CSE 513T Theory of Artificial Intelligence & Machine Learning November 03, 2016

Lecture 19

Lecturer: Brendan Juba *Scribe: Gustavo Malkomes*

1 Main topics

We will discuss how to do **reasoning with partial examples** (implicit learning), Section 2; and introduce the mechanism of **Negation-as-failure** on Section 3.

2 Incomplete information

Partial information model

Given a distribution $\mathcal{D}$ on $\mathcal{X}$ over "ground truth" examples distribution, and a distribution $\mathcal{M}$ on masking functions. We actually obtain partial examples $\rho^{(i)} = m^{(i)}(x^{(i)})$ where $m^{(i)} \in \mathcal{M}$, $x^{(i)} \in \mathcal{D}$ independently.

Definition 1 *We say that a representation φ is $(1-\epsilon)$-**testable** w.r.t. $\mathcal{M}(\mathcal{D})$ if for $\rho \in \mathcal{M}(\mathcal{D})$, we have $Pr\big[\varphi\,|_\rho = 1\big] \geq 1-\epsilon$*

Definition 2 *We say that a proof system $\mathcal{P}$ is **closed under restrictions**, if for any proof $\pi \in \mathcal{P}$ (such as π is a proof of φ from $\psi_1, \ldots, \psi_k$), and partial example ρ, there's is a proof $\pi\,|_\rho$ of $\varphi\,|_\rho$ from $\psi_1\,|_\rho, \psi_2\,|_\rho, \ldots, \psi_k\,|_\rho$.*

Reasoning with partial examples

Let's introduce an algorithm for reasoning with partial examples. Given some proof system of interest $\mathcal{P}$ (e.g., space-bounded resolution, chaining, ...), consider Algorithm 1, which decides whether or not the given set of collection of proofs has a proof of a query φ from premises $\psi_1, \cdots, \psi_k$. Our goal is to provide formal guarantees about the referred algorithm — Theorem 3.

A note on notation: we will write $\mathcal{P}(\varphi; \psi_1, \ldots, \psi_k)$ to indicate if $\mathcal{P}$, the set of proofs, contains a proof of φ from $\psi_1, \ldots, \psi_k$.

Algorithm 1 Algorithm for reasoning with partial examples

```
input query φ; KB ψ_1, ..., ψ_k; and partial examples ρ^(1), ..., ρ^(m)
S = 0
for i = 1, 2, ..., m
    if P(φ|_ρ^(i); ψ_1|_ρ^(i), ψ_2|_ρ^(i), ..., ψ_k|_ρ^(i))
        increment S
return p̂ := S/m
```

Note that as long as the proof system $\mathcal{P}(.;.)$ is efficient, so is this algorithm.

Theorem 3 *For $m \geq \frac{1}{2\gamma^2} \ln \frac{2}{\delta}$, if $\rho^{(1)}, ..., \rho^{(m)}$ are independently drawn from $\mathcal{M}(\mathcal{D})$, then with probability at least $(1-\delta)$, the value $\widehat{p}$ returned by the Algorithm 1 satisfies:*

1. *$\big[\psi_1, \wedge, \cdots \wedge \psi_k\big] \implies \varphi$ is $(\widehat{p}-\gamma)$-valid with respect to $\mathcal{D}$.*
2. *If*
   (a) *there are representations $\psi_{k+1}, ..., \psi_s$ such that $\psi_{k+1}\wedge, \ldots, \wedge\psi_s$ is $(p+\gamma)$-testable with respect to $\mathcal{M}(\mathcal{D})$.*

(b) the proof system $\mathcal{P}$ is closed under restrictions; and

(c) there is a proof φ from $\psi_1, \ldots, \psi_s$, i.e., $\mathcal{P}(\varphi; \psi_1, ..., \psi_k, \psi_{k+1}, ..., \psi_s) = 1$

then $\widehat{p} \geq p$.

That is, the algorithm is "implicitly asserting" a KB that contains every possible testable $\psi_{k+1}, ..., \psi_s$. So, in this sense, the algorithm has implicitly learned $\psi_{k+1}, ..., \psi_s$ and uses them to answer the query φ.

Proof Let's prove this theorem in two parts: more specifically, breaking the proof into items (1) and (2) of the Theorem 3. To stress the new important ideas we will start proving item (2).

