Domain Generalization :

Domain-invariant Representation Learning

Data Mining & Quality Analytics Lab.

2024. 01. 19

발표자: 정진용

발표자 소개

❖ 정진용 (Jinyong Jeong)

- 고려대학교 산업경영공학과 석·박사 통합과정(2021.09~)
- Data Mining & Quality Analytics Lab. (김성범 교수님)

❖ 관심 연구 분야

- Domain Generalization
- Semi-Supervised Learning & Class-Imbalanced Semi-Supervised Learning

❖ E-mail

- jy_jeong@korea.ac.kr

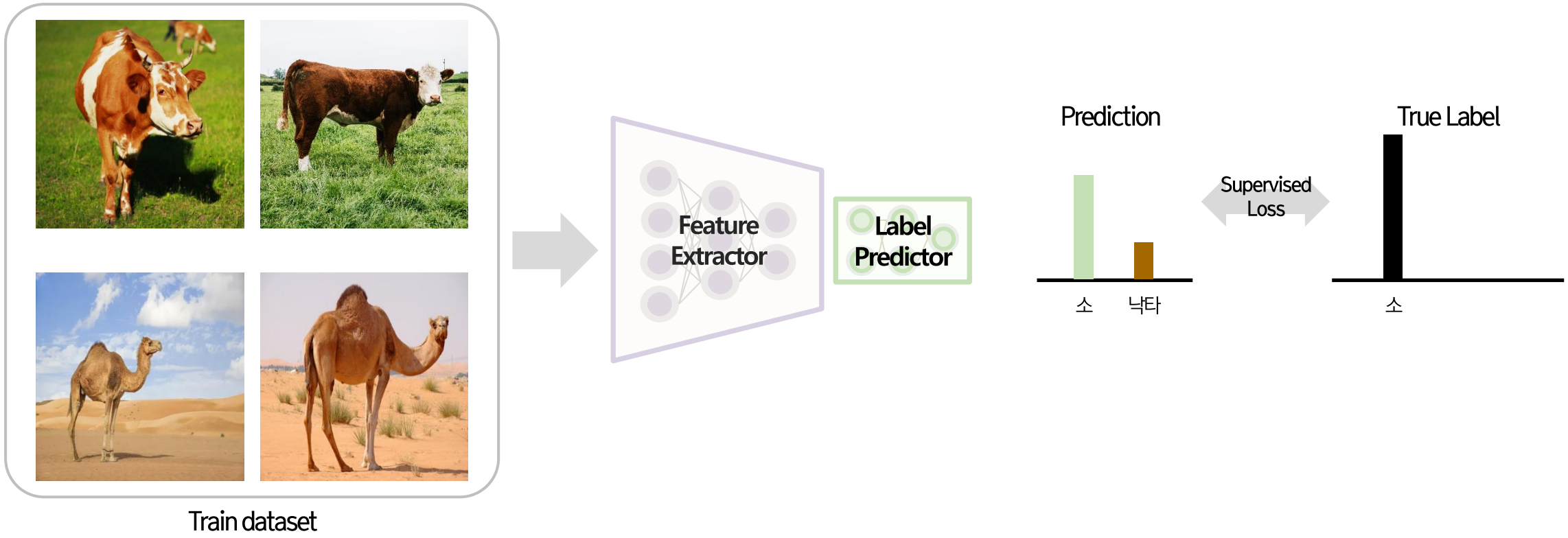
목차

1. Introduction: Domain Generalization

Introduction

Empirical Risk Minimization

❖ 지도학습을 사용하여 분류 모델($f: X \longrightarrow Y$)을 만들어보자



Introduction

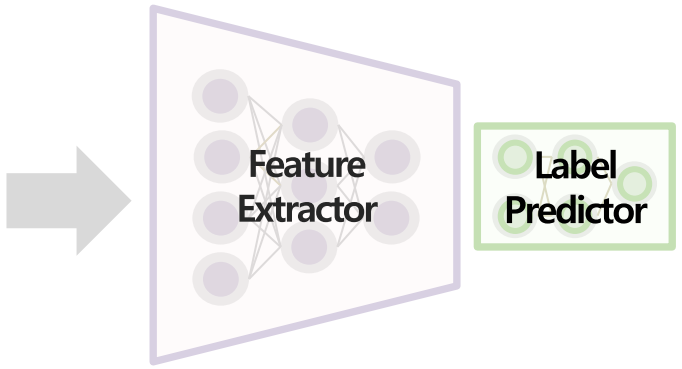
Empirical Risk Minimization

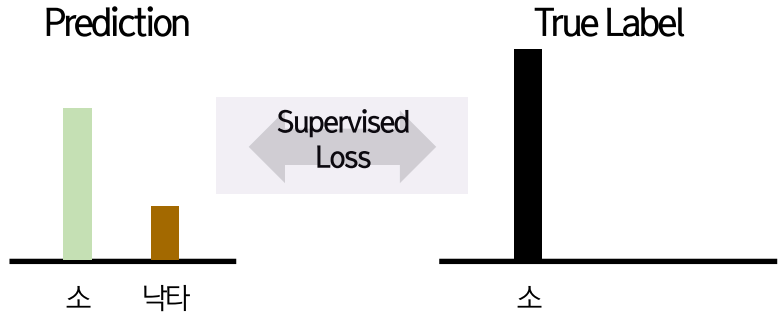
❖ 지도학습을 사용하여 분류 모델($f: X \longrightarrow Y$)을 만들어보자

