Building a Robot Judge: Data Science for Decision-Making

13. AI Regulation and Policy

Q&A Page

https://bitly.com/BRJ_Padlet13

Why not use simple models?

- Last week we discussed an array of tools for interpreting the predictions of black-box machine learning models.
- Why not just use simple models?

Why not use simple models?

- Last week we discussed an array of tools for interpreting the predictions of black-box machine learning models.
- Why not just use simple models?
- Kleinberg and Mullainathan, "Simplicity Creates Inequity" (2019):
  - simple models are strictly suboptimal in terms of equity and efficiency.

Kleinberg and Mullainathan (2019)

- Individuals have characteristics X and group membership A.
- Algorithm approximates score $f(X, A)$,
  - decide outcome (e.g. admit to college) based on threshold on $f(\cdot)$

Kleinberg and Mullainathan (2019)

- Individuals have characteristics X and group membership A.
- Algorithm approximates score $f(X, A)$,
  - decide outcome (e.g. admit to college) based on threshold on $f(\cdot)$
  - A is correlated with X (a "disadvantaged" group has lower-scored attributes, e.g. less extracurricular activities) but A doesn't independently affect suitability.

Kleinberg and Mullainathan (2019)

- Individuals have characteristics X and group membership A.
- Algorithm approximates score $f(X, A)$,
  - decide outcome (e.g. admit to college) based on threshold on $f(\cdot)$
  - A is correlated with X (a "disadvantaged" group has lower-scored attributes, e.g. less extracurricular activities) but A doesn't independently affect suitability.
- A "simple model" or "approximator" partitions X into cells, and scores each cell.
  - e.g. decision tree.

Kleinberg and Mullainathan (2019)

- Individuals have characteristics X and group membership A.
- Algorithm approximates score $f(X, A)$,
  - decide outcome (e.g. admit to college) based on threshold on $f(\cdot)$
  - A is correlated with X (a "disadvantaged" group has lower-scored attributes, e.g. less extracurricular activities) but A doesn't independently affect suitability.
- A "simple model" or "approximator" partitions X into cells, and scores each cell.
  - e.g. decision tree.

Result 1:

- assume a non-trivial (e.g. real-world) dataset (see paper)

Kleinberg and Mullainathan (2019)

- Individuals have characteristics X and group membership A.
- Algorithm approximates score $f(X, A)$,
  - decide outcome (e.g. admit to college) based on threshold on $f(\cdot)$
  - A is correlated with X (a "disadvantaged" group has lower-scored attributes, e.g. less extracurricular activities) but A doesn't independently affect suitability.
- A "simple model" or "approximator" partitions X into cells, and scores each cell.
  - e.g. decision tree.

Result 1:

- assume a non-trivial (e.g. real-world) dataset (see paper)
- starting from a simple model, there exists a more complex model (smaller cells) that improves both efficiency *and* equity.
  - Efficiency = average $f(\cdot)$ of admitted candidates.
  - Equity = relative share admitted for the disadvantaged group.

Kleinberg and Mullainathan (2019)

- Individuals have characteristics X and group membership A.
- Algorithm approximates score $f(X, A)$,
  - decide outcome (e.g. admit to college) based on threshold on $f(\cdot)$
  - A is correlated with X (a "disadvantaged" group has lower-scored attributes, e.g. less extracurricular activities) but A doesn't independently affect suitability.
- A "simple model" or "approximator" partitions X into cells, and scores each cell.
  - e.g. decision tree.

Result 1:

- assume a non-trivial (e.g. real-world) dataset (see paper)
- starting from a simple model, there exists a more complex model (smaller cells) that improves both efficiency *and* equity.
  - Efficiency = average $f(\cdot)$ of admitted candidates.
  - Equity = relative share admitted for the disadvantaged group.

Result 2:

- with a simple model (relative to a complex model), info on group membership is more likely to help the decision-maker select candidates with higher $f(\cdot)$.

“An Economic Approach to Regulating Algorithms”

Rambachan, Kleinberg, Ludwig, and Mullainathan (2020)

- Apply welfare economics to the design and regulation of algorithmic decision processes.

"An Economic Approach to Regulating Algorithms"

Rambachan, Kleinberg, Ludwig, and Mullainathan (2020)

- Apply welfare economics to the design and regulation of algorithmic decision processes.
- Algorithmic decision-making has two components:
  - (1) training a prediction function, and (2) a decision rule based on the predictions.

