A brief introduction to Stable learning

Liang Cao

2021.03.29

Zheyan Shen, Peng Cui, Kun Kuang, Bo Li. Causally Regularized Learning on Data with Agnostic Bias. ***ACM*** *, 2018.*

Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Li, Bo Li. Stable Prediction across Unknown Environments. ***KDD***, 2018.

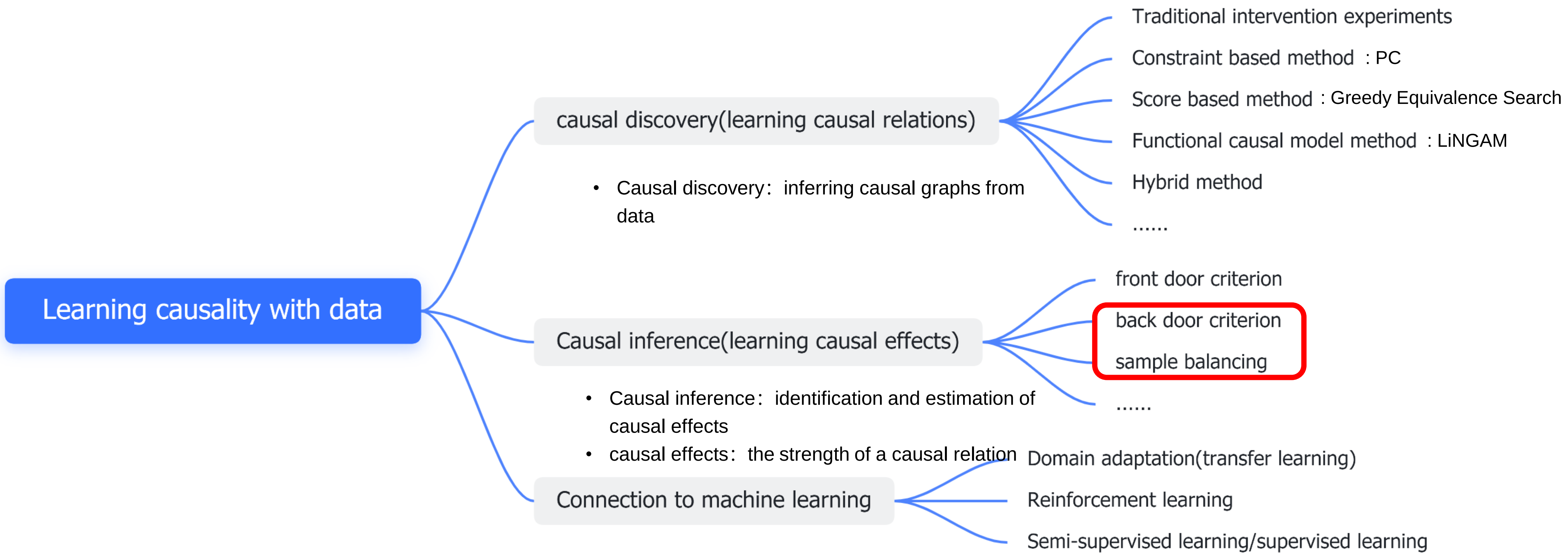
Kuang, K., Xiong, R., Cui. Stable Prediction with Model Misspecification and Agnostic Distribution Shift. ***AAAI**,* 2020

Zheyan Shen, Peng Cui, Tong Zhang, Kun Kunag. Stable Learning via Sample Reweighting. ***AAAI**,* 2020

CONTENT

/01 Introduction

Preliminary Knowledge: Learning Causality with data

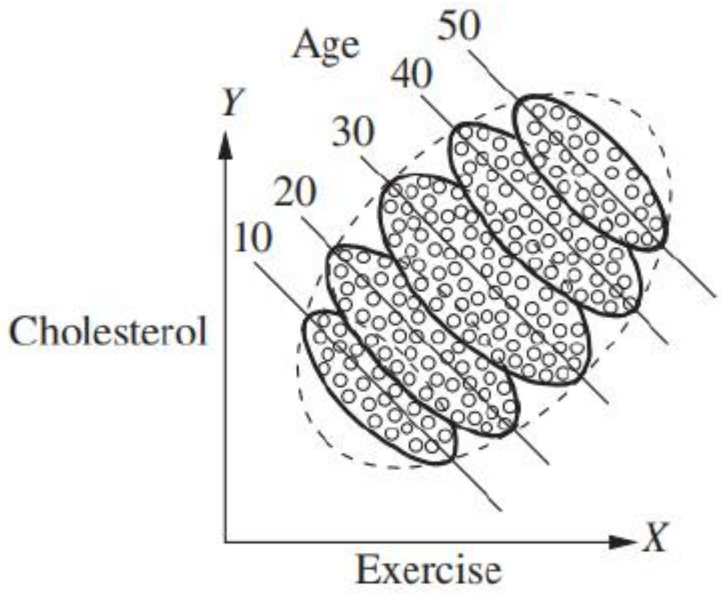


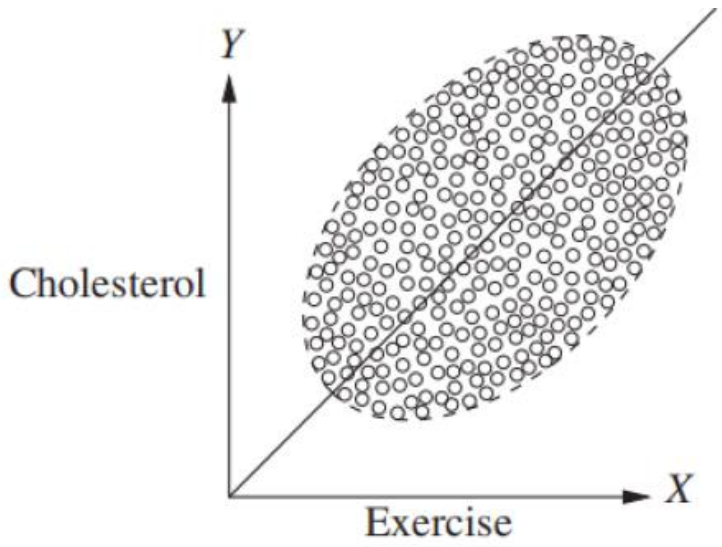
At the Very Beginning: Simpson's Paradox

Example

Consider a study that measures weekly exercise and cholesterol in various age groups.

- There is a general trend downward in each group: the more young people exercise, the lower their cholesterol is, and the same applies for middle-aged people and the elderly.

