Machine Learning CSE2510 – Lecture 5.1: Bias in Machine Learning

Odette Scharenborg

TUDelft

Welcome to week 5 - lecture 1

- Admin
- Recap previous lecture
- What is (implicit) bias and why is it wrong?
- Examples of bias in ML
- Technical aspects of bias and debiasing
- Other sources of biases

Admin

- Practice exam will be put online prior to Thursday's lab

TUDelft

Recap of the previous lectures

- Different types of classifiers
  - Generative classifiers
  - Linear classifiers

- All use a *dataset of examples* with *features* for a specific task implemented with a certain *algorithm*

Today's learning objectives

After practicing with the concepts of today's lecture you are able to:

- Explain the concept of (implicit) bias in data and algorithms in Machine Learning

TUDelft

(Implicit) bias

Question 1

Everyone please stand up

You are sitting by a camp fire. Suddenly a gust of wind comes and the pile of burning wood collapses.

Q: What do you do?

Sit down if you move away from the fire as soon as you can

Question 2

Everyone please stand up

A baseball bat and ball together cost EUR 1.10.
The bat costs EUR 1 more than the ball.

Q: What does the ball cost?

Sit down if the ball costs EUR 0.10

What did you do?

- Came to a decision, very fast, *without explicit thought*

- This decision is often right, but can be wrong

TUDelft

What is bias?

- A preference or inclination for or against something

- <mark>Can be positive, negative, or neutral</mark>

- <mark>Often accompanied by a refusal to consider the merits of other points of view</mark>

- Example negative: Treating a person (e.g., O.J. Simpson) acquitted of murder as a convicted murdered

- Example positive: Treating people who you grew up with more favourably than others

TUDelft

What is prejudice?

- <mark>An assumption made without adequate knowledge</mark>

- Most commonly used to refer to a preconceived judgement toward a person or a group of people because of a personal or specific characteristic

- Usually *un*usually resistant to rational influence

- Example: Tom Cruise – talented actor, nice guy, but his affiliation with Scientology generates lots of negative press due to people's prejudice against Scientology

What is discrimination?

- The *actions* taken based on a prejudice

- Treating a person or group of persons based solely on their membership of a certain group or category

- The behaviour of excluding or restricting members of a group from opportunities that are available to people from another group

- Example negative: slavery, Holocaust, age-discrimination
- Example positive: age-discrimination

TUDelft

Who has biases?

We all do!

[See the earlier questions and the lab for this week]

TUDelft

What influences biases?

- Families
- Churches
- Friends
- Peers
- Neighbours
- Rules, regulations, and laws
- Social media
- Newspapers
- …

Implicit vs. explicit bias

TUDelft

Implicit bias

- Expectations based on learned coincidences, which unknowingly affect everyday perceptions, judgment, memory, and behaviour

- Subconscious thought

(The earlier two questions)

Explicit bias

- Is informed by our implicit bias but is also at least in part a conscious choice

→ conscious decision

- E.g., walking on the other side of the street when you see a scary-looking person

→ implicit bias

TUDelft

Q: Implicit bias – Why is it wrong?

- It *might* lead to discrimination (behaviour)

TUDelft

Bias in ML

- AI systems or ML techniques are not inherently "bad" nor turn "bad" by themselves

- An important source for bias: **the training database**

Q: What might go wrong with the training database?

Example of bias in the training database

An example

- I will show you two pictures

For each picture

- **Stand up** if it gives you a **positive** feeling
- **Sit down** if it gives you a **negative** feeling

TUDelft

Implicit human biases

- Greenwald et al. (*J. Pers. & Soc. Psy.*, 1998)

- Implicit Association Test (IAT): people find
  - Flowers 'similar' to pleasant
  - Insects 'similar' to unpleasant

➜ Bias of *flowers* towards *pleasant* and *insects* towards *unpleasant*

TUDelft

Cultural bias in language

- Caliskan et al. (*Science*, 2017)

- Word embeddings:
  - Tool to extract (semantic) associations between words/concepts
  - Each word is a vector in a vector space of 300 dim.
  - Computed on the context it keeps in large text corpora

➔ *flowers* vector has a closer distance to *pleasant* vector and *insects* vector to *unpleasant* vector

➔ Without knowing anything about the world!

T̃U Delft

So….

