Towards a Learning Theory of Cause-Effect Inference

David Lopez-Paz[1,2], Krikamol Muandet[1], Bernhard Schölkopf[1], Ilya Tolstikhin[1]

[1]Max Planck Institute for Intelligent Systems, [2]University of Cambridge

Summary

Consider the two generative models:

causal model	anticausal model
$x \sim P$	$y \sim P$
$\epsilon \sim Q$	$\epsilon \sim Q$
$f \sim \mathcal{F}$	$f \sim \mathcal{F}$
$y \leftarrow f(x, \epsilon)$	$x \leftarrow f(y, \epsilon)$

Cause-effect inference is to decide, given samples $S = \{(x_i, y_i)\}_{i=1}^n$, whether:

i) "X causes Y" ($X \to Y$), that is, S was drawn from the **causal model**, or

ii) "Y causes X" ($X \leftarrow Y$), that is, S was drawn from the **anticausal model**.

Previous approaches rely on either

i) expensive high-dimensional conditional dependence tests (Spirtes et al., 2000),

ii) strong parametric assumptions (Hoyer et al., 2009, Daniusis et al., 2012), or

iii) hand-crafted features (Guyon, 2013).

In this paper, we pose causal inference as the problem of learning to classify probability distributions. In particular, we assume access to a collection $\{(S_i, l_i)\}_{i=1}^n$, where each $S_i \sim P^n(X_i, Y_i)$ and l_i is a binary label indicating whether "$X_i \to Y_i$" or "$X_i \leftarrow Y_i$". Given these data, we build a causal inference rule in two steps:

i) we featurize each S_i using the kernel mean embedding associated with some kernel, and

ii) we train a binary classifier on such embeddings to distinguish between causal directions.

Our framework exhibits the following features:

i) theoretical guarantees concerning learning rates and consistency,

ii) theoretically sustained approximations to deal with big-data, and

iii) state-of-the-art performance in a variety of real-world benchmarks.

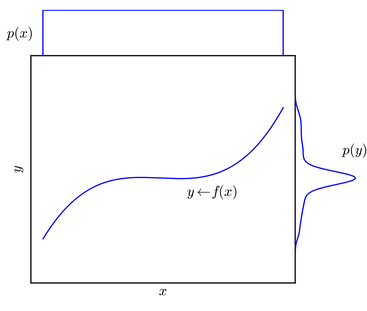
Prior art

SGS/PC (Spirtes et al., 2000)

Idea: given universe of variables $\mathcal{U} = \{X_1, \ldots, X_d\}$, X_i and X_j are causally related (for $i \neq j$) iff:

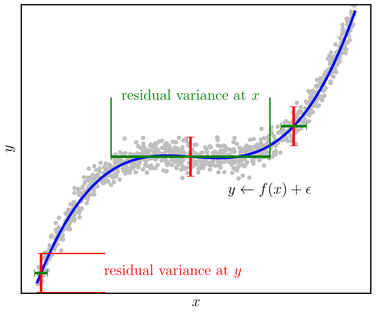
$$\nexists \mathcal{S} \subseteq (\mathcal{U} \setminus \{X_i, X_j\}), \text{ s.t. } X_i \not\perp X_j | \mathcal{S}.$$

IGCI (Daniusis et al., 2012)



"cause independent from mechanism"

ANM (Hoyer et al., 2009)



"cause independent from noise"

Hand-crafted features approach (Guyon, 2013)

i) featurize each available labeled causal sample S_i into m features $m(S_i)$, and

ii) train a binary classifier on the data $\{(m(S_i), l_i)\}_{i=1}^n$.

Kernel mean embedding of distributions

The **Kernel Mean Embedding (KME)** of a probability distribution P over $\mathcal{Z}$ associated with a measurable, bounded, and positive-definite kernel k is

$$\mu_k(P) := \int_Z k(z, \cdot) dP(z) \in \mathcal{H}_k.$$

The **Empirical Kernel Mean Embedding (EKME)** estimates $\mu_k(P)$ based on $S \sim P^n$:

$$\mu_k(P_S) := \frac{1}{|S|} \sum_{x \in S} k(x, \cdot) \in \mathcal{H}_k.$$

Problem! For many kernels k, $\mathcal{H}_k$ is an *infinite dimensional* Hilbert Space. This forces us to design learning algorithms which require the construction and inversion of $n \times n$ kernel matrices.

