T-720-ATAI-2016

Advanced Topics in Artificial Intelligence:

Reinforcement Learning

Jordi Bieger

School of Computer Science | Center for Analysis and Design of Intelligent Agents

Intelligence and learning

- General intelligence is the ability to achieve a wide variety of goals in complex environments, even when they were not anticipated.

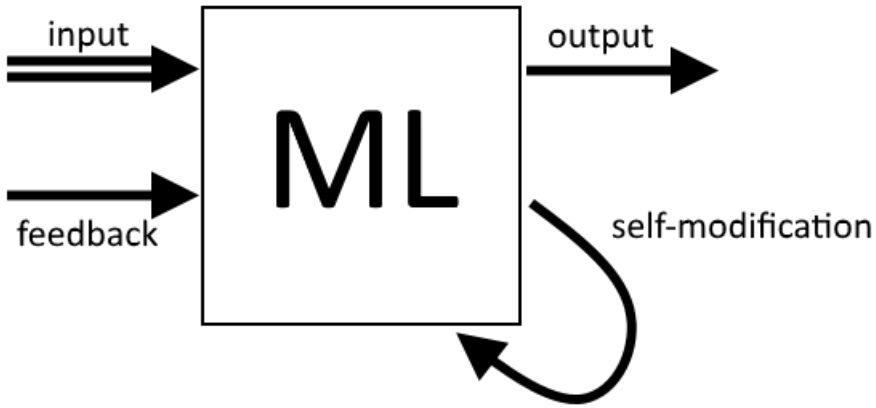
- Learning is a long term change in knowledge, skills or behavior that comes from experience or reflection.

Purpose

- Behavior can only said to be intelligent if there is some kind of purpose or reason driving it.
- The purpose of learning, in turn, is to better satisfy such a drive in the future.
- The goal of machine learning then, is to tune the algorithm's parameters (or code) in such a way that future performance is improved.
- However, in many settings current performance, and lasting consequences should not be neglected.

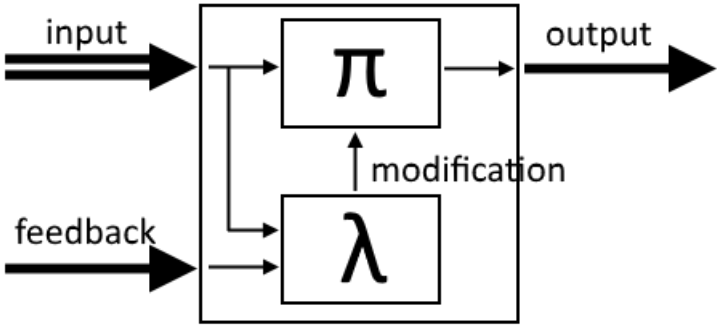
Machine learning

- A machine learning agent receives some regular inputs, as well as possibly some extra feedback.



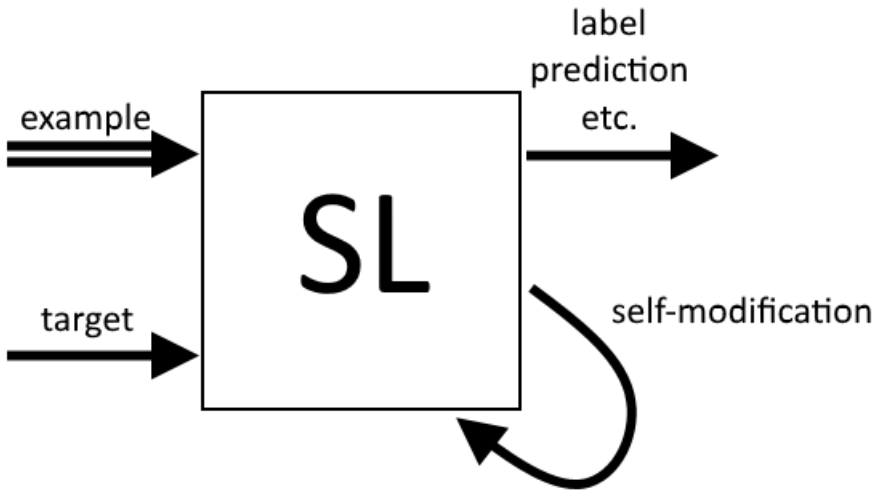
Machine learning

- The learning algorithm then uses this feedback to adjust the policy that computes the output.