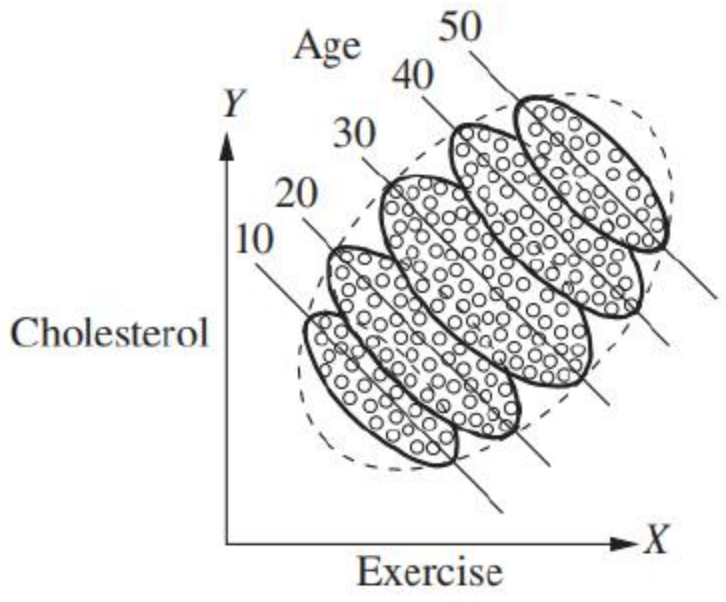


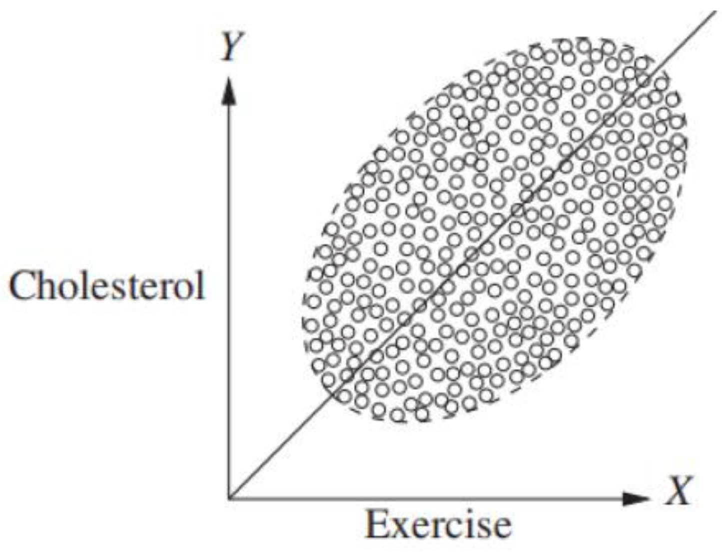


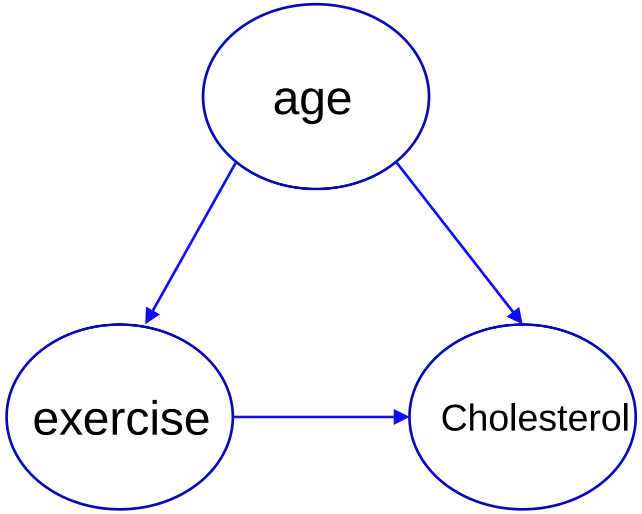
At the Very Beginning : Simpson's Paradox

Fact: Age as Confounding Factor

- Older people are more likely to exercise.
- Older people are also more likely to have high cholesterol regardless of exercise.







Introduction

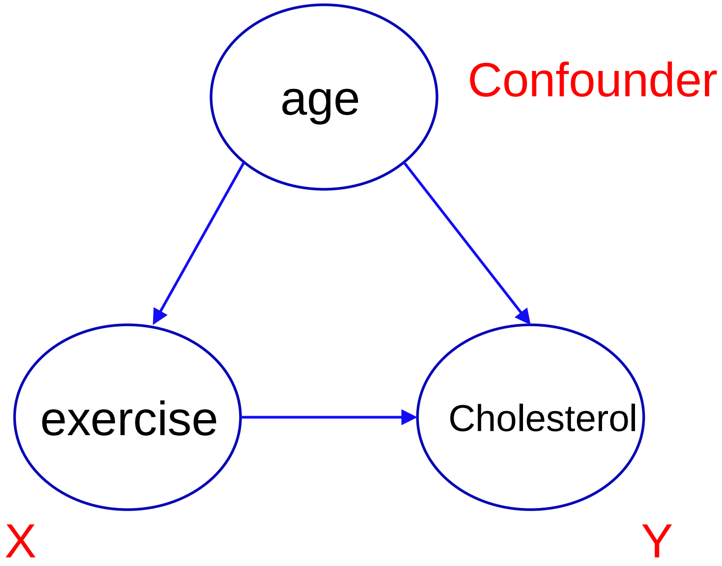
A practical causal definition

- **X is a cause of Y if and only if:**

1. **Change X leads to a change in Y**
2. **Keep everything else constant**

A manipulation/intervention directly changes only the target variable X.

$$\exists x_1 \neq x_2 \; P(Y|do(X=x_1)) \neq P(Y|do(X=x_2))$$



Correlation/dependence/association

- **X and Y are correlated/associated if and only if:**

1. **X changes, Y also changes**

$$\exists x_1 \neq x_2 \; P(Y|X=x_1) \neq P(Y|X=x_2)$$

Introduction

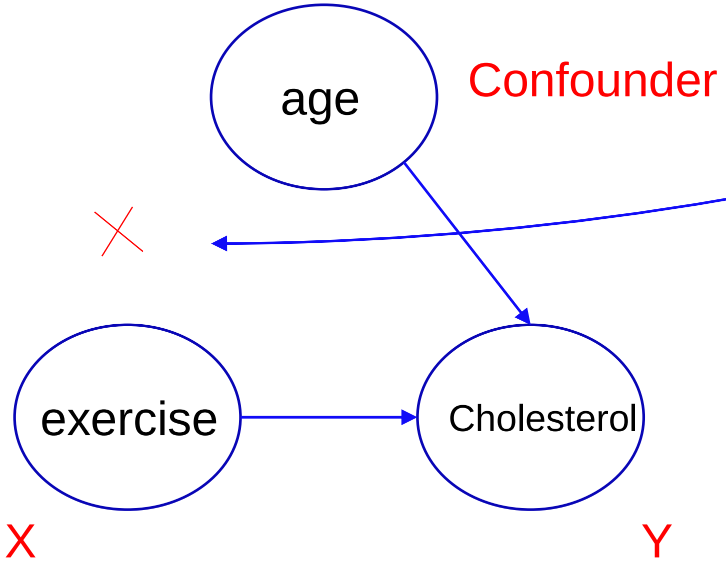
A practical definition

- X is a cause of Y if and only if:

1. Change X leads to a change in Y
2. Keep everything else constant

A manipulation/intervention directly changes only the target variable X.

$$\exists x_1 \neq x_2 \; P(Y|do\,(X{=}x_1)) \neq P(Y|do\,(X{=}x_2))$$



Correlation/dependence/association

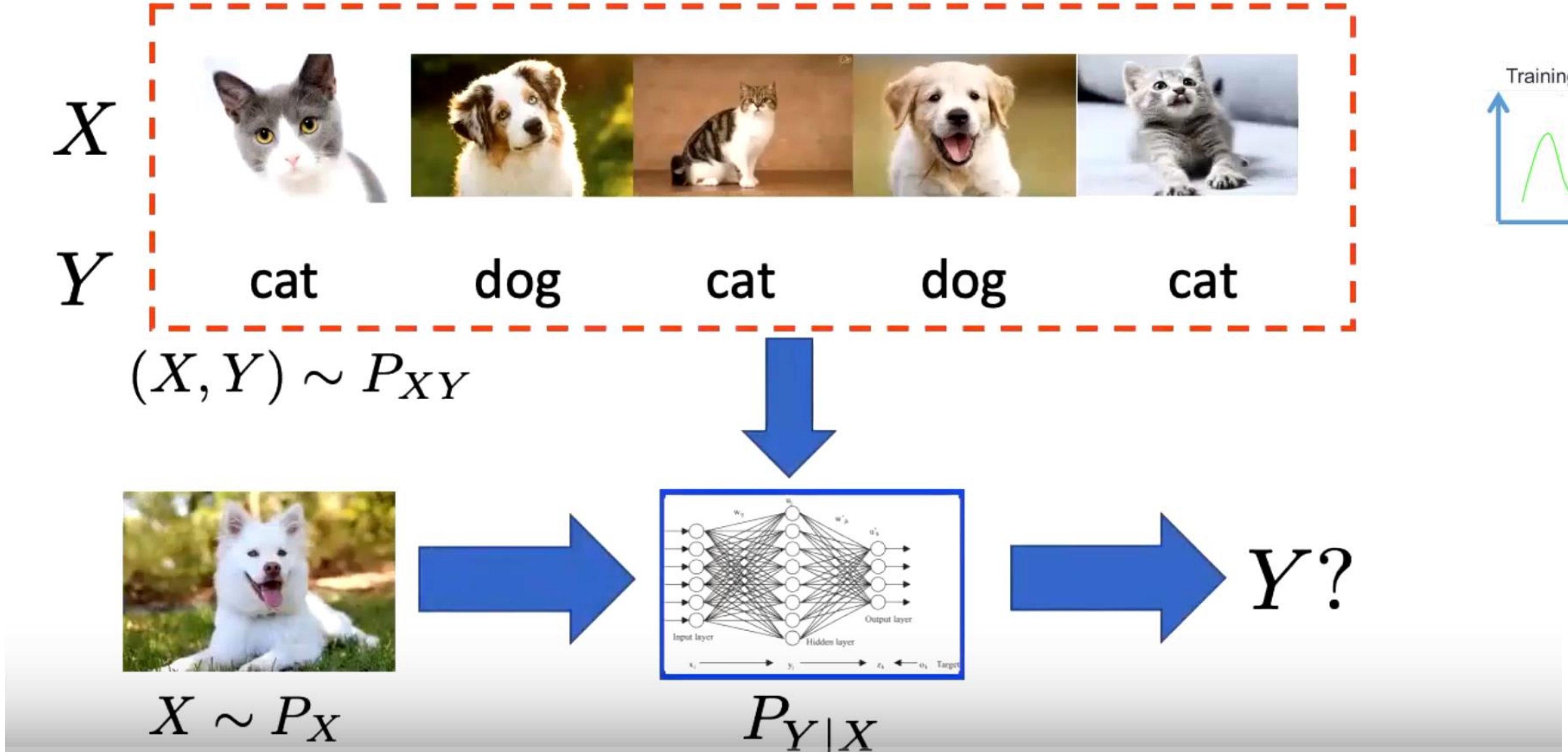
- X and Y are correlated/associated if and only if:

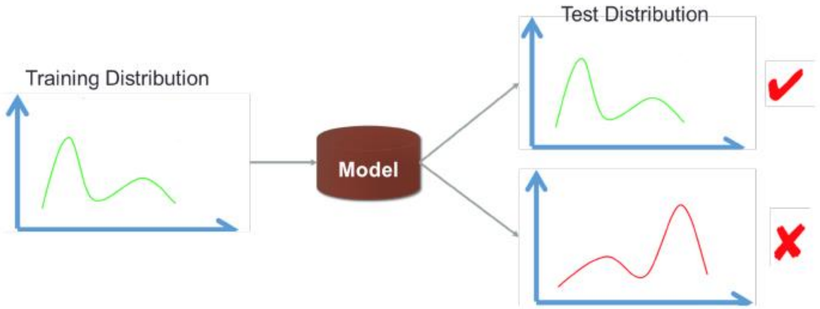
1. X changes, Y also changes

$$\exists x_1 \neq x_2 \; P(Y|X{=}x_1) \neq P(Y|X{=}x_2)$$

Introduction

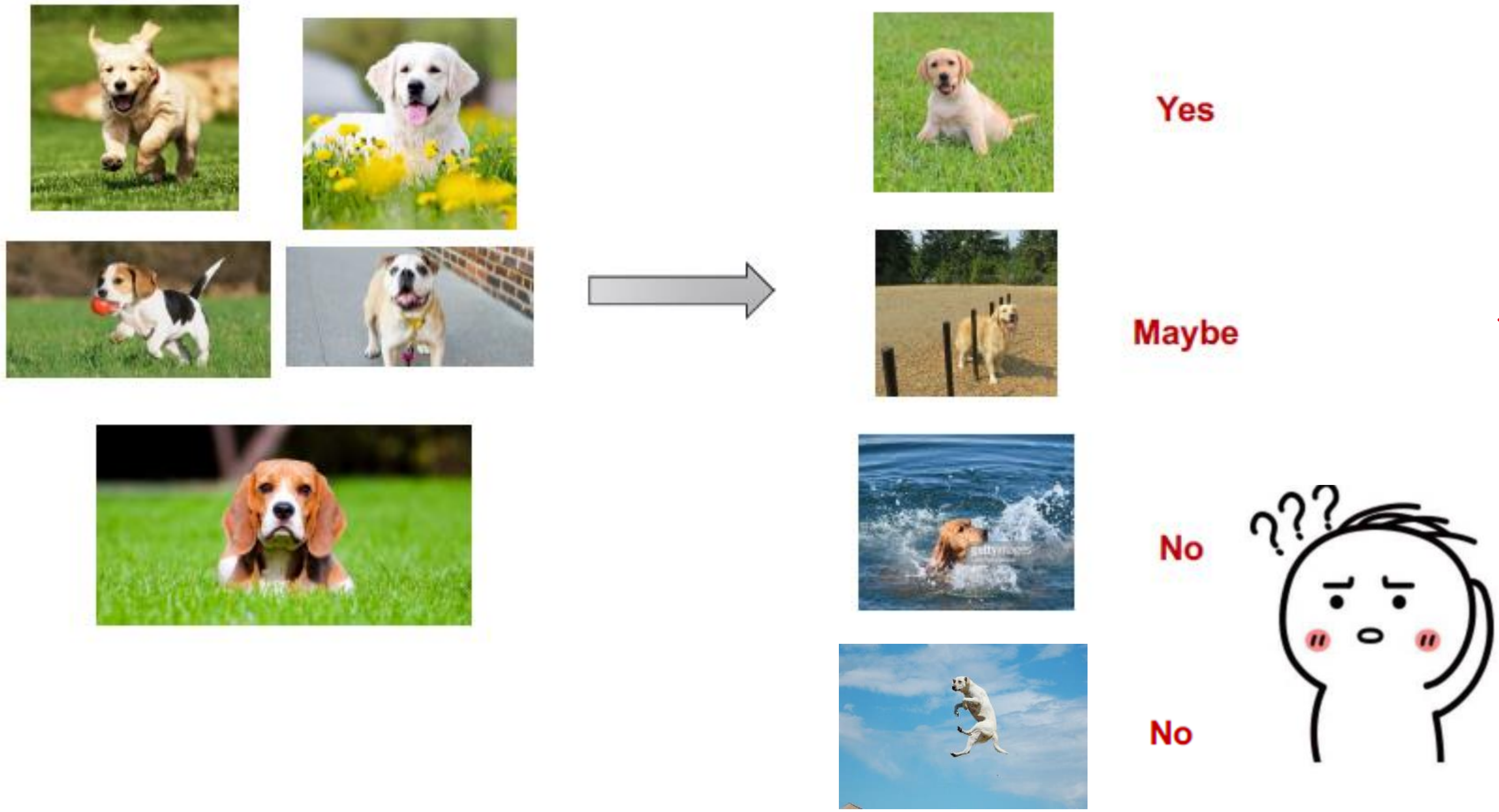
- Machine learning systems often assume training and test set have the same distribution .





Introduction

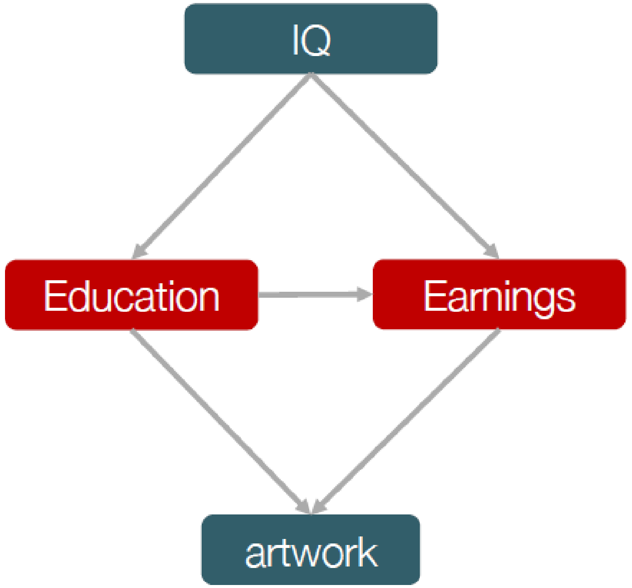
machine learning is not stable



Introduction

machine learning is not explainable

- Question: the causal effect of education attainment on earnings
- Dataset: education, earnings, IQ, spent on artwork




````
```{r}
N <- 100000

#generate data
IQ <- rnorm(N)
edu <- .5 * IQ + rnorm(N)
earnings <- .3 * IQ + .4 * edu + rnorm(N)
art <- 1.2 * edu + .6 * earnings + rnorm(N)
```
````

From which can we get an unbiased estimation?

````
```{r}
summary(lm(earnings ~ edu))
summary(lm(earnings ~ edu + IQ))
summary(lm(earnings ~ edu + IQ + art))
```
````