- Existing corpora have implicit human biases

TUDelft

Bias: example 1

Amazon scrapped 'sexist AI' tool

10 October 2018

An algorithm that was being tested as a recruitment tool by online giant Amazon was sexist and had to be scrapped, according to a Reuters report.

The artificial intelligence system was trained on data submitted by applicants over a 10-year period, much of which came from men, it claimed.

Reuters was told by members of the team working on it that the system effectively taught itself that male candidates were preferable.

[https://www.bbc.com/news/technology-45809919

Bias: example 2

Gender and racial bias found in Amazon's facial recognition technology (again)

Research shows that Amazon's tech has a harder time identifying gender in darker-skinned and female faces

By James Vincent | Jan 25, 2019, 9:45am EST

As facial recognition systems become more common, Amazon has emerged as a frontrunner in the field, courting customers around the US, including police departments and Immigration and Customs Enforcement (ICE). But experts say the company is not doing enough to allay fears about bias in its algorithms, particularly when it comes to performance on faces with darker skin.

The latest cause for concern is a study published this week by the MIT Media Lab, which found that Rekognition performed worse when identifying an individual's gender if they were female or darker-skinned.

https://www.theverge.com/2019/1/25/18197137/amazon-rekognition-facial-recognition-bias-race-gender

TUDelft

27
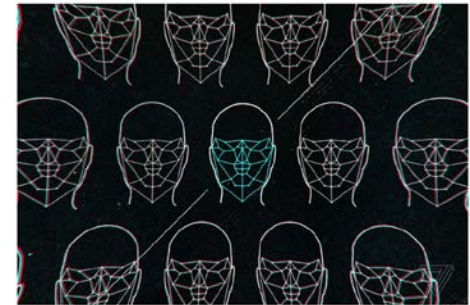
Bias: example 3

Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

By James Vincent | Mar 24, 2016, 6:43am EDT

It took less than 24 hours for Twitter to corrupt an innocent AI chatbot. Yesterday, Microsoft unveiled Tay — a Twitter bot that the company described as an experiment in "conversational understanding." The more you chat with Tay, said Microsoft, the smarter it gets, learning to engage people through "casual and playful conversation."

[https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist]

Bias: example 4

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Two Petty Theft Arrests

VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

LOW RISK 3

BRISHA BORDEN

Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

HIGH RISK 8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Technical aspects of bias and debiasing

The *algorithm* component can be unbiased and fair, if there is bias in the *data* used to make a decision, the decision itself may be biased

Debiasing → fairness

**Fairness:** "The absence of bias or discrimination on specific realms"

But definition depend on many aspects, such as domain, context, and social constructs

3 ways of quantifying fairness in ML

Fairness can be measured at different stages in ML pipeline: **training data** or **learned model**

Group level: equality between groups
Individual level: similar individuals treated similarly

Data vs. Model

Fairness

Group vs. Individual

WAE vs. WYSIWYG
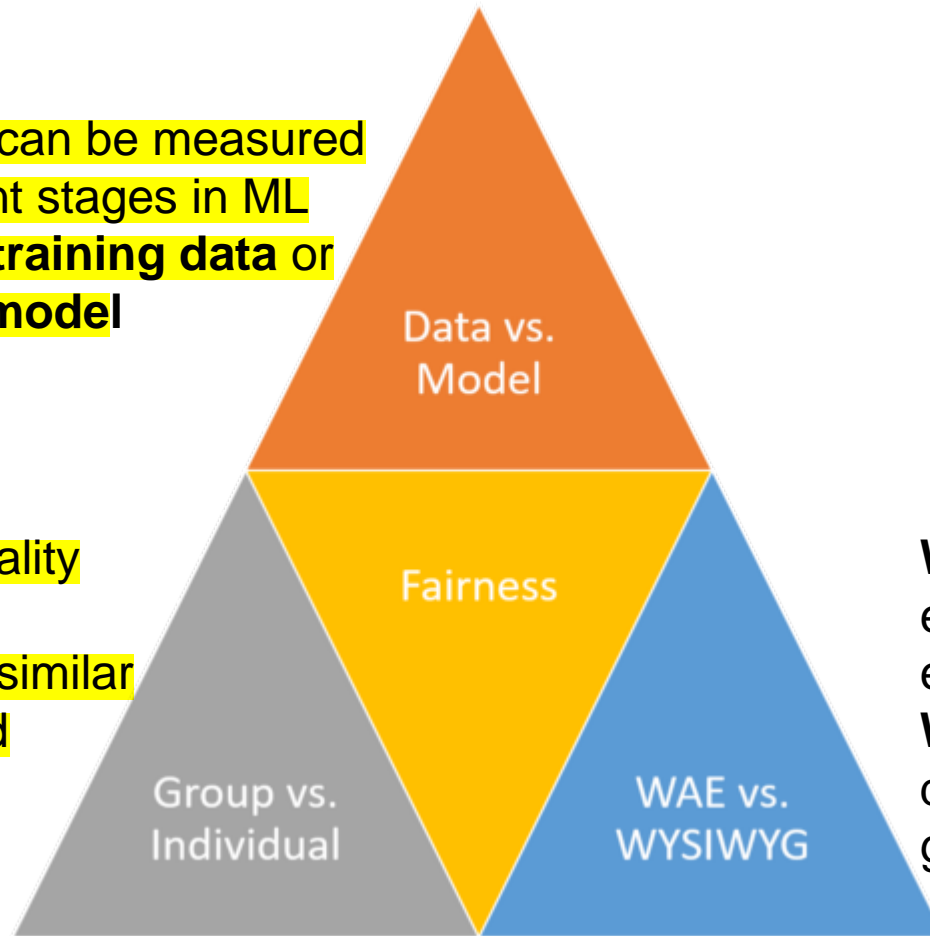
WAE (We are all equal): equal opportunities for everyone
WYSIWYG: people have different abilities for a given task

