Reinforcement Learning

Note Title

Types of learning —

Supervised learning : "Tutor" provides examples to learn from

Unsupervised learning : You're on your own

Reinforcement learning text focuses on this

Russell & Norvig text also contains other types of learning

Reinforcement learning : learning from interaction

"Learning problem"

- state of the environment
- actions that affect state
- goal(s) related to the state

- Uncertainty about the environment
- agent's actions affect environment's future state
- effects of actions cannot fully be predicted

Example: Finance - stock investing

Idea: invest in a company if you think stock will rise
sell if you think stock will fall [or money better invested elsewhere]

Example: Medicine

[See patient's symptoms, don't know state with certainty

[Medication/treatment are likely to help, but may not

Stock prices are indicators of a company's value

Uncertainty: accuracy of indicator

Stock prices fluctuate daily, based on world economy + other factors

uncertainty: predicting world events

Examples: Futures pricing

Can we learn to recognize what the state of the system really is, based on indicators ("evidence") we can collect?

Key elements:

- **policy** - mapping from state of environment to actions

  (for each possible state, specify an action.
  The set of actions you will take for each state
  is your policy.

Chess: given any state of the board, what is your move?

Blackjack: given indicators [value of your cards + observed dealer's card]

what do you do? (HIT or STAND?)

reward function — mapping from state to an **immediate reward**

for each state, what is the immediate return?

e.g. doing chores has negative immediate return in form of work, tedium, etc.
but has long-term reward (we hope)

e.g. brushing your teeth —
time consuming (if done properly)
but prevents cavities

e.g.: training animals —

dolphins — give fish for good behavior

Pavlov's dogs —

value function: mapping from state to long-term reward or penalty

e.g.: having cavities
if not using proper dental care,

car problems
if not changing oil & other maintenance

reward function is the immediate value of being in a state

value function is the ultimate value of being in a state

Example: Choice between BIG NICKEL or tiny dime

Game of Nim

m players

n stones in a pile

K : on your move, can take 1,2,...,K stones

Object: DO NOT BE THE PLAYER WHO TAKES THE LAST STONE

Example: m=2 players

n=7 stones

K=3

Player 1:

Take 2

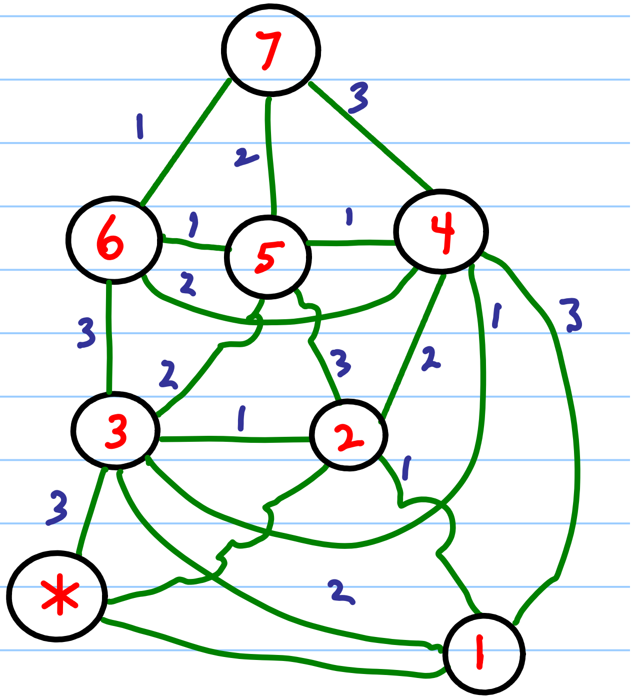
Example: m=2 players

n=9 stones

K=3

Second player always wins

[provided optimal moves are made]



state	action	reward	value
# stones	# to take		
1	1 2 3	-1	
2	1 2 3	0	
3	1 2 3	0	2 ✓
4	1 2 3	0	3 ✓ 2 ✗
5	1 2 3	0	↓
6	1 2 3	0	3 ✗
7	1 2 3	0	
8	1 2 3	0	
9	1 2 3	0	3 ✓✓ ↓

Reinforcement Learning Technique

"Exploration"

"Exploitation"

If you know that your potential moves have specific values associated with the "next states" they produce, exploit this knowledge by choosing best move

⋮

BUT

⋮

if values are only guesses, need to improve these values

⋮

"back up" values based on knowledge

Let "value" of a state be prob of a win if we are at that state.

Initially, we don't know.
Start w/ uniform weights

state	
9	.5
8	.5
7	.5
6	.5
5	.5
4	.5
3	.5
2	.5
1	🟨

Change to 0?
Chang to .4

Play game,
your first move
has state = 6
choices are
1, 2, 3
which yield
5 4 3
as next states
All have same value!

Choose move at random e.g. "2" ⟵ explore! ⟵

state = 4

opponent chooses 3

state = 1

our move — we lose.

Change value of state 1 to 0

NOTE: this game is simple; you always lose from state 1

Maybe should change weight more gradually —

Something between old value (.5) and 0

$$\text{value}(s) \leftarrow \text{value}(s) + \underset{\uparrow}{\alpha}\left(\text{value}(s') - \text{value}(s)\right)$$

learning rate $\alpha \in (0,1)$

Instead change value(1) ← .4, not 0

Playing a new game ....

say state = 6 again –

your choices yield 5 4 3

Explore try move = 1, yielding 5

opponent tries move = 1, yielding 4

state = 4

choices 1, 2, 3

yielding 3, 2, 1

val = .5 val = .5 val = .4

NOTE .4 is your value of winning

So, you are going to choose the minimum so your opponent loses

So, value of state 4 should increase —

value was .5

win: prob = 1

So back up $.5 + \alpha(1 - .5) \Rightarrow$ say .6

value(5) = .6

⋮

value(1) = .4

Exploration

Sometimes, you should choose a random move & learn what happens, rather than exploiting your moves every time.