Modeling Evaluator Confusions

- What kinds of mistakes are evaluators making?
- Can we identify patterns of mistakes in the aggregate decision making process?
- How do these patterns change over time?

**Methodology**

Criteria for Interpretability
- Distinctness: Minimal overlap of rules w.r.t the datapoints they cover
- Parsimony: Fewer rules with fewer predicates
- Class Coverage: Explain as many classes as possible

Solution
- Non-negative, non-normal, non-monotone, submodular objective
- Smooth Local Search [Feige et. al.] provides a 2/5 approximation

User Study
- Interpretable decision sets vs. Bayesian Decision Lists (Letham et. al.)
- Each user is randomly assigned one of the two models
- 10 objective and 2 descriptive questions

Text Labeling Task
- Evaluators are often confused between atheism and Christianity when documents are short.
- Female evaluators with low self-reported confidence scores are highly accurate!!

Interpretable Decision Sets

**Decision Sets**

- Respiratory Illness: Yes and Smoker = Yes and Age > 50 then Lung Cancer
- Risk: Lung Cancer: Yes and Blood Pressure = 90 then Lung Cancer
- Risk: Depression: Yes and Post Depression then Depression
- BMI ≥ 20 and Insurance: None and Blood Pressure ≥ 92 then Diabetic
- Smoker: Yes and BMI ≥ 20 and Age ≥ 60 then Diabetes
- Risk: Diabetes: Yes and BMI ≥ 20 and Post-Infections ≥ 0.2 then Diabetes
- Doctor visits: 0.4 and Childhood Obesity: Yes then Heart Disease

Given a patient with the following attributes, Respiratory Illness = Yes and Smoker = Yes, can you be absolutely sure that this patient suffers from Lung Cancer?

<table>
<thead>
<tr>
<th>Task</th>
<th>Metric</th>
<th>User Approach</th>
<th>Bayesian Decision Lists</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive</td>
<td>Human Accuracy</td>
<td>0.81</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>Avg. Time Spent (sec.)</td>
<td>113.4</td>
<td>396.86</td>
</tr>
<tr>
<td>Objective</td>
<td>Human Accuracy</td>
<td>0.97</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>Avg. Time Spent (sec.)</td>
<td>31.11</td>
<td>120.57</td>
</tr>
</tbody>
</table>

Cost-Effective Treatment Regimes

**Criteria for Cost-Effective Treatment Regimes**
- Maximal outcomes
- Minimal assessment costs
- Minimal treatment costs

Solution
- Objective function NP-Hard
- Formulate as a Markov Decision Process
- UCT algorithm with customized search space pruning

Experimental Results

Quantitative Analysis
- Experiments with bail decisions and asthma treatment recommendations
- Outcomes better than human experts in 29% of the cases
- Outcomes match state-of-the-art algorithms with 34% lesser assessment costs and 14% lesser treatment costs

Ongoing Research

- Learning unsupervised feature representations for decision making
- Can we attribute interpretability to these representations?
- Designing algorithmic frameworks which can intelligently incorporate human feedback in debugging machine learning models
- How do we discover unknown unknowns of complex models?
- Human judgements vs. Machine Predictions
  - A case study on bail decisions
  - How can machine learning algorithms help in critical decisions such as bail?

References


Contact: Himabindu Lakkaraju
Email: himalv@cs.stanford.edu