Human-Centered Machine Learning: Measuring Fairness

Dong Nguyen

2021

Utrecht University

recap!

Last time: Intro to fairness

- Dual use
- What do we mean with fairness?
- Harms: Allocative harms, representational harms
- Feedback loops
- Statistical bias and societal bias
- Model development (optimization, evaluation)

Quiz

Provide an example of a ML system that would cause allocative harm.

- a system which assesses the ‘trustworthiness’ of a person
- a system that determines someone’s insurance policy
- a system that calculates a risk score for a person being fraudulent
- a diagnostic system in hospitals
- …

Quiz

Provide an example of a ML system that would cause representational harm.

- a system that outputs images of mug shots when searching a minority sounding name
- an automatic character generator for a game, that only produces male doctors and female nurses
- the choice of smart-home assistant providers to use female/feminine sounding voices as the standard, and give it a female name. It perpetuates sexist attitudes towards women.

Quiz

Spend a few minutes exploring the Open Images dataset

- most people in the images are white
- choice of objects/categories: Baseball glove, christmas tree, croissant (vs. menora, bao buns).

Quiz

Browse through https://www.kaggle.com/datasets...

- https://www.kaggle.com/spscientist/students-performance-in-exams

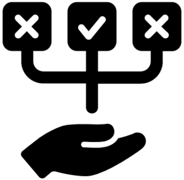
Plan for today

Today: How can we quantify the fairness of ML systems?

- Decision making
- Fairness at the group level
- Fairness at the individual level
- Beyond decision making (representations)

Decision making

Problem setup: decision making



We'll focus on decision making problems framed as *binary* classification tasks:

- Should this person be hired?
- Should this person be admitted to the university?
- Should this person receive parole?

Reminder: Allocative harms.

Human decision making

This is not a new problem!

Eren and Moren found that in the week following an upset loss suffered by the Louisiana State University (LSU) football team, judges imposed sentences that were 7% longer on average. The effect was driven by judges with undergraduate degrees at LSU (emotional impact?).

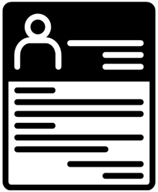
O. Eren and N. Mocan, Emotional Judges and Unlucky Juveniles, American Economic Journal: Applied Economics 10, no. 3 (2018): 171–205. [link]

Human decision making

This is not a new problem!

Example: Fictitious resume with only different names (e.g., gender, white-sounding vs. black-sounding names).

But there are caveats! And in some settings, these tests aren't possible.



See also Chapter 5 ("Testing Discrimination in Practice"); Part 1: Traditional tests for discrimination [link]

For a history of testing, see also 50 Years of Test (Un)fairness: Lessons for Machine Learning, Ben Hutchinson and Margaret Mitchell, FAT* 2019 [link]

Anti-discrimination law in the US

Disparate treatment

- *Intentional* discrimination
- Using protected attributes for classification

Disparate impact

- *Unintentional* discrimination
- *Unjustified* inequality in outcome

Protected classes in the US

- race (Civil Rights Act of 1964)
- religion (Civil Rights Act of 1964)
- national origin (Civil Rights Act of 1964)
- sex (Equal Pay Act of 1963 and Civil Rights Act of 1964)
- disability status (Rehabilitation Act of 1973 and Americans with Disabilities Act of 1990)
- ...

Netherlands

Dutch law specifies the following grounds of discrimination:

- race
- sex
- hetero- or homosexual orientation
- political opinion
- religion
- belief
- disability or chronic illness
- civil status
- age
- nationality
- working hours (full time or part time)
- type of contract (temporary or permanent)

Source: https://www.government.nl/topics/discrimination/prohibition-of-discrimination

~~Fairness through unawareness?~~

But my data doesn't contain a gender feature!

~~Fairness through unawareness?~~

But my data doesn't contain a gender feature!

Why is leaving out sensitive features not a solution?

~~Fairness through unawareness?~~

But my data doesn't contain a gender feature!

The remaining features may *correlate* with the sensitive features. This is often the case with large features spaces (most of modern ML!)

E.g., proxies (zip code for race)

~~Fairness through unawareness?~~

But my data doesn't contain a gender feature!

Amazon ditched AI recruiting tool that favored men for technical jobs

"[..] It penalized résumés that included the word "women's", as in "women's chess club captain". And it downgraded graduates of two all-women's colleges, according to people familiar with the matter."

https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine

(11 Oct 2018)

Problem setup

- Features: X
- Target variable/outcome: Y, e.g. {0,1} with binary classification
- We want to predict Y from X
- Often we have a score function R = r(X)
- We make a decision based on a threshold: $D = \mathbb{1}\{R > t\}$
- We have a sensitive attribute $A \in \{a, b\}$ (assuming two groups).

