COMP9444 Neural Networks and Deep Learning

Quiz 7 (Reinforcement Learning)

This is an optional quiz to test your understanding of the material from Week 7.

---

1. Explain the difference between the following paradigms, in terms of what is presented to the agent, and what the agent aims to do:
   - Supervised Learning
   - Unsupervised Learning
   - Reinforcment Learning

      - Supervised Learning: Each training item includes an input and a target output. The aim is to predict the output, given the input (for the training set as well as an unseen test set).
      - Unsupervised Learning: Each training item consists of only an input (no target value). The aim is to learn hidden features, or to infer whatever structure you can, from the data (input items).
      - Reinforcement Learning: An agent chooses actions in a simulated environment, observing its state and receiving rewards along the way. The aim is to maximize the cumulative reward.

2. Describe the elements (sets and functions) that are needed to give a formal description of a reinforcement learning environment. What is the difference between a deterministic environment and a stochastic environment?

   Formally, a reinforcement learning environment is defined by a set S of states, a set A of actions, a transition function δ and a reward function R.
   For a deterministic environment, δ and R are single-valued functions:
   $\delta : S \times A \rightarrow S$ and $R : S \times A \rightarrow \mathbf{R}$
   For a stochastic environment, δ and/or R are not single-valued, but instead define a probability distribution on S or $\mathbf{R}$.

3. Name three different models of optimality in reinforcement learning, and give a formula for calculating each one.

Finite horizon reward:	$\Sigma_{0 \leqslant i < h}\, r_{t+i}$
Infinite discounted reward:	$\Sigma_{i \geqslant 0}\, \gamma^i\, r_{t+i}$, $0 \leqslant \gamma < 1$
Average reward:	$\lim_{h \rightarrow \infty} (1/h)\, \Sigma_{0 \leqslant i < h}\, r_{t+i}$

4. What is the definition of:
   a. the optimal policy
   b. the value function
   c. the Q-function?

      a. The optimal policy is the function π^*: $S \rightarrow A$, which maximizes the infinite discounted reward.

b. The value function $V^{\pi}(s)$ is the expected infinite discounted reward obtained by following policy π starting from state *s*. If $\pi = \pi^*$ is optimal, then $V^*(s) = V^{\pi^*}(s)$ is the maximum (expected) infinite discounted reward obtainable from state *s*.

c. The Q–function $Q^{\pi}(s,a)$ is the expected infinite discounted reward received by an agent who begins in state s, first performs action *a* and then follows policy π for all subsequent timesteps. If $\pi = \pi^*$ is optimal, then $Q^*(s,a) = Q^{\pi^*}(s,a)$ is the maximum (expected) discounted reward obtainable from *s*, if the agent is forced to take action *a* in the first timestep but can act optimally thereafter.

5. Assuming a stochastic environment, discount factor γ and learning rate of η, write the equation for
   a. Temporal Difference learning TD(0)

$$V(s_t) \leftarrow V(s_t) + \eta\, [r_t + \gamma V(s_{t+1}) - V(s_t)]$$

   b. Q–Learning

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \eta\, [r_t + \gamma \max_b Q(s_{t+1}, b) - Q(s_t, a_t)]$$

Remember to define any symbols you use.

s_t = state at time t, a_t = action performed at time t,
r_t = reward received at time t, s_{t+1} = state at time $t+1$.