CSE 513T Theory of Artificial Intelligence & Machine Learning Mar 20, 2018

Lecture 16

Lecturer: Brendan Juba *Scribe: Zihao Deng*

Today's Topics

- Partial observability
- Testable KBs and closure under restrictions
- Reasoning with partial examples: implicit learning

1 Partial observability

"Reasoning with examples" enabled using rules with $(1-\epsilon)-$validity, where ϵ "noticeably large". It captured (some) common sense rules. However it required observing every attribute, which is unrealistic.

We've seen two extremes:

- PAC learning, completely specified examples from D. This approach learns from what is observed
- Logical reasoning have variables, rules, but *no* observations. This approach uses reasoning to cope with what we don't observe.

Now it's interesting to consider something "in between". We still assume some distribution D on X over n attributes, and we still want to know if the φ is $(1-\epsilon)-$valid with respect to distribution D. However, we only have partial observability which sees some but not all. Concretely speaking, now we are only given access to partial examples $\rho \in \{0,1,*\}^n$ generated by applying a "masking function" to a complete example $x \in X$.

Definition 1 *A masking function $m : \{0,1\}^n \to \{0,1,*\}^n$ is such that $m(x)$ is consistent with x, i.e., if $m(x)_i \neq -x_i$, then $m(x)_i = x_i$.*

In general, we'll have a *masking process* M, which is a masking-function-valued random variable. For simplicity, M and the sampling of x from D are independent, but note $m(x)$ is a function of x. We use $M(D)$ to denote the distribution over partial assignment obtained by applying a masking process M to the distribution D. The followings are some examples of masking functions.

Example 1: Observed vs. Latent. There is a fixed pattern of $*$ and non-$*$ attributes in every example; attributes are thus either never observed, a.k.a. "latent", or always observed. Intuitively speaking, whether or not $m(x)_i = *$ is independent of x. Concretely speaking, for all x, x' and $i \in \{1,\ldots,n\}$, either both $m(x)_i = *$, $m(x')_i = *$, or both $m(x)_i = x_i$, $m(x')_i = x'_i$, regardless of what x, x' are.

Example 1: "Lossy data". Suppose M_μ (where μ is a real number from $(0,1)$) is a distribution over masking functions of the previous type (i.e., $m(x)_i = m(x')_i \forall i$), s.t., for each $S \subseteq \{1,\ldots,n\}$, if $m(x)$ hides $|S|$ attributes, then m is drawn upon probability $\mu^{n-|S|}(1-\mu)^{|S|}$. Equivalently, we generate $M_\mu(x)_i$ by independently tossing a coin of bias μ for each i and setting it to $*$ if the coin comes up tails. That is, it deletes or drops the entries at random.

Example 3: "Limited visibility". Suppose the attributes of X correspond to the state of some environment, we have sensors that can only read a portion at a time, depending on our current location in the environment. Then M produces a fixed masking function m that: (i) hides attributes of locations not "visible" from current location (encoded as an attribute); (ii) passes through the rest.

2 Testable Rules

We want to say the "easy-to-see" rules ψ on partial examples from $M(D)$ are in KB. However, we must be cautious that tautologies may be "true" but not easy to check (NP-complete in general). So instead, we'll use partial evaluation.

Definition 2 *We say that a representation φ is $(1-\epsilon)-$ testable with respect to a distribution on partial examples $M(D)$ if $Pr_{\rho \in M(D)}[\varphi|_\rho = 1] \geq 1-\epsilon$.*

Recall in Homework 7, we saw that $C|_\rho$ implies C due to the "weakening rule" for the smaller clause $C|_\rho$. So for a CNF φ, each clause evaluates to 1 if any of its literals do under ρ, and evaluates to 0 only if all of its literals do under ρ; φ evaluates to 1 if all of its clauses do under ρ, and evaluates to 0 only if any of its clauses do under ρ. Here is the example from the last time. We sample x from $D|_{\texttt{bird}}$ then $\texttt{can_fly}(x) \approx \frac{5}{7}-$valid. Suppose M doesn't obscure $\texttt{can_fly}$. Then $M(D|_{\texttt{bird}})$ produces ρ such that $\texttt{can_fly}$ is $\frac{5}{7}-$testable under $M(D|_{\texttt{bird}})$.

3 Restriction-closed proof system

We'll make use of these testable rules in answering queries in restriction closed proof systems.

Definition 3 *We say that a set of proofs P for a proof system is restriction-closed if whenever there is a proof $\pi = \pi_1, \pi_2, \ldots, \pi_\ell$ of φ from premises $\psi_1, \ldots, \psi_k$ in P, then for every partial assignment ρ, there is a a proof of φ_ρ from $\psi_1|_\rho, \ldots, \psi_k|_\rho$ in P, given that none of $\psi_1|_\rho, \ldots, \psi_k|_\rho$ are 0.*

