T5 Actividades de RLHF

Índice

1 Introducción

☐ Reinforcement learning from Human Feedback (RLHF) es una técnica para incorporar información humana en sistemas IA. Su gran popularidad se debe a que gran parte del éxito de ChatGPT se atribuye al uso de RLHF en post-training de GPT. Tras ChatGPT, el post-training de LLMs producía tres modelos a partir del modelo base: el modelo SFT, el de recompensa y el alineado. En relación con estos modelos, indica la opción incorrecta o la última opción si las tres primeras son correctas:

1. El modelo SFT se entrena a partir del base mediante instruction / supervised fine-tuning (SFT). SFT adapta el modelo base para que siga un formato de respuesta predefinido, típicamente de diálogo en lenguaje natural.
2. El modelo de recompensa se entrena a partir del SFT y datos de preferencia. Por cada prompt de entrenamiento, el modelo SFT genera dos (o más) respuestas que se ordenan por preferencia humana.
3. El modelo alineado se entrena a partir del SFT y el de recompensa con aprendizaje por refuerzo. Por cada prompt de entrenamiento, el modelo SFT genera dos (o más) respuestas y el de recompensa las ordena por preferencia.
4. Las tres opciones anteriores son correctas.

Solución: La 4

☐ En el entrenamiento de LLMs actual, el post-training comprende diversas técnicas y prácticas con el fin de mejorar la utilidad de los LLMs en tareas específicas. Atendiendo al propósito de estas técnicas y prácticas, se distinguen tres tipos. En relación con los mismos, indica la opción incorrecta o la última opción si las tres primeras son correctas:

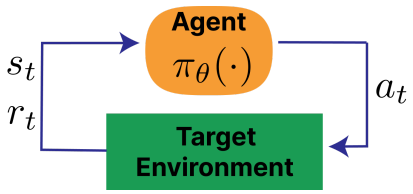
1. Instruction / supervised fine-tuning (IFT/SFT) se ocupa del aprendizaje de las características del lenguaje. El modelo pre-entrenado se adapta para predecir el siguiente token cuando el texto precedente se asemeja a ejemplos vistos. IFT es relativamente sencillo de implementar ya que es muy parecido al pre-entrenamiento.
2. Preference fine-tuning (PreFT) se orienta al aprendizaje del estilo del lenguaje según preferencias humanas, típicamente mediante reinforcement learning from human feedback (RLHF). El modelo SFT se adapta a nivel de respuesta, favoreciendo estilos de respuesta deseables y penalizando estilos indeseables. En general, PreFT es relativamente complicado de implementar ya que presenta múltiples retos sobre cómo controlar la optimización.
3. Reinforcement fine-tuning (RFT) tiene por objetivo la mejora del rendimiento en dominios verificables como la programación y las matemáticas.
4. Las tres opciones anteriores son correctas.

Solución: La 4

---

2 Reinforcement learning from human feedback (RLHF)

☐ El aprendizaje por refuerzo (reinforcement learning, RL) es una de las principales aproximaciones al aprendizaje automático, junto con el aprendizaje supervisado y el no supervisado. El modelo, agente o política a aprender, $\pi_\theta(\cdot)$, viene condicionado por el estado del entorno, s_t; tras ejecutar una acción a_t, obtiene una recompensa r_t y el entorno transita a un nuevo estado, s_{t+1}:



El objetivo del agente es aprender una política que maximice la recompensa de una trayectoria futura, esto es, de una secuencia de iteraciones estado-acción-recompensa futuras, hasta alcanzar un estado final o, en el caso de una tarea continua, indefinidamente. Formalmente, si $(r_{t+1}, r_{t+2}, \ldots)$ son las recompensas obtenidas en una hipotética trayectoria futura desde el instante t, la recompensa de dicha trayectoria se suele calcular como:

