人工智能第四次作业

西洋双陆棋的 TD-Gammon 学习方法

姓名：王丹

学号：2120151036

目录

一、实验目的

1.理解和掌握强化学习的基本原理；
2.理解 TD-Gammon，及其在 backgammon 中的应用。

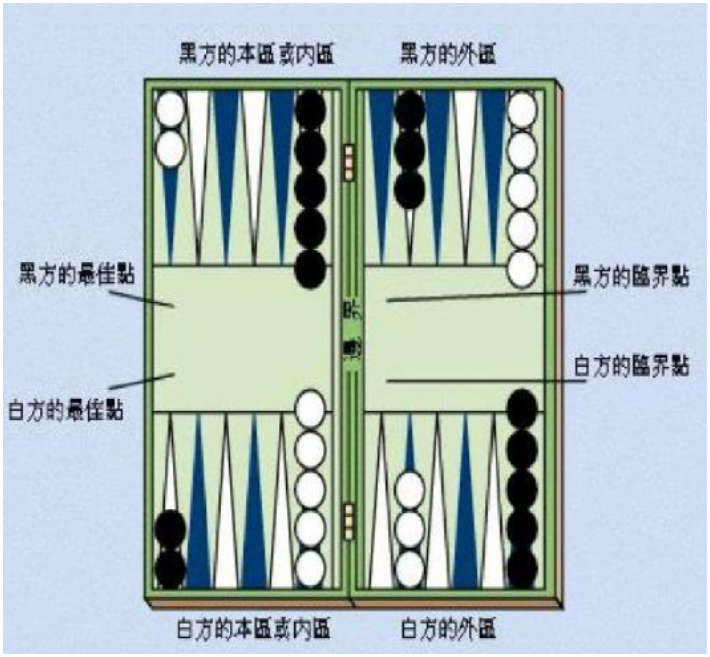
二、实验内容

阅读 Temporal Difference Learning and TD-Gammon 和 Practical Issues in Temporal Difference Learning 两篇文章，梳理脉络，总结其在西洋双陆棋上的应用。

三、原理解析

1.backgammon

西洋双陆棋是攻两人玩的一种在棋盘或桌子上走棋的游戏，靠掷两枚骰子决定走棋的步数，比赛的目的是要使自己的棋子先到达终点。棋盘分为 4 部分，或称 4 大区。每部分用黑、白颜色交替标出 6 个楔形狭长区或小据点。有一条称作



边界的垂直线把棋盘分为内区和外区。比赛时一方使用 15 枚白棋子，另一方使用 15 枚黑棋子。双方根据其所投骰子上显示的点数，从各自的内区（本区）向相反方向从一个据点到另一个据点移动自己的棋子。两枚骰子显示的可分别移动两枚棋子，也可以把它们加起来去移动一枚棋子。任何一方把所有 15 枚棋子都送到本区后，即可按骰子滚出的点数把棋子移至棋盘边界外的虚设据点，这叫做“离盘”。先把全部 15 枚棋子离盘者胜。

