ラベルの階層性を用いた不変学習

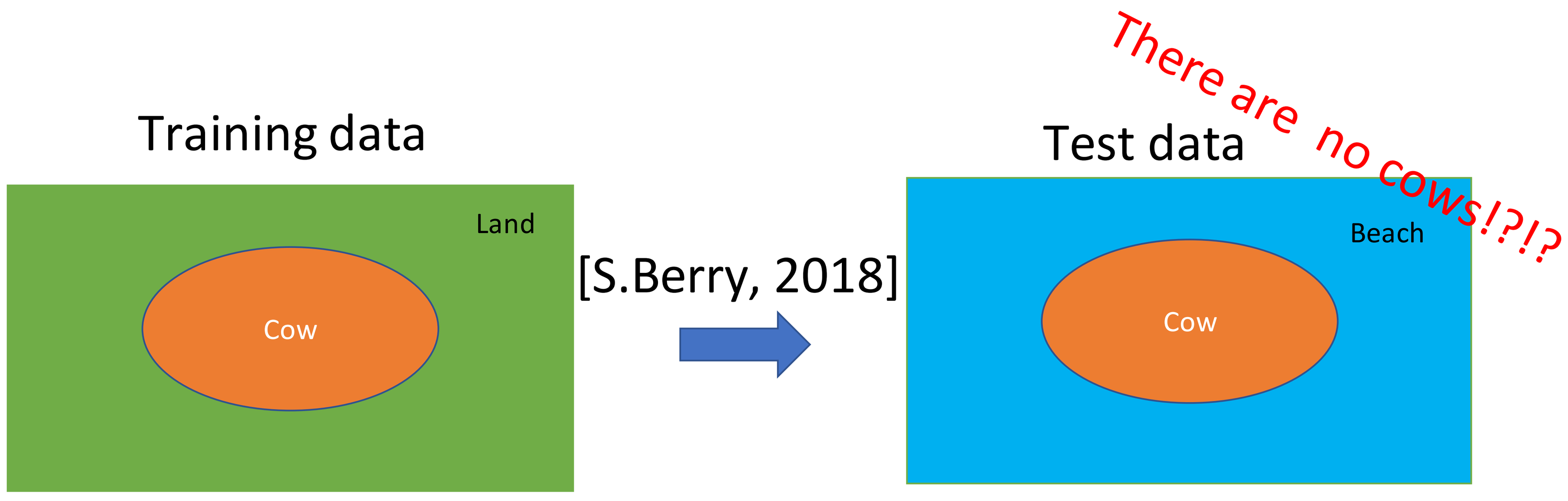
豊田　祥史　総合研究大学院大学 統計科学専攻　博士課程(三年次編入)　五年

環境外の汎化と不変学習

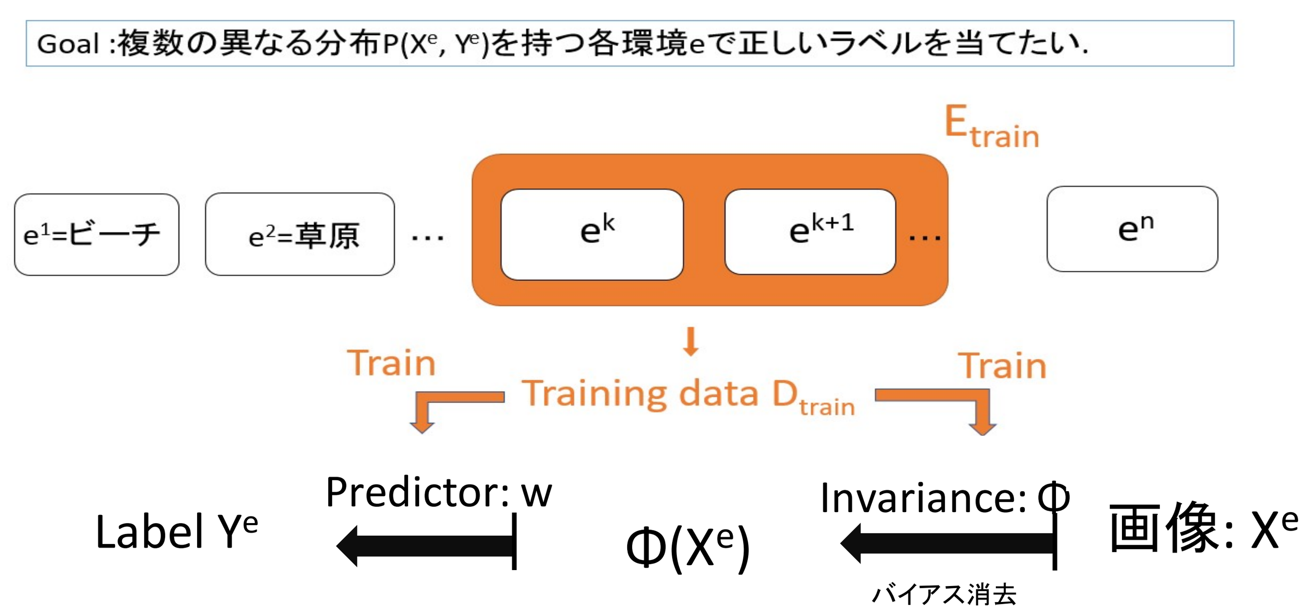
はじめに

近年の機械学習分野の重要な課題: 環境の変更により汎化性能が劣化

例: 学習時に草原を背景とした牛の画像を用いて教師あり学習.
➡ビーチを背景とした牛の画像の牛の識別精度が劣化[S. Berry, 2018]



