CAUSAL REPRESENTATION LEARNING

Ph.D. Student Chunjiang Ge
Department of Automation
6th Dec. 2020

IDEAS FROM THE FOLLOWINGS

- Yao L, Chu Z, Li S, et al. A Survey on Causal Inference[J]. arXiv preprint arXiv:2002.02770, 2020.
- Li S, Yao L, Li L, et al. Representation Learning for Causal Inference[R]. Association for the Advancement of Artificial Intelligence(AAAI), 2020.
- Schölkopf B. Causality for machine learning[J]. arXiv preprint arXiv:1911.10500, 2019.
- Guo R, Cheng L, Li J, et al. A survey of learning causality with data: Problems and methods[J]. ACM Computing Surveys (CSUR), 2020, 53(4): 1-37.

CONTENTS

NOTATION

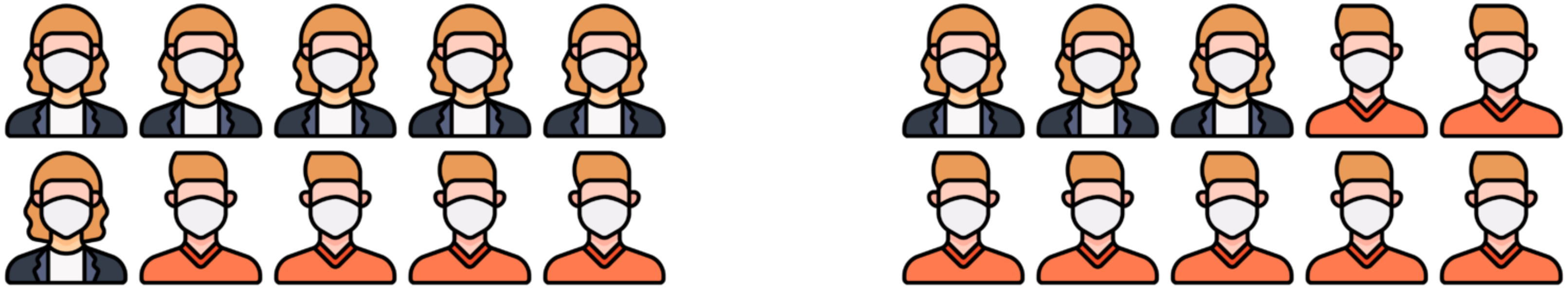
- Covariates: X_i; Assignment: T_i; Potential Outcome: Y_i
- Reweighting Weights: W_i
- Propensity Score: $e(x) = P(T = 1|X = x)$
- Neural Network Feature Extraction: $\Phi(x)$
- Classification: $h(x,\cdot)$
- Probability Distribution: P^{Φ}

TRADITIONAL CAUSAL INFERENCE METHODS

Causal Inference for Observational Study

OBSERVATIONAL STUDY

- Randomized Controlled Trials v.s. Observational study
- Study on the pesticide effect:



T=1 T=0

Covariates Shift

Kuang Kun: https://campus.swarma.org/course/2030

DE-CONFOUNDING

T=1 T=0

Kuang Kun: https://campus.swarma.org/course/2030

DE-CONFOUNDING

$$
\begin{tikzcd}
 & Z \arrow[dl] \arrow[dr] & & & & Z \arrow[dr] & \\
X \arrow[rr] & & Y \arrow[r, phantom, "\longrightarrow"] & & X \arrow[rr] & & Y
\end{tikzcd}
$$

X: medical treatment

Z: gender

Y: pesticide effect

$$
\begin{aligned}
P(Y = y|do(X = x)) &= \sum_z P(Y = y|X = x, PA = z)P(PA = z) \\
&= \sum_z \frac{P(X = x, Y = y, PA = z)}{P(X = x, |PA = z)}
\end{aligned}
$$

Kuang Kun: https://campus.swarma.org/course/2030

INVERSE PROPENSITY WEIGHTING

- Balancing Score: b(x). $W \perp\!\!\!\perp x \mid b(x)$
- Propensity Score is a kind of balancing score: $e(x) = P(W = 1 | X = x)$
- Average Treatment Effect:

$$\widehat{ATE} = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \frac{W_i Y_i^F}{\hat{e}(x)} - \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \frac{(1 - W_i) Y_i^F}{1 - \hat{e}(x)}$$

Kuang Kun: https://campus.swarma.org/course/2030

STRATIFICATION

- Stratification is another way for covariates balancing.
- Splitting the entire group into several similar groups, calculating ATE in each group.
- ATE is:

$$\widehat{ATE} = \sum_{j=1}^{J} q(j)[\bar{Y}_t(j) - \bar{Y}_c(j)]$$

MATCHING

- Matching can be viewed as a extreme version of stratification. Set each group size as 1.
- Matching estimates the counterfactuals and reduce the estimation bias brought by the confounders.

$$\hat{Y}_i(1-t_i) = argmin_{Y_j} d(\mathrm{Y_j}, \mathrm{Y_j}) \,\forall\, t_j = 1 - t_i$$

- $\mathrm{d}(\cdot)$ is a metric function.

MATCHING

- Metrics:
  - Propensity score: $\hat{e}(x) = P(y = 1|x)$
  - Euclidean distance: $d(x_1, x_2) = ||x_1 - x_2||_2^2$
  - Mahalanobis distance: $d(x_1, x_2) = [(x_1 - x_2)^T M^{-1} (x_1 - x_2)]^{0.5}$

