

CSC 301: Analysis of Algorithms

Lecture 22: Bias in Algorithms

William Rebelsky
[rebelskyw]

Grinnell College

Monday 11th March, 2024

Reminders and Announcements

- The Midterm is due Wednesday
- The Homework is due Wednesday
- The Essay will be released Wednesday and Due after break
- Please do the readings this week
- I will make summaries of how everyone is doing to send out after break (this will not include tokens)

Table of Contents

1 Bias in Algorithms

Plans for this week

- Today: <10min of lecture/questions
- Rest of today: Pseudorandomly assigned discussion groups
- Rest of week: Pseudorandomly assigned discussion groups talking about specific readings
- Goals: Be able to understand and articulate the interplay between algorithms, technology, bias and structural inequity, read the literature of computer science

Readings for this week

- Human-like Biases
- Debiasing Word Embeddings
- Racial Bias in Health Care **Wednesday Discussion**
- Big Data's Disparate Impact **Friday Discussion**
- Data Before Algorithms **Wednesday Discussion**
- Algorithms of Oppression (Introduction and 1A) **Friday Discussion**
- Feynman's Error

Essay

- Work should be written individually
- The essay should include the following:
 - A reflection on the ways in which bias can appear in the design or use of algorithms
 - A reflection on the importance of data sets
 - A reflection on the importance of considering bias when designing, implementing, and using algorithms
 - Ways in which computing professionals can address bias in algorithms
 - One paragraph on how we might improve coverage of these issues as a department
- Formatting requirements: Single Spaced, 1 inch margins, standard 8.5x11 inch paper, reasonably sized font (10-12), reasonable font choice. Length: You are all responsible students, it should be long enough to cover all of the requirements above but not so long that it loses focus.

Fairness vs Accuracy

- COMPAS: Recidivism Prediction tool. Argument about Predictive parity and accuracy equity vs real world results. Easier to read analysis
- Incompatible Fairness Definitions: Traditional definitions of fairness are mutually incompatible (e.g. balancing false negative rate means you can't balance false positive rate).
- Fairness needs to be well defined
- Trade off between accuracy and fairness:
 - *Discrimination definition 1*: differentiate between groups, the goal of predictive algorithms (e.g. should a loan be provided)
 - *Discrimination definition 2*: the unjust or prejudicial treatment of different categories of people, especially on the grounds of ethnicity, age, sex, or disability (i.e. how we normally think about it).
 - If there was discrimination evident in the training data, an accurate algorithm will perpetuate the discrimination

Examples

- Survey reference
- Hiring at Amazon
- Facial Recognition
- UK estimates of exam scores during the pandemic
- Image Generation (2 links)
- Text Generation (2 links)

Recent Example in Practice

- <https://fastsdxl.ai/>
- This tool appears to have been taken offline recently. It is still possible to fork the github repo and create your own version to test
- Most AI tools are trained on similar data sets
- Unlike other tools, this one is optimized for inference, or guessing extremely quickly what the user wants
- This makes it much easier to see the underlying bias in the training data
- “A Student” vs “A Student who is good” vs “A Student who is good at math” vs “A Student who is good at math and basketball”: (White, White woman with glasses, White student normally* male, Black male student with glasses holding a basketball)
- “bad/good girl/boy”: (good boy is always* a dog)
- “Physicist”: (white guy, normally* old)