Fairness in Machine Learning

CS60050 - (27/02/2019)
IIT Kharagpur

Bias

- Bias is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting). (ML)
- ‘Slant’-- present or view (information) from a particular angle, especially in a biased or unfair way.(English)

Bias

I do not buy damaged fruit. (Neutral)

- None of the students from the last benches will be given good marks. (Moral)

Biased System

- A system that systematically and unfairly discriminate against certain individuals or group of individuals in favor of others.
- A system discriminates unfairly if it denies an opportunity or if it assigns undesirable outcome to a group of individuals on grounds that are unreasonable or inappropriate.

Automated Credit Advisor

- System 1- Denies credit to individuals with consistently poor payment records.
- System 2- Systematically assigns poor credit ratings to individuals with ethnic surnames.

Which of the above is a biased system?

Categories of Biases

- Pre Existing Bias
- Technical Bias
- Emergent Bias

Pre Existing Biases

- Has its roots in social institutions, practices and attitudes.
- Can enter a system either through explicitly or implicitly even in spite of the best of intentions.
- For e.g. low credit ratings being given to applicants who live in ‘undesirable’ locations such as: low- income or high-crime neighborhoods as indicated by the PIN of their home address.

  (Group Redlining)

Technical Bias

- Arises from the resolution of issues in the technical design
  - Ranking
  - Click
  - Mouse Movement
  - Reading / Writing

Emergent Bias

- Not possible to identify at the time of creation or implementation: arises only in a context of use
- Result of societal knowledge, population or cultural values

Machine Learning

Given the importance of machine learning in decision making in today's world, we need to take care of these biases in ML systems. (Discover and Mitigate)

These concerns has called for a new domain of research in ML i.e. Fairness in Machine Learning.

Relevant Materials

- Bias in Computer Systems [pdf]
- Fairness in Machine Learning [slides] [talk]
- Causal Inference [article]