1. $\max_{k=0}^{\infty} r_{t+k+1}$
2. $\max_{k=0}^{\infty} \gamma^k r_{t+k+1}$, donde $\gamma \in [0, 1]$ es un factor de descuento
3. $\sum_{k=0}^{\infty} r_{t+k+1}$
4. $\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$, donde $\gamma \in [0, 1]$ es un factor de descuento

Solución: La 4

☐ Reinforcement learning from human feedback (RLHF) es una técnica popular de post-training de LLMs. Aunque RLHF se inspira en el planteamiento usual del aprendizaje por refuerzo, presenta varias diferencias esenciales. En relación con estas diferencias, indica la opción incorrecta o la última opción si las tres primeras son correctas:

1. Se usa un modelo de preferencias humanas en lugar de una función de recompensa.
2. Los estados iniciales son prompts y la "acción" consiste en completar el prompt.
3. Las recompensas se producen a nivel de respuesta, esto es, para una secuencia completa de acciones.
4. Las tres opciones anteriores son correctas.

Solución: La 4

☐ El post-training de LLMs mediante RLHF persigue la obtención de una política que maximice la recompensa esperada sin diverger excesivamente del modelo de referencia:

$$J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[r_\theta(s_t, a_t)\right] - \beta\, \mathbb{D}_{\mathrm{KL}}(\pi_{\mathrm{RL}}(\cdot \mid s_t) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid s_t))$$

donde $\beta \geq 0$ es un hiperparámetro que penaliza la divergencia Kullback-Leibler (KL) de π_{ref} respecto a π_{RL}. En relación con este planteamiento, indica la opción incorrecta o la última opción si las tres primeras son correctas:

1. La divergencia KL de π_{ref} respecto a π_{RL} es nula si y solo si ambas coinciden; si no, es positiva.
2. En general, la divergencia será menor cuanto mayor sea β.
3. Si β es muy próxima a cero, existe la posibilidad de que se produzca el fenómeno conocido como reward hacking, esto es, que la política obtenida maximice la recompensa esperada alejándose excesivamente del modelo de referencia y, en definitiva, del propósito del post-training.
4. Las tres opciones anteriores son correctas.

Solución: La 4

☐ Sean P y Q las distribuciones sobre el espacio muestral $\mathcal{X} = \{0, 1, 2\}$ siguientes:

	0	1	2
$P(x)$	$9/25$	$12/25$	$4/25$
$Q(x)$	$1/3$	$1/3$	$1/3$

En relación con las entropías relativas $\mathbb{D}_{\mathrm{KL}}(P \,\|\, Q)$ y $\mathbb{D}_{\mathrm{KL}}(Q \,\|\, P)$, indica la opción correcta:

1. $\mathbb{D}_{\mathrm{KL}}(P \,\|\, Q) < 0.9$ y $\mathbb{D}_{\mathrm{KL}}(Q \,\|\, P) < 0.9$
2. $\mathbb{D}_{\mathrm{KL}}(P \,\|\, Q) < 0.9$ y $\mathbb{D}_{\mathrm{KL}}(Q \,\|\, P) \geq 0.9$
3. $\mathbb{D}_{\mathrm{KL}}(P \,\|\, Q) \geq 0.9$ y $\mathbb{D}_{\mathrm{KL}}(Q \,\|\, P) < 0.9$
4. $\mathbb{D}_{\mathrm{KL}}(P \,\|\, Q) \geq 0.9$ y $\mathbb{D}_{\mathrm{KL}}(Q \,\|\, P) \geq 0.9$

Solución: La 2; ejemplo sacado de Wikipedia

$$\begin{aligned}
D_{\mathrm{KL}}(P \parallel Q) &= \sum_{x\in\mathcal{X}} P(x)\, \ln \frac{P(x)}{Q(x)} \\
&= \frac{9}{25} \ln \frac{9/25}{1/3} + \frac{12}{25} \ln \frac{12/25}{1/3} + \frac{4}{25} \ln \frac{4/25}{1/3} \\
&= \frac{1}{25}\left(32 \ln 2 + 55 \ln 3 - 50 \ln 5\right) \\
&\approx 0.0852996,
\end{aligned}$$

