CSDS 452 Causality and Machine Learning

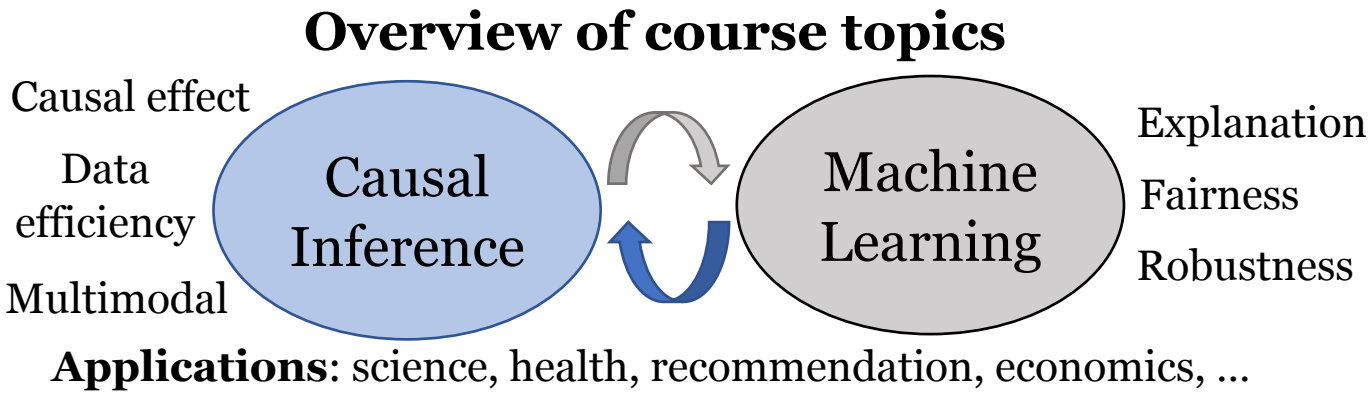
Lecture 15: Causal Generalization
1- Invariant learning

Instructor: Jing Ma

Fall 2024, CDS@CWRU

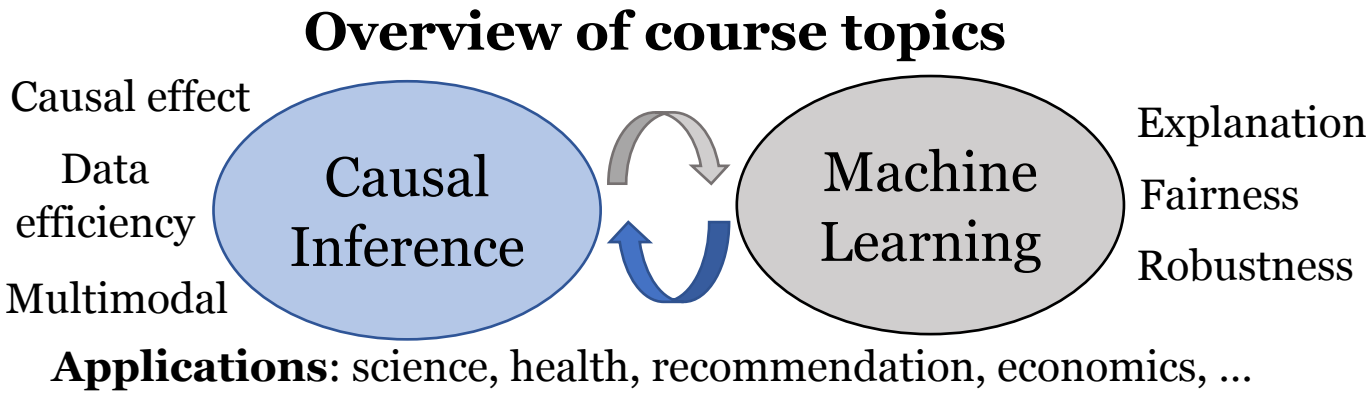
What we have covered

- Three major topics will be covered:
  - ✓ Introduction to causal inference
  - ✓ Using machine learning to help causal inference
  - ○ Using causal inference to help machine learning



What will we learn next?

- Three major topics will be covered:
  - ✓ Introduction to causal inference
  - ✓ Using machine learning to help causal inference
  - ○ **Using causal inference to help machine learning**



Causal Inference for ML

Outline

- Invariant risk minimization (IRM)
- Invariant rationalization

Invariant Risk Minimization

Martin Arjovsky, L´eon Bottou, Ishaan Gulrajani, David Lopez-Paz

2020

An example for motivation

- Classification of cows and camels

A camel?

Sandy background

5 Unusual Facts About Camels - YouTube
youtube.com

Camel Facts
thoughtco.com

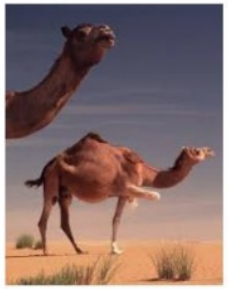
Army Bring Camels to Texas ...
texashillcountry.com

Where Do Camels Live? - WorldAtlas.com
worldatlas.com

Camel | San Diego Zoo Animals & Plants
animals.sandiegozoo.org

Camels' Humps Are Not Filled Wi...
todayifoundout.com

E-chips to be installed on camels ...
egypttoday.com



When and Where Wer...
thoughtco.com

Camels - All About Camels Facts ...
animalcorner.co.uk

Grassy background

cow | Description & Facts | Britan...
britannica.com

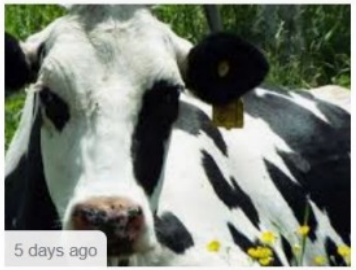
From Two Bulls, 9 Million Dairy Cows
undark.org

Dairy cattle - Wikipedia
en.wikipedia.org

Stomach Could Help Your Health ...
ucdavis.edu

Cow toilets' in Netherlands aim to cut ...
phys.org

Isis rigs cows with explosives in ...
independent.co.uk



videos will make you want to hug a c...
wbaltv.com

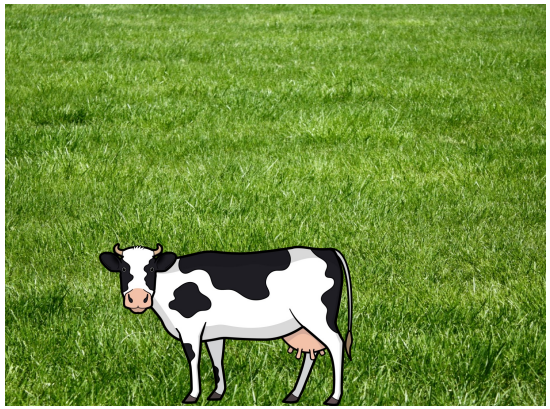
Escaped Rodeo Cow Is On The Lam In ...
huffpost.com

Atypical BSE Confirmed in FI...
dairyherd.com

Meet Knickers, the 1,400kg cow from ...
youtube.com

Correlation & Causation dilemma

- Minimization training error may lead the models to recklessly absorb all the correlations in training data.
- However, spurious correlations stemming from data biases are unrelated to the causal explanation of interest.



Cow

Camel?

?

Correlation & Causation dilemma

- Minimization training error may lead the models to recklessly absorb all the correlations in training data.
- However, spurious correlations stemming from data biases are unrelated to the causal explanation of interest.

Problem:

- identify which properties of the training data describe spurious correlations (landscapes and contexts)

- identify which properties represent the phenomenon of interest (animal shapes)

