Invariant Risk Minimization

Presenter: Zhe Wang

https://qdata.github.io/deep2Read

201909

Content

1 Introduction

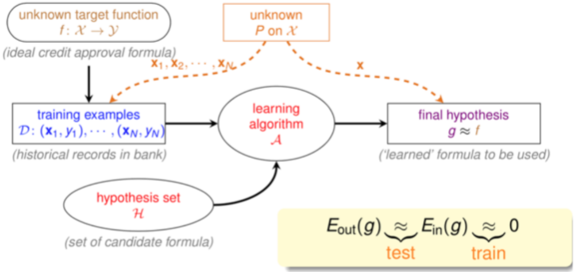
Invariant Risk Minimization

Zhe Wang

2019-9-13

Invariant Risk Minimization[1]

Unreasonable but widely-used assumptions: all training data and test data are i.i.d.



ERM principle:

$$ERM = \mathbb{E}_{e^{train}} \, l[g(x), y] \tag{1}$$

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, David Lopez-Paz

In real life? No!

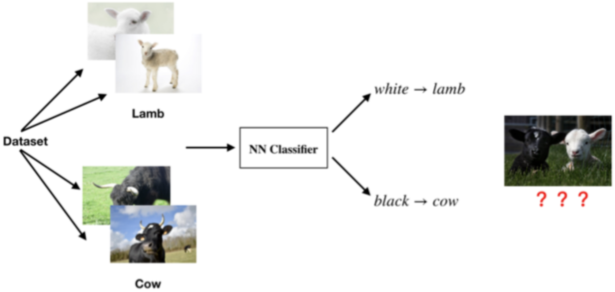
Joint data distribution:

$$P(X, Y) = P(Y|X)P(X) \tag{2}$$

Both components vary w.r.t different environments e.

Why?

correlations = spurious correlation + **causal correlation**.



X: Image, Y: {Lamb, Cow}

$$Causal\ \ correlation : \begin{cases} horn \\ fur \\ \cdots \end{cases} \qquad Spurious\ \ Correlation : \begin{cases} size \\ color \\ \cdots \end{cases}$$

Too many learned features? (Feature squeezing, Feature selection, $\cdots$)

How to separate causal correlation from spurious correlation?

Or in the language of NN, how to separate causal features from spurious features?

Invariant & Casual [2]

Intuitively,

Causal $\Longleftrightarrow$ Invariant.

- Causal reasons will always lead to the specific results, regardless of perturbation or intervention.
- If some features always accompany a phenomenon in various environments, it is reasonable to conclude they are causal features.

[2] Invariance, Causality and Robustness, Peter Bühlmann

Formally,

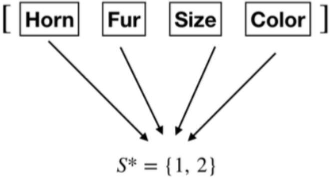
Definition (Invariant)

A subset $S^* \subset \{1, \cdots, p\}$ of variable indices, s.t. $P(Y^e | X^e_{S^*})$ remains same for all environments.

Specifically, in a linear model: $\exists S^* \subset \{1, 2, \cdots, n\}$, which is the support of β, s.t.

$$Y^e = X^e \beta_{s^*} + \epsilon^e \tag{3}$$

where $\epsilon^e \sim F^\epsilon$, F^ϵ is same for all environments.



We really care about this S^*, it is related to robustness and generalization ability.

Definition (SEMs)

$$Y \leftarrow f_Y(X_{pa(Y)}, \epsilon_Y), \quad X_j \leftarrow f_j(X_{pa(X_j)}, \epsilon_j), \tag{4}$$

where ϵ_Y, $\{\epsilon_j\}$ are all independent. $pa(Y)$, parent nodes for Y, are causal variables for Y.

Q: Under what kind of environments can we find some interesting invariant sets?

A: There are some basic assumptions:

- X and Y satisfies SEM
- perturbation doesn't perform on Y directly
- perturbation doesn't change the function f_Y

With the aforementioned assumptions, causal (parent) variables lead to invariance (Haavelmo, 1943).

causal variables $\longrightarrow$ invariant set

Reverse relation? quietly recently.

$$\begin{aligned}&\hat{S} = \bigcap_{S} \{S : \text{S passes hypothesis test of invariant with significant level } \alpha\},\\ &+\\ &\text{Gaussian Noise},\\ &\text{then, } P(\hat{S} \subset Causal(Y)) \geq 1 - \alpha.\end{aligned} \tag{5}$$

Invariant set $\longrightarrow$ causal variables

Why do we need causal correlations?