Note that it's hard to propose natural sets of proofs that aren't restriction-closed. For example, chaining proofs are closed under restrictions. Recall our example, the chaining proof of "$\texttt{mischief}$" from $KB = \{\texttt{dog}, \texttt{young}, [\texttt{dog} \wedge \texttt{young}] \Rightarrow \texttt{puppy}, \texttt{puppy} \Rightarrow \texttt{mischief}\}$. Suppose ρ sets $\texttt{dog}$ and $\texttt{puppy}$ to 1, then $KB|_\rho = \{\texttt{young}, \texttt{mischief}\}$ where $(\texttt{puppy} \Rightarrow \texttt{mischief})|_\rho = (\neg\texttt{puppy} \vee \texttt{mischief})|_\rho = \texttt{mischief}$ is a trivial proof that asserts $\texttt{mischief}$. Suppose ρ only sets $\texttt{dog}=1$, then $KB|_\rho = \{\texttt{young}, \texttt{young} \Rightarrow \texttt{puppy}, \texttt{puppy} \Rightarrow \texttt{mischief}\}$, where $([\texttt{dog} \wedge \texttt{young}] \Rightarrow \texttt{puppy})\,|_{\texttt{dog}=1} = (\neg\texttt{dog} \vee \neg\texttt{young} \vee \texttt{puppy})\,|_{\texttt{dog}=1} = (\neg\texttt{young} \vee \texttt{puppy}) = (\texttt{young} \Rightarrow \texttt{puppy})$. This $KB|_\rho$ obviously have a chaining proof for $\texttt{mischief}$. So the chaining proof for $\texttt{mischief}$ is also closed under this under this partial assignment. Now let's formally prove this claim.

Claim 4 *Chaining proofs are closed under restrictions.*

Proof Consider any chaining proof $\pi = \pi_1, \pi_2, \ldots, \pi_\ell$ of a representation φ from premises $\psi_1, \ldots, \psi_k$, and let any partial assignment ρ be given. Let π' be the sequence obtained by

- (1) replacing the ith step π_i by $\pi_i|_\rho$;
- (2) omitting any portion of the graph of π leading to some step $\pi_j|\rho$ such that $\pi_j|\rho$ is a Boolean constant (Recall from lecture 14 that for any chaining proof π there is graph associated with it).

Now π' only contains literals ("facts") that are not fixed by ρ, and only contains "rules" $[\ell_1, \ldots, \ell_r] \Rightarrow \ell$ for which:

- (a) the head literal ℓ is not fixed to a constant;
- (b) no body literal ℓ_i is fixed to 0 by ρ.

We'll argue by induction on the steps of π', that the first j steps are a legal chaining derivation of π'_j from $\psi_1|_\rho, \ldots, \psi_k|_\rho$. Indeed, if π'_j is a fact, then there are three possible cases. Case(i): π_j was among $\psi_1, \ldots, \psi_k$ and not set by ρ. Then $\pi'_j = \pi_j|_\rho = \pi_j = \psi_j|_\rho = \psi_j$ for some fact ψ_j in the premises. Case(ii): $\pi_j = [\ell_1, \ldots, \ell_r] \Rightarrow \ell$ and we obtained $\pi'_j = \ell$ because $\ell_1, \ldots, \ell_r$ were all fixed to 1 by ρ. In this case

$\pi'_j = \ell = \pi_j|_\rho = \psi_j|_\rho$ for some rule ψ_j in the premises. In these two cases it's legal to assert the fact π'_j from some $\psi_j|_\rho$. Case (iii): π_j was originally derived by chaining. That is, some rule asserted $\pi_{j,1}$ earlier with the facts in its body, which were asserted earlier as $\pi_{j,2}, \ldots, \pi_{j,r}$. But we know that none of the body literals were set to 0 by ρ according to (a). So, of the literals in the body that are not set to 1 by ρ, each appears in some earlier line of π' by construction. Likewise, if π_j is simply a rule and not all of the literals in the body are set to 1 (none set to 0 according to (a)), again we have the partially evaluated rule $\psi_j|_\rho$ in the KB, so we can assert it. So in any case, we have a legal chaining step deriving π'_j, and since the final representation in π' is $\varphi|_\rho$, this is a chaining proof of $\varphi|_\rho$ from $\psi_1|_\rho, \ldots, \psi_k|_\rho$. ■

4 Reasoning with partial examples

Suppose we can test for a proof of a query in P efficiently. We compute $\hat{p}$, which is the fraction of times (observed probability) that P contained a proof of query from KB under partial assignment $\rho \in M(D)$.

Theorem 5 *For $m \geq \frac{1}{2\gamma^2} \ln \frac{2}{\delta}$ partial examples $\rho^{(1)}, \ldots, \rho^{(m)}$ independently drawn from $M(D)$, with probability $1 - \delta$, the fraction $\hat{p}$ satisfies*

- *$[\psi_1 \wedge, \ldots, \wedge \psi_k] \Rightarrow \varphi$ is $(\hat{p} - \gamma)-$valid with respect to D. Here the set of premises $\psi_1 \wedge, \ldots, \wedge \psi_k$ is the KB.*
- *If (a): there are representations $\psi_{k+1}, \ldots, \psi_s$ such that $\psi_{k+1} \wedge, \ldots, \wedge \psi_s$ is $(p + \gamma)-$testable (these representations are "implicit KB"), (b): P is closed under restrictions, and (c): there is a proof of φ from $\psi_1, \ldots, \psi_k, \psi_{k+1}, \ldots, \psi_s$ in P, then $\hat{p} \geq p$.*

We will discuss its proof in the next lecture.