Missing Information Impediments to Learnability

Loizos Michael
Open University of Cyprus
loizos@ouc.ac.cy

Abstract

To what extent is learnability impeded when information is missing in learning instances? We present relevant known results and concrete open problems, in the context of a natural extension of the PAC learning model that accounts for arbitrarily missing information.

1 Learning from Partial Observations

In the PAC learning model (Valiant, 1984), ***examples*** are drawn from an unknown fixed distribution $\mathcal{D}$ over $\{0,1\}^n$. A label for each example is determined by applying an unknown fixed ***target*** function $f \in \mathcal{C}$ on the example; the class $\mathcal{C}$ of all such targets is the ***concept class***. Given access to labeled examples during a training phase, a learner seeks to produce, efficiently and with high probability, a ***hypothesis*** function $h \in \mathcal{H}$ that predicts, with high probability, the labels of examples drawn from $\mathcal{D}$ and labeled according to f; the class $\mathcal{H}$ of all such hypotheses is the ***hypothesis class***.

Explicit in the definition of the PAC learning model is the requirement that each example offers sufficient information to determine its label; the primary challenge of learning is, thus, to identify *how* to do so. In certain settings (e.g., in a typical medical database), however, not all information necessary to determine the label is available in an example (e.g., due to medical tests not performed). Furthermore, this happens during *both* the training *and* the testing phase, and the manner in which information is missing might critically depend on the information itself. In the spirit of supervised learning, we consider only settings where example labels are never missing during the training phase.

These partial (but noiseless) views of examples we shall call ***observations***. We represent them as ternary vectors $\texttt{obs} \in \{0,1,*\}^n$, with the value $*$ indicating that the corresponding attribute was not observed. Examples are mapped to observations through a ***masking process***, a stochastic process $\texttt{mask} : \{0,1\}^n \to \{0,1,*\}^n$ that induces a distribution over observations, which may depend on the example being mapped. The noiseless nature of observations implies that whenever an observation $\texttt{obs}$ is drawn from $\texttt{mask}(\texttt{exm})$, it holds that $\texttt{obs}[i] \in \{\texttt{exm}[i], *\}$, where $\texttt{obs}[i]$ and $\texttt{exm}[i]$ correspond, respectively, to the value of the i-th attribute according to $\texttt{obs}$ and $\texttt{exm}$. Such an observation $\texttt{obs}$ is said to ***mask*** the example $\texttt{exm}$, and each attribute with $\texttt{obs}[i] = *$ is said to be ***masked*** in $\texttt{obs}$.

Each observation is drawn from the oracle $\texttt{sense}(\mathcal{D}; f; \texttt{mask})$ in unit time, thus: *(i)* an example $\texttt{exm}$ is drawn from $\mathcal{D}$; *(ii)* the label of $\texttt{exm}$ is assigned to be $f(\texttt{exm})$; *(iii)* an observation $\texttt{obs}$ that masks $\texttt{exm}$ is drawn from $\texttt{mask}(\texttt{exm})$; *(iv)* both $\texttt{obs}$ and its label, which equals $f(\texttt{exm})$, are returned.

A learner seeks to produce, as in the PAC learning model, a hypothesis for predicting the labels. The hypothesis is a function over the *boolean attributes*, and encodes (learned) knowledge about the structure of the *underlying examples* — not knowledge about the structure of observations and the way the masking process hides information (cf. Schuurmans and Greiner, 1994). The PAC learning model can be viewed as the special case of this model when the masking process $\texttt{mask}$ is an identity.

Since a hypothesis h is defined over boolean attributes but evaluated on an observation $\texttt{obs}$, its value $h(\texttt{obs})$ may possibly remain undefined; exactly when $h(\texttt{exm})$ is not constant across all examples $\texttt{exm}$ masked by $\texttt{obs}$. In such a case, h abstains from a prediction. Abstentions are not penalized, as they are not actively chosen by the hypothesis. We shall say that a hypothesis h has a ***consistency conflict*** with an observation $\texttt{obs}$ if h does not abstain, and $h(\texttt{obs})$ differs from the label of $\texttt{obs}$.

A hypothesis h is said to be ***ε-inconsistent*** w.r.t. the oracle $\texttt{sense}(\mathcal{D}; f; \texttt{mask})$ if the probability that h has a consistency conflict with an observation $\texttt{obs}$ drawn from $\texttt{sense}(\mathcal{D}; f; \texttt{mask})$ is at most ε.