Part 2: The key observation is that the $(p+\gamma)$-testable premises $\psi_{k+1}, ..., \psi_s$ are all simultaneously simplify to '1' with probability $(p+\gamma)$ by hypothesis (a) and (b).

Let's, define I_j as:

$$I_j := \begin{cases} 1 \text{ if } (\psi_{k+1}\wedge, \ldots, \wedge\psi_s) \mid_{\rho^{(j)}} = 1 \\ 0 \text{ otherwise.} \end{cases}$$

And, for $j = 1, 2, \ldots, m$, we have:

$$\begin{aligned} \mathbb{E}\big[I_j\big] &= \Pr\left[(\psi_{k+1}\wedge, ..., \wedge\psi_s) \mid_{\rho^{(j)}} = 1\right] \\ &\geq p + \gamma. \end{aligned}$$

since the representations $\psi_{k+1}, ..., \psi_s$ are $(p+\gamma)$-testable from (a) and from (b) $\mathcal{P}$ is closed under restrictions. Now, we want to bound $\Pr[\widehat{p} < p]$. Notice that

$$S = \sum_{j=1}^{m} I_j, \quad \text{and} \quad \widehat{p} = \frac{S}{m} = \frac{1}{m}\sum_{j=1}^{m} I_j$$

Since expectation is linear,

$$\mathbb{E}[S] = \sum_{j=1}^{m} \mathbb{E}[I_j] \geq m(p+\gamma)$$

By additive Chernoff bound, and since $\widehat{p} = \frac{S}{m}$ and $p = \mathbb{E}[\widehat{p}] - \gamma = \frac{m(p+\gamma)}{m} - \gamma$,

$$\Pr\big[\widehat{p} < p\big] = \Pr\left[\frac{S}{m} < \mathbb{E}\left[\frac{S}{m}\right] - \gamma\right] \leq \exp\big(-2m\gamma^2\big) \leq \frac{\delta}{2} \quad \text{for} \quad m \geq \frac{1}{2\gamma^2}\ln\frac{2}{\delta}$$

So $\widehat{p} \geq p$ with probability at least $(1 - \frac{\delta}{2})$.

When $\psi_{k+1}, ..., \psi_s$ are simultaneous simplify to $'1'$, we have:

1. a proof of φ from $\psi_1, ..., \psi_k$, $\psi_{k+1}, ..., \psi_s$ (from (c));
2. a proof of $\varphi \mid_\rho$ from $\psi_1 \mid_\rho, ..., \psi_k \mid_\rho$, $\underbrace{\psi_{k+1} \mid_\rho, ..., \psi_s \mid_\rho}_{=1}$ since $\mathcal{P}$ is a set of restriction-closed proofs (from (b)) and the formulas $\psi_{k+1} \mid_{\rho^{(j)}}, ..., \psi_s \mid_{\rho^{(j)}}$ have all dropped out (simplified to '1').

So we actually have a proof of $\varphi \mid_{\rho^{(j)}}$ from $\psi_1 \mid_{\rho^{(j)}}, ..., \psi_k \mid_{\rho^{(j)}}$ alone, and hence we indeed actually increment S wherever $I_j = 1$. So, we return $\widehat{p} \geq \frac{S}{m}$ where $\frac{S}{m} \geq p$ w.p. $1 - \frac{S}{2}$, so $\widehat{p} \geq p$ with probability $1 - \delta/2$, which completes part 2.

Part 1: We note that when there is a proof of $\varphi \mid_{\rho^{(j)}}$ from $\psi_1 \mid_{\rho^{(j)}}, ..., \psi_k \mid_{\rho^{(j)}}$ then by the soundness of the proof system, for any setting y of the free variables that satisfy $\psi_1 \mid_{\rho^{(j)}}, ..., \psi_k \mid_{\rho^{(j)}}$, we also have $\Big(\varphi \mid \rho^{(j)}\Big)(y) = 1$.

This is true in particular for the $y = x^{(j)}$ (for the ground truth example used to produce $\rho^{(j)}$). Therefore $\Big([\psi_1 \wedge ... \wedge \psi_k] \implies \varphi\Big)(x^{(j)})$ is satisfied in this case.