Kuang Kun: https://campus.swarma.org/course/2030

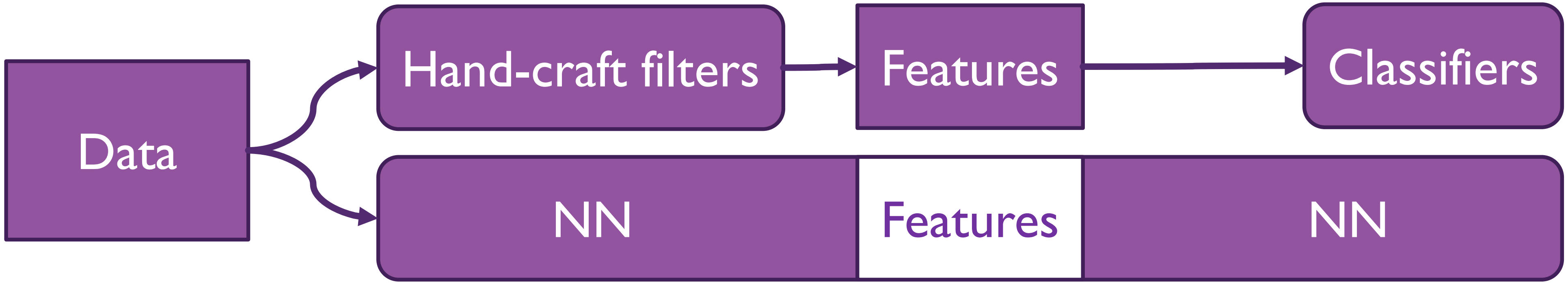
These methods balance the distribution and reduce the confounding bias.

DEEP REPRESENTATION LEARNING FOR CAUSAL INFERENCE

Representation Learning plus Causal Inference

DEEP REPRESENTATION LEARNING

- Representation learning is a set of techniques that allows a system to **automatically discover the representations** needed for feature detection or classification from raw data. This replaces **manual feature engineering** and allows a machine to both learn the features and use them to perform a specific task.



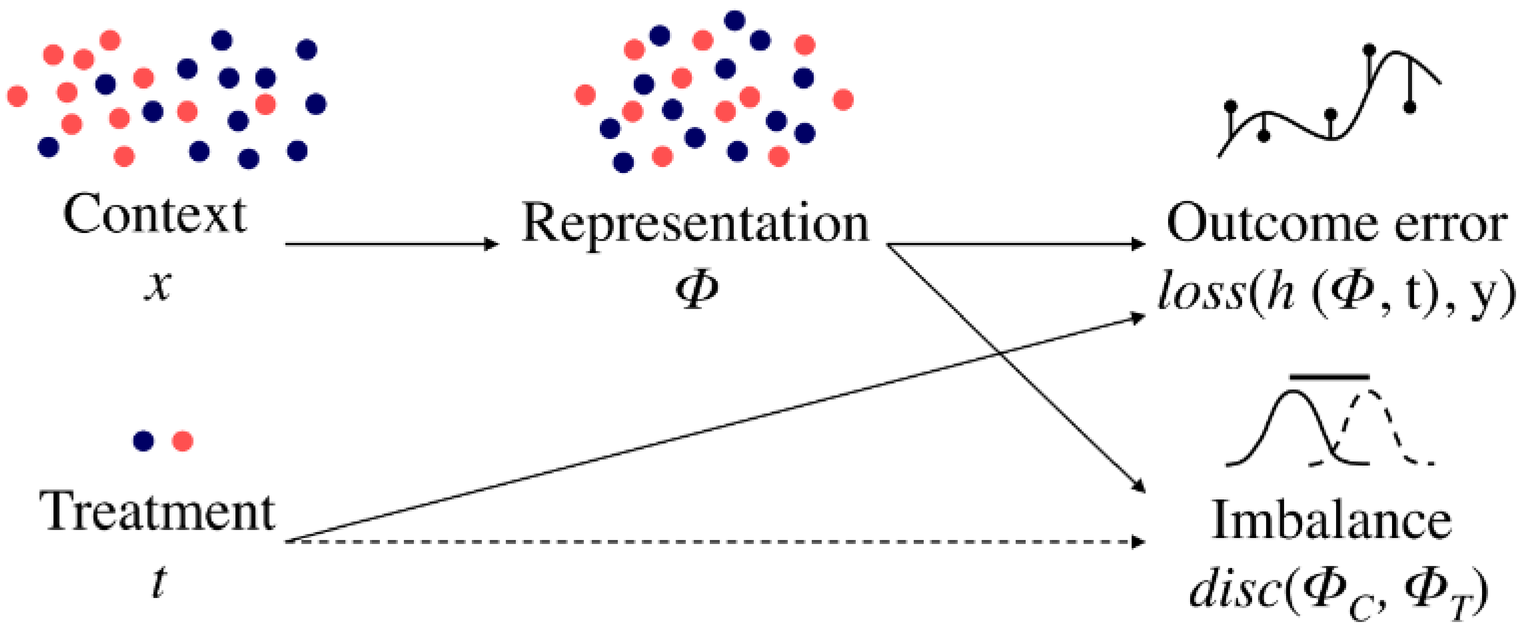
BALANCED METHODS

Balancing Distributions of Treatment and Control Group

BALANCED METHODS

- Assumption: Training data and test data are independent and identical distributed. So feature learned on the training set can be generalized to test set.
- Domain adaptation: reduce discrepancy of training and test data.
- Observational study: reduce discrepancy of treatment and control group.
- Causal Effect estimation: nn predicts counterfactual + regularization

BALANCING THE TWO GROUPS IN THE LATENT SPACE



$$\min d(\mathbf{P}(x|t=0), \mathbf{P}(x|t=1)) \Rightarrow \min d(\mathbf{P}(\Phi(x)|t=0), \mathbf{P}(\Phi(x)|t=1))$$

DISTANCES

- KL divergence

$$KL(p\|q) = \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k}$$

- MMD distance

$$MMD(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{n_1} \sum_{i=1}^{n_1} \phi(x_i) - \frac{1}{n_2} \sum_{j=1}^{n_2} \phi(x_j)_H^2$$

- Wassertein distance

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{X}} ||x - y||^p d\gamma(x, y) \right)^{1/p}$$