Problem setup

- Features: X
- Target variable/outcome: Y, e.g. {0,1} with binary classification
- We want to predict Y from X
- Often we have a score function R = r(X)
- We make a decision based on a threshold: $D = \mathbb{1}\{R > t\}$
- We have a sensitive attribute $A \in \{a, b\}$ (assuming two groups).

Should I give this person a loan?

- Features: income, debt, …
- Y: Will this person repay their loan? (1=yes, 0=no)
- D: Provide loan (1=yes, 0=no)
- A ∈ {male, female}

Confusion matrix

		Outcome (Y)	
		(+)	(−)
Decision (D)	(+)	TP $= 5$	FP $= 2$
	(−)	FN $= 3$	TN $= 5$

TP = true positive;
FP = false positive;
FN = false negative;

TN = true negative

True positive rate / Recall:
$P[D = +|Y = +] = \frac{TP}{TP+FN}$

False positive rate:
$P[D = +|Y = -] = \frac{FP}{FP+TN}$

True negative rate:
$P[D = -|Y = -] = \frac{TN}{FP+TN}$

False negative rate:
$P[D = -|Y = +] = \frac{FN}{TP+FN}$

Confusion matrix

Provide loan (D)	Pays back loan (Y) (+)	Pays back loan (Y) (−)
(+)	$\mathrm{TP} = 5$	$\mathrm{FP} = 2$
(−)	$\mathrm{FN} = 3$	$\mathrm{TN} = 5$

TP = true positive;
FP = false positive;
FN = false negative;
TN = true negative

Different stakeholders have different goals.

What would applicants find important? And what about the bank?

Plan for today

There is not one best way of measuring "fairness".

Terminology: privileged group, majority group (doesn't need to be the same, but often is).

Today: How can we quantify the fairness of ML systems?

- Decision making
- Fairness at the group level
- Fairness at the individual level
- Beyond decision making (representations)

Measuring fairness: Groups

Measuring fairness at the level of groups

Do outcomes systematically differ between different groups?

Three criteria:

equal decision measures *independence*	**conditional on outcome** *separation*	**conditional on decision** *sufficiency*
$A \perp D$	$D \perp A \mid Y$	$Y \perp A \mid D$

A=sensitive attribute; D=decision; Y=target variable/outcome

Measuring fairness at the level of groups

Do outcomes systematically differ between different groups?

Three criteria:

equal decision measures *independence*	**conditional on outcome** *separation*	**conditional on decision** *sufficiency*
$A \perp D$	$D \perp A \mid Y$	$Y \perp A \mid D$

A=sensitive attribute; D=decision; Y=target variable/outcome

Equal decision measures

$A \in \{a, b\}$ sensitive attribute; D is the decision

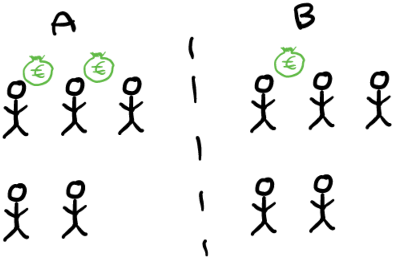
$$A \perp D$$

A generalization is: $A \perp R$.
In a binary classification scenario (e.g., $D = 1$ means hire this person):

$$P[D = 1 | A = a] = P[D = 1 | A = b]$$

The actual outcome is *not considered*
Also called: *demographic parity* or *statistical parity*.

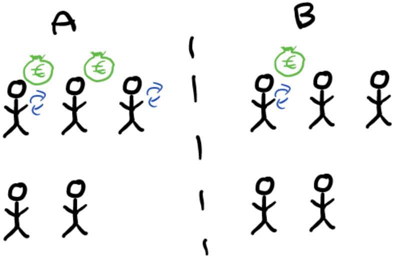
Equal decision measures



If group ***A*** and group ***B*** both apply for a loan at your bank, this is satisfied if an equal % applicants of group ***A*** and % applicants of group ***B*** are granted a loan. (Regardless of whether one group is more likely to repay.)

Here: *no*,
because: A: 2/5=0.4 vs. B: 1/5=0.2

Equal decision measures



Now, what if this classifier makes “no errors”, $(D = Y)$?

That is, all applicants who are selected indeed repay their loan and all others indeed would not have repaid their loan.

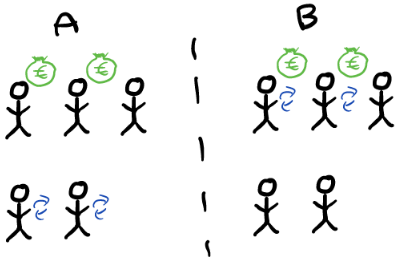
Statistical parity would not be satisfied!

