PGMs Book of Why - Judea Pearl(e)

https://kunalmenda.com/2019/02/21/causation-and-correlation/

From last time

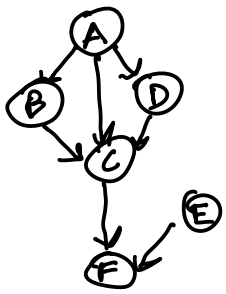
How can we determine if measuring one R.V. will reveal info about another? cond. indep.

What does "Markov" mean in "MDP"? d-separation

Def: A stochastic process is Markov w.r.t. state s iff

$$P(s_{t+1} \mid s_t, s_{t-1}, s_{t-2}, \dots s_1) = P(s_{t+1} \mid s_t)$$

Aside: Sampling from BN



Topological Sort: if $X \to Y$ then X comes before Y

A	A
B	D
D	B
C	C
E	E
F	F

Algorithm in book

Sample A from $P(A)$
Sample B from $P(B \mid A)$
D
C from $P(C \mid A, B, D)$
⋮

Markov Chain already sorted

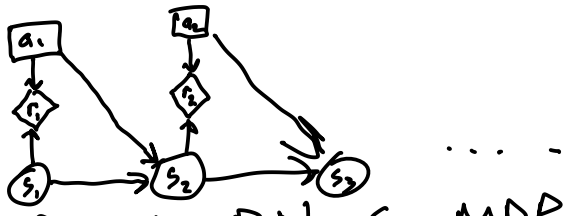
"simulate" a "trajectory" from Markov chain

sample s_1 from $P(s_i)$
while not ended
    sample s' from $P(s' \mid s)$

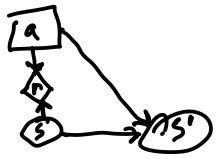
What is an MDP?
What is a policy?

Dynamic Bayes Net

Markov Decision Process



Dynamic DN for MDP



Optimization Problem

dynamics

$$\text{maximize } E\left[\sum_{t=0}^{\infty} r_t\right]$$

Decision Network

decision node

chance

utility

Keeping rewards finite

1) Finite Time

$$\sum_{t=1}^{T} r_t$$

2) Average reward

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} r_t$$

3) Discounting

γ : discount $[0,1)$

$$\sum_{t=1}^{\infty} \gamma^t r_t$$

0.9, 0.95, 0.99

if $r_t \le \bar{r}$ then $\sum_{t=0}^{\infty} \gamma^t r_t \le \frac{\bar{r}}{1-\gamma}$

4) Terminal /absorbing states

Infinite time but problem reaches terminal state (no reward, no leaving) w.p. 1

"Tuple Definition"

Markov process (S, T)

MDP defined by 5-tuple

(S, A, T, R, γ)

S - state space - set of states

e.g. $\{1,2,3\}$, {healthy, pre-cancer, cancer}, $\mathbb{R}^4$

{working, malfunctioning} $\times \mathbb{R}^6$

A - action space - set of actions

e.g. $\{1,2,3\}$, {treat, watch, wait}, $\mathbb{R}^2$,

{radar on, radar off} $\times \mathbb{R}^2$

T - transition kernel

Explicit : $T(s'|s,a)$ $\quad T: S \times A \times S \to [0,1]$ discrete

$T(s,a,s')$

Generative: $s' = G(s,a,w)$

R - reward function $\quad R: S \times A \times S \to \mathbb{R}$ or $R: S \times A \to \mathbb{R}$

$R(s,a,s')$ $\quad R(s,a)$

useful reward for distance traveled

$$R(s,a) = \mathop{\mathbb{E}}_{s' \sim T(s,a)} \left[R(s,a,s') \right]$$

γ - discount $\quad \gamma \in [0,1)$

Optimization Variable

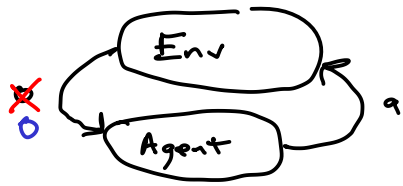
$$\underset{?}{\text{maximize}} \quad E\left[\sum \gamma^t r_t\right]$$

Open Loop

sequence of actions

Closed Loop

choose action based on observations



In MDP observations are states

"Policy" π

$$\pi : S \to A$$

Closed Loop

$$\underset{\pi: S \to A}{\text{maximize}} \; E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1}) \;\middle|\; a_t = \pi(s_t),\; s_{t+1} \sim T(s_t, a_t)\right]$$

Open Loop

$$\underset{(a_1, a_2, \dots a_T)}{\text{maximize}} \; E\left[\sum_{t=0}^{T} \gamma^t R(s_t, a_t, s_{t+1}) \;\middle|\; s_{t+1} \sim T(s_t, a_t)\right]$$

Policy Evaluation

Monte Carlo Policy Evaluation

```
function simulate (MDP, π)
    rsum = 0          sample s from s'
    for t = 1:10,000
        a = π(s)
        Sample s' from T(s,a)
        rsum += R(s,a,s')
        s = s'
    end
    return rsum
end

R = 0
for i in 1:10,000
    R += simulate(MDP, π)
end
R/N ← estimate of E[Σ r_t]
```

Stationary Policies

$(S, A, T, R, \gamma, \cancel{\mu})$

↳ $a_t = \pi(s_t)$ $\quad a_t = \pi_t(s_t)$

Stationary Policies optimal

- Infinite Horizon
- Stationary Dynamics

Example

$T = 10$

$\pi_1^*(\tilde{s}) = \text{right}$

$\pi_9^*(\tilde{s}) = \text{left}$