Introduction to Machine Learning (CS419M)
Lecture 24: Reinforcement learning & Course wrap-up

Apr 18, 2018
Slides on RL borrowed from David Silver’s lectures.
Branches of ML

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Lecture 1: Introduction to Reinforcement Learning

About RL
What is Reinforcement Learning?

- Learning by interacting with an environment to achieve a goal
- Learning by trial-and-error with only some form of evaluative feedback (aka "reward")
- Agent’s actions affect the subsequent data it receives
RL Demo
Rewards

- A reward $R_t$ is a scalar feedback signal
- Indicates how well agent is doing at step $t$
- The agent’s job is to maximise cumulative reward

Reinforcement learning is based on the reward hypothesis

Definition (Reward Hypothesis)

All goals can be described by the maximisation of expected cumulative reward
Rewards

Lecture 1: Introduction to Reinforcement Learning

The RL Problem

Reward

A reward $R_t$ is a scalar feedback signal that indicates how well the agent is doing at step $t$. The agent's job is to maximize cumulative reward.

Reinforcement learning is based on the reward hypothesis.

Definition (Reward Hypothesis)

All goals can be described by the maximization of expected cumulative reward.

Do you agree with this statement?
Agent and Environment

At each step $t$ the agent:
- Executes action $A_t$
- Receives observation $O_t$
- Receives scalar reward $R_t$

The environment:
- Receives action $A_t$
- Emits observation $O_{t+1}$
- Emits scalar reward $R_{t+1}$

$t$ increments at env. step
Major components of an RL Agent

- An RL agent may include one or more of these components:
  - Policy: agent’s behaviour function
  - Value function: how good is each state and/or action
  - Model: agent’s representation of the environment
Policy

- A policy is the agent’s behaviour
- It is a map from state to action, e.g.
- Deterministic policy: \( a = \pi(s) \)
- Stochastic policy: \( \pi(a|s) = \mathbb{P}[A_t = a|S_t = s] \)
Value Functions

- Value function is a prediction of future reward
- Used to evaluate the goodness/badness of states
- And therefore to select between actions, e.g.

$$v_\pi(s) = \mathbb{E}_\pi \left[ R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots \mid S_t = s \right]$$
Value Functions

- A value function is a prediction of future reward
  - “How much reward will I get from action $a$ in state $s$?”
- $Q$-value function gives expected total reward
  - from state $s$ and action $a$
  - under policy $\pi$
  - with discount factor $\gamma$

$$Q^\pi(s, a) = \mathbb{E}[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots | s, a]$$

- Value functions decompose into a Bellman equation

$$Q^\pi(s, a) = \mathbb{E}_{s', a'} [r + \gamma Q^\pi(s', a') | s, a]$$
Optimal Value Function

- An optimal value function is the maximum achievable value
  \[ Q^*(s, a) = \max_{\pi} Q^\pi(s, a) = Q^{\pi^*}(s, a) \]
- Once we have \( Q^* \) we can act optimally,
  \[ \pi^*(s) = \arg\max_a Q^*(s, a) \]
Exploration and Exploitation

- Reinforcement learning is like trial-and-error learning
- The agent should discover a good policy
- From its experiences of the environment
- Without losing too much reward along the way

- *Exploration* finds more information about the environment
- *Exploitation* exploits known information to maximise reward
- It is usually important to explore as well as exploit
Examples

- Restaurant Selection
  - **Exploitation**: Go to your favourite restaurant
  - **Exploration**: Try a new restaurant

- Online Banner Advertisements
  - **Exploitation**: Show the most successful advert
  - **Exploration**: Show a different advert

- Oil Drilling
  - **Exploitation**: Drill at the best known location
  - **Exploration**: Drill at a new location

- Game Playing
  - **Exploitation**: Play the move you believe is best
  - **Exploration**: Play an experimental move
Model

- A model predicts what the environment will do next
- $\mathcal{P}$ predicts the next state
- $\mathcal{R}$ predicts the next (immediate) reward, e.g.

$$
\mathcal{P}^{a}_{ss'} = \mathbb{P}[S_{t+1} = s' \mid S_t = s, A_t = a]
$$

$$
\mathcal{R}^a_s = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a]
$$
Atari: Reinforcement Learning

- Rules of the game are unknown
- Learn directly from interactive game-play
- Pick actions on joystick, see pixels and scores
Atari: Planning

- Rules of the game are known
- Can query emulator
  - perfect model inside agent’s brain
- If I take action \( a \) from state \( s \):
  - what would the next state be?
  - what would the score be?
- Plan ahead to find optimal policy
  - e.g. tree search
AlphaGo: Deep RL’s Poster child!
AlphaGo Zero

40 days – AlphaGo Zero surpasses all previous versions, becomes the best Go player in the world

36 hours – AlphaGo Zero reaches level of Alpha Go Lee, which beat world champion Lee Sedol in 2016

72 hours – AlphaGo Zero beats Alpha Go Lee, 100:0

From: “Mastering the Game of Go without Human Knowledge”
Excellent freely available textbook on RL!

Reinforcement Learning: An Introduction

Second edition, in progress

****Complete Draft****

November 5, 2017

Richard S. Sutton and Andrew G. Barto
© 2014, 2015, 2016, 2017

Remaining Logistics
Final Exam

Logistics

- Date: April 28th, 2018
- Time: 2 pm to 5 pm
- Venue: CC 103 & 105

Exam details

- Open notes (“open notes” not “open book”)
- Similar to the midsem format. Mostly testing concepts.
Final Exam Syllabus in 1-slide

- Basics of probability for machine learning
- Linear regression/Logistic regression
- Bias/Variance of classifiers
- Basics of statistical learning theory
- Model selection + Generalization errors + Regularization
- Perceptron + Naive Bayes + SVMs + kernel methods
- MLE/MAP estimates
- k-means Clustering + EM Algorithm
- Dimensionality Reduction
- Feedforward NNs + Backpropagation
- Ensemble learning (Boosting/Bagging)
- Basics of HMMs
Final Project

Deliverables

• 4-5 page final report:
  ✓ Task definition, Methodology, Prior work, Implementation Details, Experimental Setup, Experiments and Discussion, Error Analysis (if any), Summary

• Short talk summarizing the project:
  ✓ Each team will get 10 mins for their presentation and 5 minutes for Q/A
  ✓ Clearly demarcate which team member worked on what part
Final Project Schedule

- Presentations will be held on April 30th and May 1st
- The final report in pdf format should be sent to pjyothi@cse.iitb.ac.in before April 30th
- The order of presentations will be decided on a lottery basis and shared via Moodle
Final Project Grading

• Break-up of 15 points:
  • 3 points for the preliminary report
  • 3 points for the presentation
  • 3 points for Q/A
  • 4 points for the final report
  • 2 points for overall merit (awarded to exceptional projects)