基于偏好的奖励建模中重新思考 Bradley-Terry 模型

Rethinking the Bradley-Terry Models in Preference-based Reward Modeling: Foundation, Theory, and its Alternatives

https://sites.google.com/view/rewardmodels

1. Motivation

- 为什么 Bradley-Terry 模型可以用于奖励建模？其背后的原理、假设和逻辑是什么？
- 在奖励建模方面，真正重要的是什么？除了 Bradley-Terry 模型，我们还有哪些其他选择？
- 在理解了 Bradley-Terry 模型和奖励建模之后，当我们重新审视它们时，当前实践中哪些方面可能需要改进？

2. Part I: Rethinking the Bradley-Terry Models in Alignment

2.1 Strating from the Two Bradley-Terry Models

自 20 世纪 50 年代以来，BT 模型及其各种改进已被用于棋类游戏和各类体育项目的技能评估和胜率预测。最直观的解释是，当能力得分为 r_A 的团队 A 面对能力得分为 r_B 的团队 B 时，A 击败 B 的概率由下式给出：

$$P(A \text{ wins}) = \frac{r_A}{r_A + r_B}$$

在实际场景中，由于比赛结构、规则和随机性的差异，使用原始分数来预测队伍的胜率往往是不够的。这需要将分数标准化，以考虑每种比赛类型中固有的随机性。例如，在足球比赛中，随机性起着重要作用，即使是较弱的队伍有时也能战胜更强的队伍。同样，在涉及不完全信息的纸牌游戏中，获胜并不总是能反映玩家的真实技能水平。相比之下，像国际象棋和围棋这样的游戏——其中完整信息是可用的——除了比赛期间玩家的瞬间表现外，随机性要小得多。

BT 模型最直接的应用是：给定不同队伍之间的历史比赛数据，我们能否为每支队伍分配分数来评估它们的技能水平，然后使用这些分数来准确预测未来比赛的结果？我们将此称为参数估计，其中每支队伍的分数是需要估计的参数。理论上，即使在缺乏随机性的情况下，也需要大约 NlogN 次比较来对团队进行排名。考虑到随机性，目前最佳理论结果表明，为了获得较为准确的估计，至少需要 Nlog^3N 次比较。

以下是LLM中对于BT model应用的两个例子：

- 第一个例子是 LLM Chatbot Arena，也称为 LLM 梯度。在这个设置中，不同的 LLM 被视为“参与者”。在每场比赛中，两个 LLM 进行竞争，用户根据他们对回答质量的判断来决定胜者。在 LMSYS 的 Arena 中，已经进行了超过 2,000,000 场比赛，比较了超过 150 个 LLM。平均而言，每对 LLM 已经被比较了超过 26,000 次。这里，N=150，而 26000≫N(logN)^3≈1500。因此，LMSYS 可以为每个 LLM 的分数提供 95% 的置信区间。
- 另一方面，在用于 LLMs 的指令微调或人类反馈强化学习（RLHF）过程中，BT 模型用于将成对偏好标注转换为分数。在这里，每对提示-响应被视为一个玩家（尽管也可以将同一提示的两个响应视为两个玩家）。当我们有 N 个提示时，就有 2N 个响应。标注者对这些 2N 个响应进行标注，这相当于生成 N 场比赛的结果。但问题是：有 N 场比赛时，我们能可靠地给 2N 个玩家打分吗？毕竟，N≪2N(log2N)^3。

实际上，这并不是 LLM 指令微调所面临的核心问题。在体育领域，即使两支队伍没有进行过多场比赛，我们也可以使用某些特征（例如，平均身高、年龄、主要赛事经验、市场价值或平均海参消费量）来预测比赛结果。这里的重点从参数（能力）估计转移到结果预测。这在 BT 模型的历史研究中已被广泛研究，在文献中，它被称为 Bradley-Terry 回归。