Equal decision measures

Ignores the true outcome Y. Doesn't take "merit" of individuals into account. Why would we want this?

- It might very difficult or impossible to measure the actual outcome.
- We may believe that the observed relation between the attributes and outcome is unfair (e.g. historical prejudice).

Equal decision measures



Caveat: Statistical parity can be satisfied while procedure is unfair.

- E.g. having high accuracy in one group, and random predictions in the other group (as long as decision rates are equal).

Equal decision measures

We can relax this with a slack parameter:

$$|P[D = 1|A = a] - P[D = 1|A = b]| <= \epsilon$$

Or we could look at the ratio (a =unprivileged / b=privileged):

$$\frac{P[D = 1|A = a]}{P[D = 1|A = b]}$$

Relates to 80 percent rule in disparate impact law.
Example: Of the men applying at your company, you accept 60%. Of the women applying, you accept 30%. So: $0.3/0.6 = 0.5$, which is < 0.8.

Measuring fairness at the level of groups

Do outcomes systematically differ between different groups?

Three criteria:

equal decision measures	**conditional on outcome**	**conditional on decision**
independence	*separation*	*sufficiency*
$A \perp D$	$D \perp A \mid Y$	$Y \perp A \mid D$

A=sensitive attribute; D=decision; Y=target variable/outcome

Conditional on outcome

Informally: People with the same outcome should be treated the same.

$A \in \{a, b\}$ sensitive attribute; D is the decision; Y is the outcome

$$D \perp A | Y$$

A generalization is: $R \perp A | Y$.

In a binary classification setting: $D \perp A | Y = 1$ and $D \perp A | Y = 0$

Conditional on outcome

True positive rates/recall **(equal opportunity)**:

$$P[D = 1 | Y = 1, A = a] = P[D = 1 | Y = 1, A = b]$$

Example: Everyone who will repay a loan should have the same likelihood of receiving a loan (regardless of the sensitive attribute).

False positive rates:

$$P[D = 1 | Y = 0, A = a] = P[D = 1 | Y = 0, A = b]$$

Both constraints: **equalized odds**

A=sensitive attribute; D=decision; Y=target variable/outcome

Conditional on outcome

We need to know the (true) outcomes!

Often, it's hard or impossible to know the true outcomes.

- Hiring
- University admission
- ...

Conditional on outcome

True positive rate (=recall): $\frac{TP}{P}$.

		Truth (+)	**Truth** (−)
Pred	(+)	TP 1	FP 1
	(−)	FN 1	TN 2

		Truth (+)	**Truth** (−)
Pred	(+)	TP 2	FP 0
	(−)	FN 0	TN 3

What are the true positive rates?

Conditional on outcome

True positive rate (=recall): $\frac{TP}{P}$.

Pred \ Truth	(+)	(−)
(+)	TP 1	FP 1
(−)	FN 1	TN 2

TP = 0.5

Pred \ Truth	(+)	(−)
(+)	TP 2	FP 0
(−)	FN 0	TN 3

TP = 1

Measuring fairness at the level of groups

Do outcomes systematically differ between different groups?

Three criteria:

equal decision measures	conditional on outcome	conditional on decision
independence	*separation*	*sufficiency*
$A \perp D$	$D \perp A \mid Y$	$Y \perp A \mid D$

A=sensitive attribute; D=decision; Y=target variable/outcome

Conditional on decision

Informally: people with the same decision will have had similar outcomes (regardless of group).

$$Y \perp A | D$$

In a binary classification setting this means $Y \perp A | D = 0$ and $Y \perp A | D = 1$

Individuals are grouped according to the decision, not the actual outcome.

A=sensitive attribute; D=decision; Y=target variable/outcome

Conditional on decision

First case: $Y \perp A | D = 1$

$$P[Y = 1 | D = 1, A = a] = P[Y = 1 | D = 1, A = b]$$

The precision / PPV (positive predictive value) should be the same for the different subgroups.

This is also called **predictive parity**. Example: When people who are granted loans go on to repay them at the same rate (regardless of the group).

A=sensitive attribute; D=decision; Y=target variable/outcome

Conditional on decision

Second case: $Y \perp A | D = 0$

$$P[Y = 0 | D = 0, A = a] = P[Y = 0 | D = 0, A = b]$$

Example: All individuals who were denied a loan (D=0) are equally likely to have defaulted if the loan had been granted (Y=0) (regardless of the group).

A=sensitive attribute; D=decision; Y=target variable/outcome

Conditional on decision

Calibration

- We often have a **score** function R and $D = \mathbb{1}\{R > t\}$
- R is calibrated if $P[Y = 1|R = r] = r$, e.g., 80% of the people with score 0.8 indeed pay back their loan.

