Pairwise Supervision Can Provably Elicit a Decision Boundary

Han Bao[1,2,*] Takuya Shimada[1,2,*] Liyuan Xu[3] Issei Sato[1] Masashi Sugiyama[2,1] (* Equal contribution)
[1]The University of Tokyo, Japan [2]RIKEN, Japan [3]Gatsby Unit, UCL, UK

Introduction

Similarity in machine learning

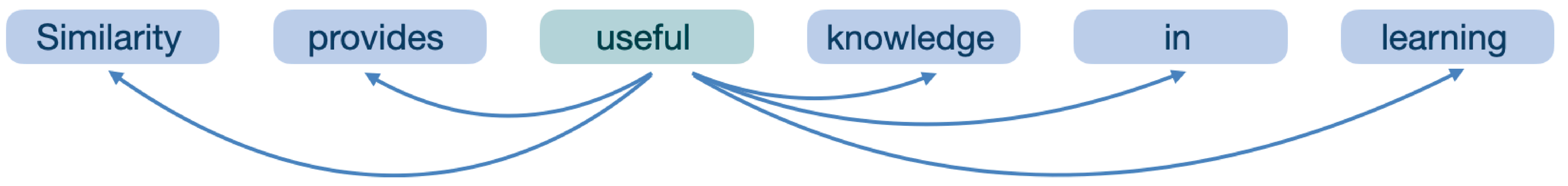


Figure 1: Co-occurring words are often regarded as "similar." [3]

- Pairwise supervision: (dis)similar pairs of patterns
- Available in many domains such as geographical analysis, chemical experiments, computer vision, natural language processing, etc.

Similarity learning in machine learning

A learning paradigm that builds a pairwise model to predict whether pairs are similar or dissimilar (in the classes that they belong to), including

- **metric learning** [5]: learn a distance metric function $d(\boldsymbol{x},\boldsymbol{x}')$ that assigns smaller/larger values to similar/dissimilar pairs
- **kernel learning** [2]: learn a kernel function $k(\boldsymbol{x},\boldsymbol{x}')$ that aligns well with yy' (product of binary labels $y,y' \in \{\pm 1\}$)

Q. Why do we need similarity learning?

Because similarity learning is *expected* to help improving downstream tasks, e.g., classification and clustering.

Can we quantitatively demonstrate the benefit of similarity learning?

Setup

Downstream task: binary classification

The label set is $\mathscr{Y} = \{\pm 1\}$. Find a classifier $h : \mathscr{X} \to \{\pm 1\}$ that minimizes the *classification risk*:

$$R_{\text{point}}(h) := \mathop{\mathbb{E}}_{(X,Y)\sim p(\boldsymbol{x},y)} [\mathbb{1}\{h(X) \neq Y\}]$$

Formal definition of pairwise supervision

Observe $X = \boldsymbol{x}$ and $X' = \boldsymbol{x}'$ independently first, then pairwise supervision T is drawn from

$$p(T = YY' | \boldsymbol{x}, \boldsymbol{x}') = \begin{cases} \eta_{+1}(\boldsymbol{x})\eta_{+1}(\boldsymbol{x}') + \eta_{-1}(\boldsymbol{x})\eta_{-1}(\boldsymbol{x}') & \text{if } YY' = +1, \\ \eta_{+1}(\boldsymbol{x})\eta_{-1}(\boldsymbol{x}') + \eta_{-1}(\boldsymbol{x})\eta_{+1}(\boldsymbol{x}') & \text{if } YY' = -1, \end{cases}$$

where $\eta_{\pm}(\boldsymbol{x}) := p(Y = \pm 1 | X = \boldsymbol{x})$.

Our formulation of similarity learning

CIPS: Classifier with Inner-Product Similarity

Find a minimizer $h : \mathscr{X} \to \{\pm 1\}$ of the *pairwise classification error*:

$$R_{\text{pair}}(h) := \mathop{\mathbb{E}}_{\substack{X,X'\sim p(\boldsymbol{x}) \\ T\sim p(T=YY'|\boldsymbol{x},\boldsymbol{x}')}} [\mathbb{1}\{h(X)\cdot h(X') \neq T\}].$$

Main theorem

Let $R_{\text{clus}}(h) := \min\{R_{\text{point}}(h), R_{\text{point}}(-h)\}$. For any classifier h, $0 \leq R_{\text{pair}}(h) \leq \frac{1}{2}$, and

$$R_{\text{clus}}(h) = \frac{1}{2} - \frac{\sqrt{1 - 2R_{\text{pair}}(h)}}{2}.$$

Corollary: $R_{\text{clus}}(h_1) < R_{\text{clus}}(h_2) \iff R_{\text{pair}}(h_1) < R_{\text{pair}}(h_2) \quad \forall h_1, h_2$

$\Longrightarrow$ Minimizer of R_{pair} is the optimal classifier up to label permutation

Proposed method

Step 1: given $\{(\boldsymbol{x}_i, \boldsymbol{x}_i', \tau_i = y_i y_i')\}_i$, obtain $h = \arg\min_h \widehat{R}_{\text{pair}}(h)$

Step 2: given $\{(\boldsymbol{x}_i, y_i)\}_i$, obtain $s = \arg\min_{s\in\{\pm 1\}} \widehat{R}_{\text{point}}(sh)$

$\Longrightarrow$ sh is the optimal binary classifier

Remark: We proposed another estimator of s *computable with only pairwise supervision* in the paper (Theorem 2).

