DPO，直接偏好学习

无需 reward modle 直接用分类 Loss 来对齐

2 背景知识

理解DPO（分布式的策略优化）的关键在于掌握其背后的几个核心概念：KL散度⁺、Bradley-Terry模型⁺和强化学习（RL）的优化目标。

2.1 KL散度

KL散度（Kullback-Leibler divergence）主要用于衡量两个概率分布之间的差异。

定义：

对于离散随机变量，假设P和Q是同一个随机变量的两个不同的概率分布，其中P通常表示真实分布，而Q表示估计或近似分布。KL散度定义为：

$D_{\mathrm{KL}}(P \parallel Q) = \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right)$ 对于连续随机变量，公式变为积分形式：

$D_{\mathrm{KL}}(P \parallel Q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$ 其中$p(x)$和$q(x)$分别是P和Q的概率密度函数。

性质：

- **非负性：** $D_{\mathrm{KL}}(P \parallel Q) \geq 0$，当且仅当P = Q时等号成立。
- **不对称性：** 一般来说，$D_{\mathrm{KL}}(P \parallel Q) \neq D_{\mathrm{KL}}(Q \parallel P)$，因此KL散度不是一个距离度量。

2.2、BT 模型：

Bradley-Terry模型主要是针对**比较关系进行建模。**

使用数据中包含的相对好坏关系对模型进行优化

让我们用下面的例子进行讲解：

假设我们有一系列篮球比赛，并有以下历史比赛结果：
A 对 B：A 胜 3 次，B 胜 1 次
A 对 C：A 胜 1 次，C 胜 1 次
B 对 C：B 胜 1 次，C 胜 1 次

似然函数

$$L(\alpha_A, \alpha_B, \alpha_C) = \left(\frac{\alpha_A}{\alpha_A+\alpha_B}\right)^3 \left(\frac{\alpha_B}{\alpha_A+\alpha_B}\right)^1 \left(\frac{\alpha_A}{\alpha_A+\alpha_C}\right)^1 \left(\frac{\alpha_C}{\alpha_A+\alpha_C}\right)^1 \left(\frac{\alpha_B}{\alpha_B+\alpha_C}\right)^1 \left(\frac{\alpha_C}{\alpha_B+\alpha_C}\right)^1$$

A胜B的概率.　α_i表i队的实力

为了简化计算，我们通常取对数似然函数：

$$\ln L(\alpha_A, \alpha_B, \alpha_C) = 3\ln\left(\frac{\alpha_A}{\alpha_A+\alpha_B}\right) + 1\ln\left(\frac{\alpha_B}{\alpha_A+\alpha_B}\right) + 1\ln\left(\frac{\alpha_A}{\alpha_A+\alpha_C}\right) + 1\ln\left(\frac{\alpha_C}{\alpha_A+\alpha_C}\right) + 1\ln\left(\frac{\alpha_B}{\alpha_B+\alpha_C}\right) + 1\ln\left(\frac{\alpha_C}{\alpha_B+\alpha_C}\right)$$

除上述对每个参数求导并令导数为0求解，通过对对数似然函数求解最大值来求解α值外，还可以通过对对数似然函数取负求其最小值，进而来估计α值，即变成了如下loss function的求解，此时可以通过梯度下降法等方式来求解参数α值：

$\Rightarrow$ 转化为二分类

$\text{Loss} = -\mathbb{E}_{(\alpha_x, \alpha_y) \sim D}\left[\ln \frac{\alpha_x}{\alpha_x + \alpha_y}\right]$ * 该公式为在给定的数据分布D下，参数α_x和α_y的比值的对数期望

RL中的应用：设 $r(x,y)$ 为 reward M.

认为 $y_w > y_l$. 则似然函数为：

$$\ln \frac{r(x,y_w)}{r(x,y_w) + r(x,y_l)}$$

$\because r(x,y_w)$ 可能为负 故取指

$$\Rightarrow \ln \frac{\exp[r(x,y_w)]}{\exp[r(x,y_w)] + \exp[r(x,y_l)]}$$

$$\Rightarrow \ln \frac{1}{1 + \exp[r(x,y_l) - r(x,y_w)]} \Rightarrow \text{sigmoid}\,[r(x,y_l - r(x,y_w)]$$

最小化 loss $= -\sigma[r(x,y_l) - r(x,y_w)]$