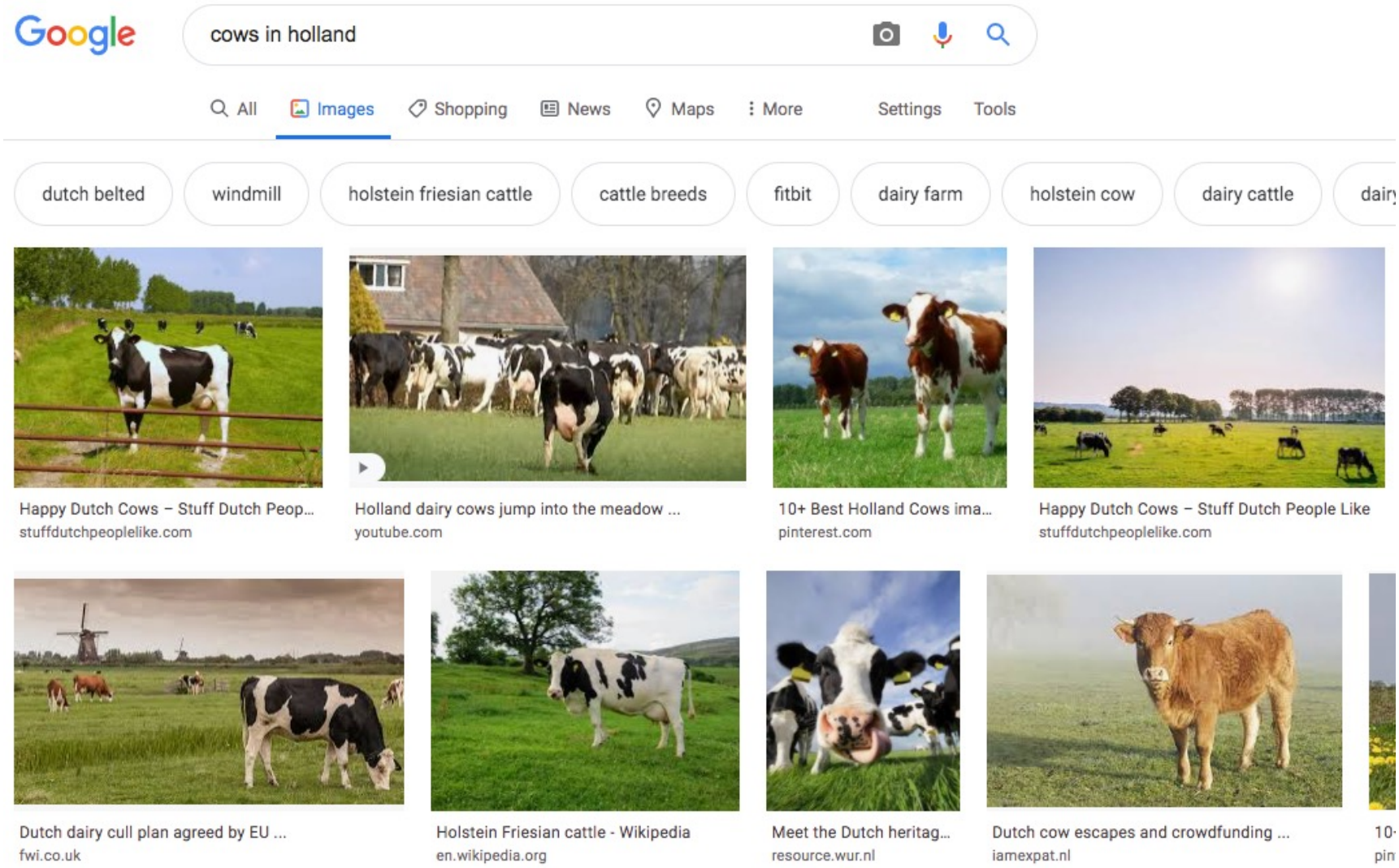
Causation -> invariance

- Spurious correlations **do not** appear to be stable properties
- There exists an intimate connection between invariance and causation useful for generalization

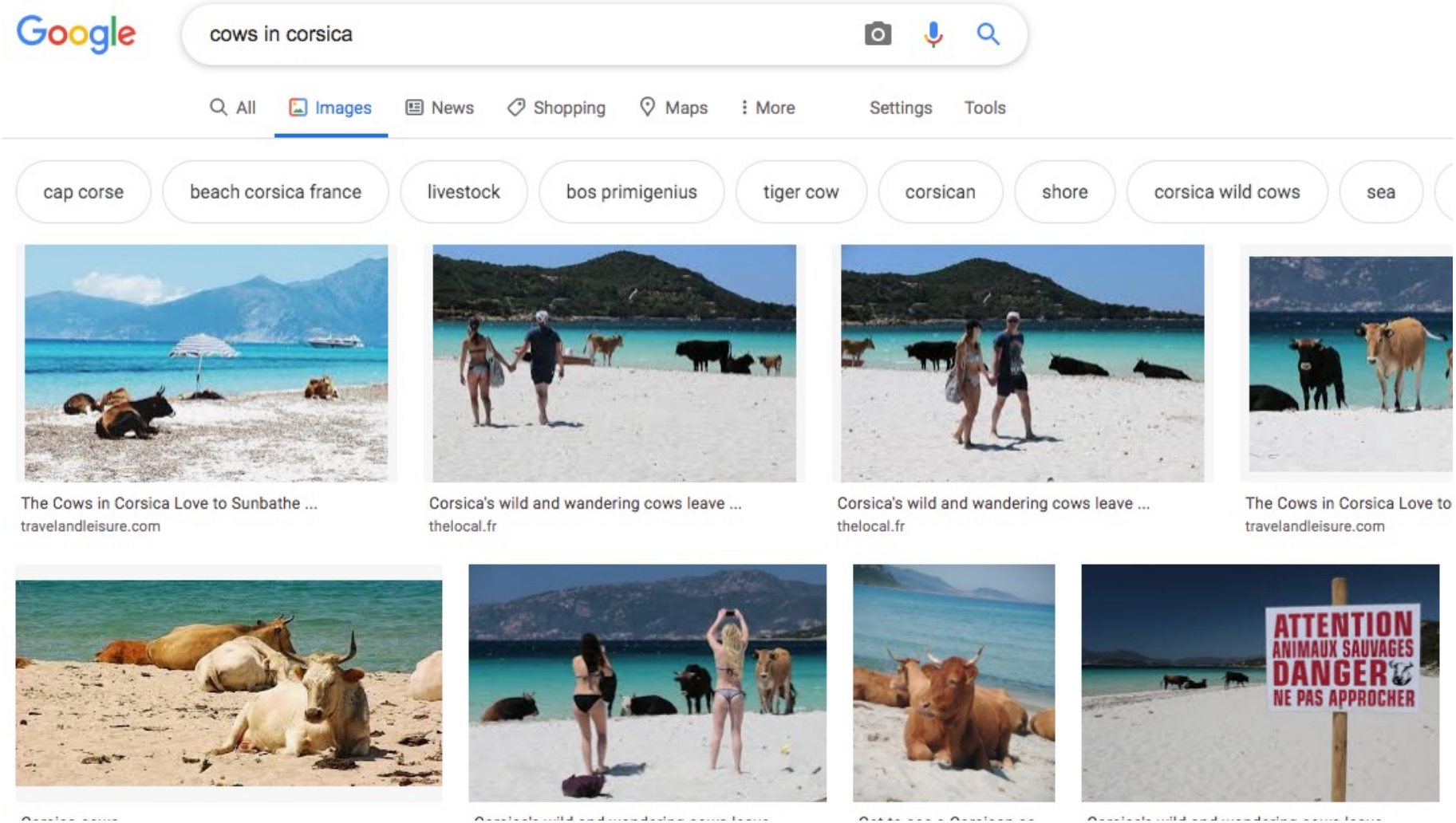
Invariant learning: Intuition

- Assume that the training data is collected into distinct, separate environments
- We promote learning correlations that are **stable** across training environments, as these should (under conditions that we will study) also hold in novel testing environments.

Different environments for cows



Different environments for cows



Invariant risk minimization (IRM) principle

To learn invariances across environments, find a data representation such that the optimal classifier on top of that representation matches for all environments.

Problem Formulation

- Consider datasets

$$D_e := \{(x_i^e, y_i^e)\}_{i=1}^{n_e} \qquad e \in \mathcal{E}_{\text{tr}} \qquad \mathcal{E}_{\text{all}} \supset \mathcal{E}_{\text{tr}}$$

- Goal: Learn a predictor

$$Y \approx f(X)$$

Problem Formulation

- Consider datasets

$$D_e := \{(x_i^e, y_i^e)\}_{i=1}^{n_e} \qquad e \in \mathcal{E}_{\text{tr}} \qquad \mathcal{E}_{\text{all}} \supset \mathcal{E}_{\text{tr}}$$

- Goal: Learn a predictor

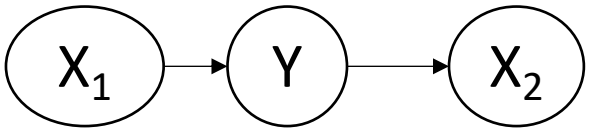
$$Y \approx f(X)$$

- We wish to minimize

$$R^{\text{OOD}}(f) = \max_{e \in \mathcal{E}_{\text{all}}} R^e(f)$$

where $R^e(f) := \mathbb{E}_{X^e, Y^e}[\ell(f(X^e), Y^e)]$ is the risk under environment *e*

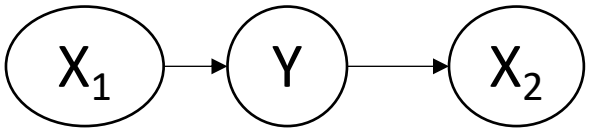
An example



$$X_1 \leftarrow \text{Gaussian}(0, \sigma^2),$$
$$Y \leftarrow X_1 + \text{Gaussian}(0, \sigma^2),$$
$$X_2 \leftarrow Y + \text{Gaussian}(0, 1).$$

