多智能体强化学习最小均值方差公式推导

2024 年 11 月 22 日

note：这里只考虑VDN和CTD论文里的vanilla公式

1 VDN

每个智能体通过DQN来根据$observation_i$选取$action_i$，并计算q_i。各个智能体的 q值之和与 reward 做loss，完成训练。

2 CTD

用第一个Q-learning预测$V^\pi(s_t)$的**均值**，用第二个Q-learning根据第一个Q-learning 的TD-error预测$V^\pi(s_t)$的方差。

3 VDN+CTD 公式推导

3.1 MDP

一个MDP可用一个五元组$(S, A, P^a_{s,s'}, R, \gamma)$表示。其中$S$（state）是状态空间，$A$（action）是动作空间， $P^a_{s,s'}$是在状态s下采取动作a可以转换到状态s'的概率， R_t是智能体在t时可以获得的奖励，在环境中，智能体每次实际获得的奖励记为r_t， γ是折扣因子。

定义策略为：

$$\pi : P(a|s)$$

策略为从状态s选择动作a的概率。

定义回报（随机变量）为：

$$G^\pi(s_t) = \sum_{k=1}^{\infty} \gamma^{k-1} R_{t+k}$$

，

回报为遵从策略π，智能体从状态s_t开始到结束，可以获得的累计折扣奖励。

在MDP中，算法的目标是得到最优的策略π^*，使得回报$G^\pi(s_t)$最大。

3.2 Value Functions

定义状态价值函数$V^\pi(s_t) = \mathbb{E}[G^\pi(s_t)]$，其中$G^\pi(s_t)$是指遵从策略$\pi$，智能体从状态$s_t$开始到结束，可以获得的累计折扣奖励。对状态价值函数可进行进一步推导，得到状态价值函数的贝尔曼方程：

$$\begin{aligned}
V^\pi(s_t) &= \mathbb{E}[G^\pi(s_t)] \\
&= \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots] \\
&= \mathbb{E}[R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \cdots)] \\
&= \mathbb{E}[R_{t+1} + \gamma G^\pi(s_{t+1})] \\
&= \mathbb{E}[R_{t+1}] + \gamma \mathbb{E}[G^\pi(s_{t+1})]
\end{aligned}$$

定义状态-动作价值函数$Q^\pi(s_t, a_t) = \mathbb{E}[G^\pi(s_t, a_t)]$，其中$G^\pi(s_t, a_t)$是指智能体从状态$s_t$开始，选取动作$a_t$，然后遵从策略$\pi$直到结束，可以获得的累计折扣奖励。同理，对状态-动作价值函数可进行进一步推导，得到状态-动作价值函数的贝尔曼方程：

$$
\begin{aligned}
Q^{\pi}(s_t, a_t) &= \mathbb{E}[G^{\pi}(s_t, a_t)] \\
&= \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots] \\
&= \mathbb{E}[R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} + \cdots)] \\
&= \mathbb{E}[R_{t+1} + \gamma G^{\pi}(s_{t+1}, a_{t+1})] \\
&= \mathbb{E}[R_{t+1}] + \gamma \mathbb{E}[G^{\pi}(s_{t+1}, a_{t+1})]
\end{aligned}
$$

由于价值函数是回报的期望，所以在最优的策略π^*时，有最优状态价值函数$V^{\pi^*}(s_t)$ 和最优动作-状态价值函数$Q^{\pi^*}(s_t, a_t)$。因此，找到使得回报$G^{\pi}(s_t)$最大的最优策略π^*的过程就相当于对价值函数进行优化的过程。

3.3 TD learning

强化学习中，一种更新价值函数的方法是TD学习。

根据状态价值函数的贝尔曼方程，可得：

$$
V^{\pi}(s_t) = r_{t+1} + \gamma V^{\pi}(s_{t+1})
$$

其中$r_{t+1} + \gamma V^{\pi}(s_{t+1})$被称为TD目标，是即时奖励和下一个状态的价值的无偏估计，用于近似贝尔曼方程中的期望值。

根据TD目标，可定义TD误差δ_t为$r_{t+1} + \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t)$。 TD误差项表示实际获得的回报与预期回报之间的差异。

利用TD误差，可以推导出更新当前的状态价值函数的公式：

$$
\begin{aligned}
V^{\pi}(s_t) &= V^{\pi}(s_t) + \alpha \times \delta_t \\
&= V^{\pi}(s_t) + \alpha (r_{t+1} + \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t))
\end{aligned}
$$

其中α是学习率，控制状态价值函数的更新速度。

同理，根据状态-动作价值函数的贝尔曼方程，可得：

$$
Q^{\pi}(s_t, a_t) = r_{t+1} + \gamma Q^{\pi}(s_{t+1}, a_{t+1})
$$

同理，可以推导出更新当前状态-动作价值函数的公式：

$$Q^{\pi}(s_t,a_t)=Q^{\pi}(s_t,a_t)+\alpha(r_{t+1}+\gamma Q^{\pi}(s_{t+1},a_{t+1})-Q^{\pi}(s_t,a_t))$$

其中α是学习率，控制状态-动作价值函数的更新速度。

3.4 todo

CTD DQN VDN