C1-17

Machine Learning
by Andrew Ng, Stanford Engineering

Xiaojie Zhou

szxjzhou@163.com

2016.10.2

第十八集：线性二次型调节控制

- 状态-动作回报(State-action Rewards)
- 有限长度的马尔可夫决策过程(Finite Horizon MDPs)
- 动力系统模型(Dynamical System Models)
- 线性二次型调节(Linear Quadratic Regulation (LQR))
- Riccati方程(Riccati Equation)

第十八集：线性二次型调节控制

- 状态-动作回报
  - 现在我们对于马尔可夫决策过程的一些细节和一些特殊的情况
    - 首先先来看看回报函数的设计
      - 之前我们设计的回报函数都基于状态，而与动作无关

$$R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \cdots$$

      - 实际上回报函数可以既与状态相关，又与动作相关

$$R(s_0, a_0) + \gamma R(s_1, a_1) + \gamma^2 R(s_2, a_2) + \cdots$$

        - 这样的回报函数称为“状态-动作回报”，是有现实意义的：比如说机器人导航，假设它的前面有个不可通行的障碍物，这样不论动作是前进还是停留最终的结果都会留在原地，但是停留比起前进来说更为省电，因此同样在这个状态下，停留的回报应该比前进的回报更大

第十八集：线性二次型调节控制

- 状态-动作回报
  - 接下来考察关于这个回报函数的一些性质

对于状态-动作回报来说，总回报为：

$$\mathrm{R}(s_0,a_0)+\gamma\mathrm{R}(s_1,a_1)+\gamma^2\mathrm{R}(s_2,a_2)+\cdots$$

对于值函数有：

$$\begin{aligned}\mathrm{V}^{\pi}(s)&=E\left[\mathrm{R}(s_0,a_0)+\gamma\mathrm{R}(s_1,a_1)+\gamma^2\mathrm{R}(s_2,a_2)+\cdots|s_0=s,\pi\right]\\&=\mathrm{R}\left(s,\pi(s)\right)+\gamma E\left[\mathrm{R}(s_1,a_1)+\gamma^2\mathrm{R}(s_2,a_2)+\cdots|\pi\right]\\&=\mathrm{R}\left(s,\pi(s)\right)+\gamma\sum_{s'\in S}p_{s,\pi(s)}\left(s'\right)\mathrm{V}^{\pi}\left(s'\right)\end{aligned}$$

因此对于最优值函数有：

$$\mathrm{V}^{*}(s)=\mathrm{R}\left(s,\pi(s)\right)+\gamma\sum_{s'\in S}p_{s,\pi(s)}\left(s'\right)\mathrm{V}^{*}\left(s'\right)$$

对于最优策略有：

$$\pi^{*}(s)=arg\max_{a\in A}\mathrm{R}(s,a)+\gamma\sum_{s'\in S}p_{s,a}\left(s'\right)\mathrm{V}^{*}\left(s'\right)$$

第十八集：线性二次型调节控制

- 有限长度的马尔可夫决策过程
  - 有限长度的马尔可夫决策过程是一种马尔可夫决策过程的常用变种之一，是一种要求持续时间为T的马尔可夫决策过程，其回报函数定义为：

$$\mathrm{R}^{(0)}(s_0, a_0) + \mathrm{R}^{(1)}(s_1, a_1) + \cdots + \mathrm{R}^{(T)}(s_T, a_T)$$

    - 在这种马尔可夫决策过程中要求要在时间T内结束，也就是需要在这段限制时间内要拿到尽可能多的收益
      - 在回报函数中一般不强调加上折现系数γ
      - 这样的决策过程是非稳定的：因为给定不同的时间T我们可能采取不同的策略，因此在这种情况下状态转换概率和回报函数可能会随着时间的推移而出现不同
      - 下面的问题是：如何在这样一个非稳定的系统中找到最优值函数和最优策略