Introduction

machine learning is not explainable

```
Call:
lm(formula = earnings ~ edu)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.2825 -0.6950 -0.0023  0.6929  4.4687 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.344e-05  3.274e-03   -0.01    0.992    
edu          5.181e-01  2.925e-03  177.12   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.035 on 99998 degrees of freedom
Multiple R-squared:  0.2388,	Adjusted R-squared:  0.2388 
F-statistic: 3.137e+04 on 1 and 99998 DF,  p-value: < 2.2e-16
```

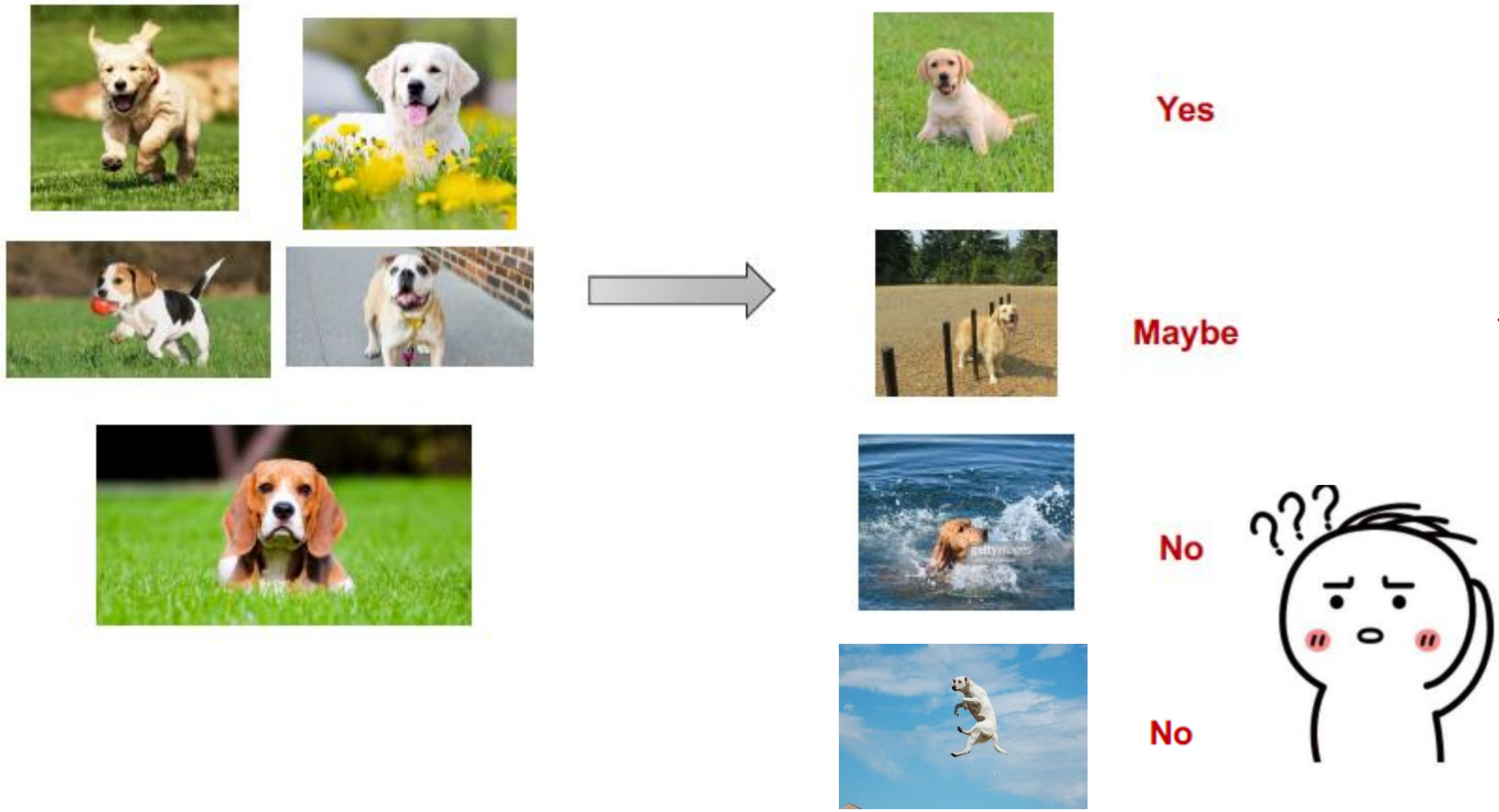
```
Call:
lm(formula = earnings ~ edu + IQ)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.2078 -0.6729 -0.0015  0.6727  3.9517 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.001230   0.003162  -0.389    0.697    
edu          0.398195   0.003158 126.088   <2e-16 ***
IQ           0.299418   0.003525  84.952   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9999 on 99997 degrees of freedom
Multiple R-squared:   0.29,	Adjusted R-squared:   0.29 
F-statistic: 2.043e+04 on 2 and 99997 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = earnings ~ edu + IQ + art)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.6666 -0.5782  0.0003  0.5773  3.7976 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.001869   0.002708   -0.69     0.49    
edu         -0.237545   0.004293  -55.33   <2e-16 ***
IQ           0.218788   0.003048   71.79   <2e-16 ***
art          0.443131   0.002324  190.68   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8563 on 99996 degrees of freedom
Multiple R-squared:  0.4793,	Adjusted R-squared:  0.4793 
F-statistic: 3.069e+04 on 3 and 99996 DF,  p-value: < 2.2e-16
```

````
```{r}
N <- 100000

#generate data
IQ <- rnorm(N)
edu <- .5 * IQ + rnorm(N)
earnings <- .3 * IQ + .4 * edu + rnorm(N)
art <- 1.2 * edu + .6 * earnings + rnorm(N)
```
````

