ゲーム理論と最適化手法
第 2 回: 強化学習手法の基礎

上田 俊

佐賀大学理工学部

Email: sgrueda@cc.saga-u.ac.jp
Web: https://www.fu.is.saga-u.ac.jp/sgrueda/

2019 年 10 月 8 日

始める前に...

- 授業の途中で時々質問します.
- 質問にこたえてくれた学生に 1 上田ポイント (略してウエポン) をプレゼント!
- ついでに飴玉もあげます.

- 第 8 回 (最適化手法終了回) までに 5 ウエポンをゲットすると，中間レポートに 20 点加算します!!

モンティ・ホール問題

- ゲームのルール
  1. 3 つの箱に (火竜の紅玉，ドキドキノコ，ドキドキノコ) がランダムに入っている.
  2. あなたは箱をひとつ選ぶ.
  3. アイルーが残りの箱のうち，ドキドキノコが入っている箱をひとつ開ける.
  4. あなたは残りふたつの箱のうち，どちらを開けるか選び直して良い.
- 変える or 変えない?

事象と確率

- 事象 X: 選んだ箱の中身
  ($X =$ 火竜の紅玉, $X =$ ドキドキノコ)
- 確率 $P(X)$: X がどのくらいの割合で発生するか.
  - $P(X =$ 火竜の紅玉$) = \frac{1}{3}$
  - $P(X =$ ドキドキノコ$) = \frac{2}{3}$

同時確率

- 1 回目の箱の中身を X_{1st}，2 回目の箱の中身を X_{2nd} とする.
- $P(X_{1st}, X_{2nd})$: 事象 X_{1st} と事象 X_{2nd} が同時に起きる確率
  - $P(X_{1st} =$ 火竜の紅玉$, X_{2nd} =$ 火竜の紅玉$)$: どちらの中身も当たりである事象の確率
  - $P(X_{1st} =$ ドキドキノコ$, X_{2nd} =$ ドキドキノコ$)$; どちらの中身も外れである事象の確率
- このときは箱を {変える，変えない} の選択は表現されていない.

条件付き確率

- あなたの行動: $A \in \{$ 変える, 変えない $\}$
- $P(X \mid A)$: 行動 A の時に事象 X が起きる確率
- では，~~物欲センサーに負ける~~ 1 回目の箱の中身は火竜の紅玉だが，箱を変えた結果 2 回目の箱の中身がドキドキノコである事象の確率は?
- $P(X_{1st} =$ 火竜の紅玉, $X_{2nd} =$ ドキドキノコ $\mid A =$ 変える$)$

知りたい答え

- $P(X_{2nd} = \text{火竜の紅玉} \mid A = \text{変える})$ と $P(X_{2nd} = \text{火竜の紅玉} \mid A = \text{変えない})$ のどちらが大きいか?
- 確率の乗法定理，周辺化を使って求める.

周辺化

- 同時確率において，一方の事象について全ての可能性を足し合わせてその変数を消去すること
- $P(X_{2nd} \mid A) = \sum_{X_{1st}} P(X_{1st}, X_{2nd} \mid A)$

乗法定理

- 同時確率 $P(X_{1st}, X_{2nd})$ と条件付き確率 $P(X_{2nd} \mid X_{1st})$ との間には，以下の関係が成立する:

$$P(X_{1st}, X_{2nd}) = P(X_{2nd} \mid X_{1st}) \cdot P(X_{1st})$$

- よって，$\sum_{X_{1st}} P(X_{1st}, X_{2nd} \mid A) = \sum_{X_{1st}} P(X_{2nd} \mid X_{1st}, A) \cdot P(X_{1st} \mid A)$ が成り立つ.

$A =$ 変える の場合

- $P(X_{2nd} = \text{火} \mid A = \text{変}) =$
  $P(X_{2nd} = \text{火} \mid X_{1st} = \text{火}, A = \text{変})$
  $\cdot\, P(X_{1st} = \text{火} \mid A = \text{変})$
  $+\, P(X_{2nd} = \text{火} \mid X_{1st} = \text{ド}, A = \text{変})$
  $\cdot\, P(X_{1st} = \text{ド} \mid A = \text{変})$
  $= 0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3}$

$A =$ 変えない の場合

- $P(X_{2nd} = \text{火} \mid A = \neg\text{変}) =$
  $P(X_{2nd} = \text{火} \mid X_{1st} = \text{火}, A = \neg\text{変})$
  $\cdot\, P(X_{1st} = \text{火} \mid A = \neg\text{変})$
  $+\, P(X_{2nd} = \text{火} \mid X_{1st} = \text{ド}, A = \neg\text{変})$
  $\cdot\, P(X_{1st} = \text{ド} \mid A = \neg\text{変})$
  $= 1 \cdot \frac{1}{3} + 0 \cdot \frac{2}{3}$

