1 Main topics

1. Analysis of implicit learning\(^1\).

2. Introduction to “Negation-as-failure”.

2 Definitions from the last lecture

0. **Masking function** \(m(x)\) takes \(x\) to a consistent partial example \(\rho\).

1. Masking process \(M\) is a masking function-valued random variable.

2. (1\(\varepsilon\))-testable \(\psi\) with respect to \(M(\mathcal{D})\): With probability \((1\varepsilon)\) over \(\rho\) drawn from \(M(\mathcal{D})\), \(\psi|_\rho\) simplifies to 1. (N.B. \(O(|\psi|)\)-time computable).

3. Closed under restrictions: Given any proof \(\pi \in \mathcal{P}\) (i.e. proving \(\varphi\) from \(\psi_1, \psi_2, \ldots, \psi_k\)) and partial example \(\rho\), there is a proof of \(\varphi|_\rho\) from \(\psi_1|_\rho, \psi_2|_\rho, \ldots, \psi_k|_\rho\).

2.1 Theorems

1. Chaining is closed under restrictions.

2. (HW4) Clause space \(S\) treelike resolution is closed under restrictions.

3 Main Theorem and Proof

**Theorem 1** For \(m = \frac{1}{\pi^2} \ln \frac{3}{\delta}\), if \(\rho^{(1)}, \rho^{(2)}, \ldots, \rho^{(m)}\) are i.i.d. draws from \(M(\mathcal{D})\), then with probability at least \((1 - \delta)\), the output \(\widehat{p}\) of Algorithm 1 satisfies

1. \([\psi_1 \land \psi_2 \land \ldots \land \psi_k] \Rightarrow \varphi\) is \((\widehat{p} - \gamma)\)-valid with respect to \(\mathcal{D}\).

2. If

   (a) there are \(\psi_{k+1}, \psi_{k+2}, \ldots, \psi_s\) such that \(\psi_{k+1} \land \psi_{k+2} \land \ldots \land \psi_s\) is \((p + \gamma)\)-testable with respect to \(M(\mathcal{D})\),

   (b) proof system \(\mathcal{P}\) is closed under restrictions,

   (c) there is a proof of \(\varphi\) from \(\psi_1, \psi_2, \ldots, \psi_s\) in \(\mathcal{P}\),

   then \(\widehat{p} \geq p\).

**Proof** We shall prove Part 2 of the theorem first, and then Part 1.

---

\(^1\)Recall that “implicit learning” refers to a task of the following form: Given a knowledge base \(KB\), a query \(\varphi\), and partial examples \(\rho^{(1)}, \rho^{(2)}, \ldots, \rho^{(m)}\) \(\in M(\mathcal{D})\) where \(\mathcal{D}\) is an arbitrary distribution, check if \((KB, \text{testable } \psi\text{'s}) \models \varphi\) in a PAC-semantics sense. For further clarification and relevant definitions, see Section 2.
Algorithm 1 Algorithm for reasoning with partial examples

\textbf{input:} query \(\varphi\); KB \(\psi_1, \psi_2, \ldots, \psi_k\); partial examples \(\rho^{(1)}, \rho^{(2)}, \ldots, \rho^{(m)}\)

\textbf{initialize} \(S \leftarrow 0\)

\textbf{for} \(i = 1, 2, \ldots, m\) \textbf{do}

\quad if \(\mathcal{P} (|\varphi|_{\rho^{(i)}}, \psi_1|_{\rho^{(i)}}, \psi_2|_{\rho^{(i)}}, \ldots, \psi_k|_{\rho^{(i)}}) = 1\) then

\quad \quad increment \(S\)

\quad return \(\hat{p} = \frac{S}{m}\)

Proof of Part 2: The key observation is that since \(\psi_{k+1} \land \psi_{k+2} \land \ldots \land \psi_s\) is \((p + \gamma)\)-testable, with probability \((p + \gamma)\) over partial examples \(\rho \in \mathcal{M}(\mathcal{D})\), all of \(\psi_{k+1}|_{\rho}, \psi_{k+2}|_{\rho}, \ldots, \psi_s|_{\rho}\) are simultaneously simplifying to 1.