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

An example

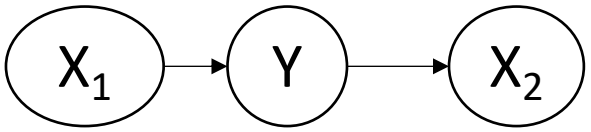


$$X_1 \leftarrow \text{Gaussian}(0, \sigma^2),$$
$$Y \leftarrow X_1 + \text{Gaussian}(0, \sigma^2),$$
$$X_2 \leftarrow Y + \text{Gaussian}(0, 1).$$

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

- $\mathcal{E}_{\text{all}}$ contains all modifications of the structural equations for X_1 and X_2, and those varying the noise of Y within a finite range

An example

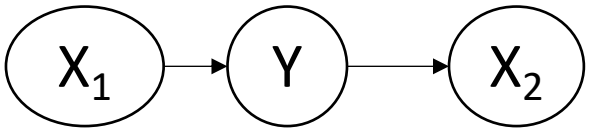


$$X_1 \leftarrow \text{Gaussian}(0, \sigma^2),$$
$$Y \leftarrow X_1 + \text{Gaussian}(0, \sigma^2),$$
$$X_2 \leftarrow Y + \text{Gaussian}(0, 1).$$

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

- $\mathcal{E}_{\text{all}}$ contains all modifications of the structural equations for X_1 and X_2, and those varying the noise of Y within a finite range
- Predict Y from (X_1, X_2) using a least-squares predictor $\hat{Y}^e = X_1^e \hat{\alpha}_1 + X_2^e \hat{\alpha}_2$

An example



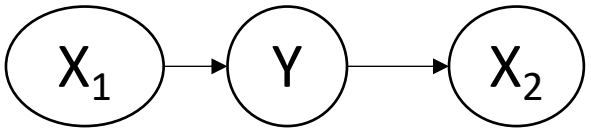
$$X_1 \leftarrow \text{Gaussian}(0, \sigma^2),$$
$$Y \leftarrow X_1 + \text{Gaussian}(0, \sigma^2),$$
$$X_2 \leftarrow Y + \text{Gaussian}(0, 1).$$

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

- $\mathcal{E}_{\text{all}}$ contains all modifications of the structural equations for X_1 and X_2, and those varying the noise of Y within a finite range
- Predict Y from (X_1, X_2) using a least-squares predictor $\hat{Y}^e = X_1^e \hat{\alpha}_1 + X_2^e \hat{\alpha}_2$

- regress from X_1^e, to obtain $\hat{\alpha}_1 = 1$ and $\hat{\alpha}_2 = 0$,

An example

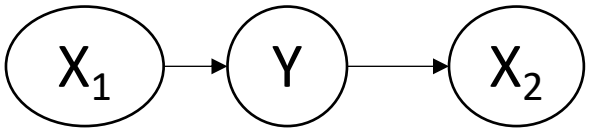


$$X_1 \leftarrow \text{Gaussian}(0, \sigma^2),$$
$$Y \leftarrow X_1 + \text{Gaussian}(0, \sigma^2),$$
$$X_2 \leftarrow Y + \text{Gaussian}(0, 1).$$

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

- $\mathcal{E}_{\text{all}}$ contains all modifications of the structural equations for X_1 and X_2, and those varying the noise of Y within a finite range
- Predict Y from (X_1, X_2) using a least-squares predictor $\hat{Y}^e = X_1^e \hat{\alpha}_1 + X_2^e \hat{\alpha}_2$
  - regress from X_1^e, to obtain $\hat{\alpha}_1 = 1$ and $\hat{\alpha}_2 = 0$,
  - regress from X_2^e, to obtain $\hat{\alpha}_1 = 0$ and $\hat{\alpha}_2 = \sigma(e)^2/(\sigma(e)^2 + \frac{1}{2})$,

An example

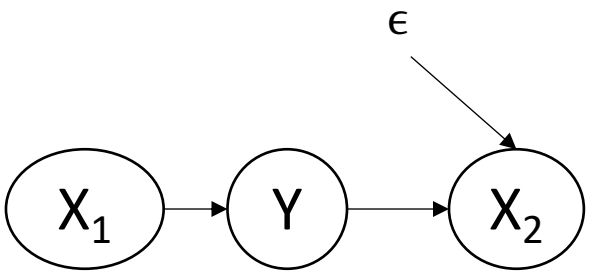


$$X_1 \leftarrow \text{Gaussian}(0, \sigma^2),$$
$$Y \leftarrow X_1 + \text{Gaussian}(0, \sigma^2),$$
$$X_2 \leftarrow Y + \text{Gaussian}(0, 1).$$

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

- $\mathcal{E}_{\text{all}}$ contains all modifications of the structural equations for X_1 and X_2, and those varying the noise of Y within a finite range
- Predict Y from (X_1, X_2) using a least-squares predictor $\hat{Y}^e = X_1^e \hat{\alpha}_1 + X_2^e \hat{\alpha}_2$
  - regress from X_1^e, to obtain $\hat{\alpha}_1 = 1$ and $\hat{\alpha}_2 = 0$,
  - regress from X_2^e, to obtain $\hat{\alpha}_1 = 0$ and $\hat{\alpha}_2 = \sigma(e)^2/(\sigma(e)^2 + \frac{1}{2})$,
  - regress from (X_1^e, X_2^e), to obtain $\hat{\alpha}_1 = 1/(\sigma(e)^2+1)$ and $\hat{\alpha}_2 = \sigma(e)^2/(\sigma(e)^2+1)$.

A common mistake

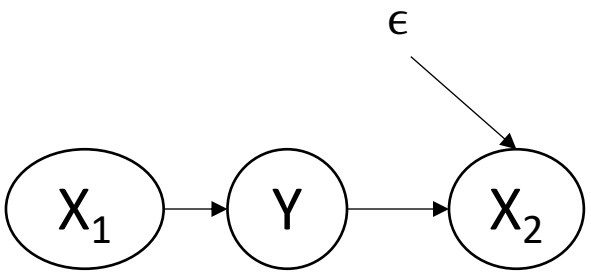


$$X_1 \leftarrow \text{Gaussian}(0, \sigma^2),$$
$$Y \leftarrow X_1 + \text{Gaussian}(0, \sigma^2),$$
$$X_2 \leftarrow Y + \text{Gaussian}(0, 1).$$

- As $X_2 = Y + \epsilon$, where $\epsilon \sim N(0,1)$, then $Y = X_2 - \epsilon$, so coefficient $\alpha_2 = 1$

Is it correct?

A common mistake



$$X_1 \leftarrow \text{Gaussian}(0, \sigma^2),$$
$$Y \leftarrow X_1 + \text{Gaussian}(0, \sigma^2),$$
$$X_2 \leftarrow Y + \text{Gaussian}(0, 1).$$

- As $X_2 = Y + \epsilon$, where $\epsilon \sim N(0,1)$, then $Y = X_2 - \epsilon$, so coefficient $\alpha_2 = 1$ (this is wrong!)
- Why? Here, as X_2 is influenced by ϵ,

$P(Y|X_2) = P(X_2 - \epsilon|X_2)$, notice that $P(\epsilon) \neq P(\epsilon|X_2)$

Analysis

- The regression using X_1 is an invariant correlation: this is the only regression whose coefficients do not depend on the environment e.

- regress from X_1^e, to obtain $\hat{\alpha}_1 = 1$ and $\hat{\alpha}_2 = 0$,
- regress from X_2^e, to obtain $\hat{\alpha}_1 = 0$ and $\hat{\alpha}_2 = \sigma(e)^2/(\sigma(e)^2 + \frac{1}{2})$,
- regress from (X_1^e, X_2^e), to obtain $\hat{\alpha}_1 = 1/(\sigma(e)^2+1)$ and $\hat{\alpha}_2 = \sigma(e)^2/(\sigma(e)^2+1)$.

Analysis

- The regression using X_1 is an invariant correlation: this is the only regression whose coefficients do not depend on the environment e.
- Conversely, the second and third regressions exhibit coefficients that vary from environment to environment.

- regress from X_1^e, to obtain $\hat{\alpha}_1 = 1$ and $\hat{\alpha}_2 = 0$,
- regress from X_2^e, to obtain $\hat{\alpha}_1 = 0$ and $\hat{\alpha}_2 = \sigma(e)^2/(\sigma(e)^2 + \frac{1}{2})$,
- regress from (X_1^e, X_2^e), to obtain $\hat{\alpha}_1 = 1/(\sigma(e)^2+1)$ and $\hat{\alpha}_2 = \sigma(e)^2/(\sigma(e)^2+1)$.

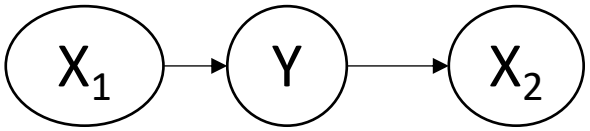
Analysis

- The regression using X_1 is an invariant correlation: this is the only regression whose coefficients do not depend on the environment e.
- Conversely, the second and third regressions exhibit coefficients that vary from environment to environment.
- The invariant rule $\hat{Y} = 1 \cdot X_1 + 0 \cdot X_2$ is the only predictor with finite R^{OOD} across $\mathcal{E}_{\mathrm{all}}$, and it is the causal explanation for Y.

