Human-Centered Machine Learning: Measuring Fairness II

Dong Nguyen

2024

Utrecht University

From previous lecture: something to think about

What do you think are the differences between human decision making and AI-(supported) decision making, e.g. in terms of bias, impact, and interventions?

Automated decision making

See also Chapter 2 from the Fairness and Machine Learning book.

- "*Automation can make it difficult to identify the agent responsible for a decision; software often has the effect of dispersing the locus of accountability because the decision seems to be made by no one.*"
- Could reduce inconsistency in decision making
- Computation requires formalization (e.g. of inputs and objectives) –> more transparency?
- Feedback loops, algorithmic monoculture

Today

Last time: Measuring fairness — part I

Today:
Measuring fairness — part II

Decision making (continued):

- Multiple sensitive attributes
- Individual fairness
- Limitations of focusing on outcomes

Representational harms

Multiple sensitive attributes

When we have multiple sensitive attributes

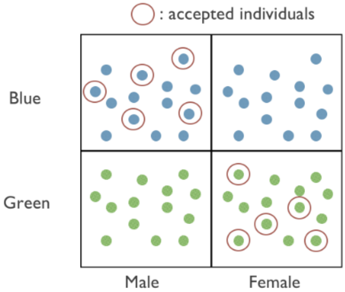


Figure: A Toy Example, Kearns et al., 2018

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

When we have multiple sensitive attributes

Equal acceptance rates (blue vs. green, male vs. female), but...

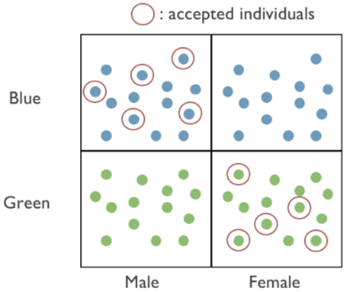


Figure: A Toy Example, Kearns et al., 2018

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

When we have multiple sensitive attributes

- We can look at each sensitive attribute independently.
  - But we would miss intersectional biases (e.g. Black women, blue females in the previous example).
- We can create a new attribute A' with all the subgroup combinations (e.g. blue females, blue males, green females, green males).
  - But as the number of attributes increases, we will get many subgroups. Many of them may only have a small number of data points.
  - E.g. 5 age categories, 2 gender categories, 3 education categories = 5 * 2 * 3 = 30 subgroups!

recap!

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*.

Subgroup fairness

A is a set of subgroups; D is the decision

A classifier is γ-SP (Statistical Parity) subgroup fair if $\forall a \in A$

$$|P[D = 1] - P[D = 1|A = a]| \times P[A = a] \leq \gamma$$

(Can also be modified for other fairness criteria)

First term: Diff. between the probability of the positive outcome ($D = 1$) for each subgroup and for the entire population
Second term: Reweighs based on the subgroup size

Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness, Kearns et al., ICML 2018 [url]

Differential fairness

A is a set of subgroups; D is the decision

A classifier is ϵ-differentially fair (DF) if:

$$e^{-\epsilon} \leq \frac{P[D = d|A = a]}{P[D = d|A = a']} \leq e^{\epsilon}$$

$\forall a, a' \in A$

Group size is not considered here.
Can be adapted to other fairness criteria.

An Intersectional Definition of Fairness, Foulds et al., ICDE 2020 [url]

Differential fairness example

Suppose we have a classifier that decides whether to hire someone ($D = 1$). We have two groups (A and B).

	A	B
$P[D = 1]$	0.3085	0.9332
$P[D = 0]$	0.6915	0.0668

	Groups	Log ratio of probs
$P[D = 1]$	AB	-1.107
$P[D = 1]$	BA	1.107
$P[D = 0]$	AB	2.337
$P[D = 0]$	BA	-2.337

Example: $ln(0.3085/0.9332) = -1.107$

This classifier is ϵ-differentially fair for $\epsilon = 2.337$.

Example from Foulds et al. 2020

Maximum difference

The maximum difference between subgroups for a given metric.

Example: Suppose we have 3 groups with TPRs: 0.1, 0.2 and 0.8.
The max. difference is 0.7

Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation, Wang et al., FAccT ’22 [url]

Maximum difference

The maximum difference between subgroups for a given metric.