R satisfies calibration by group if

$$P[Y = 1|R = r, A = a] = r$$

Calibration by group implies sufficiency.

Measuring fairness at the level of groups

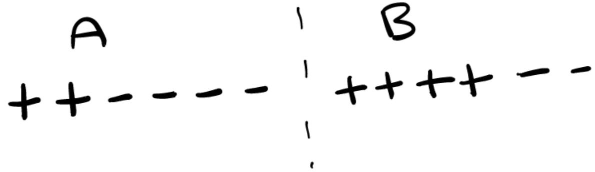
Do outcomes systematically differ between different groups?

Three criteria:

equal decision measures	**conditional on outcome**	**conditional on decision**
independence	*separation*	*sufficiency*
$A \perp D$	$D \perp A \mid Y$	$Y \perp A \mid D$

Can't we just make systems that satisfy all criteria?

A=sensitive attribute; D=decision; Y=target variable/outcome



Note: Different base rates (2/6 vs. 4/6).

Is it possible to satisfy all criteria?

Remember: statistical parity (equal % of positive outcomes), equal of opportunity (equal TPR/recall), predictive parity (equal PPV/precision)

Impossibilities

Bad news! :(

Any 2 of these 3 criteria are mutually exclusive!! (under mild assumptions).

equal decision measures	conditional on outcome	conditional on decision
independence	*separation*	*sufficiency*
$A \perp D$	$D \perp A \mid Y$	$Y \perp A \mid D$

A=sensitive attribute; D=decision; Y=target variable/outcome

So: We need to make an active choice!
Involve stakeholders and domain experts.

Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, Big Data, Special issue on Social and Technical Trade-Offs (2017) [link]
Inherent Trade-Offs in the Fair Determination of Risk Scores, Kleinberg et al., Innovations in Theoretical Computer Science (ITCS) 2017 [link]

Impossibilities

Bad news! :(

Any 2 of these 3 criteria are mutually exclusive!! (under mild assumptions).

equal decision measures	**conditional on outcome**	**conditional on decision**
independence	*separation*	*sufficiency*
$A \perp D$	$D \perp A \mid Y$	$Y \perp A \mid D$

A=sensitive attribute; D=decision; Y=target variable/outcome

So: We need to make an active choice!
Involve stakeholders and domain experts.

Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, Big Data, Special issue on Social and Technical Trade-Offs (2017) [link]
Inherent Trade-Offs in the Fair Determination of Risk Scores, Kleinberg et al., Innovations in Theoretical Computer Science (ITCS) 2017 [link]

Impossibilities

From Chouldechova, 2017. Suppose we have two groups $i \in \{A, B\}$

$$FPR_i = \boxed{\frac{p_i}{1 - p_i}} \boxed{\frac{1 - PPV_i}{PPV_i}} (1 - FNR_i)$$

Assumptions:

- the classifier makes mistakes, i.e. FPR_i and $FNR_i > 0$.
- prevalence (base rate) differs between groups, i.e. $p_A \neq p_B$

If PPV is the same across groups (predictive parity), i.e. $PPV_A = PPV_B$, then there's no way to achieve equal FPR and FNR across groups.

p: prevalence
PPV : positive predictive value (same as precision)
FPR : false positive rates
FNR: false negative rates

Chouldechova (2017) [link]

COMPAS

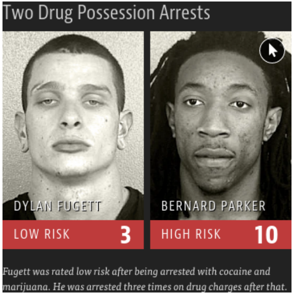


Figure: From ProPublica

COMPAS: Correctional Offender Management Profiling for Alternative Sanctions

Article by ProPublica (Angwin et al., May 23 2016) sparked a lot of debate.

You'll use the COMPAS dataset in the programming exercise.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

COMPAS

The COMPAS score: risk assessment of recidivism. Used by judges in US.

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Figure: From ProPublica

False positive rates and **false negative rates** are not equal!

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

COMPAS

The COMPAS score: risk assessment of recidivism. Used by judges in US.

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Figure: From ProPublica

False positive rates and **false negative rates** are not equal!

