Missing Information Impediments to Learnability

Loizos Michael
Open University of Cyprus
loizos@ouc.ac.cy
Given access to \((\text{exm}, \alpha)\)'s drawn during training, w.h.p. and efficiently produce \(h\) s.t. on \((\text{exm}', \alpha')\) drawn during testing, \(h(\text{exm}') \neq \alpha'\) w.p. at most \(\varepsilon\).
Given access to \((\text{obs}, \alpha)\)'s drawn during training, w.h.p. and efficiently produce \(h\) s.t. on \((\text{obs}', \alpha')\) drawn during testing, \(h(\text{obs}') \neq \alpha'\) w.p. at most \(\varepsilon\).
Given access to \((\text{obs}, \alpha')\)’s drawn during training, w.h.p. and efficiently produce \(h\) sat. on \((\text{obs'}, \alpha')\) drawn during testing, \(h(\text{obs'}) \neq \alpha'\) w.p. at most \(\varepsilon\).
Discussion: Hypotheses

- Hypotheses encode the structure of examples;
  - not observations (cf. Schuurmans & Greiner `94).

- Valid hypothesis (as in PAC): $h \equiv (x_1 \land \neg x_3 \land x_7)$
  - works on examples; extended to observations
  - e.g., on observation $10*1101$ $h$ is undefined

- Invalid hypothesis: $h \equiv (if \ x_3 = \ast \ then \ predict \ 1)$
  - works directly on observations (ternary strings)
  - e.g., on observation $10*1101$ $h$ predicts 1
Discussion: Abstentions

- Masking is arbitrary (MNAR), and stochastic.
  - e.g., $\text{mask}(\text{exm}_0) = \text{obs}_1$ w.p. 1/3, $\text{obs}_2$ w.p. 2/3

- Learning requirements are not too hard.
  - on observation $\mathbf{h} \notin \{\bot, \top\}$ undefined
  - wrong predictions penalized; not abstentions

- Learning requirements are not too easy.
  - PAC learning is necessary condition ($\text{mask} \equiv \text{ID}$)
  - abstentions not actively chosen by hypotheses
Positive Learnability Results

• Positive results for learning consistently:
  • PAC-learnable classes of monotone formulas
  • PAC-learnable classes of read-once formulas

• Straightforward algorithm and proof:
  • during training map $0**1*0 \alpha$ to $0\alpha\alpha1\alpha0 \alpha$
  • PAC-learn from resulting (complete) instances
  • typical learning reduction for read-once case
Known Result: Learnability of 1-CNF.
- using learning reduction from monotone case

Open Prob. 1: Learnability of 2-CNF.
- value((x_1 \lor \neg x_2) \land (x_2 \lor x_3)) on 0110 / 0*01 is 0
- value(x_1 \lor \neg x_2) \land value(x_2 \lor x_3) on 0110 is 0 \land 1
- value(x_1 \lor \neg x_2) \land value(x_2 \lor x_3) on 0*01 is ? \land ?

Open Prob. 2: Learnability of 3-CNF.
- value(3-CNF) on ******* is 0 \iff 3-CNF \not\in SAT
Negative Learnability Results

• Negative results for learning consistently:
  • cannot learn parities properly (if $\text{RP} \neq \text{NP}$)
  • same for monotone-term 1-decision lists

• Built on typical proofs in the literature:
  • $h$ consistent with $0**000\ 1$ iff $x_2$ or $x_3$ in $h$
  • need underlying examples for observations
  • holds even if only three attributes masked
  • similar analysis for the case of decision lists
Open Probs: Non-Proper Learning

- **Open Prob. 3:** Non-proper learnability of parities, or monotone-term 1-decision lists.
  - classes of highly non-monotone formulas

- More generally, learnability (properly or not) of formulas that are not shallow for (i.e., not “easily” reducible to) monotone formulas.
Why Care About These Problems?

• Understanding the extent to which missing information impedes the learning process.
  • Can investigate special masking processes.

• Real-world corpora have missing information in manner not easy to model (e.g., MNAR).
  • Positive results will hold in all these cases.

• Results carry to autodidactic learning model.
  • Complete missing information in instances.
Remarks on Completeness

• Require PAC-like guarantees on consistency.
  • \( h(\text{obs}) \) inconsistent with \( \text{label} \) w.p. at most \( \varepsilon \)
  • no need to worry about when \( h(\text{obs}) \) defined

• But, if one wishes to say something more...
  • cannot PAC-learn when to predict / abstain
  • cannot ensure non-negligible completeness

• Open Prob. 4: Offer agnostic-like guarantees.