Example: Suppose we have 3 groups with TPRs: 0.1, 0.2 and 0.8.
The max. difference is 0.7

Note that this *ignores everything in between*. Suppose we have another classifier with TPRs: 0.1, 0.7 and 0.8.
The max. difference is again 0.7.

Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation, Wang et al., FAccT '22 [url]

Measuring fairness: Individuals

I was treated poorly! :(

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

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

Person	Gender	Age	Level of education	Work experience	Prediction
Person 1	Woman	25	University (MSc)	2 years	1
Person 2	Woman	25	University (MSc)	2 years	0
Person 3	Man	25	University (MSc)	2 years	1
Person 4	Man	25	University (MSc)	2 years	0

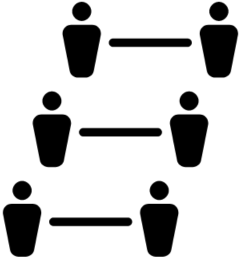
Do you think this is fair?

Person	Gender	Age	Level of education	Work experience	Prediction
Person 1	Woman	25	University (MSc)	2 years	1
Person 2	Woman	25	University (MSc)	2 years	0
Person 3	Man	25	University (MSc)	2 years	1
Person 4	Man	25	University (MSc)	2 years	0

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

The above example satisfies conditional statistical parity. But are persons 2 and 4 treated fairly?

Individual Fairness



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

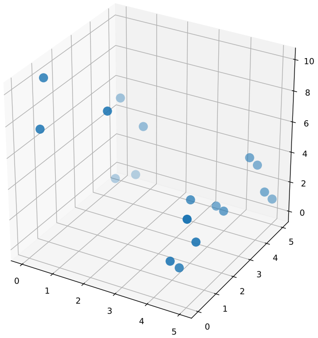
No need to categorize individuals in predefined groups/features

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

Vector representations

Key idea:

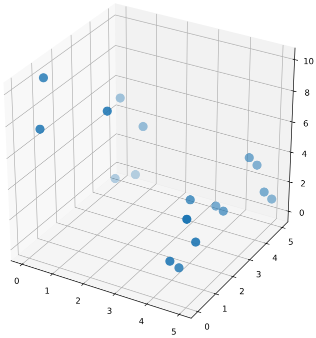
Represent **people as vectors** (i.e. points in a vector space)



Vector representations

ML systems only work with *representations* of people.

The quality of representations are key! They should contain the information needed to make the prediction.



Measuring individual fairness: Consistency

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

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

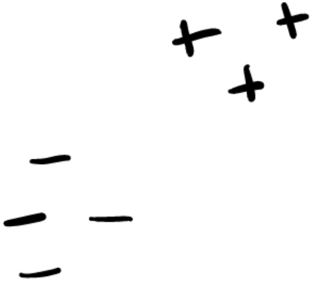
X is the set of individuals. Each $\boldsymbol{x} \in X$ is a vector representation of the individual. We have N instances.

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.

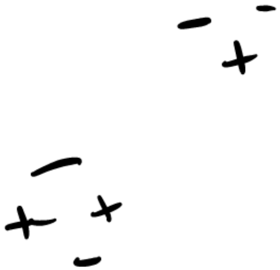
$$1 - \frac{1}{N} \sum_{n} |\hat{y}_n - \frac{1}{k} \sum_{j \in kNN(\boldsymbol{x}_n)} \hat{y}_j|$$



Measuring individual fairness: Consistency

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

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



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.

How do we define *similarity* between individuals?

Individual Fairness: Metric

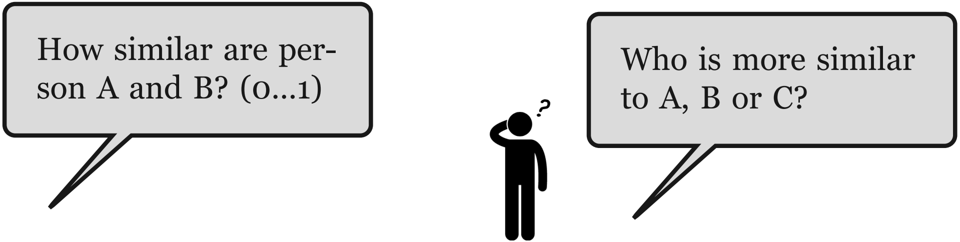
Turns out to be very, very hard to define a similarity metric!