COUNTERFACTUAL REGRESSION

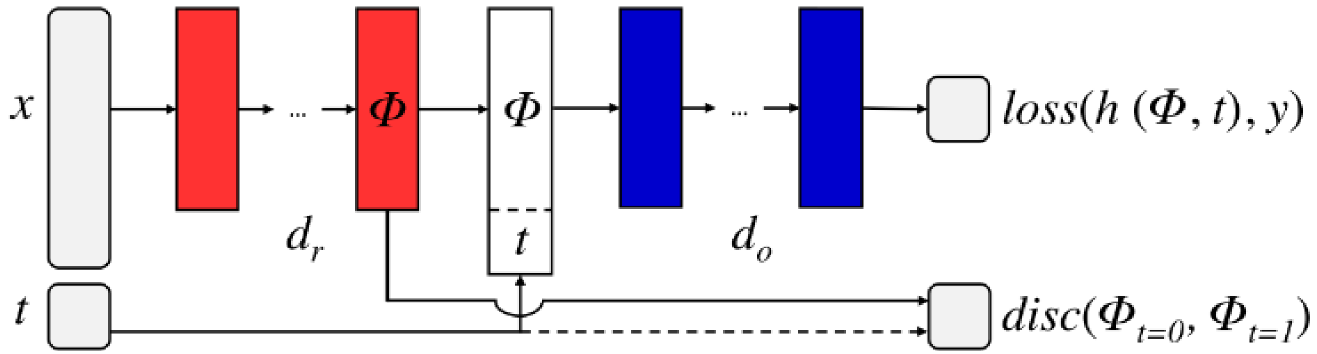
- Counterfactual Inference: e.g. $Y_i(t_i)$ is known, predicting $Y_i(1-t_1)$.
- A trivial estimation is matching.
- ITE Estimation:

$$\widehat{ITE(x_i)} = \begin{cases} y_i^F - h(x_i, 1-t_i), & t_i = 1 \\ h(x_i, 1-t_i) - y_i^F, & t_1 = 0 \end{cases}$$

COUNTERFACTUAL REGRESSION

- Loss Function

$$B_{H,\alpha,\gamma}(\Phi, h) = \frac{1}{n}\sum_{i=1}^{n}\left|h(\phi, t_i) - y_i^F\right| + \alpha\ disc_H\left(P_F^{\Phi}, P_{CF}^{\Phi}\right) + \frac{\gamma}{n}\sum_{i=1}^{n}\left|h(\Phi(x_i), 1 - t_i) - y_{j(i)}^F\right|$$



F. Johansson, U. Shalit, and D. Sontag. "Learning representations for counterfactual inference." International conference on machine learning. 2016

COUNTERFACTUAL REGRESSION

- Network composed of several parts:
  - d_r: Generate a representation $\Phi(x)$ of input feature x
  - d_o: Calculate the counter-factual inference $h(\Phi(x_0), t)$ giving x and t
  - $loss(h(\Phi, t), y)$: Loss for training inference network.
  - $\mathrm{dis}c(\Phi_{CF}, \Phi_F)$: Regularization term: minimize the distance between treatment and control group.

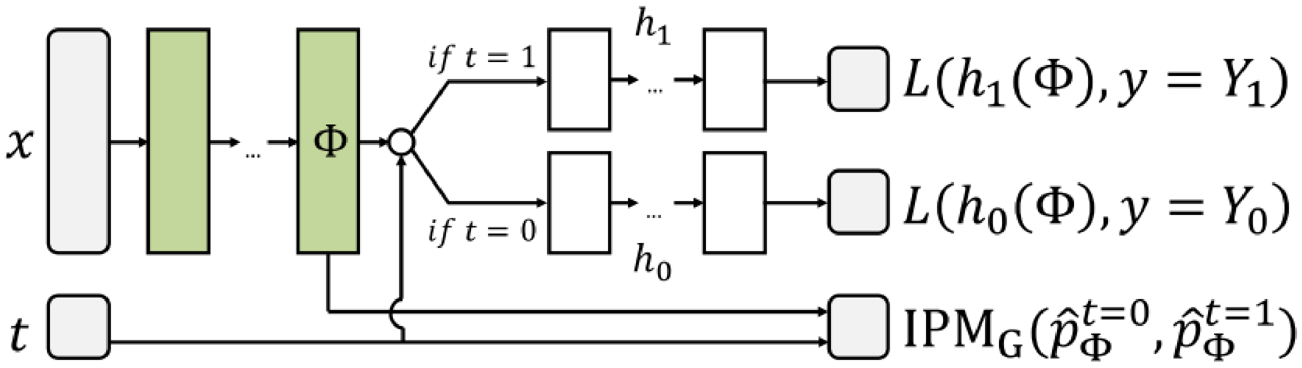
F. Johansson, U. Shalit, and D. Sontag. "Learning representations for counterfactual inference." International conference on machine learning. 2016

COUNTERFACTUAL REGRESSION

- Optimization Target

$$\min_{\substack{||\Phi||=1 \\ h,\Phi}} \frac{1}{n}\sum_{i=1}^{n} w_i\, L(h(\Phi(x_i),t_i),y_i) + \lambda\, R(h) + \alpha\, IPM_G(\{\Phi(x_i)\}_{i:t_i=0},\{\Phi(x_i)\}_{i:t_i=1})$$

$$w_i = \frac{t_i}{2u} + \frac{1-t_i}{2(1-u)}, \text{ where } u = \frac{1}{n}\sum_{i=1}^{n} t_i$$



U. Shalit, F. Johansson, and D. Sontag. "Estimating individual treatment effect: generalization bounds and algorithms." Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017