$$R(f) := \mathbb{E}_{(X,Y)\sim P(X,Y)}[L(f(X), Y)]$$

학습 목표 $$f^* = \underset{f \in \mathcal{F}}{argmin}\, R(f)$$

Train dataset





Introduction

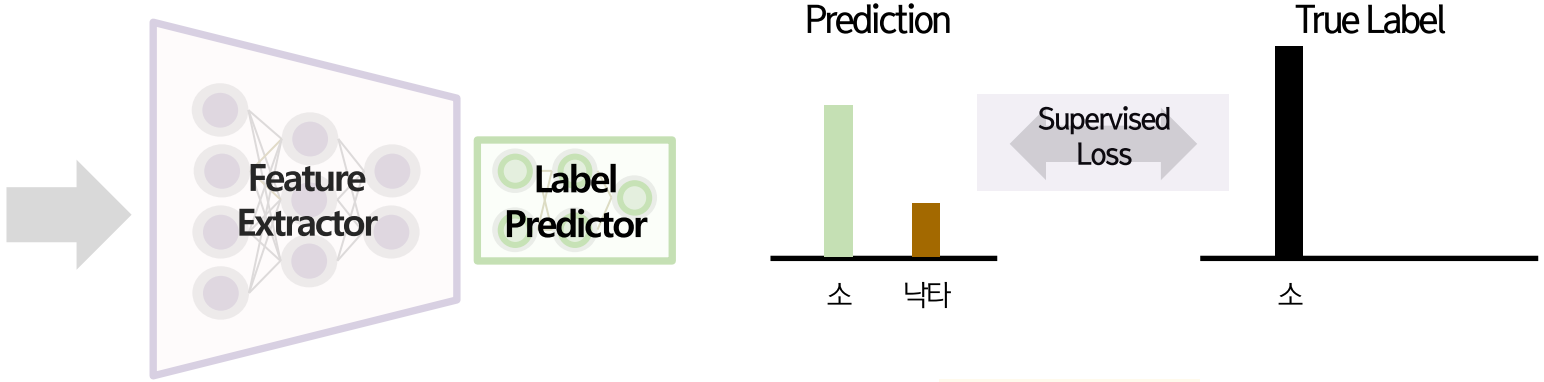
Empirical Risk Minimization

❖ 지도학습을 사용하여 분류 모델($f: X \longrightarrow Y$)을 만들어보자

$$R(f) := \mathbb{E}_{(X,Y)\sim P(X,Y)}[L(f(X), Y)]$$

학습 목표 $$f^* = \underset{f \in \mathcal{F}}{argmin}\, R(f)$$

Train dataset



실제로는 X, Y에 대한 확률 모분포 P(X, Y)를 알 수 없음

Introduction

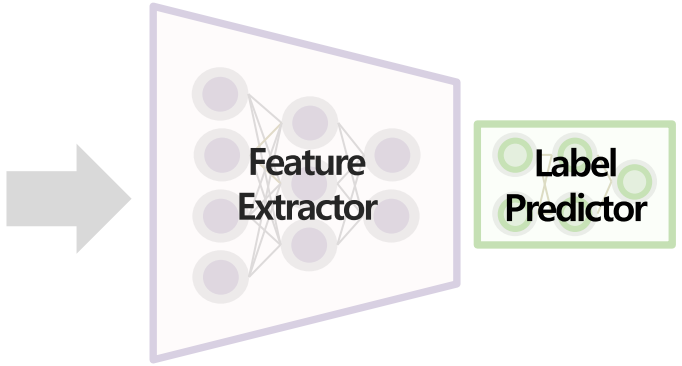
Empirical Risk Minimization

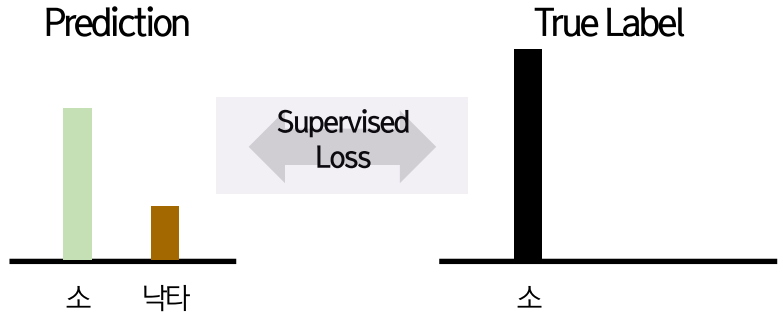
❖ 지도학습을 사용하여 분류 모델($f: X \longrightarrow Y$)을 만들어보자

$$R(f) := \mathbb{E}_{(X,Y)\sim P(X,Y)}[L(f(X), Y)]$$

학습 목표 $$f^* = \underset{f \in \mathcal{F}}{argmin}\, R(f)$$

Train dataset





실제로는 X, Y에 대한 확률 모분포 P(X, Y)를 알 수 없음
→ P(X, Y)에 속하는 데이터들을 샘플링

Introduction

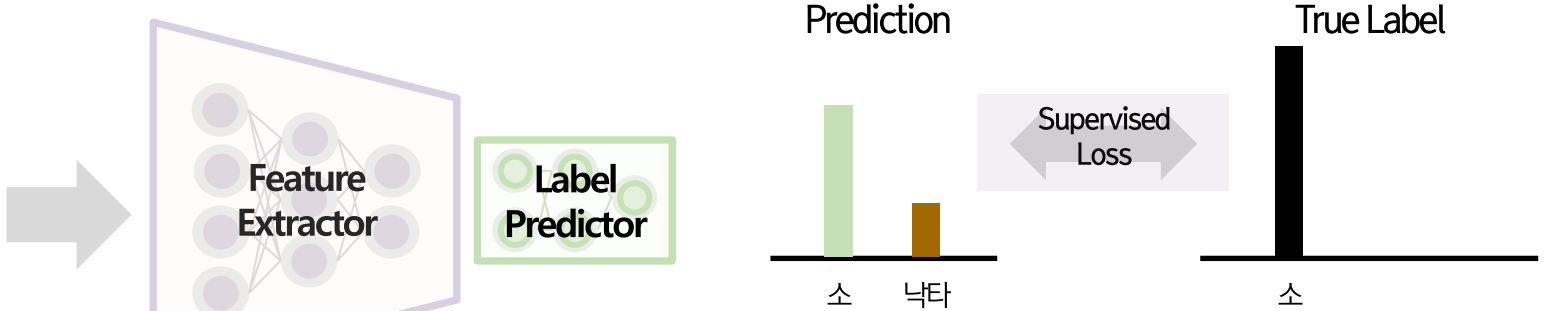
Empirical Risk Minimization

❖ 지도학습을 사용하여 분류 모델($f: X \longrightarrow Y$)을 만들어보자

$$R(f) := \mathbb{E}_{(X,Y)\sim P(X,Y)}[L(f(X), Y)]$$

학습 목표 $$f^* = \underset{f\in\mathcal{F}}{argmin}\, R(f)$$

Train dataset



실제로는 X, Y에 대한 확률 모분포 P(X, Y)를 알 수 없음
→ P(X, Y)에 속하는 데이터들을 샘플링

$$Empirical\ R(f) = \mathbb{E}_{X,Y\sim\hat{P}_{train}(X,Y)}[L(f(X), Y)] = \frac{1}{n}\sum_{i=1}^{n} L(f(x_i, y_i)$$

Introduction

Empirical Risk Minimization

❖ 학습된 분류 모델($f: X \longrightarrow Y$)을 사용하여 추론을 해보자

$\hat{P}_{train}(X, Y)$에 의해서 학습된 f

$\hat{P}_{train}(X, Y) \approx \hat{P}_{test}(X, Y)$

Train dataset

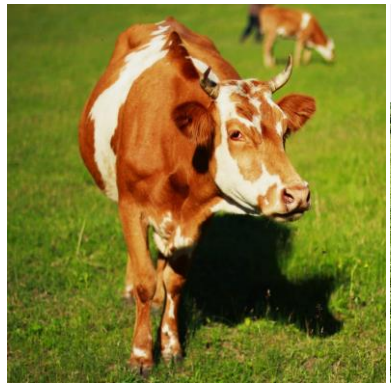
Test dataset

Introduction

Empirical Risk Minimization

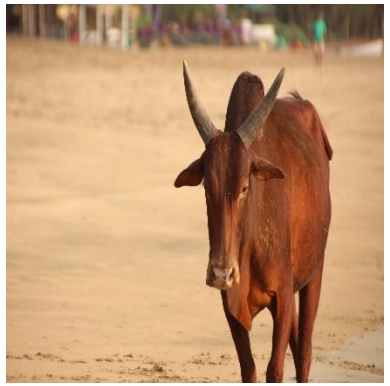
❖ 학습된 분류 모델($f: X \longrightarrow Y$)을 사용하여 추론을 해보자

$\hat{P}_{train}(X, Y)$에 의해서 학습된 f



Train dataset

$\hat{P}_{test}(X, Y)$



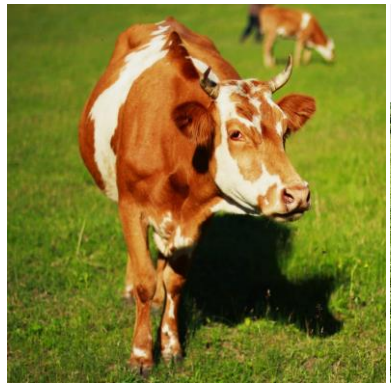
Test dataset

Introduction

Empirical Risk Minimization

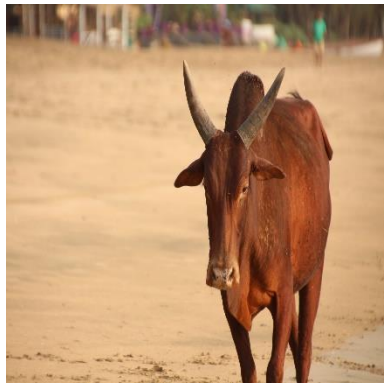
❖ 학습된 분류 모델($f: X \longrightarrow Y$)을 사용하여 추론을 해보자

