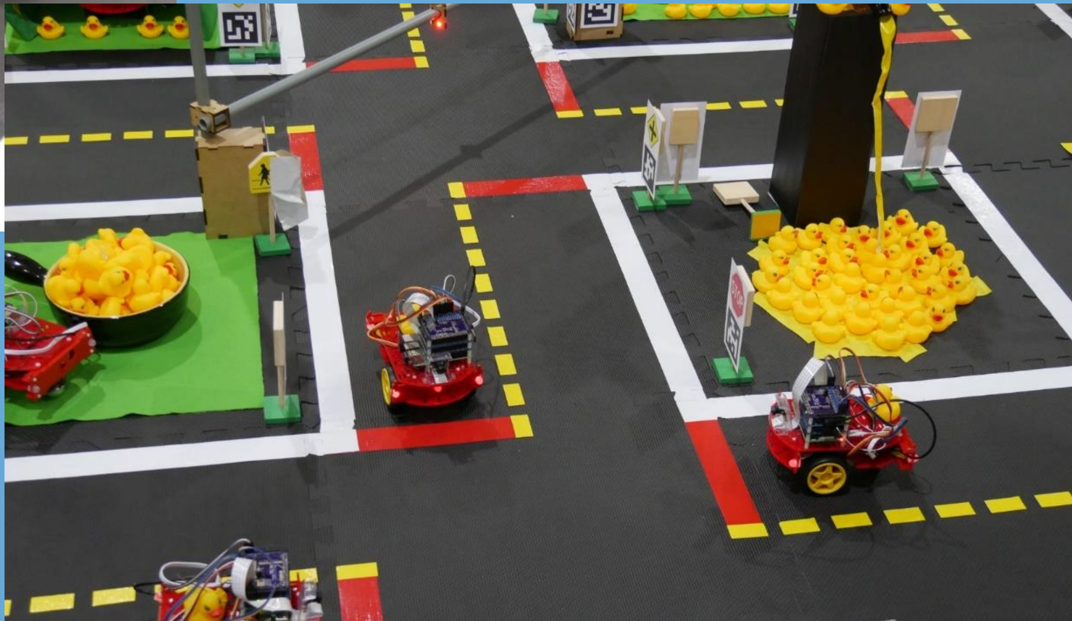
CS 3630 Reinforcement Learning: I

Markov Decision Processes



Reinforcement Learning

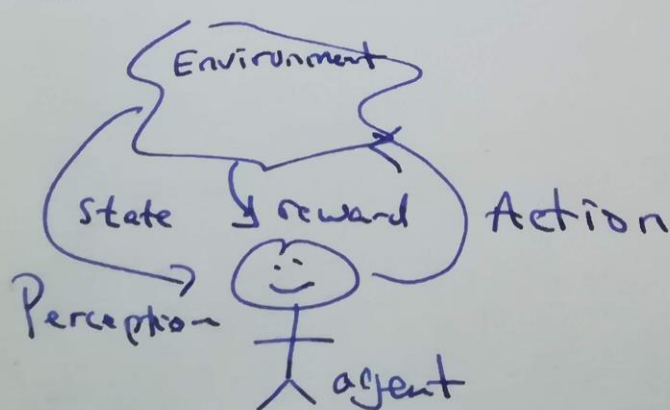
"Typical" Machine Learning

- Lots of Data
- passive Learner
- Learning = pattern analysis
  function approx.

Examples

* Find photos of cats.
* product recommendation
* Text Translation

Robots are Not Like This!



RL:

RL is a process of modifying behavior by rewarding desired outcomes.

- Dogs – Treats, learn tricks
  – punishment
  → negative reward
- Kids – 1$/A on report
- students: grades
  joy, satisfaction
  starting salary

1. Robot senses world
2. Robot decides & executes action
3. World state changes
4. Robot Receives a reward.
   Robot senses world
5. Robot update its "strategy"
6. Go to 2

Questions:

- Mathematical Model
  - states
  - actions
  - uncertainty (!)
- Mathematical formalism for reward
- Given the above How to compute an optimal strategy/policy
- Now... suppose we don't know the models for world, actions... How can we learn them?

