The Information-Theoretic Value of Unlabeled Data in Semi-Supervised Learning

Alexander Golovnev (Harvard) **Dávid Pál (Expedia, New York)** **Balázs Szörényi (Yahoo, New York)**

The Problem

- Distribution D over a domain X
- Unknown target function f from a known hypothesis class $H \subseteq \{0,1\}^X$.
- Learner receives $S = ((x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_m, f(x_m)))$.

Does knowing D help?

Sample Complexity

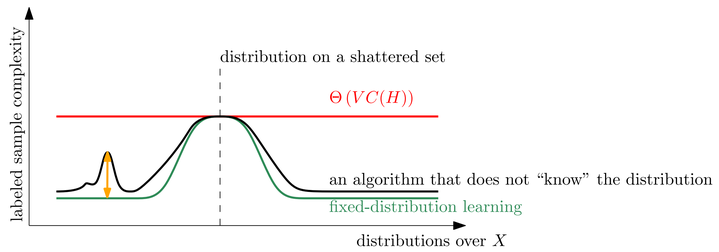
- Learning algorithm outputs a classifier $g = A(S)$.
- Its error is
$$\mathrm{err}_D(g) = \Pr_{x\sim D}[f(x) \neq g(x)]$$
- Sample complexity is the smallest $m = m(\epsilon, \delta, D)$ such that for any target $f \in H$, with probability $\geq 1 - \delta$
$$\mathrm{err}_D(g) \leq \epsilon$$

Known Results

- Without knowing D the sample complexity is
$$O\left(\frac{\mathrm{VC}(H) + \log(1/\delta)}{\epsilon}\right)$$
Modulo log factors this is achieved by any constistent algorithm, i.e. ERM.
- With knowledge of D, the sample complexity is at most
$$O\left(\frac{\log N_{D,\epsilon/2} + \log(1/\epsilon)}{\delta}\right)$$
where $N_{D,\epsilon/2}$ is $\frac{\epsilon}{2}$-covering number of H under the metric
$$d(f,g) = \Pr_{x\sim D}[f(x) \neq g(x)]\ .$$
- Covering number upper bound
$$N_{D,\epsilon} \leq \left(\frac{4e}{\epsilon}\right)^{\frac{VC(H)}{1-1/e}}$$
- There exists a distribution D such that **even with knowledge of** D, any algorithm needs at least
$$\Omega\left(\frac{\mathrm{VC}(H) + \log(1/\delta)}{\epsilon}\right)$$
labeled examples.
- There are distributions D such that any constistent algorithm has sample complexity only $O(\log(1/\delta)/\epsilon)$.

Summary of Known Results

Fix ϵ and δ.



How big is the gap between the black and the green curve?

Projections

Domain is $X = \{0,1\}^n$ and hypothesis class is $H = \{h_1, h_2, \ldots, h_n\}$ where

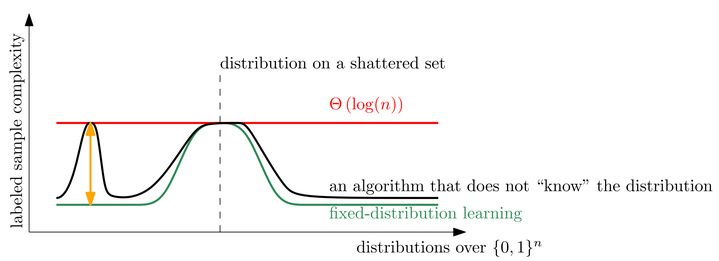
$$h_i(x) = h(x[1], x[2], \ldots, x[n]) = x[i]$$

Vapnik-Chervonenkis dimension is $\mathrm{VC}(H) = \lfloor \log_2(n) \rfloor$.

Theorem. *Fix ϵ and δ. There are distributions $D_1, D_2, \ldots, D_n$ such that*

1. With knowledge of the distribution D_i, sample complexity is $O(1)$.

2. Without knowledge of D_i, sample complexity is $\Omega(\log n)$.



Each D_i is a product distribution such that

$$\Pr_{x\sim D_i}[x[j] = 1] = \begin{cases} 1/2 & \text{if } i = j, \\ \epsilon/4 & \text{if } i \neq j. \end{cases}$$

Sketch of Proof

- D_i has $\epsilon/2$-cover of size 2. Thus $O\left(\frac{\log N_{D,\epsilon/2} + \log(1/\delta)}{\epsilon}\right) = O(1)$ samples are enough to ϵ-learn if D_i is known to the learner.
- Choose $i \in \{1, 2, \ldots, n\}$ at random.
- Choose distribution D_i and the target to be the projection h_i.
- Algorithm that does **not** know D_i and h_i, sees only the matrix
$$\left(\begin{array}{cccc|c} x_1[1] & x_1[2] & \cdots & x_1[n] & y_1 \\ x_2[1] & x_2[2] & \cdots & x_2[n] & y_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_m[1] & x_m[2] & \cdots & x_m[n] & y_m \end{array}\right).$$
- Column $x[i]$ matches column y.
- If $m \leq \log(n)$ then with constant probability at least one other column $x[j]$ matches column y.
- Learner has to pick a column i or j.

For non-proper learners, the proof is more complicated.

Conclusions

- Unlabeled data help for projections.
- For the class of **all functions**, unlabeled data do **not** help.
- The problem is open for halfspaces and axis-aligned rectangles in $\mathbb{R}^n$, and conjuctions and disjuctions in $\{0,1\}^n$. They have $\mathrm{VC}(H) = \Theta(n)$. The gap could be potentially as big as $\Omega(n)$.

References

[1] Shai Ben-David, Tyler Lu, and Dávid Pál. Does unlabeled data provably help? Worst-case analysis of the sample complexity of semi-supervised learning. In *COLT 2008*, pages 33–44, 2008.

[2] Gyora M. Benedek and Alon Itai. Learnability with respect to fixed distributions. *Theor. Comput. Sci.*, 86(2):377–389, 1991.

[3] Malte Darnstädt, Hans Ulrich Simon, and Balázs Szörényi. Unlabeled data does provably help. In *STACS 2013*, pages 185–196, 2013.

[4] Steve Hanneke. The optimal sample complexity of PAC learning. *J. Mach. Learn. Res.*, 17(38):1–15, 2016.

Thirty-sixth International Conference on Machine Learning (ICML 2019), June 9–15, 2019, Long Beach, California, USA