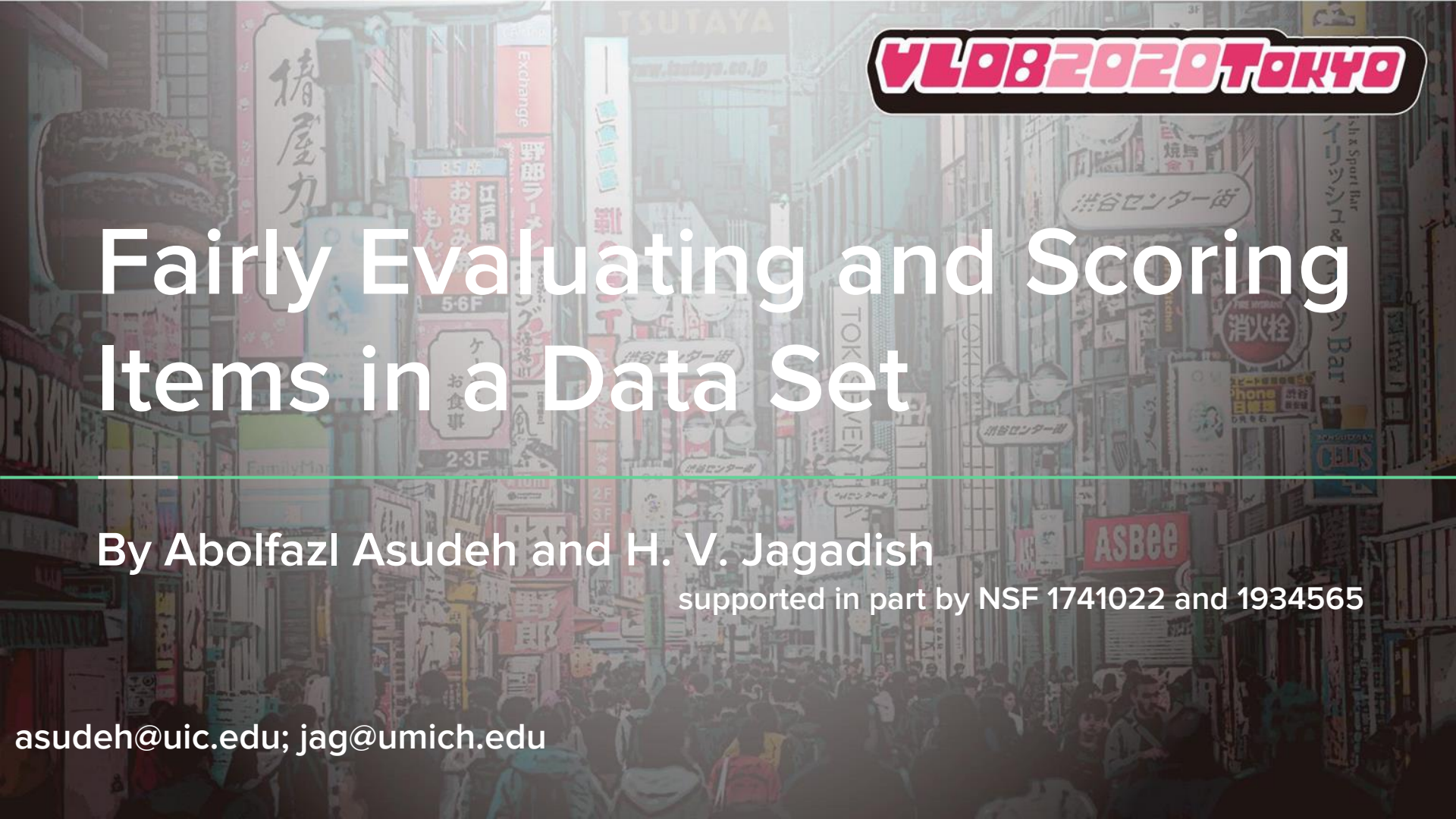


Fairly Evaluating and Scoring Items in a Data Set

By Abolfazl Asudeh and H. V. Jagadish

supported in part by NSF 1741022 and 1934565

asudeh@uic.edu; jag@umich.edu

Outline

Motivation and Definitions

Current State of the Art

Possible Directions

Scoring, Ranking, and Classification

Happens everywhere,

Scoring is a common way to perform ranking or classification. But scoring can have other uses, e.g. for selection. Similarly, ranking can be done without scoring, e.g. by pairwise comparison.

In this tutorial, we consider scoring for ranking and classification, performed each on their own and performed jointy.

Fairness

Fairness is an important requirement for any automated decision system [popularly referred to as “AI system”, whether or not this actually uses AI techniques]..