Existing formulations

SLLC [1]

Step 1: given $\{(\boldsymbol{x}_i, y_i)\}_i$, learn similarity $K(\boldsymbol{x},\boldsymbol{x}') = \boldsymbol{x}^\top A \boldsymbol{x}'$ by

$$\min_A \frac{1}{n}\sum_{i=1}^{n} \ell\left(\underbrace{y_i}_{\text{true label}}, \underbrace{\frac{1}{n'}\sum_{l=1}^{n'} y_l' K(\boldsymbol{x}_i, \boldsymbol{x}_l')}_{\text{aggregated label for } \boldsymbol{x}_i} \right)$$

Step 2: given $\{(\boldsymbol{x}_i, y_i)\}_i$, learn a kernel classifier with kernel K

Drawback: Step 2 requires the usual sample complexity $O_p(n^{-1/2})$

SD [4]

given $\{(\boldsymbol{x}_i, \boldsymbol{x}_i', \tau_i = y_i y_i')\}_i$, learn a classifier by minimizing an unbiased estimator of $R_{\text{point}}(h)$ (computable with only pairwise supervision)

Advantage: no need of Step 2

Drawback: the unbiased estimator is undefined at $p(Y = 1) = \frac{1}{2}$

Comparison of formulations

Table 1: n indicates the number of paired data in Step 1, and the number of pointwise data in Step 2.

		Sample complexity of	
	$p(Y=1) = \frac{1}{2}$	Step 1	Step 2
CIPS	✓	$O_p(n^{-\frac{1}{4}})$	$O_p(e^{-n})$
SLLC [1]	✓	$O_p(n^{-\frac{1}{4}})$	$O_p(n^{-\frac{1}{2}})$
SD [4]	undefined	$O_p(n^{-\frac{1}{2}})$	(unnecessary)

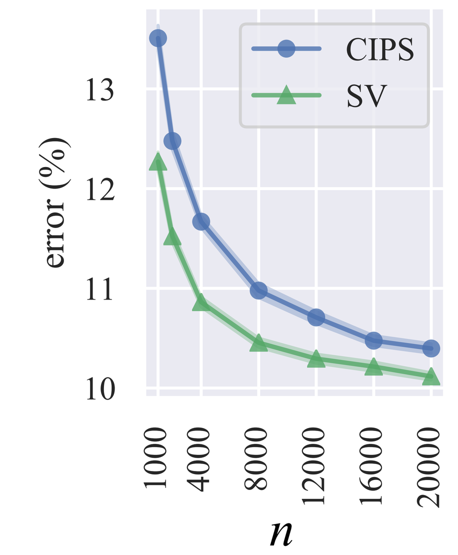
CIPS works even with $p(Y = 1) = \frac{1}{2}$ and its Step 2 is very cheap!

Experiment

We validate that binary classification is possible with R_{pair} (Step 1).

Setup: MNIST odd/even classification

- Model: $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$
- Loss: logistic loss
- Optimizer: SGD (learning rate: 10^{-2})
- Evaluation: R_{clus}



Baseline: SV (Supervised) with the same setup

Remark: n is the pairwise dataset size for CIPS and the pointwise dataset size for SV

Result: CIPS performs better than theoretically expected; because the sample complexity is $O_p(n^{-1/4})$ with n pairs, CIPS is expected to perform comparably to SV with $O(n^2)$ pairs.

Remark: To connect $f : \mathscr{X} \to \mathbb{R}$ to $h : \mathscr{X} \to \{\pm 1\}$, justification by Theorem 3 (in the paper) is needed.

References

[1] A. Bellet, A. Habrard, and M. Sebban.
Good edit similarity learning by loss minimization.
Machine Learning, 89(1-2):5–35, 2012.

[2] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. S. Kandola.
On kernel-target alignment.
In *Advances in Neural Information Processing Systems 15*, pages 367–373, 2002.

[3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean.
Distributed representations of words and phrases and their compositionality.
In *Advances in Neural Information Processing Systems 26*, pages 3111–3119, 2013.

[4] T. Shimada, H. Bao, I. Sato, and M. Sugiyama.
Classification from pairwise similarities/dissimilarities and unlabeled data via empirical risk minimization.
Neural Computation, 33(5):1234–1268, 2021.

[5] E. P. Xing, M. I. Jordan, S. J. Russell, and A. Y. Ng.
Distance metric learning with application to clustering with side-information.
In *Advances in Neural Information Processing Systems 16*, pages 521–528, 2003.