Credit Fraud Detector

Based on RLHF 2023/06/29

Introduction

上次使用 Epsilon-Greedy 時有遇到兩個問題:

- 1. 大部分 Explore & exploitation 方法只會取最大那群。(ex: Epsilon-Greedy 只會取 argmax)
- 2. 沒有讓模型參與到打分數的環節。



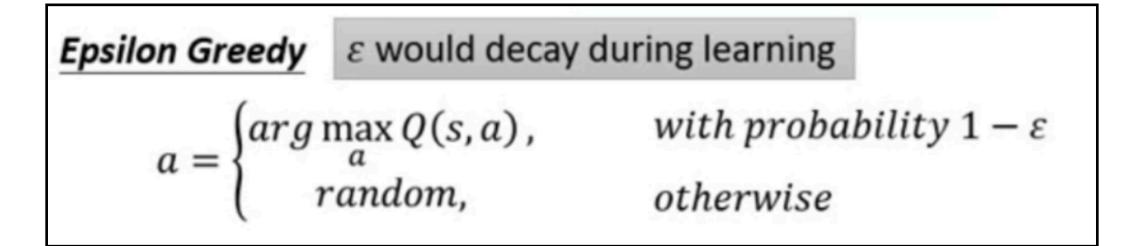
- 3. 將資料轉回 PCA 前的資料(28維度),加回閾值內資料的資料集,訓練模型查看結果
- 4. 將Epsilon乘上decay,並更新機率表如下:

Index	0	1	2	3	4	5	6	7
P	0.0	0.0	0.8	0.0	0.0	0.0	0.4	0.0

Q: Epsilon-Greedy中,會用argmax 選擇最大的機率,代表第六群會被忽視。 第二群全部抽完的機率也有86%,所以理論上會等到完全抽完第二群才抽第六群。

上禮拜的 ppt

Introduction



上次使用 Epsilon-Greedy 時有遇到兩個問題:

- 1. 大部分 Explore & exploitation 方法只會取最大那群。(ex: Epsilon-Greedy 只會取 argmax)
- 2. 沒有讓模型參與到打分數的環節。

8 - Clusters



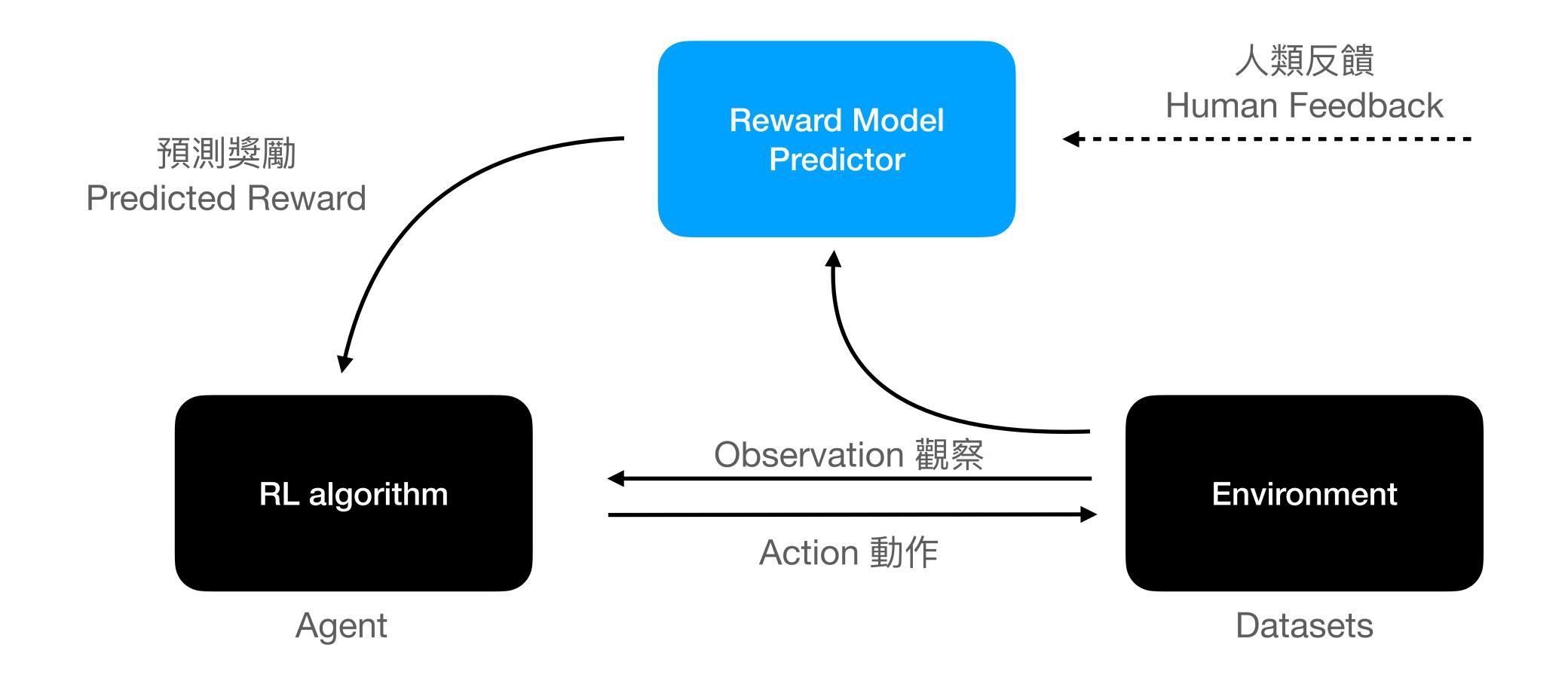
- 3. 將資料轉回 PCA 前的資料(28維度),加回閾值內資料的資料集,訓練模型查看結果
- 4. 將Epsilon乘上decay,並更新機率表如下:

Index	0	1	2	3	4	5	6	7
P	0.0	0.0	8.0	0.0	0.0	0.0	0.4	0.0

Q: Epsilon-Greedy中,會用argmax 選擇最大的機率,代表第六群會被忽視。 第二群全部抽完的機率也有86%,所以理論上會等到完全抽完第二群才抽第六群。

RLHF: 從人類反饋中學習

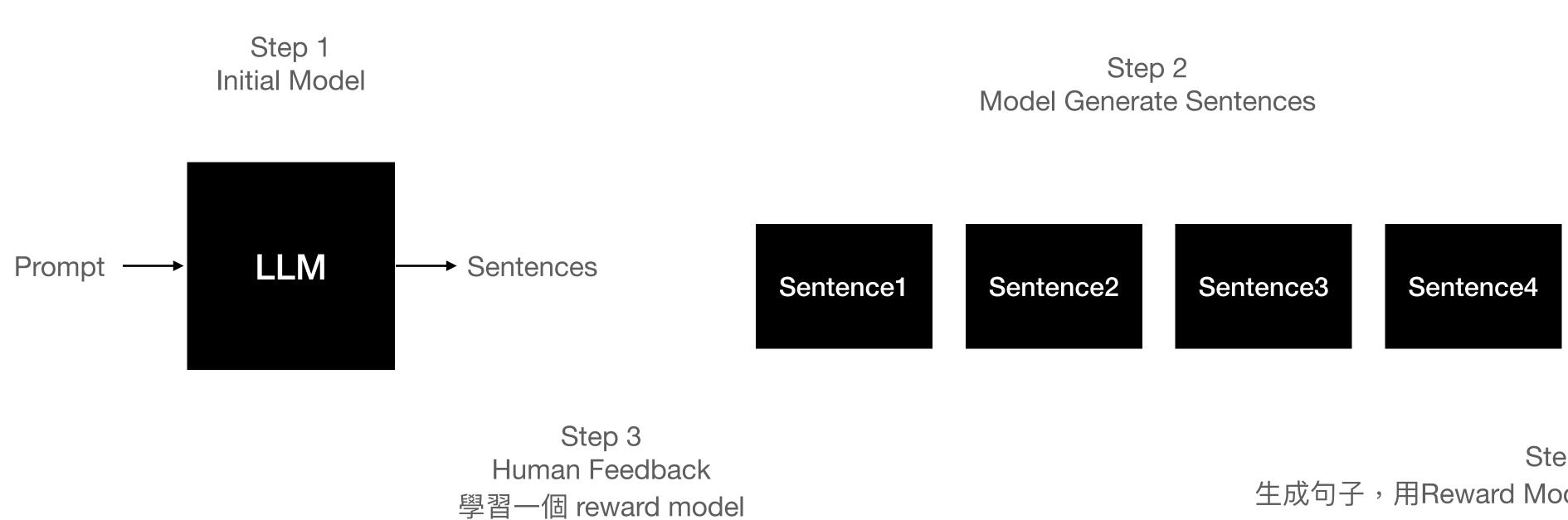
通過人類標注數據訓練得到 Reward Model (相當於是去學人類怎麼標注資料) ,有了 Reward Model後,就可以使用一般的強化學習方法去找出最佳策略。