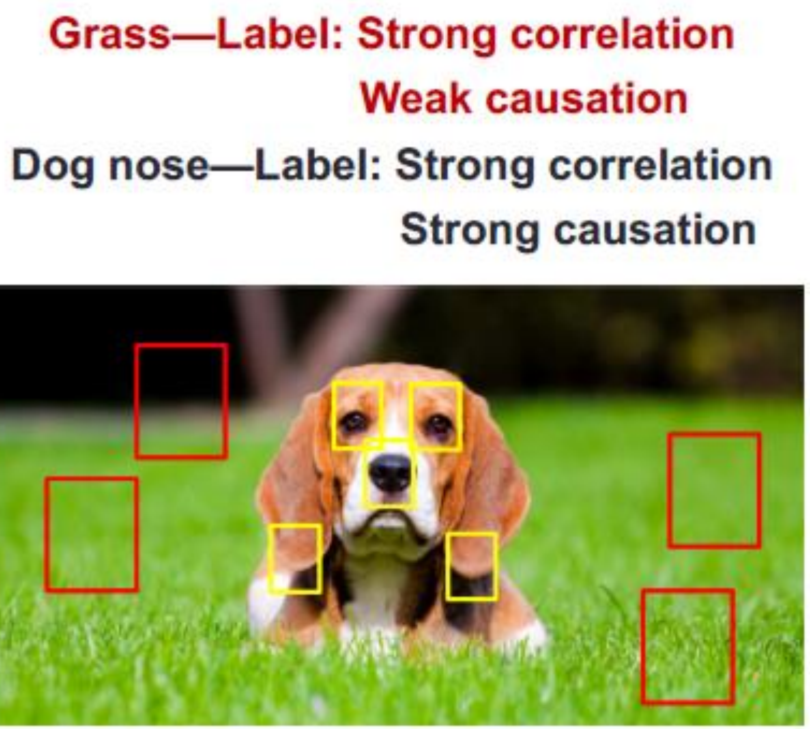
Introduction

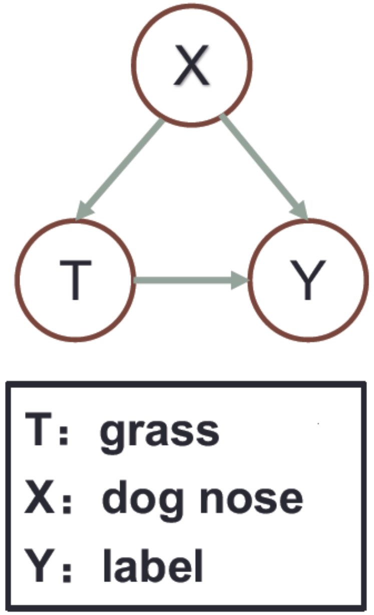
machine learning is not explainable



Introduction

The benefits of bringing causality into machine learning





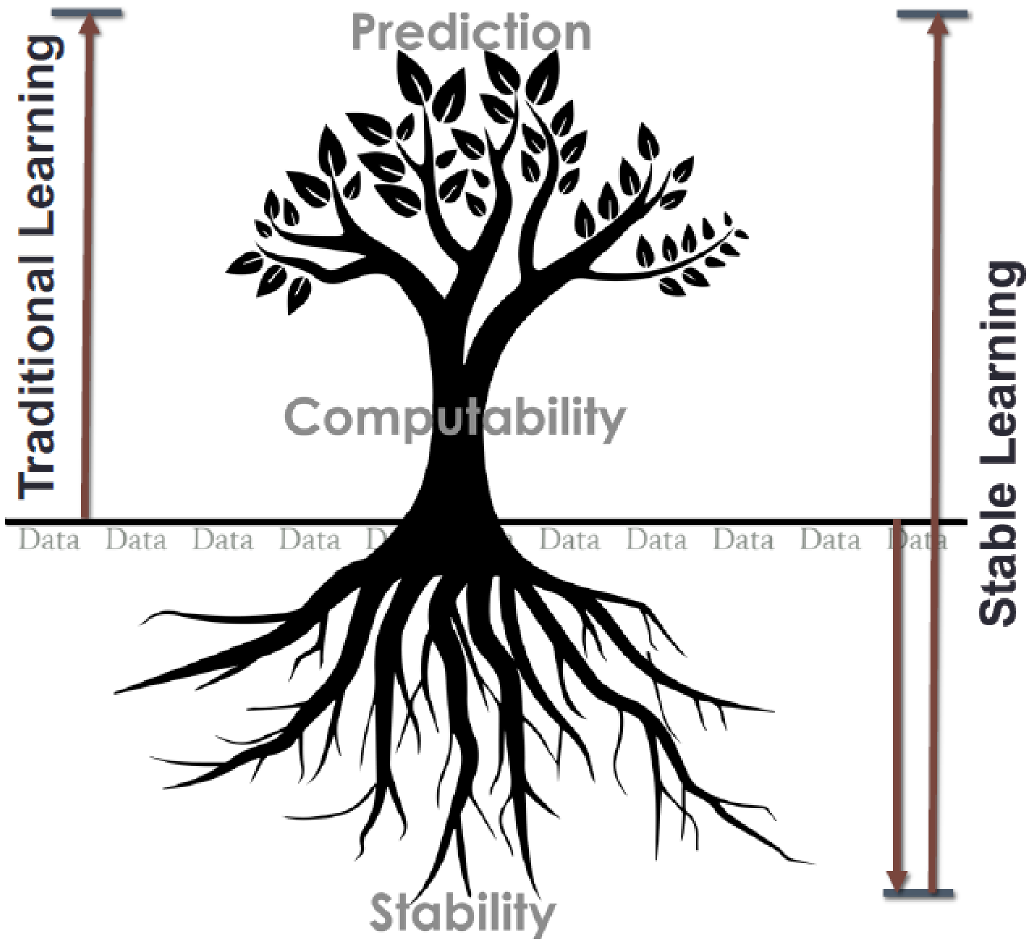
More explainable and more stable

Introduction

Prediction Performance

Learning Process

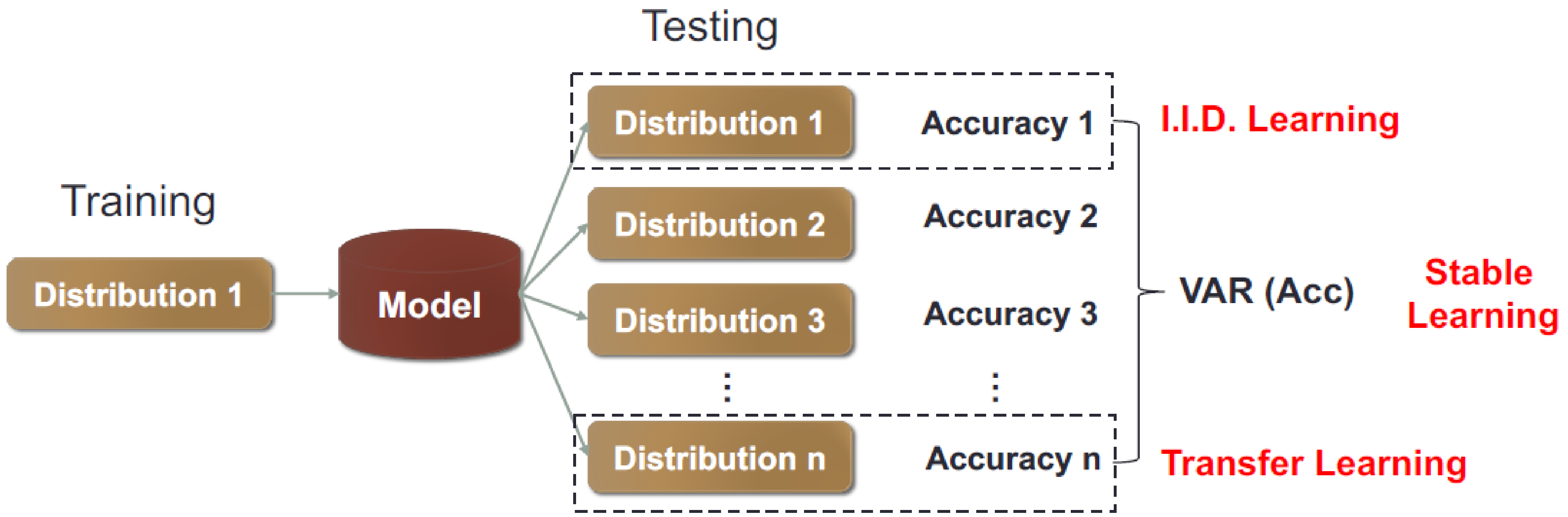
True Model



Bin Yu (2016), Three Principles of Data Science: predictability, computability, stability

Introduction

Stable Learning: Definition



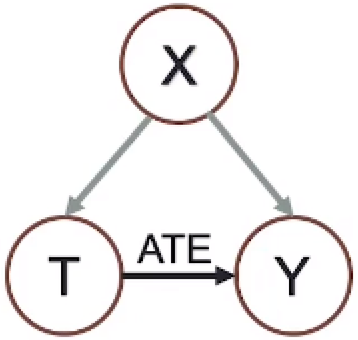
Stable Learning: Achieve uniformly good performance on any distribution