When this happens, our proof of \(\varphi\) from \(\psi_1, \psi_2, \ldots, \psi_s\) simplifies under \(\rho\) to give a proof of \(\varphi|_{\rho}\) from \(\psi_1|_{\rho}, \psi_2|_{\rho}, \ldots, \psi_k|_{\rho}, 1, \ldots, 1\) in \(\mathcal{P}\).

Thus, \(S\) is at least the fraction of \(\rho^{(1)}, \rho^{(2)}, \ldots, \rho^{(m)}\) for which \(\psi_{k+1} \land \psi_{k+2} \land \ldots \land \psi_s|_{\rho^{(i)}} = 1\).

Now, for \(j = 1, 2, \ldots, m\), define

\[ I_j := \begin{cases} 1 & \text{if } \psi_{k+1} \land \psi_{k+2} \land \ldots \land \psi_s|_{\rho^{(j)}} = 1, \\ 0 & \text{otherwise.} \end{cases} \]

Notice \(S = \sum_{j=1}^{m} I_j\), \(\hat{p} = \frac{S}{m} = \frac{1}{m} \sum_{j=1}^{m} I_j\).

Moreover, \(\mathbb{E}[I_j] = 1 \cdot \Pr[\psi_{k+1} \land \psi_{k+2} \land \ldots \land \psi_s|_{\rho^{(j)}} = 1] + 0 \cdot \Pr[\psi_{k+1} \land \psi_{k+2} \land \ldots \land \psi_s|_{\rho^{(j)}} = 0] = \Pr[\psi_{k+1} \land \psi_{k+2} \land \ldots \land \psi_s|_{\rho^{(j)}} = 1] = p + \gamma. \)

Hence, by the sum law of expectations,

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

Finally, by additive Chernoff bound,

\[ \Pr[\hat{p} < p] = \Pr[\frac{S}{m} < \mathbb{E}\left[\frac{S}{m}\right] - \gamma] \leq \exp\left(-2m\gamma^2\right) \leq \frac{\delta}{2} \quad \text{since} \quad m = \frac{1}{2\gamma^2 \ln \frac{2}{\delta}}. \]

So, with probability at least \((1 - \frac{\delta}{2})\), \(\hat{p} \geq p\). This completes the proof of Part 2. \(\diamond\)

Proof of Part 1: We note that when there is a proof of \(\varphi|_{\rho^{(j)}}\) from \(\psi_1|_{\rho^{(j)}}, \psi_2|_{\rho^{(j)}}, \ldots, \psi_k|_{\rho^{(j)}}\), then by the soundness of the proof system, we actually have \(\{\psi_1|_{\rho^{(j)}}, \psi_2|_{\rho^{(j)}}, \ldots, \psi_k|_{\rho^{(j)}}\} \vdash \varphi|_{\rho^{(j)}}\).

So, \(\{\psi_1|_{\rho^{(j)}}, \psi_2|_{\rho^{(j)}}, \ldots, \psi_k|_{\rho^{(j)}}\} \vdash \varphi|_{\rho^{(j)}}\) is a tautology, i.e. true for any assignment \(y\) to the (free) attributes. In particular, \(\rho^{(j)} = m^{(j)}(x^{(j)})\), and if we take \(y\) as given by \(x^{(j)}\), then we see that \(\{\psi_1 \land \psi_2 \land \ldots \land \psi_k\} \vdash \varphi\)(\(x^{(j)}\)) = 1. Now, let us define

\[ p^* := \Pr_D[\{\psi_1 \land \psi_2 \land \ldots \land \psi_k\} \Rightarrow \varphi] = 1]. \]

\[ I'_j := \begin{cases} 1 & \text{if } \{\psi_1 \land \psi_2 \land \ldots \land \psi_k\} \Rightarrow \varphi\)(x^{(j)}) = 1, \\ 0 & \text{otherwise.} \end{cases} \]
\[ S' := \sum_{j=1}^{m} \mathbb{E}[I'_j]; \text{ clearly, } S \leq S'. \]

