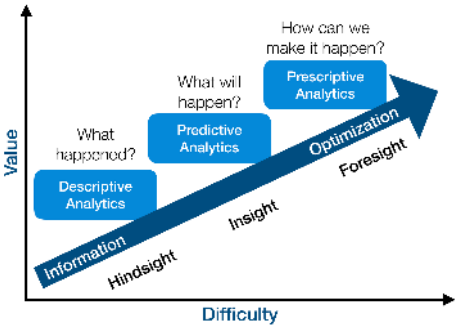
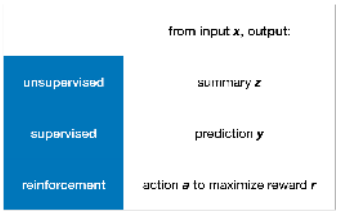
Reinforcement Learning (ch 17 Mohri et al.)

Friday, March 27, 2020 3:05 PM

Intro: sequential decision making → w/ little bits from Puterman and Sutton & Barto 2nd ed.
We'll follow notation & presentation of ch 17 Mohri et al.





ML: prediction (ex. modeling COVID-19)

RL: prescription (ex. optimal policies for governor to implement for COVID-19)

From http://www.argmin.net/2018/01/29/taxonomy/
(Ben Recht, drawing on Chris Wiggins and Gartner)

Very hot these days (Deep Mind acquired for $500m by Google in '14, made AlphaGo, beat humans)
(esp. using neural nets and Q-learning "Deep RL")

↳ for pessimistic point-of-view, see Jeffrey Funk's Dec '19 article for IEEE losses of $571m, Economist: "publicity stud"
https://spectrum.ieee.org/computing/software/ai-and-economic-productivity-expect-evolution-not-revolution

General setup

$s \in S$ Observe a state of a system/environment

$a \in A$ Choose an action

r Receive a reward based only on current state and action
(sometimes this is random, but often deterministic)

The system/environment transitions to a new state, based only on previous state and action
THIS IS RANDOM! and MARKOVIAN ← depends on previous state, but no older states
Transition probability $\mathbb{P}(s'|s,a)$

the usual setup is a MARKOV DECISION PROCESS (MDP)

The environment is like a Markov chain ...
(ex: weather tomorrow is a r.v., depends only on today, not today-and-yesterday)
... except we can influence it w/ our actions

Very broad subject, many variants. We only cover the basics

Somewhat like online learning
- there are transition probabilities, but it's not like the PAC setup w/ $x \sim D$
- we proceed in rounds/epochs
and training/testing are the same thing. No such thing as "generalization error"

Examples

- learning to play Atari or Go games only via trial-and-error (DeepMind)
  - Gerry Tesauro's backgammon program in early '90s
- teaching a robot to walk (see MuJoCo videos)

"Learning Algorithms" (environment unknown) Difficult Modern

- inventory management: when should a store ask central warehouse for inventory?
- bus engine replacement
- highway pavement maintenance
- communications routings (slotted ALOHA protocol)
- mate desertion in Cooper's Hawks
- gambling

"Planning Problem" (environment known) In basic case, this is theoretically tractable (a LP) though state-space is often HUGE. Classical

10^{20} in backgammon

does it make sense to model the environment as a Markov Chain? think of customer demand, or trucks breaking down, as random clear "action"

- toy problem: Restaurant Problem (Feynman?) You visit a new restaurant, don't know if you'll like the dish do you eat a dish you've already tried before and liked, or, do you try something new? exploration-exploitation tradeoff (not really a MDP since not really stochastic)

  Similar to the secretary problem (aka marriage problem): an optimal stopping problem
- Pig (dice game)

  Roll a dice, get to keep score, keep rolling... unless you get a 1, then lose all your pts on that round

  When to stop rolling?
  1. to maximize expectation at 1 round, this is basic probability
  2. ... but playing against an opponent, "first to 100 pts wins" — classic RL problem

     if you're losing badly, need to take a higher-risk / higher-reward strategy
- multi-armed bandit (simple slot machine)

K slot machines, each has a different (and unknown) probability of winning
Where should you put your money? clear exploration-exploitation tradeoff

We'll first setup MDP basics (→ Bellman's Eq, ...), then

1. Planning Problem (environment known), classical — operations research...

   Value iteration, Policy iteration

   "Dynamic Programming" classic texts: Puterman '94 "Markov Decision Processes"

   Filar + Vrieze '97, "Competetive M.D.P."