"An Economic Approach to Regulating Algorithms"

Rambachan, Kleinberg, Ludwig, and Mullainathan (2020)

- Apply welfare economics to the design and regulation of algorithmic decision processes.
- Algorithmic decision-making has two components:
  - (1) training a prediction function, and (2) a decision rule based on the predictions.

Result 1 (social planner):

- the equity preferences of the social planner have no effect on the training procedure for the prediction function.

“An Economic Approach to Regulating Algorithms”

Rambachan, Kleinberg, Ludwig, and Mullainathan (2020)

- Apply welfare economics to the design and regulation of algorithmic decision processes.
- Algorithmic decision-making has two components:
  - (1) training a prediction function, and (2) a decision rule based on the predictions.

Result 1 (social planner):

- the equity preferences of the social planner have no effect on the training procedure for the prediction function.
- i.e., there should be no limit on the use of sensitive attributes.

Outline

Summary

- **Internal validity**: the statistical inferences about causal effects are valid for the population and setting being studied.
- **External validity**: the statistical inferences can be generalized from the population and setting studied to other populations and settings

Internal Validity (from week 3)

Linear regression model:

$$V_i = \alpha + \beta s_i + \epsilon_i$$

- Exogeneity assumption: $\text{Cov}[s_i, \epsilon_i] = 0$
  - no omitted variable bias (unobserved confounders), no joint causality

Internal Validity (from week 3)

Linear regression model:

$$V_i = \alpha + \beta s_i + \epsilon_i$$

- Exogeneity assumption: $\mathrm{Cov}[s_i, \epsilon_i] = 0$
  - no omitted variable bias (unobserved confounders), no joint causality
  - then OLS estimates for $\hat{\beta}$ converge to β in large samples.

Internal Validity (from week 3)

Linear regression model:

$$V_i = \alpha + \beta s_i + \epsilon_i$$

- Exogeneity assumption: $\text{Cov}[s_i, \epsilon_i] = 0$
  - no omitted variable bias (unobserved confounders), no joint causality
  - then OLS estimates for $\hat{\beta}$ converge to β in large samples.
- Standard errors are correct.
  - accounting for heteroskedasticity (use the "robust" option)
  - accounting for serial correlation (clustering at level of treatment)

Internal Validity (from week 3)

Linear regression model:

$$V_i = \alpha + \beta s_i + \epsilon_i$$

- Exogeneity assumption: $\text{Cov}[s_i, \epsilon_i] = 0$
  - no omitted variable bias (unobserved confounders), no joint causality
  - then OLS estimates for $\hat{\beta}$ converge to β in large samples.
- Standard errors are correct.
  - accounting for heteroskedasticity (use the "robust" option)
  - accounting for serial correlation (clustering at level of treatment)

Under these conditions, causal inferences (statistical estimates on treatment effects) are valid for the population studied.

Internal validity (ML)

- In machine learning, we gauge internal validity by proper train/test splits, and avoidance of data leakage.
- then performance metrics are valid to that population.

External Validity: Does it generalize?

- If internal validity is satisfied, then ML metrics and causal inferences are valid for the population studied.
  - e.g., the cantons that got a tax cut.

External Validity: Does it generalize?

- If internal validity is satisfied, then ML metrics and causal inferences are valid for the population studied.
  - e.g., the cantons that got a tax cut.
  - but what about the other cantons?

External Validity: Does it generalize?

- If internal validity is satisfied, then ML metrics and causal inferences are valid for the population studied.
  - e.g., the cantons that got a tax cut.
  - but what about the other cantons?
- In general **estimates/metrics are not valid for other populations**.
  - other populations are different. so treatment effects and predictions might be different.
  - e.g., medical trials are often run with men, but medicines are then used to treat both men and women.

External Validity: Does it generalize?

- If internal validity is satisfied, then ML metrics and causal inferences are valid for the population studied.
  - e.g., the cantons that got a tax cut.
  - but what about the other cantons?
- In general **estimates/metrics are not valid for other populations**.
  - other populations are different. so treatment effects and predictions might be different.
  - e.g., medical trials are often run with men, but medicines are then used to treat both men and women.
- External validity is an issue for both causal inference and machine learning.

Outline

- Algorithms influence various aspects of life:
    - selecting tax payers for audits
    - granting or denying immigration visas
    - security screening at airports
- Besides benefits, can have risks and harms.
- Public interest requires governance to reinforce benefits and minimize risks.

