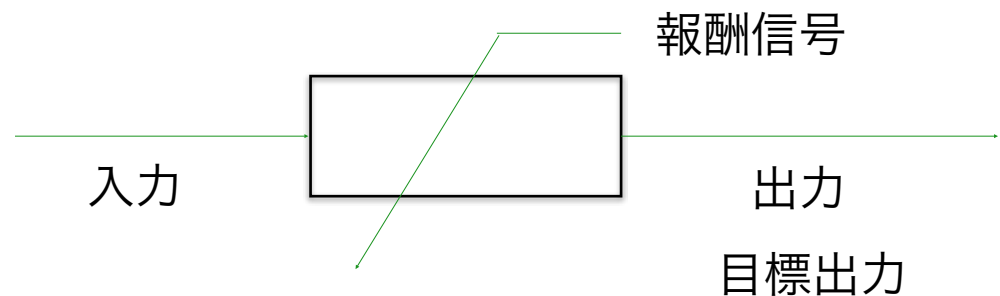


機械学習の種類

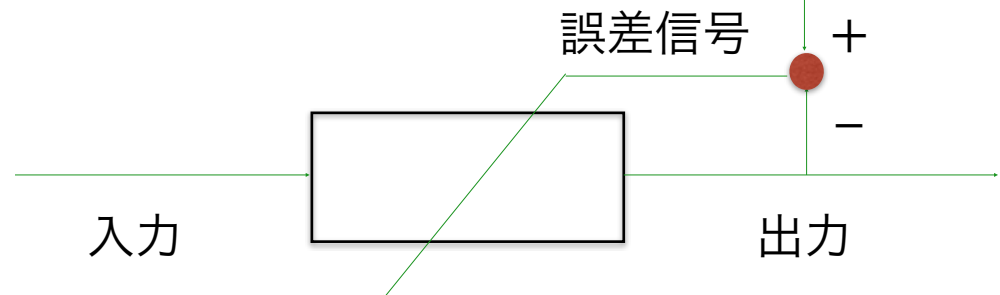
- 1) 教師無し学習
ex) Bayes 推定
ex2) 連想記憶



- 2) 報酬に基づく学習
ex1) 山登り法
ex2) 強化学習



- 3) 教師あり学習
ex) 神経回路網モデル
(最近はDeepLearning)



観戦速報・グーグルの囲碁AI「AlphaGo」が最強の棋士を破った日

グーグルの人工知能（AI）と、世界最強棋士のひとりとの5連戦。接戦となったその第1戦は、人がAIに敗れるという結果に終わった。2016年3月9日は、これからのAIを語るうえで重要な日となる。

PHOTOGRAPHS COURTESY OF GOOGLE
TEXT BY WIRED.jp_ST



AlphaGo

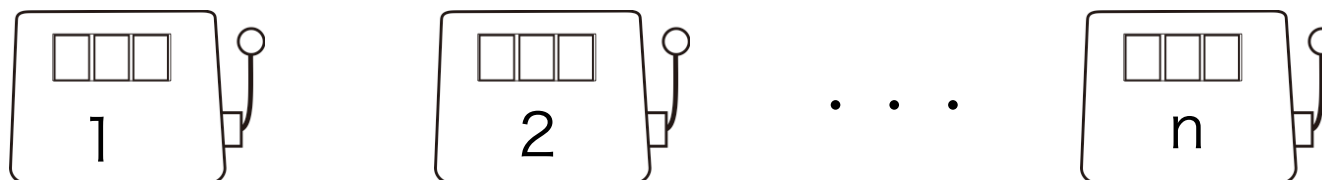
- 2016/3: トップ棋士(李世ドル, 9段)に4勝1敗
- 2017/5/24: 柯潔9段に3勝0負
- Deep Learningと強化学習を利用

強化学習

簡単な例

n-armed bandit problem (n本腕問題)

スロットマシン



どのように選ぶと一番もうかる？

アクション(action)の決め方・・・

ポリシー(policy)

