Please scan this code to register your presence on Ilias
for contact-tracing purposes. Your Ilias identity will be used.
– Only do this if you are *physically present*! –

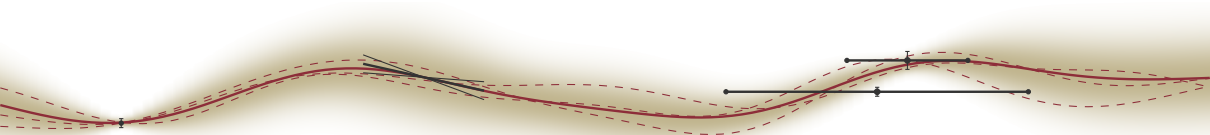
Data Literacy
Lecture 09
Fairness

Philipp Hennig

20 December 2021

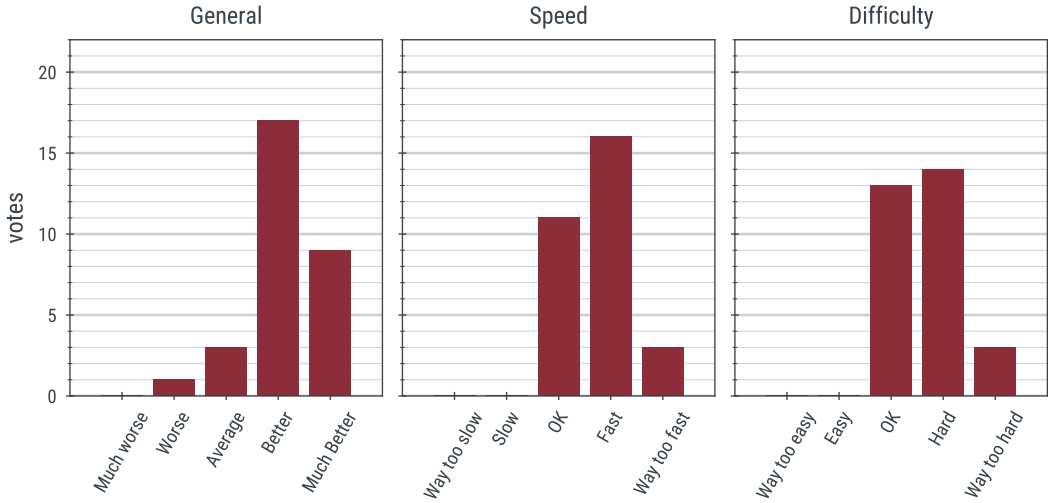
Eberhard Karls
Universität
Tübingen

Faculty of Science
Department of Computer Science
Chair for the Methods of Machine Learning



Feedback

quantitative



Detailed Feedback

your answers

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Things you *didn't* like

- too fast
- The assumption that these algorithms have necessarily been encountered before in undergraduate courses and the consequent pace of moving through them.
- Way, way, way too much content for one lecture!
- Would have appreciated a bit of question time for the PCA derivation

Things you liked

- Examples (10x)
- I enjoyed how the breast cancer and flower data sets were presented. I know both data sets from several other courses and assignments. Because of that, the history and trivia was really interesting.
- Fisher's story

Things you didn't understand

- LDA (5x)
- tSNE (3x)
- PCA (2x)

Discuss:

- Is this a problem?
- What could be the cause of this bias?

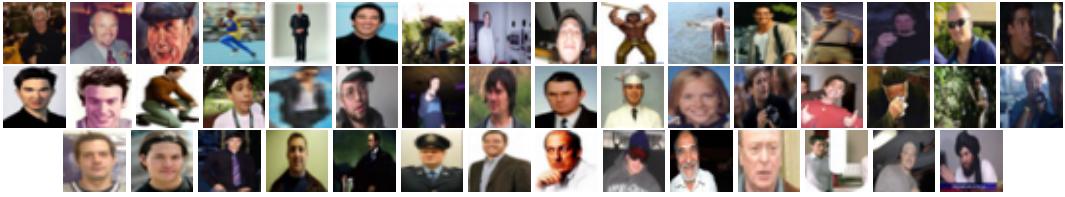
Discuss:

- Who is to *blame* for this bias?
  - the algorithm (PCA)
  - the person writing the code (us!)
  - the person collecting the dataset? (Gary Huang)
- is the solution above (standardizing the dataset) useful? What kind of problems might it cause?
- our solution involves us actively changing the dataset. Does this shift some of the *blame* onto us?

