Artificial Intelligence Methods for Social Good

M4-4 [Sequential Decision Making]:

Ecosystem Management

08-537 (9-unit) and 08-737 (12-unit)
Instructor: Fei Fang
feifang@cmu.edu
Wean Hall 4126

Outline

- Multi-Armed Bandit
- Invasive Species Management
- Wildfire Management

Learning Objective

- Understand the concept of
  - Multi-Armed Bandit (MAB)
  - Zero-regret strategy
  - Upper Confidence Bound (UCB)
  - Probably approximately correct (PAC)
- Describe how ecosystem management problems are modeled as MDPs and the key challenges
- Describe the key ideas in the solution approaches for these problems

Multi-Armed Bandit (MAB)

- K arms
- Each arm k is associated with a reward distribution R_k, with expected reward μ_k
- Gambler does not know R_k, μ_k
- In each round $t \in \{1 \dots T\}$, gambler chooses one arm k_t, and observe a reward $\hat{r}_t$ drawn from the distribution

Multi-Armed Bandit (MAB)

- Let $\mu^* = \max_k \mu_k$
- Define regret $\rho = T\mu^* - \sum_{t=1}^{T} \widehat{r_t}$
- A typical research problem in MAB: find zero-regret strategy
  - $\lim_{T\to\infty} \frac{\rho}{T} = 0$
- Probably approximately correct (PAC): with high probability, it is close to being correct

$$\Pr(error \le \epsilon) \ge 1 - \delta$$

- PAC version of zero-regret strategy

$$\Pr(\lim_{T\to\infty} \frac{\rho}{T} \le \epsilon) \ge 1 - \delta$$

Quiz 1

- If we model MAB as an MDP, which of the following representation of the state allows for the highest level of expressiveness of a policy?
- A: $s_t =< 1 >$, i.e., single state MDP
- B: $s_t =< \widehat{\mu_1}, \dots, \widehat{\mu_K} >$ where $\widehat{\mu_k}$ =average reward when k is chosen in rounds $1, \dots, t-1$
- C: $s_t =< N(1), \widehat{\mu_1}, \dots, N(K), \widehat{\mu_K} >$ where $N(k) =$ number of rounds that k is chosen in rounds $1, \dots, t-1$
- D: $s_t =< k_1, \widehat{r_1}, k_2, \widehat{r_2}, \dots, k_{t-1}, \widehat{r_{t-1}} >$ where $k_\tau =$ arm chosen in round τ

Multi-Armed Bandit (MAB)

- Model MAB as an MDP
- State $s_t =< k_1, \widehat{r_1}, k_2, \widehat{r_2}, \dots, k_{t-1}, \widehat{r_{t-1}} >$
- Action $k_t \in \{1 \dots K\}$
- Transition matrix: $P(s_{t+1}|s_t, k_t) = p_{k_t}(\widehat{r_t})$ if $s_{t+1} =< s_t, k_t, \widehat{r_t} >$
- Reward $r_t = R(s_t, a_t, s_{t+1}) = \widehat{r_t}$

Binary MAB

- K arms
- Reward is either 0 or 1, R_k: $\Pr(r = 1) = p_k, \Pr(r =$

Upper Confidence Bound in Binary MAB

- Let $N(k)$ be the number of times that k is chosen
- Let $H(k)$ be the number of times that k is chosen and reward is 1
- Let $\widehat{\mu_k} = H(k)/N(k)$, average reward when k is chosen
- Given $N(k), H(k), \widehat{\mu_k}, \delta$, we can estimate the range of μ_k, i.e., we can compute μ_{LB}^k and μ_{UB}^k such that $\Pr\left(\mu_{LB}^k \leq \mu_k \leq \mu_{UB}^k\right) \geq 1 - \delta$

Upper Confidence Bound in Binary MAB

- Chernoff-Hoeffding Bound: Let $X_1, X_2, \dots, X_n$ be independent random variables in the range $[0, 1]$ with $\mathbb{E}[X_i] = \mu$. Then for $a > 0$

$$\Pr(\frac{1}{n}\sum_{i=1}^{n} X_i \geq \mu + a) \leq e^{-2a^2 n}$$

$$\Pr(\frac{1}{n}\sum_{i=1}^{n} X_i \leq \mu - a) \leq e^{-2a^2 n}$$

- That is, with high probability, the observed average value of X_i is very close to the expected value of X_i

Upper Confidence Bound in Binary MAB

- So $\mu_{LB}^k = \widehat{\mu_k} - \sqrt{\frac{1}{2N(k)} \ln(\frac{2}{\delta})}$, $\mu_{UB}^k = \widehat{\mu_k} + \sqrt{\frac{1}{2N(k)} \ln(\frac{2}{\delta})}$ ensures $\Pr\left(\mu_{LB}^k \le \mu_k \le\right.$

