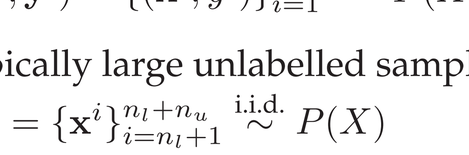
Semi-Supervised Learning, Causality, and the Conditional Cluster Assumption

Julius von Kügelgen, Alexander Mey, Marco Loog, Bernhard Schölkopf

TU Delft · Max Planck Institute for Intelligent Systems · University of Cambridge

1. Motivation

Some ML tasks involve predicting a target Y from both its causes X_C and its effects X_E:



Example: Predict disease Y from medical data.

- $X_C = \{$age, sex, diet, medical family history, genetic factors, smoking, ...$\}$
- $X_E = \{$clinical symptoms, imaging results, serum tests, tissue samples, ...$\}$

Research question: What implications does such causal knowledge of the underlying problem have for semi-supervised learning (SSL)?

2. Semi-Supervised Learning

Given:

- typically small labelled sample $(\mathbf{X}^l, \mathbf{y}^l) = \{(\mathbf{x}^i, y^i)\}_{i=1}^{n_l} \overset{\text{i.i.d.}}{\sim} P(X, Y)$
- typically large unlabelled sample $\mathbf{X}^u = \{\mathbf{x}^i\}_{i=n_l+1}^{n_l+n_u} \overset{\text{i.i.d.}}{\sim} P(X)$

Goal: improve estimate of $P(Y|X)$ from additional information about $P(X)$.

Approaches: link $P(Y|X)$ and $P(X)$ through additional assumptions [1], e.g.:

- "*points in the same cluster of $P(X)$ have the same label Y*" (cluster assumption)
- "*class boundaries of $P(Y|X)$ lie in an area where $P(X)$ is small*" (low-density separation assumption)

4. (Anti-)Causal Learning

When predicting Y from X, [4] distinguish:

Causal learning: $X \to Y$

- $P(X)$ and $P(Y|X)$ are independent causal mechanisms $\to$ no link
- SSL therefore impossible

Anticausal learning: $Y \to X$

- $P(Y)$ and $P(X|Y)$ are algorithmically independent, but $P(X)$ and $P(Y|X)$ may share information
- SSL (in principle) possible

3. Principle of Independent Causal Mechanisms

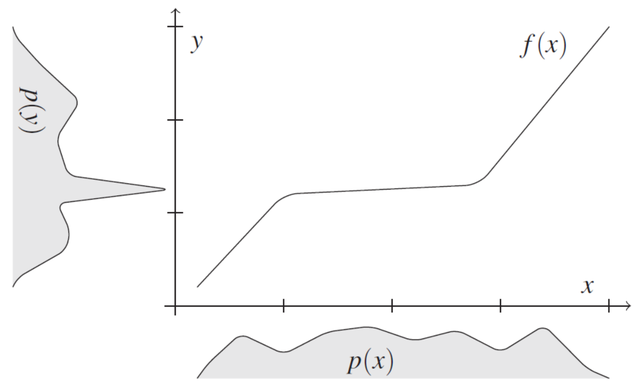
Causal factorisation: Joint distribution factorises over the underlying causal graph,

$$P(X_1, \ldots, X_d) = \prod_{i=1}^{d} P(X_i|\mathbf{PA}_i).$$

Principle of independent causal mechanisms: $P(X_i|\mathbf{PA}_i)$ are algorithmically independent modules which do not share information [3].

Algorithmic independence: Jointly encoding distributions does not admit a shorter description than describing each separately.

Independence is generally violated for non-causal factorisation (backwards model):



IGCI model for $X \to Y$ from [2]

5. SSL with Cause and Effect Features

Analogous to causal learning, $P(X_C)$ contains no information about the rest of the system.

This leaves $P(Y, X_E|X_C)$ which admits two possible factorisations:

$$P(Y, X_E|X_C) = P(Y|X_C)P(X_E|X_C, Y) \quad \text{[causal factorisation} \to \text{independent mechanisms]}$$

$$P(Y, X_E|X_C) = \underbrace{P(X_E|X_C)}_{\substack{\text{estimable from} \\ \text{unlabelled data}}} \underbrace{P(Y|X_C, X_E)}_{\substack{\text{target} \\ \text{quantity}}} \quad \text{[non-causal factorisation} \to \text{shared information]}$$

Main insight: $P(X_E|X_C)$ contains all relevant information provided by unlabelled data $(\mathbf{x}_C, \mathbf{x}_E)$ about $P(Y|X_C, X_E) \implies$ SSL should link $P(X_E|X_C)$ and $P(Y|X_C, X_E)$ via suitable assumptions.

Remark: This includes (im)possibility results for causal and anticausal learning as special cases.