Causal correlations $\Rightarrow$ Generalization ability (or to say test sets, adversarial samples in NN)

How to learn causal correlations?

Invariant $\Rightarrow$ Causal Correlations (Explore those invariant features on training environments)

How are they related to generalization ability (robustness) ?

For linear SEM:

$$\arg\min_{\beta} \max_{e} \mathbf{E}||Y^e - X^e\beta||^2 = \beta_{S(Y)} \tag{6}$$

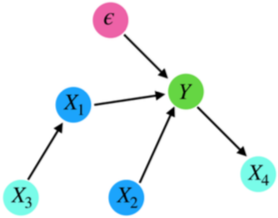
Worst case solver, robust to adversarial distributions.

Example: Regression Problem

Predict Y from $\Phi(X)$, loss function : $E||Y^e - f(\Phi(X^e))||_F^2$:

The optimal solution: $f^*(x) = \int_e yp(y|\Phi(x))dy$

If $E_{e_i}(y|\Phi(x) = h) = E_{e_j}(y|\Phi(x) = h)$, then $\Phi(x)$ elicits an invariant predictor.



Shared goal in various tasks:

$$\min_f R^{OOD}(f) = \min_f \max_{e \in \mathcal{E}_{all}} R^e(f), \tag{7}$$

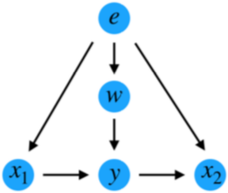
where $R^e(f) = \mathbb{E}_{X^e, Y^e}[l(f(X^e), Y^e)]$.

Method: IRM

a learning paradigm to extract nonlinear invariant predictors across multiple environments, enabling OOD generalization.

Definition:

a data representation $\Phi : \mathcal{X} \to \mathcal{H}$ elicits an invariant predictor $w \circ \Phi$ across environments $\mathcal{E}$ if there is a classifier $w : \mathcal{H} \to \mathcal{Y} : w \in \arg\min_{\bar{w}:\mathcal{H}\to\mathcal{Y}} R^e(\bar{w} \circ \Phi)$



- Y can be totally determined by $w \circ \Phi(X)$ under all environments.
- w is optimal for all environments.

Loss components:

- $\arg \min_{\Phi:\mathcal{X}\to\mathcal{H} w:\mathcal{H}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{tr}} R^e(w \circ \Phi)$,
- s.t. $w \in \arg \min_{\bar{w}:\mathcal{H}\to\mathcal{Y}} R^e(\bar{w} \circ \Phi)$ for all $e \in \mathcal{E}_{tr}$.

Mathematically,

$$Y \perp\!\!\!\perp E \mid \Phi(X) = X_1, W \tag{8}$$

While most machine learning is based on:

$$Y \perp\!\!\!\perp E \mid (X_1, X_2), W \tag{9}$$

Via Lagrange method:

$$L_{IRM}(\Phi, w) = \sum_{e \in \mathcal{E}_{tr}} R^e(w \circ \Phi) + \lambda D(w, \Phi, e) \tag{10}$$

By the normal equation:

$$w_\Phi^e = \mathbb{E}_{X^e}[\Phi(X^e)'\Phi(X^e)]^{-1}\mathbb{E}_{X^e,Y^e}[\Phi(X^e)'Y^e] \tag{11}$$

- $D(w, \Phi, e) = ||w - w_\Phi^e||^2$. Containing the inverse, can be discontinuous.
- $D(w, \Phi, e) = ||\mathbb{E}_{X^e}[\Phi(X^e)'\Phi(X^e)]w - \mathbb{E}_{X^e,Y^e}[\Phi(X^e)'Y^e]||^2$. Smooth and differentiable.

Over-parametrized:

Fix the "dummy" linear classifier $\tilde{w} = 1.0$

In general cases:

$$\min_{\Phi:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{tr}} R^e(\Phi) + \lambda||\nabla_{w|w=1.0} R^e(w\cdot\Phi)||^2. \tag{12}$$

Implement details:

For the square of gradient norm, a unbiased estimation is

$$\sum_{k=1}^{b}[\nabla_{w|w=1.0} l(w\cdot\Phi(X_k^{e,i}), Y_k^{e,i})\cdot\nabla_{w|w=1.0} l(w\cdot\Phi(X_k^{e,j}), Y_k^{e,j})], \tag{13}$$

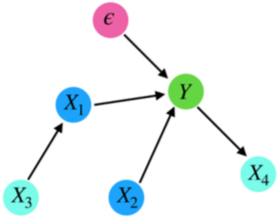
where, $X^{e,i}$ and $X^{e,j}$ are two random batches sampled from environment e.

