CSC2541: Introduction to Causality

Lecture 6 - Learning Invariant Predictors

Instructor: Rahul G. Krishnan

TA: Vahid Balazadeh-Meresht

November 14, 2022

Using ideas from causality in machine learning

- **Thus far:** Basics of causal inference
  - Assumptions for causal inference
  - Identification
  - Causal effect estimation from observational data
  - Machine learning algorithms for causal inference (TAR-Net)

- **Remainder of this class:** Advanced topics in causality
  - **This lecture:** How to improve the quality of predictive models using ideas from causality (causality for machine learning)

Out-of-distribution (OOD) generalization

- Supervised machine learning:
  - Training: capture statistical patterns from training data into a model,
  - Prediction: re-use the model to make predictions on new data.
- **Caveat:** This relies on making predictions from models that are independently and identically distributed relative to the training set.
- OOD generalization - build models that work well outside the original data distribution in training.
- **This lecture:** Using ideas from causality to build reliable predictive models with good OOD performance.

Causality for OOD generalization

Why is OOD generalization hard?

- During training, the model may pick up on a sub-set of features,
- These features' relationship to the label can change when the model is used outside of the domain it was trained on.

Why might causality help?

- Causal associations between features and labels are reliable associations (Modularity assumption).
- How can we find the causal parents of a label?

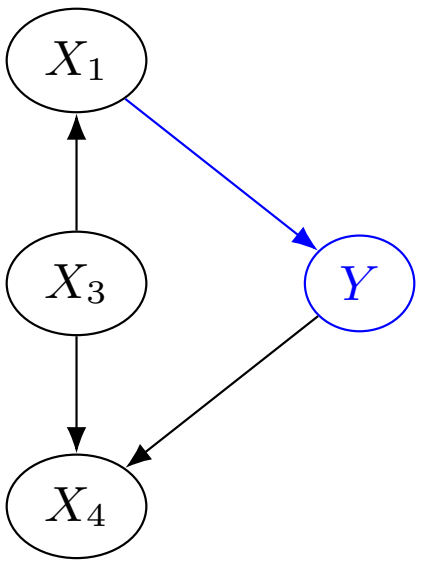
Motivating example

- Let's say we want to learn a predictive model of patient risk (re-admission to the ICU, progression of cancer).
- We have data from four different hospitals:
- **Option 1:** Pool the data together and minimize empirical risk,
- **Option 2:** Use the fact that data came from four different hospitals during learning.
- *Question:* Which of the above two options should we use?

Motivating example - Learning from hospital 1

- Consider predicting Y (patient outcome, originally X_2) from three different features X_1, X_4, X_3.
- Lets take a structural causal modeling view on the problem from hospital 1:

$$X_1 = f_1(X_3, \epsilon_1),\ Y = f_2(X_1, \epsilon_2)$$
$$X_3 = f_3(\epsilon_3),\ X_4 = f_4(X_3, Y, \epsilon_4)$$

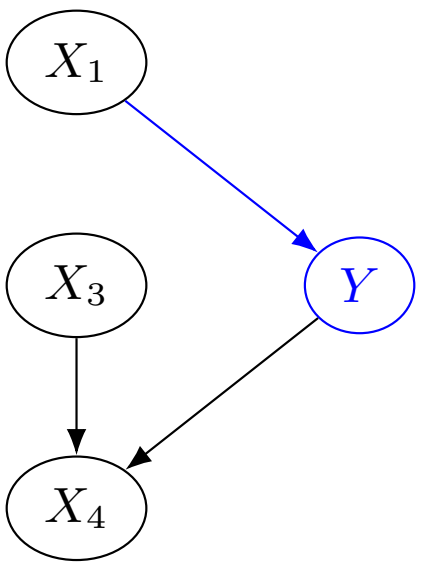


Slides Credit to Jonas Peters: `https://learning.mpi-sws.org/mlss2016/slides/cadizJonas.pdf`

Motivating example - Learning from hospital 2

- In hospital 2, there may be a slightly different relationship between features X_1 and X_3:

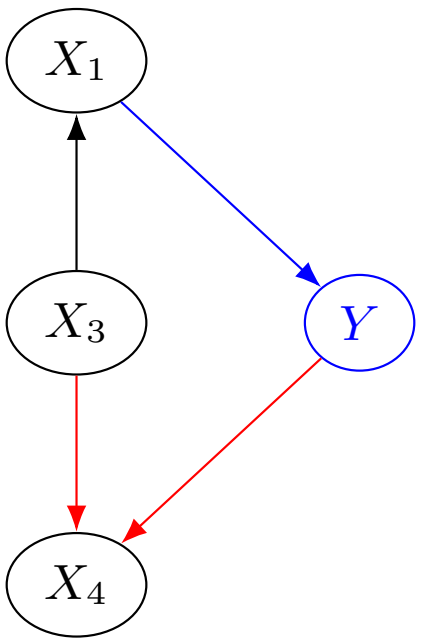
$$X_1 = \tilde{f}_1(\epsilon_1),\ Y = f_2(X_1, \epsilon_2)$$
$$X_3 = f_3(\epsilon_3),\ X_4 = f_4(X_3, Y, \epsilon_4)$$



Motivating example - Learning from hospital 3

- In hospital 3, a different function may control how feature X_4 behaves:

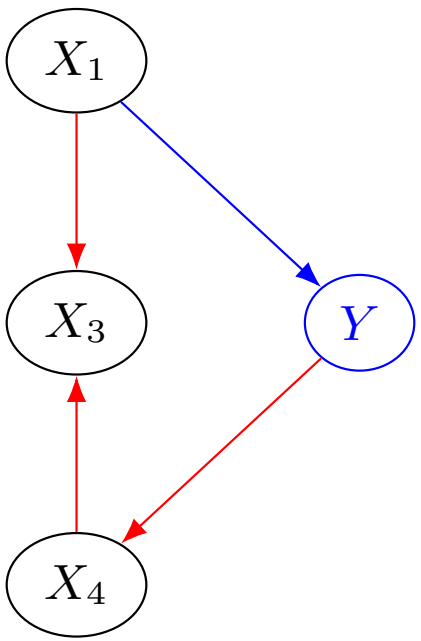
$$X_1 = f_1(\epsilon_1),\ Y = f_2(X_1, \epsilon_2)$$
$$X_3 = f_3(\epsilon_3),\ X_4 = \tilde{f}_4(X_3, Y, \epsilon_4)$$



Motivating example - Learning from hospital 4

- In hospital 4, all three functions in the SCM may differ:

$$X_1 = \tilde{f}_1(\epsilon_1),\ Y = f_2(X_1, \epsilon_2)$$
$$X_3 = \tilde{f}_3(X_1, X_4, \epsilon_3),\ X_4 = \tilde{f}_4(Y, \epsilon_4)$$



Building intuition

- Consider learning a linear or logistic regression model on data from these hospitals.
- *Discussion:* Where should the weights of the regression be highest? Why?
- If we had access to the causal graphs in each environment as domain knowledge, then we could examine it and only learn based on X_1.
- However, this is rarely the case for high-dimensional data. How can we use the available data from multiple hospitals?

Formalizing invariance

- Key insight: $P(Y|pa(Y))$ is invariant as long as the SCM for Y does not change across environments! - **Modularity Assumption**
- If we can extract these features, we can build a *causal* predictive model – more likely to function across environments.
- Instead of learning causal graphs in each environment separately, Invariant Causal Prediction (ICP)[1] says we only need to identify the subset of invariant features in order to learn a causal predictors,
- This task can (under certain assumptions) be easier than general structure learning!

[1] Peters, Bühlmann, and Meinshausen, 2016.

Linear setting - Learning goal

- Set of environments: $\mathcal{E}$
- For each $e \in \mathcal{E}$, we have $X^e \in \mathbb{R}^p, Y^e \in \mathbb{R}$
- For any $S \subseteq \{1, \dots, p\}$, X_S is the vector of $X_k, k \in S$

Assumption - Invariant prediction

There exists $\gamma^* = (\gamma_1^*, \dots, \gamma_p^*)$ with support $S^* := \{k; \gamma_k^* \neq 0\}$ such that for all environments $e \in \mathcal{E}$, X^e has an arbitrary distribution and $Y^e = X^e \gamma^* + \epsilon^e$, where $\epsilon^e \perp\!\!\!\perp X_{S^*}^e$ and $\epsilon^e \sim F_\epsilon$. We say that S^* satisfies invariant prediction.

