Causal and Anti-causal Structure for Semi-supervised Learning

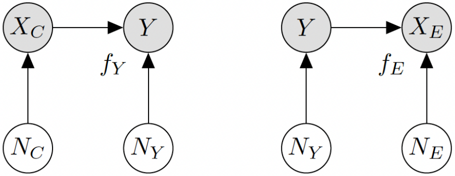
Yu (Zoey) Zhu[1]

University of California, Santa Cruz[1]

Oct 31, 2022

Causal Mechanisms

- Consider the input X and output Y
  - Causal learning: predict effect Y from cause X
  - Anticausal learning: predict cause Y from effect X



(a) Causal learning (b) Anticausal learning

- f is the deterministic mechanism
- N_C and N_Y (or N_E and N_Y) are independent noise variables

Causal learning:

$$X_C := N_C \tag{1}$$
$$Y := f_Y(X_C, N_Y) \tag{2}$$

Anticausal learning:

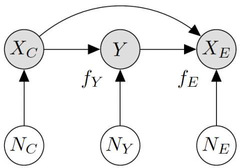
$$Y := N_Y \tag{3}$$
$$X_E := f_E(Y, N_E) \tag{4}$$

Causal Mechanism Effects on Semi-Supervised Learning (SSL)

- Training data:
  - Labeled sample: $(X^l, Y^l) = \{(x^i, y^i)\}_{i=1}^{n_l}$
  - Unlabeled sample: $X^u = \{x^i\}_{i=n_l+1}^{n_l+n_u}$
  - From same distribution P (MCAR)
- Goal: Estimate $P(Y|X)$
- SSL question: If those extra unlabeled $\{X^u\}$ can improve estimating $P(Y|X)$?
  - $\{X^u\}$ can improve the estimate of $P(X)$
  - Thus, $P(Y|X)$ can be improved if we have a link between $P(X)$ and $P(Y|X)$
  - Two common assumptions: cluster assumption; low-density separation
- Based on the Independent Causal Mechanisms (ICM) principle, $P(X_C)$ and $P(Y|X_C)$ are algorithmically independent, so intuitively:
  - Causal learning: SSL doesn't work since $P(X_C)$ contains no information about $P(Y|X_C)$
  - Anticausal learning: $P(X_E)$ may contain information about $P(Y|X_E)$

SSL with Cause and Effects Features

- Consider including both of the cause and effects features:
  - Labeled sample: $(X_C^l, Y^l, X_E^l) = \{(x_c^i, y^i, x_e^i)\}_{i=1}^{n_l}$
  - Unlabeled sample: $(X_C^u, X_E^u) = \{(x_c^i, x_e^i)\}_{i=n_l+1}^{n_l+n_u}$



$$X_C := N_C \tag{5}$$
$$Y := f_Y(X_C, N_Y) \tag{6}$$
$$X_E := f_E(Y, X_C, N_E) \tag{7}$$

- Goal: Estimate $P(Y|X_C, X_E)$ with additional information of $P(X_C, X_E)$ from unlabeled sample
- Subject to ICM, opposite to $P(X_C)$, $P(X_E|X_C)$ contains all relevant information about $P(Y|X_C, X_E)$ provided by the unlabeled sample
- Refined assumption: SSL should exploit links between two conditional distributions $P(X_E|X_C)$ and $P(Y|X_C, X_E)$ rather than the joint feature set $P(X_E, X_C)$ and $P(Y|X_C, X_E)$

SSL with Cause and Effects Features

- Assume binary classification:

$$Y := \mathbb{1}\{g(X_C) > U\} \tag{8}$$

$$X_E := Y f_1(X_C, N_E) + (1 - Y) f_0(X_C, N_E) \tag{9}$$

  - Allow arbitrary g, f_0, f_1 and N_E, without loss of generality
- Reformulate classical SSL assumptions:
  - Conditional cluster assumption: points in the same cluster of $p(X_E|X_C)$ share the same label Y
  - Low-conditional-density separation: class boundaries of $P(Y|X_C, X_E)$ should lie in regions where $P(X_E|X_C)$ is small

Algorithm 1: Semi-generative Model

- Only model the informative part of the generative process: $P(Y, X_E|X_C)$

$$argmax_{\theta} p(y^l, X_E^l|X_C^l;\theta) \sum_{y^u} p(y^u, X_E^u|X_C^u;\theta) \tag{10}$$

  where $\theta = (\theta_Y, \theta_E)$

- Minimize the negative log-likelihood (NLL) which for fixed labels decomposes into two separate terms optimized independently for θ_Y and θ_E

$$NLL(\theta|X_C, y, X_E) := -logp(y, X_E|X_C;\theta) \tag{11}$$

$$= -logp(y|X_C;\theta_Y) - logp(X_E|y, X_C;\theta_E) \tag{12}$$

Algorithm 1: Semi-generative Model

Algorithm 1: EM-like algorithm for fitting a semi-generative model by maximum likelihood