Example: Regression Problem

Predict Y from $\Phi(X)$, loss function : $E||Y^e - f(\Phi(X^e))||_F^2$:

The optimal solution: $f^*(x) = \int_e yp(y|\Phi(x))dy$

If $E_{e_i}(y|\Phi(x) = h) = E_{e_j}(y|\Phi(x) = h)$, then $\Phi(x)$ elicits an invariant predictor.

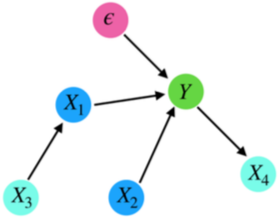


Example: Regression Problem

Predict Y from $\Phi(X)$, loss function : $E||Y^e - f(\Phi(X^e))||_F^2$:

The optimal solution: $f^*(x) = \int_e yp(y|\Phi(x))dy$

If $E_{e_i}(y|\Phi(x) = h) = E_{e_j}(y|\Phi(x) = h)$, then $\Phi(x)$ elicits an invariant predictor.



low error + invariance across $\mathcal{E}_{tr}$ = low error + invariance across $\mathcal{E}_{all}$??

Assumption: all environments share the same underlying Structural Equations.

Definition

A SEM $\mathcal{C} := (\mathcal{S}, \mathcal{N})$ governing the random vector $X = (X_1, X_2, \cdots, X_d)$ is a set of structural equations:

$$\mathcal{S}_i : X_i \leftarrow f_i(Pa(X_i), N_i) \tag{14}$$

Acyclic causal graph.

An intervention e on $\mathcal{C}$ consists of replacing some of its structural equations via manipulating the noise variable N_i

An intervention $e \in \mathcal{E}_{all}(\mathcal{C})$ is considered to be valid: the causal graph remains acyclic, $\mathbf{E}[Y^e \mid Pa(Y)] = \mathbf{E}[Y \mid Pa(Y)]$, $\mathbf{V}[Y^e \mid Pa(Y)]$ remains finite.

Generalization

Require some degree of diversity across $\mathcal{E}_{tr}$

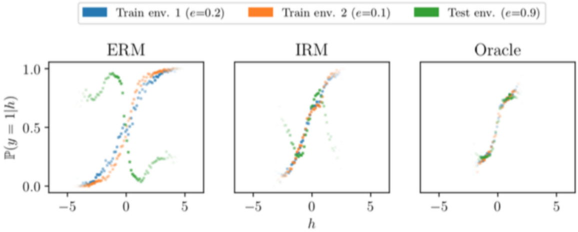
- Diversity + invariant across $\mathcal{E}_{tr}$ = invariant across $\mathcal{E}_{all}$. Basically, these environments span a high dimensional space. (These environments can't be co-linear).
- Low error across $\mathcal{E}_{tr}$ + invariant across $\mathcal{E}_{all}$ = low error across $\mathcal{E}_{all}$

Experiments on MNIST

- digits $0 \sim 4$ is assigned with label $y = 0$, others with label $y = 1$.
- flip the label with 25% probability.
- color the image.
- flip the color with a probability depends on environment $10\%, 20\%$ for train and 90% for test.

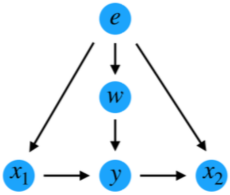
Algorithm	Acc. train envs.	Acc. test env.
ERM	86.57	14.56
IRM (ours)	70.93	**66.10**
Random guessing (hypothetical)	50	50
Optimal invariant model (hypothetical)	75	75
ERM, grayscale model (oracle)	73.52	72.90

Table 1: Accuracy (%) of different algorithms on the Colored MNIST synthetic task.



Relation to domain adaptation

Domain Adaptation, especially for adversarial DA.



- $Y \perp\!\!\!\perp E \,||\, \Phi(X)$
- $w \in \arg \min_{\bar{w} \in \mathcal{H} \to \mathcal{Y}} R^s(w \cdot f)$

Learn wrong kinds of invariant correlations.