CS 165: Foundations of Machine Learning **Winter 2019**

Lecture 17: March 16th 2019

Lecturer: Anima Anandkumar *Scribes: Qilin Li, Wen Gu*

We will label each section with the corresponding slides page number in () for your reference.

17.1 Introduction

17.1.1 Overview ($2 \sim 8$)

Machine learning, from a toy example to the applications in real life, has a broader implication in terms of bias, fairness, ethics and all the tricky issues. The mathematical model, data set used, and design of architecture for machine learning are all considered in ideal cases, but the things will change once they are brought to real life. In fact, there are more questions than answers regarding the broader implications rather than the traditional topic covered in the area of machine learning. What's more, the technology may not be able to reduce bias and prejudice as we expect.

Our history, bias, and prejudices will impact the way we create the data sets. These biased data sets will be used to train the AI models, and the biased AI models will affect the way they make decisions. Consequently, these decisions will also influence the human's social issues such as bias and prejudices. This closed loop amplifies the social problems when we do not take consideration of the side effect of technology.

For example, the ImageNet data set contains a lot of dogs and cats, but this data set cannot represent the whole picture of the real world in the sense that we have many things other than dogs and cats. In addition, modern technology is led by western civilization, and thus the value holds by western culture would also impact the way data set is created and the way technology is used.

Source of Bias
The source of bias contains but not exclusive to

- Data itself; The way people create the data set can introduce the bias.
- The objective function of machine learning; The objective function of machine learning is only to maximize the accuracy and other aspects are ignored. For example, the recommendation system only focuses on click rate, while fake news is also introduced to gain the attention from people as side effect.
- The outcome of the machine learning; The revenue maximization would also lead to unexpected political result(e.g. Facebook).

Now, AI is also used in law enforcement, such as face recognition and different kind of surveillance to determine who should be investigated and who should be charged. The statistical model has been used in the legal system to determine who is likely to conduct crime. However, the data used to train the models is highly biased. The way to overcome this problem is to gain the oversight of the whole process and enforce

the human regulations instead of overly depending on the machine.

Impact Standpoint Considerations
It gets harder to look at the impact if we look at the different stakeholders on a different scale. The impact of Al will not only be on single individuals, but on greater global scales. The increasing of scale would also induce the interaction effect in multi-agent systems which also includes Al. The way to quantify the impact of Al on different scales is still an open problem.

Systems-level Long-term Effects
On the system level, the dynamics in the real would make a measurement of impact much harder. Over the long term, the feedback loop in the system would cause the accumulative effect: a small disparity gets amplified in the system, and things get worse and worse. How to avoid this problem is a very tricky question and we do not have a clear answer.

17.1.2 Problem Description ($9 \sim 14$)

One application of the AI system is facial analysis technology. It recently gets a lot of attention not only from industry but also from the academia for people try to automate the surveillance. However, it potentially contains bias and amplifies the bias. There are two questions based on this technology: where the bias is coming from; what implication we have.

Recently, a lot of commercial facial analysis projects are available online. By looking at the photo of faces, these facial analysis systems can give the inference such as gender, age, and etc.. However, it could be wrong with high confidence sometimes due to the bias in it.

Real World Impact
In terms of surveillance, as officers of law enforcement label someone as a potential suspect, the machine inherits the bias from the human. For example, 91% of South Wales Polices' automated facial recognition matches wrongly identified out innocent people. What's more, the machine matched female as a suspect, who indeed is a male. The machine has the real world impact even it does not make the final decision.

Gold Standard Measure
All the benchmark data set are heavily skewed to the privileged population. However, celebrity face cannot even represent the population in the united states in the sense that demographics and other certain attributes are not covered in data sets. Also, there are other problems with such data set: all the photos are taken in very good quality; some faces of women have heavy makeup. These problems in benchmark data set would cause issues when it is deployed to train the model used for surveillance purpose. The problem are how we are going to overcome those biases and what the current impact of these biases.

Intersectional Performance
When we take a look at the intersection performance, the famous issues in Gold Standard Measure cast more problem. In terms of gender, the majority of benchmark data sets is male. In addition, in terms of skin shade, the data sets contain more populations with a lighter tone. As a result, the machine trained by benchmark data sets has the highest accuracy in lighter males and lowest accuracy in darker females. Not only IBM, but almost all companies have the public data sets with lighter males as majority. Thus the accuracy is very higher with respect to lighter males, but lower with respect to minority.

What's more, the national benchmarks are not immune to this problem as well: 4.4% are darker females; 20.2% are lighter females; 59.4% are lighter males; 16% are darker males. In terms of gender, 75.4% are male and 24.6% are females.

AI Model May Amplify Social Bias
The AI model may amplify social bias such as gender, race and etc..