Solution: approximate $\mathcal{H}_k$ with a random *finite-dimensional* subspaces.

Theorem (Bochner): *assume that $\mathcal{Z} = \mathbb{R}^d$ and $k(x, y) = k(x - y)$ is a shift-invariant real-valued kernel. Then Bochner's theorem states that for any $z, z' \in \mathcal{Z}$:*

$$k(z, z') = 2C_k \mathbb{E}_{w,b}\big[\cos(\langle w, z \rangle + b) \cos(\langle w, z' \rangle + b)\big], \tag{1}$$

where $\langle \cdot, \cdot \rangle$ is the Euclidean dot product in $\mathbb{R}^d$, $w \sim \frac{1}{C_k} p_k$, $b \sim \mathcal{U}[0, 2\pi]$, $p_k \colon \mathbb{R}^d \to \mathbb{R}$ is an integrable and positive Fourier transform of k, and $C_k = \int_{\mathbb{R}^d} p_k(w) dw$.

Therefore, sample $\{(w_j, b_j)\}_{j=1}^m$ (Rahimi and Recht, 2007) and approximate the EKME using the **Randomized Empirical Kernel Mean Embedding (REKME)**:

$$\mu_{k,m}(P_S) := \frac{2C_k}{|S|} \sum_{z \in S} \big(\cos(\langle w_j, z \rangle + b_j)\big)_{j=1}^m \in \mathbb{R}^m.$$

Remarks:

i) if k is characteristic, μ_k is an injective map, and

ii) $\|\mu_k(P) - \mu_k(P_S)\|_{\mathcal{H}_k} = O_P(n^{-1/2})$.

iii) $\mu_{k,m}$ is an m-dimensional vector that can be used with any learning algorithm.

iv) Replacing $\mu_k(P_S)$ with $\mu_{k,m}(P_S)$ induces a $O(m^{-1/2})$ error in risk.

v) $\|\mu_k(P) - \mu_k(P_S)\|_{\mathcal{H}_k} \geq C \frac{\sigma_{\mathcal{H}_k}}{\sqrt{n}}$ where $\sigma^2_{\mathcal{H}_k} = \sup_{\|f\|_{\mathcal{H}_k} \leq 1} \mathbb{V}_{z \sim P}[f(z)]$.

Our algorithm

Input

i) labeled causal samples $\{(S_i, l_i)\}_{i=1}^n$; $S_i = \{(x_{ij}, y_{ij})\}_{j=1}^{n_i} \sim P^{n_i}(X_i, Y_i)$, $l_i \in \{-1, +1\}$,

ii) measurable and bounded kernel function k, and

iii) number of random features m.

Training

i) featurize each S_i as $\mu_{k,m}(S_i)$ using (REKME),

ii) train *any* classifier $\hat{f}_n : \mathbb{R}^m \to \{-1, +1\}$ on the data $\{(\mu_{k,m}(S_i), l_i)\}_{i=1}^n$, and

iii) return $\hat{f}_n$.

Testing

i) featurize test sample S_0 as $\mu_{k,m}(S_0)$ as in training, and

ii) return $\hat{f}_n(\mu_{k,m}(S_0))$.

Extensions

i) easy handling of mixed attributes (continuous, discrete, categorical...)

ii) use of third label to take care of the independent "$X_i \perp Y_i$" case

iii) joint embedding of confounder candidates Z and (X_i, Y_i) for multivariate causal inference

iv) μ_k can be designed to be more complex (random forest, deep/convolutional neural network...)

