Human Decisions and Machine Predictions

Jure Leskovec (@jure)

Including joint work with Himabindu Lakkaraju, Sendhil Mullainathan, Jon Kleinberg, Jens Ludwig
Machine Learning

Humans vs. Machines
Humans vs. Machines

- Given the same data/features X, machines tend to beat human:
  - Games: Chess, AlphaGo, Poker
  - Image classification
  - Question answering (IBM Watson)

- But humans tend to see more than machines do. Humans make decisions based on (X,Z)!
Humans Make Decisions

Humans make decisions based on $X, Z$!

- Machine is trained based on $P(Y|X)$
- But humans use $P(Y|X, Z)$ to make decisions

Could this be a problem? Yes! Why? Because the data is selectively labeled and cross validation can severely overestimate model performance.
Plan for the Talk

1) Can machines make better decisions than humans?

2) Why do humans make mistakes?

3) Can machines help humans make better decisions?
Example Problem: Criminal Court Bail Assignment
Example Decision Problem

U.S. police makes >12M arrests/year

Q: Where do people wait for trial?

Release vs. Detain is high stakes:

- Pre-trial detention spells avg. 2-3 months (can be up to 9-12 months)
- Nearly 750,000 people in jails in US
- Consequential for jobs, families as well as crime
Judge’s Decision Problem

Judge must decide whether to release (bail) or not (jail)

Outcomes: Defendant when out on bail can behave badly:

- Fail to appear at trial
- Commit a non-violent crime
- Commit a violent crime

The judge is making a prediction!
Bail as a Decision Problem

- Not assessing guilt on this crime
- Not a punishment
- Judge can only pay attention to failure to appear and crime risk
The ML Task

We want to build a predictor that will based on defendant’s criminal history predict defendant’s future bad behavior.
What Characteristics to Use?

Only administrative features
- Common across regions
- Easy to get
- No “gaming the system” problem

Only features that are legal to use
- No race, gender, religion

Note: Judge predictions may not obey either of these
Data: Defendant Features

- Age at first arrest
- Times sentenced residential correction
- Level of charge
- Number of active warrants
- Number of misdemeanor cases
- Number of past revocations
- Current charge domestic violence
- Is first arrest
- Prior jail sentence
- Prior prison sentence
- Employed at first arrest
- Currently on supervision
- Had previous revocation
- Arrest for new offense while on supervision or bond
- Has active warrant
- Has active misdemeanor warrant
- Has other pending charge
- Had previous adult conviction
- Had previous adult misdemeanor conviction
- Had previous adult felony conviction
- Had previous Failure to Appear
- Prior supervision within 10 years

(only legal features and easy to get)
# Data: Kentucky & Federal

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Number of cases</th>
<th>Fraction released people</th>
<th>Fraction of Crime</th>
<th>Failure to Appear at Trial</th>
<th>Non-violent Crime</th>
<th>Violent Crime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kentucky</td>
<td>362k</td>
<td>73%</td>
<td>17%</td>
<td>10%</td>
<td>4.2%</td>
<td>2.8%</td>
</tr>
<tr>
<td>Federal Pretrial System</td>
<td>1.1m</td>
<td>78%</td>
<td>19%</td>
<td>12%</td>
<td>5.4%</td>
<td>1.9%</td>
</tr>
</tbody>
</table>

Jure Leskovec (@jure) Stanford University
Applying Machine Learning

- Just apply Machine Learning:
  - Use labeled dataset (released defendants)
  - Train a learning model to predict whether defendant when on bail will misbehave
  - Report accuracy on the held-out test set
Prediction Model

But there is a problem with this…
The Problem

- **Issue:** Judge sees factors the machine does not
  - Judge makes decisions based on \( P(Y|X,Z) \)
    - \( X \) … prior crime history
    - \( Z \) … other factors not seen by the machine
  - Machine makes decisions based on \( P(Y|X) \)
- **Problem:** Judge is selectively labeling the data based on \( P(Y|X,Z) \)!
Problem: Selective Labels

Judge is selectively labeling the dataset

- We can only train on released people
- By jailing judge is selectively hiding labels!
Selection on Unobservables

Selective labels introduce bias:

- Say young people with no tattoos have no risk for crime. Judge releases them.
- But machine does not observe whether defendants have tattoos.
- So, the machine would falsely conclude that all young people do no crime.
- And falsely presume that by releasing all young people it does better than judge!
Challenges:
(1) Selective Labels
(2) Unobservables Z
Solution: 3 Properties

Our solution depends on three key properties:

- Multiple judges
- Natural variability between judges
- Random assignment of cases to judges
- One-sidedness of the problem
Solution: Tree Observations

1) Problem is one-sided:
   - Releasing a jailed person is a problem
   - Jailing a released person is not
Solution: Three Observations

2) In Federal system cases are randomly assigned

- This means that on average all judges have similar cases
Solution: Three Observations

3) Natural variability between judges

- Due to human nature there will be some variability between the judges
- Some judges are more strict while others are more lenient
Solution: Contraction

Solution: Use most lenient judges!