Response by COMPAS developers (Northpointe): COMPAS satisfies **equal positive predictive** values (Dieterich et al. 2016, [url])

“Bias preserving” vs “bias transforming”

- **Bias preserving**: System should reflect the status quo/training data. Make society not more unequal than it currently is.
    - Quick check: A perfect classifier (zero error according to the labels in the data) satisfies these criteria.
    - Example: Equalized odds, equal opportunity.
    - Focus on *error rates*
- **Bias transforming**: Acknowledge that the status quo is a result of existing inequalities.
    - Requires making an explicit decision regarding which biases a system should exhibit.
    - Example: Demographic parity.
    - Focus on *decision rates*

Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law, Wachter et al., West Virginia Law Review, Forthcoming [link]

“Bias preserving” vs “bias transforming”

Wachter et al.: “*By design, bias preserving metrics run the risk of ‘freezing’ or locking in social injustices and discriminatory effects which does not align well with the core aim of EU non-discrimination law: to achieve substantive equality.*”

But:

- Blindly enforcing demographic parity e.g., in lending applications, can make things worse! Individuals may not be able to repay, bankruptcy, etc.
- There are settings where “bias preserving” is suitable, e.g., when we do have an unbiased “ground truth”

Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law, Wachter et al., West Virginia Law Review, Forthcoming [link]

Broader applications

Note: We have focused on decision making settings, but the same measures can also be applied to other classification problems (e.g., language identification, part-of-speech tagging, image classification).

Example:
A sentiment classification system that classifies tweets into positive and negative sentiment. We have 2 groups: older and younger Twitter users.

Is a "bias preserving" or a "bias transforming" criterion more appropriate?

Plan for today

Today: How can we quantify the fairness of ML systems?

- Decision making
- Fairness at the group level
- Fairness at the individual level
- Beyond decision making (representations)

Measuring fairness: Individuals

Fairness at the subgroup level

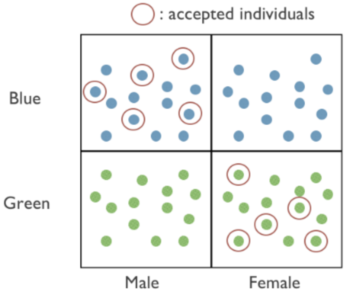


Figure: A Toy Example Kearns et al., 2018

(taken from `https://www.cis.upenn.edu/~mkearns/papers/gerryexp.pdf`)

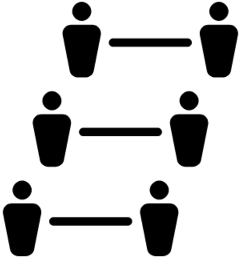
I was treated poorly! :(

Well... another person from the other group was also treated poorly!

(for example, when we focus on equal error rates)

Fairness on the group level provides *weak* guarantees for individuals.

Individual Fairness



Any two individuals that are similar with respect to the task should be treated similarly

No need to categorize individuals in predefined groups/features

Fairness through awareness, Dwrok et al., ITCS '12 [url]

recap!

Vector representations

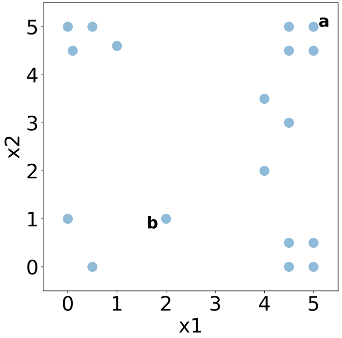


Figure: Points in a two dimensional vector space

```
a = [5, 5]
b = [2, 1]
```

a is a *two-dimensional* vector

recap!

Vector representations

```
a = [5, 5, 2]
b = [2, 1, 0]
```

a is a *three-dimensional* vector

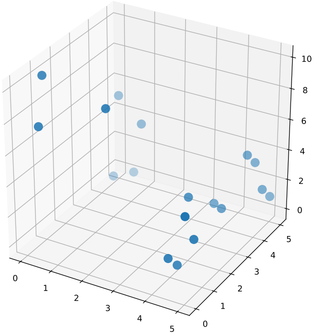


Figure: Points in a three dimensional vector space

recap!

Vector representations

```
a = [5, 5, 2]
b = [2, 1, 0]
```

a is a *three-dimensional* vector

Key idea:

Represent **people as vectors** (i.e. points in a vector space)

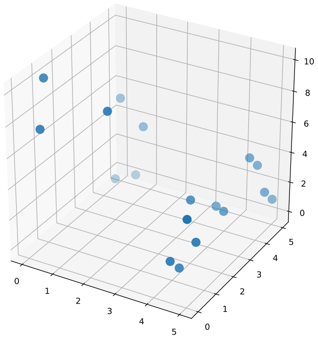


Figure: Points in a three dimensional vector space

Measuring individual fairness: Consistency

Compare the classification ($\hat{y}$) of an instance $\boldsymbol{x}$ to its k-nearest neighbors.

$$1 - \frac{1}{N} \sum_{n} |\hat{y}_n - \frac{1}{k} \sum_{j \in kNN(\boldsymbol{x}_n)} \hat{y}_j|$$

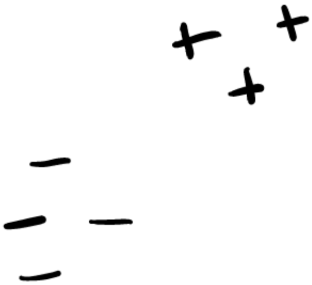
X is the set of individuals. Each $\boldsymbol{x} \in X$ is a vector representation of the individual. We have N instances.

