强化学习基础

常用随机策略

1、 贪婪策略

$$\pi_*(a|s) = \begin{cases} 1 & if \quad a = argmaxq_*(s,a) \\ 0 & \text{其他} \end{cases}$$

贪婪策略是一个确定性的策略、只在动作值函数最大的动作以概率为1的几率选择，其他动作以概率0选择。

2、$\epsilon - greedy$策略

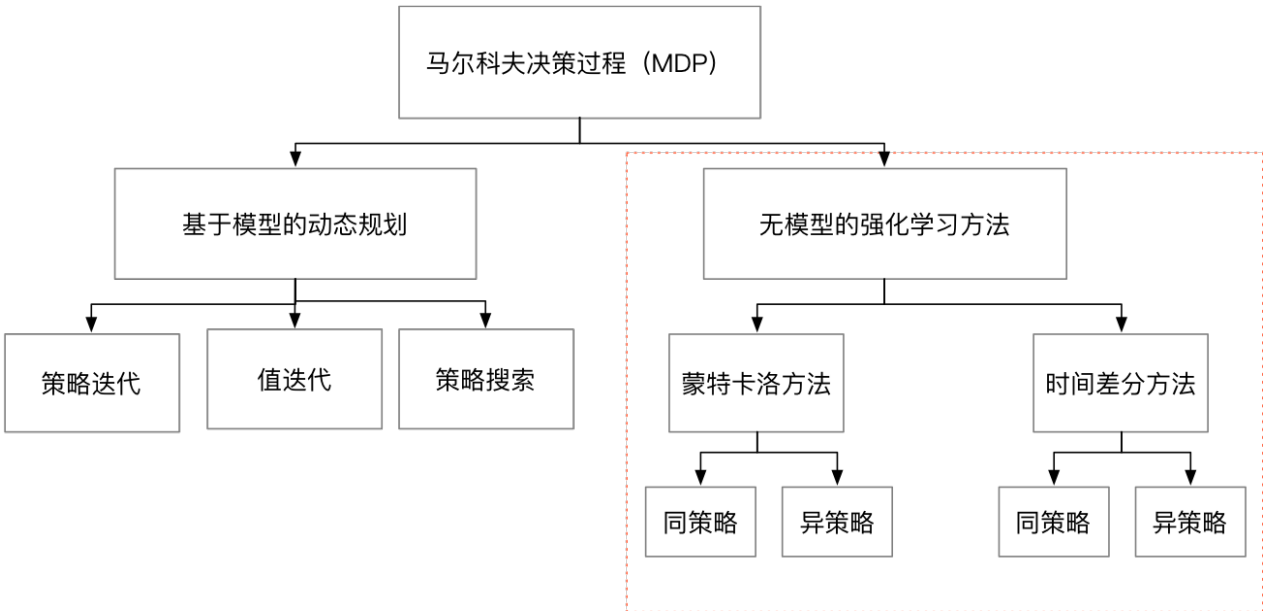
$$\pi_*(a|s) = \begin{cases} 1 - \epsilon + \dfrac{\epsilon}{A(s)} & if \quad a = argmaxq_*(s,a) \\ \dfrac{\epsilon}{A(s)} & \text{其他} \end{cases}$$

$\epsilon - greedy$策略是强化学习常用的随机策略。以比较大的概率选择动作值函数的最大的动作，以比较小的概率选择的其他动作，保证了探索性。

其他还有高斯策略、玻尔兹曼策略等。

从上面的策略来看，求解强化学习的根本需要求解或者估计状态值函数或动作值函数，因此下面主要是如何求解这两个函数。

强化学习分类



如果存在模型，可以使用动态规划的方法解决马尔科夫决策过程，类型于数学优化问题求解。强化学习的精髓是解决无模型的决策问题。我们按照前面的动作值函数计算和策略把强化学习分成两个步骤：1. 策略评估和2 策略改善。

所谓策略评估就是如何计算动作值函数，所谓策略改善就是如果等到一个更好的策略$\pi(a|s)$。

蒙特卡罗方法

要评估当前策略π，我们可以使用策略π参数产生很多次试验，每次试验都是从初始状态到结束状态，我们称之为一次试验（episode）。因此，我们可以按照统计值函数的一个平均值来代替函数期望。

1. 在没有模型的情况下，我们可以采用采样的方法计算状态值函数或者动作值函数。使用经验平均值替代期望。
2. 蒙特卡罗策略改善
    1. 蒙特卡罗利用经验平均估计策略值函数。对每个状态使用最大化动作值函数来进行策略改善。即$\pi(s) = argmax \quad q_*(s,a)$。
    2. 策略值函数递增计算方法

$$v_k(s) = \frac{1}{k}\sum_{j=1}^{k} G_j(s) = \frac{1}{k}(\sum_{j=1}^{k-1} G_j(s) + G_k(s)) = v_{k-1}(s) + \frac{1}{k}(G_k(s) - v_{k-1}(s))$$

3. 蒙特卡罗迭代

时间差分