BALANCING FROM A LOCAL VIEW

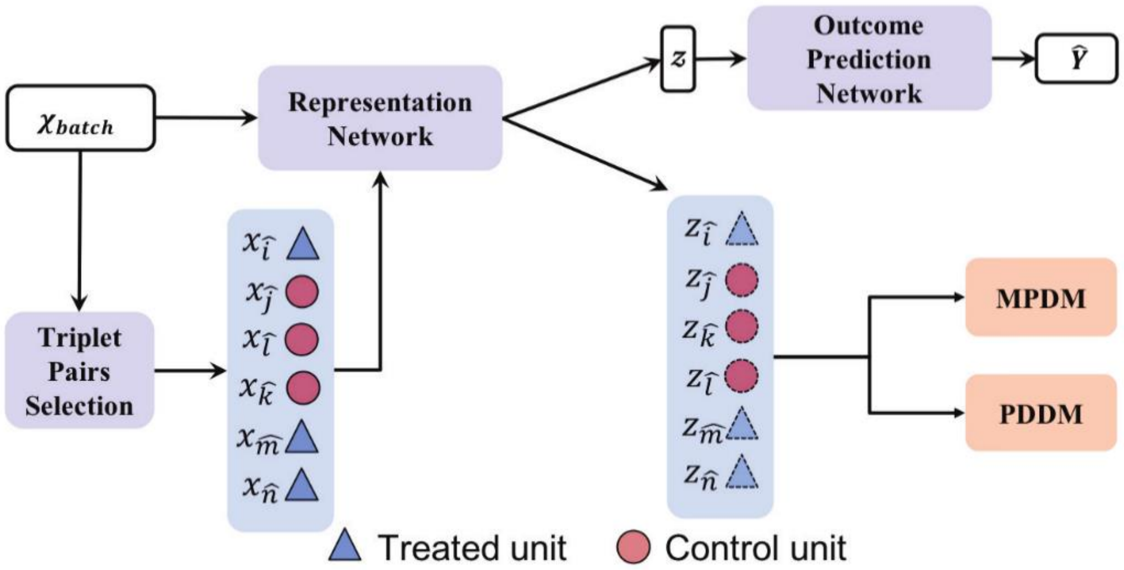
- SITE maps mini-batches of units from the covariate space to a latent space using a representation network:
  - SITE preserves the local similarity information using the Position-Dependent Deep Metric (PDDM),
  - SITE balances the data distributions with a Middle-point Distance Minimization (MPDM) strategy.

L. Yao, et al. "Representation learning for treatment effect estimation from observational data." NeurIPS 2018.

SITE

- Loss Function

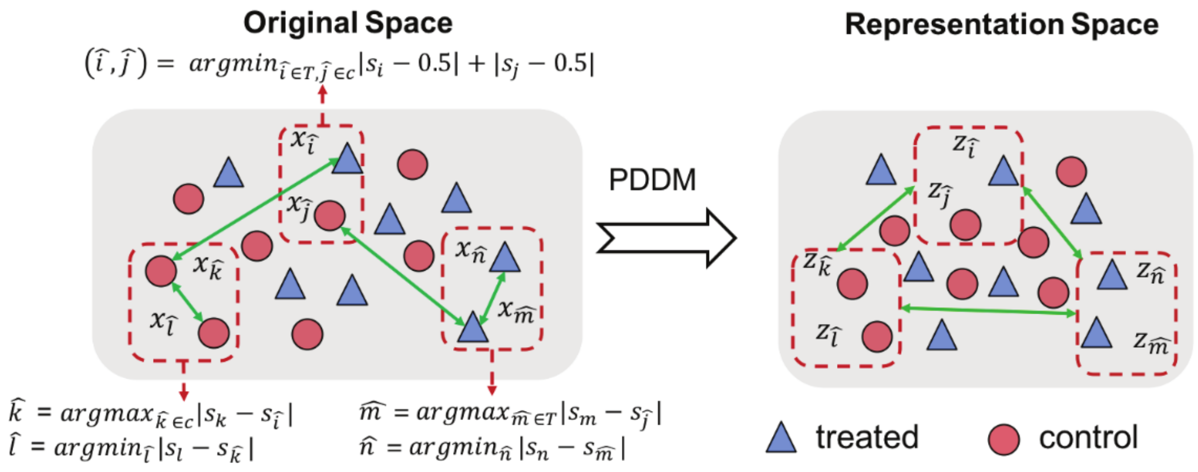
$$L = L_{FL} + \beta L_{PDDM} + \gamma L_{MPDM} + \lambda||W||_2$$



L. Yao, et al. "Representation learning for treatment effect estimation from observational data." NeurIPS 2018.

CHOICE OF SAMPLES

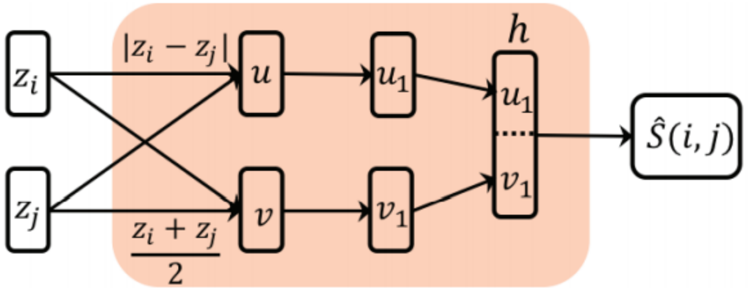
- Choose one sample from treatment and control group respectively with they lay in the intermediate region.
- Choose k farthest from i and m farthest from j.
- Choose l nearest from k and n nearest from m.