$e \in \mathcal{E}$: 背景を表す index, $X^e \in \mathcal{X}, Y^e \in \mathcal{Y}$: 背景 e での画像とそのラベル.
(X^e, Y^e) は $e \in \mathcal{E}$ に依存する確率 P_{X^e,Y^e} に従うとする.



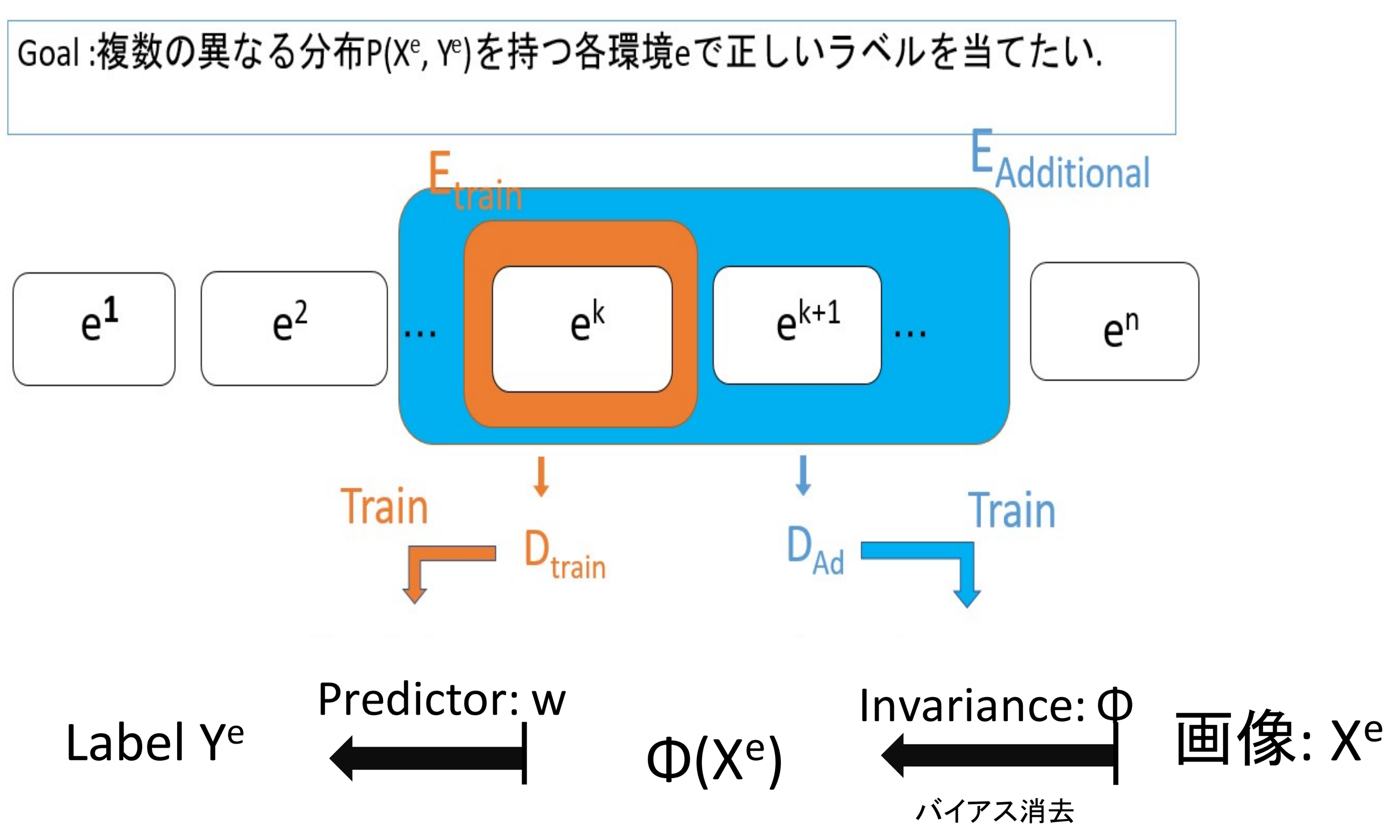
$P(Y^e|\Phi(X^e))$ の $e \in \mathcal{E}_{train}$ に関する非依存性に基づき InvarianceΦ を推定.

