Exercise Session 12

Reasoning and Decision-Making

Prepared by Liangze Jiang and Eeshaan Jain

Overview

Task 1. Human-alone, human-with-AI, and AI-alone decision-making systems.

More and more high-stakes decisions are now being made with the help of AI and algorithms: judges deciding whether to detain a suspect; doctors deciding whether to prescribe a certain medication; banks deciding whether to approve a loan. Current research mostly focuses on how accurate the AI system is and whether it is biased. But very few studies ask the more fundamental question: **Does providing AI-generated recommendations to human decision-makers improve their classification accuracy compared to human-alone or AI-alone systems?** This is an important question because the goal of deploying AI in high-stakes settings is not only to build accurate models, but also to enhance real-world decisions made by humans.

Problem setting: At a pretrial hearing, a judge must decide whether to release ($D = 0$) or detain an arrestee ($D = 1$). If released, the arrestee may ($Y(0) = 1$) or may not ($Y(0) = 0$) commit a new crime.

(a) Construct the confusion matrix of each combination of potential outcome and decision (e.g., p_{00} indicates released and not commit a new crime). Are all the cases observable in practice? If not, which ones are not observable (and why)?

(b) What metric are we typically trying to optimize (minimize) for? Suppose now we use a new weighted metric $p_{01} + l_{10}p_{10}$, which of the following systems is better for different values of $l_{10} \in (0, +\infty)$? (Here assume we have oracle access to the outcomes)

	Human-only	AI-only
Release & Would commit crime ($D = 0, Y(0) = 1$)	10%	8%
Detain & Would not commit crime ($D = 1, Y(0) = 0$)	15%	25%

(c) Suppose that we *randomly* split the arrestee into two groups of equal size. One group is evaluated by a human-only judge, while the other group is evaluated by a judge who sees an AI recommendation first (Human-with-AI). The outcomes are as follows.

	Human-only	Human-with-AI
Release & would not commit Crime ($D = 0, Y(0) = 0$)	40%	36%
Release & would commit Crime ($D = 0, Y(0) = 1$)	10%	9%

Can we determine that Human-with-AI performs better than Human-only in terms of misclassification rate?

Task 2. Chain-of-thought prompting for eliciting reasoning in LLMs.

See the Jupyter notebook for more details.

Task 3. ReAct: synergizing reasoning and acting in LLMs.

See the Jupyter notebook for more details.