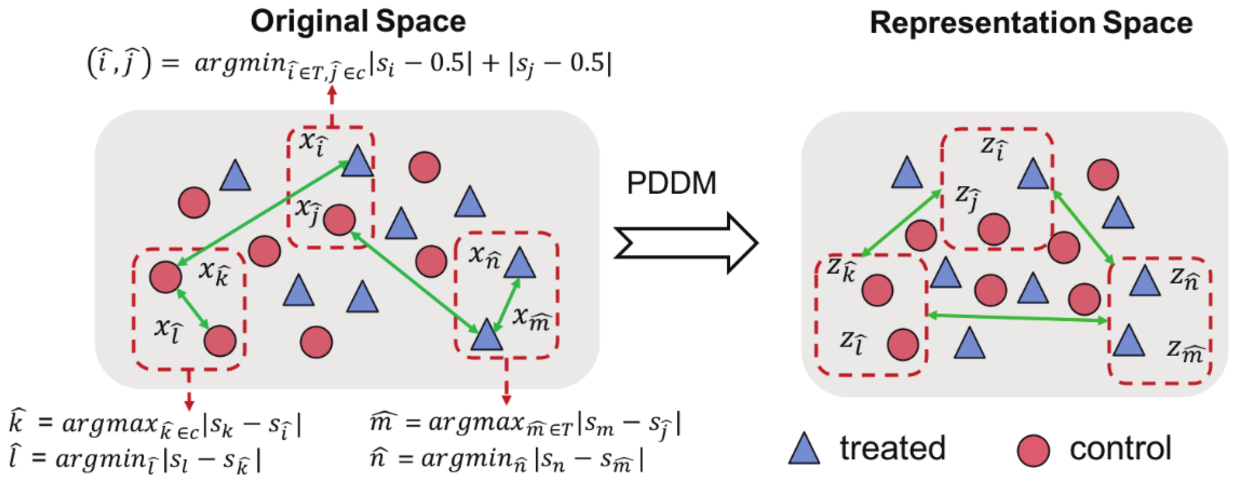


L. Yao, et al. "Representation learning for treatment effect estimation from observational data." NeurIPS 2018.

PDDM

- Position-Dependent Deep Metric (PDDM):
  - The PDDM component measures the local similarity of two units based on their relative and absolute positions in the latent space

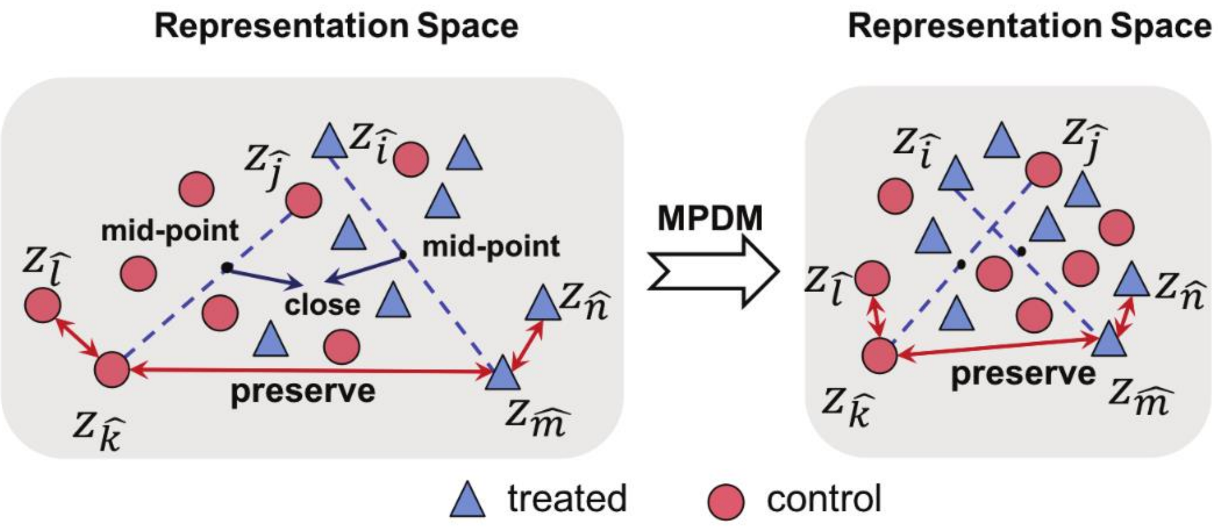




L. Yao, et al. "Representation learning for treatment effect estimation from observational data." NeurIPS 2018.

MPDM

- Middle Point Distance Minimization (MPDM):
  - Makes two mid-points close to each other.
  - Mid-point is an approximation to the center point.
  - The MPDM balances the distribution in the latent space



L. Yao, et al. "Representation learning for treatment effect estimation from observational data." NeurIPS 2018.

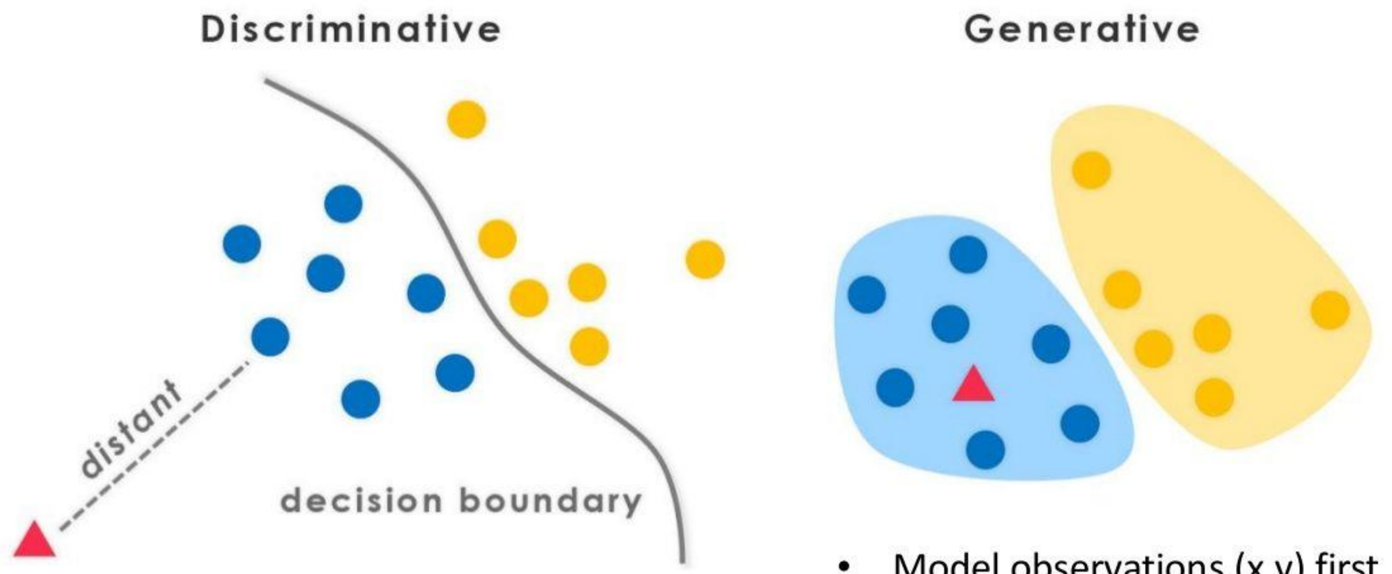
These works balance the distribution in latent space.
Can we learn the latent representation in an explicit way?