$$\begin{aligned}
D_{\mathrm{KL}}(Q \parallel P) &= \sum_{x\in\mathcal{X}} Q(x)\, \ln \frac{Q(x)}{P(x)} \\
&= \frac{1}{3} \ln \frac{1/3}{9/25} + \frac{1}{3} \ln \frac{1/3}{12/25} + \frac{1}{3} \ln \frac{1/3}{4/25} \\
&= \frac{1}{3}\left(-4 \ln 2 - 6 \ln 3 + 6 \ln 5\right) \\
&\approx 0.097455.
\end{aligned}$$

☐ En general, el post-training pionero de InstructGPT, precursor de ChatGPT, se reduce a entrenar tres modelos:

1. Modelo SFT: entrenado a partir del base (GPT-3) con unos 10K ejemplos de instrucciones humanas.
2. Modelo de recompensa: entrenado a partir del SFT con unos 100K prompts de preferencia humana, esto es, acompañados de dos o más respuestas ordenadas según preferencias humanas.
3. Modelo alineado: entrenado a partir del SFT y el de recompensa, con optimización PPO y reusando prompts.

Sin embargo, el post-training de LLMs actual presenta diferencias considerables respecto al de InstructGPT. En relación con las mismas, indica la opción incorrecta o la última opción si las tres primeras son correctas:

1. Se emplean más ejemplos de instrucciones para entrenar el modelo SFT, tanto humanas como sintéticas.
2. Se usan más prompts de preferencia en rondas sucesivas de entrenamiento de modelos de recompensa y alineados.
3. Se usa un LLM juez como proxy humano cuando no se dispone de preferencias humanas.
4. Las tres opciones anteriores son correctas.

Solución: La 4

---

3 Modelo de recompensa (RM)

☐ Sean un prompt x y dos respuestas al mismo, y_1 e y_2, dadas por un modelo $\pi(y \mid x)$. Sean $r_1 = r^*(x, y_1)$ y $r_2 = r^*(x, y_2)$ las recompensas que otorgamos a dichas respuestas y sea $P = p^*(y_1 \succ y_2 \mid x)$ la probabilidad de preferir y_1 a y_2 según el modelo Bradley-Terry. En relación con esta probabilidad, indica la opción incorrecta o la última opción si las tres primeras son correctas:

1. $P = \frac{e^{r_1}}{e^{r_1}+e^{r_2}}$
2. $P = \frac{1}{1+e^{r_2-r_1}}$
3. $P = \sigma(r_1 - r_2)$
4. Las tres opciones anteriores son correctas.

Solución: La 4

☐ Sea un modelo de recompensa parametrizado, $r_\theta(x, y) = \theta\,|y|$, donde $\theta \in \mathbb{R}$ y $|y|$ denota la longitud de la respuesta y dada por un LLM a un prompt x. Considera el modelo Bradley-Terry basado en $r_\theta(x, y)$, $p_\theta(y_1 \succ y_2 \mid x)$. En relación con el mismo, indica la opción correcta o la última opción si ninguna las tres primeras es correcta:

1. $p_\theta(y_1 \succ y_2 \mid x) = \sigma(|y_2| + |y_1|)$
2. $p_\theta(y_1 \succ y_2 \mid x) = \sigma(\theta(|y_2| - |y_1|))$
3. $p_\theta(y_1 \succ y_2 \mid x) = \sigma(\theta(|y_2| + |y_1|))$
4. Ninguna de las tres opciones anteriores es correcta.