Prior work

- Empirical Risk Minimization (ERM)
  - May capture spurious correlation

- regress from (X_1^e, X_2^e), to obtain $\hat{\alpha}_1 = 1/(\sigma(e)^2+1)$ and $\hat{\alpha}_2 = \sigma(e)^2/(\sigma(e)^2+1)$.



$$X_1 \leftarrow \text{Gaussian}(0, \sigma^2),$$
$$Y \leftarrow X_1 + \text{Gaussian}(0, \sigma^2),$$
$$X_2 \leftarrow Y + \text{Gaussian}(0, 1).$$

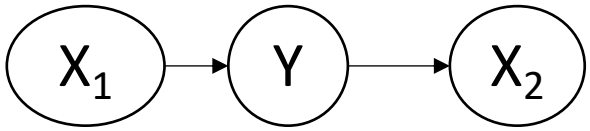
$$\mathcal{E}_{\text{tr}} = \{\text{replace } \sigma^2 \text{ by } 10,\ \text{replace } \sigma^2 \text{ by } 20\}$$

Prior work

- Empirical Risk Minimization (ERM)
  - May capture spurious correlation
- Robust learning: minimize $R^{\mathrm{rob}}(f) = \max_{e \in \mathcal{E}_{\mathrm{tr}}} R^e(f) - r_e$
  - Turns out to be equivalent to minimizing a weighted average of environment training errors

Prior work

- Empirical Risk Minimization (ERM)
  - May capture spurious correlation
- Robust learning: minimize $R^{\text{rob}}(f) = \max_{e \in \mathcal{E}_{\text{tr}}} R^e(f) - r_e$
  - Turns out to be equivalent to minimizing a weighted average of environment training errors
- Domain adaptation:
  - Estimate a data representation $\phi(X_1, X_2)$ that follows the same distribution for all environments
  - Fail to find the true invariance (e.g., in the shown example)



$$X_1 \leftarrow \text{Gaussian}(0, \sigma^2),$$
$$Y \leftarrow X_1 + \text{Gaussian}(0, \sigma^2),$$
$$X_2 \leftarrow Y + \text{Gaussian}(0, 1).$$

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

Prior work

- Empirical Risk Minimization (ERM)
  - May capture spurious correlation
- Robust learning: minimize $R^{\mathrm{rob}}(f) = \max_{e \in \mathcal{E}_{\mathrm{tr}}} R^e(f) - r_e$
  - Turns out to be equivalent to minimizing a weighted average of environment training errors
- Domain adaptation:
  - Estimate a data representation $\phi(X_1, X_2)$ that follows the same distribution for all environments
  - Fail to find the true invariance (e.g., in the shown example)
- Invariant causal prediction (ICP) techniques
  - Find the subset of variables with same residual distributions
  - Strong assumptions for the data to recover invariance: the data (i) is Gaussian, (ii) satisfies a linear SEM, and (iii) is obtained by certain types of interventions

IRM formulation

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 $e \in \mathcal{E}$.*

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

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

Minimize the empirical risk

Invariant constraint: the predictor elicited by representation ϕ is optimal over all training environments

IRM principle: "To learn invariances across environments, find a data representation such that the optimal classifier on top of that representation matches for all environments. "

IRM formulation

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 $e \in \mathcal{E}$.*

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

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

Minimize the empirical risk

Invariant constraint: the predictor elicited by representation ϕ is optimal over all training environments

Impractical: This is a challenging, bi-leveled optimization problem. Since each constraint calls an inner optimization routine

A Feasible Version -- IRMv1

- A feasible version: IRMv1

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

  - Φ becomes the entire invariant predictor
  - w = 1.0 is a scalar and fixed “dummy” classifier

From IRM to IRMv1

- **Step 1: Phrasing the constraints as a penalty**

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

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

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

A differentiable function which measures how far w is from minimizing $R^e(w \circ \Phi)$

hyper-parameter balancing predictive power and invariance.

From IRM to IRMv1

- **Step 2: Choosing a penalty D for linear classifiers w**
  - Consider linear classifier, where the optimum is:

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

  - Two different penalties:

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

Choosing a smooth penalty

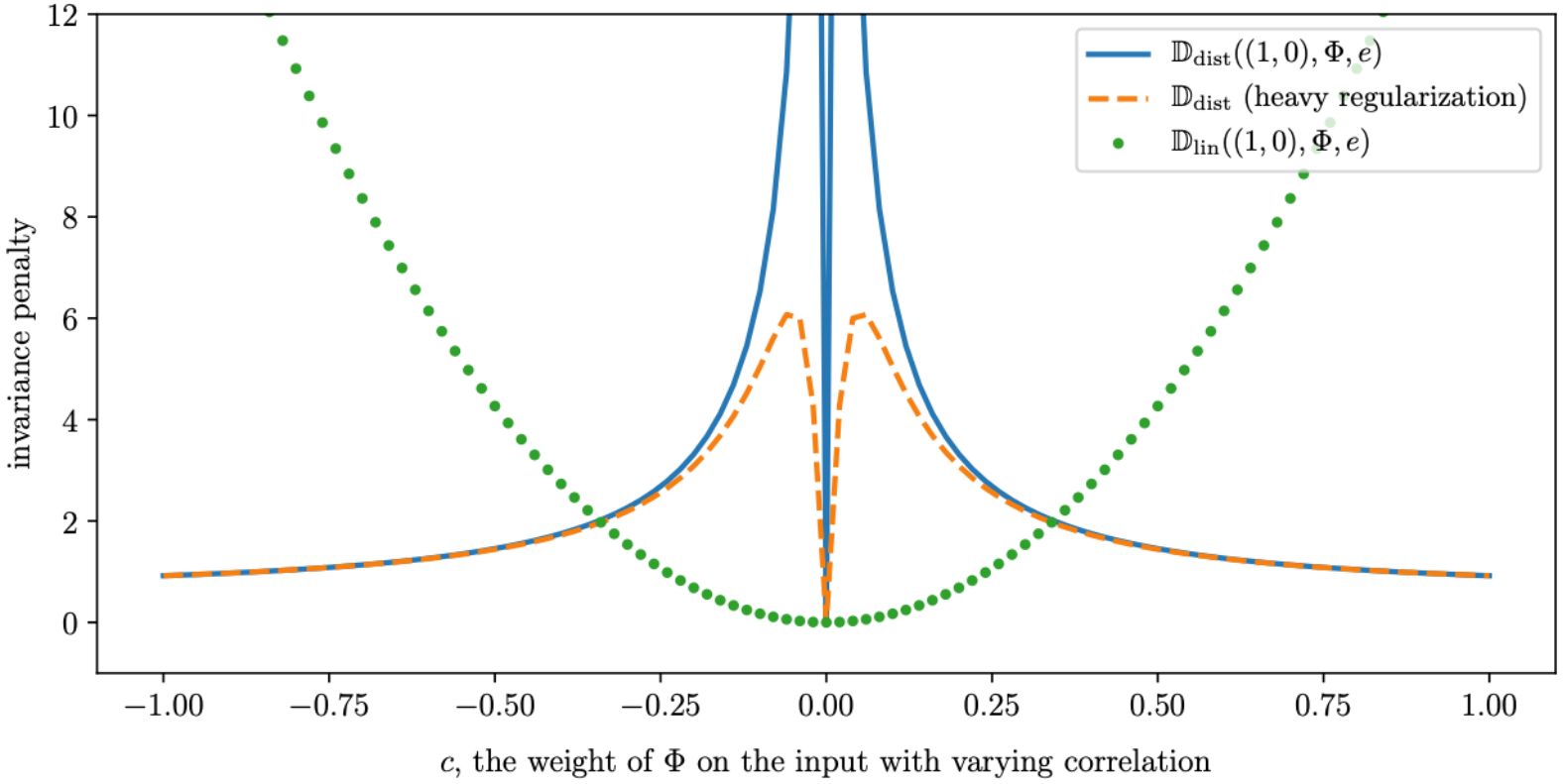


Figure 1: Different measures of invariance lead to different optimization landscapes in our Example 1. The naïve approach of measuring the distance between optimal classifiers $\mathbb{D}_{\text{dist}}$ leads to a discontinuous penalty (solid blue unregularized, dashed orange regularized). In contrast,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

From IRM to IRMv1

- **Step 2: Choosing a penalty D for linear classifiers w**
  - Consider linear classifier, where the optimum is:

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

  - Two different penalties:

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

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

More smooth

From IRM to IRMv1

- **Step 3: Fixing the linear classifier w**
  - The problem is over-parameterized
  - we can re-parametrize our invariant predictor as to give w any non-zero value $\tilde{w}$ of our choosing

$$w \circ \Phi = \underbrace{\left(w \circ \Psi^{-1}\right)}_{\tilde{w}} \circ \underbrace{\left(\Psi \circ \Phi\right)}_{\tilde{\Phi}}$$

  - restrict our search to the data representations for which all the environment optimal classifiers are equal to the same fixed vector $\tilde{w}$

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

From IRM to IRMv1

- Step 4: Scalar fixed classifiers are sufficient to monitor invariance

$$L_{\text{IRM},w=1.0}(\Phi^\top) = \sum_{e \in \mathcal{E}_{\text{tr}}} R^e(\Phi^\top) + \lambda \cdot \mathbb{D}_{\text{lin}}(1.0, \Phi^\top, e).$$

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 $\text{Ker}(\Phi)$ is orthogonal to v and contains all the $\nabla R^e(v)$.*

So, any linear invariant predictor can be decomposed as linear data representations of different ranks, in particular, we can restrict our search to matrices Φ, and w can be the fixed scalar 1.0.

