Ethics and Fairness in Data Science

COMP5310 Lecture 12

Prepared by Prof Alan Fekete

University of Sydney

Includes material based on slides by
Lisa Getoor (UCSC) and Tina Eliassi-Rad (Northeastern)

Recall: Ethics in DS process

- Rights to use datasets
  - obtaining permission, licensing rules, etc
- Care of the datasets
  - risk of privacy breach if you hold personal data
  - anonymisation is not always perfect
- Competence with the tools
  - too often, analysis is done without adequate awareness of the limitations of the approach chosen
- Built-in bias
  - eg voice recognition that doesn't handle many accents
  - eg medical datasets that have many more men than women (so predictions may be inaccurate for women)

Recall: Ethics in DS outputs

- Conclusions and automated systems can be used for good, or for ill
  - any professional has some responsibility for uses made of their work, if they can reasonably expect those uses to happen
- Eg Volkswagen detected the pattern of use in air-quality tests, and switched to lower-polluting, lower-performance mechanisms just while the test was on
  - https://www.nature.com/articles/nature.2015.18426
  - what would DS team think, when asked to build a system to detect occurrence of these tests?

Professional Ethics

- Professions typically define the expected behaviour of professionals
- Often, adherence to code is required for membership of professional associations
  - And this in turn may be required before working in the profession (eg medical practice, legal practice)
    - But, this is not yet common for data science or computing activities

Stakeholders

- You have obligations to the client or employer
- But these are not the only obligations
  - Don't obey unlawful orders (Nuremberg trials)
  - You should aim to be comfortable if people know what you did
- Think about everyone who can be impacted by what you do
  - Subjects who are described by data
  - Workers whose jobs might be displaced
  - End-users whose environment might be shaped

From Computing Professional Ethics

- Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing
- Avoid harm
- Be fair and take action not to discriminate
- Respect privacy
- Respect the work required to produce new ideas, inventions, creative works, and computing artifacts
- Perform work only in areas of competence
- Ensure that the public good is the central concern during all professional computing work

From ACM Code of Ethics and Professional Conduct, for computing professionals

See https://www.acm.org/code-of-ethics

More detail

Be fair and take action not to discriminate

"The values of equality, tolerance, respect for others, and justice govern this principle. Fairness requires that even careful decision processes provide some avenue for redress of grievances.

Computing professionals should foster fair participation of all people, including those of underrepresented groups. Prejudicial discrimination on the basis of age, color, disability, ethnicity, family status, gender identity, labor union membership, military status, nationality, race, religion or belief, sex, sexual orientation, or any other inappropriate factor is an explicit violation of the Code. Harassment, including sexual harassment, bullying, and other abuses of power and authority, is a form of discrimination that, amongst other harms, limits fair access to the virtual and physical spaces where such harassment takes place.

The use of information and technology may cause new, or enhance existing, inequities. Technologies and practices should be as inclusive and accessible as possible and computing professionals should take action to avoid creating systems or technologies that disenfranchise or oppress people. Failure to design for inclusiveness and accessibility may constitute unfair discrimination."

From ACM Code of Ethics and Professional Conduct, for computing professionals

See https://www.acm.org/code-of-ethics

What about Data Scientists?

- Many data science activities are done by people from different professional backgrounds and training, so they may belong to different professional organisations with different codes
- Data Science Association is one organisation seeking to cover data science activities
- See its code of conduct at https://www.datascienceassn.org/code-of-conduct.html
  - Some examples on following slides

Professional Ethics for Data Scientists (sample)

- " A data scientist shall provide competent data science professional services to a client. Competent data science professional services requires the knowledge, skill, thoroughness and preparation reasonably necessary for the services."
- "A data scientist shall use reasonable diligence when designing, creating and implementing algorithms to avoid harm. The data scientist shall disclose to the client any real, perceived or hidden risks from using the algorithm. After full disclosure, the client is responsible for making the decision to use or not use the algorithm. If a data scientist reasonably believes an algorithm will cause harm, the data scientist shall take reasonable remedial measures, including disclosure to the client, and including, if necessary, disclosure to the proper authorities. The data scientist shall take reasonable measures to persuade the client to use the algorithm appropriately."
- "If a data scientist reasonably believes a client is misusing data science to communicate a false reality or promote an illusion of understanding, the data scientist shall take reasonable remedial measures, including disclosure to the client, and including, if necessary, disclosure to the proper authorities. The data scientist shall take reasonable measures to persuade the client to use data science appropriately."