Learning Fair Representations, Zemel et al., ICML 2013 [link]

Measuring individual fairness: Consistency

Compare the classification ($\hat{y}$) of an instance $\boldsymbol{x}$ to its k-nearest neighbors.

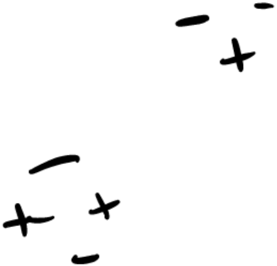
$$1 - \frac{1}{N} \sum_{n} |\hat{y}_n - \frac{1}{k} \sum_{j \in kNN(\boldsymbol{x}_n)} \hat{y}_j|$$



Measuring individual fairness: Consistency

Compare the classification ($\hat{y}$) of an instance $\boldsymbol{x}$ to its k-nearest neighbors.

$$1 - \frac{1}{N} \sum_{n} \left| \hat{y}_n - \frac{1}{k} \sum_{j \in kNN(\boldsymbol{x}_n)} \hat{y}_j \right|$$



Individual Fairness: Metric

- Judgments for every pair of individuals. Can be very nuanced and based on *human* judgements
- No need to define fairness in terms of accuracy (or stat properties)

How do we define *similarity* between individuals?

Individual Fairness: Metric

Turns out to be very, very hard to define a similarity metric!

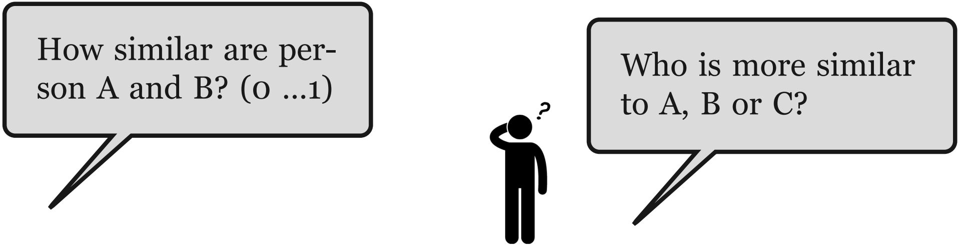
- People may differ in their opinion
- It can be hard to define a metric in a very precise way

See also work by Christina Ilvento

Individual Fairness: Metric

Turns out to be very, very hard to define a similarity metric!

- People may differ in their opinion
- It can be hard to define a metric in a very precise way



See also work by Christina Ilvento

Individual Fairness

Appealing idea, but very hard to operationalize in practice.

Some inspiration/motivation provided in the paper by Dwrok et al.:

> *[..] a decision support system for cardiology that helps a physician in finding a suitable diagnosis for a patient based on the consensus opinions of other physicians who have looked at similar patients in the past. [..] which patients are similar based on information from multiple domains such as cardiac echo videos, heart sounds, ECGs and physicians' reports.*

Less work/progress than on fairness at the group level.

Plan for today

Today: How can we quantify the fairness of ML systems?

- Decision making
- Fairness at the group level
- Fairness at the individual level
- Beyond decision making (representations)

Measuring fairness: Beyond decision making

recap!

Representational harms

Representational harms: "*when systems reinforce the subordination of some groups along the lines of identity—race, class, gender, etc.*"

Figure: Google Translate: 12th of March, 2021

NLP: Translations

Idea: Gender bias often manifests in translations when it involves co-reference resolution.

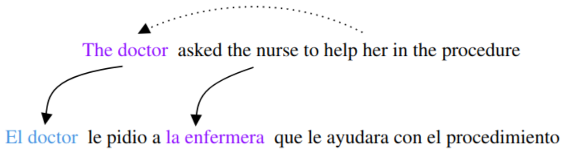


Figure: Fig 1 from Stanovsky et al.

Stanovsky et al., Evaluating Gender Bias in Machine Translation, ACL 2019. [link]

NLP: Translations

Idea: Gender bias often manifests in translations when it involves co-reference resolution.

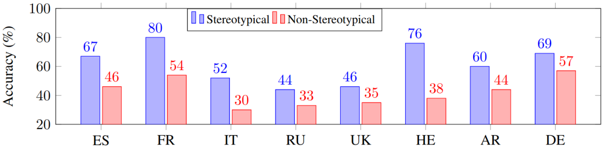


Figure 2: Google Translate's performance on gender translation on our tested languages. The performance on the stereotypical portion of WinoMT is consistently better than that on the non-stereotypical portion. The other MT systems we tested display similar trends.