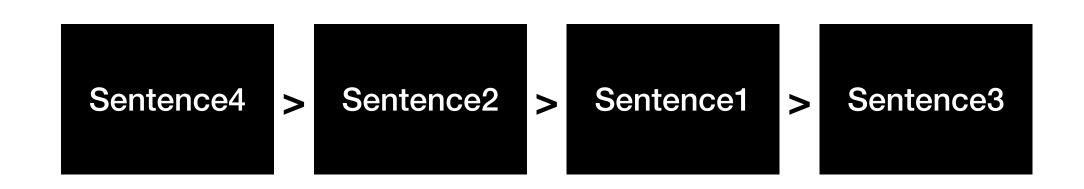


RLHF: ChatGPT

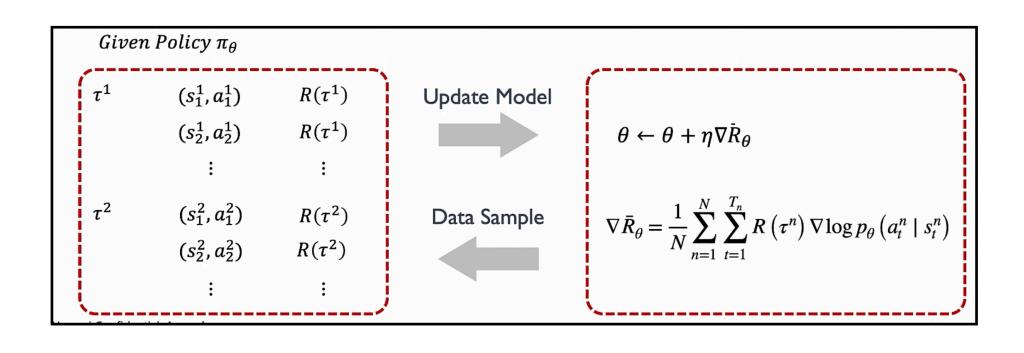
Step 2 Step 1 Step 3 Collect demonstration data Collect comparison data and Optimize a policy against the train a reward model. and train a supervised policy. reward model using the PPO reinforcement learning algorithm. A prompt is A prompt and A new prompt is () sampled from our several model sampled from Write a story Explain reinforcement Explain reinforcement prompt dataset. the dataset. outputs are learning to a 6 year old. about otters. learning to a 6 year old. sampled. B In reinforcement learning, the The PPO model is A labeler initialized from the C demonstrates the We give treats and In machine supervised policy. desired output We give treats and behavior. punishments to teach... The policy generates Once upon a time... A labeler ranks the an output. outputs from best to worst. D > G > A > B This data is used to The reward model fine-tune GPT-3.5 calculates a reward with supervised for the output. learning. This data is used to train our The reward is used reward model. to update the D > C > A > B policy using PPO.