Invariance Learning の問題点

$\Rightarrow$ 複数環境からの正確なラベル Y^e が付いた教師データの必要性...

ラベルの階層性を用いた不変学習

不完全なラベル $g(Y^e)$ がアノテーションされた $\mathcal{D}_{ad}$ で Invariance を学習
$\Rightarrow$ 正確なラベル付きのサンプルは単一環境でも Invarince Learning を!



Mathematical formulation and method

$\mathcal{E} := \{e_1, ..., e_n\}$, $X^e \in \mathcal{X}$ and $Y^e \in \mathcal{Y}$: an image and its label on e.
(X^e, Y^e) follows the joint distribution P_{X^e,Y^e}.
Given samples:

- $\mathcal{D}_* := \{(x_i^{e^*}, y_i^{e^*})\}_{i=1}^{n_*} \sim^{i.i.d.} P_{X^{e^*},Y^{e^*}}$ $(e^* \in \mathcal{E}, n_* \in \mathbb{N})$.
- $\sum_{e \in \mathcal{E}_{ad}} \mathcal{D}_e, \mathcal{D}_e := \{(x_i^e, g(y_i^e))\}_{i=1}^{n_e} \sim^{i.i.d.} P_{X^e,g(Y^e)}$ $(n_e \in \mathbb{N}, \mathcal{E}_{ad} \subset \mathcal{E})$.

Out-of-distribution Problem: f^{OOD} をどう求める?

$$f^{OOD} := \arg\min_{f:\mathcal{X}\to\mathcal{Y}} \max_{e\in\mathcal{E}} \mathcal{R}^e(f)$$

(分布の仮定) $\Rightarrow$ $\simeq \arg\min_{\substack{\Phi:\mathcal{X}\to\mathcal{H} \text{ is invariant across } \mathcal{E}, \\ w:\mathcal{H}\to\mathcal{Y}}} \max_{e\in\mathcal{E}} \mathcal{R}^e(w \circ \Phi)$.

($\mathcal{R}^e(f)$: risk of f on e, that is, $\mathcal{R}^e(f) := \int l(Y^e, f(X^e))dP_{X^e,Y^e}$)
(Φ: invariant across $\mathcal{E} \overset{def}{\Longleftrightarrow} P(Y^e|\Phi(X^e))$ does not depend on e).

