作业报告——多臂老虎机2

方鸿宇 2001213098

算法说明

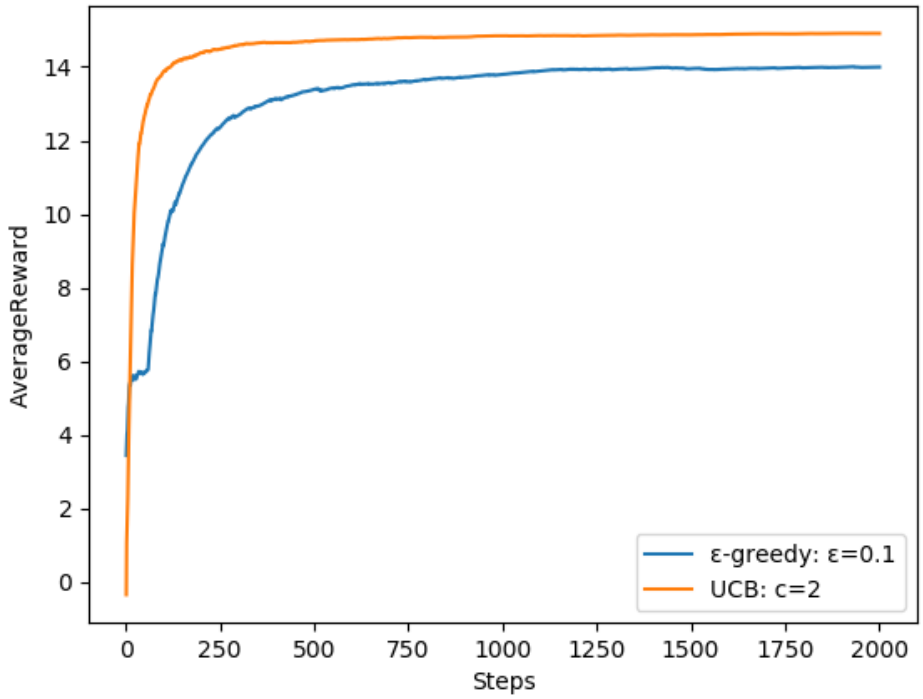
本算法实现了对多臂老虎机任务的策略学习。在训练过程对每个臂的平均奖金进行统计，并使用$\epsilon - greedy$和UCB两种动作策略进行训练。

实验设置

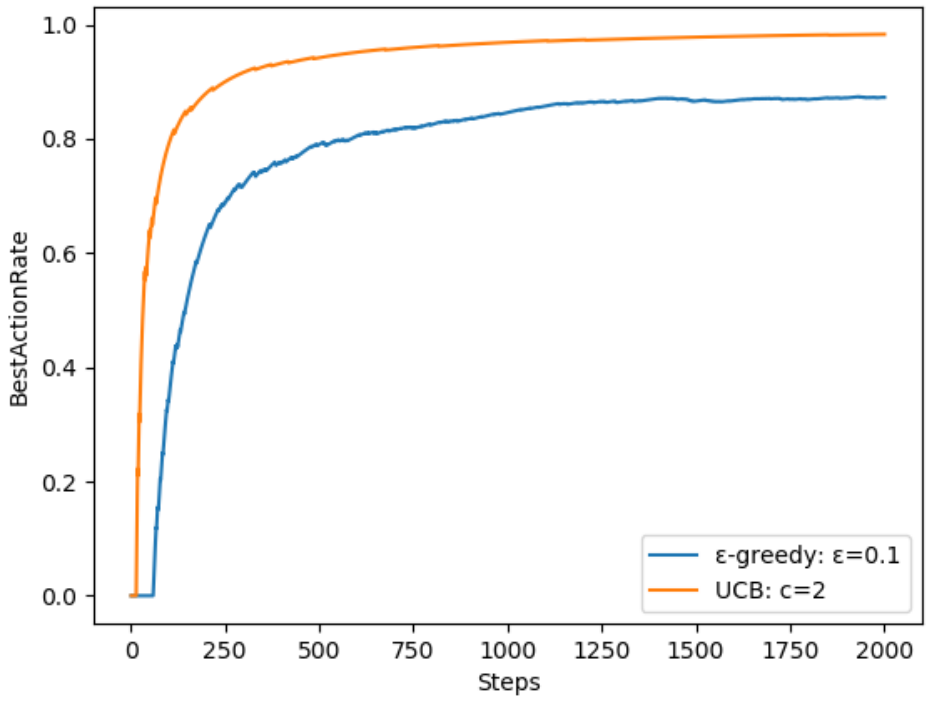
实验中设计了臂数为15的多臂老虎机，每个臂拉一次的奖金服从高斯分布，每个臂的平均奖金为1到15的一个随机数，高斯标准差为1。对于$\epsilon - greedy$实验，将ϵ设为0.1；对于UCB实验，将c设为2。实验尝试次数为2000。

实验结果

下图展示了平均奖金与训练步数的关系，可见使用UCB策略能够更快达到收敛，并收敛至更高水平。



为进一步展示UCB策略快速收敛的原因，本实验中对训练过程中的最优动作比率进行统计，并可视化。最优动作比率指训练过程中使用最优老虎机摇臂的次数占总次数的比重。统计曲线如下



相对于$\epsilon-greedy$策略，UCB策略在训练初期优先对尝试次数少的摇臂进行尝试，比起$\epsilon-greedy$的随机探索策略，这种探索策略使得模型更快找到奖励值最高的摇臂。在训练后期，$\epsilon-greedy$策略仍保持ϵ的概率探索非最佳摇臂，而UCB几乎一直选择最佳摇臂，因此前者的最优动作比率明显小于1，而后者不断趋向于1，使得后者的平均回报明显高于前者。

代码说明

代码见`code`文件夹，包含`bandit.py`、`visualize_average_reward.py`和`visualize_best_action_rate.py`三个代码文件。

`bandit.py`进行模型训练，运行时按提示在命令行中输入1或2以选择不同的动作策略，运行过程中的数据以

```
BanditID_1 Reward_1
BanditID_2 Reward_2
...
```

的格式存储。对于$\epsilon-greedy$实验，数据存储于`log/epsX.txt`文件中，文件名中的X表示$\epsilon-greedy$算法中的ϵ值；对于UCB实验，数据存储于`log/UCBX.txt`文件中，文件名中的X表示UCB算法中的参数c。

`visualize_average_reward.py`进行数据可视化，训练过程中的数据以"平均回报——步数"曲线表示。平均回报使用了全局平均回报，即模型运行过程中所有尝试所获得的回报的平均值。曲线图保存为`log/log_averageReward.png`文件中。

`visualize_best_action_rate.py`进行数据可视化，训练过程中的数据以"最优动作比率——步数"曲线表示。最优动作比率表示选取最优老虎机摇臂的次数占总次数的比重。曲线图保存为`log/log_bestActionRate.png`文件中。

代码运行方式

进入`code`文件夹，依次执行如下命令即可：

```
$ python3 bandit.py
[output]
choose a exploration strategy from:
1. epsilon-greedy
2. UCB
[input]
1 # for epsilon-greedy, or 2 for UCB
$ python3 visualize_average_reward.py
$ python3 visualize_best_action_rate.py
```