- Markov Process
- Markov Decision Process
- Value function & how to compute it
- R.L.

Markov Processes

Simplest case:

- discrete time, $k = 0, 1, 2, \dots$
- discrete states, S set of states.
- A set of transition probabilities

$$T : S \times S \rightarrow \underbrace{[0,1]}_{\text{probability}}$$

$$P\{S_{k+1} \mid S_k\} = T(s, s')$$

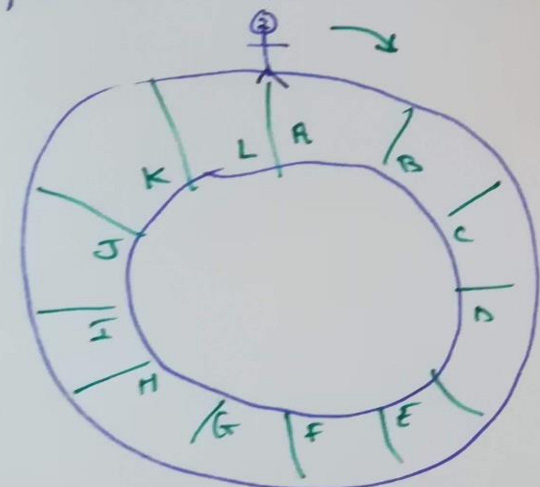
$$= T(S_k, S_{k+1})$$

$T(s, s')$ = prob {arriving to state s', given we are now in state s}

↪ M.P. evolves "autonomously"

Example:

Sisyphus has the job of, every day, throwing a large rock, on a circular track.



L_k = distance of throw on k^{th} day

$$P\{L_k = 1\} = 0.25$$

$$P\{L_k = 2\} = 0.5$$

$$P\{L_k = 3\} = 0.25$$

$$S = \{A, B, \cdots L\}$$

$T(A,B) = 0.25$ $\quad T(A,C) = 0.5$ $\quad T(A,D) = 0.25$

$\vdots$

$T(J,K) = 0.25$ $\quad T(J,L) = 0.5$ $\quad T(J,A) = 0.25$

$T(K,L) = 0.25$ $\quad T(K,A) = 0.5$ $\quad T(K,B) = 0.25$

$T(s,s') = 0$ otherwise

Properties:

- Stationary: T does not change over time

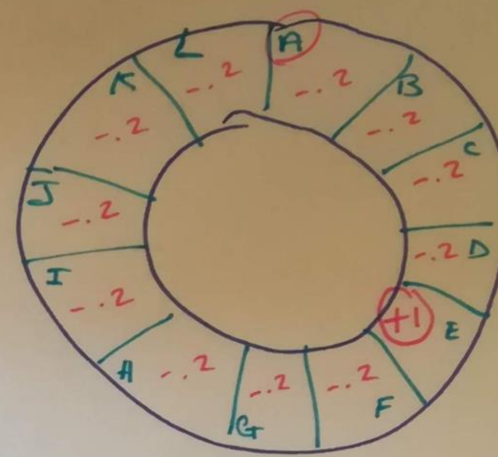
- $\boxed{P\{s_{k+1} \mid s_0, s_1, \cdots s_k\} = P\{s_{k+1} \mid s_k\}}$

Markov Property

$P\{s_3 = E \mid s_0 = A, s_1 = C, s_2 = D\} = P\{s_3 = E \mid s_2 = D\}$

$P\{S_3 = E \mid s_0 = A, s_1 = B, s_2 = D\} =$

Markov Decision Processes (MDPs)



Markov Processes evolve autonomously.