S. Berry, et al., Recognition in terra incognita, In ECCV, 2019
M. Arjovsky, et al., Invariant Risk Minimization, arXiv: 1907.02893, 2019.

目的関数の構成

本来は以下のような目的関数を設計したいが.....

$$Objective(\Phi, w) := \sum_{e\in\mathcal{E}_{tr}} \hat{\mathcal{R}}^e(w \circ \Phi) +$$
$$\lambda \cdot (\text{Dependence measure of } P(Y^e|\Phi(X^e)) \text{ on } e \in \mathcal{E}_{tr}).$$

$\mathcal{E}_{tr} := \{e^*\}$ の際は正則化項が機能しない
$\Rightarrow \mathcal{D}_{ad}$ で Invariance を推定する目的関数を構成!

$$Objective(\Phi, w) := \hat{\mathcal{R}}^{e^*}(w \circ \Phi) +$$
$$\lambda \cdot \underbrace{(\text{Dependence measure of } P(g(Y^e)|\Phi(X^e)) \text{ on } e \in \mathcal{E}_{ad})}_{\text{不完全なラベルにより計算}}.$$

Invariance の推定には, M. Arjovsky et al. 2019 による以下を援用:

$$\sum_{e\in\mathcal{E}_{ad}} \|\nabla_{\hat{w}=w} \hat{\mathcal{R}}^{(X^e,Z^e)}(\hat{w} \circ \Phi)\|^2.$$

Hyperparamter Selection

本研究に適した CV 法の提案には, $max_e \mathcal{R}^e(f)$ の, つまり, すべての $e \in \mathcal{E}$ での $\mathcal{R}^e(f)$ のシミュレート法を確立する必要があるが...

$$\mathcal{D}_e := \{(x_i^{e_1} g(y_i^e))\}_{i=1}^{n_e} \Longrightarrow \mathcal{R}^e(f) \ \times$$

w を条件付き確率 p_θ でモデリングする場合を想定し.....

- Method I $\Rightarrow$ 不完全データのリスク $\mathcal{R}_{X^e,g(Y^e)}(p_\theta \circ \Phi)$ で近似.
- Method II $\Rightarrow$ Method I に $\mathcal{R}^e(p_\theta \circ \Phi) - \mathcal{R}_{X^e,g(Y^e)}(p_\theta \circ \Phi)$ を補正.

Theorem

For some $\mathcal{Z}^{\not\to} \subset \mathcal{Z}$,

$$\mathcal{R}^e(p_\theta \circ \Phi) - \mathcal{R}_{X^e,g(Y^e)}(p_\theta \circ \Phi) = \sum_{z^{\not\to} \in \mathcal{Z}^{\not\to}} \Big\{P(g(Y^e) = z^{\not\to}) \cdot$$
$$\int -\log p_\theta(Y^e|\Phi(X^e), Y^e = g^{-1}(z^{\not\to}))dP_{(X^e,Y^e)|Y^e=g^{-1}(z^{\not\to})}\Big\}$$

Theory

Main Theorem 1: Effectiveness of Method I

Conditions (a), (b) , (c) and (d) concerning $\mathcal{E}_{ad}$ and e^* hold. Then Method II "selects a preferable hyperparameter."

(d) $\forall\lambda$ with $\mathrm{Im}\Phi_2^\lambda \neq \emptyset$, $\exists e_\lambda \in \mathcal{E}_{ad}$ such that
$P\Big(g(Y^{e^*})|\Phi^\lambda(X^{e^*})\Big) \leq e^{-\beta} - \varepsilon$, $P_{X^{e_\lambda},Y^{e_\lambda}}$-a.e.

