Towards Unbiased and Accurate Deferral to Multiple Experts

Vijay Keswani  
Yale University

Matthew Lease  
University of Texas at Austin  
Amazon AWS AI

Krishnaram Kenthapadi  
Amazon AWS AI

Human-in-the-loop frameworks
- Desirable to augment ML model predictions with expert inputs. Useful for:
  - Improving accuracy
  - Incorporating human expertise
  - Auditing models

Popular examples
- Healthcare models
- Content moderation
- Risk assessment and screening

Errors and biases in ML
- Human-in-the-loop frameworks can reflect biases or inaccuracies of the human experts. Examples:
  - Racial bias in human-in-the-loop framework for recidivism assessment (Green, Chen 2019)
  - Ethical concerns regarding audits of facial processing technologies (Raji et al. 2020)
  - Automation bias in time critical decision support systems (Cummings 2004)

Can we design human-in-the-loop frameworks that account for expertise and biases of human experts?

Prior work
- Rejection learning - Pass when not sure; experts are not explicit here (El-Yaniv et al. 2010, Cortes et al. 2016, Kamiran et al. 2012, Li et al. 2011)
- Learning to defer or joint decision-making with explicitly specified human(s)
  - Theoretical analysis limited to single expert (Madras et al. 2018, Mozannar, Sontag 2020)
  - Empirical analysis limited to studying correlations from data (Green and Chen 2019, De-Arteaga et al. 2020, Yaghi et al. 2019)

Can we design frameworks that
- can handle multiple (kinds of) experts,
- has feasible optimization formulation,
- improves accuracy and fairness of predictions?

Theoretical Properties
- Projected Gradient Descent can be used to obtain optimal classifier and deferrer
- Intuitive gradient updates- rewards good experts
- If $L_{clf}$ is Lipschitz-smooth, then projected-gradient descent converges close to optimal classifier and deferrer in time polynomial in number of experts
- Deferrer weights can be used to choose committees of smaller sizes as well

Analysis using real-world dataset
- Dataset - 1471 Twitter posts
- MTurk survey presented to 170 participants to label whether post is offensive or not
- Overall accuracy of aggregated response – 87%
- Heterogeneous expert domains - 92 participants had higher accuracy for non-AAE posts, 75 participants had higher accuracy for AAE posts

Performance of framework for this dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Overall Accuracy</th>
<th>Non-AAE Accuracy</th>
<th>AAE Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classifier only</td>
<td>.78 (.02)</td>
<td>.76 (.05)</td>
<td>.80 (.04)</td>
</tr>
<tr>
<td>Joint framework</td>
<td>.85 (.03)</td>
<td>.87 (.04)</td>
<td>.83 (.03)</td>
</tr>
<tr>
<td>Joint balanced framework</td>
<td>.84 (.03)</td>
<td>.87 (.03)</td>
<td>.81 (.04)</td>
</tr>
<tr>
<td>Joint minimax framework</td>
<td>.85 (.02)</td>
<td>.87 (.02)</td>
<td>.83 (.02)</td>
</tr>
</tbody>
</table>

Our framework improves the accuracy and fairness of the final prediction for this real-world dataset as well, despite heterogeneity in expert performances

Discussion, Limitations and Future Work
- Larger real-world datasets can be constructed for more robust analysis of hybrid frameworks
- Extension to settings where experts are replaceable or when more experts can be added
- Improved selection of smaller committees