Figure: Fig 2 from Stanovsky et al.. Accuracy: % of translations with correct gender

Stanovsky et al., Evaluating Gender Bias in Machine Translation, ACL 2019. [link]

recap!

Vector representations

```
a = [5, 5, 2]
b = [2, 1, 0]
```

a is a *three-dimensional* vector

Key idea:

Represent **linguistic units (e.g., words) as vectors** (i.e. points in a vector space)

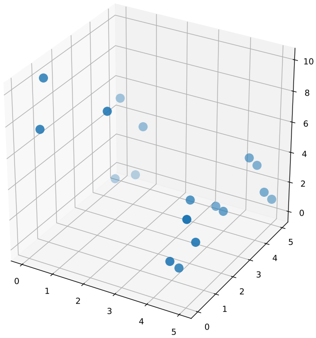


Figure: Points in a three dimensional vector space

Word as vectors

Key idea: Can we represent words as vectors?

The vector representations should:

- capture semantics
  - similar words should be close to each other in the vector space
  - relation between two vectors should reflect the relationship between the two words
- be efficient (vectors with fewer dimensions are easier to work with)
- be interpretable

Word as vectors

Key idea: Can we represent words as vectors?

The vector representations should:

- capture semantics
  - similar words should be close to each other in the vector space
  - relation between two vectors should reflect the relationship between the two words
- be efficient (vectors with fewer dimensions are easier to work with)
- be interpretable

How similar are *smart* and *intelligent?* (not similar 0–10 very similar):
How similar are *easy* and *big* (not similar 0–10 very similar):

Word as vectors

Key idea: Can we represent words as vectors?

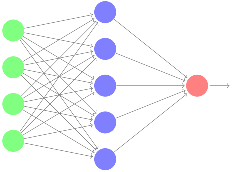
The vector representations should:

- capture semantics
  - similar words should be close to each other in the vector space
  - relation between two vectors should reflect the relationship between the two words
- be efficient (vectors with fewer dimensions are easier to work with)
- be interpretable

How similar are *smart* and *intelligent?* (not similar 0–10 very similar): 9.2
How similar are *easy* and *big* (not similar 0–10 very similar): 1.12
(*SimLex-999 dataset*)

How are they used?

How are they used?



In neural networks (text classification, sequence tagging, etc..)

cat	0.52	0.48	-0.01	···	0.28
dog	0.32	0.42	-0.09	···	0.78

As research objects

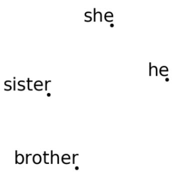
Properties

We can use cosine similarity to find similar words in the vector space.

- **dog**: *dogs, cat, man, cow, horse*
- **car**: *driver, cars, automobile, vehicle, race*
- **amsterdam**: *netherlands, rotterdam, dutch, centraal, paris*
- **chocolate**: *candy, beans, caramel, butter, liquor*

`https://projector.tensorflow.org/`

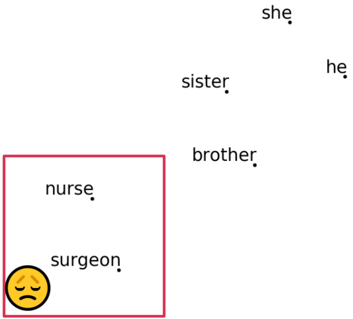
Biases in word embeddings



Man is to computer programmer as woman is to homemaker? Debiasing word embeddings, Bolukbasi et al. NIPS 2016, [link]

Semantics derived automatically from language corpora contain human-like biases, Caliskan et al., Science 2017, [link]

Biases in word embeddings



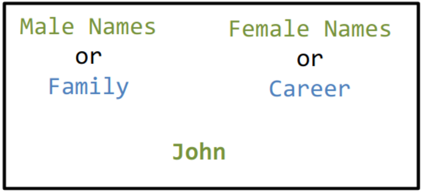
Pre-trained GloVe model on Twitter

Man is to computer programmer as woman is to homemaker? Debiasing word embeddings, Bolukbasi et al. NIPS 2016, [link]

Semantics derived automatically from language corpora contain human-like biases, Caliskan et al., Science 2017, [link]

Word-Embedding Association Test

- The Implicit Association Test (IAT) is based on response times and has been widely used.
- See https://implicit.harvard.edu/implicit/



Semantics derived automatically from language corpora contain human-like biases, Caliskan et al., Science 2017, [link]

Word-Embedding Association Test

Word-Embedding Association Test (WEAT) by Caliskan et al: use the cosine similarity between pairs of vectors as analogous to reaction time in the IAT

Were able to replicate well-known IAT findings!