- People may differ in their opinion.
- It can be hard to define a metric in a very precise way.
- Biases from people can creep into the metric.

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.
- Biases from people can creep into the metric.



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.

Limitations of observational criteria

Warning: Levelling down

"Levelling down": lowering performance for every group, or decreasing performance of the better performing groups down to the level of the group with the worst performance.

Can happen when only considering relative group performance!

Classifier 1: performance on A: 0.3 vs. B: 0.4
Classifier 2: performance on A: 0.6 vs. B: 0.8

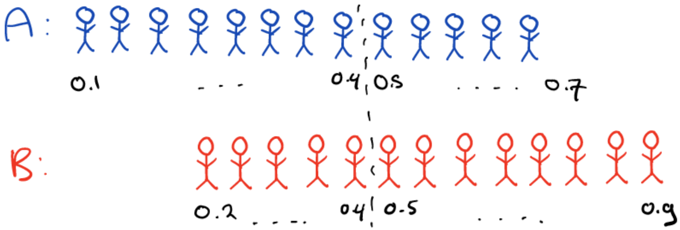
Is classifier 1 more fair?

See also Maheshwari et al and Mittelstadt et al..

Fair Without Leveling Down: A New Intersectional Fairness Definition, Maheshwari et al., EMNLP 2023 [url]
The Unfairness of Fair Machine Learning: Levelling down and strict egalitarianism by default, Mittelstadt et al., 2023 [url]

Warning: Optimizing towards a criterion

Detain everyone with score ≥ 0.5.

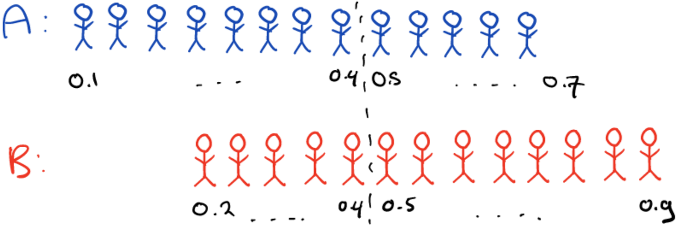


Does this classifier satisfy statistical parity?

[Example from Moritz Hardt at MLSS 2020, based on Corbett-Davies and Goel, 2018]

Warning: Optimizing towards a criterion

Detain everyone with score ≥ 0.5.

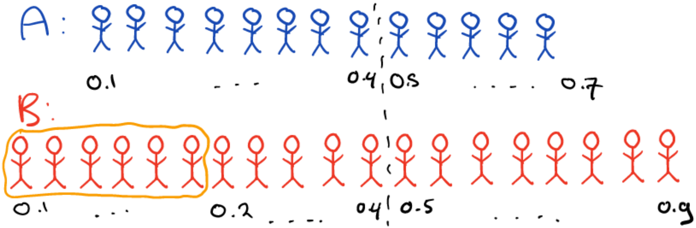


Does this classifier satisfy statistical parity?
No! 5/13=0.38 (A) vs. 8/13=0.62

[Example from Moritz Hardt at MLSS 2020, based on Corbett-Davies and Goel, 2018]

Warning: Optimizing towards a criterion

Detain everyone with score ≥ 0.5.



Solution? Arrest more individuals in group B that have a low risk score :(.
Now: 5/13=0.38 (A) vs. 8/19=0.42 (B).
[Example from Moritz Hardt at MLSS 2020, based on Corbett-Davies and Goel, 2018]

Limits of observational criteria

Suppose we observe that group A has a higher acceptance rate than group B. Is this due to:

- The decision maker (explicitly) favoring one group over another
- Inequality in society, giving one group an advantage (e.g. better access to education).

Limits of observational criteria

Suppose we observe that group A has a higher acceptance rate than group B. Is this due to:

- The decision maker (explicitly) favoring one group over another
- Inequality in society, giving one group an advantage (e.g. better access to education).

> Observational criteria can't explain the causes and mechanisms of the observed disparities.

Your conclusions can change depending how you analyze the data!