2.TD-Gammon

(1)MLP

多层感知器（MLP）是一种多层前馈网络模型，它通常由三部分组成：

a.一组感知单元组成输入层；

b.一层或多层计算节点的印藏藏；

c.一层计算节点的输出层。

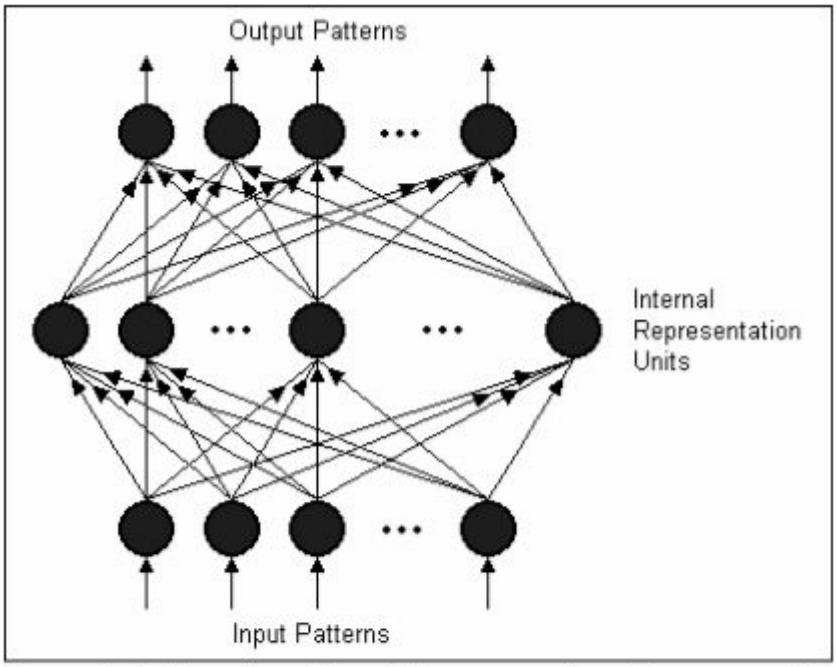


Figure 1. An illustration of the multilayer perception architecture used in TD-Gammon's neural network. This architecture is also used in the popular backpropagation learning procedure. Figure reproduced from [9].

MLP 的学习目标是不断修改网络连接权值 w，使网络输出值不断逼近实际函数值。使用反向传播算法训练多层感知器，它由两次经过网络不同层的通过组成：一次前向通过和一次反向通过。多层感知器的每一个隐藏层或输出层的神经元用来进行两种计算：计算一个神经元的输出处出现的函数信号；梯度向量的估计计算，它需要反向通过网络。

(2)TD(λ)学习

令 $x_1, x_2, \ldots, x_f$ 为一个状态序列，其中 x_t 是 t 时刻的状态（$t = 1,2,\ldots,f$ ），每个状态 x_t 对应着一个向量，而向量的每个分量对应着着一个局面特征，令 z 为最后一步棋的局面状态（此时可判断出最终胜负）对应的奖励值。若白棋取胜，则 $z = 1$；若黑棋取胜，$z = 0$。对于上述局面所构成的观测序列，有一个相应的对 z 值发估计序列 $P_1, P_2, \ldots, P_f$。 P_t 是从状态 x_t 开始白棋获胜的估计值。若机器执白子对弈，

则当白子走步时，应选择使 P_t 值最大的状态 x_t；而黑子走步时，应选择使 P_t 值最小的状态 x_t。将 P_t 作为 x_t 的函数，即 $P_t = P_t(w, x_t)$，其中，w 是权值向量。TD(λ)学习的过程可归结为通过修正 w 来拟合估计函数的过程。

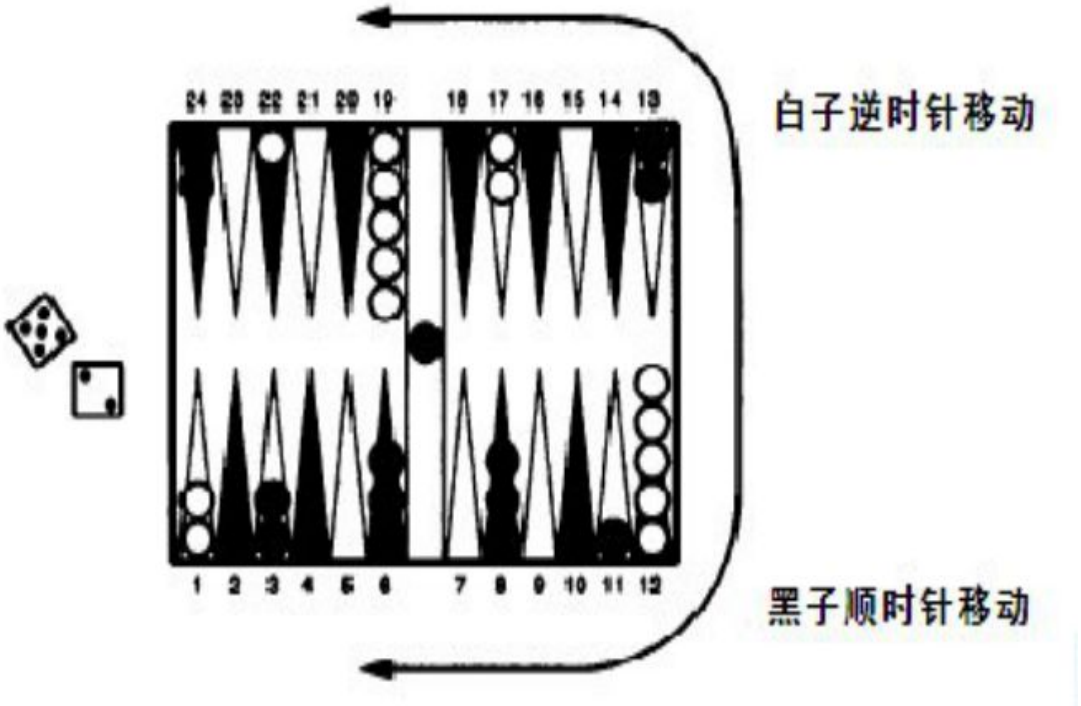
TD(λ) 利用奖励值进行学习。构造监督学习的差分序列 $(x_1, P_1), (x_2, P_2), \ldots, (x_f, P_f)$，从而对权值向量加以训练。权值向量的更新函数如下：