From IRM to IRMv1

- Step 5: Extending to general losses and multivariate outputs
  - Recall $\mathbb{D}_{\text{lin}}(w, \Phi, e) = \left\| \mathbb{E}_{X^e} \left[\Phi(X^e) \Phi(X^e)^\top \right] w - \mathbb{E}_{X^e, Y^e} \left[\Phi(X^e) Y^e \right] \right\|^2$
  - Can be written as a general function of the risk

$$\mathbb{D}(1.0, \Phi, e) = \left\| \nabla_{w|w=1.0} R^e(w \cdot \Phi) \right\|^2$$

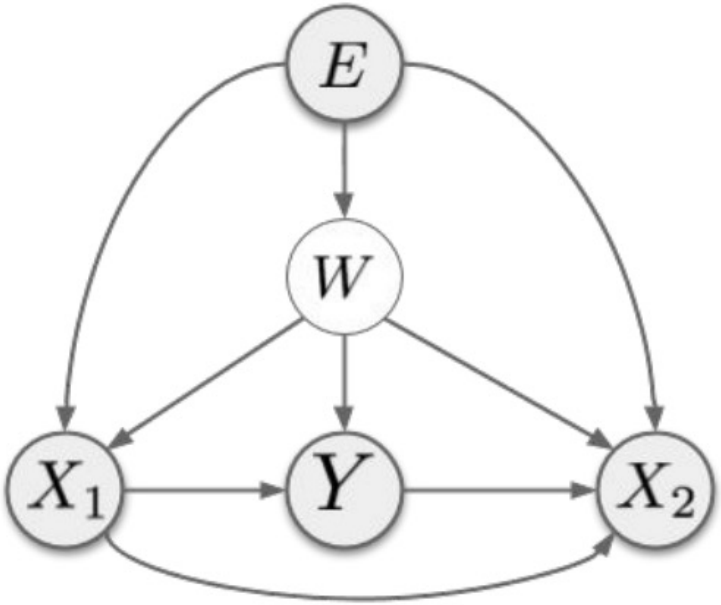
$$\min_{\Phi: \mathcal{X} \to \mathcal{Y}} \sum_{e \in \mathcal{E}_{\text{tr}}} R^e(\Phi) + \lambda \cdot \left\| \nabla_{w|w=1.0} R^e(w \cdot \Phi) \right\|^2, \qquad \text{(IRMv1)}$$

Invariance, causality and generalization

- IRM promotes low error and invariance across training environments.
- When do these conditions imply invariance across all environments?
- When do these conditions lead to low error across all environments? (Basically, OOD generalization)
- How does statistical invariance and out-of-distribution generalization relate to concepts from the theory of causation?

Please see the article for details...

An Information Theoretic View



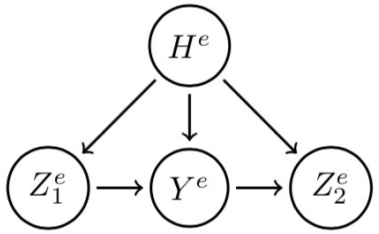
$\phi(X)$ such that:

- $Y \perp E|\phi(X)$, and
- ϕ is informative about y, i.e. we can predict y accurately from $\phi(x)$

https://www.inference.vc/invariant-risk-minimization/

Experiments

- Datasets: Synthetic data, Colored MNIST
- Synthetic data



$$H^e \leftarrow \mathcal{N}(0, e^2)$$ hidden confounder

$$Z_1^e \leftarrow \mathcal{N}(0, e^2) + W_{h\to 1} H^e$$

$$Y^e \leftarrow Z_1^e \cdot W_{1\to y} + \mathcal{N}(0, \sigma_y^2) + W_{h\to y} H^e$$

$$Z_2^e \leftarrow W_{y\to 2} Y^e + \mathcal{N}(0, \sigma_2^2) + W_{h\to 2} H^e$$

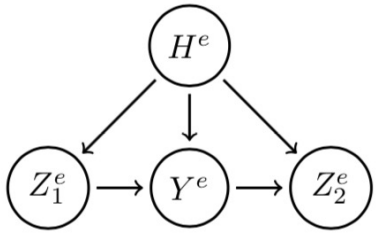
Features(may not be directly observed)

Figure 3: In our synthetic experiments, the task is to predict Y^e from $X^e = S(Z_1^e, Z_2^e)$.

https://github.com/facebookresearch/InvariantRiskMinimization

Experiments

- Datasets: Synthetic data, Colored MNIST
- Synthetic data



$$H^e \leftarrow \mathcal{N}(0, e^2)$$ hidden confounder

$$Z_1^e \leftarrow \mathcal{N}(0, e^2) + W_{h\to 1} H^e$$

$$Y^e \leftarrow Z_1^e \cdot W_{1\to y} + \mathcal{N}(0, \sigma_y^2) + W_{h\to y} H^e$$

$$Z_2^e \leftarrow W_{y\to 2} Y^e + \mathcal{N}(0, \sigma_2^2) + W_{h\to 2} H^e$$

Features(may not be directly observed)

Figure 3: In our synthetic experiments, the task is to predict Y^e from $X^e = S(Z_1^e, Z_2^e)$.

- *Scrambled* (S) observations, where S is an orthogonal matrix, or *unscrambled* (U) observations, where $S = I$.
- *Fully-observed* (F) graphs, where $W_{h\to 1} = W_{h\to y} = W_{h\to 2} = 0$, or *partially-observed* (P) graphs, where $(W_{h\to 1}, W_{h\to y}, W_{h\to 2})$ are Gaussian.
- *Homoskedastic* (O) Y-noise, where $\sigma_y^2 = e^2$ and $\sigma_2^2 = 1$, or *heteroskedastic* (E) Y-noise, where $\sigma_y^2 = 1$ and $\sigma_2^2 = e^2$.

https://github.com/facebookresearch/InvariantRiskMinimization

Synthetic data

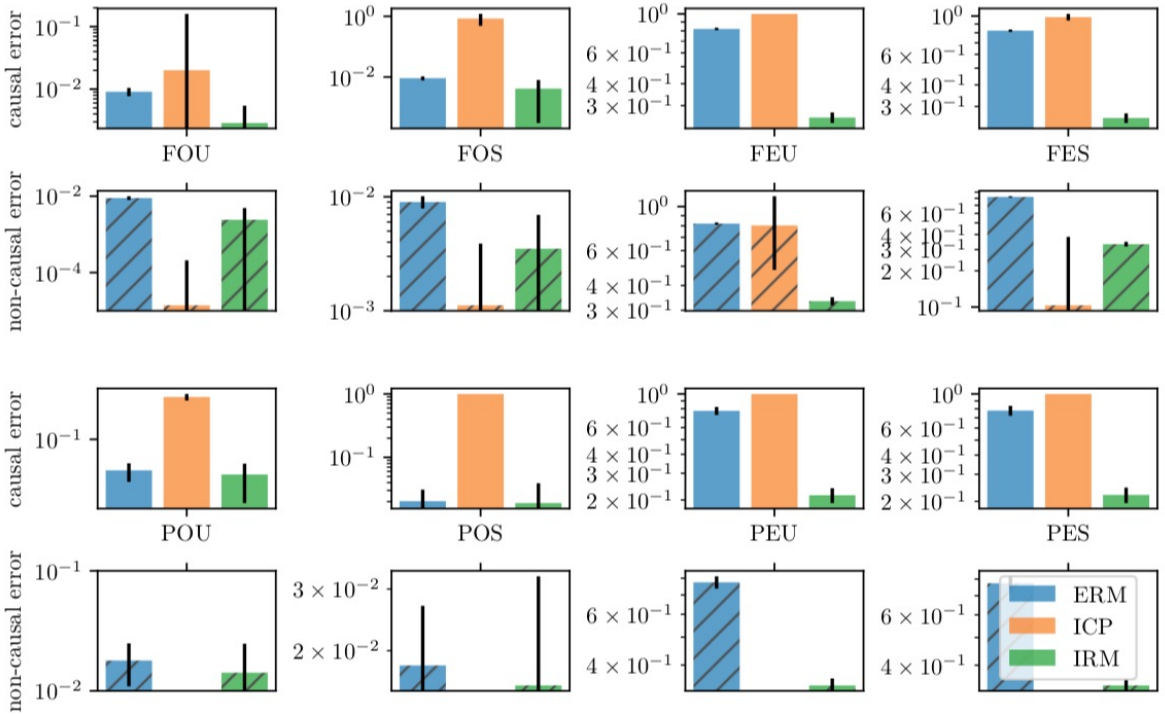


Figure 4: Average errors on causal (plain bars) and non-causal (striped bars) weights for our synthetic experiments. The y-axes are in log-scale. See main text for details.