/02

Sample Reweighting: Bridge from Causality to ML

Sample Reweighting: Bridge from Causality to ML

Causal Problem



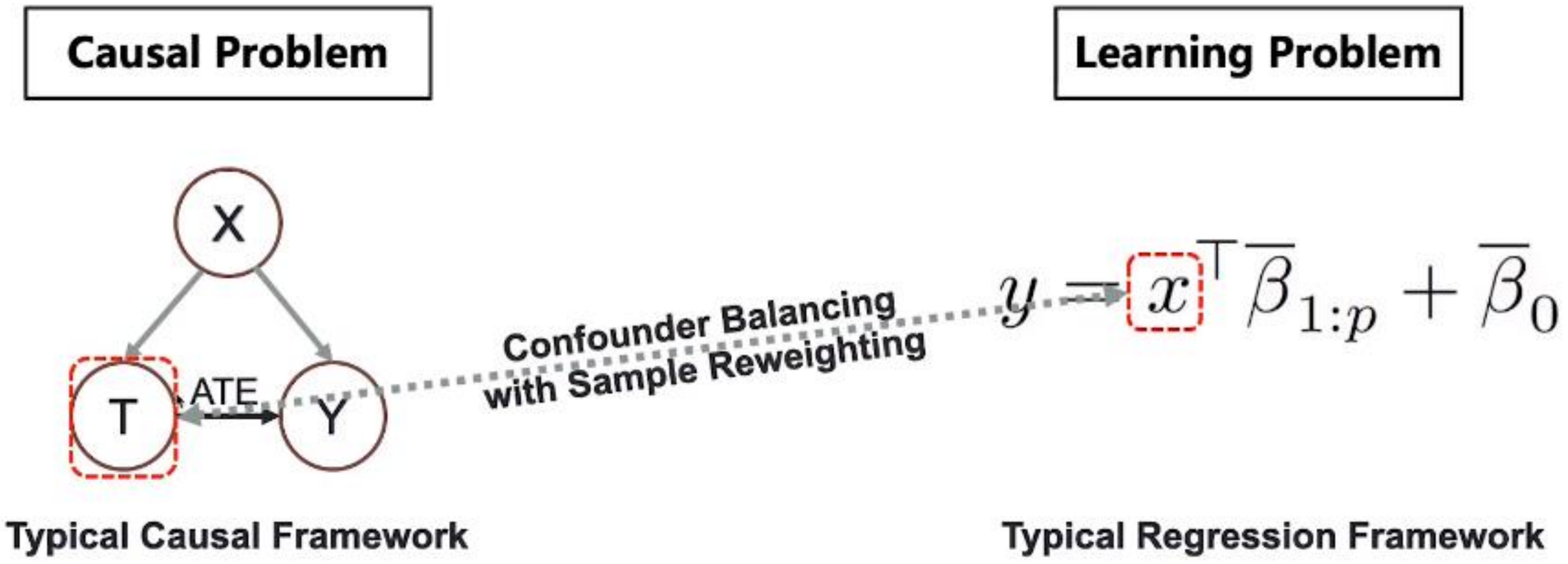
Typical Causal Framework

Learning Problem

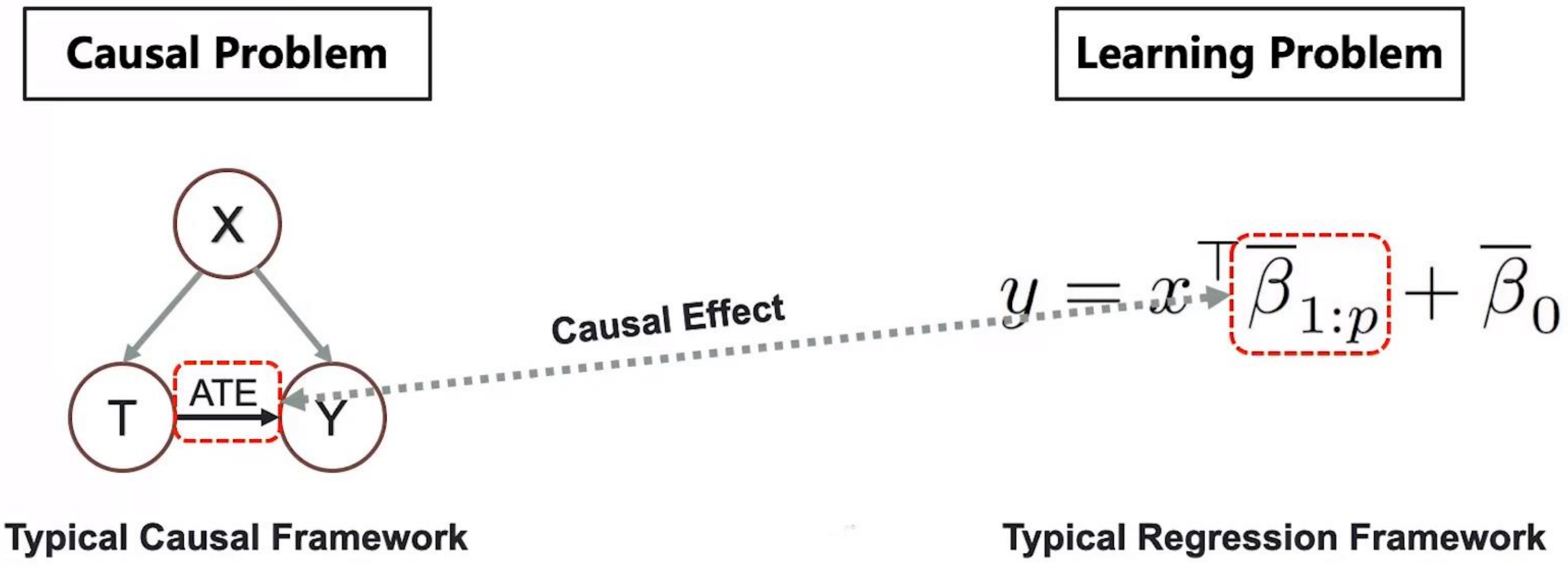
$$y = x^\top \overline{\beta}_{1:p} + \overline{\beta}_0$$

Typical Regression Framework

Sample Reweighting: Bridge from Causality to ML



Sample Reweighting: Bridge from Causality to ML



After confounder balancing, partial effect can be regarded as causal effect.
Predicting with causal variables is stable across different environments.

Sample Reweighting: Bridge from Causality to ML

Directly Confounder Balancing

Given a feature T

↓

Assign different weights to samples so that the samples with T and the samples without T have similar distributions in X

↓

Calculate the difference of Y distribution in treated and controlled groups. (correlation between T and Y)

Over-parametrization and infeasible in high-dimensional setting!

Global Balancing

Given **ANY** feature T

↓

Assign different weights to samples so that the samples with T and the samples without T have similar distributions in X

↓

Calculate the difference of Y distribution in treated and controlled groups. (correlation between T and Y)

Removing confounding bias with an unique set of global weights.

Sample Reweighting: Bridge from Causality to ML

Theoretical Guarantee

PROPOSITION 3.3. *If* $0 < \hat{P}(\mathbf{X}_i = x) < 1$ *for all* x, *where* $\hat{P}(\mathbf{X}_i = x) = \frac{1}{n}\sum_i \mathbb{I}(\mathbf{X}_i = x)$, *there exists a solution* W^* *satisfies equation (4) equals 0 and variables in* $\mathbf{X}$ *are independent after balancing by* W^*.

$$\sum_{j=1}^{p} \left\| \frac{\mathbf{X}_{\cdot,-j}^T \cdot (W \odot \mathbf{X}_{\cdot,j})}{W^T \cdot \mathbf{X}_{\cdot,j}} - \frac{\mathbf{X}_{\cdot,-j}^T \cdot (W \odot (1 - \mathbf{X}_{\cdot,j}))}{W^T \cdot (1 - \mathbf{X}_{\cdot,j})} \right\|_2^2, \quad (4)$$

↓

0

PROOF. Since $\|\cdot\| \geq 0$, Eq. (8) can be simplified to $\forall j, \forall k \neq j$

$$\lim_{n\to\infty}\left(\frac{\sum_{i:\mathbf{X}_{i,k}=1,\mathbf{X}_{i,j}=1} W_i}{\sum_{i:\mathbf{X}_{i,j}=1} W_i} - \frac{\sum_{i:\mathbf{X}_{i,k}=1,\mathbf{X}_{i,j}=0} W_i}{\sum_{i:\mathbf{X}_{i,j}=0} W_i}\right) = 0$$

with probability 1. For W^*, from Lemma 3.1, $0 < P(\mathbf{X}_i = x) < 1$, $\forall x, \forall i, t = 1$ or 0,

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n}\sum_{i:\mathbf{X}_{i,j}=t} W_i^* &= \lim_{n\to\infty} \frac{1}{n}\sum_{x:x_j=t} \sum_{i:\mathbf{X}_i=x} W_i^* \\
&= \lim_{n\to\infty} \sum_{x:x_j=t} \frac{1}{n}\sum_{i:\mathbf{X}_i=x} \frac{1}{P(\mathbf{X}_i=x)} \\
&= \lim_{n\to\infty} \sum_{x:x_j=t} P(\mathbf{X}_i = x) \cdot \frac{1}{P(\mathbf{X}_i=x)} = 2^{p-1}
\end{aligned}$$

with probability 1 (Law of Large Number). Since features are binary,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i:\mathbf{X}_{i,k}=1,\mathbf{X}_{i,j}=1} W_i^* = 2^{p-2}$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i:\mathbf{X}_{i,j}=0} W_i^* = 2^{p-1}, \quad \lim_{n\to\infty} \frac{1}{n}\sum_{i:\mathbf{X}_{i,k}=1,\mathbf{X}_{i,j}=0} W_i^* = 2^{p-2}$$

and therefore, we have following equation with probability 1:

$$\lim_{n\to\infty}\left(\frac{\mathbf{X}_{\cdot,k}^T(W^*\odot\mathbf{X}_{\cdot,j})}{W^{*T}\mathbf{X}_{\cdot,j}} - \frac{\mathbf{X}_{\cdot,k}^T(W^*\odot(1-\mathbf{X}_{\cdot,j}))}{W^{*T}(1-\mathbf{X}_{\cdot,j})}\right) = \frac{2^{p-2}}{2^{p-1}} - \frac{2^{p-2}}{2^{p-1}} = 0.$$

□

Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Li, Bo Li. Stable Prediction across Unknown Environments. ***KDD***, 2018.

