Missing Information Impediments to Learnability

Loizos Michael LOIZOS@OUC.AC.CY
Open University of Cyprus

Editor: Sham Kakade, Ulrike von Luxburg

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.
Keywords: PAC learning, missing information, masking process, open problems.

1. Learning from Partial Observations

In the PAC learning model (Valiant, 1984), ***examples*** are drawn from some unknown fixed probability distribution $\mathcal{D}$ over $\{\mathtt{0},\mathtt{1}\}^n$. A boolean-valued 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 a set of 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 efficiently identify *how* to determine the label. 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 that were not performed). Furthermore, this happens during *both* the training *and* the testing phase, and the manner in which information is missing may 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 $\mathtt{obs} \in \{\mathtt{0},\mathtt{1},*\}^n$, with the value $*$ indicating that the corresponding attribute was not observed. Examples are mapped to observations through a ***masking process***, a stochastic process $\mathtt{mask} : \{\mathtt{0},\mathtt{1}\}^n \to \{\mathtt{0},\mathtt{1},*\}^n$ that induces a probability distribution over observations, which may depend on the example being mapped. The noiseless nature of observations implies that whenever an observation $\mathtt{obs}$ is drawn from $\mathtt{mask}(\mathtt{exm})$, it holds that $\mathtt{obs}[i] \in \{\mathtt{exm}[i], *\}$, where $\mathtt{obs}[i]$ and $\mathtt{exm}[i]$ correspond, respectively, to the value of the i-th attribute according to $\mathtt{obs}$ and $\mathtt{exm}$. Such an observation $\mathtt{obs}$ is said to ***mask*** the example $\mathtt{exm}$, and each attribute with $\mathtt{obs}[i] = *$ is said to be ***masked*** in $\mathtt{obs}$.

Each observation is assumed to be drawn from the oracle $\mathtt{sense}(\mathcal{D}; f; \mathtt{mask})$ in unit time, by means of the following process: *(i)* an example $\mathtt{exm}$ is drawn from $\mathcal{D}$; *(ii)* the label of $\mathtt{exm}$ is computed to be $f(\mathtt{exm})$; *(iii)* an observation $\mathtt{obs}$ that masks $\mathtt{exm}$ is drawn from $\mathtt{mask}(\mathtt{exm})$; *(iv)* the label of $\mathtt{obs}$ is assigned to equal $f(\mathtt{exm})$; *(v)* both $\mathtt{obs}$ and $f(\mathtt{exm})$ are returned.

As in the PAC learning model, a learner in the model that we consider seeks to produce a hypothesis for predicting the labels. The hypothesis is a boolean-valued 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 `mask` is an identity.

Since a hypothesis h is defined over boolean attributes but evaluated on observations, its prediction $h(\texttt{obs})$ on observation `obs` may possibly remain undefined; this occurs exactly when $h(\texttt{exm})$ is not constant across all examples `exm` masked by `obs`. In such a case, h abstains from making 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 `obs` if h does not abstain, and $h(\texttt{obs})$ differs from the label of `obs`.

A hypothesis h is ***ε-inconsistent*** w.r.t. oracle $\texttt{sense}(\mathcal{D}; f; \texttt{mask})$ if h has a consistency conflict with an observation `obs` drawn from $\texttt{sense}(\mathcal{D}; f; \texttt{mask})$ with probability 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 probability distribution $\mathcal{D}$ over $\{\texttt{0}, \texttt{1}\}^n$, every target function $f \in \mathcal{C}$ over n attributes, every masking process* `mask` *over n attributes, and every pair of real numbers $\delta, \varepsilon \in (0, 1]$, algorithm $\mathcal{L}$ is such that:*

given the parameters n, $\mathcal{C}$, $\mathcal{H}$, δ, ε as input, and given access to the oracle $\texttt{sense}(\mathcal{D}; f; \texttt{mask})$, 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. $\texttt{sense}(\mathcal{D}; f; \texttt{mask})$.

The definition of consistent learnability insists that the typical PAC guarantees hold, but for *every* masking process. It is worth pointing out that 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), where there is no distinguished label for observations, and the aim of the learning process is to complete the values of the masked attributes. The results in the section that follows were obtained in the context of the latter model. Proofs of the results, details about that 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, the latter is a necessary condition for the former. PAC learnability in conjunction with either the monotone or the read-once property holding for the concept class is a sufficient condition for consistent learnability.

Theorem 2 *A concept class $\mathcal{C}$ that comprises either monotone or 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 what holds in the PAC learning model, a learning reduction cannot be readily employed to establish the learnability of k-CNFs for constant values of $k \geq 2$. This holds because k-CNFs cannot be evaluated *modularly* on observations (unlike on examples). Indeed, the value of a certain 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 question true for any concept class of formulas that are not modularly evaluatable?*

The case of learning 3-CNFs presents an additional challenge when compared to the case of learning 2-CNFs, since the former formulas 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 question true for any 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 class $\mathcal{H}$ and the concept class $\mathcal{C}$ are required to coincide, and the complexity condition $\texttt{RP} \neq \texttt{NP}$ is assumed.

Theorem 5 *The concept class $\mathcal{C}$ that comprises either parities or 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 holds despite $\mathcal{C}$ being PAC learnable by $\mathcal{H} = \mathcal{C}$ (Kearns and Vazirani, 1994), and even when at most three attributes are masked in each observation. Hence:

Problem 6 *Is the concept class $\mathcal{C}$ that comprises either parities or monotone-term 1-decision lists consistently learnable by a hypothesis class $\mathcal{H}$ that differs from $\mathcal{C}$?*

Refining the necessary and sufficient conditions for consistent learnability would help clarify which PAC learnability results remain true when information is missing arbitrarily, and, hence, which can be applied in realistic settings where the masking process is unknown.

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.