RLHF: ChatGPT

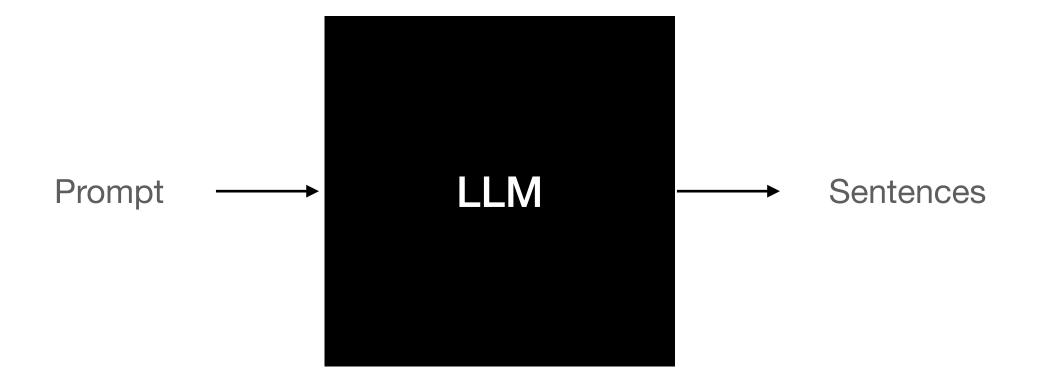




Step 4 生成句子,用Reward Model打分數,更新模型權重



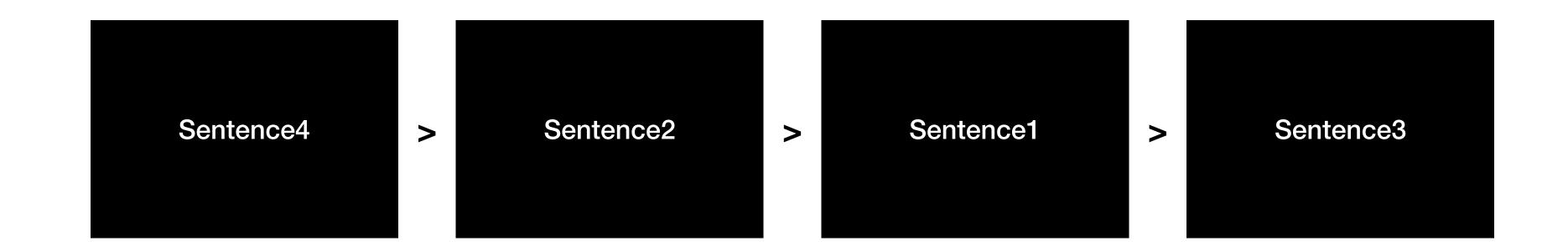
Step 1 Initial Model



