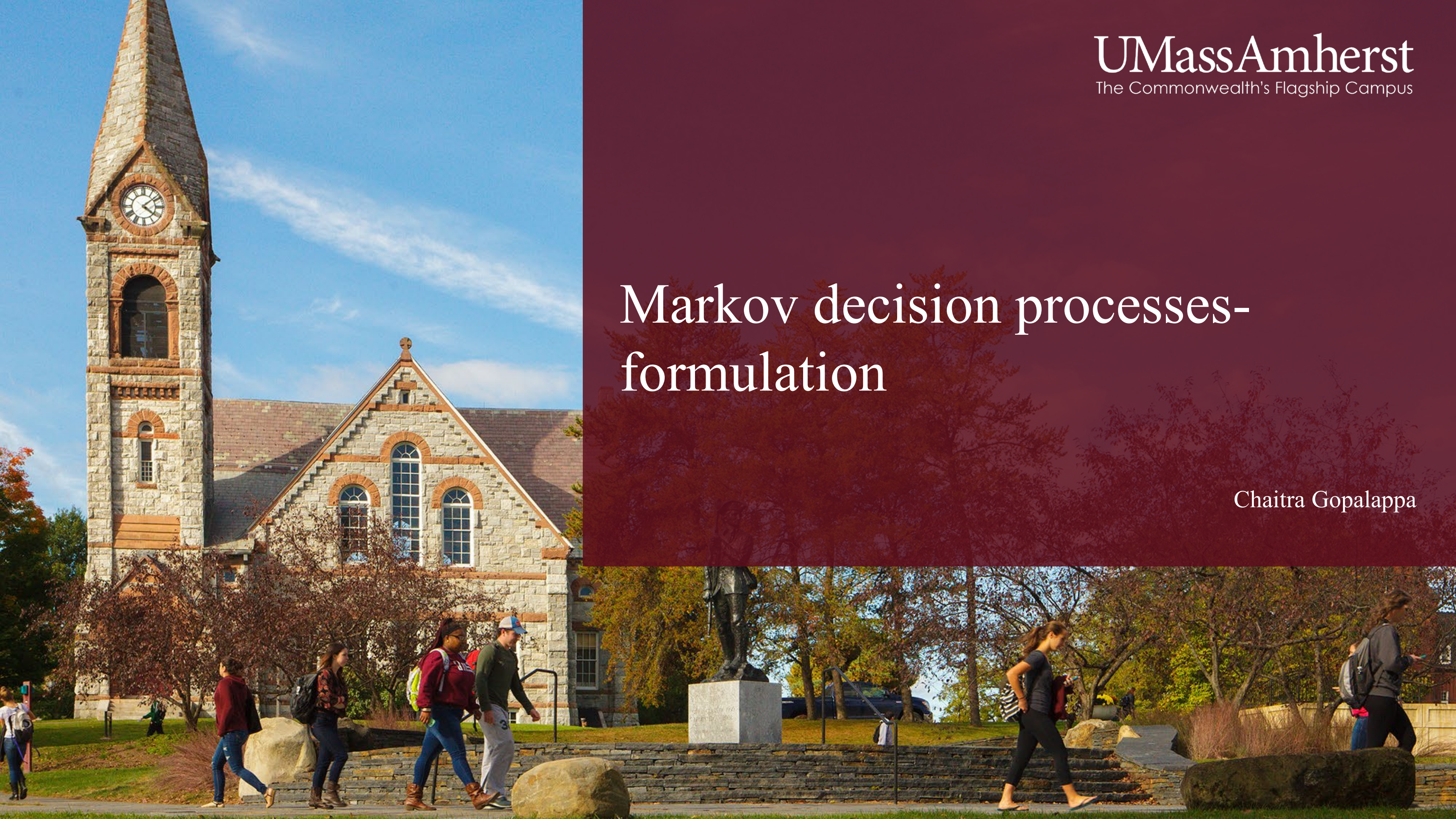

UMassAmherst
The Commonwealth's Flagship Campus

Markov decision processes-formulation

Chaitra Gopalappa

Reference

- Chapter 3, Sutton and Barto,
- https://www.andrew.cmu.edu/course/10-703/textbook/BartoSutton.pdf

Markov processes

1. A machine is inspected at the end of each day, and rated as excellent, good, fair, or inoperable. If a machine is E on day t, is found to be in G, F, or I on day t+1 50%, 30%, and 20% of the time respectively. A machine found to be in state G on day t, is found to be in G, F, and I on day t+1 30%, 40%, 30% of the times, respectively. A machine found to be in state F on day is found to be in F and I, 50% and 50% of the time respectively. A machine in I, is inoperable after.
    a. Represent the system as a Markov chain.
    b. Define the random variable, and stochastic process for this system.
    c. Write the state space, and transition probability matrix.
    d. What is average life of machine? How to calculate analytically and through simulation?

