COMP 4752
Computational Intelligence

Lecture 14
Reinforcement Learning

Reinforcement Learning

- Learning via interaction with environment
- No explicit ‘teacher’ for learning
- Agent can act with motor function, and perceive with sensors
- Learning via interaction teaches us about cause and effect, consequences of actions, what to do to achieve goals

Reinforcement Learning (RL)

- RL is learning what to do in order to maximize a numerical reward signal
  - Map states of environment to actions
- The learner is not told which actions to take, as they are in most forms of ML
- The learner must discover which actions yield the most reward by trying them
- Actions affect not only immediate reward, but transition the state and affect future reward

Reinforcement Learning (RL)

- The two most important features of RL

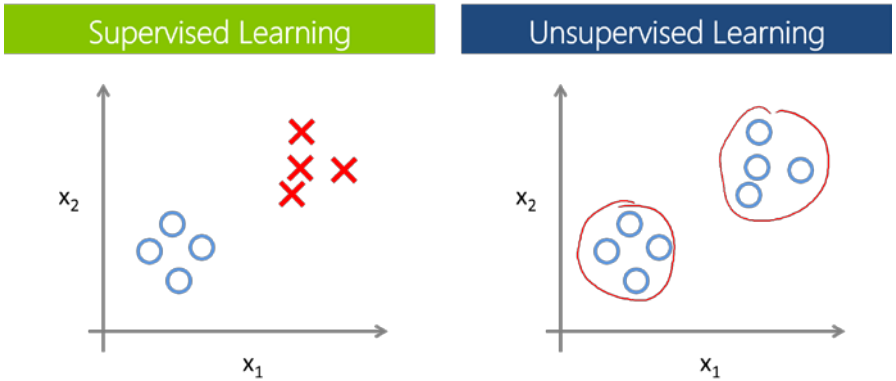
- Trial and Error
  - Trying out actions to determine their reward
- Delayed Reward
  - Actions affect the current state of the environment and future possible rewards

Reinforcement Learning

- RL is a broad class of problems
- Any method that can solve one of those problems is an RL method
- RL Problems:
  - Agent can sense the state of an environment
  - Agent can take actions in the environment
  - Agent must have a goal in the environment

RL vs. Supervised Learning

- Supervised Learning
  - Most popular machine learning problems
  - Learning from examples provided by supervisor



RL vs. Supervised Learning

- SL depends on labeled examples
- In interactive environments, it is difficult to obtain examples of desired behaviour
  - Which are ‘correctly labeled’
  - Which are representative of all situations
- In ‘uncharted territory’, an agent must be able to learn form its own experience
- In RL, we are only given possible actions, and the resulting reward from performing them

Exploration vs. Exploitation

- One of the main challenges in reinforcement learning is the trade-off between exploration and exploitation
- To obtain a lot of reward, agents must prefer actions that it knows produce good results
- In order to learn which actions produce good rewards, it must try them out first
- The agent must exploit knowledge it has, but also explore in order to gain more knowledge

Exploration vs. Exploitation Examples

- Choosing a Restaurant
  - Go to the place you know that’s alright, or try a new place you’ve never eaten at before?

- Playing Games
  - If I fold this hand of Poker I won’t lose much, but I won’t know what the opponent has unless I call and risk some money

Elements of RL

- Agent
- Environment
- Policy
- Reward Function
- Value Function
- Model of Environment (optional)

Elements of RL: Policy

- A map from perceived states of the environment to actions to be taken at those states
- Map be a simple look-up table, or a complex computation such as a search
- The core of an RL agent, defines behavior
- May be deterministic or stochastic