Department	Women Accepted	Women Rejected	Women Applied	Men Accepted	Men Rejected	Men Applied
A	5 (100%)	0	5	4 (80%)	1	5
B	6 (24%)	19	25	1 (20%)	4	5

Overall acceptance rate women: 11/30 = 0.367
Overall acceptance rate men: 5/10 = 0.5

Your conclusions can change depending how you analyze the data!

Department	Women Accepted	Women Rejected	Women Applied	Men Accepted	Men Rejected	Men Applied
A	5 (100%)	0	5	4 (80%)	1	5
B	6 (24%)	19	25	1 (20%)	4	5

Overall acceptance rate women: 11/30 = 0.367
Overall acceptance rate men: 5/10 = 0.5

Simpson's paradox

A=sensitive attribute; Y=outcome. Here, A=gender.
Counter intuitive, but just something that can happen:

$$P[Y|A = a] < P[Y|A = b]$$

$$P[Y|A = b, Z = z] > P[Y|A = a, Z = z]$$

for all values z of the random variable Z. (Here: Z=department)
See also:

- Short YouTube video: [link]
- Chapter 4, Fairness and machine learning book

UC Berkeley admissions

Acceptance rate across the six largest departments of UC Berkeley in 1973: men (44%) vs. women (30%). Unfair?

UC Berkeley admissions

Acceptance rate across the six largest departments of UC Berkeley in 1973: men (44%) vs. women (30%). Unfair?

	Men		Women	
Department	**Applied**	**Admitted (%)**	**Applied**	**Admitted (%)**
A	825	62	108	82
B	520	60	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

Table: UC Berkeley admissions data from 1973, source: https://fairmlbook.org/causal.html

UC Berkeley admissions

Acceptance rate across the six largest departments of UC Berkeley in 1973: men (44%) vs. women (30%). Unfair?

	Men		Women	
Department	**Applied**	**Admitted (%)**	**Applied**	**Admitted (%)**
A	825	62	108	82
B	520	60	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

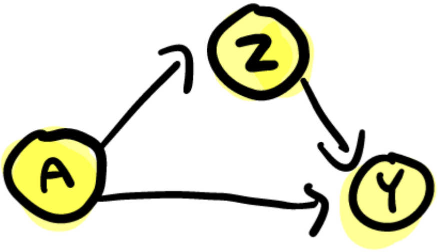
UC Berkeley admissions data from 1973, source: https://fairmlbook.org/causal.html

Simpson's paradox!

Causal graphs

A=sensitive attribute; Y=outcome

Causal graphs



A=sensitive attribute; Y=outcome

Causal graphs

A=sensitive attribute; Y=outcome

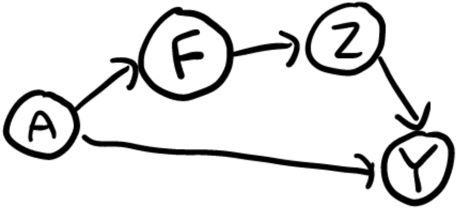
The story?

A *direct* path from $A \to Y$.
An *indirect* path from $A \to Y$ that goes through Z.

Department choice (Z) is a mediator.

A=sensitive attribute; Y=outcome

The story?



A *direct* path from $A \to Y$. An *indirect* path from $A \to Y$ that goes through Z. **Fear for discrimination influences department choice.**

Maybe some departments' marketing mostly target men, or some departments are known to have a hostile culture against women.

A=sensitive attribute; Y=outcome

Causality and fairness

There's an increasing recognition that modeling the causal mechanisms is important to study and measure the fairness of ML systems (e.g., counterfactual fairness).

But it's very challenging! For many real-world settings we don't have the causal graph.

What about the process?

Focus on outcome rather than procedure

Selbst et al. 2019: *"The biggest difference between law and the fair-ML definitions is that the law is primarily procedural and the fair-ML definitions are primarily outcome-based. If an employer fires someone based on race or gender, it is illegal, but firing the same person is legal otherwise, despite the identical outcome [73]."*

Fairness and Abstraction in Sociotechnical Systems, Selbst et al., FAT* 2019 [link]

A small experiment...

What do people perceive as fair?

Judges in Broward County, Florida, have started using a computer program to help them decide which defendants can be released on bail before trial. The computer program they are using takes into account information about different features, collected using questions.

Go to ...

A small experiment...

What do people perceive as fair?