Main Theorem 2: Effectiveness of Method II

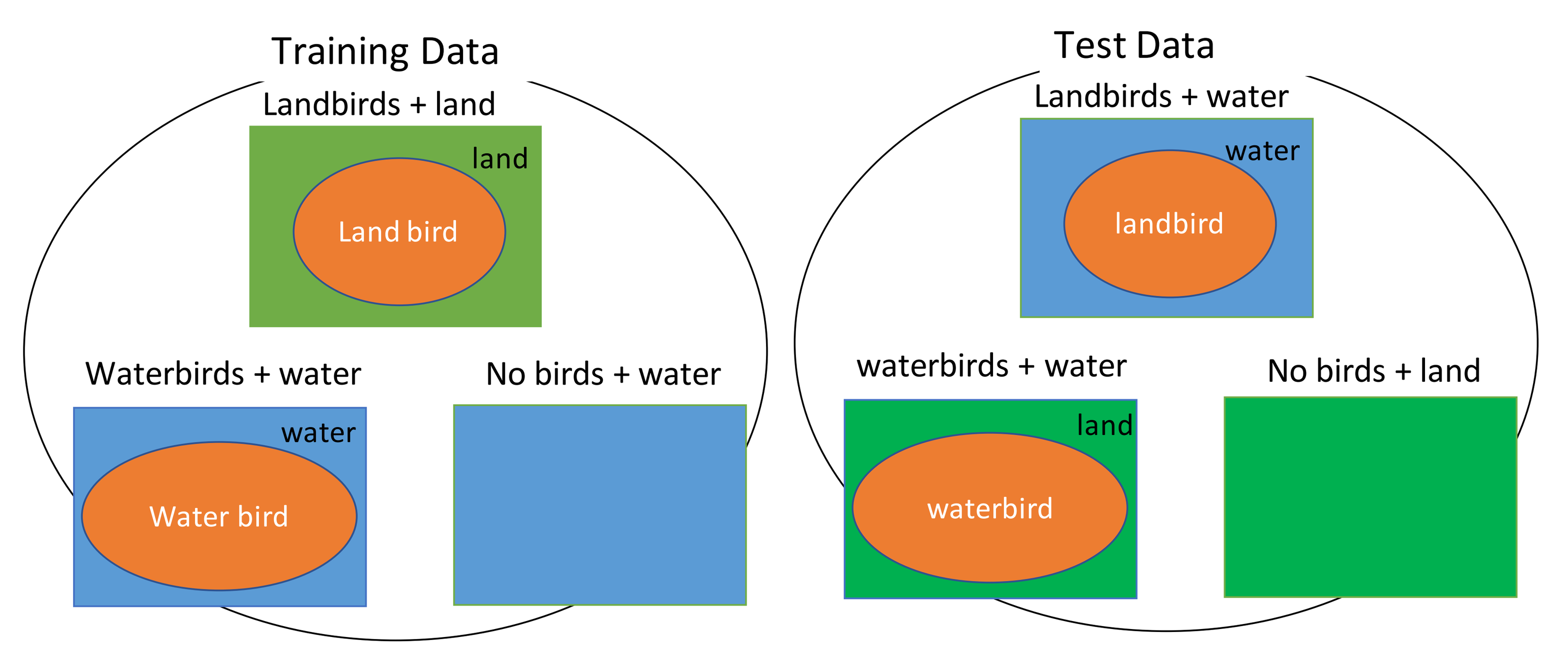
Conditions (a), (b) , (c) and (d)' concerning $\mathcal{E}_{ad}$ and e^* hold. Then Method I "selects a preferable hyperparameter."

(d)' $\forall\lambda$ with $\mathrm{Im}\Phi_2^\lambda \neq \emptyset$, $\exists e_\lambda \in \mathcal{E}_{ad}$ such that
$P\Big(g(Y^{e^*})|\Phi^\lambda(X^{e^*})\Big) \leq e^{-\beta_\lambda} - \varepsilon$, $P_{X^{e_\lambda},Y^{e_\lambda}}$-a.e.

$\beta_\lambda \leq \beta \Rightarrow$ Method II の適用範囲の広さを理論的に明らかに.

Demonstration

Objective task: landbirds *v.s.* Waterbirds *v.s.* No birds
Additional task: landbirds *v.s.* No landbirds



	ERM	Fine Tuning	Feature Extraction	Ours
Test Acc.	.317 (.044)	.364 (.028)	.052 (.013)	.727 (.062)

	Tr CV	LOD CV	CVI	CVII
Test Acc.	.651(.031)	.334 (.029)	.727 (.062)	.727 (.062)