the predictor's model is, if the model is trained using a gradient-based method.

Literature for robustness improvement in terms of fairness

2. Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2018. Fairgan: Fairness-aware generative adversarial networks. In 2018 IEEE International Conference on Big Data (Big Data). IEEE, 570–575.

Methodology: This paper introduces a network that generates synthetic fair data. The generated data is similar to real data on every aspect except they are free from discrimination. This debiased data can be used for building models which successfully ensures fair model development.

This approach does not try to remove discrimination from the dataset, unlike many of the existing approaches, but instead generate new datasets like the real one which is debiased and preserves good data utility.

FairGAN is evaluated on data generation from two perspectives, fairness and utility. Fairness is to check whether FairGAN can generate fair data, while the utility is to check whether FairGAN can learn the distribution of real data precisely.