- Take released population of a lenient judge
- Which of those would we jail to become less lenient
- Compare crime rate to a strict human judge

Note: Does not rely on judges having “similar” predictions
Solution: Contraction

Making lenient judges strict:

- Defendants (ordered by crime probability)
  - Released by a lenient judge
  - Use algorithm to make a lenient judge more strict
  - Strict human judge

What makes contraction work:

- 1) Judges have similar cases (random assignment of cases)
- 2) Selective labels are one-sided
Evaluating the Prediction

Predictions create a new release rule:

- Parameterized by % released
  - For every defendant predict $P(\text{crime})$
  - Sort them by increasing $P(\text{crime})$ and “release” bottom $k$
- Note: This is different than AUC

What is the fraction released vs. crime rate tradeoff?
Holding the release rate constant, the crime rate is reduced by 0.102 (58.45% decrease).

Release Rate vs. Overall Crime Rate

- Judges: 17%
- Machine Learning: 11%

Decision trees beat LogReg by 30% contrast with AUC ROC.

Standard errors too small to display.
Predicted Failure to Appear

Holding release rate constant, crime rate is reduced by 0.06 (60% decrease)

FTA Rate

Release Rate

Judges decisions 10%

ML algorithm 4.1%

73%
Predicted Non-Violent Crime

Holding release rate constant, crime rate is reduced by 0.025 (58% decrease)

Judges decisions
- 4.2%
- 1.7%

Notes:	Standard errors	too	small	display	on	graph
Predicted Violent Crime

Holding release rate constant, crime rate is reduced by 0.014 (50% decrease)

Judges decisions
2.8%

1.4%

Notes: Standard errors too small to display on graph
Effect of Race

<table>
<thead>
<tr>
<th>Release Rule</th>
<th>Crime Rate</th>
<th>Drop Relative to Judge</th>
<th>Percentage of Jail Population</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Black</td>
</tr>
<tr>
<td>Distribution of Defendants (Base Rate)</td>
<td>.1134</td>
<td>0%</td>
<td>.4877</td>
</tr>
<tr>
<td>Judge</td>
<td>.1134</td>
<td>0%</td>
<td>.573</td>
</tr>
<tr>
<td></td>
<td>(.0010)</td>
<td>(.0029)</td>
<td>(.0027)</td>
</tr>
<tr>
<td>Algorithm</td>
<td></td>
<td></td>
<td>Black</td>
</tr>
<tr>
<td>Usual Ranking</td>
<td>.0854</td>
<td>-24.68%</td>
<td>.5984</td>
</tr>
<tr>
<td></td>
<td>(.0008)</td>
<td>(.0029)</td>
<td>(.0027)</td>
</tr>
<tr>
<td>Match Judge on Race</td>
<td>.0855</td>
<td>-24.64%</td>
<td>.573</td>
</tr>
<tr>
<td></td>
<td>(.0008)</td>
<td>(.0029)</td>
<td>(.0027)</td>
</tr>
<tr>
<td>Equal Release Rates for all Races</td>
<td>.0873</td>
<td>-23.02%</td>
<td>.4877</td>
</tr>
<tr>
<td></td>
<td>(.0008)</td>
<td>(.0029)</td>
<td>(.0028)</td>
</tr>
</tbody>
</table>

ML does not beat the judge by racially discriminating
Source of Errors

Judges err on riskiest defendants
Predicted Risk vs. Jailing

Least Likely to Jail by judge

<table>
<thead>
<tr>
<th>Predicted risk</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Most Likely to Jail</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Predicted risk vs. density
Why do Judges do Poorly?

Odds are against algorithm: Judge sees many things the algorithm does not:

- Offender history:
  Observable by both

- Demeanor:
  - “I looked in his eyes and saw his soul”
  - Unobservable by the algorithm

Can we diagnose judges’ mistakes and help them make better decisions?
Why do Judges do Poorly?