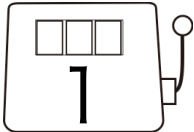
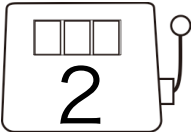
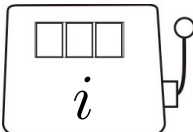
- 1) はじめはランダムポリシー
・・・知識の探索(exploration)
- 2) 探索が進んだらポリシーを変更
・・・知識の活用(exploitation)

Exploration or Exploitation

探索をいつまで続けるか } は学習における大問題
経験をどのくらい信じるか }

exploration or exploitation
経験 or 活用

探索 (exploration)

選んだマシン			...	
	r_1^1	r_1^2	...	r_1^i
出たコイン数	r_2^1	r_2^2		r_2^i
(報酬 : reward)	\vdots	\vdots		\vdots
	r_n^1	r_n^2	...	r_n^i
マシン i の 価値(value)	$V(i) = \bar{r}_n^i = \frac{r_1^i + r_2^i + \dots + r_n^i}{n}$			

注) 各マシンから出たコインの平均値を求めるには
履歴 $r_1 \sim r_n$ を全て記憶しておくこと必要はない!

平均値の求め方

平均値とサンプル数のみ記憶しておけばOK!!

$$\bar{r}_1 = r_1$$

$$\bar{r}_{n+1} = \bar{r}_n + \frac{1}{n+1} (r_{n+1} - \bar{r}_n) \quad (n > 2)$$

これまでの平均値との差

n が大きくなると r_{n+1} が \bar{r}_n に及ぼす影響は小

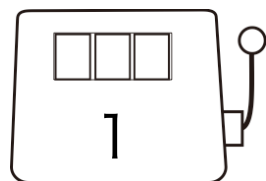
問: 上式を証明しなさい

変形版

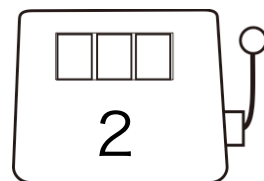
$$\bar{r}_{n+1} = \bar{r}_n + \frac{1}{k} (r_{n+1} - \bar{r}_n) \quad (k: \text{定数})$$

- 最近の約 k サンプルの平均!
- 真の平均値が時間とともに変わる場合はこちらがbetter

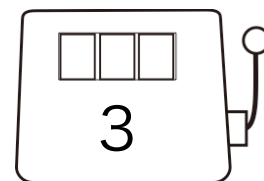
例) n-armed bandit problem (n本腕問題)



マシンの
価値(value) $V(1)=?$



$V(2)=?$



$V(3)=?$

どれを選ぶ?

1) はじめはランダムポリシー

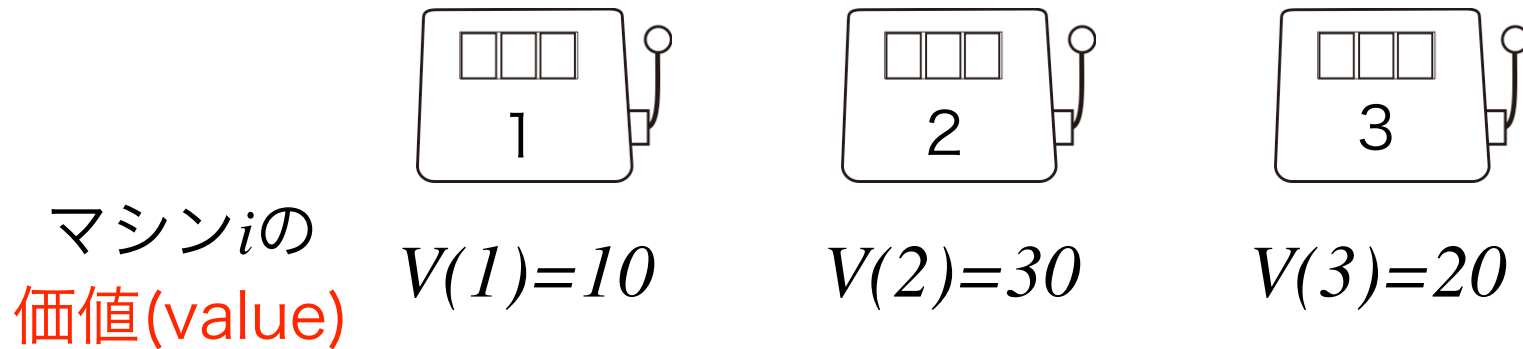
…知識の探索(exploration)

