Reinforcement Learning Standford Notes

Thamu Mnyulwa

March 2020

1 Introduction

This is a list of notes made on a reinforcement learning course offered at Stanford university I found on the stanford.edu website [1]. According to the instructor it is a entry level graduate student or Phd course on reinforcement learning. I copy the slides followed by my own notes.

Figure 1: The Universe

2 Lecture 1

Plan

- Overview of Reinforcement Learning
- Introduction to sequential decision making under uncertainty

Repeated Interactions with the world

- Learn to make **good sequences of decisions**
- *How do we define goodness?*

What we mean by **good** here is some notion of optimality we have some utility measure over the decisions are being made. Fundamental challenge in artificial intelligence and machine learning is learning to make good decisions under uncertainty.

Example from Yael Niv

- Childhood: primitive brain & eye,swims around, attaches to a rock
- Adulthood: digests brain, sits
- Suggests brain is helping guide decisions(no more decisions, no need for brain?)

A species evolves over time, when it is a child it's brain is primitive. It sits to a rock and sits there, when its a adult it digests its brain, and sits there. Perhaps this is an indication that the point of intelligence or at least the point of having a brain in part is helping to guide decisions. So once all the decisions in an agents life are completed maybe we no longer need a brain.

Atari

- Video games are a complex task that take human players often a while to learn where we don't know the results in advance, hence a great medium for RL
- David Silver 2015 Atari : A paradigm shift in RL as the game learned directly from pixels and eventually played better than people.

Robotics

- Sergey Levine and Pieter Abbel's lab, UC Berkeley

2.1 Reinforcement learning Involves

Reinforcement learning Involves

- Optimization
- Delayed consequences
- Exploration
- Generalization

1. Optimization

- Goal is to find an optimal way to make decisions
- Either Yielding best outcomes
- Or at least a very good strategy

2. Delayed consequences

- Decisions now can impact things much later ... (e.g. saving for retirement)
- Introduces two challenges
    1. When planning: Decisions involve reasoning about not just immediate benefit of a decision but also its longer term ramifications
    2. When learning: temporal credit assignment is hard (what caused high or low rewards?)

One of the challenges to doing this is, because you do not necessarily receive immediate feedback, how do you determine the credit assignment problem. This is, "how do you determine the causal relationship between the decisions you made in the past and the outcome in the immediate future"? This is a big problem in reinforcement learning.

3. Exploration

- Learning about the world making decisions
    - Agent as scientist
    - Learn to ride a bike by trying (and failing)
- Censored data
    - Only get a reward (label) for decision made
    - E.g You don't know what would have happened if we had taken red pill instead of blue pill (matrix movie reference).
- Decisions impact what we learn about
    - E.g Depending on what University we decide to go to, we will have different later experiences ...

We try to get the agent to learn from experiences. One of the big problems is that data is censored.What we mean is that you only get to learn from what you try to do. By making a decision, you have the opportunity cost of not getting to take on another decision.

Policy

- **Policy** is mapping from past experience to action
- Why not just pre-program a policy?

4. **Generalization**

   - **Policy** is mapping from past experience to action
   - Why not just pre-program a policy?
   - Deep-Mind Nature 2015 game

AI Planning (RL)

- **Optimization**
- **Generalization**
- Exploration
- **Delayed consequences**

1. Computes good sequences of decisions
2. **But given model of how decisions impact the world**

Why are RL problems different to other types of ML? A topic that comes about in a lot of AI problems is planning. Planning involves Optimization, Generalization and Delayed consequences. You might take a move and go early but it might not be immediate that that was a good move until someone has gone minutes later. It does not involve exploration (main point). The idea in planning is that you are given a model of how the world works, you are given the rules of the game for example, you are given what the reward is and the important part is what you should do given a model of the world. It does not require exploration.

Supervised Machine Learning (vs RL)

- **Optimization**
- **Generalization**
- Exploration
- Delayed consequences

1. Learns from experience
2. **But provided correct labels**

Supervised ML often involves optimization and generalization . However, frequently it does not involve either exploration or delayed consequences. It does not tend to include exploration because typically in supervised learning you are given a data set (i.e your agent is not collecting its experience or data about the world, it is given the experience with which it can use). Similarly it is typically given one decision (*cross sectional data*) instead of say making a decision with which one would need to make say now rather than later on make another decision as in RL.

Unsupevised Machine Learning (vs RL)

- **Optimization**
- **Generalization**
- Exploration
- Delayed consequences

1. Learns from experience
2. **But no labels from the world**

Unsupervised machine learning also involves optimization and generalization , but generally does not involve exploration or delayed consequences. Typically, you have no labels about the world. In RL you typically get in between that (semi-supervised) which is a utility of the label you want. You don't get a true label of the world in RL.

Imitation Learning (vs RL)

- **Optimization**
- **Generalization**
- Exploration
- **Delayed consequences**

1. Learns from experience . . . **of others**
2. **Assumes input demos of good policies**

Imitation learning is similar to RL, but different. It includes Optimisation, generalisation and delayed consequences. However, here the idea is that we are going to be learning from the experiences of others. Instead of our intelligent agent going to go into the world and take experiences to make its own decisions, it may watch another intelligent agent (which may be a person), observe the outcomes and then use that experience to figure out how it wants to act. There are a lot of benefits to doing this but it is slightly different to RL because we don't have to directly figure out the exploration problem. Immitation learning is becoming increasingly important and was introduced by "Abbel, Coates and Ng helicopter team, Standford". In a paper to see how quickly you could imitate experts flying toy helicopters [2].

-
-
-

3 Conclusion

"I always thought something was fundamentally wrong with the universe"

Bibliography

[1] Emma Brunskill. Standford cs234: Reinforcement learning lecture 1.

[2] Jessica Taylor, Eliezer Yudkowsky, Patrick LaVictoire, and Andrew Critch. Alignment for advanced machine learning systems. *Machine Intelligence Research Institute*, 2016.