北京理工大學

数学与统计学院学术报告

A Statistical Framework for Alignment with Biased AI Feedback

报告人： 蔡占锐　香港大学经管学院

时间： 2026年5月15日，11:00-12:00

地点： 北京理工大学良乡校区文萃楼E207

摘要： Modern alignment pipelines are increasingly replacing expensive human preference labels with evaluations from large language models (LLM-as-Judge). However, AI labels can be systematically biased compared to high-quality human feedback datasets. In this paper, we develop two debiased alignment methods within a general framework that accommodates heterogeneous prompt-response distributions and external human feedback sources. Debiased Direct Preference Optimization (DDPO) augments standard DPO with a residual-based correction and density-ratio reweighting to mitigate systematic bias, while retaining DPO's computational efficiency. Debiased Identity Preference Optimization (DIPO) directly estimates human preference probabilities without imposing a parametric reward model. We provide theoretical guarantees for both methods: DDPO offers a practical and computationally efficient solution for large-scale alignment, whereas DIPO serves as a robust, statistically optimal alternative that attains the semiparametric efficiency bound. Empirical studies on sentiment generation, summarization, and single-turn dialogue demonstrate that the proposed methods substantially improve alignment efficiency and recover performance close to that of an oracle trained on fully human-labeled data.

个人简介： 蔡占锐，现任香港大学经管学院创新与信息管理系助理教授。于2021年于宾夕法尼亚州立大学获得统计学博士学位，并于之后在卡内基梅隆大学进行博士后研究，以及在爱荷华州立大学统计系担任助理教授。研究兴趣包括统计在大模型中的应用，统计推断中的隐私保护问题，以及机器学习在统计方法中的应用等。