For example, in Google translator, the translation from gender-neutral language to gender-sensitive language is problematic. It is shown that males are highly correlated to words such as maestro, skipper, protege, philosopher, and captain, while females are highly correlated to words such as homemaker, nurse, receptionist, librarian, and socialite.

Most recently, auto-complete feature of Gmail drops the gender-based pronouns due to risk that the incorrect inference of gender would offend users.

Also, OpenAI uses the Reddit as the text source to train the new language model GPT-2. Despite the performance improved due to large quantity and the diversity of Reddit data, bias on the Reddit would also be infused in this new language model.

What's more, using the pre-trained model is a very popular trend since it improves the efficiency of training process. Nonetheless, the trained new model would inherit the bias from the pre-trained model.

Now, consider a case where people are using AI to process the resume to determine if we would hire someone. It is not the case that we have fields and structures for people to fill in, and all the resume are free text. In processing resumes, a deep learning model may make the bias worse in the sense that the model might be using the gender feature or races to determine if a person is qualified for the job. This kind of bias cannot be eliminated since we have no way to interpret the feature vectors in the model. Even if we regularize the way people present the information by creating fields and structures for people to fill in and fix the gender bias problem by decorrelating the gender words in model training, the problem still exists since other information(e.g. race, age, and appearance) can be used to infer the gender.

All of these examples show the problem of AI inheriting or amplifying bias.

17.1.3 Proposed Solution ($15 \sim 18$)

We can try to use the regulation to resolve the problem mentioned. However, we are not sure if regulation is good or not.

For example, big Bank companies such as Capital One and Bank of America are using machine learning for loan and credit approval, meanwhile they keep policy teams to scrutinize the model. However, deep learning model would be hard to have an impact on these tasks because people still tend to hand engineer all the features. Even if deep learning is a better algorithm to improve the way banks give loans and credits, it is still hard for people to make the transition due to the black box property of deep learning.

Identifying Basic Obligations for Algorithmic Decision Aids

- Primary: reflect relevant ground truth (an accuracy condition).
- Secondary: respect contextual normative constraints e.g.
  - Transparency: Can the algorithmic process be audited or reviewed?
  - Explainability: Does the algorithmic process provide an appropriate explanations for its results?
  - Privacy: Does the algorithmic process protect the privacy of impacted persons?
  - Balance of Error Burdens: how are the impacts of mistakes distributed?

Barriers to Model Accuracy in practice

- Data Missingness: Imperfect attention or surveillance practices
- Objective Ground Truth Meas.:
  - No access to evaluative data (e.g. false negatives in vetting)
  - Does not exist (job fit).
  - Poisoned historical data
- Discretion Overrides: human intervention causing otherwise perfect algorithms leading to go astray (e.g. NC, PA AFST)

In the past, people overly trusted the machine to be correct. In fact, the machine also makes mistakes but in a different way from the human. Thus the human intervention is also important in the regulation.

Currently, the self-driving system still cannot take over human on driving tasks, and we still need human on the road for a long time. For example, there is a Tesla accident where the system asks human to pay attention several seconds before the crash but people do not have the time to respond. Thus the timing for human intervention is very critical. If we still overly trust the machine, then it is meaningful to think about how to design human intervention. This requires the expertise of psychology to answer the questions: what cognitive bias and issue people have when they making decisions; especially whether people trust AI system or not is based on the how those biases going to affect the decisions of AI system. Based on those questions, we should know how to design the intervention interface. However, those are still open questions.

Algorithmic Audit
Require algorithm audits via:

- Data worksheets
- Algorithmic Impact Assessments
- Open algorithm validation tests
- Ex-Post/post-harm compensations

The Regulation Game Framework

- Explicit implementation of the regulatory framework that divides responsibility for the separate sets of competing normative goals and incentives.

- Actor vs. Regulator.
- Regulate actors output/behavior, not inputs

Industry Standards & Task-forces
Industry-led coalition on:

- Setting standards
- Accrediting model
- Punishing bad actors
- Accountable to government

Regulatory Infrastructure for Ethics
Minimal needs:

- Fluency: capable of understanding what an AI system is doing (we are far from this in deep learning case)
- Procedural Transparency: the way to train the model; the objective function imposed; the constraint used
- Avenues for (frictionless) Dissent: the way to accept the different voice
- Accountable Redress: reliable compensation for mistakes

17.2 Mathematical Formulations

How to formulate the fairness is still an open question. We are not sure which model is good. One should consult other people for the specific problem.