Solución: La 4; es $p_\theta(y_1 \succ y_2 \mid x) = \sigma(\theta(|y_1| - |y_2|))$

---

4 Instruction fine-tuning (IFT)

☐ El primer paso del post-training de LLMs es el instruction / supervised fine-tuning (IFT/SFT). Típicamente, se trata de adaptar un modelo base para que se comporte como un asistente (rol `assistant`) conversacional de un usuario (rol `user`). El modelo debe responder a un prompt con una consulta del usuario, posiblemente precedida por los mensajes intercambiados entre usuario y asistente durante turnos conversacionales previos. Con el fin de identificar roles y separar mensajes, se usan tokens especiales de acuerdo con algún `chat template`. Aunque IFT es técnicamente muy parecido al pre-entrenamiento de modelos base, presenta algunas diferencias sustanciales. En relación con las mismas, indica la opción incorrecta o la última opción si las tres primeras son correctas:

1. En IFT se emplean muchos menos datos de entrenamiento que en pre-entrenamiento.
2. Dado un par prompt-respuesta de entrenamiento, los tokens del prompt se tienen en cuenta para predecir los de la respuesta (en el paso forward), pero se excluyen de la pérdida (y paso backward) ya que el modelo adaptado debe aprender respuestas y no prompts.
3. Por lo general se emplean tamaños de batch menores que en pre-entrenamiento.
4. Las tres opciones anteriores son correctas.

Solución: La 4

---

5 Rejection sampling (RS)

☐ Rejection sampling es una técnica de preference fine-tuning (PreFT) sencilla y popular. En general, se trata de hacer instruction / supervised fine-tuning (IFT/SFT) con respuestas seleccionadas por un modelo de recompensa de entre múltiples respuestas generadas por el modelo en adaptación (u otros). Dados M prompts y N respuestas generadas por prompt, los criterios usuales de selección se basan en una matriz de recompensas $R \in \mathbb{R}^{N \times M}$, donde r_{ij} la recompensa obtenida por la respuesta j al prompt i. En relación con estos criterios, indica la opción incorrecta o la última opción si las tres primeras son correctas:

1. Top-K Per Prompt escoge las K mejores por prompt.
2. Best-of-N Sampling equivale a top-1 por prompt.
3. Top Overall Prompts escoge las K mejores globalmente.
4. Las tres opciones anteriores son correctas.

Solución: La 4

---

6 Direct preference optimization (DPO)

☐ Direct preference optimization (DPO) es una técnica de preference fine-tuning (PreFT) relativamente sencilla ya que no requiere entrenar un modelo de recompensa explícito para alinear un modelo de referencia dado, $\pi_{\text{ref}}(y \mid x)$, mediante RLHF. Dado un dataset de preferencias $\mathcal{D} = \{(x^{(i)}, y_w^{(i)}, y_l^{(i)})\}$, DPO se deriva a partir del objetivo RL usual, esto es, la búsqueda de una política óptima asociada a un modelo de recompensa dado, $r(x, y)$,

$$\pi_r(y \mid x) = \underset{\pi}{\text{argmax}}\ \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi(y|x)}\left[r(x,y)\right] - \beta\, \mathbb{D}_{\text{KL}}(\pi(y \mid x) \,||\, \pi_{\text{ref}}(y \mid x))$$

donde $\beta \geq 0$ es un hiperparámetro que penaliza la divergencia Kullback-Leibler (KL) de π_{ref} respecto a π_{RL}. Se puede comprobar que una política óptima asociada a $r(x, y)$, debe ser de la forma

$$\pi_r(y \mid x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y \mid x)\ \exp\left(\frac{1}{\beta} r(x, y)\right)$$

donde Z es la **función partición**, independiente de π, pero costosa de estimar,

$$Z(x) = \sum_y \pi_{\text{ref}}(y \mid x)\ \exp\left(\frac{1}{\beta}\, r(x, y)\right)$$

Además, $r(x, y)$, puede expresarse en función de $\pi_r(y \mid x)$, como sigue:

1. $r(x,y) = \beta \log \dfrac{\pi_{\text{ref}}(y \mid x)}{\pi_r(y \mid x)} + \beta\, \log Z(x)$
2. $r(x,y) = \beta \log \dfrac{\pi_r(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta\, \log Z(x)$
3. $r(x,y) = \beta \log \dfrac{\pi_{\text{ref}}(y \mid x)}{\pi_r(y \mid x)}$
4. $r(x,y) = \beta \log \dfrac{\pi_r(y \mid x)}{\pi_{\text{ref}}(y \mid x)}$