- 8 (2x2x2) settings
- 2 Metrics:
  - Causal error
  - non-causal error
- Observations:

  IRM has lowest causal error

Colored MNIST

- Goal: predict a **binary** label assigned to each image based on the digit
  - Y=0 for digits 0-4, Y=1 for digits 5-9, then flip with prob=0.25
- Color: green (0)/ red(1) by flipping y with prob p_e
  - (Strong) spurious correlation between color and Y
- Two training environments: {p_e=0.1, 0.2}, test environment: {p_e=0.9}
  - The spurious correlation is **reversed** from train to test (non-causal features fail)
- Model: MLPs + different objectives

Colored MNIST

Algorithm	**Acc. train envs.**	**Acc. test env.**
ERM	87.4 ± 0.2	17.1 ± 0.6
IRM (ours)	70.8 ± 0.9	$\mathbf{66.9 \pm 2.5}$
Random guessing (hypothetical)	50	50
Optimal invariant model (hypothetical)	75	75
ERM, grayscale model (oracle)	73.5 ± 0.2	73.0 ± 0.4

Table 1: Accuracy (%) of different algorithms on the Colored MNIST synthetic task. ERM fails in the test environment because it relies on spurious color correlations to classify digits. IRM detects that the color has a spurious correlation with the label and thus uses only the digit to predict, obtaining better generalization to the new unseen test environment.

- ERM model classifies mainly based on color, so fails on the test data
- IRM's performance is just slightly lower than the optimal case

Colored MNIST

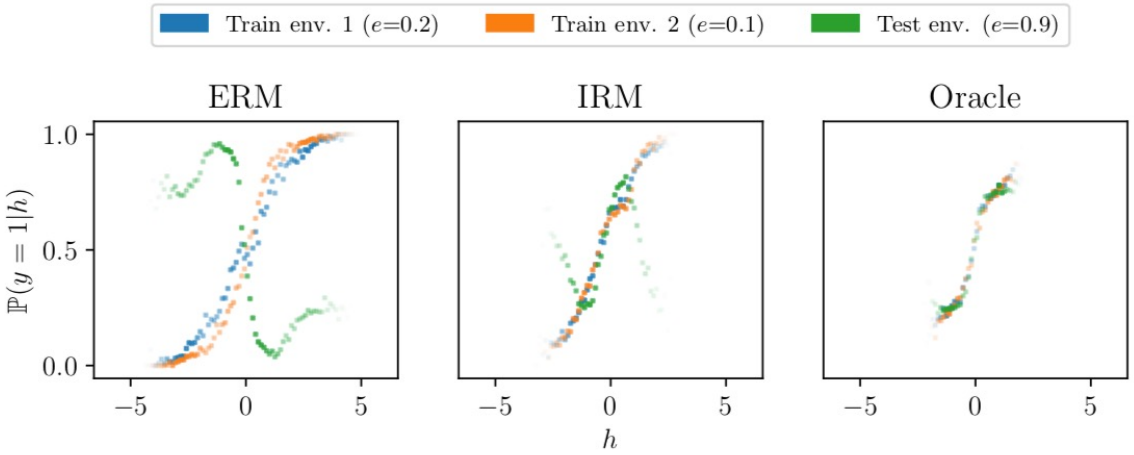


Figure 5: $P(y=1|h)$ as a function of h for different models trained on Colored MNIST: (left) an ERM-trained model, (center) an IRM-trained model, and (right) an ERM-trained model which only sees grayscale images and therefore is perfectly invariant by construction. IRM learns approximate invariance from data alone and generalizes well to the test environment.

- ERM has different distributions of P(y=1|h) on training / test data
  - Do not catch the causal features

Smells like information bottleneck…

$$\max_{\phi} \{I[Y, Z] - \beta I[X, Z]\}$$

Take the least from input and generate most of the output

Smells like information bottleneck…

Take the least from input and generate most of the output

$$\max_{\phi} \{I[Y, Z] - \beta I[X, Z]\}$$

$$\max_{\phi} \{I[Y, Z] - \beta I[Y, E|Z]\} \qquad Z = \phi(X)$$

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

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

Summary and Takeaways

- IRM learns a **robust** predictor based on causal associations between variables, rather than spurious correlations.
- Causal predictors enable **out-of-distribution (OOD)** generalization.
- Assume that data are sampled from **different environments**.
- **IRM principle**: find a representation of features, such that the optimal predictor is simultaneously optimal in all environments.

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

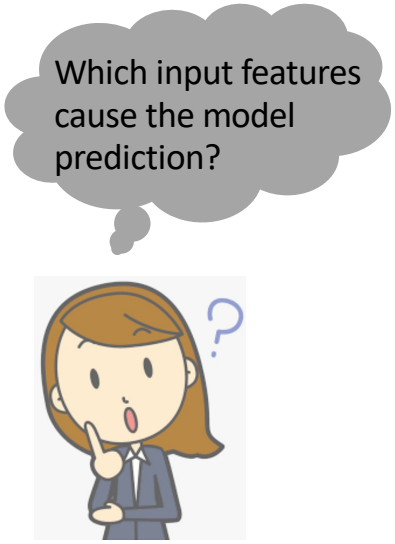
Invariant Rationalization

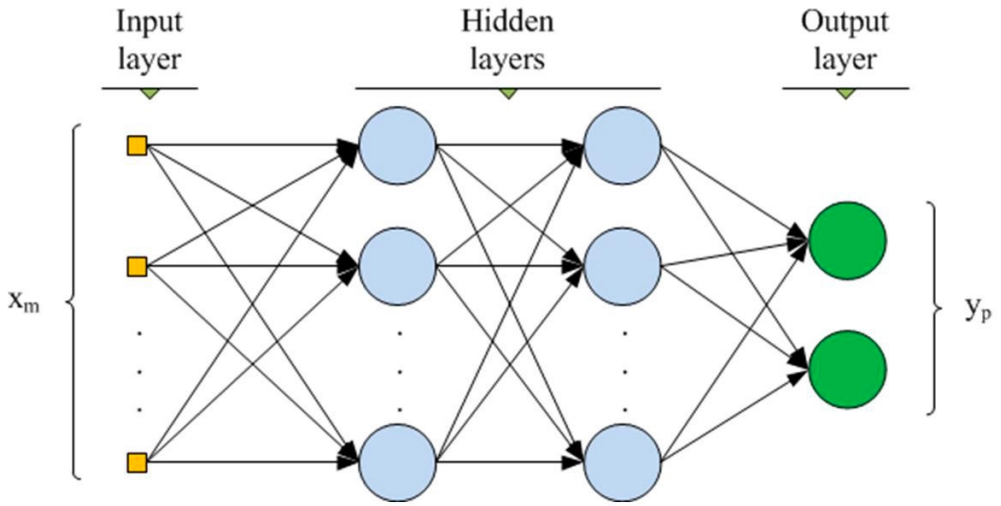
2020

Shiyu Chang, Yang Zhang, Mo Yu, Tommi S. Jaakkola
MIT-IBM Watson AI Lab, IBM Research, CSAIL MIT

Motivation

- A general aim: Explain the predictions of complex neural models





Background

- **Selective rationalization**： identifying a small subset of input features that best explains or supports the prediction.
  - Key idea： find a small subset of the input features – *rationale* – that suffices on its own to yield the same outcome.
- The most commonly-used criterion： maximum mutual information (**MMI**) criterion.
  - Goal: maximizes the mutual information between the subset and the model output. $Z = Z(X)$
  - Constraint: the selected subset remains within a prescribed length.

Drawback of MMI Criterion

- It is prone to highlighting spurious correlations between the input features and the output as valid explanations.
  - These spurious correlations represent statistical relations present in the training data, but may change when test.