$\hat{P}_{train}(X, Y)$에 의해서 학습된 f



Train dataset

$\hat{P}_{test}(X, Y)$



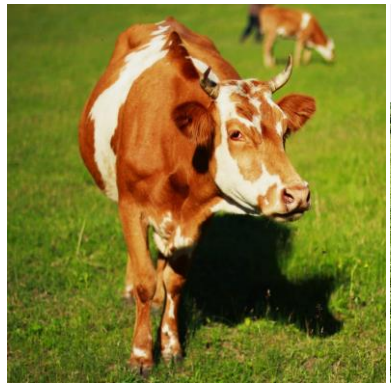
Test dataset

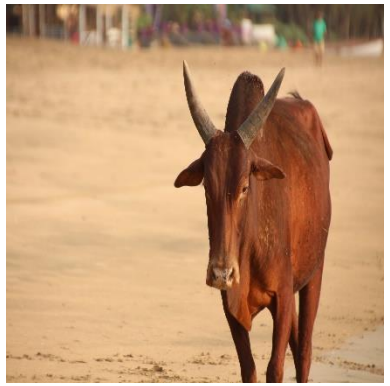
Introduction

Empirical Risk Minimization

❖ 학습된 분류 모델($f: X \longrightarrow Y$)을 사용하여 추론을 해보자

$\hat{P}_{train}(X, Y)$에 의해서 학습된 f



Train dataset

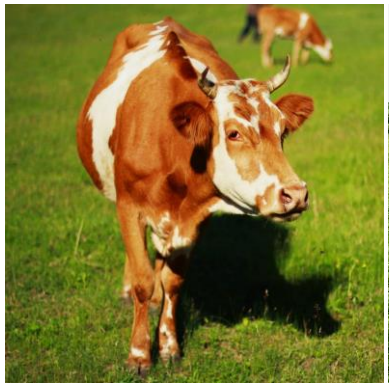
Test dataset

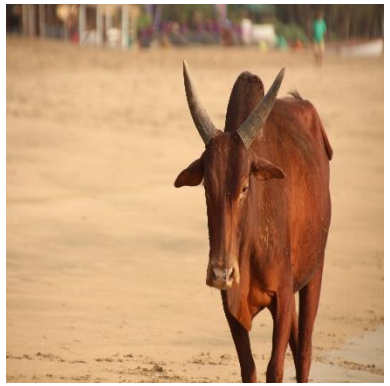
Introduction

Empirical Risk Minimization

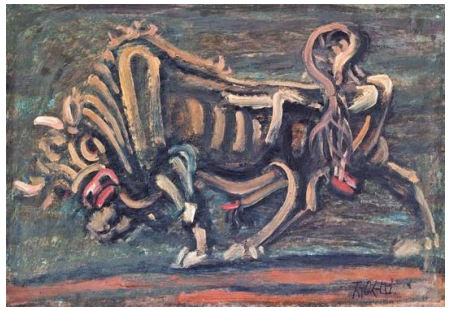
❖ 학습된 분류 모델($f: X \longrightarrow Y$)을 사용하여 추론을 해보자

$\hat{P}_{train}(X, Y)$에 의해서 학습된 f



Train dataset

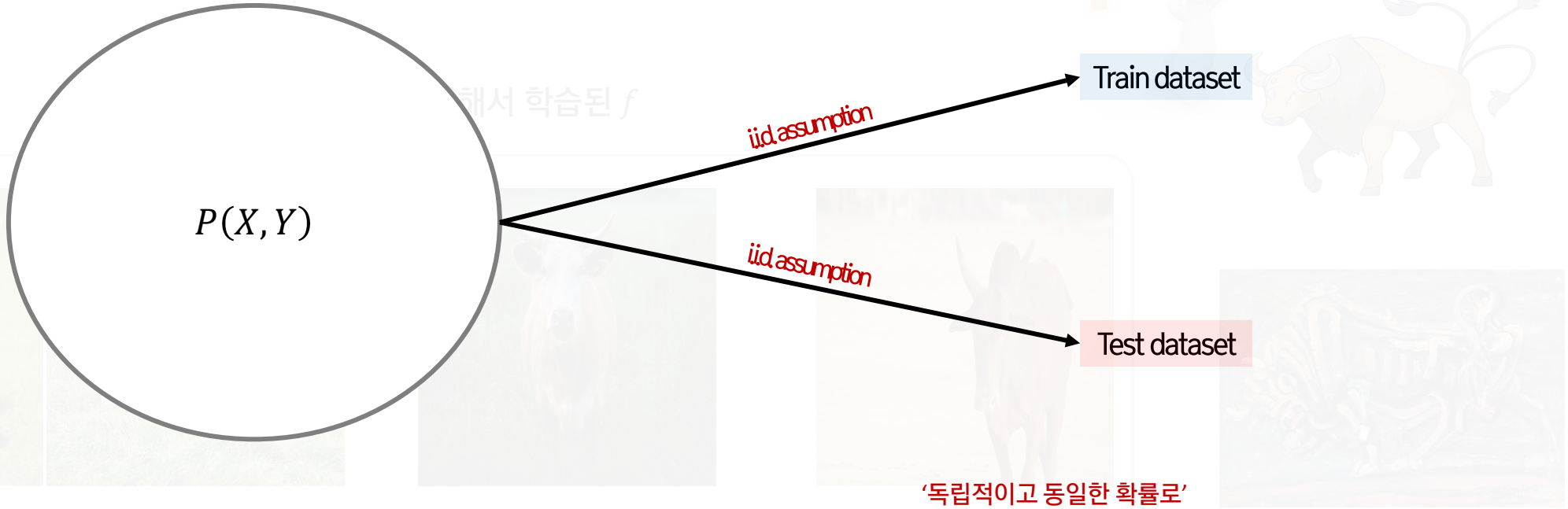


Test dataset

Introduction

Empirical Risk Minimization

❖ 학습된 분류 모델($f: X \longrightarrow Y$)을 사용하여 추론을 해보자



'독립적이고 동일한 확률로'

기존 머신러닝 및 딥러닝 학습에서는 train dataset과 test dataset이 같은 분포에서 샘플링 되었다는 가정이 있음

→ 만약 i.i.d 가정이 깨진다면, Empirical Risk Minimization을 통해 학습된 모델은 예측 성능이 저하될 수 있음

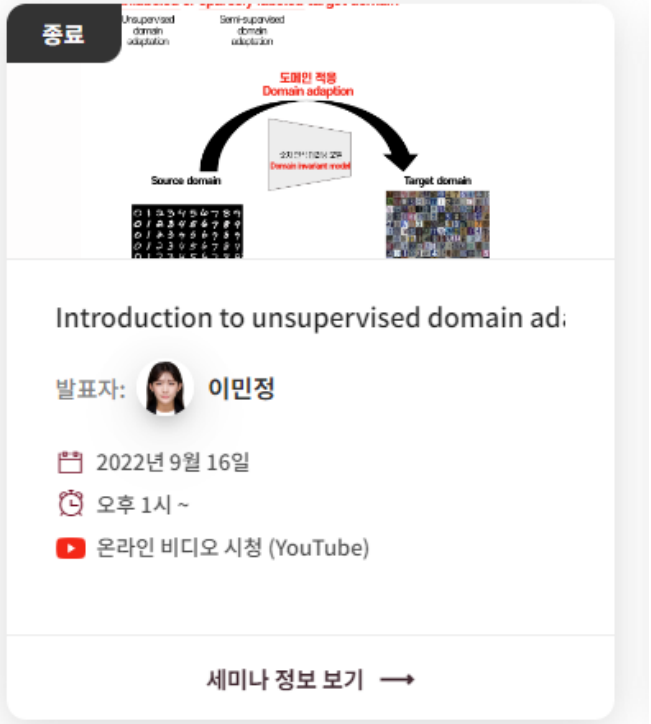
'Distribution shift'