Figure from: https://towardsdatascience.com/whats-new-in-deep-learning-research-reducing-bias-and-discrimination-in-machine-learning-models-be116d3a571c

Debiasing training set / model

- Check the distribution of class labels in training set: group A >> group B?

➔ <mark>Equalize distribution in training set</mark>

TUDelft

Debiasing group vs. Individual level

- Similar outcomes for different groups, e.g.,
  - Smart algorithms
  - Carefully selecting the features used for the ML task, e.g., zip codes are well-known proxies for race, are often eliminated

  ➔ Be aware of correlations of variables with other variables that the algorithm uses, e.g., surnames with geographical census data

- Similar outcomes for similar individuals: much harder
  - 2 individuals on either side of the line are very similar but different outcome

T̃U Delft

WAE / WYSIWYG

- Through smart implementing of the algorithm

TUDelft

Debiasing tools

Example:

- AI Fairness 360: http://aif360.mybluemix.net

TUDelft

Other sources of biases

Q: What other sources besides biases in the training data could there be?

[3 mins] To discuss this with your neighbours

[The following slides are based on Bryson, "Three very different sources of bias in AI, and how to fix them"]

Other sources of biases

1. Lack of diversity in ML developers

2. Implicit human biases in our culture

3. Evil programmers

TUDelft

Lack of diversity in ML developers

- See example of the facial recognition software

- <mark>Solution? Diversify your developers team</mark>

TUDelft

Implicit human biases in our culture

- AI and ML have made so much progress because we know better how to transfer and represent the world around us into AI

- This view is dependent on culture → human bias

- Solution? Change the culture…..?

Evil programmers

- No ML/AI programme is self-learning
➔ All ML/AI programs are implemented by people

- Evil people can build in biases

- Solution? Check all software? By whom?

TUDelft

Biases are hard to discover: why?

A case study [Zouridis et al., 2019]

In the 'old' days:

- When you wanted a loan you went to see a person and explained your case

- Person used rules and common sense to make a decision

➔ "Street-level bureaucracy"

TUDelft [The following slides and case study are based on Zouridis, van Eck, Bovens, "Automated discretion"]

Introduction of IT and ML on decision-making

- Fill in a form on the internet
- Automated decision based on rules
- No room for your personal story
- No room for 'bending the rules'
➔ "System-level bureaucracies"

"Computer says no"

TUDelft

I, Daniel Blake

- Elderly carpenter suffers from heart attack
- Work capability assessment: fit for work
- Doctor: not fit for work

➔ Lost in bureaucracy
  - Computer illiterate
  - Forms need to be filled in online, processed digitally
  - Case managers:
    - Pre-programmed decision systems
    - Unwilling to sympathise and empathise

Human judgements based on rules of thumb

Automated decisions based on algorithms

== " The computer says 'yes' or 'no' "

TUDelft

Street-level bureaucracy

- Direct interaction/personal contact with individuals

- Rules and regulations leave room for professional discretion

- Case-by-case basis

- There is a scarcity of resources compared to the task to be done, i.e., more people are needed to check every single request