Markov processes

1. A machine is inspected at the end of each day, and rated as excellent, good, fair, or inoperable. If a machine is E on day t, is found to be in G, F, or I on day t+1 50%, 30%, and 20% of the time respectively. A machine found to be in state G on day t, is found to be in G, F, and I on day t+1 30%, 40%, 30% of the times, respectively. A machine found to be in state F on day is found to be in F and I, 50% and 50% of the time respectively. A machine in I, is inoperable after.
   a. Represent the system as a Markov chain.
   b. Define the random variable, and stochastic process for this system.
   c. Write the state space, and transition probability matrix.
   d. What is average life of machine? How to calculate analytically and through simulation?

- Let X_t be the state of the system at time t
- $\{X_t\}|_{t=0}^{\infty}$ is a stochastic process defined by the n-tuple $\{\Omega, P\}$
  - Ω is the state space; $\Omega = \{\mathrm{E, G, F, I}\}$
  - $P = \begin{bmatrix} 0 & 0.5 & 0.3 & 0.2 \\ 0 & 0.3 & 0.4 & 0.3 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 1 \end{bmatrix}$
  - $\Pr\{X_t = i | X_{t-1}, X_{t-2}, \ldots, X_0\} = \Pr\{X_t = i | X_{t-1}\}$. Thus $\{X_t\}|_{t=0}^{\infty}$ is Markov chain.
  - Average life of machine assuming a new machine starts in E, is the first passage time from E to I
    - $m_{ij} = 1 + \sum_{k \neq j} p_{ik} m_{kj}$; solve for m_{EI} as system of linear equations

Markov processes

1. A machine is inspected at the end of each day, and rated as excellent, good, fair, or inoperable. If a machine is E on day t, is found to be in G, F, or I on day t+1 50%, 30%, and 20% of the time respectively. A machine found to be in state G on day t, is found to be in G, F, and I on day t+1 30%, 40%, 30% of the times, respectively. A machine found to be in state F on day is found to be in F and I, 50% and 50% of the time respectively. A machine in I, is inoperable after.
   a. Represent the system as a Markov chain.
   b. Define the random variable, and stochastic process for this system.
   c. Write the state space, and transition probability matrix.
   d. What is average life of machine? How to calculate analytically and through simulation?

2. Same as #1, except that when machine is in I it is replaced the next day. Also, based on the condition of the system, there is a certain cost due to defective items created by the machine. G and F are associated with a cost of \$1000 and \$3000 respectively. New machines cost \$6000. What is the average cost to the system?

•

$$P = \begin{bmatrix} 0 & 0.5 & 0.3 & 0.2 \\ 0 & 0.3 & 0.4 & 0.3 \\ 0 & 0 & 0.5 & 0.5 \\ 1 & 0 & 0 & 0 \end{bmatrix}; c = [0, \$1000, \$3000, \$6000]$$

Average cost of maintaining the system = $\boldsymbol{\pi c^T}$;
$\boldsymbol{\pi}$ is the steady state vector;
$\boldsymbol{c^T}$ is the transpose of the cost vector;

Markov decision process

1. A machine is inspected at the end of each day, and rated as excellent, good, fair, or inoperable.
2. Every day the operator can choose from 3 possible actions: A{do noting, maintain, replace}

- What is optimal policy?