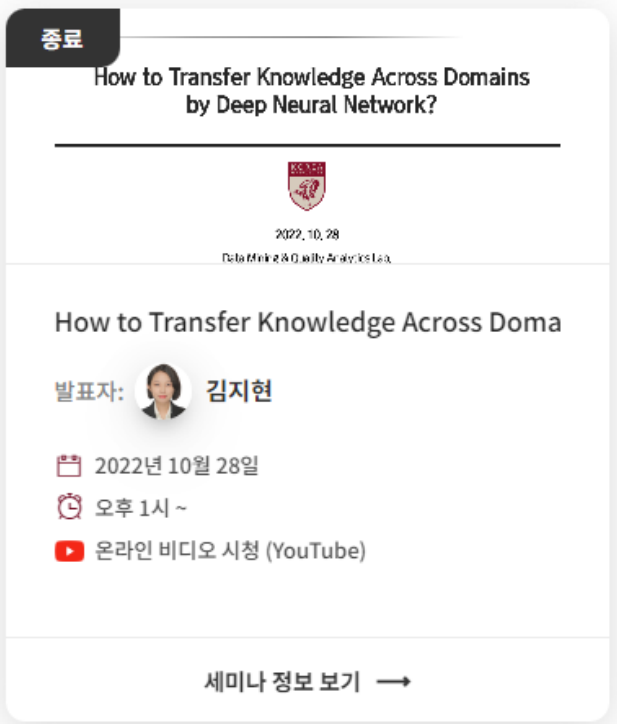
Introduction

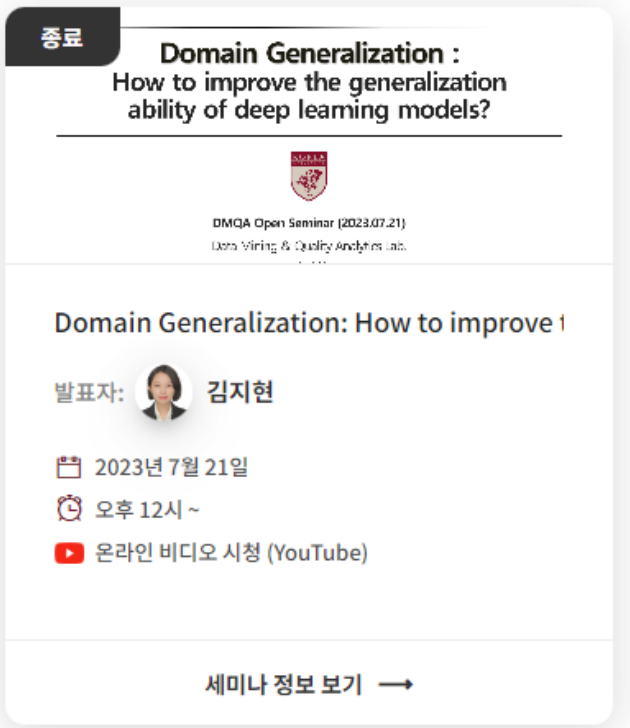
Background of Domain Generalization

❖ Distribution shift 상황에서도 모델 일반화 성능을 향상 시킬 수 있는 다양한 연구들이 존재

- Test data에 대한 성능을 높이자!





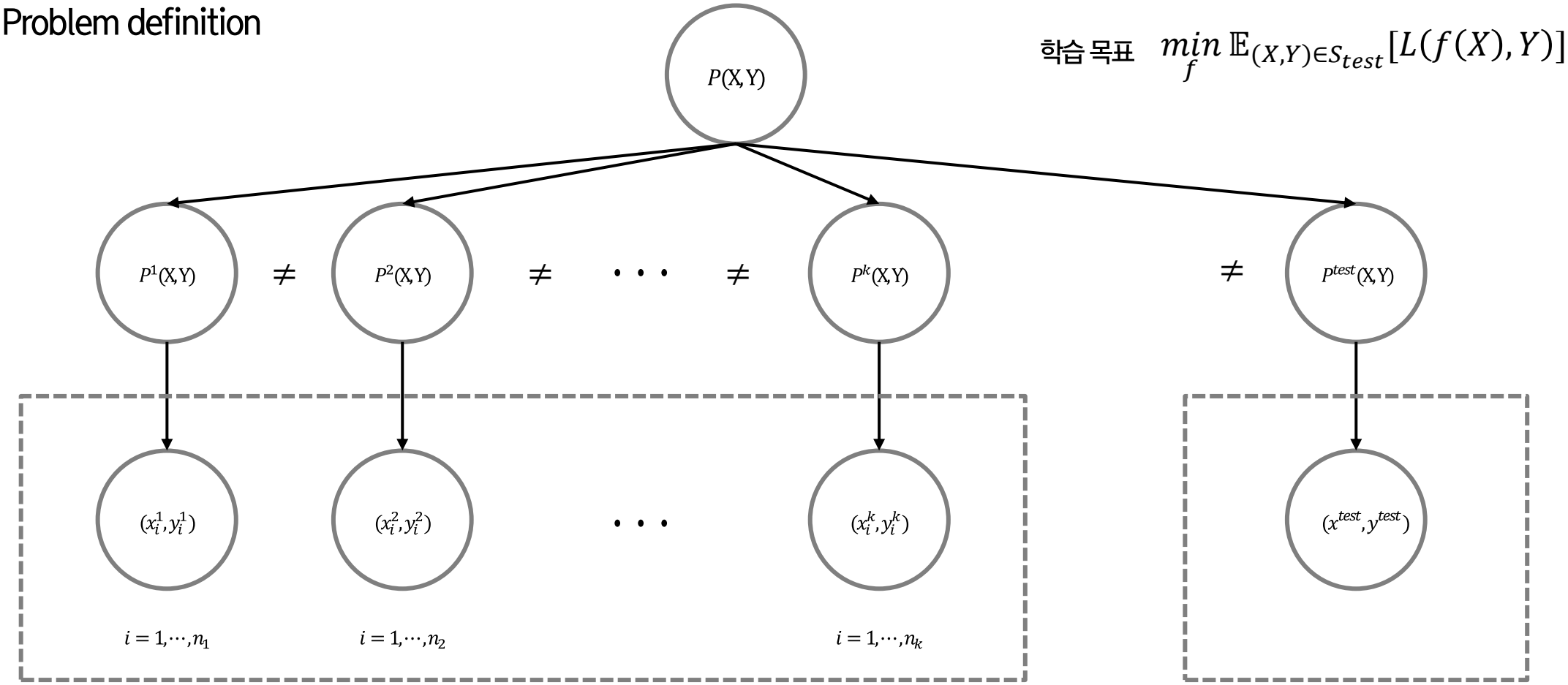


Introduction

Background of Domain Generalization

❖ Problem definition

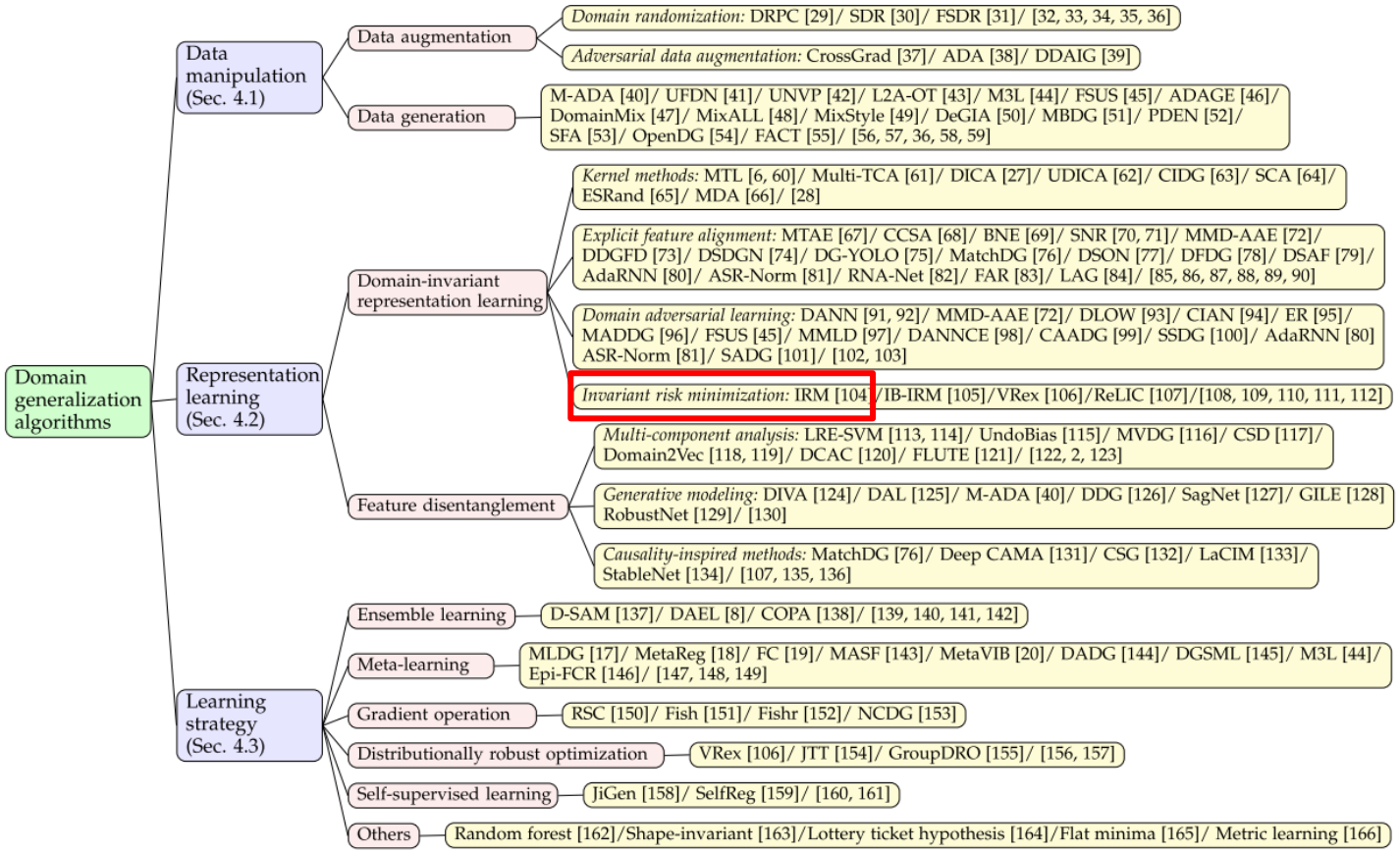
학습 목표 $\min_f \mathbb{E}_{(X,Y)\in S_{test}}[L(f(X),Y)]$