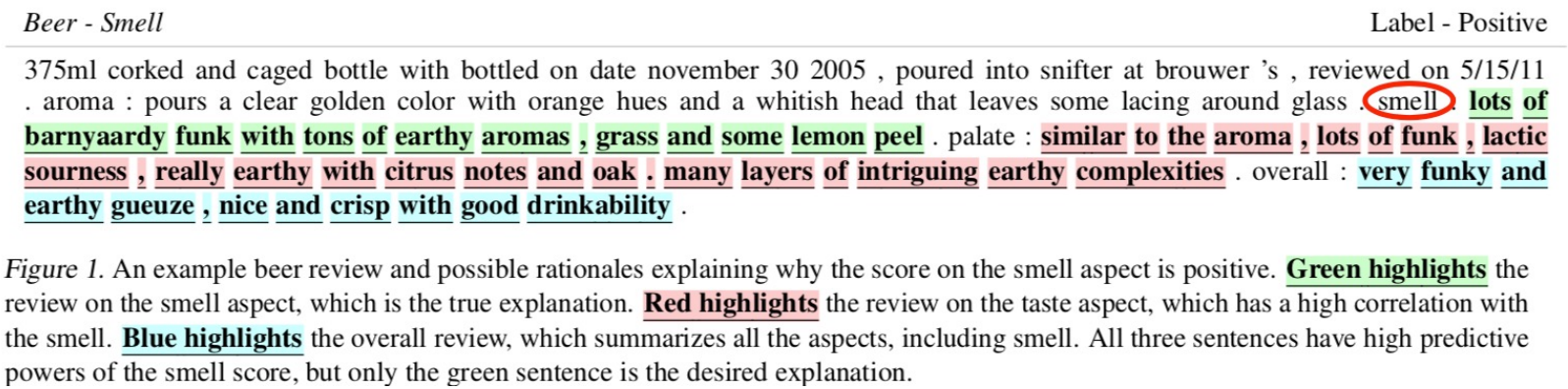


Figure 1. An example beer review and possible rationales explaining why the score on the smell aspect is positive. **Green highlights** the review on the smell aspect, which is the true explanation. **Red highlights** the review on the taste aspect, which has a high correlation with the smell. **Blue highlights** the overall review, which summarizes all the aspects, including smell. All three sentences have high predictive powers of the smell score, but only the green sentence is the desired explanation.

Causal features

- New goal: Design a rationalization criterion that approximates finding causal features.
- Challenges & solutions: assessing causality is challenging => approximate the task by searching features that are *invariant*.
- This line of work is referred as **invariant risk minimization (IRM)**
  - highlight spurious (non-causal) variation by dividing the data into different environments.
  - The same predictor, if based on causal features, should remain **optimal** in each environment separately

A toy example

$$p_{\boldsymbol{X}_1}(1) = 0.5$$

$$p_{Y|\boldsymbol{X}_1}(1|1) = p_{Y|\boldsymbol{X}_1}(0|0) = 0.9$$
$$p_{\boldsymbol{X}_2|Y}(1|1) = p_{\boldsymbol{X}_2|Y}(0|0) = 0.9.$$
$$p_{Y|\boldsymbol{X}_2}(1|1) = p_{Y|\boldsymbol{X}_2}(0|0) = 0.9,$$

In this case, there is no reason for MMI to favor X1 over the others.

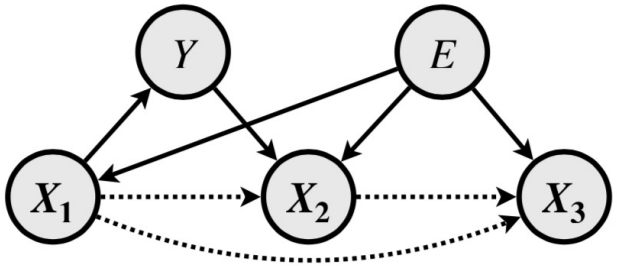


Figure 2. A probabilistic model illustrating different parts of an input that have different probabilistic relationships with the model output Y. A sentence $\boldsymbol{X}$ can be divided into three variables $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$. All $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$ can be highly correlated with Y, but only $\boldsymbol{X}_1$ is regarded as a plausible explanation.

Adversarial Invariant Rationalization

- Exclude rationales with spurious correlations by utilizing the extra information provided by an environment variable.

- If E = e1, all prior distributions are same as before.
- If E = e2, the priors are almost the same, except for the prior of X1

$$q_{\boldsymbol{X}}(\cdot) = p_{\boldsymbol{X}|E}(\cdot|e_2)$$
$$q_{\boldsymbol{X}_1}(1) = 0.6.$$
$$q_{Y|\boldsymbol{X}_2}(1|1) \approx 0.926,$$
$$q_{Y|\boldsymbol{X}_2}(0|0) \approx 0.867,$$

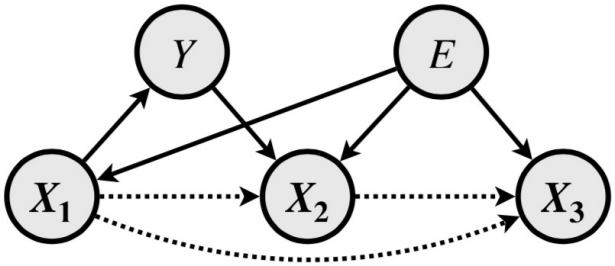
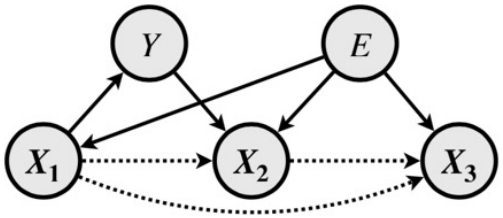


Figure 2. A probabilistic model illustrating different parts of an input that have different probabilistic relationships with the model output Y. A sentence $\boldsymbol{X}$ can be divided into three variables $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$. All $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$ can be highly correlated with Y, but only $\boldsymbol{X}_1$ is regarded as a plausible explanation.

Adversarial Invariant Rationalization

- Key: Y is independent of E *only when* conditioned on X1 , so $p_{Y|X1}(\cdot|\cdot)$ would not change with E. We call this property *invariance*.



- Invariant rationalization objective

$$\max_{\boldsymbol{m}\in\mathcal{S}} I(Y;\boldsymbol{Z}) \quad \text{s.t.} \quad \boldsymbol{Z}=\boldsymbol{m}\odot\boldsymbol{X}, \quad Y \perp E \mid \boldsymbol{Z},$$

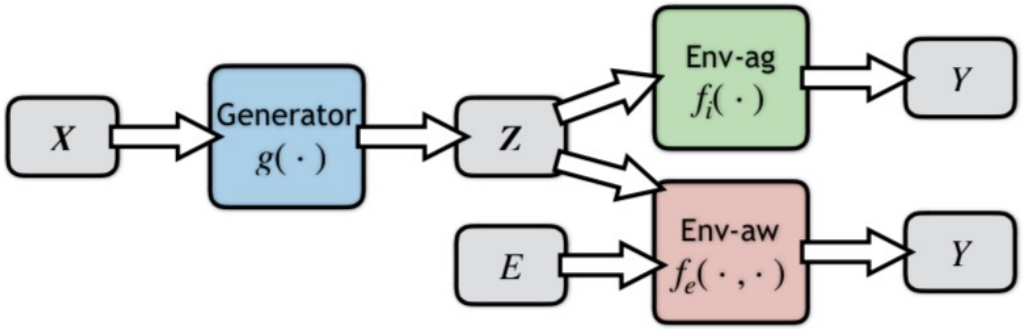
mutual information

m: a binary mask
S: a subset of $\{0, 1\}^N$

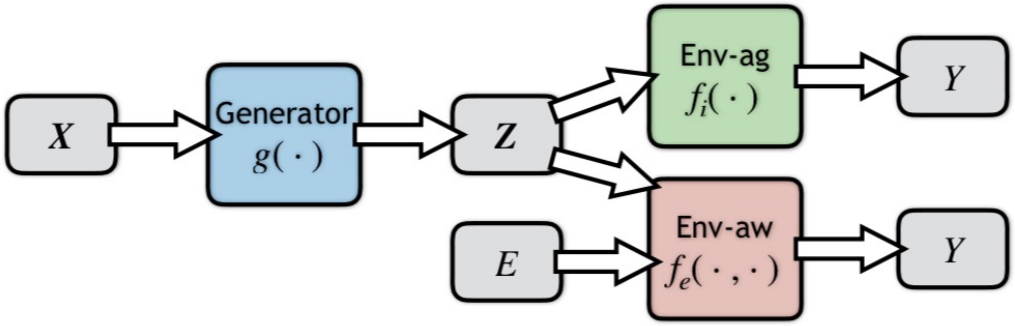
invariance constraint

The INVRAT Framework

- Extend the IRM principle to neural predictions with a game-theoretic framework to impose invariance.
- Three players:
  - the rationale generator,
  - environment-agnostic predictor,
  - environment-aware predictor.