Input: labelled data $(\mathbf{X}_C^l, \mathbf{y}^l, \mathbf{X}_E^l)$; unlabelled data $(\mathbf{X}_C^u, \mathbf{X}_E^u)$; parametric models $p(y|\mathbf{x}_C; \boldsymbol{\theta}_Y)$ and $p(\mathbf{x}_E|\mathbf{x}_C, y; \boldsymbol{\theta}_E)$

Output: fitted labels $\mathbf{y}^u$; estimates $\hat{\boldsymbol{\theta}}_Y, \hat{\boldsymbol{\theta}}_E$

$t \leftarrow 0$

$\hat{\boldsymbol{\theta}}_Y^{(0)} \leftarrow \arg\min \mathrm{NLL}(\boldsymbol{\theta}_Y|\mathbf{y}^l)$

$\hat{\boldsymbol{\theta}}_E^{(0)} \leftarrow \arg\min \mathrm{NLL}(\boldsymbol{\theta}_E|\mathbf{y}^l)$

while *not converged* **do**

- $\mathbf{y}^{(t)} \leftarrow \mathbb{I}\{p(\mathbf{y}|\mathbf{X}_C^u, \mathbf{X}_E^u; \boldsymbol{\theta}_Y^{(t)}, \boldsymbol{\theta}_E^{(t)}) > 0.5\}$
- $\hat{\boldsymbol{\theta}}_Y^{(t+1)} \leftarrow \arg\min \mathrm{NLL}(\boldsymbol{\theta}_Y|\mathbf{y}^l, \mathbf{y}^{(t)})$
- $\hat{\boldsymbol{\theta}}_E^{(t+1)} \leftarrow \arg\min \mathrm{NLL}(\boldsymbol{\theta}_E|\mathbf{y}^l, \mathbf{y}^{(t)})$
- $t \leftarrow t + 1$

end

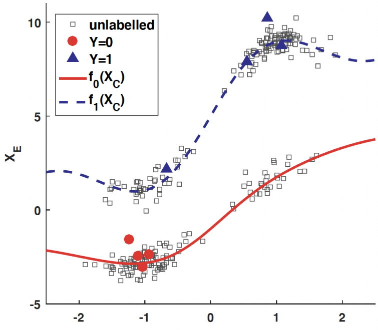
return $\mathbf{y}^{(t-1)}, \boldsymbol{\theta}_Y^{(t)}, \boldsymbol{\theta}_E^{(t)}$

Algorithm 2: Conditional Self-learning

- Extract information from $P(X_E|X_C)$ instead of propagating labels based on similarities between points in the joint feature space (X_C, X_E).
- Assume an additive noise model:

$$f_i(X_C, N_E) = f_i(X_C) + N_{E,i}, i = 0, 1 \tag{13}$$

- Assumption for the noise: mean zero and unimodal. Ensure the one-to-one function from X_C to X_E for each label.
- Learn functions $\hat{f}_0$ and $\hat{f}_1$.



Algorithm 2: Conditional Self-learning

Algorithm 2: Conditional self-learning

Input: labelled data $(\mathbf{X}_C^l, \mathbf{y}^l, \mathbf{X}_E^l)$; unlabelled data;
$(\mathbf{X}_C^u, \mathbf{X}_E^u)$; `regress()` method

Output: fitted labels $\mathbf{y}^u$; functions $\hat{f}_0, \hat{f}_1$

$t \leftarrow 0$
while *unlabelled data left* **do**
  for $i = 0, 1$ **do**
    $\hat{f}_i^{(t)} \leftarrow \texttt{regress}(\mathbf{X}_{E,i}^l, \mathbf{X}_{C,i}^l)$
    $\mathbf{r}_i \leftarrow ||\mathbf{X}_E^u - \hat{f}_i^{(t)}(\mathbf{X}_C^u)||^2$
  end
  $(i, j) \leftarrow \arg\min\{\mathbf{r}_{i,j} : i = 0, 1; j = 1, ..., n_u\}$
  $y^{n_l+j} \leftarrow i$
  $\mathbf{X}_{E,i}^l, \mathbf{X}_{C,i}^l \leftarrow \texttt{append}(\mathbf{x}_E^{n_l+j}, \mathbf{x}_C^{n_l+j})$
  $t \leftarrow t + 1$
end
return $\mathbf{y}^u, \hat{f}_0^{(t-1)}, \hat{f}_1^{(t-1)}$