第十八集：线性二次型调节控制

- 有限长度的马尔可夫决策过程
  - 找到有限长度的马尔可夫决策过程的最优值函数和最优策略

由于在不稳定序列中状态转换概率和回报函数可能会随着时间的推移而出现不同，因此最优值函数与当前所在时刻有关，因此我们不妨假设t时刻给定初始状态为s的最优值函数为：

$$V_t^*(s) = E\left[R^{(t)}(s_t,a_t) + \cdots + R^{(T)}(s_T,a_T)|\pi^*, s_t = s\right]$$

$$= R^{(t)}\left(s,\pi^*(a)\right) + E\left[R^{(t+1)}(s_{t+1},a_{t+1}) + \cdots + R^{(T)}(s_T,a_T)|\pi^*\right]$$

$$= R^{(t)}\left(s,\pi^*(a)\right) + \sum_{s'\in S} p_{s,\pi^*(a)}^{(t)}\left(s'\right)V_{t+1}^*\left(s'\right) = \max_{a\in A} R^{(t)}(s,a) + \sum_{s'\in S} p_{s,a}^{(t)}\left(s'\right)V_{t+1}^*\left(s'\right)$$

从中我们发现当前时刻的最优值函数与后一时刻的最优值函数相关，而且我们可知最后一个时刻当t＝T时给定初始状态为s的最优值函数为（t＝T为最后一个状态，没有后续）：

$$V_T^*(s) = \max_{a\in A} R^{(T)}(s,a)$$

由此我们可以通过动态规划的方法解决在有限长度的马尔可夫决策过程中对于最优值函数的求解问题，具体算法为：

1、 计算t＝T时给定初始状态为s的最优值函数：

$$V_T^*(s) = \max_{a\in A} R^{(T)}(s,a)$$

2、 计算t＝T－1…0时给定初始状态为s的最优值函数：

$$V_t^*(s) = \max_{a\in A} R^{(t)}(s,a) + \sum_{s'\in S} p_{s,a}^{(t)}\left(s'\right)V_{t+1}^*\left(s'\right)$$

在得到t时刻的最优值函数后又可以通过下面的办法求出t时刻的最优策略：

$$\pi_t^*(s) = arg\max_{a\in A} R^{(t)}(s,a) + \sum_{s'\in S} p_{s,a}^{(t)}\left(s'\right)V_{t+1}^*\left(s'\right)$$

第十八集：线性二次型调节控制

- 动力系统模型
  - 之前讲述的有限长度的马尔可夫决策过程的最优值函数和最优策略求解建立在状态和动作均在有限离散空间内取值的情况
    - 现在来考虑状态和动作在连续空间内取值的情况，具体来看就是在给定连续空间内取值的状态集合S、连续空间内取值的动作集合A、状态转移概率p_{sa}、持续时间T、回报函数R的情况下求解对应最优值函数和最优策略
      - 状态转移概率可以借助上一节中讲到的模拟器思想通过采样和监督学习算法加以解决，但是由于有限长度的马尔可夫决策过程是个不稳定过程，因此此时的目标函数和所在时刻t有关
        - 比如说假设t时刻的下一状态s_{t+1}是t时刻当前状态s_t和动作a_t的线性组合加上一个随机噪声ω_t（其中A_t,B_t为系数）

$$s_{t+1} = A_t s_t + B_t a_t + \omega_t$$

      - 而对于回报函数R来说一种经典的假设是
        - 其中矩阵U_t,V_t为半正定矩阵，这使得任意时刻回报函数R的值必定小于等于0

$$R^{(t)}(s_t, a_t) = -\left(s_t^T U_t s_t + a_t^T V_t a_t\right)$$

      - 对于这样的状态转换为线性的而回报函数为二次的模型称为动力系统模型

第十八集：线性二次型调节控制

- 线性二次型调节
  - 首先考虑对于状态转换中系数A_t,B_t的求解

假设t时刻的下一状态s_{t+1}是t时刻当前状态s_t和动作a_t的线性组合，即：

$$s_{t+1} = A_t s_t + B_t a_t$$

对此可以采用采样和监督学习算法加以解决，首先我们可以先得到m次有限长度的马尔可夫决策过程序列：

$$s_0^{(1)} \xrightarrow{a_0^{(1)}} s_1^{(1)} \xrightarrow{a_1^{(1)}} \dots \xrightarrow{a_{T-1}^{(1)}} s_T^{(1)}$$

