Predictive Algorithms

Widely used in:
- Criminal Justice (pretrial release, sentencing, parole decisions)
- Healthcare (choosing among treatment options)
- Consumer finance (issuing loans)

Typically designed to predict observable outcomes:
- Recidivism
- Health outcomes
- Default on a loan

Fairness also often defined in terms of observables
- Error rates (False Positives, False Negatives)
- Calibration, predictive parity
- Equalized odds, equal opportunity

Problem: Observable outcomes confound risk and the effect of interventions.
- Limited use to decision-makers.
- Hard to evaluate performance or fairness.

Solution: Use potential outcomes under various intervention(s) instead.
- More useful information for decision-makers.
- More sensible definitions of fairness.

Observable and potential outcomes

Notation

Observable variables
- \( A \): Exposure (e.g. incarceration)
- \( Y \): Outcome (e.g. recidivism)
- \( R \): Race (0 = black, 1 = white)
- \( S \): Other covariates
- \( \hat{Y} \): Predictor of outcome \( Y \)

Potential outcomes
- \( Y^{A=a} \): Outcome under treatment \( a \)

Assumption:
- \( Y = \sum Y^{A=a} \mid (A = a) \)
Potential outcome \( Y^a \) is observed when treatment is set to \( A = a \).

Confounding with observable outcomes

Example: predicting outcome for pneumonia patients.

Predicting observable outcomes:
- \( A \in \{0, 1\} = \) hospitalization indicator
- \( Y \in \{0, 1\} = \) death indicator

Doctors treat older patients more aggressively.
Result: **Spurious negative correlation** between age and death.

Predicting potential outcomes:
- \( Y^0 \): death indicator under no hospitalization
- \( Y^1 \): death indicator under hospitalization

By assumption, \( Y = AY^0 + (1-A)Y^1 \)

COMPAS: Potential Outcomes Reanalysis

The COMPAS recidivism prediction tool (Northpointe, inc.) predicts recidivism for a crime within 2 years.

Data (ProPublica, 2016):
- 5,278 arrest cases from Broward County, FL
- 3,175 black; 2,103 white
- Jail durations, recidivism outcomes, covariates

\[ S \in \{0, 1\} = \text{COMPAS score, } 1 = \text{“high risk”} \]

\[ \text{2 scores: General (G) and recidivism risk (G)} \]

Violent recidivism risk (V)

Previous Analyses
- ProPublica (2016): Found different error rates and predicted score ratios based on race.
- Northpointe (2016): Found different error rates and predicted score ratios based on race.

Analysis 1: False Positive Rates

ProPublica:
- Assumes \( A \parallel Y^{0=R}=0, R=r, X \)

Results:

<table>
<thead>
<tr>
<th>( (G) )</th>
<th>( (\hat{V}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>Black</td>
</tr>
<tr>
<td>ProPublica</td>
<td>0.33</td>
</tr>
<tr>
<td>Reanalysis</td>
<td>0.24</td>
</tr>
</tbody>
</table>

Counterfactual scores show similar bias.

Analysis 2: False Negative Rates

ProPublica:
- Assumes \( A \parallel Y^{0=S}=0, R=r, X \)

Results:

<table>
<thead>
<tr>
<th>( (G) )</th>
<th>( (\hat{V}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>Black</td>
</tr>
<tr>
<td>ProPublica</td>
<td>0.48</td>
</tr>
<tr>
<td>Reanalysis</td>
<td>0.54</td>
</tr>
</tbody>
</table>

Counterfactual scores show similar bias.

Analysis 3: Positive Predictive Values

Northpointe:
- Assumes \( A \parallel Y^{0=S}=1, R=r, X \)

Results:

<table>
<thead>
<tr>
<th>( (G) )</th>
<th>( (\hat{V}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>Black</td>
</tr>
<tr>
<td>Northpointe</td>
<td>0.59</td>
</tr>
<tr>
<td>Reanalysis</td>
<td>0.65</td>
</tr>
</tbody>
</table>

COMPAS Reanalysis Conclusions

- Error rates (FPR, FNR) similar to ProPublica results.
- Approximate predictive parity, similar to Northpointe results.
- Slightly higher PPVs for blacks than whites.
- General recidivism: Slightly higher PPVs than from observed outcomes.
- Bias in score ratios, but much less than in ProPublica results (not shown).

Predicting Recidivism in Pennsylvania

Background
- State of Pennsylvania currently developing a recidivism prediction instrument.
- Mandated by 2010 legislation.
- Goal: identify low- and high-risk defendants for further analysis.

Data
- Records from 131,076 criminal defendants from 2004-2006 (Pennsylvania Commission on Sentencing)
- \( A \in \{0, 1\} = \) indicator for minimum sentence served
- \( X = \) Covariates
- \( Y = \) Recidivism

Analyses
- Compare “naive” modeling approach to potential outcomes-based approach

Naive model

\[ S = \tilde{E}[Y|X] \]

Potential outcomes model

\[ S = \tilde{E}[Y^{0}=X] \]

Identifiable under assumptions of consistency, exchangeability, and positivity.

Pennsylvania Recidivism Conclusions

- Non-trivial proportions of changes in predicted outcome.
- Error rates nearly identical.
- Unclear if incarceration has an effect beyond aging.
- Positivity violation means we can’t estimate \( Y^1 \).
- Further work:
  - Are there systematic differences in the two models for certain subpopulations?
  - Comparing the models on fairness criteria.