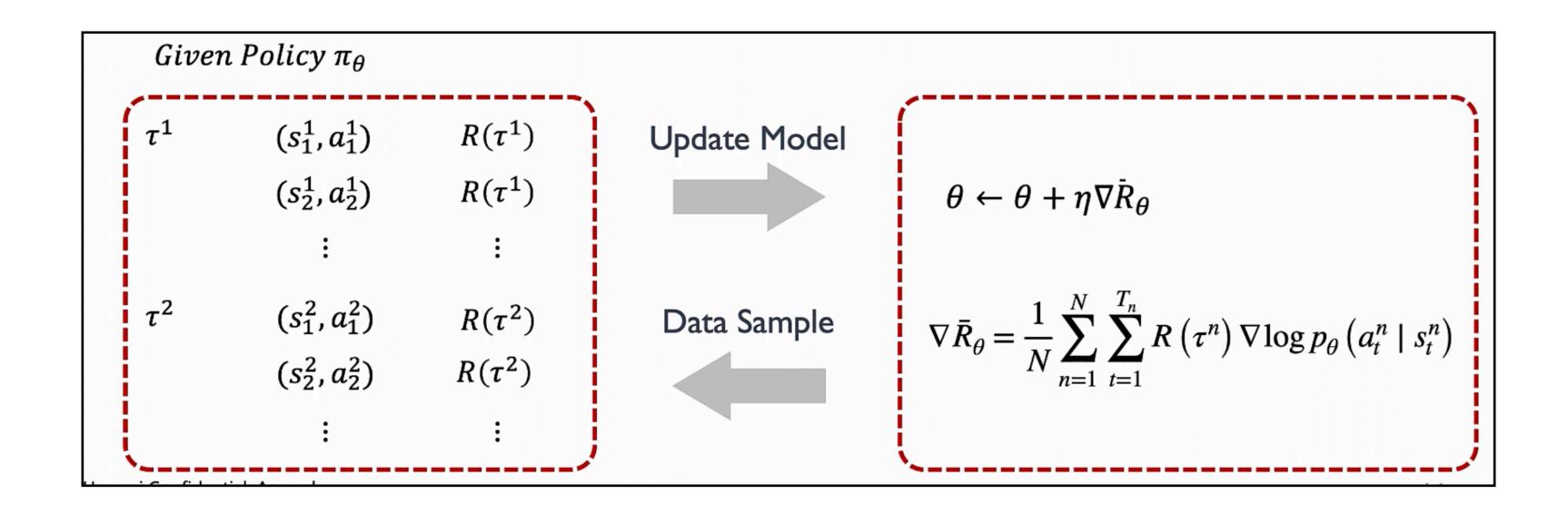
Step 2 Model Generate Sentences



Step 3 Human Feedback

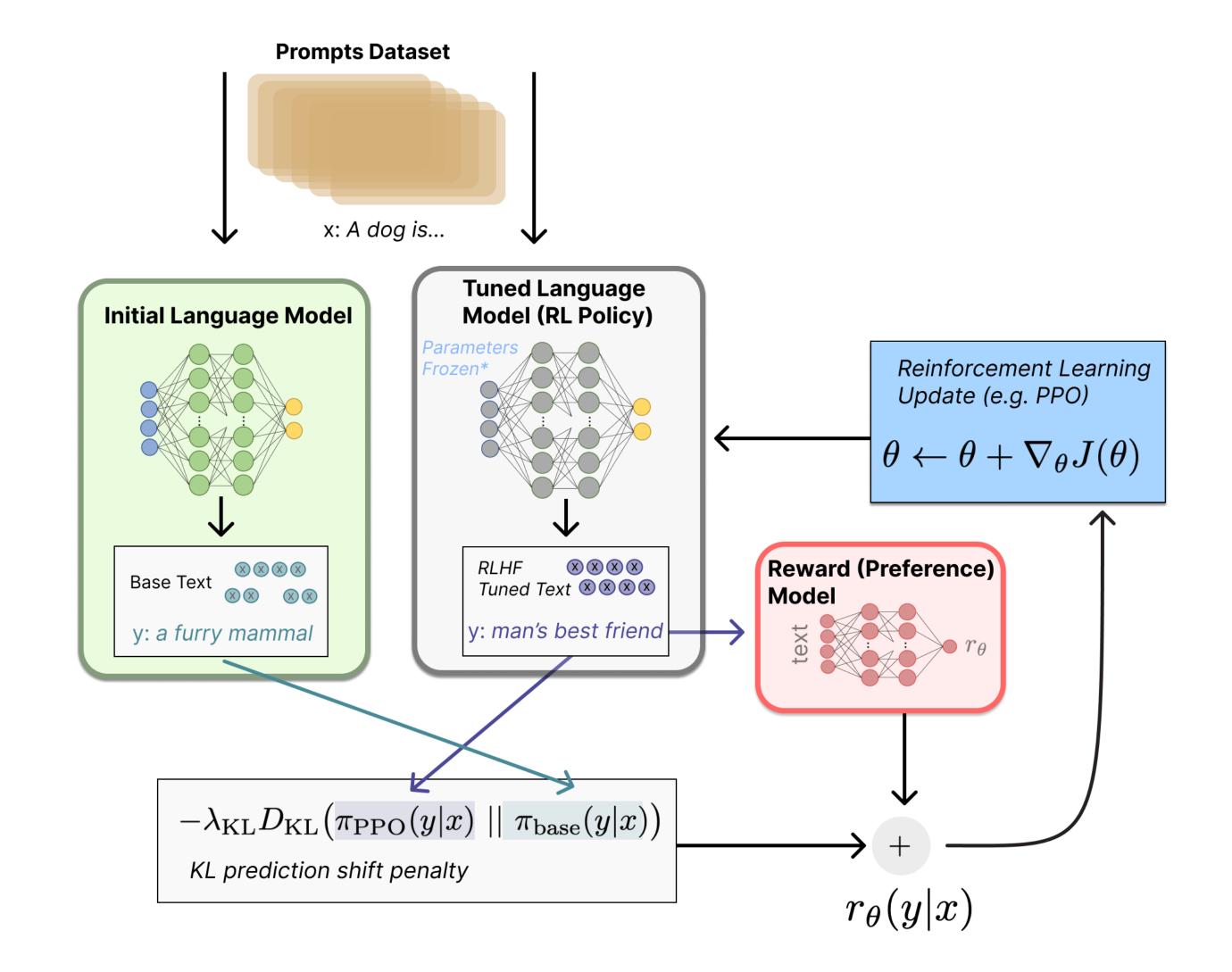


Step 4 更新模型權重