Benefits

- Efficiency, accuracy, scalability
- Algorithms can be a boon to due process
  - Consistent decision making
  - Making bias evident
- Growing digital economy

Principles and Objectives

Principles

- Justice, equality, non-discrimination
- Privacy, surveillance
- Safety and reliability

Principles and Objectives

Principles

- Justice, equality, non-discrimination
- Privacy, surveillance
- Safety and reliability

Objectives

- Accuracy
- Equity
- Explainability

Principles and Objectives

Principles

- Justice, equality, non-discrimination
- Privacy, surveillance
- Safety and reliability

Objectives

- Accuracy
- Equity
- Explainability
- Auditability, transparency
- Responsibility, accountability

Challenges to developing standards

- Collective decision processes
  - tradeoffs among various stakeholders
  - distortions from lobbying
  - technical issues $\rightarrow$ politicians and voters have low information

Challenges to developing standards

- Collective decision processes
  - tradeoffs among various stakeholders
  - distortions from lobbying
  - technical issues $\rightarrow$ politicians and voters have low information
- Global coordination needed for digital tech
  - accounting for different cultures and contexts

Challenges to developing standards

- Collective decision processes
  - tradeoffs among various stakeholders
  - distortions from lobbying
  - technical issues $\rightarrow$ politicians and voters have low information
- Global coordination needed for digital tech
  - accounting for different cultures and contexts
- How to assign responsibility for risks/harms
  - creator / owner / operator/ user?
  - how to understand / determine intentions
  - balance accountability with innovation and growth

Governance Strategies

- Industry-driven approach;
  - Reduces regulatory red tape, could help innovation
  - No central authority to enforce best-practices;
  - Expands the power of large corporations.
  - Significant externalities, tendency to concentration

Governance Strategies

- Industry-driven approach;
  - Reduces regulatory red tape, could help innovation
  - No central authority to enforce best-practices;
  - Expands the power of large corporations.
  - Significant externalities, tendency to concentration
- Regulator-driven approach:
  - significant technical knowledge/skills needed to be effective
  - bad actors always a step ahead.
  - limits innovation and expansion of digital economy.
  - could collude with industry leaders

Transparency

- Closed-source algorithms result in "black box justice" and could be abused by insiders.
- But open-source algorithms are prone to gaming: savvy attorneys could "trick" the algorithm.

Transparency

- Closed-source algorithms result in “black box justice” and could be abused by insiders.
- But open-source algorithms are prone to gaming: savvy attorneys could “trick” the algorithm.
- Understanding the code/model not the same as understanding behavior
  - ML processes not understandable by non-experts
  - Sometimes even experts don’t understand the model

Enforcement

- How can we make sure that the decision maker is not merely claiming to follow the rules?
  - Disclose the code?
  - Disclose the logs?
- Idea:
  - technical tools for verifying correctness
  - ensure that appropriate evidence exists for later oversight.
  - can be decentralized on blockchain

"An Economic Approach to Regulating Algorithms"

Rambachan, Kleinberg, Ludwig, and Mullainathan (2020)

- Apply welfare economics to the design and regulation of algorithmic decision processes.
- Algorithmic decision-making has two components:
  - (1) training a prediction function, and (2) a decision rule based on the predictions.

Result 1 (social planner):

- the equity preferences of the social planner have no effect on the training procedure for the prediction function.
- i.e., there should be no limit on the use of sensitive attributes.

"An Economic Approach to Regulating Algorithms"

Rambachan, Kleinberg, Ludwig, and Mullainathan (2020)

- Apply welfare economics to the design and regulation of algorithmic decision processes.
- Algorithmic decision-making has two components:
  - (1) training a prediction function, and (2) a decision rule based on the predictions.

Result 1 (social planner):

- the equity preferences of the social planner have no effect on the training procedure for the prediction function.
- i.e., there should be no limit on the use of sensitive attributes.

Result 2 (private actors):

- key factor is disclosure of decision process (data, ML training, and decision rule), which, unlike human decision-making, allows prejudicial treatment to be detected.

"An Economic Approach to Regulating Algorithms"

Rambachan, Kleinberg, Ludwig, and Mullainathan (2020)

- Apply welfare economics to the design and regulation of algorithmic decision processes.
- Algorithmic decision-making has two components:
  - (1) training a prediction function, and (2) a decision rule based on the predictions.

Result 1 (social planner):

- the equity preferences of the social planner have no effect on the training procedure for the prediction function.
- i.e., there should be no limit on the use of sensitive attributes.

Result 2 (private actors):

- key factor is disclosure of decision process (data, ML training, and decision rule), which, unlike human decision-making, allows prejudicial treatment to be detected.
- without disclosure, algorithms will be just as biased as humans.

