计算机科学188：人工智能导论
2024年春季 笔记25

作者（其他所有笔记）：尼基尔·夏尔马

作者（贝叶斯网络笔记）：乔希·胡格和杰基·梁，由王瑞吉娜编辑

作者（逻辑笔记）：亨利·朱，由考佩林编辑

致谢（机器学习与逻辑笔记）：部分内容改编自教材《人工智能：一种现代方法》。

最后更新时间：2023年8月26日

探索与利用

我们现在已经介绍了几种不同的方法，让智能体学习最优策略，并且一直强调“充分探索”对于此过程是必要的，但没有真正详细说明“充分”到底意味着什么。在接下来的两节中，我们将讨论两种在探索和利用之间分配时间的方法：ε - 贪婪策略和探索函数。

ε - 贪婪策略

遵循ε-贪婪策略的智能体定义某个概率ε，并以概率ε随机行动并进行探索。相应地，它们以概率1 - ε遵循当前已确立的策略并进行利用。这是一个非常简单的可实施策略，但仍然可能相当难以处理。如果选择一个较大的ε值，那么即使在学习到最优策略之后，智能体仍将大多随机行动。类似地，选择一个较小的ε值意味着智能体很少进行探索，导致Q学习（或任何其他选定的学习算法）非常缓慢地学习到最优策略。为了解决这个问题，必须手动调整ε并随着时间的推移降低它以看到效果。

探索函数

探索函数避免了手动调整ε这个问题，探索函数使用修改后的Q值迭代更新来对访问较少的状态给予一些偏好。修改后的更新如下：

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha \cdot [R(s,a,s') + \gamma \max_{a'} f(s',a')]$$

其中 f 表示一个探索函数。在设计探索函数时存在一定程度的灵活性，但常见的选择是使用

$$f(s,a) = Q(s,a) + \frac{k}{N(s,a)}$$

其中 k 是某个预先确定的值， $N(s,a)$ 表示 Q 状态 (s,a) 被访问的次数。处于状态 s 的智能体总是从每个状态中选择具有最高 $f(s,a)$ 的动作，因此永远不必在探索和利用之间做出概率性决策。

相反，探索由探索函数自动编码，因为术语 $\frac{k}{N(s,a)}$ 可以给一些不常采取的动作足够的“奖励”，使得这些动作被选中，而不是选择具有更高Q值的动作。随着时间的推移，状态被更频繁地访问，每个状态的这个奖励会朝着0减少，并且 $f\left(s,a\right)$ 会朝着 $Q\left(s,a\right)$ 回归，使得利用变得越来越占主导。

总结

记住强化学习有一个潜在的马尔可夫决策过程（MDP）非常重要，并且强化学习的目标是通过推导最优策略来解决这个MDP。使用强化学习与使用诸如值迭代和策略迭代等方法的区别在于缺乏对潜在MDP的转移函数 T 和奖励函数 R 的了解。因此，智能体必须通过在线试错来学习最优策略，而不是通过纯离线计算。有很多方法可以做到这一点：

- 基于模型的学习——运行计算以估计转移函数 T 和奖励函数 R 的值，并使用诸如值迭代或策略迭代等MDP求解方法结合这些估计值。
- 无模型学习——避免估计 T 和 R ，而是使用其他方法直接估计状态的值或Q值。
  - 直接评估——遵循策略 π ，简单地计算从每个状态获得的总奖励以及每个状态被访问的总次数。如果获取了足够的样本，这将收敛到在 π 下状态的真实值，尽管速度较慢且浪费了关于状态之间转换的信息。
  - 时序差分学习——遵循策略 π ，并使用指数移动平均值结合采样值，直到收敛到在 π 下状态的真实值。TD学习和直接评估是在线学习的示例，它们在确定特定策略是否次优并需要更新之前，先学习该策略的值。
  - Q学习 - 通过Q值迭代更新直接通过试错学习最优策略。这是离策略学习的一个例子，即使采取次优行动也能学习到最优策略。
  - 近似Q学习 - 与Q学习做相同的事情，但使用基于特征的状态表示来进行泛化学习。
- 为了量化不同强化学习算法的性能，我们使用遗憾的概念。遗憾捕捉了如果我们从一开始就在环境中最优地行动所积累的总奖励与通过运行学习算法所积累的总奖励之间的差异。