Definition 1 *A concept class $\mathcal{C}$ is **consistently learnable** by a hypothesis class $\mathcal{H}$ if there exists an algorithm $\mathcal{L}$ such that for every natural number n, every distribution $\mathcal{D}$ over $\{0,1\}^n$, every target*

function $f \in \mathcal{C}$ over n attributes, every masking process $\texttt{mask}$ *over n attributes, every real number $\delta \in (0, 1]$, and every real number $\varepsilon \in (0, 1]$, algorithm $\mathcal{L}$ has the following property:*

given the parameters n, $\mathcal{C}$, $\mathcal{H}$, δ, ε as input, and given access to the oracle $\texttt{sense}(\mathcal{D}; f; \texttt{mask})$ *to obtain observations, algorithm $\mathcal{L}$ runs in time polynomial in n, $1/\delta$, $1/\varepsilon$, and the size of f, and returns, with probability at least $1 - \delta$, a hypothesis $h \in \mathcal{H}$ that is ε-inconsistent w.r.t.* $\texttt{sense}(\mathcal{D}; f; \texttt{mask})$.

Consistent learnability asks that the typical PAC guarantees hold for *each* masking process. The resulting learning requirements are not overly demanding, since exactly when learnability may suffer due to less information in observations, hypotheses may abstain more and avoid consistency conflicts. Abstentions cannot, however, be abused, as they cannot be actively invoked. It is the masking process that effectively determines when hypotheses abstain, and this is beyond the learner's control.

The model of consistent learnability presented herein is a special case of the *autodidactic learning model* (Michael, 2008, 2010). The results in the section that follows were obtained in the context of the latter model. Proofs of the results, details about the model, and comparison to other extensions of the PAC learning model that accommodate missing information, can be found in the cited works.

2 Known Results and Open Problems

Since consistent learnability implies PAC learnability, PAC learnability is a necessary condition for the consistent learnability of any given concept class. The PAC learnability property in conjunction with either the monotone or the read-once property is a sufficient condition for consistent learnability.

Theorem 2 *A concept class $\mathcal{C}$ that comprises either only monotone formulas or only read-once formulas, is consistently learnable by a hypothesis class $\mathcal{H}$, assuming that $\mathcal{C}$ is PAC learnable by $\mathcal{H}$.*

Thus, the concept classes of conjunctions and linear thresholds (Kearns and Vazirani, 1994) are consistently learnable. Unlike in the PAC learning model, a learning reduction cannot be readily employed to establish the learnability of k-CNFs for constant k. This holds because k-CNFs cannot be evaluated *modularly* on observations (unlike on examples). Indeed, the value of the conjunction of two subformulas on some observation may not be determinable only by the values of the subformulas (e.g., when they are undefined), but may require knowledge of the subformulas themselves. Hence:

Problem 3 *Is the concept class $\mathcal{C}$ of 2-CNFs consistently learnable by a hypothesis class $\mathcal{H}$? Is the same question true for some other concept class of formulas that are not modularly evaluatable?*

The case of 3-CNFs presents an additional challenge, as they are not believed to be evaluatable *efficiently*. Indeed, their evaluation on the observation $*^n$ implies deciding their satisfiability. Hence:

Problem 4 *Is the concept class $\mathcal{C}$ of 3-CNFs consistently learnable by a hypothesis class $\mathcal{H}$? Is the same question true for some other concept class of formulas that are not efficiently evaluatable?*

Despite being a necessary condition, PAC learnability is not, by itself, a sufficient condition for consistent learnability — at least not when the hypothesis and concept classes coincide, and $\texttt{RP} \neq \texttt{NP}$.

Theorem 5 *The concept class $\mathcal{C}$ that comprises either only parities or only monotone-term 1-decision lists, is not consistently learnable by the hypothesis class $\mathcal{H} = \mathcal{C}$, unless* $\texttt{RP} = \texttt{NP}$.

The negative result above holds despite $\mathcal{C}$ being PAC learnable by $\mathcal{H} = \mathcal{C}$ (Kearns and Vazirani, 1994), and even assuming that at most three attributes are masked in each observation. Hence:

Problem 6 *Is the concept class of parities consistently learnable by a hypothesis class $\mathcal{H} \neq \mathcal{C}$? Is the concept class of monotone-term 1-decision lists consistently learnable by a hypothesis class $\mathcal{H} \neq \mathcal{C}$?*

Closing the gap between what is necessary and sufficient for consistent learnability would help in clarifying which PAC learnability results remain applicable when information is missing arbitrarily.

References

Michael Kearns and Umesh Vazirani. *An Introduction to Computational Learning Theory.* The MIT Press, Cambridge, Massachusetts, U.S.A., 1994.

Loizos Michael. *Autodidactic Learning and Reasoning.* PhD thesis, School of Engineering and Applied Sciences, Harvard University, U.S.A., May 2008.

Loizos Michael. Partial Observability and Learnability. *Artificial Intelligence*, 174(11):639–669, July 2010.

Dale Schuurmans and Russell Greiner. Learning Default Concepts. In *Proceedings of the Tenth Canadian Conference on Artificial Intelligence (AI'94)*, pages 99–106, May 1994.

Leslie Valiant. A Theory of the Learnable. *Communications of the ACM*, 27(11):1134–1142, November 1984.