$$s_0^{(2)} \xrightarrow{a_0^{(2)}} s_1^{(2)} \xrightarrow{a_1^{(2)}} \dots \xrightarrow{a_{T-1}^{(2)}} s_T^{(2)}$$

$$\dots$$

$$s_0^{(m)} \xrightarrow{a_0^{(m)}} s_1^{(m)} \xrightarrow{a_1^{(m)}} \dots \xrightarrow{a_{T-1}^{(m)}} s_T^{(m)}$$

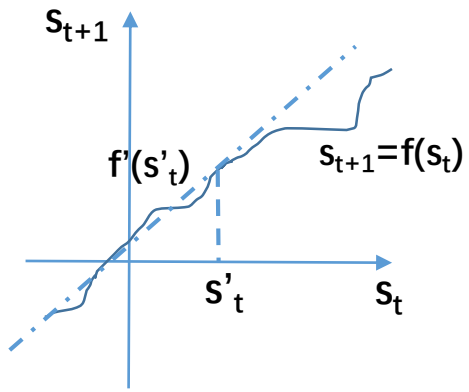
那么我们就可以通过如下的目标函数通过线性回归的监督学习算法对参数A_t, B_t进行学习

$$A_t, B_t = arg \max_{A_t, B_t} \frac{1}{2} \sum_{i=1}^{m} \left\| s_{t+1}^{(i)} - \left(A_t s_t^{(i)} + B_t a_t^{(i)} \right) \right\|^2$$

这里我们认为下一状态s_{t+1}是t时刻当前状态s_t和动作a_t的线性组合，实际上还有另外一种方法，就是先假设一个非线性模型然后再将这个非线性模型转换为线性模型

第十八集：线性二次型调节控制

- 线性二次型调节
  - 非线性模型的线性化



左图曲线显示了下一状态s_{t+1}与当前状态s_t之间的关系曲线$s_{t+1}=f(s_t)$（这种曲线可以通过物理仿真模拟软件得出，为了更好地说明问题，在这里我们假设下一状态s_{t+1}只与当前状态s_t有关，而与动作a_t无关）。从图中可以看出这是一条非线性的曲线，表达形式复杂甚至可能找不到对应的函数表达，不方便进行后续的参数学习，因此需要对其线性化。具体方法为在当前状态s_t的取值空间上找一个点s'_t（这个点要根据实际问题进行选取，要选在实际情况下状态s_t几乎都在该点附近取值的，比如说倒立摆的控制中倒立摆和小车的角度几乎都控制在90度上下，那么我们就需要在这附近进行取值），然后在这一点上做切线（如图中点划线所示），该直线即为对应线性化后的结果，其表达式为（其中由于s'_t为常数，因此$f'(s'_t)$和$f(s'_t)$也均为常数，因此这相当于s_{t+1}为s_t的线性组合）：

$$s_{t+1} \approx f'(s'_t)(s_t - s'_t) + f(s'_t)$$

而在加入动作a_t后可以得到更一般的形式（其中s'_t与a'_t均为常数）：

$$s_{t+1} \approx \left(\nabla_s f(s'_t, a'_t)\right)^T (s_t - s'_t) + \left(\nabla_a f(s'_t, a'_t)\right)^T (a_t - a'_t) + f(s'_t, a'_t)$$