- Given: Data from different environments,
- Goal: Identify S^*, γ^*.

Strategy

- The assumptions describe the existance of one S^*, γ^*.
- The goal of ICP is to gradually identify both sets.
- We'll use the strategy of performing hypothesis tests over different subsets of index combinations and use (in the linear case) regressions to identify the coefficients.
- The basic algorithm relies on the following test:

1. For any $\gamma \in \mathbb{R}^p$ and resulting choice of S, we define the null hypothesis as

$$H_{0,\gamma,S}(\mathcal{E}) : \gamma_k = 0 \text{ if } k \notin S \text{ and } \begin{cases} \exists F_\epsilon \text{ such that for all } e \in \mathcal{E} \\ Y^e = X^e\gamma + \epsilon^e \text{ with } \epsilon^e \perp\!\!\!\perp X^e_S,\ \epsilon^e \sim F_\epsilon \end{cases}$$

2. Given such a test we can use it iteratively over subset of features to figure out which one are the causal parents of Y.

Plausible and Identifiable causal predictors

- *Any* variables $S \subseteq \{1, \dots, p\}$ are *plausible* predictors under $\mathcal{E}$ if $\exists \gamma \in \mathbb{R}^p$ that $H_{0,\gamma,S}(\mathcal{E})$ is true.
- Then, we'll define *identifiable* predictors as:

$$S(\mathcal{E}) = \bigcap_{S: H_{0,S}(\mathcal{E}) \text{ is true}} S \qquad S = \bigcap_{\gamma \in \Gamma(\mathcal{E})} \{k : \gamma_k \neq 0\}$$

- Enlarging the environment results in a larger set of causal predictors i.e. $S(\mathcal{E}_1) \subseteq S(\mathcal{E}_2)$ for two sets of environments $\mathcal{E}_1 \subseteq \mathcal{E}_2$.

Causal coefficients

- We similarly can define a plausible set of causal coefficients for a set of indices S as $\Gamma_S(\mathcal{E}) = \{\gamma \in \mathbb{R}^p : H_{0,\gamma,S}(\mathcal{E}) \text{ is true}\}$,
- Across all the subsets we have $\Gamma(\mathcal{E}) = \bigcup_{S \subseteq \{1,\dots,p\}} \Gamma_S(\mathcal{E})$.
- Enlarging the environment results in shrinking the coefficients of causal predictors i.e. $\Gamma(\mathcal{E}_1) \supseteq \Gamma(\mathcal{E}_2)$ for two sets of environments $\mathcal{E}_1 \subseteq \mathcal{E}_2$.

Algorithm

1. Given (X^e, Y^e) from a finite number of environments $e \in \mathcal{E}$ where $|X^e| = p$,
2. For each $S \subseteq \{1, \dots, p\}$ test if $H_{0,S}(\mathcal{E})$ holds at level α,
3. Set
$$\hat{S}(\mathcal{E}) = \bigcap_{S: H_{0,S}(\mathcal{E}) \text{ not rejected.}} S$$
4. For obtaining confidence sets:
$$\hat{\Gamma}(\mathcal{E}) = \bigcup_{S \subseteq \{1,\dots,p\}} \hat{\Gamma}_S(\mathcal{E})$$

where

$$\hat{\Gamma}_S(\mathcal{E}) = \begin{cases} \emptyset \text{ if } H_{0,S}(\mathcal{E}) \text{ can be rejected at level } \alpha \\ \hat{C}(S) \text{ otherwise} \end{cases}$$

$\hat{C}(S)$ is a $(1-\alpha)$-confidence set for regressing Y on X_S on the pooled data