ex) 各マシンの価値(value)はこれまでの報酬の平均値で決定

2) 探索が進んだらポリシー変更

…知識の活用(exploitation)

知識の活用(exploitation): greedy法



例1) greedy法：一番いいと思う行動を選択

$$a = \arg \max_i V(i)$$

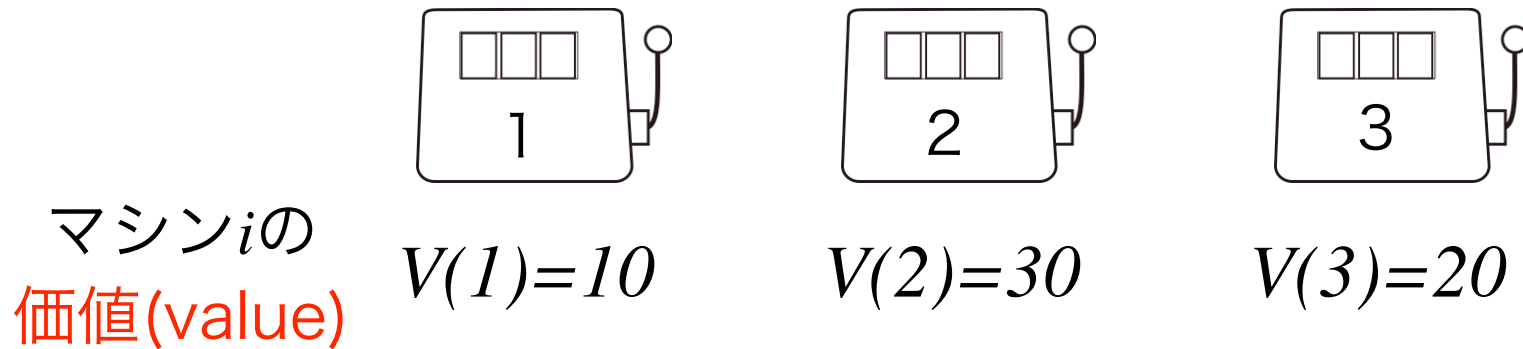
arg max は最大の時の*i*

問1: a を求めなさい

問2: $\max V(i)$ はいくらか?

問3: greedy法の欠点を述べなさい

知識の活用(exploitation): ε -greedy法

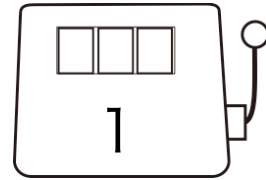


例2) ε -greedy法 : 主にgreedy法, たまに random探索

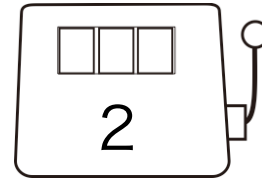
$$a = \begin{cases} \arg \max_i V(i) & \dots (\text{確率 } 1 - \varepsilon) \\ \text{random from } i \in \{1, \dots, n\} & \dots (\text{確率 } \varepsilon) \end{cases}$$

知識の活用(exploitation): softmax法

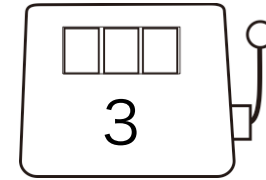
よく使われる



$V(1)=10$



$V(2)=30$



$V(3)=20$

マシン i の
価値(value)