Professional Ethics for Data Scientists (some more)

- “A data scientist shall rate the quality of data and disclose such rating to client to enable client to make informed decisions. The data scientist understands that bad or uncertain data quality may compromise data science professional practice and may communicate a false reality or promote an illusion of understanding. The data scientist shall take reasonable measures to protect the client from relying and making decisions based on bad or uncertain data quality.”
- “A data scientist shall rate the quality of evidence and disclose such rating to client to enable client to make informed decisions. The data scientist understands that evidence may be weak or strong or uncertain and shall take reasonable measures to protect the client from relying and making decisions based on weak or uncertain evidence.”

Professional Ethics for Data Scientists (some more)

- "A data scientist shall protect all confidential information, regardless of its form or format, from the time of its creation or receipt until its authorized disposal."
- "A data scientist may reveal information relating to the representation of a client to the extent the data scientist reasonably believes necessary:
  - (1) to prevent reasonably certain death or substantial bodily harm;
  - (2) to prevent the client from committing a crime or fraud that is reasonably certain to result in substantial injury to the financial interests or property of another and in furtherance of which the client has used or is using the data scientist's services."
- "A data scientist shall make reasonable efforts to prevent the inadvertent or unauthorized disclosure of, or unauthorized access to, information relating to the representation of a client, which means:
  - (1) Not displaying, reviewing or discussing confidential information in public places, in the presence of third parties or that may be overheard;
  - (2) Not e-mailing confidential information outside of the organization or professional practice to a personal e-mail account or otherwise removing confidential information from the client by removing hard copies or copying it to any form of recordable digital media device; and
  - (3) Communicating confidential information only to client employees and authorized agents (such as attorneys or external auditors) who have a legitimate business reason to know the information."
- "A data scientist shall comply with client policies that apply to the acceptance, proper use and handling of confidential information, as well as any written agreements between the data scientist and the client relating to confidential information."
- "A data scientist shall protect client confidential information after termination of work for the client."
- "A data scientist shall return any and all confidential information in possession or control upon termination of the data scientist - client relationship and, if requested, execute an affidavit affirming compliance with obligations relating to confidential information."

Data quality

- Data quality is essential for results to be useful
  - "garbage in, garbage out"
- Importance of getting good data sources
  - The whole provenance chain needs to be trusted
  - Data should be representative of the actual domain
- Then cleaning data
  - doing so in ways that improve data quality as much as possible, for the particular uses

Tool quality

- Aim to automate the analysis, and test a lot
  - Easier in Python than in Excel
    - Many spreadsheets have un-noticed errors in formulas and data
- Know the limitations of any technique you use
  - Check the applicability
    - Eg make sure data looks linear, before using linear regression
    - Scatterplots with all the data are very helpful in spotting overall trends, at least for each pair of attributes
  - Be sceptical, and ready to step away from your initial findings
  - Be open about limitations when communicating results
    - Don’t overclaim

Asset management

- Effective management of project assets (data and code) is important for
  - Quality outcomes
  - Confidence in results (reproducible etc)
  - Privacy and confidentiality
- Especially, access control (both policy and effective mechanisms) has huge impact on client's business interests, and on subjects whose data is stored

Privacy

- People have a legal and moral right to privacy
  - This is often considered to include the right to keep others from knowing personal data
    - including contact information, medical records, financial records
    - What about location, browsing history,
  - In some jurisdictions, this also includes the right to know what information about you is held by others, and the right to require them to correct or remove that information
  - See https://en.wikipedia.org/wiki/Privacy

The value of data about people

- Data about people can be very helpful in improving computer-based services
  - Eg personalization
  - Eg effective user interfaces
- Also, data about people is essential to effective business decision-making and public policy choices
  - But maybe the decisions can be done without individual data that is tied to a particular person, but rather to aggregates or trained predictive models