The most common use of ADS is for classification, which could sometimes be done by first computing a score. There are many other uses of ADS.

You heard a lot about fairness, and other important aspects of responsibility, in the keynote by Julia Stoyanovich.

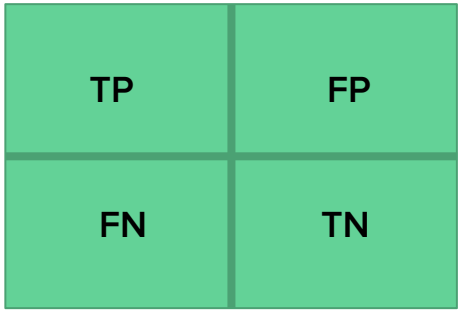
Our focus is score-based ranking and classificaiton.

What is Fairness

Surprisingly hard to define!!!

One very simple set up is to assume a classification task with a known correct answer, at least after the fact. Every classification algorithm is likely to have some error.

So, we define a matrix as in this figure:



An ideal classifier has only TP and FP.

But a real classifier has non-zero FN and TN

21 metrics

Positive Predictive Value: TP/(TP+FP)

Of those labeled positive, how many are truly positive

False Positive Rate: FP/(FP+TN)

How likely is an individual to be mistakenly labeled positive

Error Rate: (FP+FN)/(FP+FN+TP+TN)

How frequently does the system produce the wrong label

....

21 definitions of fairness (Narayanan)

For every protected group, compute metric of choice, and compare against the same metric for the population as a whole (Or another subgroup).

Not all definitions can be satisfied simultaneously, in general.

Even 3 can be impossible (Chouldechova, and several follow on papers).

21 fairness definitions and their politics

Arvind Narayanan

FAT* Tutorial, 2018

Fairness categories

- Notation:
  - $X = \{X_1, \ldots, X_m\}$: scoring attributes (aka features)
  - S: sensitive attribute(s) such as race and gender that identify demographic groups such as male, black, etc
  - Y: target variable (true label)

- At a high level, fairness definitions* fall into three categories:
  1. Independence
  2. Separation
  3. Sufficiency

* Excluding individual fairness

Independence

- A model satisfies independence if $f_\theta(X) \perp S$. That is, the outcome of the model is independent from the sensitive attribute(s)
- In a binary classification setting: $\mathrm{P}(f_\theta(X) = 1|S = a) = \mathrm{P}(f_\theta(X) = 1|S = b)$
- Fairness notions such as demographic parity and statistical parity follow the independence model

Separation

- f_θ satisfies separation, if its outcome is independent from the sensitive attribute(s) conditional on the target variable: $f_\theta(X) \perp S \mid Y$

- In case of binary classifier, separation is equivalent to requiring for all demographic groups *a, b* the two constraints

$$\mathrm{P}(f_\theta(X) = 1 | Y = 1, S = a) = \mathrm{P}(f_\theta(X) = 1 | Y = 1, S = b)$$
$$\mathrm{P}(f_\theta(X) = 1 | Y = 0, S = a) = \mathrm{P}(f_\theta(X) = 1 | Y = 0, S = b)$$

- Fairness notions such as Equalized Odds fall under the separation category.

Equalized Odds

Both True Positive Rate and False Positive Rate should be equal for protected subgroup and others.

Sufficiency

- f_θ satisfies sufficiency, under the same model outcomes, sensitive attribute(s) and the true outcome are independent: $Y \perp S \mid f_\theta(X)$. That is, the sensitive attribute and target variable are clear from the context.
- In case of binary classifier, separation is equivalent to requiring for all demographic groups *a, b*

$$\mathrm{P}(Y = 1 | f_\theta(X) = 1, S = a) = \mathrm{P}(Y = 1 | f_\theta(X) = 1, S = b)$$

- Fairness notions such as Predictive Parity fall under the sufficieny category.

Predictive Parity

Equal positive predictive value over subgroups.

We usually compare a protected subgroup (e.g. african-american) vs. the rest of of the population. But, we could equally well compute the PPV for each value of the protected attribute `race`.

Equal can never mean exactly equal in practice. We can have a ratio reported as a score. A threshold is used to declare unfairness where binary classification labels are required. Usually set at 80%.

Causality

If A affects B, which in turn affects C.

Then transitively, A affects C. However, A and C are independent given the value of B.

For example, race can affect socioeconomic status, which in turn can affect whether one is hired. If there is no other way that race affects hiring, then we can get hiring separated from race given socioeconomic status.

Disparate Treatment v. Disparate Outcome

Historically, and in law, we find two common “definitions” of fairness.

Each is really a class, since many things are left unspecified.

