CSE 513T Theory of Artificial Intelligence & Machine Learning March 22, 2018

Lecture 18

Lecturer: Brendan Juba *Scribe: John Dunagan*

1 Analysis of Reasoning With Partial Examples

Let us notate

$$\mathbb{P}(\varphi, \psi_1, \ldots, \psi_k) = \begin{cases} 1 & \text{if there is a proof of } \varphi \text{ from } \psi_1, \ldots, \psi_k \in \mathbb{P} \\ 0 & \text{otherwise} \end{cases}$$

Consider the following algorithm for reasoning with partial examples

ReasonWithPartialExamples$(\varphi, \psi_1, \ldots, \psi_k, \rho^{(1)}, \ldots, \rho^{(m)})$
  initialize $s \leftarrow 0$
  for i from 1 to m **do**
    $s \leftarrow s + \mathbb{P}(\varphi|_{\rho^{(i)}}, \psi_1|_{\rho^{(i)}}, \ldots, \psi_k|_{\rho^{(i)}})$
  return $\hat{p} = \frac{s}{m}$

Theorem 1 *For $m = \frac{1}{2\gamma^2} \ln \frac{2}{\delta}$, if $\rho^{(1)}, \ldots, \rho^{(m)}$ are independently drawn from $M(\mathcal{D})$, then with probability $1 - \delta$, $\hat{p}$ satisfies*

1. *$[\psi_1 \wedge \cdots \wedge \psi_k] \Rightarrow \varphi$ is $(\hat{p} - \gamma)$-valid*
2. *If*
   (a) *there is a set of $\psi_{k+1}, \ldots, \psi_s$ such that $\psi_{k+1} \wedge \cdots \wedge \psi_s$ is $(p+\gamma)$-testable with respect to $M(\mathcal{D})$*
   (b) *$\mathbb{P}$ is closed under restrictions*
   (c) *there is a proof of φ from $\psi_1, \ldots, \psi_k, \psi_{k+1}, \ldots, \psi_s$ in $\mathbb{P}$*

   then $\hat{p} \geq p$

Proof We will begin by addressing the second part of the theorem, and then address the first

Key Observation: if $\psi_{k+1} \wedge \cdots \wedge \psi_s$ is $(p+\gamma)$-testable, this means that for $\rho \in M(\mathcal{D})$, $\psi_{k+1}|_\rho, \ldots, \psi_s|_\rho$ all simultaneously simplify to 1 with probability $p + \gamma$

For $j \in [1, \ldots, m]$, define the indicator

$$I_j = \begin{cases} 1 & \text{if } \psi_{k+1} \wedge \cdots \wedge \psi_s|_{\rho^{(j)}} \text{ simplifies to 1} \\ 0 & \text{otherwise} \end{cases}$$

Then we have that

$$\mathbb{E}\Big[\frac{1}{m} \sum_{j=1}^{m} I_j\Big] = p + \gamma$$

and thus, by a Chernoff Bound

$$\Pr\Big[\frac{1}{m} \sum_{j=1}^{m} I_j < p\Big] < e^{-2m\gamma^2}$$

Notice, when $\psi_{k+1}, \ldots, \psi_s$ simultaneously simplify to 1 under $\rho^{(i)}$, a proof of φ from $\psi_1, \ldots, \psi_s$ simplifies to a proof of $\varphi|_{\rho^{(i)}}$ from $\psi_1|_{\rho^{(i)}}, \ldots, \psi_k|_{\rho^{(i)}}$.

Since $\mathbb{P}$ is closed under restrictions and there was a proof of φ from $\psi_1, \ldots, \psi_k$, then there is also a proof of $\varphi|_{\rho^{(i)}}$ from $\psi_1|_{\rho^{(i)}}, \ldots, \psi_k|_{\rho^{(i)}}$. Therefore $I_j = 1$ if and only if we increment s for $\rho^{(i)}$, so

$$s \geq \sum_{j=1}^{m} I_j$$

and hence we have that $\frac{s}{m} \geq p$ with probability $1 - \frac{\delta}{2}$. Thus, the second part of the theorem is true with probability $1 - \frac{\delta}{2}$.

For the first part of the theorem, note that when there is a proof of $\psi_1|_{\rho^{(i)}}, \ldots, \psi_k|_{\rho^{(i)}}$ for any assignment to the masked attributes j such that $\rho_j^{(i)} = *$, if this complete assignment y satisfies $\psi_1|_y, \ldots, \psi_k|_y$, then y must also satisfy φ by the soundness of the proofs. In particular, this is true of $y = x^{(i)}$ (where $x^{(i)}$ is the "ground truth" used to generate $\rho^{(i)}$). Therefore, in this case,

$$([\psi_1 \wedge \cdots \wedge \psi_k] \Rightarrow \varphi)(x^{(i)}) = 1$$

and in particular

$$\hat{p} \leq \frac{1}{m} \# \Big\{ i : ([\psi_1 \wedge \cdots \wedge \psi_k] \Rightarrow \varphi)(x^{(i)}) = 1 \Big\}$$

because whenever we'd increment s, $x^{(i)}$ must be satisfied. Since...

$$\mathbb{E}\Big[\frac{1}{m} \# \Big\{ j : ([\psi_1 \wedge \cdots \wedge \psi_k] \Rightarrow \varphi)(x^{(j)}) \Big\} \Big] = \Pr\Big[[\psi_1 \wedge \cdots \wedge \psi_k] \Rightarrow \varphi \Big]$$