Anonymization

- A common approach in data science projects which use data about people, is to produce anonymized datasets
  - remove the “personal identification” but keep the other relevant details
    - Eg replace the customer-name by a random other id
      - Anonymized-id is not linked to name, address, etc
- A challenge is from “re-indentification”, often there are only one or few people who share certain attributes
  - Eg if you know birthdate, postcode, gender, this might be enough to find the name by matching with public records
  - See https://en.wikipedia.org/wiki/Data_re-identification especially work of Professor Latanya Sweeney

Aggregate summaries

- Using census datasets etc for social science analysis
  - Usually, social scientists want to study trends over groups of people, rather than facts about individuals
  - So release data on aggregates
    - Eg average income, in each suburb
    - Eg number of people who own their own home, in each (income-bin,family-size,suburb) combination
- Clearly, an aggregate where only one person is included, reveals facts about that person
  - So a common approach is to refuse to release any summary over a group with too few people
- But, someone can deduce about one person, by combining aggregates if they know the set of people differ only in that one person

Data perturbation

- Introduce random errors into the data
  - Eg store salary which is actual salary plus a random amount
- Aggregates over big collections will be quite accurate, as errors cancel out
- A special form of this is now being used in US Census, to ensure “differential privacy”
  - See https://www.npr.org/2021/05/19/993247101/for-the-u-s-census-keeping-your-data-anonymous-and-useful-is-a-tricky-balance

Fairness in ML

- “Algorithmic systems are being adopted in a growing number of contexts, fueled by big data. These systems filter, sort, score, recommend, personalize, and otherwise shape human experience, increasingly making or informing decisions with major impact on access to, e.g., credit, insurance, healthcare, parole, social security, and immigration. Although these systems may bring myriad benefits, they also contain inherent risks, such as codifying and entrenching biases; reducing accountability, and hindering due process; they also increase the information asymmetry between individuals whose data feed into these systems and big players capable of inferring potentially relevant information.”

From https://fatconference.org

ML -> action

- When a computer system plays a role in decisions or actions, any unfairness in the computer's operation can lead to unfair experiences for people
  - This holds even if the computer's output is simply information considered by a human decision-maker, alongside other sources of information
- Some examples:
  - algorithms that influence what ads are shown to someone, can impact on the opportunities they get
  - algorithms that predict whether a credit card payment is fraudulent, can impact whether someone can purchase goods
  - Algorithms that predict whether a bank loan will be repaid, can lead someone to pay a higher interest rate and thus lose money

Terminology

- Informally, we often speak of bias in behaviour
  - But “bias” has (very different) technical meaning in statistics
- Legal rules are often expressed about “discrimination” or “unfairness”

Law

- Fundamental human rights principles, including being treated fairly, come from international treaties eg International Covenant on Economic, Social and Cultural Rights
- Different countries takes these principles into law with different criteria and different circumstances
- In Australia, “Discrimination can be against the law if it is based on a person’s:
  - age
  - disability, or
  - race, including colour, national or ethnic origin or immigrant status
  - sex, pregnancy, marital or relationship status, family responsibilities or breastfeeding
  - sexual orientation, gender identity or intersex status.
- Discrimination on these grounds is against the law in a number of areas of public life, including: employment, education, getting or using services or renting or buying a house or unit. Some limited exceptions and exemptions apply.”
  - [https://humanrights.gov.au/quick-guide/12030]

Issues

- How might discrimination arise in use of computer-based data-driven systems?
  - Unlikely that a ML algorithm explicitly considers the value of a protected attribute, and deliberately builds different models for the different groups
  - But the data that is used to build models, may not be unfairly chosen
    - Perhaps more samples from one group than its share of the population
      - Eg if data is gathered from a source that is more used by one group
  - The data may include the value of a protected attribute, and the model produced by training may use that attribute in its predictions
    - Thus treating cases differently based on that protected characteristic, even if otherwise identical
  - Even if protected attributes are not in data, perhaps other attributes correlate with protected ones
    - Eg names often provide hints about ethnicity and gender
    - Eg browsing to particular sites, may correlate with sexual orientation
  - The predictive model may reflect facts that arise in a discriminatory society, or from past discriminatory behaviors