第十八集：线性二次型调节控制

- 线性二次型调节
  - 接下来继续讨论如何对最优值函数和最优策略进行求解

由于有限长度的马尔可夫决策过程的回报函数定义为：

$$R^{(0)}(s_0,a_0)+R^{(1)}(s_1,a_1)+\cdots+R^{(T)}(s_T,a_T)$$

由于此时状态集为连续空间，因此对应的最优值函数为：

$$V_t^*(s)=\max_{a\in A}R^{(t)}(s,a)+\int_{s'}p_{s,a}^{(t)}(s')V_{t+1}^*(s')$$

由于在之前我们用动态规划的办法来求解最优值函数，因此我们在这里还是采用这个思想，先考察$\mathrm{t}=\mathrm{T}$时给定初始状态为s的最优值函数：

$$V_T^*(s)=\max_{a\in A}R^{(T)}(s,a)$$

由于之前已经定义过回报函数为$R^{(t)}(s_t,a_t)=-\left(s_t^TU_ts_t+a_t^TV_ta_t\right)$，代入后可得：

$$V_T^*(s)=\max_{a\in A}-\left(s_T^TU_Ts_T+a_T^TV_Ta_T\right)$$

由于U_T,V_T均为半正定矩阵，因此$s_T^TU_Ts_T\geq 0,a_T^TV_Ta_T\geq 0$，因此对于$\mathrm{t}=\mathrm{T}$来说最好的结果是令$a_T=0$（也就意味着在$\mathrm{t}=\mathrm{T}$时不做任何决策是最优的），从而$a_T^TV_Ta_T=0$，有：

$$V_T^*(s)=-s_T^TU_Ts_T$$

那么对于$\mathrm{t}=\mathrm{T}-1\ldots 0$时，可采用下面这个式子进行递推：

$$V_t^*(s)=\max_{a\in A}R^{(t)}(s,a)+\int_{s'}p_{s,a}^{(t)}(s')V_{t+1}^*(s')=\max_{a\in A}R^{(t)}(s,a)+E_{s'\sim p_{s_t,a_t}}\left[V_{t+1}^*(s')\right]$$

回报函数为$R^{(t)}(s_t,a_t)=-\left(s_t^TU_ts_t+a_t^TV_ta_t\right)$，代入后可得：

$$V_t^*(s)=\max_{a\in A}-\left(s_t^TU_ts_t+a_t^TV_ta_t\right)+E_{s'\sim p_{s_t,a_t}}\left[V_{t+1}^*(s')\right]$$

此时我们考察$\mathrm{t}=\mathrm{T}-1$时的递推情况，由于$\mathrm{t}=\mathrm{T}$时的最优值函数为$V_T^*(s)=-s_T^TU_Ts_T$，因此：

$$V_{\mathrm{T}-1}^*(s)=\max_{a\in A}-\left(s_{\mathrm{T}-1}^TU_{\mathrm{T}-1}s_{\mathrm{T}-1}+a_{\mathrm{T}-1}^TV_{\mathrm{T}-1}a_{\mathrm{T}-1}\right)+E_{s'\sim p_{s_t,a_t}}\left[V_T^*(s')\right]$$

$$=\max_{a\in A}-\left(s_{\mathrm{T}-1}^TU_{\mathrm{T}-1}s_{\mathrm{T}-1}+a_{\mathrm{T}-1}^TV_{\mathrm{T}-1}a_{\mathrm{T}-1}\right)+E_{s_T}\left[-s_T^TU_Ts_T\right]$$

此时由于$-s_T^TU_Ts_T$中U_T已知，因此这是一个关于s_T的函数，而$E_{s_T}\left[-s_T^TU_Ts_T\right]$相当于对这个函数中的唯一变量$s_T$求均值，因此均值的结果变成了常数。而在$-\left(s_{\mathrm{T}-1}^TU_{\mathrm{T}-1}s_{\mathrm{T}-1}+a_{\mathrm{T}-1}^TV_{\mathrm{T}-1}a_{\mathrm{T}-1}\right)$中由于要求$a\in A$下的最大值，最终的结果又使得$a_{\mathrm{T}-1}^TV_{\mathrm{T}-1}a_{\mathrm{T}-1}$变成了常数，因此最终这个式子只是一个关于$s_{\mathrm{T}-1}$的不带一次项的二次函数，因此可将其表示为下面的形式：

$$V_{\mathrm{T}-1}^*(s)=s_{\mathrm{T}-1}^T\Phi_{\mathrm{T}-1}s_{\mathrm{T}-1}+\Psi_{\mathrm{T}-1}$$