LATENT REPRESENTATION METHODS

Generate Latent Causal and Non-Causal Variables

GENERATIVE MODELS

Discriminative vs. Generative

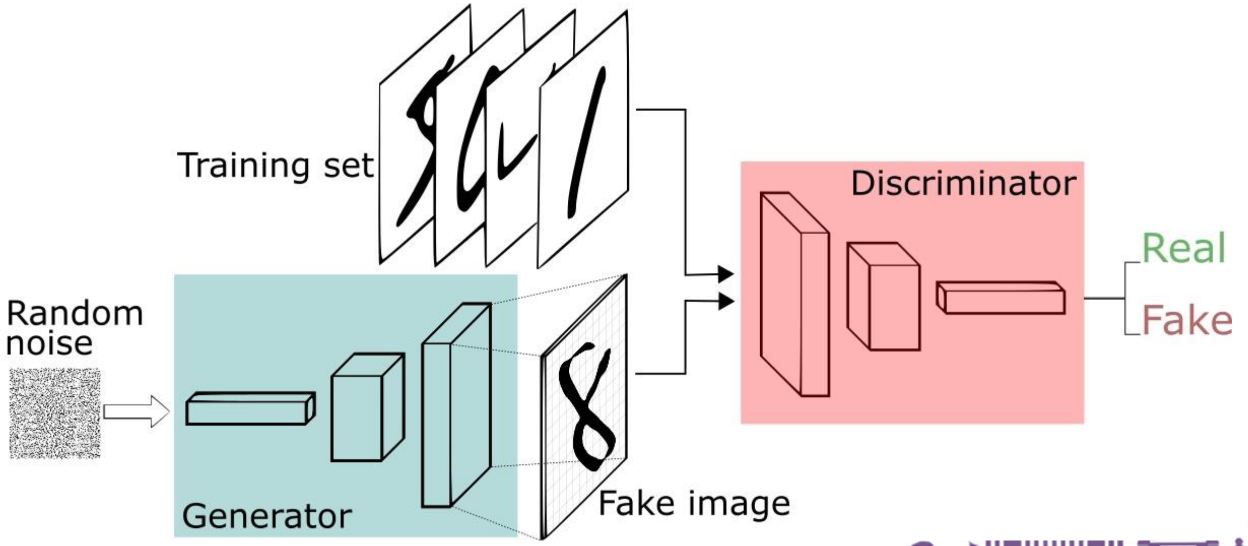


- Only care about estimating the conditional probabilities
- Very good when underlying distribution of data is really complicated (e.g. texts, images, movies)

- Model observations (x,y) first, then infer p(y|x)
- Good for missing variables, better diagnostics
- Easy to add prior knowledge about data

GENERATIVE MODELS

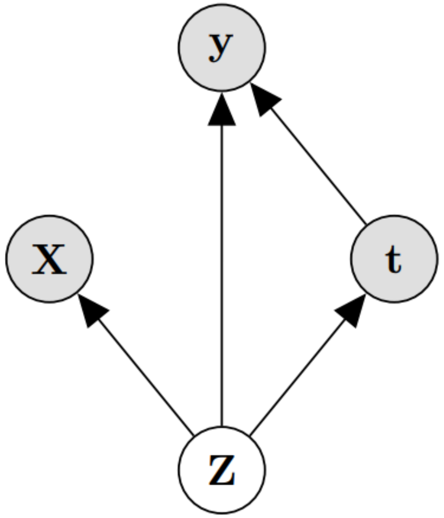
- Generative Adversarial Network(GAN) and Variational Auto-Encoder(VAE) are commonly used deep learning generative models.
- VAE defines the probability $P(Z)$ explicitly and applies variational inference.
- GAN samples data from the distribution $P(X)$ directly without defining the probability explicitly.