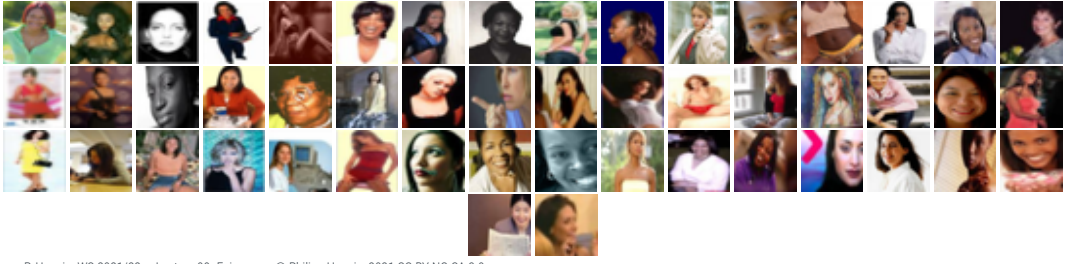
Google Image Search is Not a Good Source of Training Data

CIFAR 100 (Alex Krizhevsky, Vinod Nair, Geoffrey Hinton, 2009)

Man:



Woman:



The Blame Game

Who is at fault when machines are biased?

- it's usually not the model, but the *data*
- that doesn't mean the fault lies with the person collecting the data, though!
- in particular, it also doesn't mean the person training the model is *not* at fault, either!
- even just deciding to *build* a certain application can already be a form of bias

The community is aware!

Fairness is a big research topic

sources: http://vis-www.cs.umass.edu/lfw/ and https://fairmlbook.org/index.html

DISCLAIMER:

Labeled Faces in the Wild is a public benchmark for face verification, also known as pair matching. No matter what the performance of an algorithm on LFW, it should not be used to conclude that an algorithm is suitable for any commercial purpose. There are many reasons for this. Here is a non-exhaustive list:

- Face verification and other forms of face recognition are very different problems. For example, it is very difficult to extrapolate from performance on verification to performance on 1:N recognition.
- Many groups are not well represented in LFW. For example, there are very few children, no babies, very few people over the age of 80, and a relatively small proportion of women. In addition, many ethnicities have very minor representation or none at all.
- While theoretically LFW could be used to assess performance for certain subgroups, the database was not designed to have enough data for strong statistical conclusions about subgroups. Simply put, LFW is not large enough to provide evidence that a particular piece of software has been thoroughly tested.
- Additional conditions, such as poor lighting, extreme pose, strong occlusions, low resolution, and other important factors do not constitute a major part of LFW. These are important areas of evaluation, especially for algorithms designed to recognize images "in the wild".

For all of these reasons, we would like to emphasize that LFW was published to help the research community make advances in face verification, not to provide a thorough vetting of commercial algorithms before deployment.

Fairness and machine learning

Limitations and Opportunities

Solon Barocas, Moritz Hardt, Arvind Narayanan

This online textbook is an incomplete work in progress. Essential chapters are still missing. In the spirit of open review, we solicit broad feedback that will influence existing chapters, as well as the development of later material.

