CSE 513T Theory of Artificial Intelligence & Machine Learning April 10th, 2018

Lecture 23

Lecturer: Brendan Juba *Scribe: Yiqi Lin*

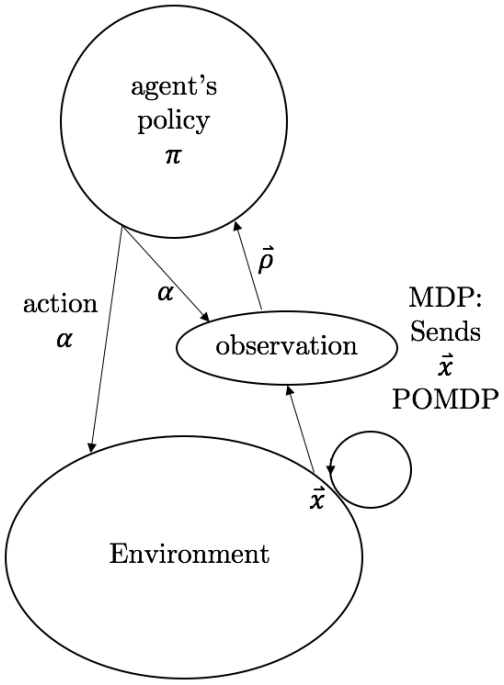
CSE 513T 4/10

1. Pitfalls in learning unknown environment
2. Formulation of learning from example histories

1 Note

1. Alex Durgin will give a guest lecture on Thursday
2. Final project presentation sign-up sheet is circulating

2 Recap



	$\bar{x} =$	$\partial t_{r,0}$	$\partial t_{r,1}$	$\partial t_{r,2}$	$\partial t_{b,0}$	$\partial t_{b,1}$	$\partial t_{b,2}$	$holds_{r,b}$
$t=5$	$\vec{\rho_5} =$	0	0	1	*	*	1	0
$t=6$	$\vec{\rho_6} =$	1	0	0	0	*	*	0

Note:

Goal $\partial t_{b,2} \wedge \neg \partial t_{r,2}$

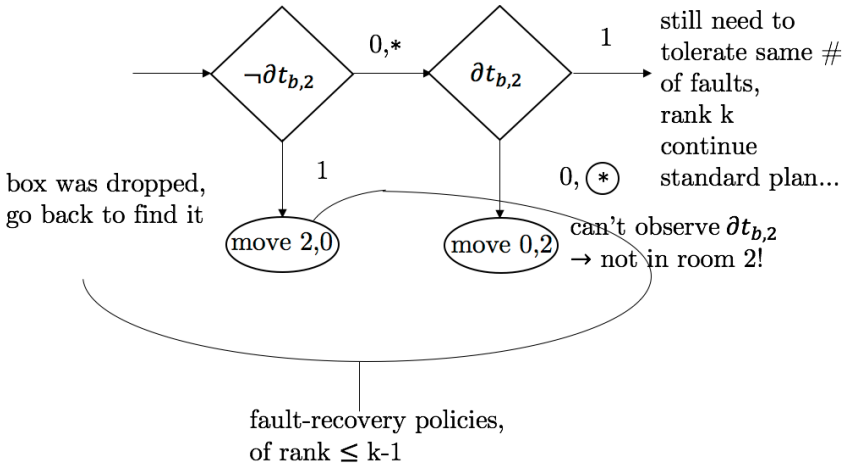
can't simultaneously observe

both satisfied

Determining if $G(\vec{x_t}) = 1$

Problem of reasoning in PAC-semantics: want to determine $G(\vec{x_t})$ is $1 - \varepsilon$ valid.

Policies: bounded-rank decision trees (captured fault—tolerant plans, and others...)



3 Today's Topic:

Success ≠ ability to predict

Instead ability to produce plans to achieve goals in the environment

4 Motivation:

1. Model of development of intelligent being, e.g. a baby
2. Knowledge engineering is difficult — even the toy robot & box example was challenging. We tend to overlook the "obvious", as when we discussed "common sense". Need agents we build to share this "common sense" — learning can be effective way to acquire this knowledge (or, some of it...)

5 Pitfalls in learning environment

Many "obvious" formulations of this problem are intractable.

Ex:

"given access to a POMDD encoding a noisy STRIPS domain, find a complete description of how the environment changes in response to actions"

Difficulty #0: STRIPS planning is NP-hard

Difficulty #1: May be hard to detect conditional effects...

Eg:

- "Combination lock domain" — there is a secret combination $\bar{s} \in \{0,1\}^n$
- Step-by-step the agent enters $\bar{x} \in \{0,1\}^n$
- Lock opens (conditional effect) if $\bar{x} = \bar{s}$
- Otherwise, nothing happens
- A simple O(n) rule, O(n)- attribute, 2-action STRIPS domain
- If goal is "open the lock" —this is hard!
- In fact, without hitting the secret on some trial, we don't see that it's possible to hit the lock.

Suppose s chosen uniformly at random, if the agent does not enter $\bar{x} = \bar{s}$, the environment behaves identically. After m trials, if $\bar{s}$ was not one of the m combinations entered, our algorithm choice of what $\bar{x}$ to try next is identical.