17.2.1 Formulation$(19 \sim 24)$

Variables definition (22)

- **X**: input feature
- **Z**: hidden property or sensitive feature e.g. gender
- **Y**: task e.g. classification, accept/deny your loan application

Categories of parity

1. Impact parity

$$P(y|z=0) = p(y|z=1)$$

   For example, z=0 represents female and z=1 represent males. The probability distribution of whether a school accepts/rejects a student should be the same for male and female.

2. Treatment parity
   The output of y depends only on x, not on z.
   The model will not use z when making decision y. For example, in some music contest, the performer is hidden from the referees to avoid appearance from creating any impressions.
   Problem: z can be hidden in other places in x. e.g. zip code reveals the house location, which is related to oneś race, income, etc.

3. Representational parity
   Map x to r(x) such that r(x) is independent of z (indistinguishable representations). This entails impact parity to train r(x).
   Problem: There is a trade-off between accuracy and decorrelation. The performance could get poorer.

4. Opportunity parity
   False positive and false negative rates match.

$$P(Y'=1|Y=y,z=0) = P(Y'=1|Y=y,z=1), y \in \{0,1\}$$

   (Hardt et al., 2016)

All parity constraints can be formulated as optimization problems. In this lecture, we focus on discussing 1. Impact parity and 2. Treaty parity.

Implementation

Disparate learning process (DLP) (23) only use z during training but not testing. Besides, z is not used as an input feature but only applied when calculating the loss to maintain a level of impact parity.

Examples (24):

- Optimization-based
  - Adding constraints ML optimization
  - Same idea but with regularizes
- Representation-based

- Probabilistic mappings to fair representations
- GAN-Based learning setups

However, do we indeed achieve better parity by imposing those constraints?

17.2.2 Limitation($25 \sim 30$)

Toy Example: Hair length and getting hired.

The variables are defined as:

- **X**: 1) length of work experience 2) hair length
- **Z**: gender (protected variable). 0:woman, 1:man
- **Y**: getting hired

In this example, we want to hide the gender variable when deciding whether to hire a person or not. However, accidentally we introduce the input feature: hair length. Since we would like to achieve impact parity: $P(y|gender = 0) = p(y|gender = 1)$ and treatment parity $P(y|gender, X) = P(y|X)$. According to Bayes theorem, we need to enforce $\frac{P(gender=0|X,Y)}{P(gender=0|X)} = \frac{P(gender=1|X,Y)}{P(gender=1|X)}$. And we know $P(gender = 0|X) \neq P(gender = 1|X)$ because men and women have different distribution of hair length. We have to adjust $P(gender|X, Y)$ manually in order to achieve parity.

The result is shown in (26)(Lipton et al., 2018). The green solid line is the decision boundary when X only contains the length of work experience. We can see that it is reasonable since we hire guys with longer work experience. But we cannot achieve impact parity since men are more likely to have longer work experience. Therefore, we apply DLP to ensure parity. The blue dashed line is the new decision boundary including the hair length feature with parity condition enforced. We can see that the decision actually is not fairer. The short-haired women are disadvantaged by DLP, as well as the men with long working experience.

Besides, from the legend of the plot, we can see that when p increases from 26% to 100%, the accuracy actually drops from 0.96 to 0.75, where the tradeoff between performance and fairness happens.

One may ask: In real life, we will never use a feature of hair length to decide whether to hair a person or not. You are correct. But as we discussed before, it is very difficult to judge whether a feature X relates to the protected feature Z. For example, zip code reveals the house location, which is related to oneś race, income, etc.

Findings (29)

1. For reconciling impact disparity and treatment disparity, treatment disparity is optimal (theoretical).
2. When x fully encodes z, for a sufficiently powerful model, DLP indistinguishable from treatment disparity (theoretical). E.g. If all men have short hair and all women have long hair, then the model can completely remove the effect of hair length because the model knows this feature directly encodes gender information.
3. When x partially encodes z, DLP results in side effects. (empirical)
   (a) Re-orders within-group based on otherwise irrelevant characteristics.
   (b) Produces a potentially bizarre incentive to conform to stereotype.

17.3 Summary $(30 \sim 32)$

Parity in machine learning is still an open question. (30)
Dynamics & equilibrium effects

- These problems consider going beyond a static view of classification.
- Must consider the impact of policies on real-world dynamics.
- Subsequent data gathered, incentives created etc..

Fairness also leads to better AI/ML (31)

1. Learning with long tail distribution. Real life is long tail, the average accuracy is not accurate for real life.
2. Reducing bias amplification
3. Incorporating model constraints
4. Understanding domain adaption
5. Designing interpretable models such as robust learning

References

Hardt, M., E. Price, and N. Srebro
2016. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413.

Lipton, Z. C., J. McAuley, and A. Chouldechova
2018. Does mitigating ml's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, Pp. 8136–8146.