Two possible reasons why judges are making mistakes:

- Misuse of observable features
  - These are the administrative features available to the algorithm
    - E.g.: Previous crime history, etc.

- Misuse of unobservable features
  - Features not seen by the algorithm
    - “I looked in his eyes and saw his soul”
A Different Test

Predict judge’s decision
- Training data does NOT include whether defendant actually commits a crime

Gives a release rule
Release if predicted judge would release
- This weights features the way judge does

What does it do differently then?
- Does not use the unobservables!
Predicting the Judge

Build a model to predict judge’s behavior

- We predict what the judge will do (and NOT if a defendant will commit a crime)

Mix judge and the model:

\[(1 - \gamma) \text{Judge's real decision} + \gamma \text{ Predicted decision}\]

Does this beats the judge (for some \(\gamma\))?

- Model will do better only if the judge misweighs unobservable features
Artificial Judge Beats the Judge

Model of the judge beats the judge by 35%
Putting It All Together
Recap so far…

Algorithm beats human judge in the bail assignment problem

Artificial judge beats the real judge
   This means judge is misweighing unobservable features

Can we model human mistakes?
Judge j

Decisions of j

- Release
- Release
- Jail
- Jail

Items i

- Crime
- No crime
- Crime
- No crime

True labels
Modeling Human Decisions

Judge j

Decision of judge j on defendant i

Defendant i

True label \( t_i \)

Decision

\[
\begin{array}{cc|c}
\text{Truth} & \text{Decision} & \text{Probability} \\
\hline
p_{c,c} & p_{c,n} & \text{Probability that j decided “no crime” to a defendant who will do “crime”}
\end{array}
\]
Problem Setting

Input:
- Judges and defendants described by features

Output:
- Discover groups of judges and defendants that share common confusion matrices
Joint Confusion Model

Judge \( j \)

Cluster \( c_j \)

Judge attributes

\( a_1, a_2, a_3, a_4, a_5 \)

Decision of judge \( j \) on item \( i \)

Confusion matrix

Truth

Decision

Similar judges share confusion matrices when deciding on similar items
What do we learn on Bail?

Judge attributes:
- Year, County, Number of Previous Cases, Number of Previous Felony Cases, Number of Previous Misd. Cases, Number of Previous Minor Cases

Defendant attributes:
- Previous Crime History, Personal Attributes (Children/Married etc.), Social Status (Education/House Own/Moved a lot in past 10 years), Current Offense Details
Why judges make mistakes

Less experienced judges make more mistakes

- Left: Judges with high number of cases
- Right: Judges with low number of cases
Why judges make mistakes

Judges with many felony cases are stricter

- Left: Judges with many felony cases
- Right: Judges with few felony cases
Why judges make mistakes

Defendants who are single, did felonies, and moved a lot are accurately judged

Defendants who have kids are confusing to judges
Conclusion

New tool to understand human decision making

The framework can be applied to various other questions:

- Decisions about patient treatments
- Assigning students to intervention programs
- Hiring and promotion decisions
Power of Algorithms

Algorithms can help us understand if human judges are biased

Algorithms can be used to diagnose reason for bias

Key insight

- More than prediction tools
- Serve as behavioral diagnostic tools
Focusing on Decisions

Not just about prediction
Key is starting with decision:

- Performance benchmark: Current “human” decisions
  - Not ROC but “human ROC”
- What are we really optimizing?

Focus on decision also raises new concerns
Reflections: New Concerns

1) Selective labels

- We don’t see labels of people that are jailed
- Extremely common problem: Prediction -> Decision -> Action
- Which outcomes we see depends on our decisions
Reflections: New Concerns

2) Data responds to prediction

- Whenever we make a prediction/decision we affect the labels/data we see in the future
- Creates a self-reinforcing feedback loop
3) Constraints on input features:

For example, race is an illegal feature
- Our models don’t use it

But, is that enough?
- Many other characteristics correlate with race

How do we deal with this additional constraint?
Reflections: New Concerns

4) Omitted payoff bias

- Is minimizing the crime rate really the right goal?
- There are other important factors
  - Consequences of jailing on the family
  - Jobs and the workplace
  - Future behavior of the defendant
- How could we model these?
Conclusion

Bail is not the only prediction problem

Many important problems in public policy are predictive problems!

Potential for large impact
References