Let's give Sisyphus a tiny bit of free will:

R : Throw rock counterclockwise

L : Throw rock clockwise

Same throwing abilities for L + R

~~P{R}~~ $P\{L_k\}$ unchanged

$\Rightarrow T(A, L, B) = 0.25 \quad T(A, L, C) = 0.5 \quad T(A, L, D) = 0.25$

$\hookrightarrow T(s, a, s') = P\{s' \mid s, a\}$ $\quad T : S \times A \times S \rightarrow [0,1]$

A = set of possible actions.

L = action

$\Rightarrow T(A, R, L) = 0.25 \quad T(A, R, K) = 0.5 \quad T(A, R, J) = 0.25$

⋮

Rewards : $R : S \rightarrow \mathbb{R}$

For Sisyphus: $R(E) = +1$

$R(s) = -0.2$ for $s \neq E$

Suppose Sisyphus executes two actions

$a_1 = L$, $a_2 = L$

$k=1$ $\quad$ $k=2$

Expectation

Suppose a random variable X takes values from the set $\{c_1, c_2, \cdots, c_n\}$. The expected value of X is defined as

$$E[X] = \sum_{i=1}^{n} P\{X = c_i\} \times c_i$$

Example: roll one die, X shows

$$E[X] = \sum_{i=1}^{6} \frac{1}{6} \times i = \underline{3.5}$$

all values equally likely.

Intuition: Perform this many times

Average of X's $\longrightarrow E[X]$

Generalization

$$E[X_1 + X_2] = \sum_i \sum_j P\{\underline{X_1 = c_i, X_2 = c_j}\}(c_i + c_j)$$

Two dice

$$E[X_1 + X_2] = \sum \sum \frac{1}{36}(i+j) = 7$$

$$E[\sum X_i] = \sum P\{X_\bullet = c\} \sum(c_i)$$

For Sisyphus: $E[R(s_0) + R(s_1) + R(s_2)]$ given $a_1 = L$, $a_2 = L$??

We know

$$\begin{cases} R(S_0) = -0.2 \text{ because } S_0 = A \\ R(S_1) = -0.2, \; S_1 \in \{B, C, D\} \\ R(S_2) = \begin{cases} +1 & S_2 = E \\ -0.2 & \text{Else} \end{cases} \end{cases}$$

Expected return

Define return $\underline{\underline{r_h}} = \sum_{i=0}^{h} R(s_i)$. ⟵ (*)

$\underline{r_2}$ for sisyphus $= R(s_0) + R(s_1) + R(s_2)$

$E[r_2] = E[R(s_0) + R(s_1) + R(s_2)]$

$= \underbrace{P\{r_h = .6\}}_{P\{S_3=E\}} \times .6 + \underbrace{P\{r_h = -.6\}}_{P\{S_3 \neq E\}} \times -.6$ *

because $\underline{\underline{r_h}}$ has only two possible values

$-0.2 + -0.2 + -0.2 = -0.6 \Leftarrow S_2 \neq E$

$-0.2 + -0.2 + 1 = 0.6 \Leftarrow S_2 = E$

To arrive $S_2 = E$

S_0	S_1	S_2	Probability
A	B	E	$(0.25) \times (0.25) = 0.0625$
A	C	E	$(0.5) \times (0.5) = 0.25$
A	D	E	$(0.25) \times (0.25) = 0.0625$

$P(S_3 = E) = 0.375$

$P\{S_3 \neq E\} = 1 - P\{S_3 = E\}$
$= 0.625$

This is great for finite horizons (i.e. only consider a finite # of stages).

Suppose $h \to \infty$??

To deal with this, use discounted rewards. discount factor

$r_h = \sum_{i=0}^{h} \gamma^i R(s_i)$

for $0 < \gamma < 1$.

$\lim_{h\to\infty} \sum_{i=0}^{h} \gamma^i R(s_i) \leq \sum_{i=0}^{\infty} \gamma^i R_{max}$

$\sum_{i=0}^{\infty} \gamma^i = \frac{1}{1-\gamma} \quad 0 < \gamma < 1$

$\sum_{i=0}^{\infty} \gamma^i R(s_i) \leq \frac{R_{max}}{1-\gamma}$