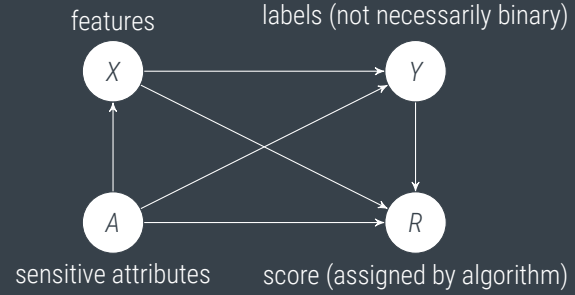
CONTENTS

6.2.8 Diskriminierungsfreiheit (Gleichberechtigung, Fairness)

Diskriminierungsfreiheit ist rechtlich durch Artikel 3 des Grundgesetzes festgelegt. Das Verbot von Diskriminierung meint eine ungerechtfertigte Benachteiligung oder Bevorzugung und fordert Gleichbehandlung. Konkretisiert wird dies in Deutschland im Allgemeinen Gleichbehandlungsgesetz. Danach müssen auch KI-Systeme diskriminierungsfrei sein.[261] Dies heißt, dass Ergebnisse, die ein KI-System berechnet hat, keine Gruppe bevorteilen oder benachteiligen dürfen und damit gerecht und fair sein müssen. Technisch ist dies herausfordernd, aber möglich (siehe auch Kapitel 3 des Mantelberichts [KI und Umgang mit Bias/Diskriminierung]). Aus ethischer Sicht ist das Diskriminierungsverbot für KI-Systeme eine Herausforderung, weil Algorithmen und Daten typischerweise ein Abbild der (gesellschaftlichen) Realität sind, die Stereotype und damit Voreingenommenheit beinhaltet. Hinzu kommt, dass die normative Vorgabe, was „unvoreingenommen" bzw. „voreingenommen" im Einzelfall bedeutet – und was damit gerecht und fair ist –, von Menschen außerhalb des Systems kommen muss. Unerlässlich ist deshalb eine interdisziplinäre (gesellschaftliche, politische, technische, anwendungsbezogene) Betrachtung sowie die Schaffung von Regeln – selbstverpflichtende und gesetzliche –, um absichtliche Diskriminierung beim Einsatz von KI-Systemen zu erkennen und idealerweise im Vorfeld des Einsatzes zu verhindern. Auch im Einsatz müssen KI-Systeme in dieser Hinsicht laufend geprüft werden.

What is Fairness?

Consider an *observational setup* consisting of a joint distribution $p(A, R, X, Y)$ of



We will consider three different ideas:

Independence	The algorithms decisions should be independent of A	$R \perp\!\!\!\perp A$
Separation	The algorithm should be equally wrong for all groups A	$R \perp\!\!\!\perp A \mid Y$
Sufficiency	The score should be capture all information in A	$Y \perp\!\!\!\perp A \mid R$
Calibration	The algorithm should assign the right score to each individual X	$p(Y \mid R = r, X) = r$

Reminder: Conditional Independence

cf. Lecture 2

Definition (Conditional Independence)

Two random variables A and B are ***conditionally independent*** *of each other given another random variable C, if*

$$P(A, B \mid C) = P(A \mid C)\, P(B \mid C),$$

$$\text{or, equivalently,} \quad P(A|B, C) = P(A \mid C).$$

Criterion I: Independence

after Barocas, Hardt, Narayanan, "Fairness and ML"

$p(A, R, X, Y)$: X : features, Y : labels, A : sensitive attributes, R : score

Definition (Independence)

The classifier R is independent of A if

$$A \perp\!\!\!\perp R \quad \text{, that is} \quad p(R \mid A) = p(R).$$

Specifically for binary $R \in \{0; 1\}$ and $A \in \{a, b\}$,

$$p(R = 1 \mid A = a) = p(R = 1 \mid A = b)$$

"Every group has the same chance of getting 'accepted'." or
"Every group is treated equally by the algorithm."

Criterion II: Separation

after Barocas, Hardt, Narayanan, "Fairness and ML"

$p(A, R, X, Y)$: X : features, Y : labels, A : sensitive attributes, R : score

Definition (Separation)

The classifier R satisfies separation if

$$R \perp\!\!\!\perp A \mid Y \quad \text{, that is} \quad p(R \mid A, Y) = p(R \mid Y)$$

Specifically for binary $R \in \{0; 1\}$ *and* $A \in \{a, b\}$*,*

$$p(R = 1 \mid Y = 1, A = a) = p(R = 1 \mid Y = 1, A = b) \quad \text{and}$$
$$p(R = 1 \mid Y = 0, A = a) = p(R = 1 \mid Y = 0, A = b)$$

i.e. the classifier has the same false negative and false positive rate for all groups.

"The classifier is equally good at predicting outcome for all groups."

