

## 实验报告：强化学习实践

### 一、实验目的

理解强化学习基本原理，实现 Q-learning 算法在 Gym 环境中的应用。

### 二、实验内容

1. 使用 Gym 的 CartPole-v1 环境
2. 实现 Q-learning（离散化状态空间）
3. 观察智能体随训练逐步学会平衡杆子

### 三、实验步骤与完整代码

```
import numpy as np
import gymnasium as gym
import matplotlib.pyplot as plt

# 创建环境
env = gym.make('CartPole-v1')

# 状态离散化设置
buckets = (1, 1, 6, 3)
upper_bounds = [env.observation_space.high[0], 0.5, env.observation_space.high[2],
np.radians(50)]
lower_bounds = [env.observation_space.low[0], -0.5, env.observation_space.low[2], -
np.radians(50)]

def discretize(obs):
    ratios = [(obs[i] - lower_bounds[i]) / (upper_bounds[i] - lower_bounds[i]) for i in
range(len(obs))]
    new_obs = [int(round((buckets[i] - 1) * ratios[i])) for i in range(len(obs))]
    new_obs = [min(buckets[i] - 1, max(0, new_obs[i])) for i in range(len(obs))]
    return tuple(new_obs)

# Q-learning 参数设置
q_table = np.zeros(buckets + (env.action_space.n,))
alpha = 0.1
gamma = 0.99
epsilon = 1.0
epsilon_decay = 0.995
epsilon_min = 0.01
episodes = 1000
max_steps = 200
rewards = []

# Q-learning 主循环
for episode in range(episodes):
    state = discretize(env.reset()[0])
    total_reward = 0
```

```

for t in range(max_steps):
    #  $\epsilon$ -greedy 策略选择动作
    if np.random.random() < epsilon:
        action = env.action_space.sample()
    else:
        action = np.argmax(q_table[state])

    # 执行动作
    obs, reward, done, _, _ = env.step(action)
    next_state = discretize(obs)

    # Q 值更新
    q_table[state][action] += alpha * (reward + gamma * np.max(q_table[next_state])
    - q_table[state][action])

    state = next_state
    total_reward += reward

    if done:
        break

    #  $\epsilon$  衰减
    epsilon = max(epsilon_min, epsilon * epsilon_decay)
    rewards.append(total_reward)

    if episode % 100 == 0:
        print(f"Episode {episode}, Total Reward: {total_reward}")

# 可视化训练结果
plt.plot(rewards)
plt.xlabel('Episode')
plt.ylabel('Total Reward')
plt.title('Q-learning on CartPole-v1')
plt.grid(True)
plt.show()

env.close()

```

#### 四、实验结果与分析

训练初期奖励较低 (<50)，随着训练轮次增加，奖励逐渐上升，约在 300 轮后稳定在 200（环境最大步数），表明智能体已学会长时间平衡杆子。奖励曲线呈典型强化学习趋势：探索→利用→收敛。离散化方法虽然简化了问题，但损失了精度；深度 Q 网络（DQN）可直接处理连续状态空间，效果更优。

#### 五、思考题

探索 (Exploration) 与利用 (Exploitation) 如何平衡?

答: 常用  $\epsilon$ -greedy 策略: 以概率  $\epsilon$  随机探索, 以  $1-\epsilon$  选择当前最优动作。训练初期  $\epsilon$  较高以充分探索环境, 后期逐渐衰减以聚焦最优策略的执行。

深度强化学习相比传统 RL 有哪些优势?

答: 传统 RL (如 Q-table) 无法处理高维或连续状态空间; 深度 RL 使用神经网络近似 Q 函数或策略, 可泛化到未见过的状态, 适用于 Atari 游戏、机器人控制等复杂任务。

## 六、实验总结

本实验成功实现了基于 Q-learning 的 CartPole 平衡控制任务, 掌握了状态离散化、Q 值更新、 $\epsilon$ -greedy 策略等强化学习核心概念。通过奖励曲线的变化, 直观观察到智能体从随机探索到掌握平衡策略的学习过程, 验证了强化学习在控制任务中的有效性。