遵循这一思路，本论文针对 LLM 对齐的具体场景，并提供了一个关于将 Siamese MLP 结构应用于实现 BT 回归的收敛性证明。

2.2 Assumptions under the Bradley-Terry Models in LLM Alignment

当我们把偏好视为匹配，并尝试使用 Bradley-Terry（BT）模型来模拟结果时，需要做出哪些假设？

不同的响应是参与者，偏好代表匹配结果。这些匹配中的随机性来自哪里？

一个自洽的解释是我之前提到的那个。如果我们假设每个玩家的"表现"围绕其真实技能水平呈高斯分布，并且进一步假设这个高斯分布对所有玩家具有相同的方差，那么玩家的分数可以通过这个方差进行标准化。这将引导我们得到一个误差函数（erf）版本（而不是 tanh）的（伪）BT 模型。为了得到经典的 BT 模型，我们假设每个玩家的表现遵循 Gumbel 分布，其位置参数等于玩家的技能水平。即便如此，仍然有一个关键细节需要澄清——这个比赛中的方差或随机性来自哪里？

BT 模型假设不同标注者在不同回答的确定性方面存在偏差。这些偏差遵循 Gumbel 分布（因此它们之间的差异遵循逻辑分布）。类似地，我们可以假设这些偏差遵循高斯分布，在这种情况下它们的差异也将是高斯的——从而得到 BT 模型的 erf 版本。

此外，我们可以从另一个角度（受一些认知心理学文献中关于认知瓶颈的讨论启发）来分析。标注的正确性（即是否正确地排序真实奖励值）取决于奖励之间的绝对差异。直观地讲，如果响应 1 和响应 2 的真实得分分别为 r1 和 r2，并且如果 r1 和 r2 非常接近，那么错误标注这两个响应之间偏好的概率会增加，使得标注更像是随机猜测。同时，不同的标注者区分这些细微差异的能力也不同。理想情况下，一个完美的标注者能够区分任何微小的奖励差异，并且根据真实奖励，这个标注者总能以 100%的准确率正确排序偏好。在另一极端，如果标注者能力很差，即使奖励差异很大，他们也可能无法正确区分响应，他们的标签会是随机的（正确率为 0.5）。

3. Rethinking the Reward Modeling Objective

3.1 The Concept of Order Consistency

之前，我们讨论了 BT 模型背后的假设以及使用 BT 模型将偏好数据转换为分数的逻辑。由于我们在嵌入空间中进行回归，不同提示-响应对之间的排序关系可以在一定程度上推广到其他提示-响应对，因此我们不需要像经典的 BT 模型那样使用更多的样本来估计参数。相反，我们可以使用相对较少的标注来对新提示-响应对进行预测。

这里，我们再次可以比较奖励模型与传统运动赛事中 BT 回归的差别。在运动中，我们关心每场比赛的获胜概率（至少，那些下注的人是关心的）。在这种情况下，为了做出准确的预测，知道每支队伍的精确得分变得至关重要。然而，在奖励模型的背景下，我们并不关心一个提示-响应对战胜另一个提示-响应对的精确概率。我们更关心的是它们之间的排序。当使用奖励模型（例如，在推理时优化期间），我们会对单个提示获得多个响应。我们不需要精确预测每个响应战胜其他响应的概率；我们只需要识别出最佳响应。

从这个角度来看，使用 BT 模型进行奖励建模似乎过于关注细节。通过重新审视数据，我们提出了一个更通用的高层次优化目标：顺序一致性。形式上，我们将其定义为如下：

Definition 13 (Order Consistency) *We consider the loss over an ordering model* $\hat{H}$

$$\mathcal{L}_{\mathrm{oc}}(\hat{r}) = \mathbb{E}_{x_1,x_2,y_1,y_2,h} \mathbb{1}\left[h = \hat{H}\right] \tag{16}$$

That is, the probability that a reward model ordering agrees with annotation.

