Markov Decision Process

Outline

Introduction

Parts of MDP

Policy & Rewards

Partially Observable MDP

Applications

10% of MSc is in KLA Tencor!

In fact, it’s a combination of 2 things:

- Random chance
- Actions of the student

– – –

- Situations like this can still be analyzed and optimized.
- You can optimize your actions to find the best outcome in spite of the randomness.
- One way to do this is by using a **Markov Decision Process.**

Parts of MDP

State

The range (possible values) of the random variables in a **stochastic process** is called the **state space** of the **process**.

Action

Just like there is a large set of possible states, there is also a large set of possible actions that might be taken. The current state often influences which actions are available.

Probability Distribution

It is used to determine the transition from the current state to the next state(often contained in a matrix).

Reward

Artificially generated value calculated based on value of next state compared to current state. More favourable states generate better rewards.

Time Step

- The initial state is chosen randomly from the set of possible states.
- Based on that state, an action is chosen.
- The next state is determined based on the probability distribution for the given state and the action chosen.
- A reward is granted for the next state.
- The entire process is repeated from step 2.

Cumulative Reward

To decide what to do, the agent compares different sequences of rewards. The most common way to do this is to convert a sequence of rewards into a number called the **value** or the **cumulative reward**.

To do this, the agent combines an immediate reward with other rewards in the future.

Total Reward

The value is the sum of all the rewards.

$$V = \sum_{i=1}^{\infty} r_i$$

Average Reward

In this case, the agent's value is the average of its rewards, averaged over for each time period.

$$V = lim_{n \to \infty} (r_1 + \cdots + r_n)/n.$$

Discounted Reward

Under this criterion, future rewards are worth less than the current reward.

$$V = r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots + \gamma^{i-1} r_i + \cdots$$

where γ, the **discount factor**, is a number in the range

$0 \leq \gamma < 1$.

Policy

A policy β is a set of numbers $\beta = \{\beta i(a), a \in A, i = 1,...,M\}$ with the interpretation that if the process is in state i, then action a is to be chosen with probability $\beta i(a)$.

$0 <= \beta i(a) <= 1$, for all i, a

$\sum_a \beta i(a) = 1$, for all i

Under any given policy $\boldsymbol{\beta}$, the sequence of states $\{X_n,\ n = 0, 1, \ldots\}$ constitutes a Markov chain with transition probabilities $P_{ij}(\boldsymbol{\beta})$ given by

$$P_{ij}(\boldsymbol{\beta}) = P_{\boldsymbol{\beta}}\{X_{n+1} = j | X_n = i\}^*$$

$$= \sum_a P_{ij}(a)\beta_i(a)$$

For any policy $\boldsymbol{\beta}$, let π_{ia} denote the limiting (or steady-state) probability that the process will be in state i and action a will be chosen if policy $\boldsymbol{\beta}$ is employed. That is,

$$\pi_{ia} = \lim_{n\to\infty} P_{\boldsymbol{\beta}}\{X_n = i, a_n = a\}$$

The vector $\pi = (\pi_{ia})$ must satisfy

(i) $\pi_{ia} \geqslant 0$ for all i, a,

(ii) $\sum_i \sum_a \pi_{ia} = 1$,

(iii) $\sum_a \pi_{ja} = \sum_i \sum_a \pi_{ia} P_{ij}(a)$ for all j (4.33)

$$\beta_i(a) = P\{\boldsymbol{\beta} \text{ chooses } a | \text{state is } i\}$$
$$= \frac{\pi_{ia}}{\sum_a \pi_{ia}}$$

Optimal Policy

Let $V^{\pi}(s)$ be the expected value of following π in state s. This specifies how much value the agent expects to receive from following the policy in that state.

Policy π is an **optimal policy** if there is no policy π' and no state s such that $V^{\pi'}(s) > V^{\pi}(s)$.

Partially Observable Markov Decision Process

In these scenarios, the system does not know exactly what state it is currently in, and therefore has to guess.

“I’m going in the right direction”

vs

“I think I’m going in the right direction”

MDP vs POMDP

Belief state : The state the system believes it is in. The belief state is a probability distribution.

“I think I’m going in the right direction” might really mean:
- 80% chance this is the right direction
- 15% mostly-right direction
- 5% completely wrong direction

MDP vs POMDP

Observations : After the system takes an action based on its belief state, it observes what happens next and updates its belief state accordingly.

If you took a right turn expecting a freeway and didn’t see the freeway you expected, you would then change your probability distribution for whether you were going in the right direction.

Applications

Robotics

Automated systems

Medicine

Biology

Networks

Linguistics

Machine Learning

Queries

References

Introduction to Probability Models, Sheldon M Ross

www.artint.info

www.web.stanford.edu

Thank You!