"An Economic Approach to Regulating Algorithms"

Rambachan, Kleinberg, Ludwig, and Mullainathan (2020)

- Apply welfare economics to the design and regulation of algorithmic decision processes.
- Algorithmic decision-making has two components:
  - (1) training a prediction function, and (2) a decision rule based on the predictions.

Result 1 (social planner):

- the equity preferences of the social planner have no effect on the training procedure for the prediction function.
- i.e., there should be no limit on the use of sensitive attributes.

Result 2 (private actors):

- key factor is disclosure of decision process (data, ML training, and decision rule), which, unlike human decision-making, allows prejudicial treatment to be detected.
- without disclosure, algorithms will be just as biased as humans.
- with disclosure, discrimination decreases relative to humans, and government should impose no constraints on the use of sensitive attributes as predictors.

"An Economic Approach to Regulating Algorithms"

Rambachan, Kleinberg, Ludwig, and Mullainathan (2020)

- Apply welfare economics to the design and regulation of algorithmic decision processes.
- Algorithmic decision-making has two components:
  - (1) training a prediction function, and (2) a decision rule based on the predictions.

Result 1 (social planner):

- the equity preferences of the social planner have no effect on the training procedure for the prediction function.
- i.e., there should be no limit on the use of sensitive attributes.

Result 2 (private actors):

- key factor is disclosure of decision process (data, ML training, and decision rule), which, unlike human decision-making, allows prejudicial treatment to be detected.
- without disclosure, algorithms will be just as biased as humans.
- with disclosure, discrimination decreases relative to humans, and government should impose no constraints on the use of sensitive attributes as predictors.
  - caveat: disclosure must include the data and ML training process, not just the decision rule.

Outline

Incentive Responses

- Decisions today change features tomorrow.
- Take the case of ML-based credit scoring.
- Some strategic responses are benign/helpful:
  - e.g., pay back existing debts to improve score

Incentive Responses

- Decisions today change features tomorrow.
- Take the case of ML-based credit scoring.
- Some strategic responses are benign/helpful:
  - e.g., pay back existing debts to improve score
- Others could be costly manipulation
  - e.g., open more credit accounts to increase score, but at some risk
  - more generally, ML subjects can pay some cost and manipulate their features to improve their predicted label.

Incentive Responses

- Decisions today change features tomorrow.
- Take the case of ML-based credit scoring.
- Some strategic responses are benign/helpful:
  - e.g., pay back existing debts to improve score
- Others could be costly manipulation
  - e.g., open more credit accounts to increase score, but at some risk
  - more generally, ML subjects can pay some cost and manipulate their features to improve their predicted label.
- Milli et al, "The Social Cost of Strategic Classification" (2019)
  - model sequential decision of modeler and subject as Stackelberg Competition, a classic model from game theory on the interaction between duopolists.

Strategic Classification (Milli et al 2019)

- Each individual has features X and a label $Y \in \{0,1\}$.
- Institution gets utility from a classifier $f : X \to Y$ equal to $\Pr(f(X) = Y)$.
- Individual utility gets utility when $\hat{Y} = 1$ and can change to features X' at cost $c(X, X')$. So
$$u(X'; X) = f(X') - c(X, X')$$

Strategic Classification (Milli et al 2019)

- Each individual has features X and a label $Y \in \{0,1\}$.
- Institution gets utility from a classifier $f : X \to Y$ equal to $\Pr(f(X) = Y)$.
- Individual utility gets utility when $\hat{Y} = 1$ and can change to features X' at cost $c(X,X')$. So
$$u(X';X) = f(X') - c(X,X')$$
- The individual reports
$$\Delta(X) = \arg\max_{X'} u(X';X)$$
- and therefore the institution's objective is
$$\max_{f(\cdot)} \Pr(f(\Delta(X)) = Y).$$

Strategic Classification (Milli et al 2019)

- Each individual has features X and a label $Y \in \{0,1\}$.
- Institution gets utility from a classifier $f : X \to Y$ equal to $\Pr(f(X) = Y)$.
- Individual utility gets utility when $\hat{Y} = 1$ and can change to features X' at cost $c(X, X')$. So
$$u(X'; X) = f(X') - c(X, X')$$
- The individual reports
$$\Delta(X) = \arg\max_{X'} u(X'; X)$$
- and therefore the institution's objective is
$$\max_{f(\cdot)} \Pr(f(\Delta(X)) = Y).$$
- Equilibrium:
  - features x_j that are costly to change (high $\frac{\partial c}{\partial x_j}$) will be used by the designer. features that are less costly to change will not be used.
  - in strategic context, designer chooses overall more conservative decision threshold.