Introduction

Background of Domain Generalization

❖ Taxonomy of domain generalization

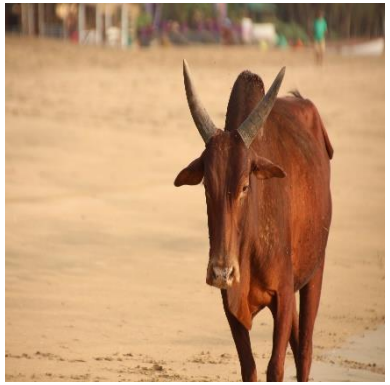


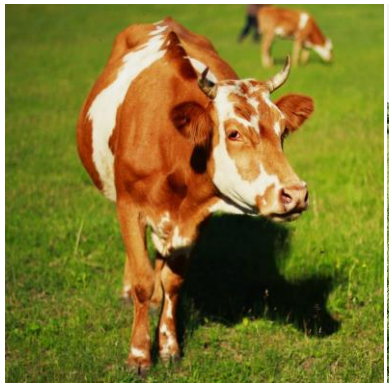
Domain-invariant representation learning

Invariant Risk Minimization

❖ Why is it a cow?

- '소' 라는 것과 'green or glassy background' 는 실제로 연관이 없음
  - ➢ Spurious correlation, not stable property
- '소의 형태' 와 같은 invariant correlations을 배운다면 새로운 도메인에도 적용될 것임





Train domains

Test domains

Domain-invariant representation learning

Invariant Risk Minimization

❖ Invariant risk minimization (arxiv, 2019)

- Google Deepmind, Facebook AI, Stanford, Meta AI 연구원들에 의해 연구되었으며, 2024년 1월 19일 기준 1,681회 인용됨
- Domain-invariant representation learning에서 새로운 학습 패러다임을 제시한 논문

Invariant Risk Minimization

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, David Lopez-Paz

1 Introduction

Machine learning suffers from a fundamental problem. While machines are able to learn complex prediction rules by minimizing their training error, data are often marred by selection biases, confounding factors, and other peculiarities [49, 48, 23]. As such, machines justifiably inherit these data biases. This limitation plays an essential role in the situations where machine learning fails to fulfill the promises of artificial intelligence. More specifically, minimizing training error leads machines into recklessly absorbing all the correlations found in training data. Understanding which patterns are useful has been previously studied as a correlation-versus-causation dilemma, since spurious correlations stemming from data biases are unrelated to the causal explanation of interest [31, 27, 35, 52]. Following this line, we leverage tools from causation to develop the mathematics of spurious and invariant correlations, in order to alleviate the excessive reliance of machine learning systems on data biases, allowing them to generalize to new test distributions.

As a thought experiment, consider the problem of classifying images of cows and camels [4]. To address this task, we label images of both types of animals. Due to a selection bias, most pictures of cows are taken in green pastures, while most pictures of camels happen to be in deserts. After training a convolutional neural network on this dataset, we observe that the model fails to classify easy examples of images of cows when they are taken on sandy beaches. Bewildered, we later realize that our neural network successfully minimized its training error using a simple cheat: classify green landscapes as cows, and beige landscapes as camels.

To solve the problem described above, we need to identify which properties of the training data describe spurious correlations (landscapes and contexts), and which properties represent the phenomenon of interest (animal shapes). Intuitively, a correlation is spurious when we do not expect it to hold in the future in the same manner as it held in the past. In other words, spurious correlations do not appear to be stable properties [54]. Unfortunately, most datasets are not provided in a form amenable to discover stable properties. Because most machine learning algorithms depend on the assumption that training and testing data are sampled independently from the same distribution [51], it is common practice to shuffle at random the training and testing examples. For instance, whereas the original NIST handwritten data was collected from different writers under different conditions [19], the popular MNIST training and testing sets [8] were carefully shuffled to represent similar mixes of writers. Shuffling brings the training and testing distributions closer together, but

Domain-invariant representation learning

Invariant Risk Minimization

❖ Invariant predictor와 invariant correlation을 간단한 수식으로 확인

- Example : Structural equation model

$$X_1 \leftarrow Gaussian(0, \sigma^2) \quad \text{(noise)}$$

$$Y \leftarrow X_1 + Gaussian(0, \sigma^2) \quad \text{(noise)}$$

$$X_2 \leftarrow Y + Gaussian(0,1) \quad \text{(fixed noise)}$$

Noisy observation of Y

$e \in \mathcal{E}_{tr} = \{replace\ \sigma^2\ by\ 10, replace\ \sigma^2\ by\ 20\}$

이때, ~~environment~~ domain e에 대해서 Least squares predictor $\hat{Y}^e = X_1^e\hat{\alpha}_1 + X_2^e\hat{\alpha}_2$를 사용하여 Y를 예측하면?

- Regress from X_1^e : $\hat{\alpha}_1 = 1,\ \hat{\alpha}_2 = 0$
- Regress from X_2^e : $\hat{\alpha}_1 = 0,\ \hat{\alpha}_2 = \sigma(e)^2/(\sigma(e)^2 + \frac{1}{2})$
- Regress from (X_1^e, X_2^e) : $\hat{\alpha}_1 = 1/(\sigma(e)^2 + 1),\ \hat{\alpha}_2 = \sigma(e)^2/(\sigma(e)^2 + 1)$

regress from X_2

$$SSR = \sum (Y - \hat{\alpha}_1 X_1 - \hat{\alpha}_2 X_2)^2$$

$$\frac{\partial}{\partial \hat{\alpha}_2} = -2\sum (Y - \hat{\alpha}_2 X_2) X_2 = 0$$

$$\hat{\alpha}_2 = \frac{\sum X_2 Y}{\sum (X_2)^2} = \frac{Cov(X_2, Y)}{Var(X_2)}$$

$$Cov(X_2, Y) = E[(X_2 - E[X_2])(Y - E[Y])]$$
$$= E[(X_2 - 0)(Y - 0)]$$
$$= E[Y^2 + (Y \times N(0,1))] \quad (N(0,1) = 0)$$
$$= E[Y^2]$$
$$= Var(Y) \quad (\because E[Y] = 0)$$