Let's define:

$$p^* := \Pr_{\mathcal{D}}\Big[\big([\psi_1 \wedge \psi_2 \ldots \wedge \psi_k] \implies \varphi\big) = 1\Big].$$

$$I'_j := \begin{cases} 1 \text{ if } \big([\psi_1 \wedge \psi_2 \ldots \wedge \psi_k] \implies \varphi\big)(x^{(j)}) = 1 \\ 0 \text{ otherwise.} \end{cases}$$

$$S' := \sum_{j=1}^{m} \mathbb{E}\big[I'_j\big]$$

As before, $\mathbb{E}[S'] = mp^*$ so that $p^* = \mathbb{E}\Big[\frac{S'}{m}\Big]$, and with another Chernoff bound we have:

$$\Pr\left[\widehat{p} - \gamma > p^*\right] = \Pr\left[\frac{S}{m} - \gamma > \mathbb{E}\left[\frac{S'}{m}\right]\right] \leq \Pr\left[\frac{S'}{m} > \mathbb{E}\left[\frac{S'}{m}\right] + \gamma\right] \leq \frac{\delta}{2}.$$

Hence, with probability at least $(1 - \frac{\delta}{2})$, $\widehat{p} - \gamma \leq p^* = \Pr\left[([\psi_1 \wedge \psi_2 \wedge \ldots \wedge \psi_k] \implies \varphi) = 1\right]$. By the union bound, we complete the proof. ■

Remark

Important part as with reasoning with complete examples, this allows us to tolerate some counterexamples, say, if we accept φ as following from KB when $\widehat{p} \geq 75\%$. $\widehat{p}$ is just a conservative estimate of $\Pr\left[KB \implies \varphi\right]$. In particular, if $\Pr\left[KB\right] = 1$, then $\widehat{p}$ is an estimate of $Pr[\varphi]$. Also, if we only know $\Pr\left[KB\right] \geq 1 - \epsilon$, then (by a union bound) we find $\Pr\left[\varphi\right] \geq \widehat{p} - \epsilon - \gamma$.

3 Negation as failure (NAF)

Back to common sense. There are common sense rules for reasoning in particular absent things we don't observe directly. (e.g. "law of inertia", mental states and intentions of others, ...). Learning from examples does not seem to give a mechanism for acquiring or representing this kind of knowledge. We need to be able to make guesses that could turn out to be wrong in face of new evidence. A prominent mechanism for representing and reasoning with such rules is known as "negation as failure" (or NAF) .

Essentially NAF allows us to state rules that hold given a lack of some other knowledge, i.e., *"unless I have reason to believe (otherwise)."* Since acquiring such knowledge blocks these rules this is non-monotonic reasoning.

We will start with chaining, and we add a new "version" of each literal ℓ, "NOT ℓ", that may only appear in the body of a rule. So a rule,

$$[\ell_1 \wedge ... \wedge \ell_r \wedge \text{NOT } \ell_{r+1} \wedge ... \wedge \text{NOT } \ell_{r+s}] \implies \ell$$

Should mean, if $\ell_1, ...\ell_r$ have been proven and $\ell_{r+1}, ..., \ell_{r+s}$ can't be proven; then we infer ℓ.

A naive approach to NAF would simply try to prove $\ell_{r+1}, \cdots, \ell_{r+s}$ to decide whether or not the rule can be applied. But consider (first) the KB:

- $[\text{NOT } x_1] \implies x_2$
- $[\text{NOT } x_2] \implies x_1$

Can we prove x_1 ?

If so, then we can't prove x_2, and sure x_1 is provable. If not, then we can prove x_2, and sure, x_1 is not provable. We have no reason to prefer one over the other, since they are renamings of one another. Suppose we add

- $[\text{NOT } x_3] \implies x_3$

to the KB.

There is no consistent judgment of whether or not x_3 is possible.

Now suppose we add the rule

- $[x_1] \implies x_3$

There is, now, a unique consistent choice of which facts are provable and which aren't. That's it x_1 and x_3 are provable and x_2 is not provable. (Note: changing any one of these leads to a contradiction)

The worst part is that it's NP-complete to decide whether or not a consistent setting exists.

Elegant way out. Only commit to some things being provable/unprovable and say "I don't know"about the rest — we find a partial judgment about what is provable. We need to some guarantees that some things will be provable — starting point, and anything provable using chaining should be provable.

Next time, what should we guarantee is marked as "unprovable"?