B4M36SMU

Reinforcement learning 3 - MDP

Monday 22$^{\text{nd}}$ May, 2017

Passive reinforcement learning agent

- Evaluates a fixed policy
- Observes rewards and calculates expected utility
- Model is not known
- *Model-based* and *model-free* learning

Direct utility estimation

- Learn expected reward-to-go from the observations

Adaptive dynamic programming

- Estimate probability of transitions $P(s' \mid s, a)$
- Store state-action frequency table N_{sa}
- and state-action-state frequency table $N_{saa'}$
- Evaluate policy the same way as in policy iteration

Temporal difference learning

$$U^{\pi}(s) = U^{\pi}(s) + \alpha \left(R(s) + \gamma U^{\pi}(s') - U^{\pi}(s) \right)$$

Active reinforcement learning agent

- No fixed policy, policy calculated online
- Exploration vs. exploitation
- SARSA algorithm learns Q-function

$$Q^{\pi}(s,a) = Q^{\pi}(s,a) + \alpha\left(R(s) + \gamma Q^{\pi}(s',a') - Q^{\pi}(s,a)\right)$$

Now implement a passive RL agent using TD

- Download a jupyter notebook with instructions from the CW.

Recommended literature

Stuart Russell and Peter Norvig
Artificial Intelligence: A Modern Approach, third edition.
http://aima.cs.berkeley.edu/
Chapter 21