Semantics derived automatically from language corpora contain human-like biases, Caliskan et al., Science 2017 [link]

Word-Embedding Association Test

Let X and Y be two sets of target words of equal size and A, B the two sets of attribute words.
For a given target word w we get a score:

$$s(w, A, B) = mean_{a \in A} cos(\vec{w}, \vec{a}) - mean_{b \in B} cos(\vec{w}, \vec{b})$$

Target words X—flowers: *aster, clover, hyacinth, crocus, rose, …*
Target words Y—insects: *ant, caterpillar, flea, spider, bedbug, …*
Attribute words A—pleasant: *freedom, love, peace, cheer, …*
Attribute words B—unpleasant: *abuse, crash, filth, murder, divorce,…*

Semantics derived automatically from language corpora contain human-like biases, Caliskan et al., Science 2017
[link]

Word-Embedding Association Test

Let X and Y be two sets of target words of equal size and A, B the two sets of attribute words.
For a given target word w we get a score:

$$s(w, A, B) = mean_{a \in A} cos(\vec{w}, \vec{a}) - mean_{b \in B} cos(\vec{w}, \vec{b})$$

Target words X—math: *math, algebra, numbers, calculus, …*
Target words Y—arts: *poetry, art, dance, literature, …*
Attribute words A—male: *male, man, boy, brother, he, him, …*
Attribute words B—female: *female, woman, girl, sister, she, her,…*

Semantics derived automatically from language corpora contain human-like biases, Caliskan et al., Science 2017
[link]

Word-Embedding Association Test

Let X and Y be two sets of target words of equal size and A, B the two sets of attribute words.
For a given target word w we get a score:

$$s(w, A, B) = mean_{a \in A} cos(\vec{w}, \vec{a}) - mean_{b \in B} cos(\vec{w}, \vec{b})$$

These scores are then aggregated:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

Semantics derived automatically from language corpora contain human-like biases, Caliskan et al., Science 2017 [link]

Word-Embedding Association Test

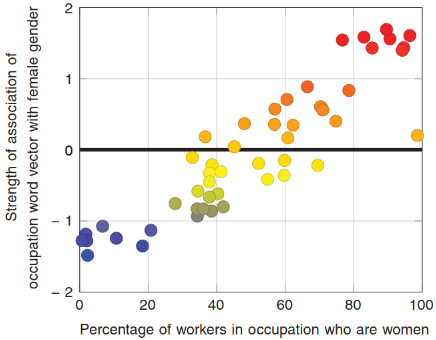


Fig. 1. Occupation-gender association. Pearson's correlation coefficient $\rho = 0.90$ with $P < 10^{-18}$.

Semantics derived automatically from language corpora contain human-like biases, Caliskan et al., Science 2017 [link]

Perpetuation of bias in sentiment analysis

> *"I had tried building an algorithm for sentiment analysis based on word embeddings [..]. When I applied it to restaurant reviews, I found it was ranking Mexican restaurants lower. The reason was not reflected in the star ratings or actual text of the reviews. It's not that people don't like Mexican food.* ***The reason was that the system had learned the word "Mexican" from reading the Web****."*

(emphasis mine)

http://blog.conceptnet.io/posts/2017/conceptnet-numberbatch-17-04-better-less-stereotyped-word-vectors/

Reflection and outlook

Fairness criteria don’t capture everything! They can’t be “proof” that a system is fair!

Literature

- Chapter 2 "*Classification*" of https://fairmlbook.org/ "Fairness and machine learning" book, by Solon Barocas, Moritz Hardt, Arvind Narayanan
  - p1—p18, up to (including) "Independence versus Sufficiency"
  - p25—p30, "Case study: Credit scoring"
  - p36—p37, "What is the purpose of a fairness criterion?"
- "*Semantics derived automatically from language corpora contain human-like biases*", Caliskan et al., Science 2017 [link]
- "*Machine Bias*", Angwin et al., ProPublica, 2016 [link]

Next time

Do the quiz by Thursday 10am

Next time:

- We'll look at approaches to make ML models more fair
- It's important that you're familiar with the criteria discussed today!

Recap:

- vectors, linear algebra
- gradients
- loss function (e.g., in logistic regression)

Announcements

- Programming assignment has been posted
- Group assignments for paper presentations + programming will be released later today.

Something to think about

So, people are biased.
Machine learning systems are biased.

What do you think are the differences between biased humans and biased ML systems, e.g. in terms of impact, or interventions?

Part of these slides are inspired by talks by Moritz Hardt ([url]) and Arvind Narayanan ([url]).