当给定一个标注数据集时，我们所能做的只是复制数据集中的（不完美的）标注。我们用\hat{H}表示顺序模型。只要我们足够优化\hat{H}，这个\hat{H}就不会偏离真实标注太远。

Proposition 14 (Lower bound on population level order consistency) *Suppose a learned model* $\hat{H}$ *achieves objective equation 16 up to* $1 - \delta\epsilon$ *error for some small* $0 < \delta < 1$ *and* $\epsilon < 3/20$*, i.e.,*

$$\mathbb{E}_{x_1,x_2,y_1,y_2,h} \mathbb{1}\left[h = \hat{H}\right] \geq 1 - \delta\epsilon \tag{17}$$

3.2 BT Model and Order Consistency

显然，BT 模型是一种优化顺序一致性的方法。它明确地将\hat{H}表示为两个奖励估计值的差。这种表述方式具有反对称性——如果我们交换两个奖励估计值的位置，排序估计的符号将相应地反转。

3.3 Classification and Order Consistency

此外，直接对标记为正/负的样本进行二元分类也是对顺序一致性的优化，但它优化的是顺序一致性的上限。具体来说，它要求对正样本提示-响应样本的预测值大于 0，对负样本的预测值小于 0。在这里，反对称性的明确要求不再存在——我们期望分类器能够从数据中学习这一特性。

理论上，BT 模型优化的是一对提示-响应在与其他对比时获胜的概率。另一方面，分类模型优化的是正样本是“好”样本的概率。这种直接的分类方法本质上消除了负样本的影响，将其视为该样本在对比中获胜的边际概率。

本文提出了以下结果：BT 奖励和分类奖励可以通过一个与 iii 无关的常数连接起来。当加上一个常数时，分类奖励可以上界 BT 奖励。

Proposition 31 (Classification reward) *Suppose data actually coming from BT model Equation* (1)*, and the score* $s_i := \operatorname{logit} P(i \text{ wins})$ *is connected to BT reward that for a constant* C *does not depends on* i

$$s_i \geq r_i - C$$

Proof We condition on which j that i competed with and apply Jensen's inequality

$$\mathbb{P}(i \text{ wins}) = \mathbb{E}_j[\mathbb{P}(i \succ j | j)] = \mathbb{E}_j\left[\frac{u_i}{u_i + u_j}\right] \geq \frac{u_i}{u_i + \mathbb{E}[u_j]}$$

With some straightforward algebra, we have that

$$\frac{\mathbb{P}(i \text{ wins})}{1 - \mathbb{P}(i \text{ wins})} \mathbb{E}[u_j \geq u_i]$$

Take log at each side and substitute $u_i = \exp(r_i)$ then rearrange

$$s_i := \operatorname{logit} \mathbb{P}(i \text{ wins}) \geq r_i - \log \mathbb{E}[\exp(r_j)]$$

we have $\log \mathbb{E}[\exp(r_j)]$ is a constant. ■

4. Rethinking the Annotation Strategy for Global Reward Approximators

在之前的分析中，我们观察到所有推导过程都不需要提示-响应对来自相同的提示。这是因为，在 BT 回归的假设下，比较发生在嵌入空间中，该空间表示提示和响应的联合嵌入。直观上，在嵌入空间的不同点之间进行顺序一致性学习本质上是给这些点分配分数。这些分数是全局适用的，而不是特定于某些提示的。

实际上，比较不同的提示-响应对的能力是奖励建模的隐含前提条件。这种能力使我们能够学习一个奖励模型，该模型可以预测任何提示-响应对的分数。我们正在学习的奖励是一个通用函数逼近器（受**多目标强化学习**中通用价值函数逼近器（UVFA）概念的启发）。

4.1 Why Compare Different Prompt-Response Pairs?

可以从体育竞赛中得出一个有用的类比：想象我们有很多年轻的足球运动员，他们住在不同的城市。这些球员在他们自己的城市内比赛，并根据当地比赛获得排名。然而，如果没有跨城市的比赛，就很难评估这些球员在更广泛、全球层面的表现。同样，仅使用本地数据训练的评分模型将局限于本地比较。如果我们想预测一个来自未知城市的球员的表现，我们需要更多样化的比赛数据——理想情况下，不仅仅是城市内的比赛，还应包括城市间的竞争。

在经典的分类问题中，选择用于标注的数据至关重要：我们不能只选择明显的正面或负面样本，因为训练任务会变得过于简单，导致在面对更复杂的样本时性能不佳。同样，如果我们只关注分类边界附近的样本，过拟合的风险就会增加。对于奖励建模来说，这一挑战转化为奖励操纵的风险。