Strategic Classification (Milli et al 2019)

- Each individual has features X and a label $Y \in \{0,1\}$.
- Institution gets utility from a classifier $f: X \to Y$ equal to $\Pr(f(X) = Y)$.
- Individual utility gets utility when $\hat{Y} = 1$ and can change to features X' at cost $c(X, X')$. So
$$u(X'; X) = f(X') - c(X, X')$$
- The individual reports
$$\Delta(X) = \arg\max_{X'} u(X'; X)$$
- and therefore the institution's objective is
$$\max_{f(\cdot)} \Pr(f(\Delta(X)) = Y).$$
- Equilibrium:
  - features x_j that are costly to change (high $\frac{\partial c}{\partial x_j}$) will be used by the designer. features that are less costly to change will not be used.
  - in strategic context, designer chooses overall more conservative decision threshold.
- Social costs:
  - the costs $c(\cdot)$ are socially wasteful, but responses to manipulation increase them.
  - $c(\cdot)$ could be different across groups, causing inequity

Outline

Internal vs External Validity

AI Governance

Incentive Responses to AI Decisions

What can and should AI decide?

AI for legal decisions

Recap and Conclusion

Perception Tasks

- Content identification (Shazam, reverse image search)
- Face recognition
- Medical diagnosis from scans
- Speech to text
- Deepfakes

Perception Tasks

- Content identification (Shazam, reverse image search)
- Face recognition
- Medical diagnosis from scans
- Speech to text
- Deepfakes

High accuracy causes risk of privacy violations.

Perception Tasks

- Content identification (Shazam, reverse image search)
- Face recognition
- Medical diagnosis from scans
- Speech to text
- Deepfakes

High accuracy causes risk of privacy violations.

Systems are sometimes more accurate/effective for some groups, e.g. most-frequent customers.

Perception Tasks

- Content identification (Shazam, reverse image search)
- Face recognition
- Medical diagnosis from scans
- Speech to text
- Deepfakes

High accuracy causes risk of privacy violations.

Systems are sometimes more accurate/effective for some groups, e.g. most-frequent customers.

Overall, problems seem straightforward to solve.

Human Judgment Annotation Tasks

- Spam detection
- Detection of copyrighted material
- Automated essay grading
- Hate speech detection
- Content recommendation

Human Judgment Annotation Tasks

- Spam detection
- Detection of copyrighted material
- Automated essay grading
- Hate speech detection
- Content recommendation

These tasks are subjective, so some error is inevitable.
But human judgments are correlated enough that predictions are useful.

Human Judgment Annotation Tasks

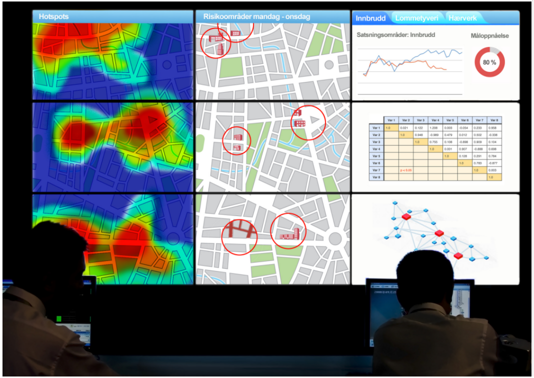
- Spam detection
- Detection of copyrighted material
- Automated essay grading
- Hate speech detection
- Content recommendation

These tasks are subjective, so some error is inevitable.
But human judgments are correlated enough that predictions are useful.

Labels are past behavior, so model is stable and incentive responses are constrained.

- compare: predicting how someone will score on these predictions in the future.

Predictive Policing



Predictive policing poses discrimination risk, thinktank warns

Machine-learning algorithms could replicate or amplify bias on race, sexuality and age

▲ One officer said human biases including more stop and searches of black men were likely to be introduced into algorithm data sets. Photograph: Carl Court/Getty Images https://www.theregister.com/2020/12/08/texas_compsci_phd_ai/

{* ARTIFICIAL INTELLIGENCE *}

Uni revealed it killed off its PhD-applicant screening AI – just as its inventors gave a lecture about the tech

Fears of bias put compsci dept into damage-limitation mode after years of using it to analyze applications

Katyanna Quach Tue 8 Dec 2020 // 12:04 UTC SHARE

A university announced it had ditched its machine-learning tool, used to filter thousands of PhD applications, right as the software's creators were giving a talk about the code and drawing public criticism.