No disparate treatment is measured at the individual level, and requires that an individual not be treated differently on account of a protected attribute.

No disparate outcome is a group measure, and requires that the aggregate over the group of all individuals with a particular value of the protected attribute, the outcomes be similar. E.g. fraction of women selected for a job corresponds to fraction of women who applied (or to fraction of women in the population).

Subgroup fairness

The term “subgroup fairness” is often used to refer to the idea of “Demographic parity” in outcome. In other words, measure disparateness of impact for each protected group.

Subgroup fairness has also been used to refer to equalized error rates across groups.

Fairness in Machine Learning
Limitations and Opportunities

Solon Barocas, Moritz Hardt, Arvind Narayanan

Book available at:
https://fairmlbook.org

Fairness in Ranking

Ranking is a much more complex output than a binary class label.

Defining fairness is correspondingly more involved.

The simplest measures consider the top-k, and then address it as if it were a classification task -- items ranked in the top-k have one label and the rest have another. Now we can use the entire fairness framework for classification.

This, of course, begs the question of choosing k.

Answers obtained could be very different depending on the value chosen.

Exposure-based assessment

In many ranking situations, such as in information retrieval or recommender systems, higher ranked items get more attention than lower ranked ones.

There often is not a hard cut-off. But it is possible to define a monotonically decreasing function, such as inverse rank, that quantifies how much attention an item gets.

Now we can aggregate the attention received per protected group, and make that into a criterion against which we assess fairness.

Fairness of Exposure in Rankings

Ashudeep Singh and Thorsten Joachims:

KDD 2018: 2219-2228

Probability-Based Assessment

If ordered lists were created separately for each protected group, and the lists were then merged at random, how likely is it that we will observe the ordering reported?

In a random merge of these sorted lists, it should be the case that for any pair of groups, if we consider all pairs of items in the list from these groups, there should be approximately as many pairs items with the group A item ranked higher as the number with group B items ranked higher. The difference in this number is a measure of bias, related to the probability measure discussed above.

Measuring Fairness in Ranked Outputs

Ke Yang and Julia Stoyanovich

SSDBM 2017: 22:1-22:6

Bias can arise for many reasons

- Bias in the world
- Bias in our representation of the world
  - Bias in modeling choices
  - Bias in sampling

Bias in the World

Prejudice is common

Most of us, even if we try to be fair, have implicit biases

There is a long history of discrimination against people based on sex, religion, race, ethnicity, sexual orientation, and so on. These biases get reflected in the training data we see.

Bias in Modeling

If we use a standardized test as a proxy for intellectual ability, we are selecting particular aspects of intelligence to focus on, and we are ignoring the impact of test-taking skills and test preparation.

If exactly two values are allowed for gender, then we so not have the ability to represent other genders.

The list of modeling choices is very long, with personal biases getting reflected in these choices, often without even the modeler aware of it.

Bias in Sampling

Often we cannot get the data we want, so we use the data we can get.

We need opinions for all citizens, we use Twitter as a proxy, knowing that not everyone tweets, and this is a biased sample, which skews younger, more tech-savvy, better off, and so on.

We want to know the number of crimes committed, which is really unknown. So we use the number of crimes recorded by the police, which is not a random sample of the crimes committed.

Diversity

Merely having a good representative sample is not enough when the sizes of protected groups vary. If a learning system has a total error metric, it may do best by ignoring a small group altogether.

Most famously, this occurred in computer vision systems, repeatedly, since they did not get trained on enough dark skinned faces.

Diversity constraints require that there be enough representation of even small protected groups so that they are not ignored.

Diversity in Big Data: A Review.

Marina Drosou, H. V. Jagadish, Evaggelia Pitoura and Julia Stoyanovich:

Big Data 5(2): 73-84 (2017)

Coverage

There are several mathematical definitions of diversity. The simplest, and most popular of these is coverage.

Coverage is simply the count of elements in a given group.

Where intersectional fairness is considered, on multiple criteria, such as race, sex, and nationality, we get a combinatorial number of intersected groups. Coverage thresholds apply to the membership in each of these.

Stability

Traditionally, minimizing total (or average) error has been the single criterion used for any modeling task.

With recent efforts on fairness, additional criteria are considered, such as fairness and diversity.

One additional criterion of importance is stability -- that the results will not change if the model parameters are changed slightly. Without stability, the results can be very different with small changes to model parameters, leading one to suspect the modeling results.

Challenges

Responsible scoring and ranking is hard because

- Representative data are hard to find
- Bias in the world and in the modeling affect results
- There is no universal definition of fairness, and impossibility theorems show that not all definitions can be satisfied simultaneously.
- Many of these notions are not even defined, for ranking, since much of the existing work has been for classification.