HIDDEN VARIABLE GENERATION

T=1

T=0



t: medical treatment

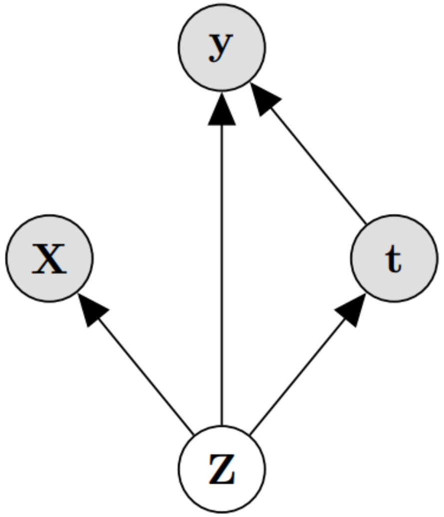
y: pesticide effect

X: co-variates

Z: latent variables e.g. gender

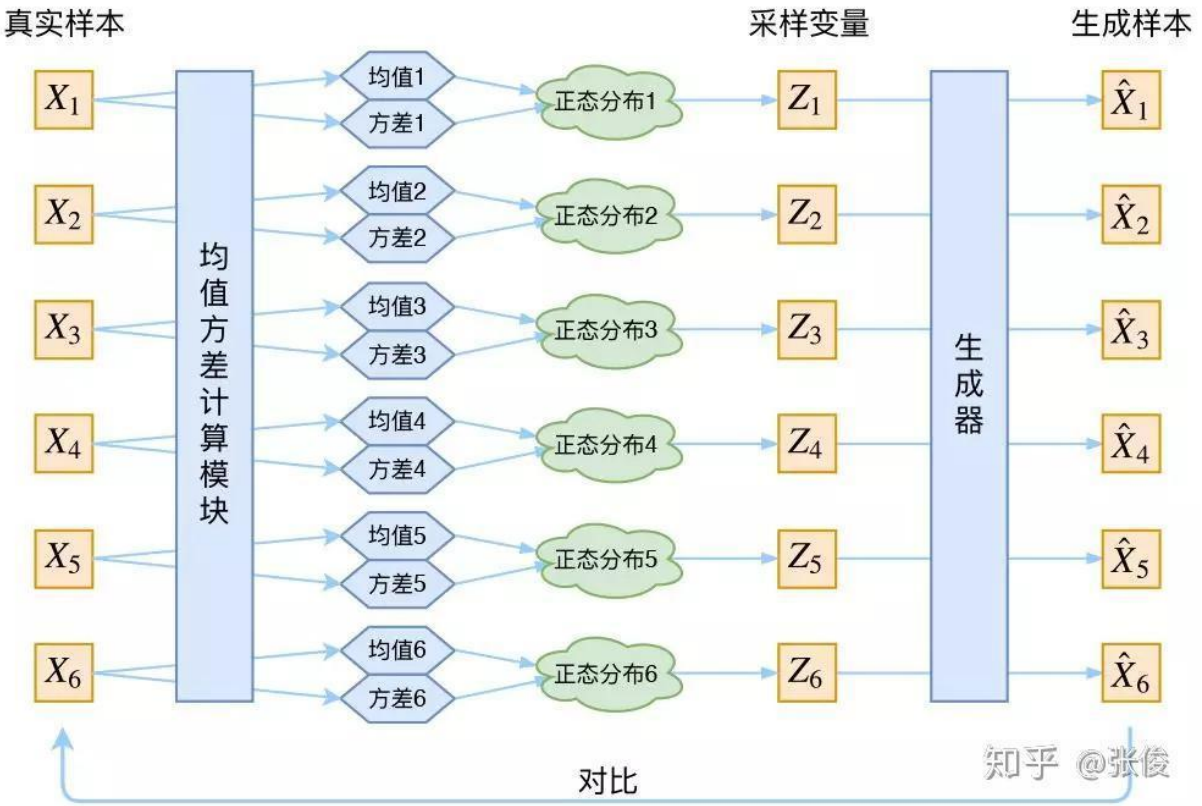
MOTIVATION FOR CEVAE

- Hidden confounder could affect the causal effect, for example, socio-economic status.
- X: observed features; Y: outcome; t: treatment; Z: unobserved features.
- We assume that $\mathrm{P}(Z, X, y, t)$ can be approximately recovered from $P(X, y, t)$.



Louizos, Christos, et al. "Causal effect inference with deep latent-variable models." Advances in Neural Information Processing Systems. 2017.

VAE

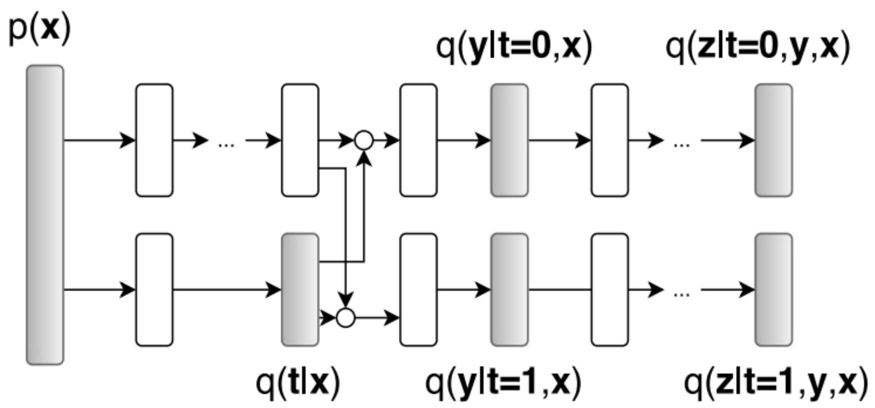


- Assumptions:
  - $P(Z_i|X_i) = \mathcal{N}(\mu_i, \sigma_i^2)$
- Variational Inference:
  - $KL\big(P(Z|X), \mathcal{N}(0, \mathrm{I})\big)$

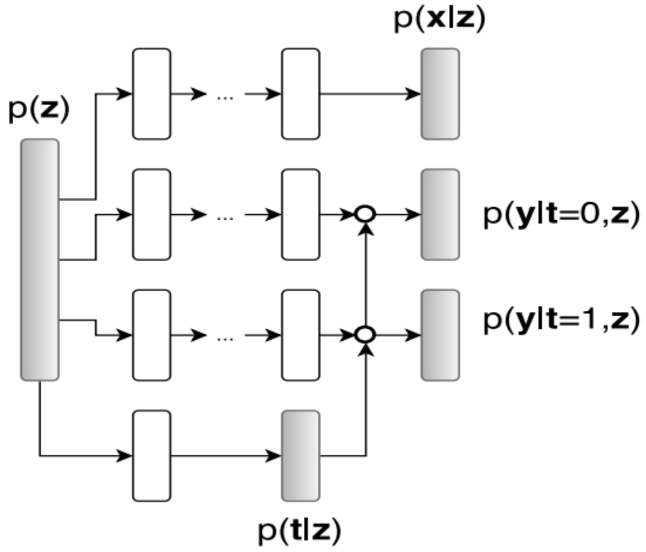
$$= \frac{1}{2}\sum_{i=1}^{d}(\sigma^2 - \log\sigma^2 - 1) + \frac{1}{2}\sum_{i=1}^{d}\mu^2$$

变分自编码器VAE：原来是这么一回事 | 附开源代码
https://zhuanlan.zhihu.com/p/34998569

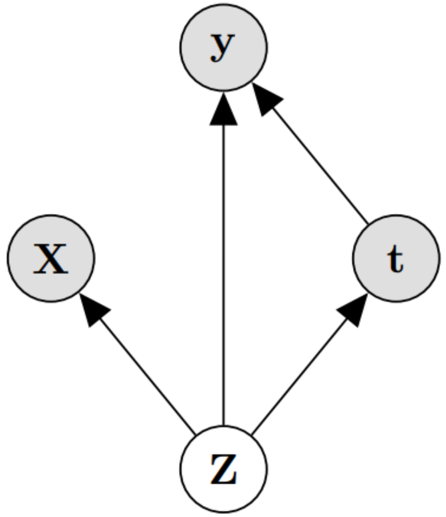
CEVAE



(a) Inference network, $q(\mathbf{z}, t, y|\mathbf{x})$.