// MOST READ

Apple fires warning shot at Facebook and Google on privacy, pledges fight

Predicting future choices and social outcomes

- Predicting criminal recidivism
- Predictive policing
- Predicting terrorist risk
- Predicting at-risk kids

Predicting future choices and social outcomes

- Predicting criminal recidivism
- Predictive policing
- Predicting terrorist risk
- Predicting at-risk kids

These systems are risky and can have unintended consequences.

Predicting future choices and social outcomes

- Predicting criminal recidivism
- Predictive policing
- Predicting terrorist risk
- Predicting at-risk kids

These systems are risky and can have unintended consequences.

Predictions influence availability of labels and subsequent behavior.

Predicting future choices and social outcomes

- Predicting criminal recidivism
- Predictive policing
- Predicting terrorist risk
- Predicting at-risk kids

These systems are risky and can have unintended consequences.

Predictions influence availability of labels and subsequent behavior.

Outcomes are in future so models lack external validity.

Predicting future choices and social outcomes

- Predicting criminal recidivism
- Predictive policing
- Predicting terrorist risk
- Predicting at-risk kids

These systems are risky and can have unintended consequences.

Predictions influence availability of labels and subsequent behavior.

Outcomes are in future so models lack external validity.

Errors are costly. Strong incentive responses.

Overview of problems relevant to ML fairness

- Accuracy issues:
  - model stability
  - selective labeling

Overview of problems relevant to ML fairness

- Accuracy issues:
  - model stability
  - selective labeling
- Equity issues:
  - (relative) error rate
  - (relative) costs of errors

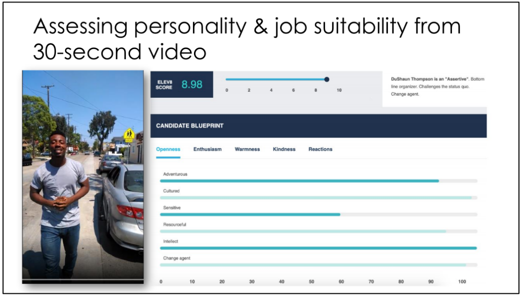
Overview of problems relevant to ML fairness

- Accuracy issues:
  - model stability
  - selective labeling
- Equity issues:
  - (relative) error rate
  - (relative) costs of errors
- Social problems from introducing system:
  - externalities (e.g. privacy violations)
  - asymmetric information (AI company knows your preferences (price point) $\rightarrow$ they have information advantage and can capture more surplus).

Overview of problems relevant to ML fairness

- Accuracy issues:
  - model stability
  - selective labeling
- Equity issues:
  - (relative) error rate
  - (relative) costs of errors
- Social problems from introducing system:
  - externalities (e.g. privacy violations)
  - asymmetric information (AI company knows your preferences (price point)$\rightarrow$ they have information advantage and can capture more surplus).
- Incentive responses:
  - subjects try to manipulate features to game system
  - systems (e.g. essay grading) perceived as biased/unfair are discouraging.

Why don't algorithmic hiring systems work? (Raghavan et al, 2019)



Zoom private chat: Identify a problem with algorithmic hiring, explain why, and what would have to change to fix that problem.

Additional issues with using AI for predicting social outcomes

Narayanan slides

- Hunger for personal data
- Transfer of power from domain experts & workers to unaccountable tech companies
- Veneer of objectivity
- Lack of explainability
- ...

Outline

Internal vs External Validity

AI Governance

Incentive Responses to AI Decisions

What can and should AI decide?

AI for legal decisions

Recap and Conclusion

What about legal decisions?

- So far, in the legal context, we have focused mainly on parole and bail decisions.
    - there aren't many papers/systems out there for determining "guilty" vs "innocent"

What about legal decisions?

- So far, in the legal context, we have focused mainly on parole and bail decisions.
  - there aren't many papers/systems out there for determining "guilty" vs "innocent"
- There are two main reasons for this:
  - there is a measurable/"true" label that we can predict: whether someone is arrested again in some period of time.
  - the factors that judges are supposed to use are also measured: factors that predict recidivism.

What about legal decisions?

- So far, in the legal context, we have focused mainly on parole and bail decisions.
  - there aren't many papers/systems out there for determining "guilty" vs "innocent"
- There are two main reasons for this:
  - there is a measurable/"true" label that we can predict: whether someone is arrested again in some period of time.
  - the factors that judges are supposed to use are also measured: factors that predict recidivism.
- In contrast, for the liability decision (guilty or not):
  - the label is not observed directly, we just have a human judge's decision to go on.
  - the factors are part of a specific circumstance, and not part of a standard data set.