	ADVANCED BLACKJACK STRATEGY TABLE									
	Dealer's First Card									
Your Hand	2	3	4	5	6	7	8	9	10	A
18+	STAND	STAND	STAND	STAND	STAND	STAND	STAND	STAND	STAND	STAND
17	STAND	STAND	STAND	STAND	STAND	STAND	STAND	STAND	STAND	STAND
16	STAND	STAND	STAND	STAND	STAND	HIT	HIT	HIT	HIT	HIT
15	STAND	STAND	STAND	STAND	STAND	HIT	HIT	HIT	HIT	HIT
14	STAND	STAND	STAND	STAND	STAND	HIT	HIT	HIT	HIT	HIT
13	STAND	STAND	STAND	STAND	STAND	HIT	HIT	HIT	HIT	HIT
12	HIT	HIT	STAND	STAND	STAND	HIT	HIT	HIT	HIT	HIT
11	DOUBLE	DOUBLE	DOUBLE	DOUBLE	DOUBLE	DOUBLE	DOUBLE	DOUBLE	DOUBLE	HIT
10	DOUBLE	DOUBLE	DOUBLE	DOUBLE	DOUBLE	DOUBLE	DOUBLE	DOUBLE	HIT	HIT
9	HIT	DOUBLE	DOUBLE	DOUBLE	DOUBLE	HIT	HIT	HIT	HIT	HIT
8	HIT	HIT	HIT	HIT	HIT	HIT	HIT	HIT	HIT	HIT
7	HIT	HIT	HIT	HIT	HIT	HIT	HIT	HIT	HIT	HIT
6	HIT	HIT	HIT	HIT	HIT	HIT	HIT	HIT	HIT	HIT
5	HIT	HIT	HIT	HIT	HIT	HIT	HIT	HIT	HIT	HIT
Soft 20	STAND	STAND	STAND	STAND	STAND	STAND	STAND	STAND	STAND	STAND
Soft 19	STAND	STAND	STAND	STAND	STAND	STAND	STAND	STAND	STAND	STAND
Soft 18	STAND	DOUBLE	DOUBLE	DOUBLE	DOUBLE	STAND	STAND	HIT	HIT	HIT
Soft 17	HIT	DOUBLE	DOUBLE	DOUBLE	DOUBLE	HIT	HIT	HIT	HIT	HIT
Soft 16	HIT	HIT	DOUBLE	DOUBLE	DOUBLE	HIT	HIT	HIT	HIT	HIT
Soft 15	HIT	HIT	DOUBLE	DOUBLE	DOUBLE	HIT	HIT	HIT	HIT	HIT
Soft 14	HIT	HIT	HIT	DOUBLE	DOUBLE	HIT	HIT	HIT	HIT	HIT
Soft 13	HIT	HIT	HIT	DOUBLE	DOUBLE	HIT	HIT	HIT	HIT	HIT
Pair A	SPLIT	SPLIT	SPLIT	SPLIT	SPLIT	SPLIT	SPLIT	SPLIT	SPLIT	SPLIT
Pair 10	STAND	STAND	STAND	STAND	STAND	STAND	STAND	STAND	STAND	STAND
Pair 9	SPLIT	SPLIT	SPLIT	SPLIT	SPLIT	STAND	SPLIT	SPLIT	STAND	STAND
Pair 8	SPLIT	SPLIT	SPLIT	SPLIT	SPLIT	SPLIT	SPLIT	SPLIT	SPLIT	SPLIT
Pair 7	SPLIT	SPLIT	SPLIT	SPLIT	SPLIT	SPLIT	HIT	HIT	HIT	HIT
Pair 6	SPLIT	SPLIT	SPLIT	SPLIT	SPLIT	HIT	HIT	HIT	HIT	HIT
Pair 5	DOUBLE	DOUBLE	DOUBLE	DOUBLE	DOUBLE	DOUBLE	DOUBLE	DOUBLE	HIT	HIT
Pair 4	HIT	HIT	HIT	SPLIT	SPLIT	HIT	HIT	HIT	HIT	HIT
Pair 3	SPLIT	SPLIT	SPLIT	SPLIT	SPLIT	SPLIT	HIT	HIT	HIT	HIT
Pair 2	SPLIT	SPLIT	SPLIT	SPLIT	SPLIT	SPLIT	HIT	HIT	HIT	HIT

Reward Function

- Defines the goal in a RL problem
- Maps each perceived state (or state-action pair) to a single number (reward)
- Reward is the desirability of a state
- RL agent’s goal is to maximize reward
- Biology: Reward = Pleasure / Pain

Reward Function Examples

- Path-Finding
  - Goal has positive reward
  - Possible negative reward on hazards
  - Non-goal states have no reward
- Blackjack
  - Winning State = Positive Reward
  - Losing State = Negative Reward

Value Function

- Rewards: what are immediately good
- Value function indicates what may be eventually be good
- Value of a state is the total amount of reward an agent can expect to accumulate in the future starting from it
- Values indicate long-term desirability

Value Function

- Reward: Pleasure or Pain
- Value: How pleased or displeased we are to be here
- We seek actions that have the highest value, those will bring the highest eventual reward
- Value is harder to determine than reward, so we must calculate and estimate them
- Our policy should lead us to high value states

Value Function Examples

- Path-Finding
  - Non-goal state on the best path to the goal has high desirability, and a high value
- Blackjack
  - Me having 20 with dealer showing 6 has high probability of me winning, and high value

Model of Environment

- Mimics the behaviour of environment
- Ex: Given state and action, the model may attempt to predict the next state
- Models are used for planning (A*, AB)
- Not necessary for RL in general
- Won’t use models in this course