$$Var(Y) = Var(X_1) + var(N(0, \sigma^2)) = 2 \cdot \sigma^2$$

$$Var(X_2) = Var(Y) + Var(N(0,1))$$
$$= 2\sigma^2 + 1$$

$$\therefore \hat{\alpha}_2 = \frac{\sigma^2}{\sigma^2 + \frac{1}{2}}$$

Domain-invariant representation learning

Invariant Risk Minimization

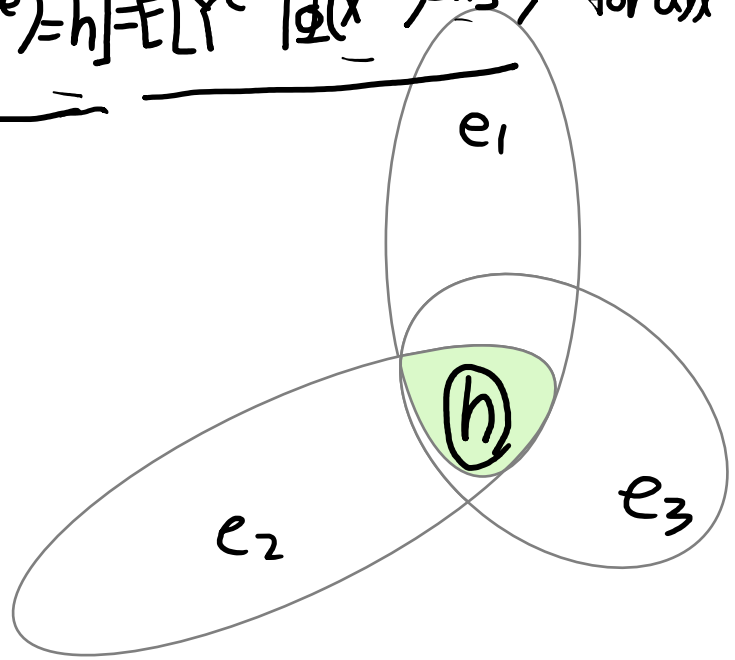
❖ Empirical data로부터 invariances를 얻는 원리

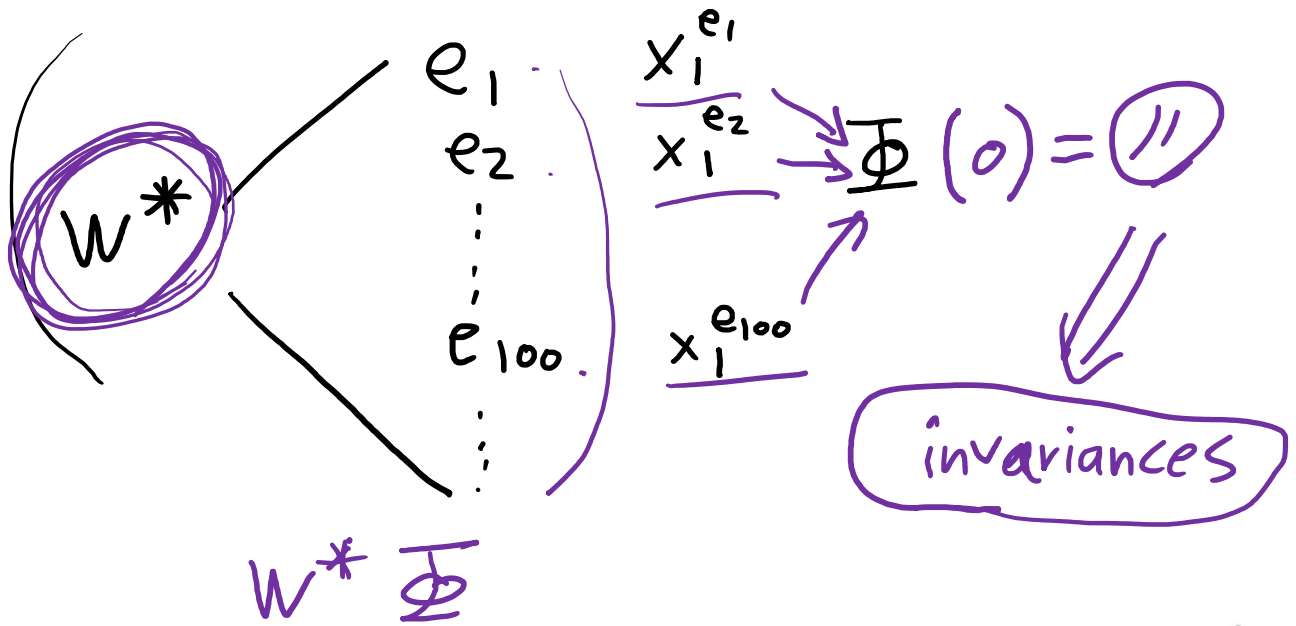
Definition 3. *We say that a data representation $\Phi : \mathcal{X} \to \mathcal{H}$ elicits an invariant predictor $w \circ \Phi$ across environments $\mathcal{E}$ (if there is a classifier $w : \mathcal{H} \to \mathcal{Y}$ simultaneously optimal for all environments, that is, $w \in \arg\min_{\bar{w}:\mathcal{H}\to\mathcal{Y}} R^e(\bar{w} \circ \Phi)$ for all $e \in \mathcal{E}$.)*

MSE, cross-ent.

$E[Y^e | \Phi(X^e) = h] = E[Y^{e'} | \Phi(X^{e'}) = h]$, for all $e, e' \in \mathcal{E}$

Φ





Domain-invariant representation learning

Invariant Risk Minimization

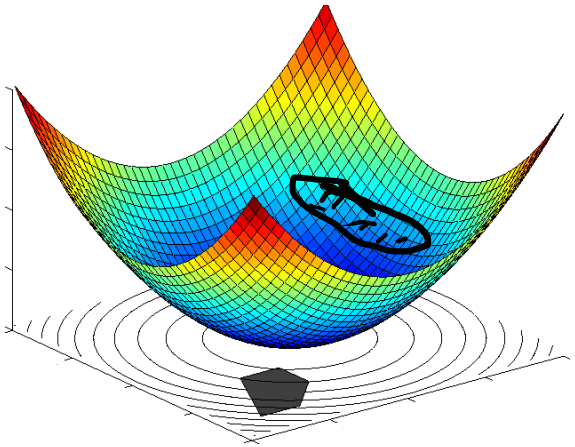
❖ Empirical data로부터 invariances를 얻는 원리

Definition 3. *We say that 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* $w : \mathcal{H} \to \mathcal{Y}$ *simultaneously optimal for all environments, that is,* $w \in \arg\min_{\bar{w}:\mathcal{H}\to\mathcal{Y}} R^e(\bar{w} \circ \Phi)$ *for all* $e \in \mathcal{E}$.

$$\min_{\substack{\Phi:X\to\mathcal{H} \\ w:H\to Y}} \sum_{e\in\mathcal{E}_{tr}} R^e(w \circ \Phi)$$

subject to $w \in argmin_{\bar{w}:\mathcal{H}\to Y} R^e(\bar{w} \circ \Phi)$ for all $e \in \mathcal{E}_{tr}$

(IRM)



(Inner routine) : 고정된 Φ로부터 모든 학습 환경에서 loss의 기대값이 가장 작은 w를 찾고,

(Outer routine) : 현재 모든 학습 환경에서 최적인 w를 사용해서 Φ를 업데이트

Domain-invariant representation learning

Invariant Risk Minimization

❖ Empirical data로부터 invariances를 얻는 원리

Definition 3. *We say that 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* $w : \mathcal{H} \to \mathcal{Y}$ *simultaneously optimal for all environments, that is,* $w \in \arg\min_{\bar{w}:\mathcal{H}\to\mathcal{Y}} R^e(\bar{w} \circ \Phi)$ *for all* $e \in \mathcal{E}$.

$$\min_{\substack{\Phi:X\to\mathcal{H} \\ w:H\to Y}} \quad \sum_{e\in\mathcal{E}_{tr}} R^e(w \circ \Phi)$$

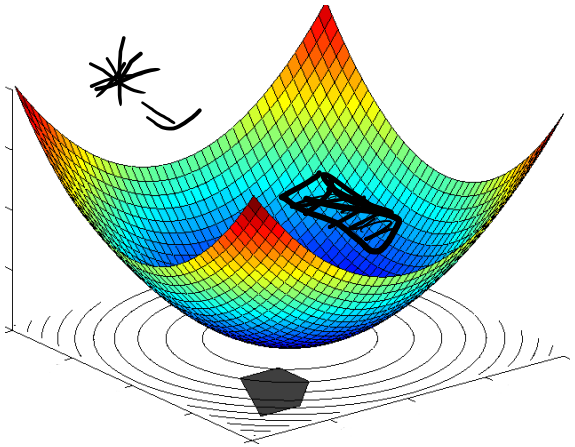
$$\text{subject to} \quad w \in argmin_{\bar{w}:\mathcal{H}\to Y} R^e(\bar{w} \circ \Phi) \text{ for all } e \in \mathcal{E}_{tr}$$