4.2 Cross-Prompt Comparisons in Reward Modeling

在之前的工作中，例如 RPO [3]，已经提出了比较来自不同提示的响应的想法。例如，在评估一个有帮助的聊天机器人的上下文中，我们可以比较来自 LLM 对不同提示的响应，以标记一个提示的响应是否比另一个提示的响应更有帮助。虽然有些提示可能“更容易”，使得模型更有可能生成有帮助的答案，但关键的想法是跨提示比较帮助我们学习使响应变得有帮助的通用描述。

此外，跨提示比较可以带来更高质量的标注。直观地看，随机选择两个提示及其对应响应进行比较，能创造更大的多样性，并使识别出更好响应变得更容易。这与比较同一提示的两个响应的更受限制的情况形成对比，其中差异可能更细微。因此，我们提出以下结果

Proposition 15 (Cross-Prompt Comparisons Increase Utility Diversity) *When data for pairwise annotation is generated through random sampling of two responses $y_1, y_2 \sim \ell(x)$, and the utility of those two responses are sampled from a Gaussian distribution with variance σ_x^2, i.e., $y \sim \ell(x), r_{x,y} \sim \mathcal{N}(\mu_x, \sigma_x^2)$, when there are multiple prompts x, we have*

$$\mathbb{E}_x \mathbb{E}_{y_1,y_2|x} \left[|r_{x,y_1} - r_{x,y_2}| \right] \leq \mathbb{E}_{x_1,x_2} \mathbb{E}_{y_1|x_1,y_2|x_2} \left[|r_{x_1,y_1} - r_{x_2,y_2}| \right] \tag{26}$$

需要更严谨的证明来考虑标注的准确性。为了证明由于随机选择提示，标注质量在期望上会得到提升，我们推导了以下结果。

Theorem 16 (Cross-Prompt Annotation Improves Annotation Quality) *When data for pairwise annotation is generated through random sampling of two responses $y_1, y_2 \sim \ell(x)$, and the utility of those two responses are sampled from a location-scale family with probability density function $g_x(x) = f((x - \mu_x)/\sigma_x)$ for f being unimodal and symmetric to 0. For any $\xi : \mathbb{R}_+ \to [1/2, 1]$, first order differentiable, monotone increasing and concave, we have*

$$\mathbb{E}_x \mathbb{E}_{y_1,y_2|x} \left[\xi(|r_{x,y_1} - r_{x,y_2}|) \right] \leq \mathbb{E}_{x_1,x_2} \mathbb{E}_{y_1|x_1,y_2|x_2} \left[\xi(|r_{x_1,y_1} - r_{x_2,y_2}|) \right]. \tag{27}$$

因此，作者好奇地通过实验来验证，是否允许不同提示之间进行标注会对奖励模型更有益，实验对这一方面进行了深入探索。

5. Experiments: Building Strong Reward Models with Classifiers

主要进行了三个方面的实验：

- 探讨了在使用嵌入作为输入时，BT 奖励模型和分类奖励模型之间的性能差异。在大多数实验中，分类奖励模型的性能可以与 BT 奖励模型相媲美。然而，分类奖励模型更加灵活——它可以使用任何现有的分类器来实现，例如 MLP，或者像 LightGBM/XGBoost 这样的基于树的模型。
- 研究了在不同标注质量和数量下，不同奖励建模方法的性能变化。随着标注质量下降或数据量减少，与分类奖励模型相比，BT 奖励模型的性能表现较差。

- 比较了跨提示标注和单个提示的多个响应标注在奖励建模方面的性能，并对跨提示场景进行了压力测试，展示了跨提示能带来更大性能提升的条件。当响应的多样性较低时，跨提示标注可以显著提高性能。在实践中的 RLHF 工作流程（通常默认采用在线方法，因为它们能显著获得更好的结果），其中为标注者随机生成两个响应进行标注，跨提示标注可以显著提高奖励建模的有效性。