例3) softmax法: 価値 (value)に応じて選択確率を変える

方法1)

$$\underbrace{P(a)}_{a \text{ を選ぶ確率}} = \frac{V(a)}{\sum_{i=1}^n V(i)}$$

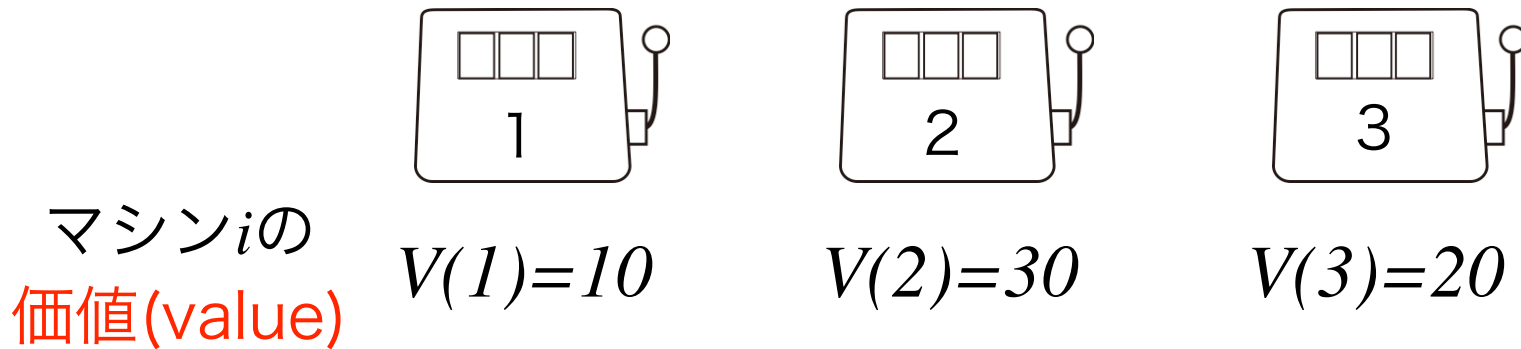
問1: $P(1)$ を求めなさい

問2: $P(1) + P(2) + P(3) = 1$ であることを確認しなさい

問3: この計算方法の欠点は?

$V(i) > 0$ となるタスクにしか使えない

知識の活用(exploitation): softmax法



例3) softmax法：価値 (value)に応じて選択確率を変える

方法2)

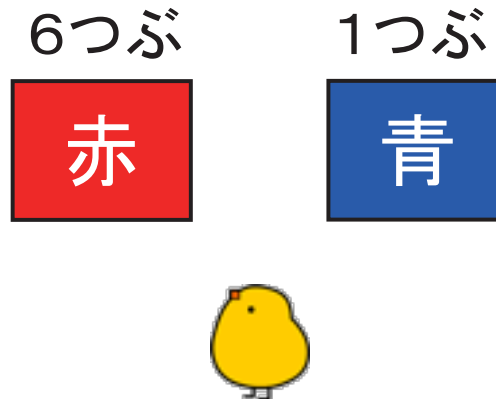
$$P(a) = \frac{e^{\frac{V(a)}{T}}}{\sum_{i=1}^n e^{\frac{V(i)}{T}}} \quad (T \text{ は正の定数})$$

問1: $P(1)$ を求めなさい

問2: $P(1) + P(2) + P(3) = 1$ であることを確認しなさい

問3: $n=2$, $V(2)=30$ のとき、 $P(1)$ が $V(1)$ に応じてどう変化するかグラフに示しなさい。また、 T の値に応じて $P(1)$ がどう変化するか述べなさい。

Exploration or Exploitation



選択確率は 赤 : 青 = 6 : 1

Q. なぜ100% 赤を選ばない??

進化の上で得た知恵?

- ・ヒトや多くの動物の行動決定は softmax法
- ・マッティング則：多くの動物が示す行動選択確率は報酬比で決まる