Invasive Species Management

https://www.cbp.gov/travel/clearing-cbp/bringing-agricultural-products-united-states

- Invasive Species
  - Reduce biodiversity
  - E.g., Tamarisk: Native in Middle East, Outcompete native vegetation in US for water

https://www.nasa.gov/vision/earth/environment/invasive_species_MM.html

Invasive Species Management

- Manage spatially-spreading organism
- Tamarisk spread along rivers
- Seed travel along rivers (mostly downstream)
- Interventions: eradicate the invasive species and/or plant native species

Published Rule of Thumb Policies

- Intuition: upstream is important, severity of invasion is important

- Triage policy
  - Treat most-invaded edge (river reach) first
  - Break ties by treating upstream first
- Leading edge
  - Eradicate along the leading edge of invasion
- Chades, et al.
  - Treat most-upstream invaded edge first
  - Break ties by amount of invasion

MDP Model for Invasive Species Management

- State $s_t \in S$: current status of invasion
  - Tree-structured river network
    - Directed
    - Each edge $e \in E$ has H sites for trees to grow
  - Status of each site ∈ {empty, occupied by native, occupied by invasive}
  - s_t: status of all sites
- Action $a_t \in A$: management action for the invasive species
  - Action for each edge ∈ {do nothing, eradicate, plant, eradicate + plant}
  - a_t: action on all edges
  - Practical constraint: at most one edge has a non "do-nothing" action → Feasible action set A

MDP Model for Invasive Species Management

- Transition probability $P(s_{t+1}|s_t, a_t)$: describes the change of state due to the management action and natural dynamics
  - Nature
    - Natural death
    - Seed production: every occupied site may generate seed
    - Seed dispersal: generated seeds dispersed to downstream sites (upstream also possible, but less likely)
    - Seed competition: seeds dispersed to the same site compete to become established
      - Couple all edges together
      - Make probabilistic inference intractable: with current observation, infer status of sites
  - Encapsulated with an (expensive) simulator
- Reward $r_t = R(s_t, a_t)$: cost of action + penalty of invasion
  - More Tamarisk trees → higher penalty
- Policy $\pi: S \rightarrow A$:

Quiz 2

- If we use a table to store the non-zero transition probabilities $P(s_{t+1}|s_t, a_t)$ in this model, at least how many entries are needed (roughly)?
- A: $3^{2EH} \cdot EH$
- B: $3^{2EH} \cdot 4^E$
- C: $3^{EH} \cdot EH \cdot 3^H$

MDP Model for Invasive Species Management

- Optimization problem: choose optimal policy π^* to maximize discounted cumulative reward

$$J(\pi) = \mathbb{E}[\sum_{\tau=0}^{\infty} \gamma^{\tau} r_{\tau} \,|s_0, \pi]$$

- Value function $V^{\pi}(s_t) = \mathbb{E}[\sum_{\tau=t}^{\infty} \gamma^{\tau-t} r_{\tau} \,|s_t, \pi]$

MDP Model for Invasive Species Management

- Why MDP is an appropriate model for the problem?
  - MDP policy balances short-term and long-term impact of intervention
    - We can set the discount factor γ to control the balance: US Forest Service set the discount factor to be 0.96
  - MDP models uncertainty of environment

Solve the MDP

- If all elements are known: Value iteration
- Challenge: $P(s_{t+1}|s_t, a_t)$ is not given in a table, instead, we only have access to a simulator
  - Simulator: given s, a, provide a sample of s'
- Option 1: run enough simulations to get P, then run value iteration
  - Too slow, Too many samples needed (exponential)
- Option 2: directly interact with the simulator when update policy

Solve the MDP with Access to Simulator

- Slightly change the goal: Find policy $\hat{\pi}$ that is near optimal with high probability without running too many simulations
  - $\Pr\left(\left|V^*(s_0) - V^{\hat{\pi}}(s_0)\right| \leq \epsilon\right) \geq 1 - \delta$
  - Draw a polynomial number of samples from the simulator
  - Called PAC-RL (Probably approximately correct reinforcement learning)
- Equivalently: $V_{UB}(s_0) - V_{LB}(s_0) \leq \epsilon$

Solve the MDP with Access to Simulator

- Key problem: How to sample from the simulator to reduce confidence level?
- Algorithm 1: DDV
- Algorithm 2: LGCV

DDV Algorithm

- Idea 1: Optimism Principle
  - For every state s, only consider action with highest upper confidence level $Q_{UB}(s,a)$ (similar to MCTS)
- Idea 2: Value of Information
  - $\Delta V(s_0) = V_{UB}(s_0) - V_{LB}(s_0)$
  - DDV=$\Delta_{s,a}\Delta V(s_0)$=$\Delta V(s_0)$-$\Delta_{\mathrm{s,a}} V'(s_0)$
    - For every (s,a), how much $\Delta V(s_0)$ will change as a result of sampling (s,a)
  - Compute/Estimate DDV for every (s,a) pair satisfying Optimism Principle, choose (s,a) with highest DDV
- The key is to estimate $V(s_0)$!

