Ethics in Data Science (Part 3, Fairness and Interpretability)

Daniel Lawson — University of Bristol

Lecture 12.1.3 (v1.0.2)

Signposting

- Part 1 covers ethics and the law,
- Part 2 covers Privacy and disclosure,
- This is part 3 covering Fairness and interpretability.

Interpretable Data Science

- How can we attribute interpretability to decisions?
- Two main classes of solution:
  - Interpretable algorithms,
  - Explaining black-box decisions through counter-factuals.
- Book: "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable." Christoph Molnar 2019.

What is Interpretability?

- Molnar, chapter 2:
  - “Interpretability is the degree to which a human can **understand the cause** of a decision.”
  - “Interpretability is the degree to which a human can **consistently predict** the model’s result.”
- There is a continuous trade-off between:
  - **specific explanations** regarding individuals (“the decision would change if. . . ”)
  - more **general explanations** for many decisions (“these factors are important. . . ”)

Interpretable algorithms

- Many algorithms such as decision trees, linear classifiers, explicit statistical models, all incorporate explicit notions of **why** a decision was made.
- It is then a “simple” matter of **examining the algorithm** to attribute the why to a particular decision.
- Examples:
  - Why did the decision tree refuse me a loan? *Specifically*: Because my income was less than £50k, and my postcode was disfavorable. *Generally*: Because income is a strong predictor of repayment.
  - Why did the Bayesian model refuse me a loan? *Specifically*: Because the posterior probability of repayment was less than 75%. *Generally*: Because income is a strong predictor of repayment.

Interpreting black box algorithms

- If we only have a black box, we can provide it with **different inputs** and see how it responds.
- Example: why did the neural network refuse me a loan? **Counterfactually**, it would have accepted if:
    - I had earned at least £50K...
    - I lived in a neighboring postcode...
    - I had repaid a credit card debt of at least £10K...
- We can also peel back the black box, for example, attributing **local differentials** to each attribute.
- Neural networks are not quite black boxes. There is a growing literature on interpretability.
- This is currently inconclusive and can be model dependent.
    - For example, there may be non-monotonicity ("earning more makes you more likely to receive a loan, unless...")
    - Interpretability can therefore require changes or constraints to algorithms.

Algorithmic Fairness

- Are algorithms fair? To find out we have to try to interpret them.
- Algorithms can be sexist, racist, ageist, and many other types of -ist.
- They do this by observing **associations** between variables and the outcome, in the **training data**.
  - hypothetically: non-whites may historically have failed to pay back their loans more than whites.
  - race becomes a predictor of repayment failure.
- So should we omit race from the data?
- Big data can facilitate **proxy discrimination** by means of non-protected attributes (e.g. postcode) that correlate strongly with protected attributes (e.g. race)
- It has been shown robustly that **protected attribute data** must be collected, in order to test algorithms for fairness. The algorithm must still not use them.

Why is algorithmic fairness a problem?

- Besides the **legal** problem, there is an important **ethical** problem in algorithmic bias
  - Current algorithms don't understand **causation**, only **correlation**
  - They certainly don't understand **sampling bias**
  - Therefore, they will tend to penalize **any** historically marginalized group!
  - If algorithms affect life, this leads to a cycle of bias that, without intervention, may never stop
- Example: Consider a historically poor city, B. Being from B was historically associated with failed loan repayments. Fewer mortgages are given in B and on worse terms. B therefore remain a poor city, attracting fewer businesses and fewer upwardly mobile people.

Counterfactuals and proxy data

- Counterfactuals are useful for understanding bias.
- But it is **not enough to replace** one attribute with another, in order to generate a counterfactual. All attributes that are correlated with that attribute, but are not considered meaningful for the decision, must be updated.
- For example, suppose we are testing our algorithm for racism. Race can be predicted from postcode, friendship groups, facebook likes, retweets, skin reflectance, socio-economic status, etc. Whatever is in the data needs to be re-examined.
- i.e. we need a **counterfactual model**.

How to address algorithmic fairness?

- **Data, data, data!** As with all data science, data is key. If the data are biased the answers will be too.
- **Algorithm choice.** There will be biases in your data, no matter now hard you try. You can model sources of bias, use counterfactual reasoning, etc.
- **Monitor performance.** Collect the sensitive data and check that your algorithm is actually fair with respect to race, gender, etc.
- It is not a solved problem!

Measures of algorithmic fairness

- Do two people, who are the same in all **meaningful** respects but R,
  - have the same **Equality of outcome**? i.e. have the same rate of success in outcome, e.g. receive the same loan when they applied for it?
  - have the same **Equality of opportunity**? have the same opportunity, e.g. **without applying**, would they still receive the same loan if they wished to?
- These can be quite different because there are many processes preventing certain groups from desiring a particular outcome.
- For example, there are fewer women in data science.
  - Do women have the same success rate as men, on application?
  - Do women have the same opportunity to enter it?
  - These may differ if e.g. women do not choose data science unless they are excellent at it,,,, (selection bias)
  - Or if they are poorly prepared due to previous choices of training.

Example

Hardt, Price and Srebo Equality of Opportunity in Supervised Learning 2016 explored in https://blog.acolyer.org/2018/05/07/equality-of-opportunity-in-supervised-learning/.

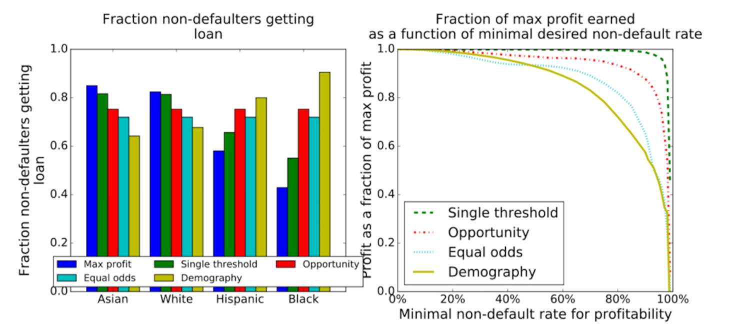

Figure 11: On the left, we see the fraction of non-defaulters that would get loans. On the right, we see the profit achievable for each notion of fairness, as a function of the false positive/negative trade-off.

Discussion

- Interpretability is a key component in ensuring fairness.
- Interpretability is typically created through either interpretable models, or counterfactual exploration.
- Equality is a very important concept:
  - Equality of opportunity is a better measure than equality of outcome.
  - This does not need to be costly with respect to a loss function.
- This is an active area of research.

Reflection

- What are the benefits and challenges surrounding interpretability?
- How would you go about justifying the decisions of your own Neural Networks?
- How does this differ to justifying the decisions of a linear regression?
- What responsibility does the data scientist have for algorithmic fairness?
- By the end of the course you should:
  - Be able to describe the main ways algorithms are interpreted,
  - Be able to use the two main definitions of algorithmic fairness.

References

- Algorithmic Bias Tutorial by Francesco Bonchi with Slides from KDD 2016
- Book: "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable." Christoph Molnar 2019.
- Hardt, Price and Srebo Equality of Opportunity in Supervised Learning 2016 explored in https://blog.acolyer.org/2018/05/07/equality-of-opportunity-in-supervised-learning/.