而这又可以继续往下推导出$\mathrm{t}=\mathrm{T}-2$时有：

$$V_{\mathrm{T}-2}^*(s)=\max_{a\in A}-\left(s_{\mathrm{T}-2}^TU_{\mathrm{T}-2}s_{\mathrm{T}-2}+a_{\mathrm{T}-2}^TV_{\mathrm{T}-2}a_{\mathrm{T}-2}\right)+E_{s_{T-1}}\left[s_{\mathrm{T}-1}^T\Phi_{\mathrm{T}-1}s_{\mathrm{T}-1}+\Psi_{\mathrm{T}-1}\right]=s_{\mathrm{T}-2}^T\Phi_{\mathrm{T}-2}s_{\mathrm{T}-2}+\Psi_{\mathrm{T}-2}$$

由此可以得到一般性结论：

$$V_{\mathrm{t}+1}^*(s)=s_{\mathrm{t}+1}^T\Phi_{\mathrm{t}+1}s_{\mathrm{t}+1}+\Psi_{\mathrm{t}+1}\Rightarrow V_{\mathrm{t}}^*(s)=s_{\mathrm{t}}^T\Phi_{\mathrm{t}}s_{\mathrm{t}}+\Psi_{\mathrm{t}}$$

由此可得：

$$V_t^*(s)=\max_{a\in A}-\left(s_t^TU_ts_t+a_t^TV_ta_t\right)+E_{s'\sim p_{s_t,a_t}}\left[s_{\mathrm{t}+1}^T\Phi_{\mathrm{t}+1}s_{\mathrm{t}+1}+\Psi_{\mathrm{t}+1}\right]$$

而由于$s_{\mathrm{t}+1}=A_ts_t+B_ta_t$有：

$$V_t^*(s)=\max_{a\in A}-\left(s_t^TU_ts_t+a_t^TV_ta_t\right)+\left(A_ts_t+B_ta_t\right)^T\Phi_{\mathrm{t}+1}\left(A_ts_t+B_ta_t\right)+\Psi_{\mathrm{t}+1}$$

其中：

$$-\left(s_t^TU_ts_t+a_t^TV_ta_t\right)+\left(A_ts_t+B_ta_t\right)^T\Phi_{\mathrm{t}+1}\left(A_ts_t+B_ta_t\right)+\Psi_{\mathrm{t}+1}$$

$$=a_t^T\left(-V_t+B_t^T\Phi_{\mathrm{t}+1}B_t\right)a_t+2s_t^TA_t^T\Phi_{\mathrm{t}+1}B_ta_t+\left(-s_t^TU_ts_t+s_t^TA_t^T\Phi_{\mathrm{t}+1}A_ts_t+\Psi_{\mathrm{t}+1}\right)$$

此时不难发现$-\left(s_t^TU_ts_t+a_t^TV_ta_t\right)+\left(A_ts_t+B_ta_t\right)^T\Phi_{\mathrm{t}+1}\left(A_ts_t+B_ta_t\right)+\Psi_{\mathrm{t}+1}$为关于$a_t$的二次函数，因此其最优值在对$a_t$的偏微分为 0 处取得，有：

$$\nabla_{a_t}\left(a_t^T\left(-V_t+B_t^T\Phi_{\mathrm{t}+1}B_t\right)a_t+2s_t^TA_t^T\Phi_{\mathrm{t}+1}B_ta_t+\left(-s_t^TU_ts_t+s_t^TA_t^T\Phi_{\mathrm{t}+1}A_ts_t+\Psi_{\mathrm{t}+1}\right)\right)=0$$

$$\Leftrightarrow 2\left(-V_t+B_t^T\Phi_{\mathrm{t}+1}B_t\right)a_t+2s_t^TA_t^T\Phi_{\mathrm{t}+1}B_t=0\Leftrightarrow a_t=\frac{s_t^TA_t^T\Phi_{\mathrm{t}+1}B_t}{V_t-B_t^T\Phi_{\mathrm{t}+1}B_t}$$