Solución: La 2

☐ Sea $r_\theta(x, y)$ un modelo de recompensa con modelo de preferencias Bradley-Terry asociado

$$p_\theta(y_1 \succ y_2 \mid x) = \frac{\exp(r_\theta(x, y_1))}{\exp(r_\theta(x, y_1)) + \exp(r_\theta(x, y_2))}$$

DPO expresa Bradley-Terry en función de la política óptima asociada a $r_\theta(x, y)$, $\pi_\theta(y \mid x)$, como

$$p_\theta(y_1 \succ y_2 \mid x) = \sigma(\mu_\theta(y_1 \succ y_2 \mid x))$$

donde $\mu_\theta(y_1 \succ y_2 \mid x)$ es:

1. $\beta \log \dfrac{\pi_{\text{ref}}(y_1 \mid x)}{\pi_\theta(y_1 \mid x)} - \beta \log \dfrac{\pi_{\text{ref}}(y_2 \mid x)}{\pi_\theta(y_2 \mid x)}$
2. $\beta \log \dfrac{\pi_{\text{ref}}(y_2 \mid x)}{\pi_\theta(y_2 \mid x)} - \beta \log \dfrac{\pi_{\text{ref}}(y_1 \mid x)}{\pi_\theta(y_1 \mid x)}$
3. $\beta \log \dfrac{\pi_\theta(y_1 \mid x)}{\pi_{\text{ref}}(y_1 \mid x)} - \beta \log \dfrac{\pi_\theta(y_2 \mid x)}{\pi_{\text{ref}}(y_2 \mid x)}$
4. $\beta \log \dfrac{\pi_\theta(y_2 \mid x)}{\pi_{\text{ref}}(y_2 \mid x)} - \beta \log \dfrac{\pi_\theta(y_1 \mid x)}{\pi_{\text{ref}}(y_1 \mid x)}$

Solución: La 3

☐ Dado un dataset de preferencias $\mathcal{D} = \{(x^{(i)}, y_w^{(i)}, y_l^{(i)})\}$, DPO minimiza el riesgo empírico con log-pérdida según un modelo Bradley-Terry parametrizado a través de su política óptima π_θ

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \ \log \sigma(\mu_\theta(y_w \succ y_l \mid x))$$

donde la logodds se expresa directamente en función de π_θ y el modelo de referencia,

$$\mu_\theta(y_w \succ y_l \mid x) = \beta((\log \pi_\theta(y_w \mid x) - \log \pi_\theta(y_l \mid x)) - (\log \pi_{\text{ref}}(y_w \mid x) - \log \pi_{\text{ref}}(y_l \mid x)))$$

y $\beta \geq 0$ es un hiperparámetro que penaliza la divergencia Kullback-Leibler (KL) de π_{ref} respecto a π_θ. El gradiente de la pérdida DPO, $\nabla_\theta \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}})$, es:

1. $-\beta \, \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\sigma \left(\mu_\theta(y_w \succ y_l \mid x) \right) \left(\nabla_\theta \log \pi(y_w \mid x) - \nabla_\theta \log \pi(y_l \mid x) \right) \right]$
2. $-\beta \, \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\sigma \left(\mu_\theta(y_w \succ y_l \mid x) \right) \left(\nabla_\theta \log \pi(y_l \mid x) - \nabla_\theta \log \pi(y_w \mid x) \right) \right]$
3. $-\beta \, \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\sigma \left(\mu_\theta(y_l \succ y_w \mid x) \right) \left(\nabla_\theta \log \pi(y_w \mid x) - \nabla_\theta \log \pi(y_l \mid x) \right) \right]$
4. $-\beta \, \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\sigma \left(\mu_\theta(y_l \succ y_w \mid x) \right) \left(\nabla_\theta \log \pi(y_l \mid x) - \nabla_\theta \log \pi(y_w \mid x) \right) \right]$

Solución: La 3