Sample Reweighting: Bridge from Causality to ML

Causally Regularized Logistic Regression (CRLR)

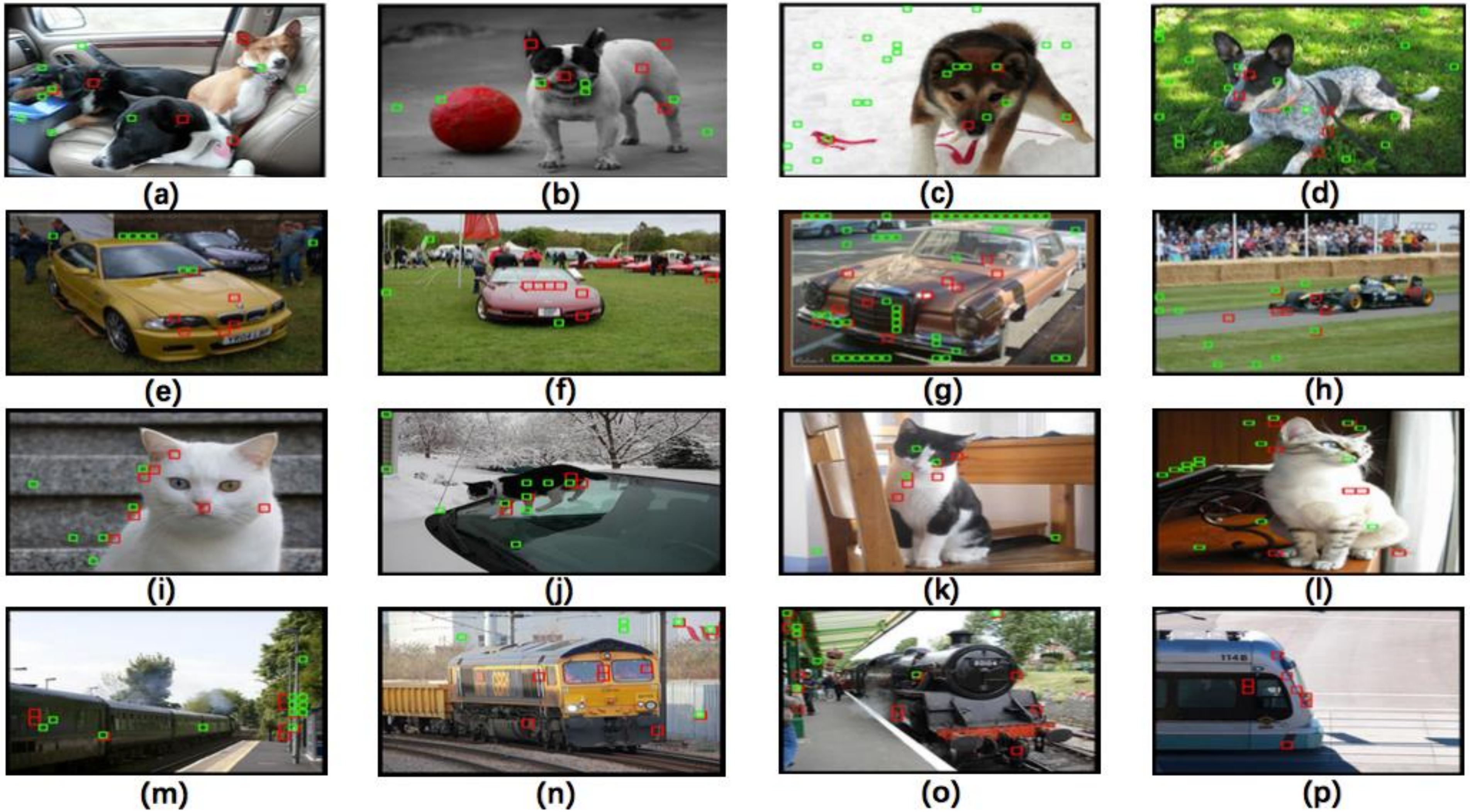
$$
\begin{aligned}
\min \quad & \sum_{i=1}^{n} W_i \cdot \log(1 + \exp((1 - 2Y_i) \cdot (x_i \beta))), \\
s.t. \quad & \sum_{j=1}^{p} \left\| \frac{X_{-j}^T \cdot (W \odot I_j)}{W^T \cdot I_j} - \frac{X_{-j}^T \cdot (W \odot (1 - I_j))}{W^T \cdot (1 - I_j)} \right\|_2^2 \le \lambda_1, \\
& W \succeq 0, \quad \|W\|_2^2 \le \lambda_2, \quad \|\beta\|_2^2 \le \lambda_3, \quad \|\beta\|_1 \le \lambda_4, \\
& \left(\sum_{k=1}^{n} W_k - 1\right)^2 \le \lambda_5,
\end{aligned}
$$

Sample reweighted logistic loss

Causal Contribution

Zheyan Shen, Peng Cui, Kun Kuang, Bo Li. Causally Regularized Learning on Data with Agnostic Bias. ***ACM MM**, 2018.*

Sample Reweighting: Bridge from Causality to ML



/03

Stable Learning: From Statistical Learning Perspective

Stable Learning: From Statistical Learning Perspective

Sample reweighting

$$y = x^\top \bar{\beta}_{1:p} + \bar{\beta}_0 + b(x) + \epsilon,$$

Linear part

Bias cannot be modeled by linear part

Noise term

Assumption

1) the linear part of generation model is stable and invariant to unknown distribution shift

2) the misspecification bias could be unstable and bounded $|b(x)| \leq \delta$.

Un-stability

1) Bias term

2) Input variables without causality

Estimate parameters as accurately as possible and make the error uniformly small for all x

Stable Learning: From Statistical Learning Perspective

Sample reweighting

Least squares regression
$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left(x_i^\top \beta_{1:p} + \beta_0 - y_i\right)^2$$

Solutions without collinearity: $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$

However, the estimation error caused by misspecification term can be as bad as $\|\hat{\beta} - \overline{\beta}\|_2 \le 2(\delta/\gamma) + \delta,$ *where* γ^2 *is the smallest eigenvalue of* $\mathbf{E}(x - \mathbf{E}x)(x - \mathbf{E}x)^\top$.

A small γ implies high collinearity, which means **high collinearity** leads to poor solution