Simulation Study

- Compare the semi-generative model and conditional self-learning with the baseline methods:
  - Supervised Logistic Regression (SLR): only use the labeled data and ignore the causal structure.
  - Transductive Support Vector Machine (T-SVM) with linear and RBF kernels: conventional SSL methods.
- Simulate the data in the following three different cases: Draw $X_C \in \mathbb{R}^{d_C}$ from a mixture of m d_C-dimensional Gaussian.

$$Y := \mathbb{1}\{\sigma(a'X_C + b) > N_Y\} \tag{14}$$

$$X_E := \{f_i(X_C) + D_i N_E\}\mathbb{1}\{Y = i\} \tag{15}$$

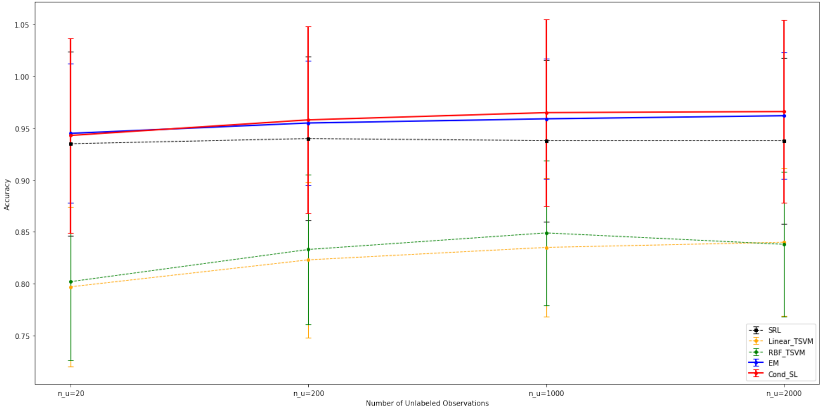
  with $N_Y \sim U[0,1]$, $N_E \sim N_{d_E}(0, I)$, $a \in \mathbb{R}^{d_C}$, $D_i \in \mathbb{R}^{d_E \times d_E}$, $\sigma(x)$ is the sigmoid function, $i = 0, 1$.
  - Case I: Linear additive noise model with one-dimensional feature dimensions
  - Case II: Linear additive noise model with high dimensional feature dimensions
  - Case III: Non-linear additive noise model with high dimensional feature dimensions
- Generate the data in each case with 10 labeled data and 20, 200, 1000 and 2000 unlabeled data (with the increase of missing proportions).

Case I

$$f_i(X_C) = {A_i}' X_C + b_i \tag{16}$$

- feature dimensions: $d_C = d_E = 1$
- $X_C \sim \sum_{m=1}^{3} \omega_m N(\mu_{C_m}, \sigma^2_{C_m})$: $\omega = [0.3, 0.4, 0.3]$, $\mu_C = [-5, 0, 5]$, $\sigma_C = [0.5, 0.5, 0.5]$
- Y: $a = 0.5$, $b = 0$
- X_E: $A_0 = -A_1 = 1$, $b_0 = -b_1 = 2$, $D_0 = D_1 = 0.25$
- Use the linear regression for $P(X_E|X_C, y; \theta_E)$ in the semi-generative model and $P(X_E^l|X_C^l)$ in the conditional self-learning.

Case I



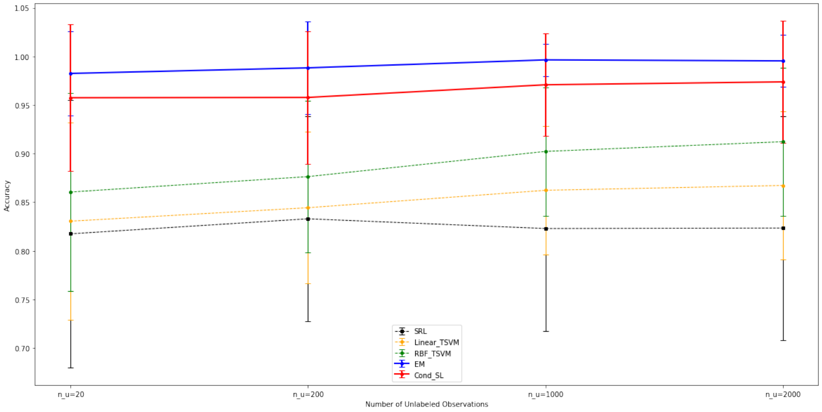
- Causally-motivated methods outperform the other three and have no significant difference.
- Two T-SVM methods provide similar results and perform even worse than the supervised logistic regression.