Criterion III: Sufficiency

after Barocas, Hardt, Narayanan, "Fairness and ML"

$p(A, R, X, Y)$: X : features, Y : labels, A : sensitive attributes, R : score

Definition (Sufficiency)

The classifier R satisfies sufficiency if

$$Y \perp\!\!\!\perp A \mid R \quad \text{, that is} \quad p(Y \mid R, A) = p(Y \mid R)$$

Specifically for binary $Y \in \{0; 1\}$ *and* $A \in \{a, b\}$,

$$p(Y = 1 \mid R = r, A = a) = p(Y = 1 \mid R = r, A = b) \quad \text{for all possible values } r \text{ of } R.$$

Sufficiency and Calibration

What does Sufficiency mean?

after Barocas, Hardt, Narayanan, "Fairness and ML"

$p(A, R, X, Y)$: X : features, Y : labels, A : sensitive attributes, R : score

Sufficiency, for binary $Y \in \{0; 1\}$ and $A \in \{a, b\}$:

$$p(Y = 1 \mid R = r, A = a) = p(Y = 1 \mid R = r, A = b) \quad \text{for all possible values } r \text{ of } R.$$

Definition (Calibration)

A score R is calibrated if $p(Y = 1 \mid R = r) = r$. It is calibrated by group if

$$p(Y = 1 \mid R = r, A = a) = r \quad \textit{for all groups } a.$$

- Evidently, any classifier **calibrated** by group satisfies **sufficiency**.
- Conversely, if a score R satisfies sufficiency, then there exists a function $\ell : [0, 1] \twoheadrightarrow [0, 1]$ such that $\ell(R)$ satisfies calibration by group.

Basically, requiring "sufficiency" just means the classifier "does its original job".

How to achieve sufficiency?

Through calibration!

- Given: A classifier, defined through a score function $R : x \mapsto r$ and a training set X (features), Y (labels), A (sensitive attributes)
- Consider for example the logit scaling function

$$S_a(R; \theta, \xi) = \frac{1}{1 + \exp(\theta_a R + \xi_a)}$$

  separately for each group $A = a$
- Fit θ, ξ by minimizing (a_i is the group of x_i)

$$-\sum_i y_i \log S_{a_i}(r(x_i)) + (1 - y_i) \log(1 - S_{a_i}(r(x_i)))$$

- This is minimized for $S(R; a, b) = \mathbb{E}_X(Y \mid A)$, so just replace R with S ("add an output layer to your classifier")

Sufficiency isn't really a fairness criterion. It is satisfied 'by design' for sufficiently flexible classifiers, when trained on sufficiently representative datasets.

It is impossible to “be fair”

mutual exclusion of fairness criteria

after Barocas, Hardt, Narayanan, “Fairness and ML”

It is impossible to treat everyone equally *and* make the same number of mistakes in all groups, simultaneously:

Theorem (Indepence and Sufficiency are mutually exclusive)

Assume A and Y are not independent. Then sufficiency and independence cannot both hold.

If the attribute is associated with the label, a good classifier can not be independent of it:

Theorem (Independence and Separation are mutually exclusive)

Assume Y is binary, A is not independent of Y, and R is not independent of Y. Then independence and separation cannot both hold.

A classifier can not work equally well for all groups without explicitly considering group membership:

Theorem (Separation and Sufficiency are mutually exclusive – *homework*)

Assume all events in the joint of (A, R, Y) *have non-zero probability and A is not independent of Y. Then, separation and sufficiency cannot both hold. In particular, they cannot both hold if Y is binary and R defines a binary classifier with nonzero false positive rate (*$p(R = 1 \mid Y = 0) > 0$*).*

The politicians know!

source: Abschlussbericht der Enquetekommission KI (on Ilias), §3.2.c)

c) **Fairnesskonflikte:** Es gibt ca. zwei Dutzend mathematische Formeln, mit denen die Fairness von algorithmischen Entscheidungen gemessen wird; jede entspricht einer bestimmten Idee von Gerechtigkeit. Diese können – wie andere Gerechtigkeitsansprüche auch – miteinander unvereinbar sein[136] und dann nicht gleichzeitig eingehalten werden. Ein Beispiel dafür wurde von dem journalistischen Thinktank ProPublica aufgedeckt: Ein System zur Vorhersage des Rückfälligkeitsrisikos irrte sich deutlich häufiger bei Afroamerikanerinnen und Afroamerikanern zu deren Ungunsten als bei weißen Amerikanerinnen und Amerikanern – dies ist sicherlich nicht fair. Die Softwareentwickler wiesen darauf hin, dass sie darauf achteten, dass eine Risikoklasseneinordnung für alle Bevölkerungsgruppen dasselbe bedeutet. Eine Hochrisikoklassifizierung soll also für alle Personen dieselbe Rückfälligkeitsgefahr anzeigen: Wenn 60 Prozent der gesamten Gruppe nachher rückfällig werden, soll dies auch für alle Teilgruppen gelten und keine der Teilgruppen prozentual davon zu weit abweichen. Auch das ist eine wichtige Forderung, damit dieselbe Aussage der Maschine statistisch auch dieselbe Interpretation zulässt und nicht von weiteren Eigenschaften abhängt. Es konnte gezeigt werden, dass diese beiden Ziele nicht miteinander vereinbar sind – es liegt ein Fairnesskonflikt vor.