$$w_{t+1} - w_t = \Delta w_t = \alpha(P_{t+1} - P_t)\sum_{k=1}^{t} \lambda^{t-k} \nabla_w P_k$$

其中，α是一个小值常量，一般 $\alpha \in [0.0, 1.0]$，是学习速率。α决定着参数调整的幅度。α越大，权值的调整幅度越大；反之，幅度越小。梯度 $\nabla_w P_k = \frac{\partial P_k}{\partial w}$ 是 P_k 关于向量 w 的偏导数。λ是衰减因子，是一个启发式的参数，决定着学习过程是从短期的预测中学习，还是从长期的预测中学习。当 $\lambda = 0$ 时，$\Delta w_t = \alpha P_{t+1}$，仅从下一个状态中学习；当 $\lambda = 1$ 时，仅从最终结果中学习。

(3)TD-Gammon=TD(λ)+MLP

TD-Gammon 中的学习算法是 TD(λ)与使用多层神经网络通过反向传播 TD 误差来进行训练的非线性函数逼近的直接结合。



网络的输入层：使用一共 16 个输入单元编码整个局面特征：其中 6 个单元编码黑子在 1-6 号位置的数目特征，6 个单元编码白子在 19-24 号位置的数目特征，2 个单元编码黑子和白子的离盘数，2 个单元编码当前可移动棋子的一方。

网络的隐藏层：TD-Gammon2.0 使用 40 个隐藏单元，TD-Gammon2.1 使用 80 个隐藏单元，TD-Gammon3.0 使用了 160 个隐藏单元和可选择的 3 层搜索。

隐藏单元 j 的输出 h(j)是一个非线性曲线函数的加权和：

$$h(j)=\sigma(\sum_i w_{ij}\phi(i))=\frac{1}{1+e^{-\sum_i w_{ij}\phi(i)}}$$

网络的输出层：白棋从当前状态开始能获胜的奖励值。（在 Temporal Difference Learning and TD-Gammon 中，输出 Y[t]是一个四元向量代表白方或黑方赢棋的四种可能性。）

参数的选择：学习速率 $\alpha=0.1$，衰减因子 $\lambda=0.7$。初始权重 $w.\in[-0.5,0.5]$，对网络的各连接权值 w_{ij} 赋予一个-0.5 到+0.5 间的随机数。

棋盘中每走一步，棋盘状态就会改变，形成一个新的输入状态 x_i，一盘棋如果进行了 n 步，就有 n 个代表不同棋盘状态的向量 $x_1,x_2,\ldots,x_n$。我们给最后一个状态（已判定输赢的状态）一个实际评估值 z。当白棋取胜时，$z=1$；当黑棋取胜时，$z=0$。

对局过程中，程序对棋局状态进行记录。下完一局后，程序根据这局棋对局记录对网络进行训练。

权值调整的过程：

a.按 n 递减顺序取 x_{i+1} 和 x_i；

b.计算 P_{i+1}（如果 i=n,则白棋赢为 1，黑棋赢为 0）和 P_i

c.根据误差反向传播法调整权值 $w\leftarrow w+\sum_{i=1}^{n}\alpha(P_{i+1}-P_i)\sum_{k=1}^{i}[\lambda^{i-k}\frac{\partial P_k}{\partial w}]$；

d.取下一记录，返回 a，直到棋局记录全部训练完毕。

四、小结

两篇论文对西洋双陆棋的 TD-Gammon 学习算法进行了介绍，表明从零知识开始进行自学习训练的方案能够在西洋双陆棋中取得良好效果。通过这个应用，我明白了强化学习在机器博弈中的优越性，在读两篇文章的过程中，我还读了 TD 学习算法在其它种类的棋（如五子棋、六子棋）中的应用。此外，还回顾了 MLP 相关知识。