(IRM)

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

ERM과 penalty를 조절하는 hyper-parameter

(IRMv1)



Domain-invariant representation learning

Invariant Risk Minimization

$$\min_{\substack{\Phi:X\to\mathcal{H}\\ w:H\to Y}} \sum_{e\in\mathcal{E}_{tr}} R^e(w\circ\Phi) \quad \text{ERM}$$

$$\text{subject to} \quad w\in \operatorname{argmin}_{\bar{w}:\mathcal{H}\to Y} R^e(\bar{w}\circ\Phi) \text{ for all } e\in\mathcal{E}_{tr} \quad \text{(IRM)}$$

❖ From (IRM) to (IRMv1)

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

Soft regularization term

w^e_Φ

Classifier w를 linear로 가정 ✔

Fixed Φ, $w^e_\Phi \in \operatorname{argmin}_{\bar{w}} R^e(w\circ\Phi)$: $w^e_\Phi = \mathbb{E}_{X^e}[\Phi(X^e)\Phi(X^e)^T]^{-1}\,\mathbb{E}_{X^e,Y^e}[\Phi(X^e)Y^e]$

$$\mathbb{D}_{dist}(w,\Phi,e) = \|w - w^e_\Phi\|^2$$

고정된 Φ에서의 optimal classifier w^e_Φ와 선택된 w 사이의 거리

[Normal equations]

$$\frac{\partial}{\partial w}(Xw-Y)^T(Xw-Y)$$

$$= 2X^TXw - 2X^TY = 0.$$

$$\therefore \quad X^TXw = X^TY$$

$$\therefore w = (X^TX)^{-1}(X^TY)$$

Domain-invariant representation learning

Invariant Risk Minimization

이때, environment e에 대해서 Least squares predictor $\hat{Y}^e = X_1^e\hat{\alpha}_1 + X_2^e\hat{\alpha}_2$를 사용하여 Y를 예측하면?

- Regress from X_1^e : $\hat{\alpha}_1 = 1,\ \hat{\alpha}_2 = 0$
- Regress from X_2^e : $\hat{\alpha}_1 = 0,\ \hat{\alpha}_2 = \sigma(e)^2/(\sigma(e)^2 + \frac{1}{2})$
- Regress from (X_1^e, X_2^e) : $\hat{\alpha}_1 = 1/(\sigma(e)^2 + 1),\ \hat{\alpha}_2 = \sigma(e)^2/(\sigma(e)^2 + 1)$

$Y = 1 \cdot X_1 + 0 X_2$

Example 1에서 w = (1, 0), data representation $\Phi = \begin{bmatrix} 1 & 0 \\ 0 & c \end{bmatrix}$ 추가

$Y = [X_1 \ X_2] \begin{bmatrix} 1 & 0 \\ 0 & c \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

$\Phi(X^e) = \begin{bmatrix} 1 & 0 \\ 0 & c \end{bmatrix} \begin{bmatrix} X_1^e \\ X_2^e \end{bmatrix} = \begin{bmatrix} X_1^e \\ c \cdot X_2^e \end{bmatrix}$

$\Phi(X^e)\Phi(X^e)^T = \begin{bmatrix} X_1^e \\ c \cdot X_2^e \end{bmatrix} \begin{bmatrix} (X_1^e)^T & (c \cdot X_2^e)^T \end{bmatrix} = \begin{bmatrix} (X_1^e)(X_1^e)^T & c\sim \\ c\sim & c^2\sim \end{bmatrix}$

$$\mathbb{D}_{dist}(w, \Phi, e) = \|w - w_\Phi^e\|^2$$

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

$\mathbb{D}_{dist}$는 c = 0인 지점에서 불연속

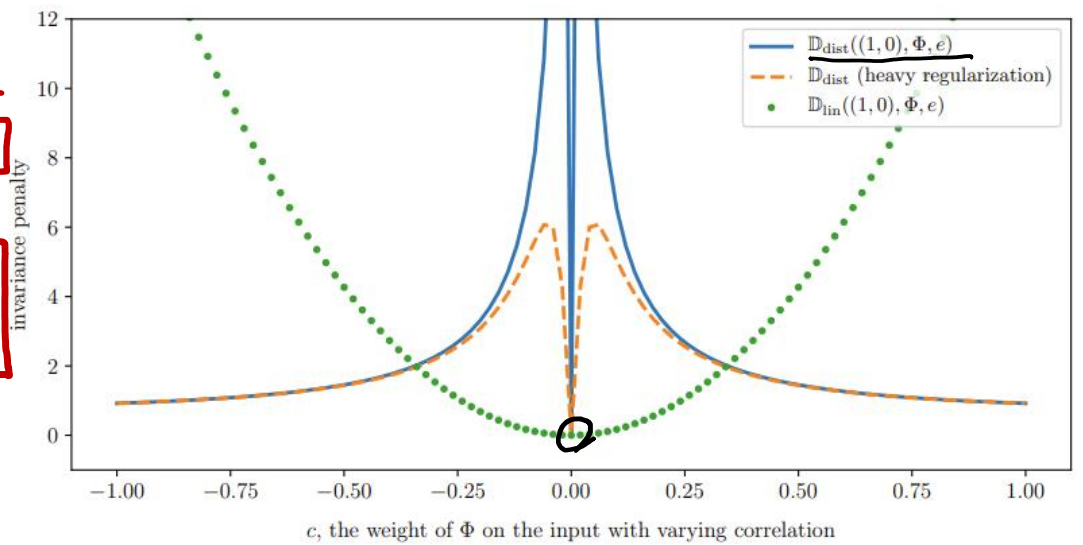


Figure 1: Different measures of invariance lead to different optimization landscapes in our Example 1. The naïve approach of measuring the distance between optimal classifiers $\mathbb{D}_{\text{dist}}$ leads to a discontinuous penalty (solid blue unregularized, dashed orange regularized). In contrast, the penalty $\mathbb{D}_{\text{lin}}$ does not exhibit these problems.

$$\mathbb{D}_{lin}(w, \Phi, e) = \left\|\mathbb{E}_{X^e}[\Phi(X^e)\Phi(X^e)^T]w - \mathbb{E}_{X^e,Y^e}[\Phi(X^e)Y^e]\right\|^2$$

$X^T X W = 2 X^T Y$

⇓

정규방정식 위반하는 정도

Domain-invariant representation learning

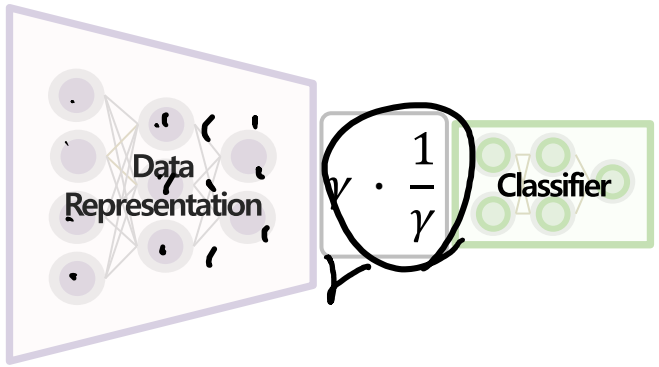
$$L_{IRM}(\Phi, w) = \sum_{e \in \mathcal{E}_{tr}} (R^e(w \circ \Phi) + \lambda \cdot \mathbb{D}_{lin}(w, \Phi, e))$$