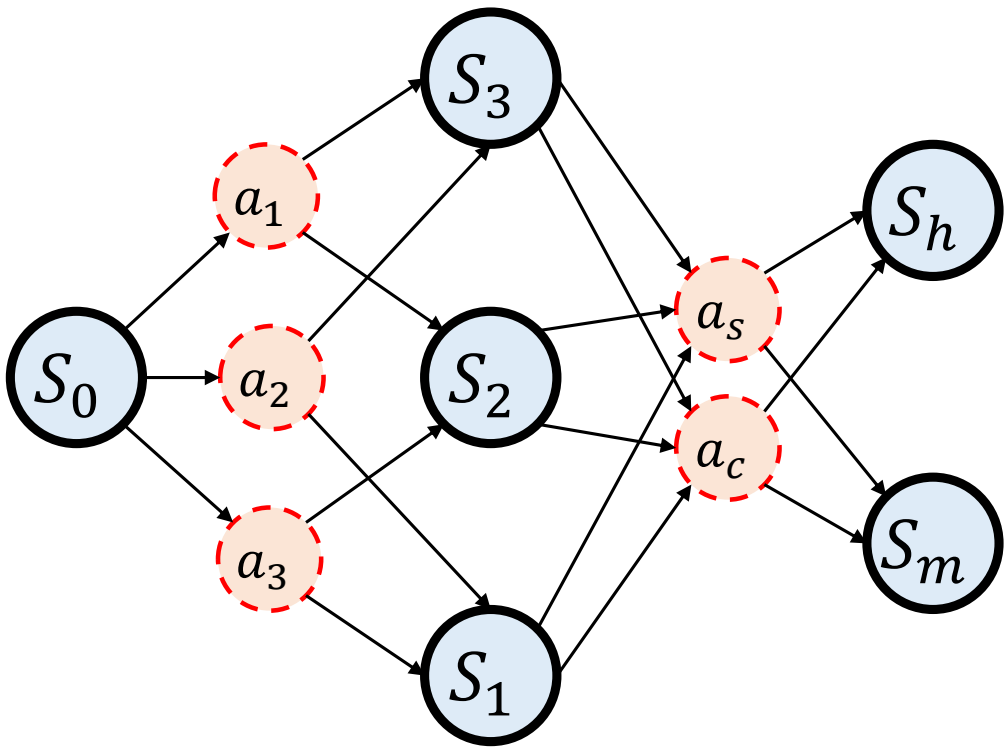
変えた方が良い!

- $P(X_{2nd} = \text{火} \mid A = \text{変}) = \dfrac{2}{3}$

  $> P(X_{2nd} = \text{火} \mid A = \neg\text{変}) = \dfrac{1}{3}$
- これをコンピュータに学習させるにはどうすればよいだろうか?
- 機械学習 $=$ マルコフ決定過程 $+$ 政策の試行錯誤による修正

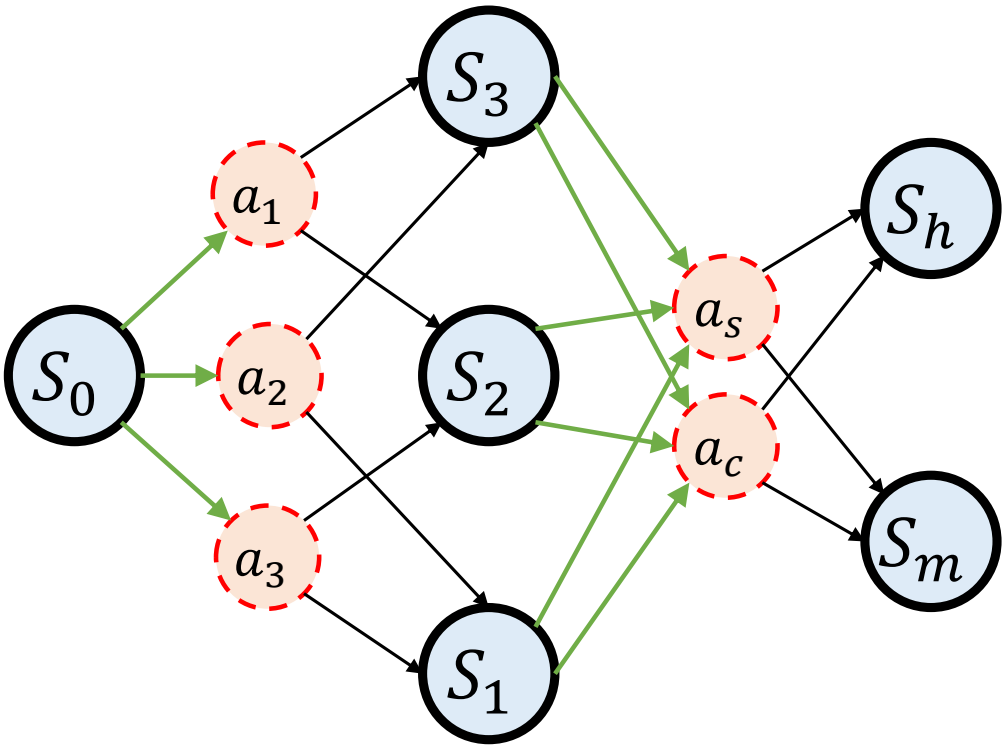
マルコフ決定過程

- Markov decision process
- 状態遷移が確率的に生じる動的システム (確率システム) の確率モデル



政策

- $\pi(s, a)$: 状態 s のときに，行動 a を取る確率
- $\pi(s_i, a_s) = 0$, $\pi(s_i, a_c) = 1$ を学習する．



25 を言った方が負け

- 2 人で交代に，1 から順に 25 までの数を言う.
- 言う数の個数は，1 個，2 個，3 個のいずれか好きなのを選んでよい.
- 最後に 25 を言った方が負け.

来週

- Q 学習を使って，先ほどのゲームの勝ち方を学習する.
- Excel を使います.
  - https://www.cc.saga-u.ac.jp/mees/entry2.html に自分の PC にインストールする方法が載っています.

第 2 回小レポート課題

- $P(X_{2nd} = \text{火} \mid X_{1st} = \text{ド}, A = \text{変}) = 1$ となる理由を説明しなさい.
  - 式変形はしない. よく状況を考えるとわかる.
  - 図や表を書いてみるとわかりやすくなるかもしれない.
- 「25 を言った方が負け」ゲームの途中で相手は 22 と言ってきた. あなたは何と答えますか？理由とともに説明しなさい.