

Project Work Example

Caden Hewlett

2024-03-22

The *Bellman Equation* for the frequentist Q-learning update model is given by.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_t + \gamma \operatorname{argmax}_{a \in A} \{Q(s_{t+1}, a)\} - Q(s_t, a_t) \right]$$

In the literature, a 4-tuple MDP $\langle S, A, p_t, p_r \rangle$ is considered; where S is the state set (assumed to be discrete) and A is the action set. Critically, 4-tuple model includes p_t which is the probability that action a_t at time t actually sends you to the target s_{t+1} , where $(s_t \xrightarrow{a_t \in A} s_{t+1})$. A real-world example of this would be training a reinforcement model on an investment portfolio; there is some probability p_t that investing in a given company (a given $a \in A$) will actually result in that company actually using that money on productive R&D ($s_{t+1} \in S$.) In a simplified pathfinding model, however, we assume that the transition probabilities given action a at time t is not stochastic. In other words,

$$\forall t \in \mathbb{Z}^+, \forall (s_i, s_j) \in S \text{ s.t. } \exists a \in A \text{ s.t. } (s_i \xrightarrow{a} s_j), \mathbb{P}(S_{t+1} = s_{t+1} \mid \{a_t, s_t\}) \equiv \underbrace{p_t(s \xrightarrow{a} t)}_{\text{from literature}} = 1$$

Further, in the four-tuple model there is p_r ; similar to p_t , it is the probability that we receive reward r in the rewards set R given we arrive at state s_{t+1} after taking $(s_t \xrightarrow{a_t \in A} s_{t+1})$ at time t . In the investment example, this means that given the investment and increased R&D, p_r is the probability of a given return on investment $r \in R \subseteq \mathbb{R}$.

in a Bayesian sense

$$V^*(s) = \operatorname{argmax}_{a \in A} \{Q(s_{t+1}, a)\}$$