Notes on Prompt Risk Control: A Rigorous Framework for Responsible Deployment of Large Language Models

Catherine Chen cyc2152

December 2023

T. P. Zollo, T. Morrill, Z. Deng, J. C. Snell, T. Pitassi, and R. Zemel, "Prompt Risk Control: A Rigorous Framework for Responsible Deployment of Large Language Models." arXiv, Nov. 22, 2023. Accessed: Dec. 17, 2023. [Online]. Available: http://arxiv.org/abs/2311.13628

Abstract: Propose algorithm to select a prompt based on rigorous upper bounds on families of informative risk measures that accommodate for the possibility of distribution shifts in deployment.

1 Introduction

Definition 1.1 (Loss)**.** A particular scoring notion that can be calculated for a single instance, such as, ROUGE score, and top-1 accuracy.

Definition 1.2 (Risk)**.** A population-level measure of scores, such as, mean, median, or Conditional Value-at-Risk (CVaR).

2 Setting and Notation

Consider $S = \{(x_i, y_i)\}_{i=1}^n$, a validation dataset drawn from a joint distribution $\mathcal{D}$ over user queries $x \in \mathcal{X}$ and gold standard responses $y \in \mathcal{Y}$, with a generator model $G : \mathcal{X} \to \mathcal{O}$, a LLM. To improve the response to query x, a prompt $p \in \mathcal{P}$ may be added to the input to G. For a given prompt p, G_p is a model that produces a response to x using p. Here, $\mathcal{X}, \mathcal{Y}, \mathcal{O}$ and $\mathcal{P}$ are spaces of text strings.

Assume we are given a bounded loss function $l : \mathcal{O} \times \mathcal{Y} \to \mathbb{R}$ that captures the generation quality of G, with a lower score denoting a better response. A loss function scores the quality of a generation for a single example.

A risk function measures some aspect of the distribution of loss across the population. Define a general notion of risk as a function $R : l \to \mathbb{R}$, where l, the loss value, is treated as the distribution of a random variable. In general, $l = l(O, Y)$ represents the distribution of loss scores over random subsets of paired responses $O \subseteq \mathcal{O}$ and labels $Y \subseteq \mathcal{Y}$ (which may be dummy labels if not required by the loss function).

$R(G_p, l)$ is a shorthand for $R\left(l\left(O_{G_p}, Y\right)\right)$, where O_{G_p} denotes the outputs produced by generator G using prompt p.

3 Prompt Risk Control

3.1 Bounding the Mean: Learn Then Test (LTT)

3.2 Quantile Risk Control (QRC)

4 Extending Bounds for Distribution Shifts