تمارین سری چهارم درس مدل‌های زبانی بزرگ — حمیدرضا امیرزاده — ۴۰۱۲۰۶۹۹۹

---

۱) آ) استفاده از log و sigmoid باعث می‌شود که اگر اختلاف winner و loser زیاد باشد، پدیده gradient vanishing اتفاق بیفتد و در نتیجه تابع پاداش طوری آموزش نمی‌بیند که اختلاف winner و loser بسیار زیاد باشد.

$$\frac{\partial \log(\sigma(z))}{\partial z} = 1 - \sigma(z) \underset{z\to\infty}{=} 0$$

اگر جلوی اختلاف زیاد winner و loser را بگیریم، ممکن است باعث شود که به جز یک دسته به سایر داده‌ها پاداش کمی اختصاص داده شود، با وجود اینکه ممکن است داده loser در مقابل سایر داده‌ها، داده خوبی باشد و بنابراین نباید پاداش پایینی دریافت کند. به همین علت تابع پاداش مرتباً به سمت تعداد کمی از داده‌ها بایاس می‌شود.

ب) $r_\theta(x,y)$: زیرا می‌خواهیم داده تولیدی توسط مدل زبانی که پارامترهای ϕ را دارد، تابع پاداش را ماکسیمم کند. به عبارت دیگر Policy را آموزش دهیم که خروجی آن تابع پاداش را ماکسیمم کند.

$-\beta \log\left(\pi_\phi^{RL}(y|x) / \pi^{SFT}(y|x)\right)$: این ترم حاصل شده از KL به این دلیل است که اجازه ندهد در مسیر آموزش توزیع داده خروجی Policy بهینه‌سازی شده بر اساس تابع پاداش، در هر گام از توزیع خروجی Policy اولیه فاصله زیادی بگیرد. زیرا Policy با داده‌های تولید شده توسط خودش در هر گام آموزش می‌بیند و اگر اجازه دهیم از خروجی Policy اولیه فاصله زیادی بگیرد، ممکن است که داده‌های مهملی تولید کند (garbage) که این داده مرتباً می‌تواند تابع پاداش را ماکسیمم کند اما خروجی مدل را به سمت یک توزیع بد ببرد. علاوه بر این از آنجایی که داده آموزشی noisy است، می‌تواند این فاصله گرفتن به خاطر همان نویز به خروجی Policy منتقل شود.

$\gamma E_{x\sim D_{pretrain}}\left[\log(\pi_\phi^{RL}(x))\right]$: از آنجایی که داریم مدل را با ارزش‌های انسانی منطبق می‌کنیم، ممکن است این امر باعث شود که توانایی مدل در انجام تسک‌های معمول NLP کاهش یابد. بنابراین با امید ریاضی گرفتن روی داده‌های دیتاست‌های NLP از تمام احتمالاتی که Policy آموزش داده شده اختصاص می‌دهد به آن داده‌ها و ماکسیمم کردن این امید ریاضی می‌توان عملکرد مدل روی تسک‌های NLP را نیز خوب نگه داشت.

ج) زیرا y درون $r_\theta(x,y)$ با استفاده از سیاست ϕ به دست می‌آید، بنابراین تابعی از آن است. بنابراین مشتق $r_\theta(x,y)$ نسبت به ϕ صفر نیست و مقدار دارد.

د)

$$L_\theta = E_{\pi_\theta}[G_t] \qquad \nabla_\theta L_\theta = \nabla_\theta E_{\pi_\theta(\tau),\tau}[G_t] = \nabla_\theta \int \pi_\theta(\tau) G_\tau \, d\tau = \int \nabla_\theta \pi_\theta(\tau) G_\tau \, d\tau$$

$$= \int \pi_\theta(\tau) \frac{1}{\pi_\theta(\tau)} \nabla_\theta \pi_\theta(\tau) G_\tau \, d\tau = \int \pi_\theta(\tau) \nabla_\theta \log \pi_\theta(\tau) G_\tau \, d\tau$$

$$= E_{\pi_\theta,\tau}\left[\nabla_\theta \log \pi_\theta(\tau) G_\tau\right]$$

(۲) ۴) تابع هدف زیر را به صورت یک لاگرانژین نوشته و از آن نسبت به π_θ که توابع پیچیده‌ای هستند مشتق می‌گیریم و برابر صفر قرار می‌دهیم.

$$L = \int \left(r_\phi(x,y) - \beta \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)} \right) \pi_\theta(y|x)\, p(x)\, dx\, dy + \lambda \left(1 - \int \pi_\theta(y|x)\, p(x)\, dx\, dy \right)$$

↓

قید نرمالیزه بودن توابع

برای مشتق‌گیری از تابع $\pi_\theta(y|x)$ مقدار x, y را ثابت فرض می‌کنیم.

$$\frac{\partial L}{\partial \pi_\theta(y|x)} = \left(r_\phi(x,y) - \beta \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)} \right) p(x) - \beta\, p(x) - \lambda\, p(x) = 0$$

$$\Rightarrow \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)} = \exp\left(\frac{1}{\beta} r_\phi(x,y) \right) \cdot \exp\left(-\frac{\lambda+\beta}{\beta} \right)$$

$$\Rightarrow \pi_\theta(y|x) = \exp\left(-\frac{\lambda+\beta}{\beta} \right) \pi_{ref}(y|x) \exp\left(\frac{1}{\beta} r_\phi(x,y) \right) \quad : \quad Z(x) = \frac{1}{\exp\left(-\frac{\lambda+\beta}{\beta} \right)}$$

$$\Rightarrow \pi_\theta(y|x) = \frac{1}{Z(x)} \pi_{ref}(y|x) \exp\left(\frac{1}{\beta} r_\phi(x,y) \right)$$

ب)

$$\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)} Z(x) = \exp\left(\frac{1}{\beta} r_\phi(x,y) \right) \Rightarrow \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)} + \beta \log Z(x) = \frac{r_\phi(x,y)}{\beta}$$

$$\Rightarrow \beta \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)} + \beta \log Z(x) = r_\phi(x,y) \qquad L_R(r_\phi, D) = -E_{(x,y_w,y_l)\sim D}\left[\log \sigma(r_\phi(x,y_w) - r_\phi(x,y_l)) \right]$$

$$L_{DPO}(\pi_\theta : \pi_{ref}) = -E_{(x,y_w,y_l)\sim D}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)} \right) \right]$$

ج)

$$\nabla_\theta L_{DPO}(\pi_\theta : \pi_{ref}) = -E_{(x,y_w,y_l)\sim D}\left[\frac{\partial \log(\sigma(A))}{\partial \theta} \right]$$

$$A = \beta \left(\log \pi_\theta(y_w|x) - \log \pi_{ref}(y_w|x) \right) - \beta \left(\log \pi_\theta(y_l|x) - \log \pi_{ref}(y_l|x) \right)$$

$$= -E_{(x,y_w,y_l)\sim D}\left[(1 - \sigma(A)) \frac{\partial A}{\partial \theta} \right] = -\beta E_{(x,y_w,y_l)\sim D}\left[\sigma\left(\beta \log \frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)} - \beta \log \frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} \right) \right.$$
$$\left. \left(\nabla_\theta \log \pi_\theta(y_l|x) - \nabla_\theta \log \pi_\theta(y_w|x) \right) \right]$$

$$= -\beta E_{(x,y_w,y_l)\sim D}\left[\sigma\left(\hat{r}_\theta(x,y_l) - \hat{r}_\theta(x,y_w) \right) \left(\nabla_\theta \log \pi_\theta(y_w|x) - \nabla_\theta \log \pi_\theta(y_l|x) \right) \right]$$