(b) Model network, $p(\mathbf{x}, \mathbf{z}, t, y)$.



Louizos, Christos, et al. "Causal effect inference with deep latent-variable models." Advances in Neural Information Processing Systems. 2017.

CAUSALITY IN CLASSIFICATION TASKS

Grass, Yes!

Dog, Yes!

Dataset#1

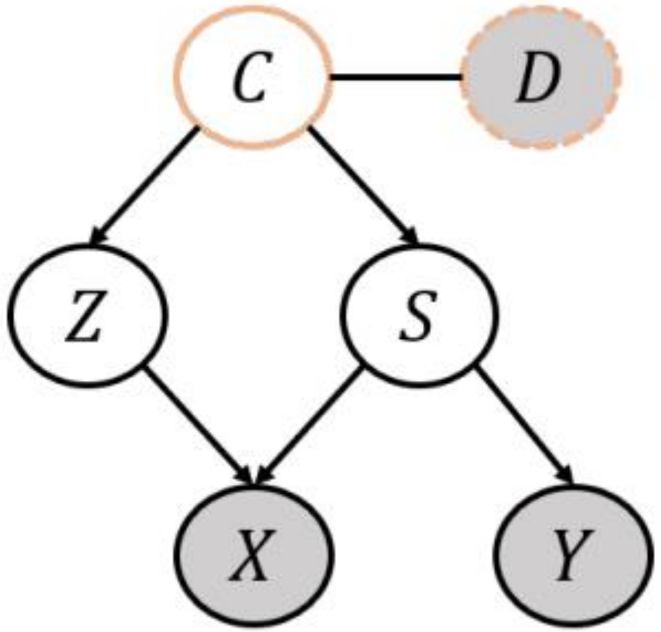
Grass, No!

Dog, Yes!

Dataset#2

MOTIVATION FOR LACIM

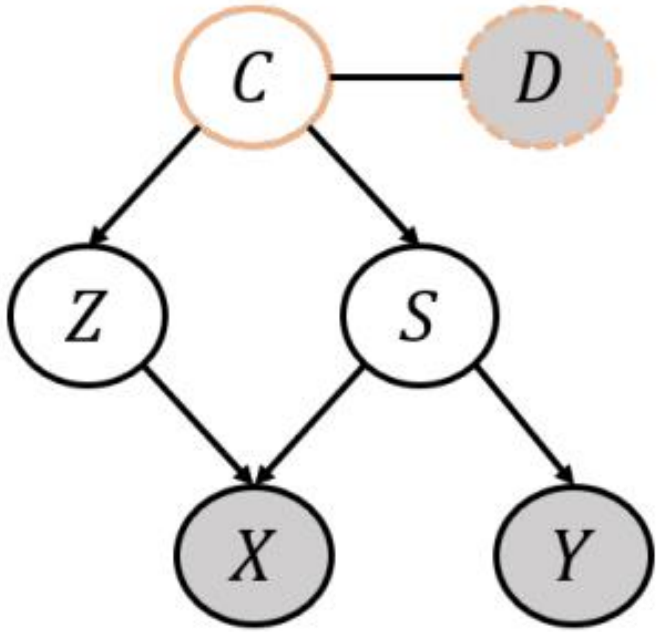
- Current supervised learning can learn spurious correlation.
  - X: picture
  - Y: label
  - Z: background(grass)
  - S: foreground(dog)
  - C: domain
  - D: index variable



Sun, Xinwei, et al. "Latent Causal Invariant Model." arXiv preprint arXiv:2011.02203 (2020).

LACIM

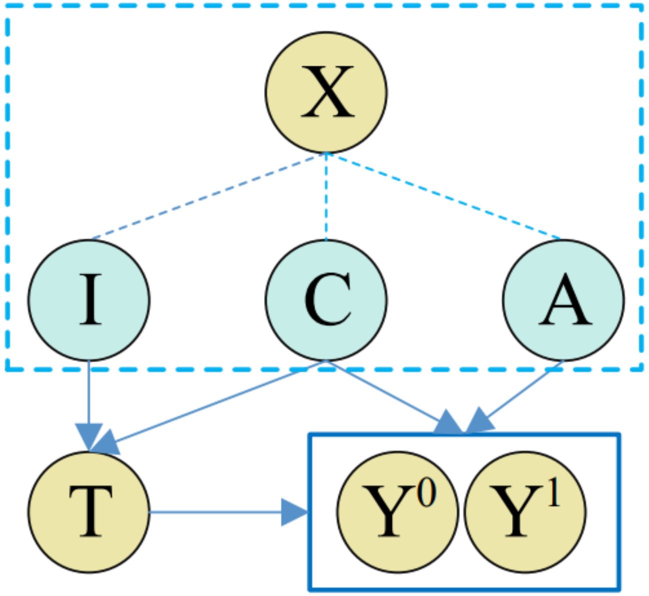
- The confounder C blocks the back-door path from Z to Y , making the Z spuriously correlated with Y.
  - S: causality connections
  - Z: spurious connections
- We denote LaCIM-C and LaCIM-D as two versions of LaCIM respectively with C observed and not. Given C or D, Z and Y would become independent.



Sun, Xinwei, et al. "Latent Causal Invariant Model." arXiv preprint arXiv:2011.02203 (2020).

DECOMPOSED REPRESENTATION

- Back-door criteria demonstrated that the controlling of the confounding factor is sufficient for removing that bias.
- I: instrumental factor I, which only affect the treatment T;
- C: confounding factor, which is the common cause of treatment T and the outcome Y;
- A: adjustment factor, which only determine the outcome Y.



Wu, Anpeng, et al. "Learning Decomposed Representation for Counterfactual Inference."

SUMMARY

- Balancing the distribution of treatment and control group.
- Generating variables representation in latent space and de-confounding.
- Disentangling the representation of features.

Q&A

THANKS

Thanks to D.A.THU, BAAI and 集智