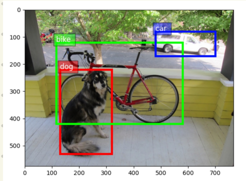
# RF 基础

Supervised Learning 有标签

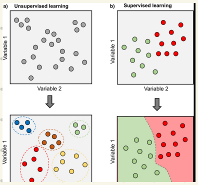



bounding box 坐标，标签

↓ 模型

结果

Unsupervised Learning 无标签



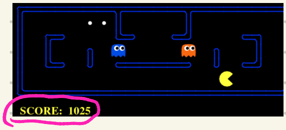

仅原始数据

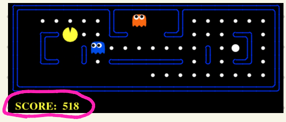
↓ 模型

结果

Reinforcement Learning 学习环境，优化行为

（无数连续标签的监督学习）



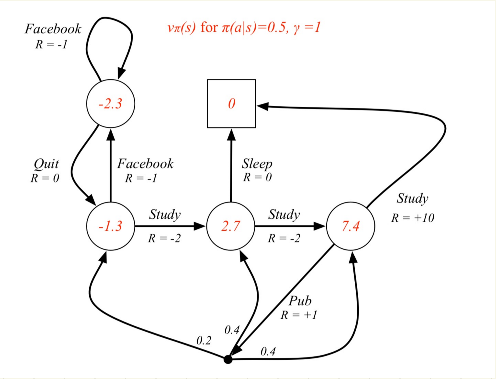




最优动作

↑ 模型

环境

## Markov Decision Process（从感性认知到量化）




$\langle S, A, P, R, \gamma \rangle$

A Markov decision process (MDP) is a Markov reward process with decisions. It is an *environment* in which all states are Markov.

**Definition**

A *Markov Decision Process* is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$

- $\mathcal{S}$ is a finite set of states
- $\mathcal{A}$ is a finite set of actions
- $\mathcal{P}$ is a state transition probability matrix, $\mathcal{P}^a_{ss'} = \mathbb{P}[S_{t+1} = s' \mid S_t = s, A_t = a]$
- $\mathcal{R}$ is a reward function, $\mathcal{R}^a_s = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a]$
- $\gamma$ is a discount factor $\gamma \in [0, 1]$.

矩阵表示 P

$$\mathcal{P} = \begin{array}{c} \\ C1 \\ C2 \\ C3 \\ Pass \\ Pub \\ FB \\ Sleep \end{array} \begin{array}{c} \begin{array}{ccccccc} C1 & C2 & C3 & Pass & Pub & FB & Sleep \end{array} \\ \left[\begin{array}{ccccccc} & 0.5 & & & & 0.5 & \\ & & 0.8 & & & & 0.2 \\ & & & 0.6 & 0.4 & & \\ & & & & & & 1.0 \\ 0.2 & 0.4 & 0.4 & & & & \\ 0.1 & & & & & 0.9 & \\ & & & & & & 1 \end{array}\right] \end{array}$$

Agent 系列动作

- C1 C2 C3 Pass Sleep
- C1 FB FB C1 C2 Sleep
- C1 C2 C3 Pub C2 C3 Pass Sleep
- C1 FB FB C1 C2 C3 Pub C1 FB FB FB C1 C2 C3 Pub C2 Sleep

量化系列动作

Sample returns for Student MRP:
Starting from $S_1 = C1$ with $\gamma = \frac{1}{2}$

$$G_1 = R_2 + \gamma R_3 + ... + \gamma^{T-2} R_T$$

Bellman Equation

$V = R + \gamma P V$

| | | | |
|---|---|---|---|
| C1 C2 C3 Pass Sleep | $v_1 = -2 - 2 * \frac{1}{2} - 2 * \frac{1}{4} + 10 * \frac{1}{8}$ | = | −2.25 |
| C1 FB FB C1 C2 Sleep | $v_1 = -2 - 1 * \frac{1}{2} - 1 * \frac{1}{4} - 2 * \frac{1}{8} - 2 * \frac{1}{16}$ | = | −3.125 |
| C1 C2 C3 Pub C2 C3 Pass Sleep | $v_1 = -2 - 2 * \frac{1}{2} - 2 * \frac{1}{4} + 1 * \frac{1}{8} - 2 * \frac{1}{16} \ldots$ | = | −3.41 |
| C1 FB FB C1 C2 C3 Pub C1 ... FB FB FB C1 C2 C3 Pub C2 Sleep | $v_1 = -2 - 1 * \frac{1}{2} - 1 * \frac{1}{4} - 2 * \frac{1}{8} - 2 * \frac{1}{16} \ldots$ | = | −3.20 |

如同人性：延时满足与即时满足，每人有不同 γ 值

State Value function

$V^\pi(s) = E_\pi\{G_T \mid S_t = s\}$ expected return starting from state $s$ following policy $\pi$

Action value function

$Q^\pi(s,a) = E_\pi\{G_T \mid S_t = s, a_t = a\}$ expected return starts from state $s$, following policy $\pi$ taking action $a$

V & Q 关系 $V^\pi(s) = \sum_{a\in A} \pi(a|s) \cdot Q^\pi(s,a)$ 所有Q 加权平均后是 V

---

如果 MDP 的元组信息缺失 $\langle S, A, P, R, \gamma \rangle$

→ Monte-Carlo Learning
Temporal-Difference Learning
预测 Value function

---

- Value based — learn values of states and actions
- policy-based — learn policy directly, which completely by-passes learning values or actions all together (因为 state space or action space too large)
  比如之前 MDP 无限个状态 动作
- actor critic — combination of value-based and policy-based