TUDelft

Introduction of IT

time

- Motivations for decisions were written down using word processors
- Then became standard blocks of text
- Decision-making algorithms were developed as a tool for the professional who
  - Input the data into a form
  - Checked the decision spit out by the algorithm
- → Professionals could twiddle and tweak the input to obtain the desired decision
- Automation of the input forms
- Individuals had to fill in their own forms

TUDelft

System-level bureaucracy

- The management of the organisation no longer checks whether applications are processed in a legitimate way

- 'Production' is checked quantitatively

- Most decisions based on rules and regulations implemented in the algorithm

- *Programmers* control the system

- Independent judge reviews individual cases *only* when there is an *appeal*
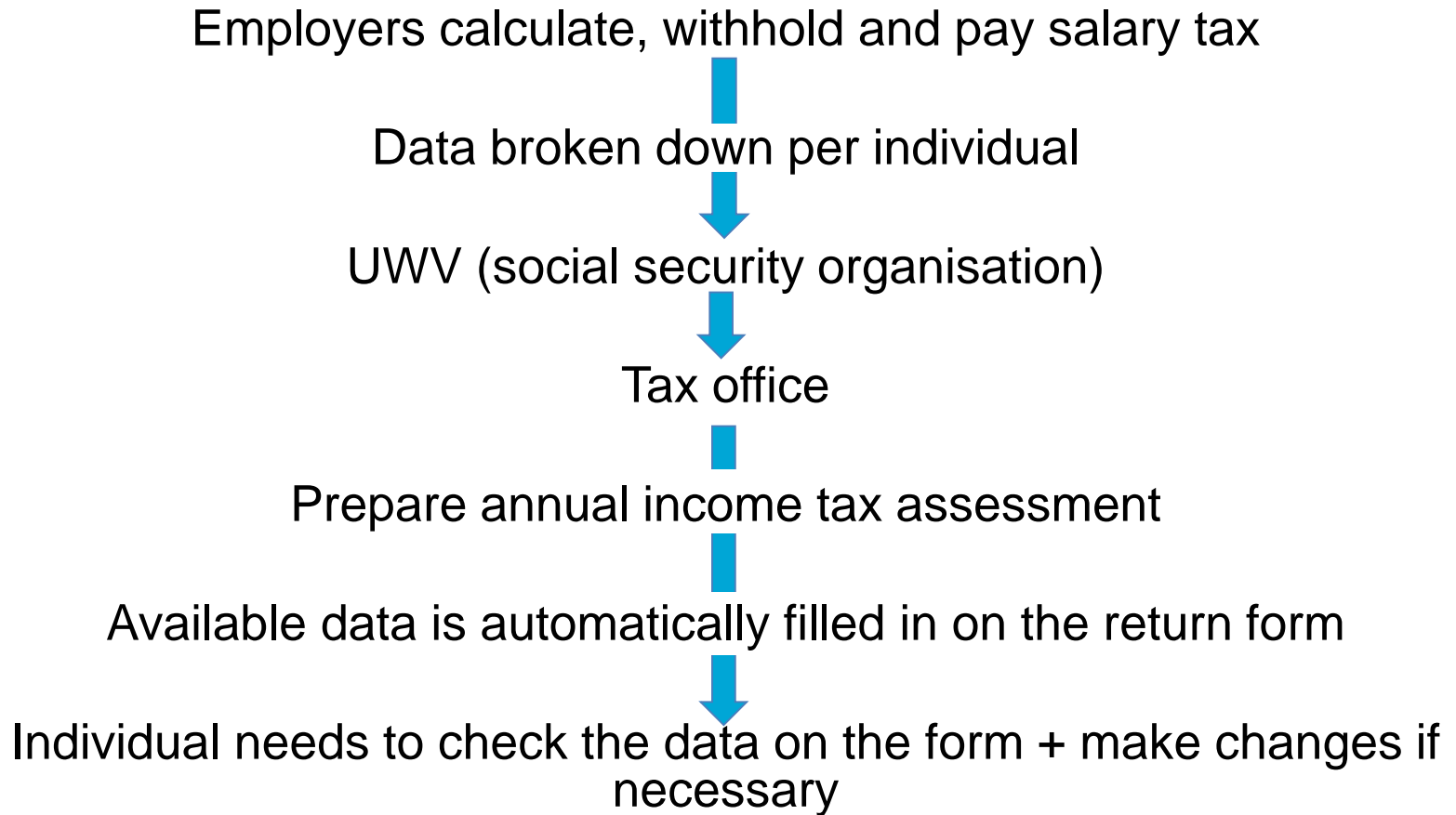
TUDelft

Chains of system-level bureaucracies

- Who is responsible for the decisions?