Learning rate and consistency

Assumptions:

i) $\exists$ *Mother distribution* $\mathcal{M}$ on {cause-effect measures P on $\mathcal{Z}$} $\times$ $\{-1, 1\}$,

ii) $\{(P_i, l_i)\}_{i=1}^n \sim \mathcal{M}^n$; with l_i indicating $X_i \to Y_i$ or $X_i \leftarrow Y_i$ for P_i,

iii) training data of the form $S_i = \{(X_{i,j}, Y_{i,j})\}_{j=1}^{n_i} \sim P_i^{n_i}$,

iv) measurable and bounded kernel k with $\sup_{z \in \mathcal{Z}} k(z, z) \leq 1$,

v) class $\mathcal{F}_k$ of functionals mapping $\mathcal{H}_k$ to $\mathbb{R}$ with Lipschitz constants uniformly bounded by $L_{\mathcal{F}}$,

vi) minimization of surrogate risk $R_\varphi(f) := \mathbb{E}_{(P,l) \sim \mathcal{M}} \big[\varphi\big(-f\big(\mu_k(P)\big) l\big)\big]$ in $\mathcal{F}_k$,

vii) $\varphi \colon \mathbb{R} \to \mathbb{R}^+$ is L_φ-Lipschitz s.t. $\varphi(z) \geq \mathbb{1}_{z > 0}$ and $\varphi\big(z\big) \leq B$ for all z.

Theorem:

With probability not less than $1 - \delta$ over all sources of randomness

$$R_\varphi(\tilde{f}_n) - R^*_{\varphi, \mathcal{F}_k} \leq 4 L_\varphi R_n(\mathcal{F}_k) + 2B \sqrt{\frac{\log(2/\delta)}{2n}} + \frac{4 L_\varphi L_{\mathcal{F}}}{n} \sum_{i=1}^n \left(\sqrt{\frac{\mathbb{E}_{z \sim P_i}[k(z, z)]}{n_i}} + \sqrt{\frac{\log(2n/\delta)}{2n_i}} \right).$$

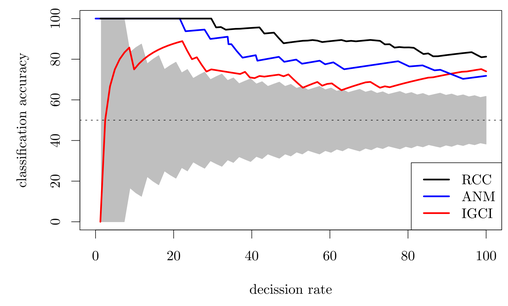
Numerical simulations

We term our method the *Randomized Causation Coefficient* (RCC).

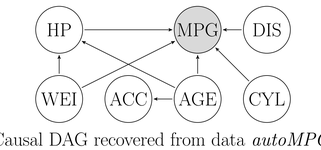
Each causal sample is featurized as $\nu(S) = (\mu_{k,m}(P_{S_x}), \mu_{k,m}(P_{S_y}), \mu_{k,m}(P_{S_{xy}}))$ using a mixture of three Gaussian kernels with respective bandwidths $(0.1\gamma, 1\gamma, 10\gamma)$, where γ is set according to the median heuristic. We use $m = 1,000$ random features, and a random forest as our binary classifier.

We synthesize our training data $\{(S_i, l_i)\}_{i=1}^n$ using a simple generative model detailed in the paper.

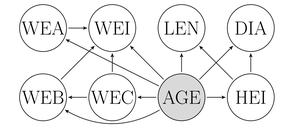
State-of-the-art on Tüebingen real-world cause-effect pairs:



Application to recovery of **multivariate** causal DAGs by jointly embedding potential confounders:



Causal DAG recovered from data *autoMPG*.



Causal DAG recovered from data *abalone*.

RCC **ranked third** in Chalearn's cause-effect inference competition (Guyon, 2014)

More details, experiments, and source available in the paper!

References

P. Daniusis, D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. Inferring deterministic causal relations. *UAI*, 2012.

I. Guyon. Cause-effect pairs kaggle competition, 2013. URL `https://www.kaggle.com/c/cause-effect-pairs/`.

I. Guyon. Chalearn fast causation coefficient challenge, 2014. URL `https://www.codalab.org/competitions/1381`.

P. O. Hoyer, D. Janzing, J. M. Mooij, J. R. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. *NIPS*, 2009.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. *NIPS*, 2007.

P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT Press, 2000.