What can legal AI achieve?

- Perception tasks:
  - speeding cameras
  - gunshot detection
  - facial recognition for fare dodging / trespassing

What can legal AI achieve?

- Perception tasks:
  - speeding cameras
  - gunshot detection
  - facial recognition for fare dodging / trespassing
- Human judgement annotation on structured data:
  - copyright infringement
  - detecting corruption in budget accounts
  - detecting evasion in income / tax accounts

What can legal AI achieve?

- Perception tasks:
  - speeding cameras
  - gunshot detection
  - facial recognition for fare dodging / trespassing
- Human judgement annotation on structured data:
  - copyright infringement
  - detecting corruption in budget accounts
  - detecting evasion in income / tax accounts
- Human judgment annotation on unstructured data?
  - determining liability from trial documents
  - e.g. affidavits, police reports, witness testimony

What can legal AI achieve?

- Perception tasks:
  - speeding cameras
  - gunshot detection
  - facial recognition for fare dodging / trespassing
- Human judgement annotation on structured data:
  - copyright infringement
  - detecting corruption in budget accounts
  - detecting evasion in income / tax accounts
- Human judgment annotation on unstructured data?
  - determining liability from trial documents
  - e.g. affidavits, police reports, witness testimony

  $\uparrow$ *would require a lot of (sophisticated) NLP tools*

What can legal AI not achieve

What can legal AI not achieve

- Algorithm has severe evidence constraints:
  - can only use evidence that appears in a lot of cases; it ignores special/mitigating circumstances.
  - cannot (easily) contextualize evidence that is more or less trustworthy

What can legal AI not achieve

- Algorithm has severe evidence constraints:
  - can only use evidence that appears in a lot of cases; it ignores special/mitigating circumstances.
  - cannot (easily) contextualize evidence that is more or less trustworthy
- Would not work in many types of cases:
  - murder cases, where only evidence is witness testimony
  - antitrust violations
  - tax avoidance through accounting tricks

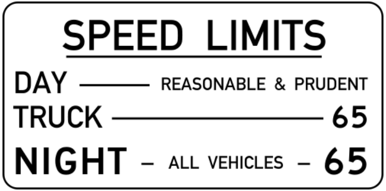
What can legal AI not achieve

- Algorithm has severe evidence constraints:
  - can only use evidence that appears in a lot of cases; it ignores special/mitigating circumstances.
  - cannot (easily) contextualize evidence that is more or less trustworthy
- Would not work in many types of cases:
  - murder cases, where only evidence is witness testimony
  - antitrust violations
  - tax avoidance through accounting tricks
- Would not work on new types of cases.
  - In particular, would not account for new laws/legislation.

What can legal AI not achieve

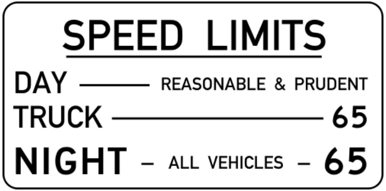
- Algorithm has severe evidence constraints:
  - can only use evidence that appears in a lot of cases; it ignores special/mitigating circumstances.
  - cannot (easily) contextualize evidence that is more or less trustworthy
- Would not work in many types of cases:
  - murder cases, where only evidence is witness testimony
  - antitrust violations
  - tax avoidance through accounting tricks
- Would not work on new types of cases.
  - In particular, would not account for new laws/legislation.
- Teaching the algorithm to understand rare evidence, discount suspicious evidence, and to understand new laws, would require something much closer to **legal artificial intelligence**.

Legal Vagueness and Value Judgments



- Even if the AI could read new laws, there is the problem of legal vagueness:
  - How will the AI decide in this circumstance?

Legal Vagueness and Value Judgments



- Even if the AI could read new laws, there is the problem of legal vagueness:
  - How will the AI decide in this circumstance?

- Making choices in the presence of vagueness or indeterminacy requires value judgements.

What counts as a "good" outcome? Is it even measurable?

Philosophical Issues

- What does it mean to surrender the implementation of law enforcement and judicial decision making to machines?
- What are the long-term implications for the system and its adaptiveness to change?
  - what are the political and cultural impacts?
  - how does it affect motivation to appeal?

Outline

Building a Robot Judge

- This course has focused on **machine learning** and **causal inference** for **decision-making**.
  - **expert** decision-making requiring **judgment** – not just legal but also medical, political, etc.

