MachineLearnAthon - Microlecture
Fairness in Machine Learning

Recorded by Lara Kuhlmann

MachineLearnAthon
A project Co-funded by the Erasmus+ programme of the European Union

Learning outcomes of today

After successfully completing this micro-lecture, you are able to….

- Define fairness in machine learning and explain why it is essential in high-stakes domains
- Identify and compare different fairness metrics
- Recognize the limitations and incompatibilities among fairness metrics in practice
- Comprehend bias as the opposite of fairness in machine learning and recognize its various forms and sources.
- Describe bias mitigation strategies
- Understand the ethical and contextual dimensions of fairness

Co-funded by the Erasmus+ programme of the European Union

Agenda for today

- Definition of Fairness in ML
- Fairness Metrics
- Bias in ML
- Forms of Bias
- Bias Mitigation Strategies
- Ethical & Practical Challenges

Co-funded by the Erasmus+ programme of the European Union

Fairness in ML

- Definition: Fairness is the absence of any prejudice or favouritism toward an individual or group based on their inherent or acquired characteristics.[1]
- Why it matters: ML systems influence hiring decisions[2], lending[3], policing, and healthcare[4]. Unfair models may amplify social bias and discrimination.
- Example: A biased model denies loans more frequently to applicants from a certain ethnicity due to skewed training data.

[1] Mehrabi et al. 2019
[2] Miranda Bogen and Aaron Rieke. 2018
[3] Mukerjee et al. 2002
[4] De Fauw et al 2018

Fairness Metrics

- Demographic Parity:
  The likelihood of a positive outcome should be the same regardless of whether the person is in the protected group.[1]
- Equal Opportunity:
  This means that the probability of a person in a positive class being assigned to a positive outcome should be equal for both protected and unprotected group members.[1]
- Equalized Odds:
  The probability of a person in the positive class being correctly assigned a positive outcome and the probability of a person in a negative class being incorrectly assigned a positive outcome should both be the same for the protected and unprotected group members.[1]
- Calibration:
  Given a prediction score, its meaning should be consistent across groups.

[1] Sahil Verma and Julia Rubin. 2018

Co-funded by the Erasmus+ programme of the European Union

Bias in ML

- Definition: Bias in ML is defined as a systematic error in decision-making processes that results in unfair outcomes.[1] It is the opposite of fairness.
- Why it matters: Biased models can amplify existing societal biases and discrimination in high-stakes domains.
- Sources of Bias
  - Data Collection: Data bias occurs when the data used to train machine learning models are unrepresentative or incomplete, leading to biased outputs.[1]
  - Feature Selection: Including or excluding features that inadvertently capture or amplify societal biases.
  - Model Design: Algorithmic choices or objective functions that prioritize overvall accuracy over fairness across subgroups.
  - Human Decisions: This can happen when users provide biased training data or when they interact with the system in ways that reflect their own biases.[1]

[1] Ferrara 2023

Co-funded by the Erasmus+ programme of the European Union

Bias Mitigation Strategies

- Pre-processing:
  Modify training data (e.g., reweighting, resampling, obfuscating sensitive features).
- In-processing:
  Incorporate fairness constraints into the learning algorithm (e.g., fairness-aware loss functions).
- Post-processing:
  Adjust predictions after training (e.g., by altering thresholds).

Mehrabi et al. 2019

Co-funded by the Erasmus+ programme of the European Union

Ethical & Practical Challenges

- Representation harm:
  Stereotypical outputs may emerge even without direct bias.
- Trade-offs:
  Fairness vs. accuracy —improving one may harm the other.[1]
- Context matters:
  What's fair can vary by culture, law, or stakeholder perspective.
- Explainability:
  Addressing the complexity of human behavior and decision-making.[1]

[1] Ferrara 2023

Co-funded by the Erasmus+ programme of the European Union

Recap this lecture

After successfully completing this lecture, you are able to….

- Define fairness in machine learning and explain why it is essential in high-stakes domains
- Identify and compare different fairness metrics
- Recognize the limitations and incompatibilities among fairness metrics in practice
- Comprehend bias as the opposite of fairness in machine learning and recognize its various forms and sources.
- Describe bias mitigation strategies
- Understand the ethical and contextual dimensions of fairness

Co-funded by the Erasmus+ programme of the European Union

The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

This material was created with the assistance of an AI language model.

Co-funded by the Erasmus+ programme of the European Union