⇩

Reducing collinearity by sample reweighting

Stable Learning: From Statistical Learning Perspective

Toy example

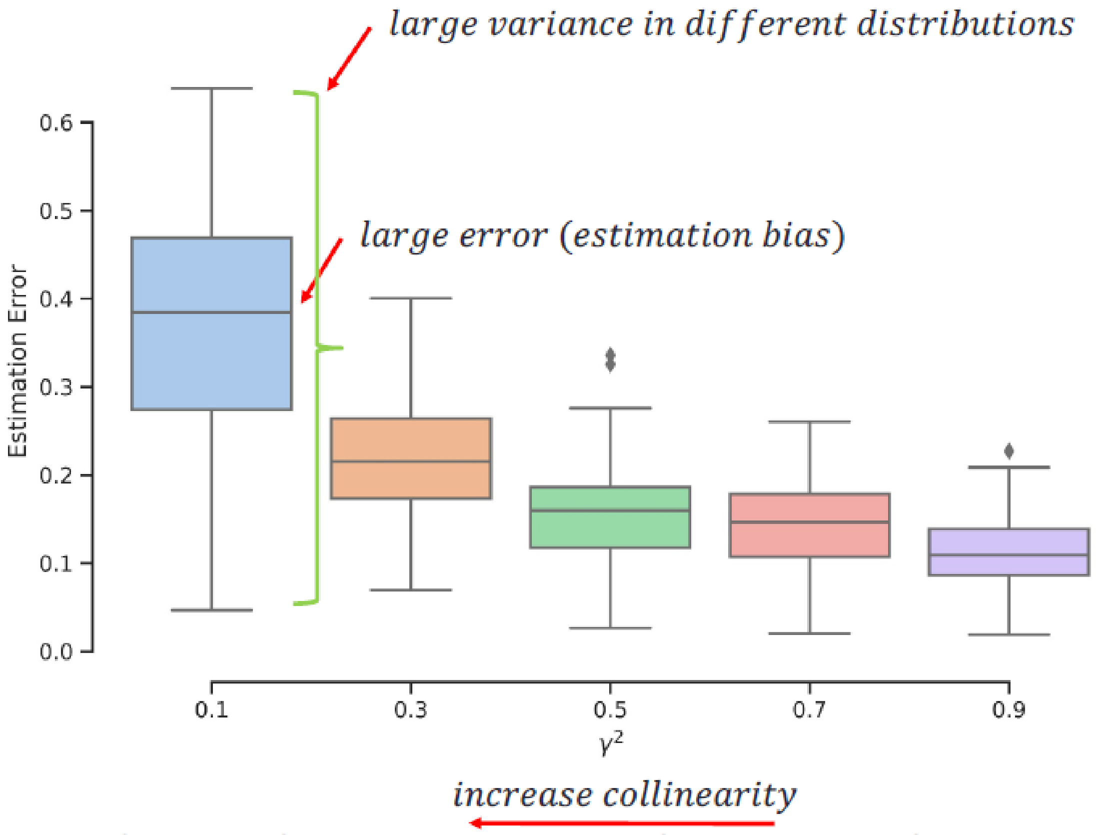
- Assume the design matrix X consists of two variables X_1, X_2, generated from a multivariate normal distribution:

$$X \sim N(0, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

- By changing ρ, we can simulate different extent of collinearity.
- To induce bias related to collinearity, we generate bias term $b(X)$ with $b(X) = Xv$, where v is the eigenvector of centered covariance matrix corresponding to its smallest eigenvalue γ^2.
- The bias term is sensitive to collinearity.

Stable Learning: From Statistical Learning Perspective

Toy example



Stable Learning: From Statistical Learning Perspective

Idea: Learn a new set of ***sample weights*** $w(x)$ to decorrelate the input variables and increase the smallest eigenvalue

For regression:

$$\hat{\beta}_{WLS} = \arg\min_{\beta} \sum_{i=1}^{n} \hat{W}_i \cdot (Y_i - \mathbf{X}_{i,}\beta)^2.$$

For classification:

$$\sum_{i=1}^{n} w(x_i) \ln\left(1 + \exp\left(-\beta^{\top} x_i y_i\right)\right).$$

Stable Learning: From Statistical Learning Perspective

Sample reweighting

Algorithm 1 Sample Reweighted Decorrelation Operator (SRDO)

Require: Design Matrix $\mathbf{X}$

```
for i = 1 ... n do
    Initialize a new sample x̃_i ∈ R^p with empty vector
    for j = 1 ... p do
        Draw the j^th feature of new sample x̃_{i,j} from X_{,j} at random
    end for
end for
```

By treating the different columns independently while performing **random resampling**, we can obtain a column-decorrelated design matrix with the same marginal as before.

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \xrightarrow{\text{Decorrelation}} \tilde{\mathbf{X}} = \begin{pmatrix} x_{i1} & \cdots & x_{rl} & \cdots \\ x_{j1} & \cdots & x_{sl} & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & \cdots & x_{tl} & \cdots \end{pmatrix}$$

where i, j, k, r, s, t are drawn from $1 \dots n$ at random

Stable Learning: From Statistical Learning Perspective

Sample reweighting

Algorithm 1 Sample Reweighted Decorrelation Operator (SRDO)

Require: Design Matrix $\mathbf{X}$

for $i = 1 \ldots n$ **do**
  Initialize a new sample $\tilde{x}_i \in \mathbb{R}^p$ with empty vector
  for $j = 1 \ldots p$ **do**
    Draw the j^{th} feature of new sample $\tilde{x}_{i,j}$ from $\mathbf{X}_{,j}$ at random
  end for
end for
Set $\tilde{x}_i$ as positive samples and x_i as negative samples, then train a binary classifier.
Set $w(x) = \frac{p(Z=1|x)}{p(Z=0|x)}$ for each sample x_i in $\mathbf{X}$, where $p(Z = 1|x)$ is the probability of sample x been drawn from $\tilde{D}$ estimated by the trained classifier.

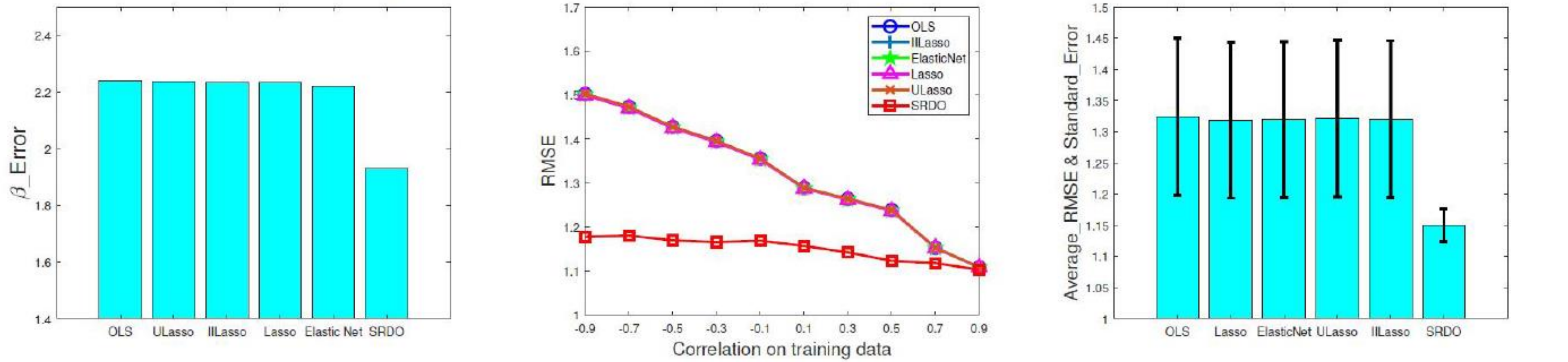
Ensure: A set of sample weights $w(x)$ which can deccorelate $\mathbf{X}$

By treating the different columns independently while performing random resampling, we can obtain a column-decorrelated design matrix with the same marginal as before.

get sample weight by using density ratio estimation

Stable Learning: From Statistical Learning Perspective

- Simulation Study



(a) Estimation error (b) Prediction error over different test environments (c) Average prediction error&stability

/05 Conclusion

Conclusion

1. Stable Learning cares about not only the prediction accuracy but also the prediction stability across different distributions.

2. Causality provide firm soil for the understanding intrinsic mechanism of stable learning.

Thanks!

Q&A?

Sample Reweighting: Bridge from Causality to ML

Causal Regularizer for Continuous Variable

$$\min_{W} \sum_{j=1}^{p} \left\| \mathbb{E}[\mathbf{X}_{,j}^T \mathbf{\Sigma}_W \mathbf{X}_{,-j}] - \mathbb{E}[\mathbf{X}_{,j}^T W] \mathbb{E}[\mathbf{X}_{,-j}^T W] \right\|_2^2$$

Decorrelated Weighted Regression:

$$\min_{W,\beta} \sum_{i=1}^{n} W_i \cdot (Y_i - \mathbf{X}_{i,}\beta)^2$$

$$s.t \quad \sum_{j=1}^{p} \left\| \mathbf{X}_{,j}^T \mathbf{\Sigma}_W \mathbf{X}_{,-j}/n - \mathbf{X}_{,j}^T W/n \cdot \mathbf{X}_{,-j}^T W/n \right\|_2^2 < \lambda_2$$

$$|\beta|_1 < \lambda_1, \quad \frac{1}{n}\sum_{i=1}^{n} W_i^2 < \lambda_3,$$

$$\left(\frac{1}{n}\sum_{i=1}^{n} W_i - 1\right)^2 < \lambda_4, \quad W \succeq 0,$$

Kuang, K., Xiong, R., Cui. Stable Prediction with Model Misspecification and Agnostic Distribution Shift. *AAAI,* 2020

https://github.com/KunKuang/Decorrelated-Weighted Regression