2. Learning algos (environment unknown), modern — robotics...

   "Monte Carlo"

   "Evolutionary Strategies"

   "Temporal-Differencing" (TD), incl. Q-learning

http://incompleteideas.net/book/the-book-2nd.html

classic texts: Sutton & Barto ('98 1st ed, '18 2nd ed) "RL: an Intro" → free PDF at authors' website

more math: Bertsekas, Tsitsiklis '96 "Neuro-Dynamic Programming"
Szepesvári '10 "Algo. for RL"

Brief history

MDP's codified w/ books by Bellman ('57), Howard ('60)

roots: Cayley 1875
Wald, during WWII ← note Dantzig's simplex algo for LP created for very similar problems at same time
Massé
RAND Corp.
Arrow, Bellman, Blackwell, ..., Shapley, ...
↑ Nobel prize '72 econ ↑ Nobel prize '12 econ

mostly economics, game theory, operations research → OR = math behind running a business (supply chain decisions ...)
and pure math (stochastic processes / finance)
↑ fun things like time-inconsistent stopping problems:
according to the model, you should stop smoking tomorrow
...but tomorrow, it will tell you to stop the next day! etc.

Learning algos: see Sutton, Barto late '90s

① Background math

MDP: Set of States S, actions A } if both sets finite, call it a "finite MDP"

initial state s_0

transition prob: $\mathbb{P}(s'|s,a)$ ← a matrix in a finite MDP

reward prob: $\mathbb{P}(r'|s,a)$, often deterministic $r' = r(s,a)$

time: → cts time (possible but we won't discuss)
→ discrete → finite horizon rounds/epochs $\{0,1,\dots,T\}$ $\sum_{t=0}^{T} r_t$
→ infinite horizon → discounted $\sum_{t=0}^{\infty} \gamma^t r_t$, $\gamma \in [0,1)$ ← has nicest theory
→ avg. $\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} r_t$ (if it exists)

(why discount? ① we have to, since $\sum_{t=0}^{\infty} r_t = \infty$
② $100 today is worth less than $100 tomorrow (inflation... or, I could have invested it and received a return)
③ for large T, "it's not my problem" (change jobs, politician out-of-office, death, technology/society will have changed the rules of the game) and, since T isn't known exactly, the discount sort of covers it)

Our job is to decide on the actions, i.e.,
at time t, a decision rule mapping the state to an action
(so a matrix for a finite MDP)

but often we want a policy (Mohri's "policy" is called a "stationary policy" in Puterman)
meaning the decision rule doesn't depend on t
(and we'll prove that for ∞-horizon discounted MDP, $\exists$ an optimal stationary policy)

In particular, a deterministic (stationary) policy is a mapping $\pi : S \to A$
i.e., $a = \pi(s)$ → non-deterministic also possible, $\pi : S \to \Delta(A)$ (prob. distr. over A)
but we'll also prove $\exists$ optimal deterministic policy

Objective: maximize expected return/reward (for a deterministic policy)

$$\sum_{t=0}^{T} r(s_t, \underbrace{\pi(s_t)}_{a_t}) \qquad \text{(Finite horizon)}$$

or

$$\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(s_t))$$

Def The policy value $V_\pi(s)$ of policy π at state $s \in S$ is

$$V_\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s\right]$$

(π dependence is implicit)

(similar def'n for finite-time (undiscounted) but we'll focus on ∞-time discounted)

over usual s_t Markov chain randomness (always done, implicit)
and over $a_t \sim \pi(s_t)$ (only necessary if not a deterministic policy)

Def An optimal policy π^* means it has the highest value among all policies, regardless of the initial state s

i.e., $\forall \pi, \forall s, \; V_{\pi^*}(s) \geq V_\pi(s)$

Not obvious an optimal policy even exists! → ex: s = sport, Michael Phelps is maximal for s = swimming, Usain Bolt is maximal for s = 100m dash, but by "optimal", we mean an athlete who can beat anyone at any sport (clearly such an optimal athlete doesn't exist)

We'll show 1) an optimal policy does exist
1') ... and it is truly a (stationary) policy
2) ... and there is an optimal deterministic policy!