TPM : if do nothing

	E	G	F	I
E	0	0.5	0.3	0.2
G	0.	0.3	0.4	0.3
F	0	0	05	0.5
I	0	0	0	1

TPM : if maintain

	E	G	F	I
E	0.3	0.2	0.3	0.2
G	0.1	0.2	0.4	0.3
F	0	0.4	0.5	0.1
I	0	0	0	1

TPM : if replace

	E	G	F	I
E	1	0	0	0
G	1	0	0	0
F	1	0	0	0
I	1	0	0	0

Cost : if do nothing

	E	G	F	I
	0	1000	3000	\infty

Cost : if maintain

	E	G	F	I
	2000	3000	4000	\infty

Cost : if replace

	E	G	F	I
	6000	6000	6000	6000

- Let X_t be the state of the system at time t
- Let D_t be the decision at time t
- $\{X_t, D_t\}|_{t=0}^{\infty}$ is a Markov decision process defined by the n-tuple $\{\Omega, A, P_a, R_a\}$
  - Ω is the state space; $\Omega = \{\text{E}, \text{G}, \text{F}, \text{I}\}$
  - A is the action space ($A = \{do\ nothing\ (d), maintain(m), replace(r)\}$)
    - each element in A denoted by a
  - P_a is the TPM corresponding to an action 'a'
  - R_a is the immediate reward matrix corresponding to an action 'a'
- A policy (ρ) is a vector of size $|\Omega|$, referring to the action to be taken in corresponding state
  - e.g., $\rho = [d, d, m, r]$ implies take action d if system is state E, d if system is state G, m if system is state F, and r if system is state I
- Every policy has a value, which can be interpreted as follows. Suppose $\rho = [d, d, m, r]$
  - $P_{\rho=[d,d,m,r]} = \begin{bmatrix} use\ \ row\ correpsonding\ to\ P_d \\ use\ \ row\ correpsonding\ to\ P_d \\ use\ \ row\ correpsonding\ to\ P_m \\ use\ \ row\ correpsonding\ to\ P_r \end{bmatrix}$; similarly create cost vector $c_{\rho=[d,d,m,r]}$
  - Value of policy $\rho = \boldsymbol{\pi}_\rho \boldsymbol{c}_\rho^T$
- **To find optimal policy find the policy with the least cost (or maximum reward)**
  - Solution methods: exhaustively enumeration (In above example number of policies = 3^4)
  - Other efficient approaches: dynamic programming (model-based); reinforcement learning (model-free)

Rewriting into MDP terminologies

$P_{do-nothing}$ =

	E	G	F	I
E	0	0.5	0.3	0.2
G	0.	0.3	0.4	0.3
F	0	0	05	0.5
I	0	0	0	1

$P_{maintain}$ =

	E	G	F	I
E	0.3	0.2	0.3	0.2
G	0.1	0.2	0.4	0.3
F	0	0.4	0.5	0.1
I	0	0	0	1

$P_{replace}$ =

	E	G	F	I
E	1	0	0	0
G	1	0	0	0
F	1	0	0	0
I	1	0	0	0

Immediate reward ($r(i, a)$: reward of taking action a when system is in state i ; notice, values have been changed to negative values as we are now calling the cost components as 'reward'

$r(., a = do\_nothing)$:

	E	G	F	I
	-0	-1000	-3000	\infty

$r(., a = maintain)$:

	E	G	F	I
	-2000	-3000	-4000	\infty

$r(., a = replace)$:

	E	G	F	I
	-6000	-6000	-6000	-6000

Rewriting into MDP terminologies

$P_{do-nothing}$ =

	E	G	F	I
E	0	0.5	0.3	0.2
G	0.	0.3	0.4	0.3
F	0	0	05	0.5
I	0	0	0	1

$P_{maintain}$ =

	E	G	F	I
E	0.3	0.2	0.3	0.2
G	0.1	0.2	0.4	0.3
F	0	0.4	0.5	0.1
I	0	0	0	1

$P_{replace}$ =

	E	G	F	I
E	1	0	0	0
G	1	0	0	0
F	1	0	0	0
I	1	0	0	0

Immediate reward ($r(i, a)$: reward of taking action a when system is in state i ; notice, values have been changed to negative values as we are now calling the cost components as 'reward'

$r(., a = do\_nothing)$:

	E	G	F	I
	-0	-1000	-3000	∞

$r(., a = maintain)$:

	E	G	F	I
	-2000	-3000	-4000	∞

$r(., a = replace)$:

	E	G	F	I
	-6000	-6000	-6000	-6000

$R_{a=do_{nothing}}$:

	E	G	F	I
E	-0	-1000	-3000	∞
F	-0	-1000	-3000	∞
G	-0	-1000	-3000	∞
I	-0	-1000	-3000	∞

$R_{a=maintain}$:

	E	G	F	I
E	-2000	-3000	-4000	∞
F	-2000	-3000	-4000	∞
G	-2000	-3000	-4000	∞
I	-2000	-3000	-4000	∞

$R_{a=replace}$:

	E	G	F	I
E	-6000	-6000	-6000	-6000
F	-6000	-6000	-6000	-6000
G	-6000	-6000	-6000	-6000
I	-6000	-6000	-6000	-6000

- Let X_t be the state of the system at time t
- Let D_t be the decision at time t
- $\{X_t, D_t\}|_{t=0}^{\infty}$ is a Markov decision process defined by the 4-tuple $\{\Omega, A, P_a, r_a\}$
  - Ω is the state space; $\Omega = \{\mathrm{E}, \mathrm{G}, \mathrm{F}, \mathrm{I}\}$
  - A is the action space ($A = \{do\ nothing\ (d), maintain(m), replace(r)\}$)
    - each element in A denoted by a
  - P_a is the TPM corresponding to an action 'a'
- A policy (ρ) is a vector of size $|\Omega|$, referring to the action to be taken in corresponding state
  - e.g., $\rho = [d, d, m, r]$ implies take action d if system is state E, d if system is state G, m if system is state F, and r if system is state I

- Every policy has a value, which can be interpreted as follows. Suppose $\rho = [d, d, m, r]$

  - $P_{\rho=[d,d,m,r]} = \begin{bmatrix} use\ \ row\ correpsonding\ to\ P_d \\ use\ \ row\ correpsonding\ to\ P_d \\ use\ \ row\ correpsonding\ to\ P_m \\ use\ \ row\ correpsonding\ to\ P_r \end{bmatrix}$; similarly create cost vector $c_{\rho=[d,d,m,r]}$
  - Value of policy $\rho = \boldsymbol{\pi}_\rho \boldsymbol{c}_\rho^T$

- To find optimal policy find the policy with the least cost (or maximum reward)
  - Solution methods: exhaustively enumeration (In above example number of policies = 3^4)
  - Other efficient approaches: dynamic programming (model-based); reinforcement learning (model-free)

Problem

- University campus: People can belong to one of three disease stages susceptible, infected, recovered.
- On any given decision-making step, the university needs to decide what action to take, test once a week, test every 3 day, test every day.
- Formulate this as a MDP

- Let X_t be the state of the epidemic at time t
- Let D_t be the decision at time t
- $\{X_t, D_t\}|_{t=0}^{\infty}$ is a Markov decision process defined by the 4-tuple $\{\Omega, A, P_a, R_a\}$
- Ω is the state space;
  - we have a multivariate state $[S, I, E]$, if there are 1000 people in a population, 700 are S(susceptible), 200 are I(infected), and 100 are E(recovered), then the state of the system is $[700,200,100]$
  - $\Omega = \{[S, I, E]\}; S + I + E = N$;
- A is the action space
  - $A = \{weekly, twice\ a\ week, daily\}$)
- P_a is the TPM
  - An element $\mathrm{p}(i, a, j) =$ probability of transitioning to state j when system is in state i and action a is taken
- In addition cost of testing (action), there is an additional cost associated with the state it transitions to, so we have a reward matrix R_a
  - An element $r(i, a, j) =$ immediate reward of taking action a when system is in state i and transitioning to state j

General anatomy of MDP formulation

- Let X_t be the state of the system at time t
- Let D_t be the decision at time t
- $\{X_t, D_t\}|_{t=0}^{\infty}$ is a Markov decision process defined by the n-tuple $\{\Omega, A, P_a, R_a\}$
- Ω is the state space;
- A is the action space
- P_a is the TPM
  - An element $\mathrm{p}(i, a, j)$ = probability of transitioning to state j when system is in state i and action a is taken
- R_a is the TRM
  - An element $r(i, a, j)$ = immediate reward of taking action a when system is in state i and transitioning to state j

Inventory problem

- A factory determines, at the end of each week, whether to order inventory or not (yes/no decision) based on the inventory at the time. If the decision is yes, it orders upto K.
  - Demand ~Poisson(8000 per week)
  - Maximum inventory capacity = 50000 (K)
  - Maximum backorder capacity = 300 (B)
  - No inventory cost
  - Fixed shipping and ordering cost
  - Product varies by number of orders
- Formulate as MDP

UMassAmherst
The Commonwealth's Flagship Campus