- Who *should be* responsible for the decisions?

An example of a system-level bureaucracy

- Tax authorities:
  - Levy salaries tax
  - National insurance contributions

- By cooperating with:
  - Employers
  - Benefits agencies (e.g., UWV)
  - Statistics office (Centraal Bureau van de Statistiek)

T̃U Delft

Processing and data stream

Employers calculate, withhold and pay salary tax

Data broken down per individual

UWV (social security organisation)

Tax office

Prepare annual income tax assessment

Available data is automatically filled in on the return form

Individual needs to check the data on the form + make changes if necessary

TUDelft

- 90% of the tax returns are decided by computer
  - Amount of tax to be paid
  - Determines a citizen's official income
  - ➔ National income database

- 10% handled by a tax official

➔ Based on pre-defined set of fraud risk rules

Q: How and where can problems occur?

TUDelft

Data

- Are provided by individuals or other companies
- Are given uniform definitions by the programmers
- Are re-used if they mean *approximately* the same
- Are not substantially checked before being included in the administrations of the individual

- Errors are hard to change by the individual
- Decisions are hard to revert

TUDelft

Decision-making algorithms

- Have not been developed with the aim of processing knowledge or performing analyses, but for calculations

- Data are applied as variables in mathematical formulas

- Information managers
  - Determine which data are used
  - Program the algorithms
  - Content of decision rules is untraceable

- IT applications are typically old, and have grown throughout the years
➔ The decision rules cannot be isolated from the administrative process even by IT experts

TUDelft

Decisions are thus

- Based on data from *other* organisations

- Made by the programmers of the software
  - But these programmers
    - Never see individuals
    - Never see individual cases
    - Are not decision-making experts

TUDelft

To return to our questions

- Who is responsible for the decisions?
  - Programmers

- Who *should be* responsible for the decisions?
  - ….

Imagine

Belastingdienst

shutterstock.com • 227718901

- Build algorithms
- Build new features
- Build a new skin, etc
- And then….

BY INVITATION only

TUDelft

Creative software developers team

- A separate group
- Freedom to develop your own ideas for new applications or improvements of existing system

Freedom

- Ideas are presented to the management
- After agreement, implementation

TUDelft

Q: What is not to like about that?

This group

- Has no formal decision-making authority

- Has never seen a case

- Has never met an individual

- Yet, develop software that will make important, sometimes life-changing, decisions for people

TUDelft

In case of errors or biases

- Who is responsible?

- How do you find out where the problem originates?

- How do you resolve the problem?

TUDelft

Data analysts

- Look for patterns in the data
- Can help detect fraud
- Can help distinguish between errors and fraud
- Identify people who are likely not to commit fraud
- Can suggest changes to the algorithms

➜ Can influence the decision making process with their analyses and suggestions

➜ What is not to like about that…?

Conclusion

Automatic systems:

- More efficient
- Processing of cases much faster
- Biases in the individual street-level bureaucrats no longer play a role in decision-making

But

- Little room to correct errors
- IT determines which cases are to be dealt with automatically/by hand
- Important roles for programmers with little background knowledge on decision making
- Cultural bias might enter the algorithms
- Important roles for parameters in algorithms!

➔New ethical rules are needed

TU Delft

Take-home message

Building fairness and non-discriminatory behaviour into AI models is not only a matter of technological advantage but of social responsibility

TUDelft

"I sure hope we can sign up for health care before we die of natural causes."

Final words

Paraphrasing Lily Hu (Harvard):

"The dominant perception within the machine-learning/AI community is that everything is fundamentally an *optimization* problem, or a *prediction* problem, or a *classification* problem. And when you do that—if you treat it in a standard machine-learning way—you will end up reinforcing those inequalities."

[Source and highly recommended read: https://www.harvardmagazine.com/2019/01/artificial-intelligence-limitations]

Suggested readings

Cathy O'Neil

Weapons of mass destruction: How Big Data Increases Inequality and Threatens Democracy

Virginia Eubanks

Automating inequality: How high-tech tools profile, police, and punish the poor

TUDelft