Coverage guarantees

Theorem - Coverage guarantees

Assume the estimator $\hat{S}(\mathcal{E})$ is constructed with a valid test for $H_{0,S}(\mathcal{E})$ for all sets $S \subseteq \{1, \dots, p\}$ at level α in the sense that for all S, $\sup_{P: H_{0,S}(\mathcal{E}) \; true} P[H_{0,S}(\mathcal{E}) \; rejected] \leq \alpha$. Then, under the invariant prediction assumption, for any P over (Y, X), the following holds:

$$P[\hat{S}(\mathcal{E}) \subseteq S^*] \geq 1 - \alpha$$

If, moreover, the confidence set $\hat{C}(S)$ satisfies $P[\gamma \in \hat{C}(S)] \geq 1 - \alpha$ for any (γ, S) that satisfy the invariant prediction assumption, then

$$P[\gamma^* \in \hat{\Gamma}(\mathcal{E})] \geq 1 - 2\alpha$$

A concrete hypothesis test

Let's first simplify the null hypothesis. Define

$$\beta^{\text{pred},e}(S) := \text{argmin}_{\beta \in \mathbb{R}^p;\ \text{if } k \notin S} \mathbb{E}(Y^e - X^e\beta)^2$$

Then, we have

$$H_{0,S}(\mathcal{E}) : \begin{cases} \exists \beta \in \mathbb{R}^p \text{ and } \exists F_\epsilon \text{ such that for all } e \in \mathcal{E} \\ \beta^{\text{pred},e}(S) \equiv \beta \text{ and } Y^e = X^e\beta + \epsilon^e \text{ with } \epsilon^e \perp\!\!\!\perp X_S^e,\ \epsilon^e \sim F_\epsilon \end{cases}$$

The goal is to have a test such that

$$P[H_{0,S^*}(\mathcal{E}) \; rejected] \leq \alpha$$

A concrete hypothesis test - Cont.

Hypothesis test for $H_{0,S}(\mathcal{E})$:

- Fit a linear regression model on **all data** to get an estimate $\hat{\beta}^{\text{pred}(S)}$ using set S as the features. Let $R = Y - X\hat{\beta}^{\text{pred}}(S)$.
- Test the null hypothesis that the **mean** of R is identical for all environments: Use a **two-sample t-test** for residuals in environment e against residuals in other environments with Bonferroni correction.
- Test the null hypothesis that the **variance** of R is identical for all environments: Use an **F-test** for residuals in environment e against residuals in other environments with Bonferroni correction.
- Combine the two p-values by taking twice the smaller of the two values.
- If the p-value is smaller than α, reject the set S.

Identifiability results for linear Gaussian SCMs

- We saw coverage guarantees for $\hat{S}(\mathcal{E}) \subseteq S^*$
- But trivial solutions also work here like $\hat{S}(\mathcal{E}) = \emptyset$
- Can we identify S^*? i.e., $\hat{S}(\mathcal{E}) = S^*$?
- We'll give identifiability results for linear Gaussian SCMs, where the observational data (environment 1) is generated from the following (Assume $Y := X_1$):

$$X_j^1 = \sum_{k \neq j} \beta_{j,k}^1 X_k^1 + \epsilon_j^1$$
$$\epsilon_j^1 \sim \mathcal{N}(0, \sigma_j^2)$$

- Data in other environments is created with *do*-interventions, i.e., in the e-th experiment, we intervene on variables $A^e \subseteq \{2, \ldots, p+1\}$ and set them to values $a_j^e \in \mathbb{R}, j \in A^e$

Identifiability results for linear Gaussian SCMs

Theorem - Identification of causal predictors

Consider a linear Gaussian SCM with interventions. Then, all causal predictors are identifiable, i.e.,

$$S(\mathcal{E}) = S^*$$

if the interventions are *do*-interventions with $a_j^e \neq \mathbb{E}(X_j^1)$ and there is at least one single intervention on each variable other than Y, that is for each $j \in \{2, \dots, p+1\}$ there is an experiment with $A^e = \{j\}$.

 Peters, Jonas, Peter Bühlmann, and Nicolai Meinshausen (2016). "Causal inference by using invariant prediction: identification and confidence intervals". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.5, pp. 947–1012.