Case II

$$f_i(X_C) = A_i{}' X_C + b_i \tag{17}$$

- feature dimensions: $d_C = d_E = 10$
- $X_C \sim \sum_{m=1}^{2} \omega_m MVN_{d_C}(\mu_{C_m}, \Sigma_{C_m})$: $\omega = [0.5, 0.5]$, $\mu_{C_1} = [-\frac{1}{2}, -\frac{2}{2}, ..., -\frac{d_C}{2}]$, $\mu_{C_2} = [\frac{1}{2}, \frac{2}{2}, ..., \frac{d_C}{2}]$, $\Sigma_{C_1} = \Sigma_{C_2} = diag_{d_C}(0.5)$
- Y: $a = 0.5$, $b = 0$
- X_E: $A_0 = -A_1 = 0.5J_{10}$, $b_0 = b_1 = 0$, $D_0 = D_1 = diag_{d_E}(0.25)$
- Use the ridge regression with penalty $\lambda = 1$ for $P(X_E|X_C, y; \theta_E)$ in the semi-generative model and $P(X_E^l|X_C^l)$ in the conditional self-learning.

Case II



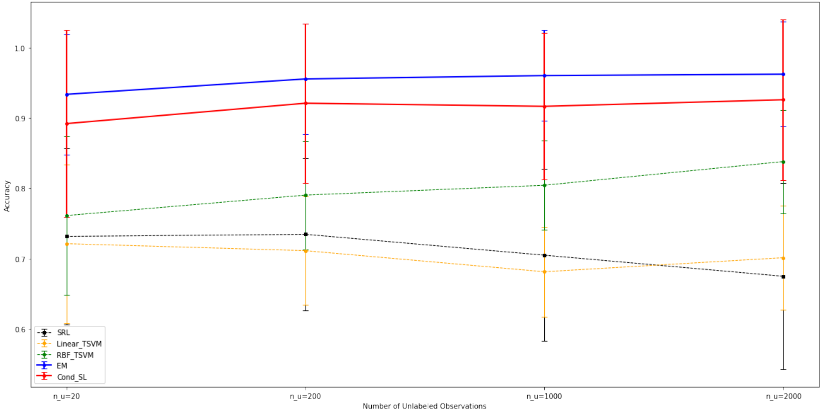
- Causally-motivated methods still outperform the other three.
- Most of the SSL methods perform better than the low-dimensional Case I, and overwhelm the supervised logistic regression.

Case III

$$f_0(X_C) = A_0{}'X_C + sin(X_C) + b_0 \tag{18}$$
$$f_1(X_C) = A_1{}'X_C + cos(X_C) + b_1 \tag{19}$$

- feature dimensions: $d_C = d_E = 10$
- $X_C \sim \sum_{m=1}^{2} \omega_m MVN_{d_C}(\mu_{C_m}, \Sigma_{C_m})$: $\omega = [0.5, 0.5]$, $\mu_{C_1} = [-\frac{1}{2}, -\frac{2}{2}, ..., -\frac{d_C}{2}]$, $\mu_{C_2} = [\frac{1}{2}, \frac{2}{2}, ..., \frac{d_C}{2}]$, $\Sigma_{C_1} = \Sigma_{C_2} = diag_{d_C}(0.5)$
- Y: $a = 0.5$, $b = 0$
- X_E: $A_0 = -A_1 = 0.5J_{10}$, $b_0 = b_1 = 0$, $D_0 = D_1 = diag_{d_E}(0.25)$
- Use the ridge regression with penalty $\lambda = 1$ for $P(X_E|X_C, y; \theta_E)$ in the semi-generative model and the kernel ridge regression with penalty $\lambda = 1$ for $P(X_E^l|X_C^l)$ in the conditional self-learning.

Case III



- All the methods perform worse compared with Case II.
- EM-like approach still provides stable performance.
- T-SVM with RBF kernel can work better compared with linear T-SVM and supervised logistic regression.

Conclusions

- Exploration of the conditional distribution $X_E|X_C$ instead of the joint distribution of (X_E, X_C) can help improve the classification performance.
- The increase of unlabeled data can slightly contribute to higher accuracy for SSL methods (T-SVM and proposed methods) with both cause and effect features, but also increase the standard deviations.
- Connections to domain adaptation: the proposed approaches are robust to changes in $P(X_C)$. $P(Y|X_C)$ and $P(X_E, Y|X_C)$ remain stable.
- Model flexibility: for structured data such as natural images or text, we can use GANs or VAEs to model the additive noise functions.

Reference

Schölkopf, Bernhard, et al. "On causal and anticausal learning." arXiv preprint arXiv:1206.6471 (2012).

von Kügelgen, J., Mey, A., Loog, M., Schölkopf, B. (2019). Semi-Supervised Learning, Causality and the Conditional Cluster Assumption. arXiv. https://doi.org/10.48550/arXiv.1905.12081

Hoyer, P. O., D. Janzing, J. M. Mooij, J. Peters, and B. Sch¨olkopf (2009). "Nonlinear causal discovery with additive noise models". In: Advances in Neural Information Processing Systems 21, pp. 689– 696.