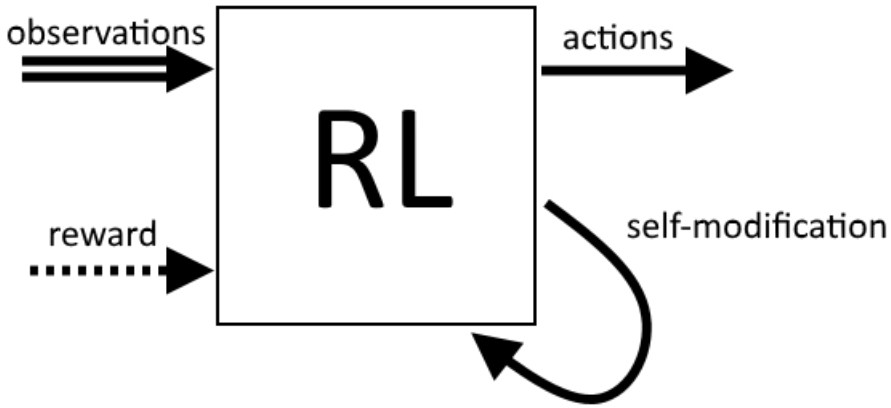
Supervised learning

- In supervised learning the feedback is usually the output that should have accompanied the input: i.e. the ground truth.



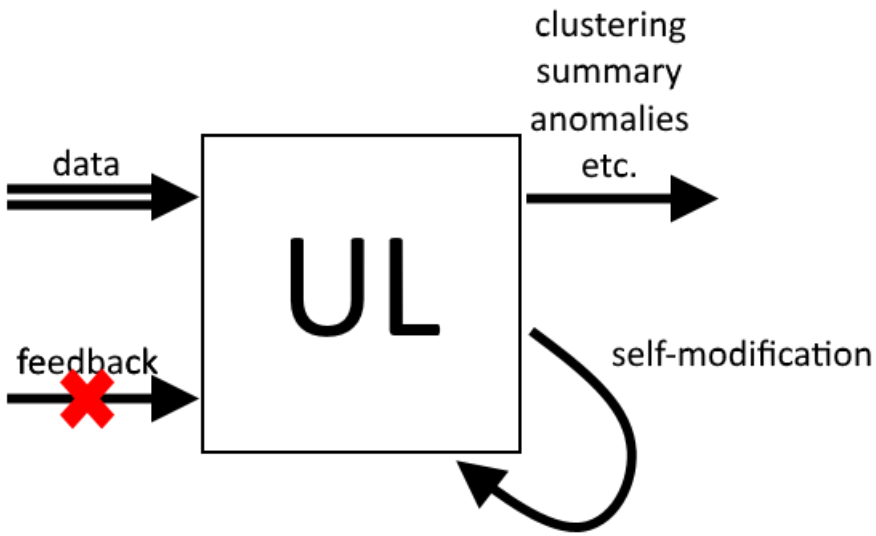
Reinforcement learning

- In reinforcement learning, the feedback merely consists of a scalar value that resembles a reward (or punishment).



Unsupervised learning

- In unsupervised learning, there is no external feedback. The objective function takes only data into account.



Machine learning paradigms

	Supervised	Reinforcement	Unsupervised
Feedback	target	reward	none
Clarity	yes	no	yes
Online	rarely	yes	sometimes
Stochastic	rarely	regularly	rarely
Delays	no	yes	-
Distribution	i.i.d.	sequential	i.i.d.
Autonomous	no	mostly	yes

Reinforcement learning and AGI

- Reinforcement learning is often considered the most natural paradigm for AGI.
  - Supervised learning requires too much supervision.
  - Unsupervised learning is hard to evaluate/control externally.

