Aprendizagem por Reforço - Q Learning

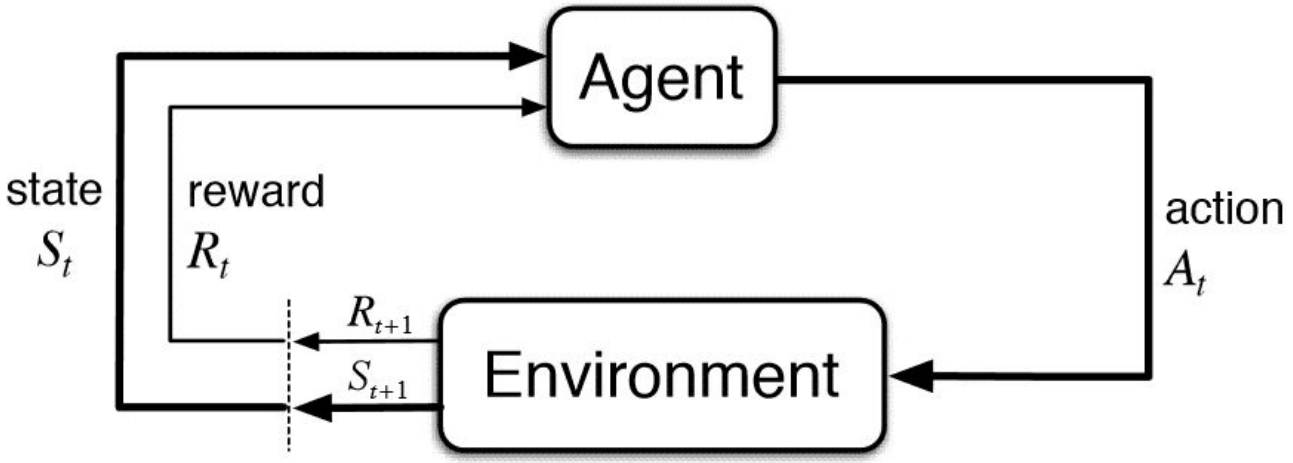
Relembrando… Algoritmo *Value Iteration*

- **Utilidade de um estado é dado pela equação de Bellman:**
  - $U(s) = R(s) + \gamma \max_a \Sigma_{s'} T(s,a,s') U(s')$

- **Necessário conhecer a priori o modelo de transição T e as recompensas**
- **Quando não se conhece, o agente pode explorar o ambiente e aprender a política**

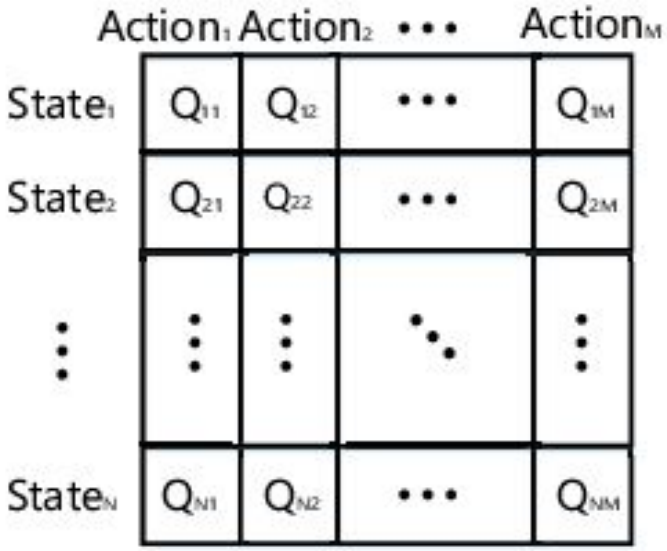
Aprendizado por Reforço

- **Idéia Básica:**
  - Explorar o ambiente para calcular qualidade das ações para cada estado



Q-Learning

- **Objetivo: aprender a qualidade das ações para cada estado, armazenada na tabela Q**



Q-Learning

- **Procedimento:**
  - Inicializar valores de qualidade Q(s,a) de forma arbitrária para cada estado e ação
  - Dado estado inicial
  - Repita
    - Escolher uma ação e executá-la
    - Aplicar a recompensa recebida
    - Observar o novo estado s'
    - Atualizar o valor de Q(s,a) conforme regra de aprendizagem
  - Procedimento pode ser repetido até alcançar uma convergência dos valores da tabela Q

Q-Learning

- **Equação de atualização**

Erro: estimativa de Q(a.s) menos valor atual

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma max_{a'} Q(s',a') - Q(s,a)]$$

Taxa de aprendizagem

Estimativa de Q(s,a) pela Equação de Belman

Q-Learning

- **Equação de atualização (rationale)**

  - Bellman Equation:

$$Q(s,a) = r + \gamma max_{a'} Q(s',a')$$

  - Loss function (squared error):

$$L = \mathbb{E}[(\underbrace{\boldsymbol{r + \gamma max_{a'} Q(s',a')}}_{\textbf{target}} - Q(s,a))^2]$$

Q-Learning

- **Ver exemplo do código**