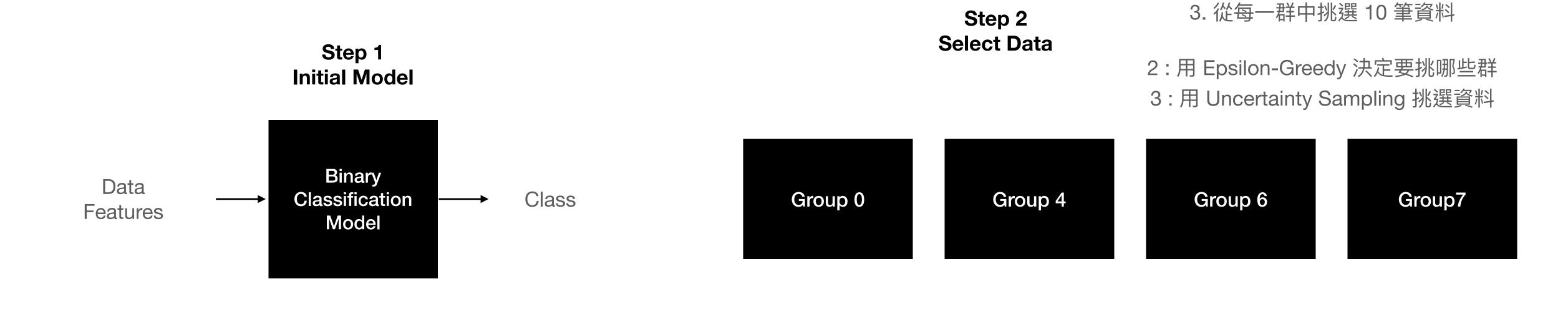


在實際環境中,讓 Agent 和 Environment互動,產生一系列採樣數據。 即獲得很多的 (s, a) 和R(t) 的 Pair 。(在狀態 s 下,採取動作 a 得到的獎勵 R(t)) 最後將數據送到訓練過程中計算,更新模型的參數 theta。