...and because the number of j's is a sum over indicator variables...

$$\# \Big\{ j : ([\psi_1 \wedge \cdots \wedge \psi_k] \Rightarrow \varphi)(x^{(j)}) \Big\} = \sum_{j=1}^{m} I\Big[([\psi_1 \wedge \cdots \wedge \psi_k] \Rightarrow \varphi)(x^{(j)}) \Big]$$

... we can use a Chernoff Bound to show that

$$\Pr\Big[\frac{1}{m} \# \Big\{ j : ([\psi_1 \wedge \cdots \wedge \psi_k] \Rightarrow \varphi)(x^{(j)}) \Big\} > \Pr[[\psi_1 \wedge \cdots \wedge \psi_k] \Rightarrow \varphi] \Big] < e^{-2m\gamma^2}$$

By our choice of m, this probability is $\leq \frac{\delta}{2}$. Therefore, $[\psi_1 \wedge \cdots \wedge \psi_k] \Rightarrow \varphi$ is at least $(\hat{p} - \gamma)$-valid with probability $\geq 1 - \frac{\delta}{2}$. So because the conditions of Part 1 and Part 2 each hold individually with probability $\geq 1 - \frac{\delta}{2}$, by the union bound, both hold simultaneously with probability $\geq 1 - \delta$.

■

2 Introduction to Negation-As-Failure

Knowledge learned from observations does not encompass all common sense. We are willing to draw conclusions about unobserved attributes. **Negation-as-failure** is another mechanism for "leaping to conclusions". Roughly, it allows us to write rules that are contingent on *lack* of knowledge. More precisely:

We will extend the representations used by chaining to include a new kind of literal that may only appear in the body of rules. For each of our original "fact" literals l, we add a new literal "not l", where "not " is the symbol for negation-as-failure (a.k.a. "naf"). The intended meaning of $[l_1 \wedge \cdots \wedge l_r \wedge \text{not } l_{r+1} \wedge \cdots \wedge \text{not } l_{r+s}] \Rightarrow l$ is that if we have proven $l_1, \ldots, l_r$ and we cannot prove $l_r, \ldots, l_{r+s}$, then we may infer l.

2.1 Naive Attempt at Supporting Negation-As-Failure

Naive strategy: Try to prove each of $l_1, \ldots, l_{r+s}$. If we succeed at proving $l_1, \ldots l_r$, but could not prove $l_{r+1}, \ldots, l_{r+s}$, then we conclude l. Unfortunately, negation as failure is subtle. Consider the following examples:

Example 1 Support that our knowledge base is given

$$KB = \left\{ \begin{array}{ll} \text{not } x_1 & \Rightarrow x_2 \\ \text{not } x_2 & \Rightarrow x_1 \end{array} \right\}$$

Note that there are two sets of consistent answers: either x_1 is provable and x_2 is not, or x_2 is provable and x_2 is not. Furthermore, a justification of x_1 would rely on circular reasoning: x_2 needs to be not provable for x_1 to be provable, and x_1 needs to provable for x_2 to be not-provable. So there is no way to prove one answer over the other.

Example 2 Now suppose we have

$$KB = \left\{ \begin{array}{ll} \text{not } x_1 & \Rightarrow x_2 \\ \text{not } x_2 & \Rightarrow x_1 \\ \text{not } x_3 & \Rightarrow x_3 \end{array} \right\}$$

Can we prove x_3? Well if we can't prove x_3 then we can prove x_3, which means we can't infer x_3 after all, which means we can prove x_3, which means we can't, which means we can, which means we can't, which means we can, etc... So there is no consistent set of judgments on what is provable and what isn't

Example 3 Now suppose we have

$$KB = \left\{ \begin{array}{ll} \text{not } x_1 & \Rightarrow x_2 \\ \text{not } x_2 & \Rightarrow x_1 \\ \text{not } x_3 & \Rightarrow x_3 \\ x_1 & \Rightarrow x_3 \end{array} \right\}$$

Thanks to the last rule, there's now a unique consistent choice. x_1 and x_3 are provable, but x_2 is not.

Unfortunately, finding a consistent set of judgments on what is provable is NP-Complete. ☹

2.2 Partial Models

Partial models are an elegant way out of these problems. We take a weaker interpretation of "not l" where we allow ourselves to sometimes avoid committing to whether or not l is provable (e.g. by not declaring any of x_1, x_2, or x_3 to be provable in any of the above three examples). A consistent set of judgments on a subset of the literals is called a "partial model". Formally,

Definition 2 *A **partial model** is given by disjoint sets of literals $L_\top$ and $L_\bot$, such that for every rule $[l_1 \wedge \cdots \wedge l_r \wedge not\ l_{r+1} \wedge \cdots \wedge not\ l_{r+s}] \Rightarrow l$ in the KB, if $l_1, \ldots, l_r$ and not $l_{r+1}, \ldots, not\ l_{r+s} \in L_\top$, then $l \in L_\top$ also, and if $l \in L_\bot$, then at least one of $l_1, \ldots, l_r, not\ l_{r+1}, \ldots, not\ l_{r+s}$ must also be in $L_\bot$. We treat facts as rules with empty bodies, so we must have all facts in $L_\top$.*

If we simply take $L_\top$ to be the result of all forward chaining derivations on KB, and take $L_\bot$ to be empty, then this is a partial model of KB. A *too conservative* choice–nothing is declared unprovable. Next time we'll see a better partial model.