Aprendizaje reforzado

IIC2613 2021-1

Qué es?

- Uno de los 3 paradigmas del Machine Learning.
- Aprendizaje a través de la experiencia (interacción con el entorno).
- Algoritmos “goal oriented”: enfocados en maximizar una “recompensa”.

Modelo del ambiente

- MDP: Markov Decision Process
  - Espacio de estados S
  - Espacio de acciones A(s)
  - Probabilidades de transición
  - Recompensas

- Propiedad markoviana: “el futuro depende del pasado únicamente a través del presente”
  - Dado un estado, la acción a tomar no depende de los estados anteriormente observados

Modelo del ambiente

- Se busca determinar una política óptima de comportamiento.
  - S -> A(S)

- Las acciones que tomamos en el presente reciben una recompensa inmediata, pero también resultan en recompensas a futuro.

- Factor de descuento.

Función de valor

$$V^*(s) = \max_{\pi} E\left\{\sum_{t=0}^{\infty} \gamma^t r_t^{\pi}\right\}$$

$$V^*(s) = \max_{a}\left\{R(s,a) + \gamma \sum_{s' \in S} T(s,a,s')V^*(s')\right\}, \quad \forall s \in S$$

Función Q

$$Q(s,a) := R(s,a) + \gamma \sum_{s' \in S} T(s,a,s') \hat{V}(s')$$

Model-based y model-free learning

- Model based
  - Tenemos un modelo de la realidad, en forma de probabilidades de transición.
  - Value iteration, policy iteration.

- Model free:
  - No conocemos las probabilidades de transición.
  - Actuamos en función de la experiencia que vamos adquiriendo.
  - Q-learning.

Iteración de valor

Initialize V(s) arbitrarily
loop until policy good enough
    loop for $s \in S$
        loop for $a \in A$
            $Q(s,a) := R(s,a) + \gamma \sum_{s' \in S} T(s,a,s') \hat{V}(s')$
        end loop
        $\hat{V}(s) := \max_a Q(s,a)$
    end loop
end loop
return $\{\hat{V}(s)\}$

Iteración de política

Choose an arbitray policy π'

Loop

$\pi := \pi'$

Compute value function of policy π:

#solve linear equations

$V_\pi(s) := R(s, \pi(s)) + \gamma \sum_{s' \in S} T(s, \pi(s), s') V_\pi(s')$

Improve the policy at each state

$\pi'(s) := \arg\max_a \left(R(s, a) + \gamma \sum_{s' \in S} T(s, a, s') V_\pi(s') \right)$

until $\pi = \pi'$

Q-learning

Initialize $Q(s, a)$ arbitrarily
Repeat (for each episode):
    Initialize s
    Repeat (for each step of episode):
        Choose a from s using policy derived from Q (e.g., ε-greedy)
        Take action a, observe r, s'
        $Q(s, a) \leftarrow Q(s, a) + \alpha\left[r + \gamma \max_{a'} Q(s', a') - Q(s, a)\right]$
        $s \leftarrow s'$;
    until s is terminal

Aprendizaje reforzado

IIC2613 2021-1