Building a Robot Judge

- This course has focused on **machine learning** and **causal inference** for **decision-making**.
  - **expert** decision-making requiring **judgment** – not just legal but also medical, political, etc.
- Engineering goals:
  - Develop tools for "building a robot judge" – machine prediction and support of expert decisions.

Building a Robot Judge

- This course has focused on **machine learning** and **causal inference** for **decision-making**.
  - **expert** decision-making requiring **judgment** – not just legal but also medical, political, etc.
- Engineering goals:
  - Develop tools for "building a robot judge" – machine prediction and support of expert decisions.
- Scientific goals:
  - Understand the factors underlying decisions of judges.

Building a Robot Judge

- This course has focused on **machine learning** and **causal inference** for **decision-making**.
  - **expert** decision-making requiring **judgment** – not just legal but also medical, political, etc.
- Engineering goals:
  - Develop tools for "building a robot judge" – machine prediction and support of expert decisions.
- Scientific goals:
  - Understand the factors underlying decisions of judges.
  - Assess the real-world impacts of decisions on society – e.g. defendants, patients.

Learning objectives

Learning objectives

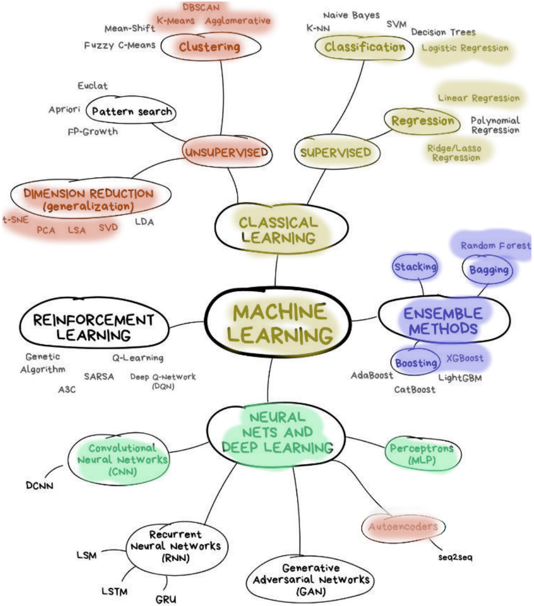
1. **Implement and evaluate machine learning pipelines.**

Learning objectives

1. **Implement and evaluate machine learning pipelines.**
2. **Implement and evaluate causal inference designs.**

Learning objectives

1. **Implement and evaluate machine learning pipelines.**
2. **Implement and evaluate causal inference designs.**
3. **Understand how (not) to use data science tools (ML and CI) to support expert decision-making.**


MACHINE LEARNING
CLASSICAL LEARNING
UNSUPERVISED
SUPERVISED
Clustering
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Pattern search
Euclat
Apriori
FP-Growth
DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA
Classification
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Regression
Linear Regression
Polynomial Regression
Ridge/Lasso Regression
ENSEMBLE METHODS
Stacking
Bagging
Random Forest
Boosting
XGBoost
AdaBoost
LightGBM
CatBoost
REINFORCEMENT LEARNING
Genetic Algorithm
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C
NEURAL NETS AND DEEP LEARNING
Convolutional Neural Networks (CNN)
DCNN
Perceptrons (MLP)
Autoencoders
seq2seq
Recurrent Neural Networks (RNN)
LSM
LSTM
GRU
Generative Adversarial Networks (GAN)

Review Questions

https://bit.ly/BRJ-A13-Qs

Exam

- I will provide more detail in the coming weeks, and we will have a review session in early January.
- Please post questions here and we will try to answer them regularly, or post links to answers:

`https://padlet.com/eash44/1gunge0ijdx2bc0c`

Next Term: NLP Course

- In the spring term, I teach a complementary course in natural language processing:
  - "Sequencing Legal DNA: NLP for Law and Political Economy" (851-0739-01L)

Next Term: NLP Course

- In the spring term, I teach a complementary course in natural language processing:
  - "Sequencing Legal DNA: NLP for Law and Political Economy" (851-0739-01L)
- Not a lot of overlap, and in many ways it builds on the content in this course.
  - i.e., focus on sequence data, and on transformer architectures (e.g. BERT, GPT-3)
- Similar setup in terms of course credits:
  - 3 credits for the lectures/assignments, 2 additional credits for a project.

Stay in touch

- e.g. add me on LinkedIn
- let me know if anything in this course helps you later on!
- can provide references for your work in the course.

Stay in touch

- e.g. add me on LinkedIn
- let me know if anything in this course helps you later on!
- can provide references for your work in the course.

Thanks!