So, we can simply compute: for any fixed "trajectory" (sequence of m combinations) what is the probability one of them is $\bar{s}$?

$$\Pr_{\bar{s}}[\bar{x} = \bar{s}] = \frac{1}{2n} \text{ for any fixed } \bar{x}$$

$$\Pr_{\bar{s}}\left[\bar{x}^{(1)} = \bar{s} \vee \bar{x}^{(2)} = \bar{s} \vee \cdots \vee \bar{x}^{(n)} = \bar{s}\right] = \frac{m}{2n}$$

Even if we go for $m = 2^{n/2}$ rounds, prob we hit combinations is $\frac{1}{2^{\frac{n}{2}}}$ — so, with probability $1 - \frac{1}{2^{\frac{n}{2}}}$, we may spend $2^{\frac{n}{2}}$ rounds and still not see the lock open.

Fix:

We'll assume that we're given a training set of example histories of interaction, we will only promise to do wee w.r.t. whatever portion of environment we were likely to see during training.

Difficulty #2: "Explicit representations" In general, noisy STRIPS poses an agnostic learning problem — we must find rules that approximately capture the environment, that may be violated sometimes. Moreover, the updated rules for attributes may be given by a DNF if multiple rules have that effect, may be hard to learn even without noise!

Fix:

Avoid producing explicit representations of the rules, use implicit learning.

6 Formulation of Learning from Example Histories.

Capturing fix #1:

Assume there was an "exploration policy" π_0 used to collect the training set

Ex:

π_0 might just choose uniform random actions

Ex:

if we collected a training set using arbitrary policies $\pi^{(1)}, \pi^{(2)}, \cdots \pi^{(m)}$, π_0 might choose one $\pi^{(1)}$ uniformly at random and follow it.

Basically, gives a $\pi_0$that generates the training set we got

Training set $x^{(1)}, x^{(2)}, \cdots x^{(m)}$, where $\bar{x}^{(j)} = \bar{x}_0^{(j)}, \bar{\rho}_0^{(j)}, \alpha_1^{(j)}, \bar{x}_1^{(j)}, \bar{\rho}_1^{(j)}, \alpha_2^{(j)}, \cdots, \alpha_T^{(j)}, \bar{x}_T^{(j)}, \bar{\rho}_T^{(j)}$

Def^n:

We say the sequence of actions $\alpha_1, \cdots, \alpha_t$ $(t \leq T)$ is μ-typical $(\mu \in [0,1])$ for the reference policy π_0 and POMDP if the probability of π_0 takes precisely the sequence of actions $\alpha_1, \cdots, \alpha_t$ in the POMDP is at least μ.

Ex:

For the "random action" policy, we only find that the initial, short sequences are μ-typical for non-negligible μ. (the probability of hitting $\alpha_1, \cdots, \alpha_t$ decays exponentially in t for this policy...)

Note:

Any sequence that we see repeated at least tlog|A| times in our training set must be $\Omega(1/m)$ typical.

(then in expectation $O(1/m)$-typical is observed $<<1$ time in the training set, independent trials,

Chernoff bound gives $\Pr[\text{observe} \geq \sqrt{tlog|A|}\ \text{times}] < e^{-tlog|A|} = \frac{1}{|A|^t} \cdot \delta$

Union bound over $\leq |A|^t$ sequences $\rightarrow$ doesn't happen)

Note:

to learn the effect of a sequence of actions, must observe them some # of times...

Conversely, to observe the μ-typical sequences, only need to take $m = \tilde{O}(\frac{1}{\mu})$

As on HW9 #1, to make use of training set to estimate the environment's behavior, we'll need to also assume that we know the probability / likelihood with which π_0 chose each of its actions in the training set.

What we'll get into next lecture: how to efficiently construct bounded-rank decision tree policies in, e.g., noisy STRIPS in spite of the fact that planning in the domain is NP-hard.

Catch:

we only promise to find policies that only take typical sequences of actions (with respect to π_0) on every branch.

Reduces the number of possible sequences of actions we can consider to estimate size of training set.

Alt:

there are at most $1/\ \mu$ μ-typical sequences for any π_0 and POMDP

Since (for fixed t) distinct μ-typical sequences are disjoint events, each comprising at least μ prob. So, $>\mu$ events of $\geq \mu$ probability $\rightarrow >1$ probability total. *

7 Sketch of task

Given:

access to histories of interaction between exploration policy π_0 and a POMDP goal DNF G

Parameters μ (typicality), ε (error tolerance in rules), γ (accuracy), δ (confidence)

If:

there is a rank-k decision tree for time horizon T ($1-\varepsilon+\gamma-$testable) and a testable (implicit KB) formula ψ s.t. on each branch of the policy

1. The sequence of action is μ-typical
2. There is a (space-bounded, treelike) resolution proof that G is satisfied given ψ and the literals tested on the branch

Then we find such a policy achieving G with probability $1-\alpha(n,T)(\varepsilon+\gamma)$

(2) is satisfied, e.g., if ψ encodes the STRIPS rules describing the environment.