Invariant Risk Minimization

❖ Over-parameterized model과 linear classifier w 고정

- Over-parameterized model → $\mathbb{D}_{lin} \approx 0$ 가능할 수 있음
- 따라서 classifier w를 고정한 뒤, data representation Φ만 활용하여 invariant predictor 구축

$$w \circ \Phi = (\gamma\Phi, \frac{1}{\gamma} w)$$



차수

$$\mathbb{D}_{lin}(w, \Phi, e) = \left\| \mathbb{E}_{X^e}[\Phi(X^e)\Phi(X^e)^T] w - \mathbb{E}_{X^e, Y^e}[\Phi(X^e) Y^e] \right\|^2$$

Over-parameterized model에서는 가중치들이 작은 값을 가져도 표현을 할 수 있음

Domain-invariant representation learning

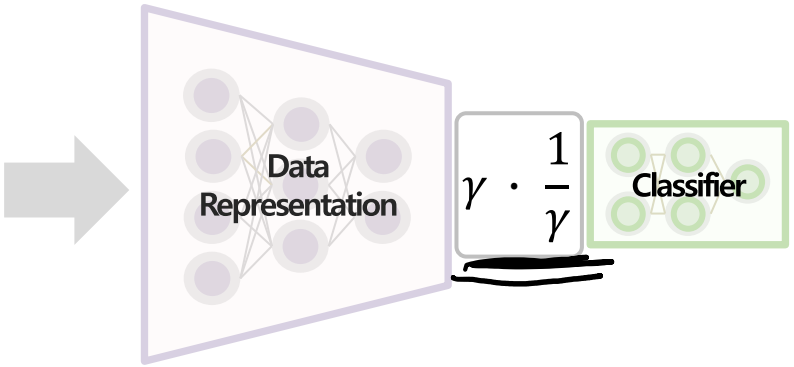
$$L_{IRM}(\Phi, w) = \sum_{e \in \mathcal{E}_{tr}} (R^e(w \circ \Phi) + \lambda \cdot \mathbb{D}_{lin}(w, \Phi, e))$$

Invariant Risk Minimization

❖ Over-parameterized model과 linear classifier w 고정

- Over-parameterized model → $\mathbb{D}_{lin} \approx 0$ 가능할 수 있음
- 따라서 classifier w를 고정한 뒤, data representation Φ만 활용하여 invariant predictor 구축

$$w \circ \Phi = (\gamma\Phi, \frac{1}{\gamma} w)$$



Domain-invariant representation learning

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

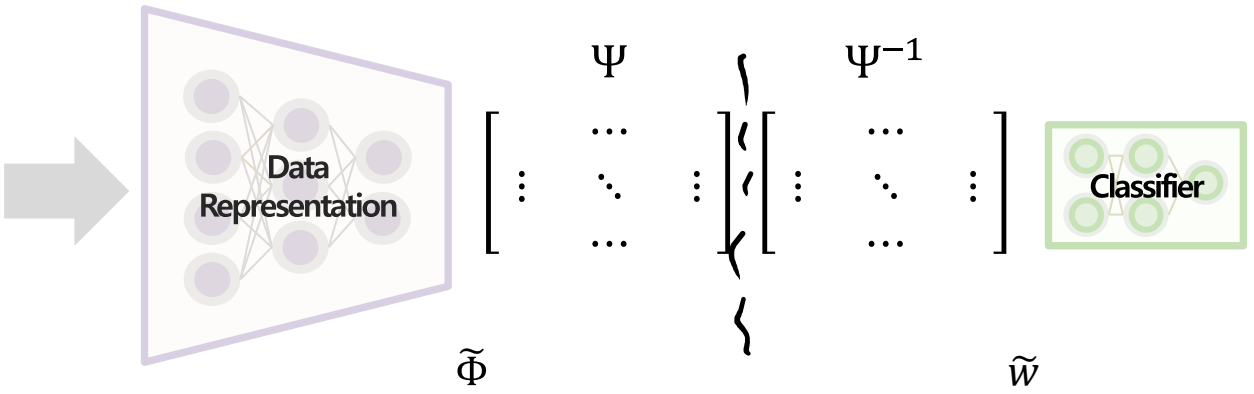
Invariant Risk Minimization

❖ Over-parameterized model과 linear classifier w 고정

- Over-parameterized model → $\mathbb{D}_{lin} \approx 0$ 가능할 수 있음
- 따라서 classifier w를 고정한 뒤, data representation Φ만 활용하여 invariant predictor 구축

$$(w \circ \Psi^{-1}) \circ (\Psi \circ \Phi)$$

$$w \circ \Phi = (\gamma\Phi, \frac{1}{\gamma}w)$$



1. w를 고정해서, $\mathbb{D}_{lin}(w, \Phi, e) \approx 0$을 조금 개선할 수 있음.
2. 모든 e에 항상 '고정' W.

고정된 $\tilde{w}$가 모든 environments에 대해서 optimal classifier가 될 수 있는 Φ를 찾음으로써 invariant predictor 구축

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

Domain-invariant representation learning

Invariant Risk Minimization

❖ $\tilde{w}$를 scalar값으로 고정시켜도 invariance를 학습할 수 있음

Theorem 4. *For all $e \in \mathcal{E}$, let $R^e : \mathbb{R}^d \to \mathbb{R}$ be convex differentiable cost functions. A vector $v \in \mathbb{R}^d$ can be written $v = \Phi^\top w$, where $\Phi \in \mathbb{R}^{p \times d}$, and where $w \in \mathbb{R}^p$ simultaneously minimize $R^e(w \circ \Phi)$ for all $e \in \mathcal{E}$, if and only if $v^\top \nabla R^e(v) = 0$ for all $e \in \mathcal{E}$. Furthermore, the matrices Φ for which such a decomposition exists are the matrices whose nullspace $\mathrm{Ker}(\Phi)$ is orthogonal to v and contains all the $\nabla R^e(v)$.*

$w^* \in \underset{\bar{w} \in \mathbb{R}}{\arg\min}\, R^e(\bar{w} \circ \Phi), \forall e$

$\frac{\partial}{\partial w} R^e(\Phi^T w) = \Phi \nabla R^e(\Phi^T w) = \Phi \nabla R^e(v) = 0$

$w^T \Phi \nabla R^e(v) = w^T \cdot 0 = \vec{0}$

$\therefore v^T \nabla R^e(v) = 0$

$w \to 1.0$

$\Phi \to$ non-linear

$$\min_{\Phi: X \to Y} \sum_{e \in \mathcal{E}_{tr}} R^e(\Phi) + \lambda \cdot \left\| \nabla_{w|w=1.0} R^e(w \cdot \Phi) \right\|^2 \quad \text{(IRMv1)}$$

Domain-invariant representation learning

Invariant Risk Minimization

❖ 훈련 도메인에서 학습한 invariance를 모든 도메인으로 확장

- 필요 조건 1 : 훈련 도메인이 충분한 다양성을 가지고 있어야 함
- 필요 조건 2 : 다양한 도메인이 invariance를 가지고 있어야 함

Assumption 8. *A set of training environments $\mathcal{E}_{tr}$ lie in* linear general position *of degree r if $|\mathcal{E}_{tr}| > d - r + \frac{d}{r}$ for some $r \in \mathbb{N}$, and for all non-zero $x \in \mathbb{R}^d$,*

훈련 도메인 수

$$\dim\left(\text{span}\left(\left\{\mathbb{E}_{X^e}\left[X^e X^{e\top}\right]x - \mathbb{E}_{X^e,\epsilon^e}\left[X^e\epsilon^e\right]\right\}_{e\in\mathcal{E}_{tr}}\right)\right) > d - r.$$

span

훈련 도메인으로부터 생성된 벡터 공간의 차원

Conclusion

❖ Summary

- Empirical risk minimization은 학습과 테스트 데이터가 같은 분포에서 i.i.d. 샘플링 되었다는 가정이 필요함
- 이러한 가정이 깨져서 발생하는 두 데이터셋 사이의 distribution shift 상황에서 ERM은 좋지 못한 성능을 보임
- Distribution shift 상황에서도 테스트 데이터에 대한 성능을 개선할 수 있는 domain generalization 소개
  - ➢ 서로 분포가 다른 다양한 도메인에서 invariance correlation을 학습하는 방법론
  - ➢ Invariance를 복잡한 최적화 상황에서 그래디언트 기반 relaxed version 제안

고맙습니다