Step 4 更新模型權重



RLHF: OUR



作法

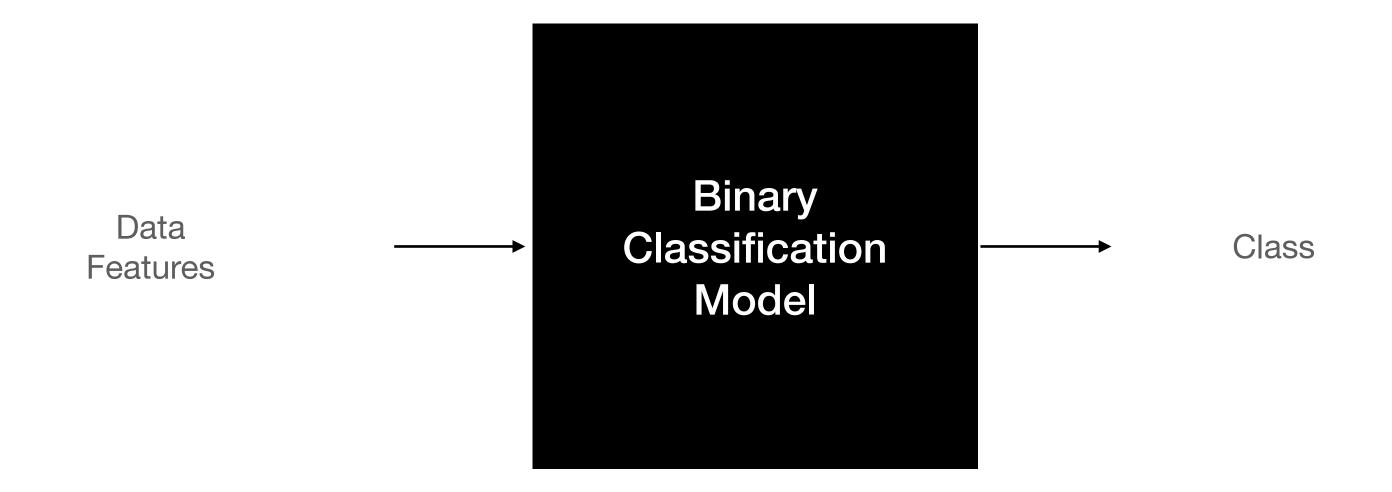
1. PCA + Cluster 分成一些群

2. 先挑選一些群

Step 4 Step 3 Re-train model 更新模型權重 **Human Feedback Training Reward Model** Binary Class Group 6 Group 4 Group 6 Group7 Classification **Data Features** (Label by Reward Model) Model 70分 5分 0分 30 分

以人工的方式標注這些資料, 首先更新機率表,其次為每一群打分數

Step 1 Initial Model



Step 2
Select Data

Group 0

Group 4

Group 6

Group7

作法

- 1. PCA + Cluster 分成一些群
 - 2. 先挑選一些群
- 3. 從每一群中挑選 10 筆資料

舉例來說:

目前我們根據機率表選到 group2, group4, group6, group7。 我們會在這四群裡面各選10筆具有重要性的資料出來。

Exploration-Exploitation

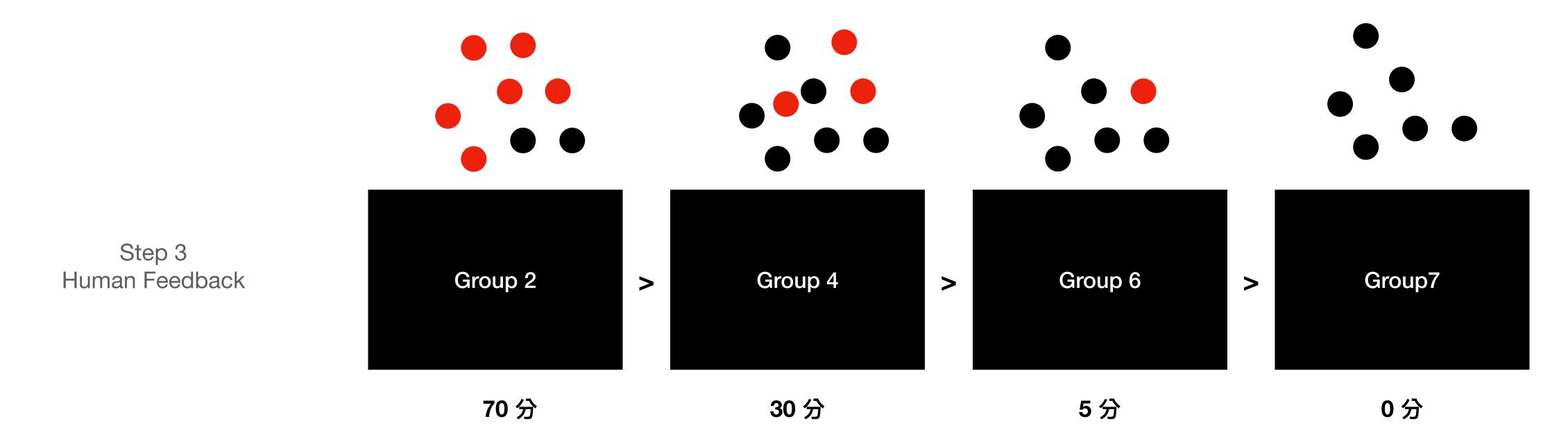
- 1. 朴素bandit
- 2. Thompson sampling
- 3. Upper Confidence Bound
- 4. Epsilon-Greedy