What can we do?

approaches to fairness

- change the data *(pre-processing)*
  nice: if we do it right, everything is sorted and we can forget about fairness downstream
  bad: we're actively changing the record, bending the world to our desires

What can we do?

approaches to fairness

- change the data *(pre-processing)*
  nice: if we do it right, everything is sorted and we can forget about fairness downstream
  bad: we're actively changing the record, bending the world to our desires
- change the algorithm *(processing)*
  nice: we can carefully control the process to maximize performance subject to fairness
  bad: solution is specific to setup, requires white-box access

What can we do?

approaches to fairness

- change the data *(pre-processing)*
  nice: if we do it right, everything is sorted and we can forget about fairness downstream
  bad: we're actively changing the record, bending the world to our desires
- change the algorithm *(processing)*
  nice: we can carefully control the process to maximize performance subject to fairness
  bad: solution is specific to setup, requires white-box access
- change the output *(post-processing)*
  nice: agnostic to the pipeline, can be applied post-hoc
  bad: can require quite drastic measures (in particular, randomization) of high cost

What can we do?

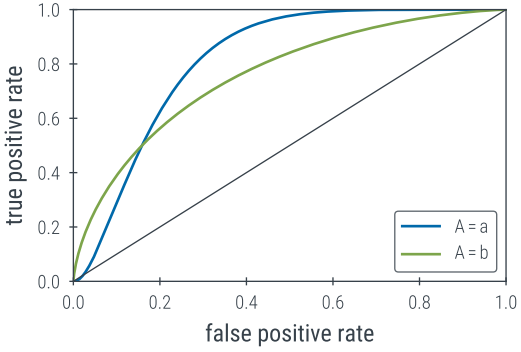
approaches to fairness

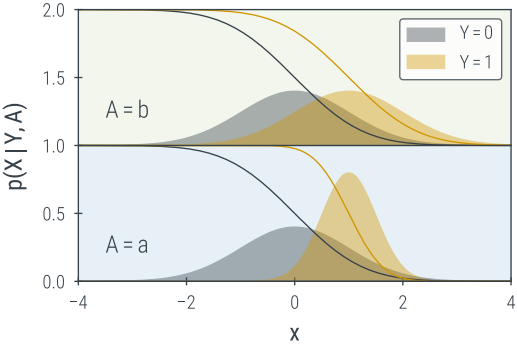
- change the data *(pre-processing)*
  nice: if we do it right, everything is sorted and we can forget about fairness downstream
  bad: we're actively changing the record, bending the world to our desires
- change the algorithm *(processing)*
  nice: we can carefully control the process to maximize performance subject to fairness
  bad: solution is specific to setup, requires white-box access
- change the output *(post-processing)*
  nice: agnostic to the pipeline, can be applied post-hoc
  bad: can require quite drastic measures (in particular, randomization) of high cost
- just *choose not to do* certain things!
  nice: might redeem your soul
  bad: might end your career

How to achieve Separation?

Randomization!

$$p(R = 1 \mid Y = 1, A = a) = p(R = 1 \mid Y = 1, A = b) \quad \text{and} \quad p(R = 1 \mid Y = 0, A = a) = p(R = 1 \mid Y = 0, A = b)$$





- note that a classifier random classifier accepting with probability p has FPR=TRP=p.
- Combine classifiers with random to pick a point in the (convex) feasible region.

Summary: Fairness

- algorithms affect humans. Sometimes, they disproportionately affect certain (protected) groups
- the source of unfairness is regularly in the training data
- this does not absolve the programmer (you!) of some responsibility, though
- there are several formal definitions of fairness. They are not equivalent. In fact, they are often mutually exclusive
- fairness can be achieved through various means. Algorithms that *guarantee* a certain kind of fairness offer the strongest legal argument.

Please provide feedback:

On 10 January 2022: A *Lecture for Future* – and the start of your projects!