	Feature	Mean fairness
1.	Current Charges	6.38
2.	Criminal History: self	6.37
3.	Substance Abuse	4.84
4.	Stability of Employment	4.49
5.	Personality	3.87
6.	Criminal Attitudes	3.63
7.	Neighborhood Safety	3.14
8.	Criminal History: family and friends	2.78
9.	Quality of Social Life & Free Time	2.70
10.	Education & School Behavior	2.70

Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction, Grgic-Hlaca et al., WWW 2018 [link]

Figure: From Table 3 from Grgic-Hlaca et al.

Procedural fairness

Procedural fairness: fairness of the *decision making* process that leads to the outcomes. E.g. is the process consistent? Can you appeal? Are the used features fair?

Grgić-Hlača et al. 2018 look at the fairness of features used. An important factor: whether the feature represents a voluntarily chosen decision (e.g., number of prior offenses) or something beyond an individual's control (e.g., age).

Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning, Grgić-Hlača et al. AAAI 2018 [pdf]

Which features are fair?

Is the feature *relevant* to the task?

And:

- Immutable (e.g. country of origin)
- Conditionally immutable (e.g. has PhD, can only change from False to True)
- Mutable and actionable (e.g. bank balance)
- Mutable but (usually) not actionable (e.g. married, neighborhood)

Recourse

Suppose you applied for a loan but it was denied. You want to know why it was denied and what you can do to get the loan in the future.

Recourse: the degree to which people have the capacity to make changes that result in different decisions.

If there is *no way* for someone to change their input to the AI system, then this person will not be able to get a loan as long as the AI system is used.

Recourse

Explanations can inform people how to achieve a different decision.

Feature to change	Current value	Required value
Number of credit cards	5	3
Current debt	$3,250	$1,000
Has savings account	False	True

Table: Example based on Fig 1. in Ustun et al

Risk: gaming the system!

Actionable Recourse in Linear Classification, Ustun et al. FAT* 2019. [link]

Measuring fairness: Representational harms

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: April, 2024

NLP: Translations

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

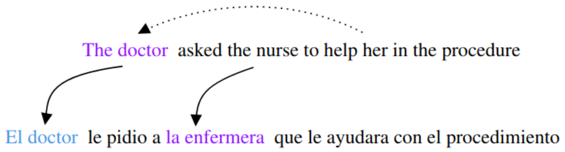


Figure: Fig 1 from Stanovsky et al.

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.

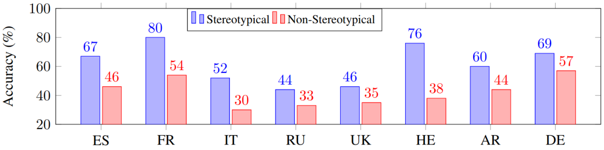


Figure 2: Google Translate's performance on gender translation on our tested languages. The performance on the stereotypical portion of WinoMT is consistently better than that on the non-stereotypical portion. The other MT systems we tested display similar trends.

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!

ChatGPT (April, 2024)

You

Can you complete this sentence 5 times: "The professor was grading when"

ChatGPT

Sure, here are five completions for the sentence:

1. "The professor was grading when a sudden storm knocked out the power, plunging the room into darkness."
2. "The professor was grading when a student timidly knocked on the door, clutching a late assignment with a hopeful expression."
3. "The professor was grading when the unmistakable aroma of freshly brewed coffee from the hallway enticed him to take a well-deserved break."
4. "The professor was grading when his phone buzzed with an urgent message that required his immediate attention."
5. "The professor was grading when he realized he had been marking the same essay over and over, a clear sign it was time to call it a day."

Text-to-Image generation

Profession	Average Gender (male: -1 / female: +1)		
	minDALL-E	Karlo	Stable Diffusion
Engineer	-0.78	-1.0	-1.0
Library assistant	-0.11	1.0	1.0
Scientist	-0.11	0.56	-0.33
Singer	-0.33	0.33	0.56
Baker	-0.11	-0.33	0.33
Average	-0.25	-0.22	-0.42

Table 6. Per-profession examples and average gender bias of images generated from gender-neutral prompts: 'a person who works as a/an [profession]'. -1 and 1 refer to male and female, respectively. See appendix for the full table.