Uncertainty Sampling

- 1. Random
- 2. Least Confident
- 3. Margin Sampling
- 4. Entropy

2:用 Exploration-Exploitation 決定要挑哪些群

3:用 Uncertainty Sampling 挑選資料



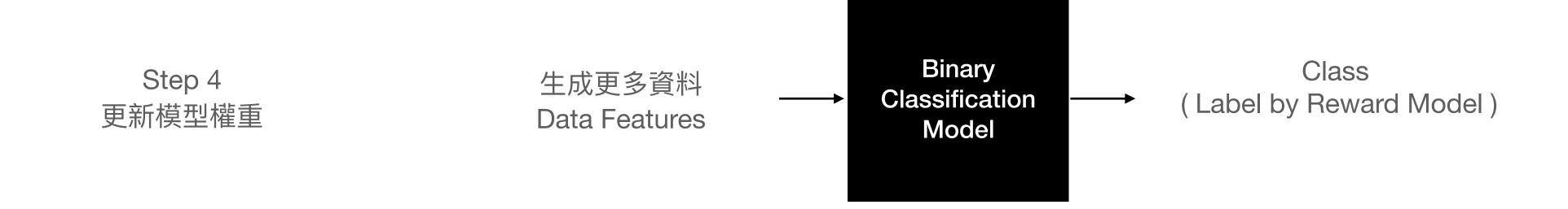
以人工的方式標注這些資料,首先更新機率表,其次為每一群打分數。

舉例來說:

把每一個 group 挑出來的資料當成是一個句子要打分數,裡面有幾個 Fraud data 就加幾分。 這些標注的資料會幫助我們更新機率表,並且用於訓練 Reward Model。

Generative Model 來作為 Reward Model。

將剛剛選到的其中 20 筆資料當作 Test ,來判斷何時停止訓練 Reward Model。



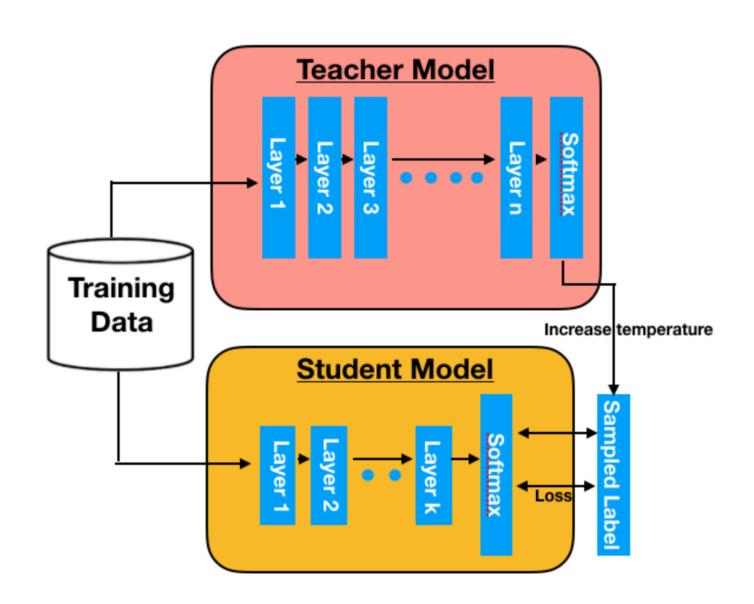
當 Reward Model 被訓練好,就可以去生成資料來訓練模型。

舉例來說:

假設標注到 1000筆時,發現Reward Model 效果不錯已經學會打分數, 這時就使用 Reward Model 對剩下的 Unlabeled 資料加上標籤。 把這些資料拿去給最終模型 Train,查看效果。(由於這個資料集有所有資料的label)

Knowledge Distillation 知識蒸餾

考量到銀行可能需要較簡單的模型,因此最後可以把我們訓練的大模型轉換到小模型上。



知識蒸餾是抽取複雜模型訓練出的精華給小模型,讓這個小的簡單模型也能達到跟複雜模型一樣的效果。

其中 student 模型從可以更複雜的 teacher 模型中 「學習」 。如果已經透過複雜的結構建構出不錯的模型,可以用知識蒸餾訓練出較簡易版本的模型,準確度不會差太多。