由此可见t时刻的最优策略为s_t的函数：

$$a_t=\frac{s_t^TA_t^T\Phi_{\mathrm{t}+1}B_t}{V_t-B_t^T\Phi_{\mathrm{t}+1}B_t}\Leftrightarrow a_t=L_ts_t$$

第十八集：线性二次型调节控制

- Riccati方程

由于

$$a_t = \frac{s_t^T A_t^T \Phi_{t+1} B_t}{V_t - B_t^T \Phi_{t+1} B_t} \Leftrightarrow a_t = L_t s_t$$

因此：

$$\begin{aligned} V_t^*(s) &= -\left(s_t^T U_t s_t + a_t^T V_t a_t\right) + \left(A_t s_t + B_t a_t\right)^T \Phi_{t+1}\left(A_t s_t + B_t a_t\right) + \Psi_{t+1} \\ &= s_t^T\left(-U_t + A_t^T \Phi_{t+1} A_t\right) s_t + 2 s_t^T A_t^T \Phi_{t+1} B_t a_t + \left(-a_t^T V_t a_t + a_t^T B_t^T \Phi_{t+1} B_t a_t + \Psi_{t+1}\right) \\ &= s_t^T\left(-U_t + A_t^T \Phi_{t+1} A_t\right) s_t + 2 s_t^T A_t^T \Phi_{t+1} B_t L_t s_t + \left(-s_t^T L_t^T V_t L_t s_t + s_t^T L_t^T B_t^T \Phi_{t+1} B_t L_t s_t + \Psi_{t+1}\right) \\ &\Leftrightarrow s_t^T \Phi_t s_t + \Psi_t \end{aligned}$$

由此可得Φ_t和Ψ_t的递推式：

$$\begin{aligned} \Phi_t &= -U_t + A_t^T \Phi_{t+1} A_t + 2 A_t^T \Phi_{t+1} B_t L_t - L_t^T V_t L_t + L_t^T B_t^T \Phi_{t+1} B_t L_t \\ &= -U_t + A_t^T \Phi_{t+1} A_t + 2 A_t^T \Phi_{t+1} B_t \frac{B_t^T \Phi_{t+1} A_t}{V_t - B_t^T \Phi_{t+1} B_t} - A_t^T \Phi_{t+1} B_t \frac{B_t^T \Phi_{t+1} A_t}{V_t - B_t^T \Phi_{t+1} B_t} \\ &= -U_t + A_t^T \left(\Phi_{t+1} + \frac{\Phi_{t+1} B_t B_t^T \Phi_{t+1}}{V_t - B_t^T \Phi_{t+1} B_t}\right) A_t \end{aligned}$$

$$\Psi_t = \Psi_{t+1}$$

其中$\Phi_t = A_t^T \left(\Phi_{t+1} + \frac{\Phi_{t+1} B_t B_t^T \Phi_{t+1}}{V_t - B_t^T \Phi_{t+1} B_t}\right) A_t - U_t$被称为离散时间 Riccati 方程，而$\Psi_t$在确定性系统下只需要等于$\Psi_{t+1}$，但在非确定性系统中还要加上噪声的影响

而这也给出了在有限长度的马尔可夫决策过程的最优值函数和最优策略求解的一般性方法：

1、初始化Φ_t和Ψ_t

$$\Phi_t = \Phi_T = -U_t, \Psi_t = \Psi_T = 0$$

2、对于$t = T - 1 \ldots 0$，通过下式计算Φ_t, Ψ_t

$$\Phi_t = A_t^T \left(\Phi_{t+1} + \frac{\Phi_{t+1} B_t B_t^T \Phi_{t+1}}{V_t - B_t^T \Phi_{t+1} B_t}\right) A_t - U_t$$

$$\Psi_t = \Psi_{t+1}$$

3、通过下式计算$t = T - 1 \ldots 0$的最优值函数和最优策略：

$$V_t^*(s) = s_t^T \Phi_t s_t + \Psi_t$$

$$a_t = \frac{s_t^T A_t^T \Phi_{t+1} B_t}{V_t - B_t^T \Phi_{t+1} B_t} = L_t s_t$$