DDV Algorithm

- Idea 3: Optimal Sampling for Policy Evaluation
  - Goal: Estimate $\mathrm{V}^{\pi}(\mathrm{s}_0)$ through simulator so that the estimated value $\hat{V}^{\pi}(s_0)$ satisfy
  $$\Pr(\left|\hat{V}^{\pi}(s_0) - \mathrm{V}^{\pi}(\mathrm{s}_0)\right| \leq \epsilon) \geq 1 - \delta$$
  - Compute occupancy measure $u^{\pi}(s)$: the discounted probability that a policy π visits state s
  - Use Extended Value Iteration: Sample (s, a) in proportion to $u^{\pi}(s)^{\frac{2}{3}}$
  - Or use Monte Carlo Trials: Sample (s, a) in proportion to $u^{\pi}(s)$

DDV Algorithm

- Repeat
  - Sample (s, a) with highest estimated DDV
- Until width of estimated confidence interval $\leq \epsilon$

- Confidence interval is estimated using Extended Value Iteration algorithm based on optimal sampling

LGCV Algorithm

- Key idea: Improve DDV by improving the way to compute confidence intervals

- Two different ways to compute confidence interval
  - Extended Value Iteration (EVI)
  - Monte Carlo (MC) samples drawn according to a fixed policy

- LGCV
  - Use EVI to compute $V_{UB}(s_0)$
  - Use EVI+MC to compute $V_{LB}(s_0)$
  - In each iteration
    - Either Draw a minibatch of samples to improve EVI interval
    - Or Draw a minibatch of samples to improve MC interval

Evaluate the algorithms

- Evaluate different policies with the simulator: MDP based policies improves rule-of-thumb policies by $\approx 25\%$!

Wildfire Management

- Ideal state: a natural state with large pine trees, open understory, frequent ground fires that remove understory plants but do not damage trees
- Lack of controllable fires leads to densely distributed pine trees, heavy accumulation fuels in understory, high risk of large catastrophic fires that kill all trees and damage soils
- Selectively extinguish natural wildfires or even conduct prescribed burns to reduce risk

https://www.fs.usda.gov/detail/r6/landmanagement/resourcemanagement/?cid=stelprdb5423597

https://www.tahoedailytribune.com/news/lake-tahoe-forest-service-to-conduct-fall-prescirbed-burns-and-wildfire-management/

Wildfire Management

- Study area: Deschutes National Forest
- Management question: When lightning ignites a fire, should we let it burn or extinguish it?

Wildfire Management

- How can AI help?
  - Develop simulators
  - Evaluate rule-of-thumb policies
  - Design better policies

Wildfire Management

- Formulate the problem as an MDP
  - State s_t:
    - Grid representation of the area (4000 cells)
    - For each grid cell: # and age of trees, fuel load
    - s_t: state of all cells, 25^{4000} states!
  - Action a_t: {LetBurn, Suppress} when there is a fire ignition
  - Reward $r_t = R(s_t, a_t, l_t)$: cost of lost timber value, cost of fire suppression
  - Transition function $P(s_{t+1}|s_t, a_t) = P(l_t|s_t, a_t)P(s_{t+1}|s_t)$
  - Optimization goal: $\max_{\pi} \mathbb{E}[\sum_t \gamma^t r_t]$

Solve the MDP

- Possible approaches
  - Policy Gradient
    - Represent policy as a parameterized function $\pi(s;\theta)$
    - Estimate gradient $\nabla_\theta J(\pi(s;\theta))$ via Monte Carlo trials
    - Perform gradient ascent
    - Does work well: noisy gradient, hard to stabilize with limited samples
  - Bayesian Optimization with regression tree (SMAC)

Practical Challenge

- Visualize rollout policies of MDP (MDPVis.github.io)
  - How Cumulative Timber Loss increases over time in different trials given the policy
  - Debug the system
  - Interpret policies and communicate with stakeholders

Extensions

- Multiple owners of forest, multiple fire mangers

Acknowledgment

- Slides are prepared based on Prof. Thomas Dietterich's presentation slides