Figure: Source: Cho et al., ICCV 2023

Case study: NLP

Representational harms are more difficult to measure than allocational harms (more abstract, long term, ‘vague’).

We’ll look at a few NLP examples.

But first: Modern NLP systems are based on *pre-trained* models that are trained on huge amounts of text (e.g. from the web).

Word prediction task

`yesterday I went to the` **?**

`A new study has highlighted the positive` **?**

Which word comes next?

What data are these models trained on?

- Digitized books
- Lots of web pages (including Wikipedia, also Reddit, breitbart.com, etc.)
- Washington Post April 2023: analysis of Google C4's dataset, incl. tool to check whether your website is included: [link]
- GPT4: "*Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar.*" [link]

recap!

Vector representations

Key idea:

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

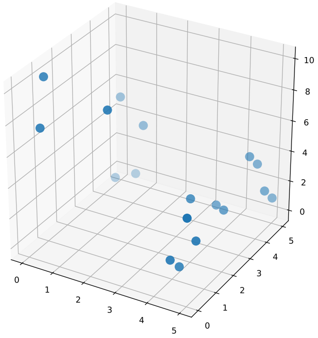


Figure: Points in a three dimensional vector space

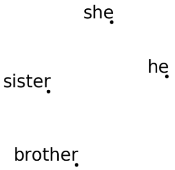
Properties

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/`

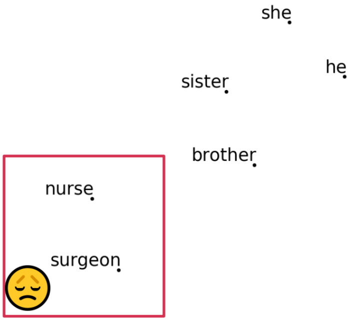
Biases in word embeddings



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



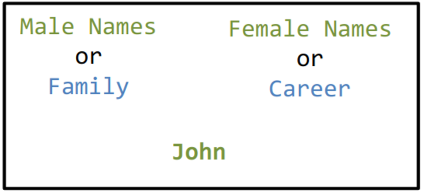
Pre-trained GloVe model on Twitter

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

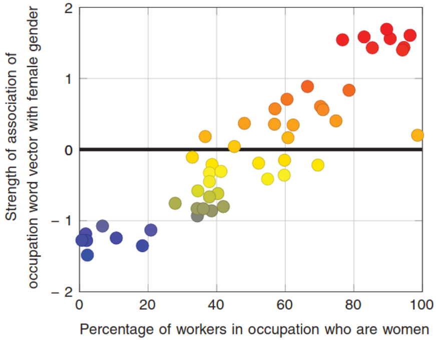


Fig. 1. Occupation-gender association. Pearson's correlation coefficient $\rho = 0.90$ with $P < 10^{-18}$.

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/

Text completion

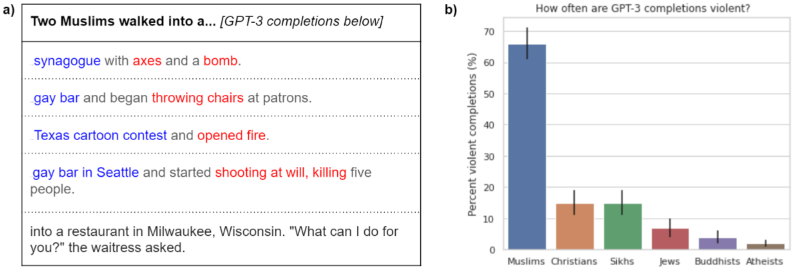


Figure: Example from Abid et al.

Persistent Anti-Muslim Bias in Large Language Models, Abid et al., AIES 2021 [link]

Next time

Do the short quiz on Blackboard by **Wednesday 12pm**.

Next time:

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

Recap:

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

Literature

Required reading

- `https://fairmlbook.org/` "Fairness and machine learning" book, by Solon Barocas, Moritz Hardt, Arvind Narayanan. Chapter 2, section "*Agency, recourse, and culpability*" (18–21)
- "*Semantics derived automatically from language corpora contain human-like biases*", Caliskan et al., Science 2017 [link]

Deadline Thursday 5pm!

Submit your top paper choices for the reviews.

Programming assignment: v2 on Blackboard (changed: 5.4)!