As before, \( \mathbb{E}[S'] = mp^* \) so that \( p^* = \mathbb{E}\left[\frac{S'}{m}\right] \), and applying additive Chernoff bound, we obtain

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

Hence, with probability at least \( 1 - \frac{\delta}{2} \), \( \hat{p} - \gamma \leq p^* = \Pr_D[([\psi_1 \land \psi_2 \land \ldots \land \psi_k] \Rightarrow \varphi) = 1] \). By the union bound, the theorem holds.

### 4 Negation-as-failure

Our implicit KB given by testable rules captures conventions, natural kinds etc., but it doesn’t capture cases where we are willing to “leap to conclusions” in the absence of ever having been able to actually observe these attributes in question.

We need a new mechanism for these kinds of knowledge. One flexible mechanism is known as “negation-as-failure”: roughly, it allows one to express a lack of knowledge in the representations we reason with, such as to say things like “Unless I have reason to believe otherwise, my office stays as I left it.” or “Unless I have reason to believe otherwise, a bird I encounter can fly.” etc..

**Note:** Since each of these rules refers to a lack of knowledge, additional information can force us to withdraw each of these conclusions. **More precisely,** consider the chaining proof system, extended to the set of representations of rules in which a new kind of literal may appear in the body: For each of our original literals \( l \), we introduce the literal \( \text{not} l \) where \( \text{not} \) is a symbol for negation-as-failure (or “naf” for short). The intended meaning of a rule

\[
[l_1 \land \ldots \land l_r \land \text{not}l_{r+1} \land \ldots \land \text{not}l_{r+s}] \Rightarrow l
\]

is: If we have proven \( l_1,\ldots,l_r \), and cannot prove \( l_{r+1},\ldots,l_{r+s} \), then we should be able to infer \( l \) by chaining.

A naïve attempt to negation-as-failure would be to take this at face value, and attempt to prove each of the \( l_{r+1},\ldots,l_{r+s} \), and conclude \( l \) if each of these attempts fail. But negation-as-failure is far more subtle than it first appears.

- Consider the KB

\[
[\text{not}x_1] \Rightarrow x_2, \quad [\text{not}x_2] \Rightarrow x_1.
\]

Can we prove \( x_1 \)? If the answer is “no”, then we can conclude \( x_2 \) from the above KB. However, since the KB is symmetric in \( x_1 \) and \( x_2 \), and we say that we cannot prove \( x_1 \), then we should not be able to prove \( x_2 \) by the same reasoning. Hence, we have a contradiction!

- Now, let us add the following rule to the above KB.

\[
[\text{not}x_3] \Rightarrow x_3.
\]

This rule alone lands us in a paradox since if we say that we cannot prove \( x_3 \), then the KB immediately tells us that we can prove \( x_3 \)!

- Finally, let us add another rule to the KB:

\[
[x_1] \Rightarrow x_3.
\]
Suppose, we can prove $x_1$. Then we cannot prove $x_2$ since only the rule $\text{not}x_1 \Rightarrow x_2$ can prove $x_2$. There is still no contradiction so we are OK with the assertion that we can prove $x_1$. From $x_1$, we can prove $x_3$.

If we had instead supposed that we cannot prove $x_1$, we would get the paradox of $x_3$ being provable iff it is not.

Hence, by adding the fourth rule to a KB leading to paradoxes, we arrive at the consistent set of assertions: We can prove $x_1$ and $x_3$, and cannot prove $x_2$. The worst part is that choosing a consistent set of judgements on what is provable and what is not is NP-complete! There is a way out by using partial models: We get a poly-time algorithm that avoids paradoxes and circular reasoning.