6. The Conditional Cluster Assumption

Conditional cluster assumption: "*Points in the same cluster of $P(X_E|X_C)$ share the same label Y.*"

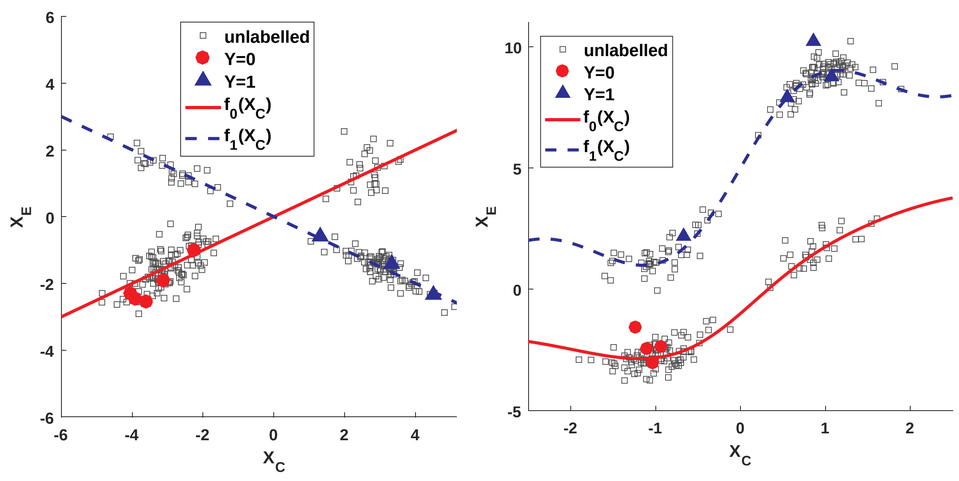
$\to$ Think of clusters of $P(X_E|X_C)$ as clusters in the space of functions $f : \mathcal{X}_C \to \mathcal{X}_E$.

Example: Binary classification with additive noise model for X_E:

$$Y := \mathbb{I}\{g(X_C) > U\}$$

$$X_E := \begin{cases} f_0(X_C) + \sigma_0 N_E, & (Y = 0) \\ f_1(X_C) + \sigma_1 N_E, & (Y = 1) \end{cases}$$

where $U \sim \mathcal{U}[0, 1]$, $N_E \sim \mathcal{N}(0, 1)$.



7. Algorithms & Experimental Results

Semi-generative model [5] + EM:

1. Initialise $p(Y, X_E|X_C; \boldsymbol{\theta})$ from labelled data.
2. E-Step: $\boldsymbol{y}^u = p(\mathbf{y}|\mathbf{X}_C^u, \mathbf{X}_E^u; \hat{\boldsymbol{\theta}})$
3. M-Step:

$$\hat{\boldsymbol{\theta}}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{y}^l, \mathbf{y}^u, \mathbf{X}_E^l, \mathbf{X}_E^u|\mathbf{X}_C^l, \mathbf{X}_C^u; \boldsymbol{\theta})$$

4. Repeat 2. and 3. until convergence.

Conditional self-learning:

1. Initialise f_0 and f_1 from labelled data by regressing $\mathbf{X}_{E,0}^l$ on $\mathbf{X}_{C,0}^l$ and $\mathbf{X}_{E,1}^l$ on $\mathbf{X}_{C,1}^l$.
2. Compute prediction errors of f_0 and f_1 on unlabelled data, $\mathbf{e}_i = ||\mathbf{X}_E^u - f_i(\mathbf{X}_C^u)||^2$.
3. Label the data point with smallest prediction error and retrain the corresponding f_i.
4. Repeat 2. and 3. until all points are labelled.

Method	S1 (lin)	S2 (non-lin)	S3 (4-dim)	Diabetes	Heart
Lin. log. reg.	.968 ± .023	.823 ± .080	.945 ± .039	.626 ± .058	.526 ± .066
Lin. T-SVM	.865 ± .093	.878 ± .074	.822 ± .117	.602 ± .065	**.746 ± .060**
RBF T-SVM	.863 ± .094	.876 ± .075	.821 ± .116	.601 ± .064	.745 ± .060
RBF label propag.	.924 ± .082	.909 ± .065	-	.650 ± .030	.528 ± .068
Semi-gen. (sup.)	.968 ± .076	.935 ± .074	.949 ± .082	.669 ± .064	.550 ± .096
Semi-gen.+soft EM	**.986 ± .081**	**.989 ± .024**	**.991 ± .067**	.661 ± .063	.518 ± .050
Semi-gen.+hard EM	.985 ± .079	.972 ± .058	.987 ± .076	**.695 ± .064**	.518 ± .050
Cond. self-learning	.980 ± .052	.923 ± .090	.961 ± .069	.659 ± .079	.719 ± .076

References

[1] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. The MIT Press, 1st edition, 2010.

[2] D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

[3] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

[4] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[5] J. von Kügelgen, A. Mey, and M. Loog. Semi-generative modelling: Covariate-shift adaptation with cause and effect features. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1361–1369, 2019.