  - Reinforcement learning is a common paradigm in nature.
  - It makes it relatively easy to specify tasks that we don't know how to solve, and even give hints.
  - The rewards correspond roughly to an innate drive that all intelligent systems must have.
  - Reinforcement learning deals with autonomy, delayed rewards, online learning and active learning while requiring no initial information about the task-environment.

Markov decision process (MDP)

- Task environments are usually modelled as Markov decision processes in the RL paradigm.
- An MDP contains:
  - a set of environment states $\mathcal{S}$
  - a set of actions $\mathcal{A}$
  - (optionally) an action availability function $\mathcal{A}(s)$
  - a transition function $\mathcal{T}$
  - a reward function $\mathcal{R}$

Synchronous interaction

- Interaction between the agent and environment is usually assumed to be synchronous.
  1. The agent senses the state of the environment.
  2. The agent responds with an action.
  3. The environment computes the next state and a reward.
  4. The reward is communicated to the agent.
  5. And the cycle repeats…

Determinism vs. stochasticity

- The transition and reward functions can either be deterministic or stochastic.
- Deterministic functions directly compute the next state or the reward, whereas stochastic functions compute distributions over next states and rewards.
- Stochastic transition functions are often notated as functions that take three parameters: $\mathcal{T}(s, a, s') = \mathbb{P}(S_{t+1} = s' | S_t = s \ \wedge A_t = a)$.

Reward sources

- Rewards can be influenced from different sources:
  - R(s): e.g. in chess you can judge how well you're doing solely from the board positions.
  - R(s,a) or rarely R(a): e.g. in our mouse environment we might say that every action takes a certain amount of energy. We could augment it by adding actions to sprint or jump that might use more energy.
  - R(s,a,s'): e.g. in a coin flipping game where you have to bet if the next flip will be the same as the last.

Markov property

- The Markov property says that the current state contains all of the relevant information about the history of the world in order to determine the next state and reward.
- “the future is independent of the past given the present”
  - $\mathbb{P}(S_{t+1}|S_t, A_t) = \mathbb{P}(S_{t+1}|S_0, A_0, \ldots, S_t, A_t)$

Goal of learning

- To find a policy that optimizes future expected reward.
- Approaches:
  - Policy search: $\pi(s) \rightarrow a$
  - Value iteration: $Q(s,a) \rightarrow v$
  - Model-based: $T(s,a) \rightarrow s'$ and $R(s,a) \rightarrow r$

- Bellman equations:
  - $V_\pi(s) = \sum_{a \in \mathcal{A}(s)} \pi(a|s) Q_\pi(s,a)$
  - $Q_\pi(s,a) = \mathbb{E}_\pi[r + \gamma Q_\pi(s',a')]$

Temporal difference learning

- Greedy policy:
  - $\pi(a|s) = f(x) = \begin{cases} 1, & a = \max\limits_{a^*} Q(s, a^*) \\ 0, & \text{otherwise} \end{cases}$
- Q-learning (off-policy):
  - $Q(s,a) \overset{\Delta}{\rightarrow} \alpha \left(r + \gamma \max\limits_{a'} Q(s', a') \right)$

Exploration vs. exploitation

- ε-greedy policy:
  - $\pi(a|s) = \begin{cases} 1-\epsilon, & a = \max_{a^*} Q(s,a^*) \\ \frac{\epsilon}{|\mathcal{A}(s)|}, & otherwise \end{cases}$

On-policy vs. off-policy

- Q-learning (off-policy):
  - $Q(s,a) \overset{\Delta}{\rightarrow} \alpha \left(r + \gamma \max_{a'} Q(s',a') \right)$
- Sarsa (on-policy):
  - $Q(s,a) \overset{\Delta}{\rightarrow} \alpha \left(r + \gamma Q(s',a') \right)$

Pros and cons

- On-policy vs. off-policy?
- Exploration vs. exploitation?
- Markov decision processes?
- Reinforcement learning?

Questions?