Case study I

- Voice recognition software
- In 2016, a study found that Google's voice recognition identified more words in male speakers, than in female speakers
  - Most likely, because training sets had more male examples
- Using female persona for voice-based assistants, may perpetuate gender stereotypes

See https://makingnoiseandhearingthings.com/2016/07/12/googles-speech-recognition-has-a-gender-bias/ and https://theconversation.com/artificial-intelligence-has-a-gender-bias-problem-just-ask-siri-123937

Case study II

- Amazon automated resume processing
  - gave higher scores to resumes from men than from women
- If "ground truth" comes from a biased setting, then ML will learn the same bias
- If ML system is deployed and shapes future actions, then future datasets will also have inbuilt bias

See https://theconversation.com/amazons-sexist-hiring-algorithm-could-still-be-better-than-a-human-105270

Case study III

- Recividism risk assessment
- COMPAS system
  - "Correctional Offender Management Profiling for Alternative Sanctions"
  - Developed by Northpointe using ML techniques
    - Intended to reduce racial bias of humans
  - Used by judges in parole and sentencing decisions in several USA jurisdictions (including California)
- Criticised for racial bias by ProPublica in 2016

See https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing and https://www.technologyreview.com/s/607955/inspecting-algorithms-for-bias/

COMPAS operation

- “Northpointe’s core product is a set of scores derived from 137 questions that are either answered by defendants or pulled from criminal records. Race is not one of the questions. The survey asks defendants such things as: “Was one of your parents ever sent to jail or prison?” “How many of your friends/acquaintances are taking drugs illegally?” and “How often did you get in fights while at school?” The questionnaire also asks people to agree or disagree with statements such as “A hungry person has a right to steal” and “If people make me angry or lose my temper, I can be dangerous.””

Quote from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Effectiveness of COMPAS

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Image from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Measurement of bias

- Recall confusion matrix

Predicted \ Actual	actually has property	Actually doesn't have property
Positive (predicted to have property)	TP (True Positive)	FP (False Positive)
Negative (predicted as not-with-property)	FN (False Negative)	TN (True Negative)

- Ideal parity
  - The proportions in each cell would be (nearly) the same in each group of subjects
- Accuracy parity
  - Look for (nearly) equal values of (TP+TN)/(TP+FP+FN+TN) in each group of subjects
- Precision parity (positive predictive parity)
  - Look for (nearly) equal values of TP/(TP+FP) in each group of subjects
- Equalized odds
  - Look for both (nearly) equal values of TP/(TP+FN), and (nearly) equal values of FP/(FP+TN) in each group of subjects

Impossibility!

- If the groups have different rates for the actual condition, and an algorithm isn't perfect, then the algorithm can't have BOTH equalized odds AND precision parity!
  - Certainly can't have ideal with all proportions (nearly) the same in different groups
- "Fairness" of a system depends on how we decide to define fairness

But what is the alternative?

- People are also biased in decision-making
- Should we ask only that algorithm is less discriminatory than people?
  - Which people?
  - Could we reduce discrimination in people (training etc)?
- Is it worth having more discrimination, if the predictions are also more accurate overall?

Change the prediction, or the decision?

- What should we do if, for example, an algorithm predicts that blacks are less likely to repay a loan than whites with similar income etc?
  - One proposal: the model is discriminatory, so adjust the predictive model till it treats different races the same
  - An alternative: aim for most accurate prediction, and then introduce affirmative action to make loans anyway to disadvantaged groups
  - Lots of controversy: see eg https://spectatorworld.com/life/militant-liberals-politicizing-artificial-intelligence/

At least, think about it!

- Top ML research conferences now require each paper to discuss the “broader impacts”, and these discussions are considered by experts on ethics
- See https://statmodeling.stat.columbia.edu/2020/12/21/the-neurips-2020-broader-impacts-experiment/