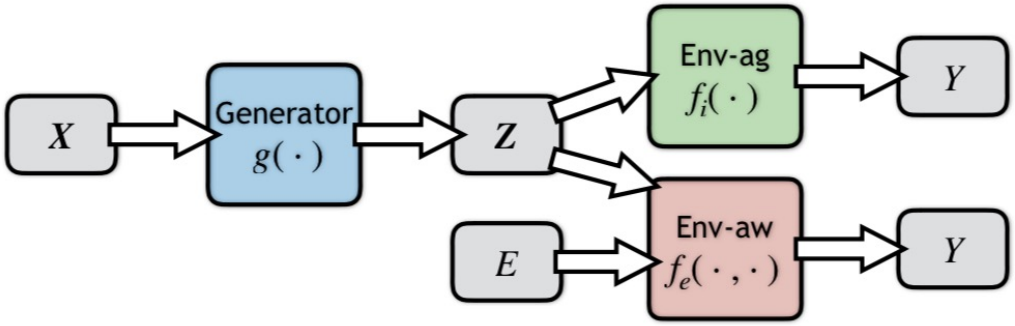


The INVRAT Framework



$$\min_{g(\cdot), f_i(\cdot)} \max_{f_e(\cdot,\cdot)} \mathcal{L}_i(g, f_i) + \lambda h(\mathcal{L}_i(g, f_i) - \mathcal{L}_e(g, f_e)),$$

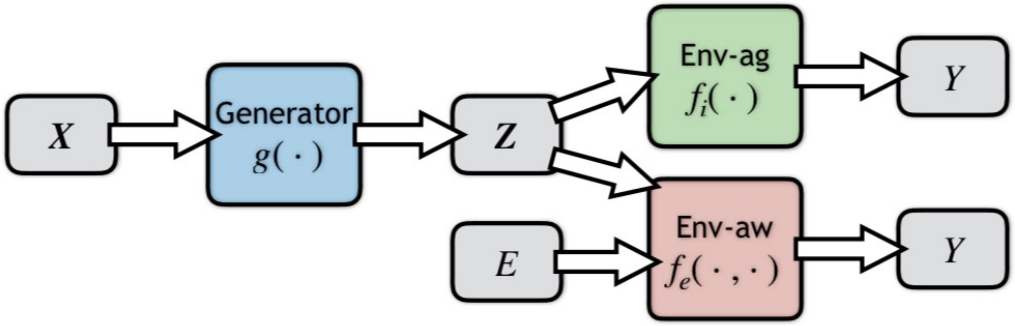
The INVRAT Framework



$$\min_{g(\cdot),f_i(\cdot)} \max_{f_e(\cdot,\cdot)} \mathcal{L}_i(g,f_i) + \lambda h(\mathcal{L}_i(g,f_i) - \mathcal{L}_e(g,f_e)),$$

Loss of environment-aware predictor

The INVRAT Framework



$$\min_{g(\cdot),f_i(\cdot)} \max_{f_e(\cdot,\cdot)} \mathcal{L}_i(g,f_i) + \lambda h(\mathcal{L}_i(g,f_i) - \mathcal{L}_e(g,f_e)),$$

Loss of environment-agnostic predictor

Invariance and Generalizability

- Does keeping the invariant rationales and dropping the non-invariant rationales improve the generalizability in the unknown test environment?

Theorem 1. *Assume the probabilistic graph in figure 2 and that there are two environments e_t and e_a. $\boldsymbol{Z} = \boldsymbol{X}_1$ achieves the saddle point of the following minimax problem*

$$\min_{\boldsymbol{Z} \in \mathcal{X}} \max_{\pi_1, \pi_2, \pi_3} \mathcal{L}^*_{test}(\boldsymbol{Z}; \pi_1, \pi_2, \pi_3),$$

where $\mathcal{X}$ denotes the power set of $[\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3]$.

The invariance rationale has a nice property which minimizes the risk under the most adverse test environment

Experiments

- Dataset
  - Multi-aspect beer reviews
  - IMDB
- Two environments created
  - manually injects tokens with false correlations with Y
  - The goal is to validate if the proposed method *excludes* these tokens from rationale selections

Experiments

Table 1. Results on the synthetic IMDB dataset. The last column is the percentage of testing examples with the injected punctuation selected as a part of the rationales. The best test results are **bolded**.

	Dev Acc	Test Acc	Bias Highlighted
RNP	78.90	72.25	78.24
INVRAT	86.65	**87.05**	**0.00**

- RNP relies on these injected punctuation, whose probabilistic distribution varies drastically between training set and test set => poor generalizability, low predictive accuracy on test set.

Experiments

Table 2. Experimental results on the multi-aspect beer reviews. We compare with the baselines on highlight lengths of 10, 20 and 30. For each aspect and length, we report the best accuracy on the validation set and its corresponding performance on the human annotation set. The best precision (P), recall (R) and F1 score are **bolded**.

Methods	Len	Appearance				Aroma				Palate			
		Dev Acc	P	R	F1	Dev Acc	P	R	F1	Dev Acc	P	R	F1
RNP	10	75.20	13.51	5.75	8.07	75.30	30.30	15.26	20.30	75.00	28.20	17.24	21.40
3PLAYER	10	77.55	15.84	6.78	9.50	80.75	**48.85**	**24.43**	**32.57**	76.60	14.15	8.54	10.65
INVRAT	10	75.65	**49.54**	**20.93**	**29.43**	77.95	48.21	24.36	32.36	76.10	**32.80**	**20.01**	**24.86**
RNP	20	77.70	13.54	11.29	12.31	78.85	34.32	34.18	34.25	77.10	19.80	23.78	21.60
3PLAYER	20	82.56	15.63	13.47	14.47	82.95	35.73	35.89	35.81	79.75	20.73	24.91	22.63
INVRAT	20	81.30	**58.03**	**49.59**	**53.48**	81.90	**42.72**	**42.52**	**42.62**	80.45	**44.04**	**52.75**	**48.00**
RNP	30	81.65	26.26	33.10	29.29	83.10	39.97	60.13	48.02	78.55	19.18	33.81	24.47
3PLAYER	30	80.55	12.56	15.90	14.03	84.40	33.02	49.66	39.67	81.85	21.98	39.27	28.18
INVRAT	30	82.85	**54.03**	**69.23**	**60.70**	84.40	**44.72**	**67.35**	**53.75**	81.00	**26.51**	**46.91**	**33.87**

Experiments

Beer - Appearance Rationale Length - 20

into **a pint glass , poured a solid black , not so much head but enough , tannish in color ,** decent lacing down the glass . as for aroma , if you love coffee and beer , its the best of both worlds , a very fresh strong full roast coffee blended with (and almost overtaking) a solid , classic stout nose , with the toasty , chocolate malts . with the taste , its even more coffee , and while its my dream come true , so delicious , what with its nice chocolate and burnt malt tones again , but i almost say it <unknown> any <unknown> , and takes away from the beeriness of this beer . which is n't to say it is n't delicious , because it is , just seems a bit unbalanced . oh well ! the mouth is pretty solid , a bit light but not all that unexpected with a coffee blend . its fairly smooth , not quite creamy , well carbonated , thoroughly , exceptionally drinkable .

Beer - Aroma Rationale Length - 20

into a pint glass , poured a solid black , not so much head but enough , tannish in color , decent lacing down the **glass . as for aroma , if you love coffee and beer , its the best of both worlds ,** a very fresh strong full roast coffee blended with (and almost overtaking) a solid , classic stout nose , with the toasty , chocolate malts . with the taste , its even more coffee , and while its my dream come true , so delicious , what with its nice chocolate and burnt malt tones again , but i almost say it <unknown> any <unknown> , and takes away from the beeriness of this beer . which is n't to say it is n't delicious , because it is , just seems a bit unbalanced . oh well ! the mouth is pretty solid , a bit light but not all that unexpected with a coffee blend . its fairly smooth , not quite creamy , well carbonated , thoroughly , exceptionally drinkable .

Beer - Palate Rationale Length - 20

into a pint glass , poured a solid black , not so much head but enough , tannish in color , decent lacing down the glass . as for aroma , if you love coffee and beer , its the best of both worlds , a very fresh strong full roast coffee blended with (and almost overtaking) a solid , classic stout nose , with the toasty , chocolate malts . with the taste , its even more coffee , and while its my dream come true , so delicious , what with its nice chocolate and burnt malt tones again , but i almost say it <unknown> any <unknown> , and takes away from the beeriness of this beer . which is n't to say it is n't delicious , because it is , just seems a bit unbalanced . oh well ! the **mouth is pretty solid , a bit light but not all that unexpected with a coffee blend . its fairly** smooth , not quite creamy , well carbonated , thoroughly , exceptionally drinkable .

Figure 4. Examples of INVRAT generated rationales on the multi-aspect datasets. Human annotated words are underlined. Appearance, aroma and palate rationales are in bold text and highlighted in **green**, **red**, and **blue** respectively.

References & Reading Materials

- Arjovsky M, Bottou L, Gulrajani I, et al. Invariant risk minimization[J]. arXiv preprint arXiv:1907.02893, 2019.
  - https://arxiv.org/pdf/1907.02893.pdf;
- Chang S, Zhang Y, Yu M, et al. Invariant rationalization[C]//International Conference on Machine Learning. PMLR, 2020: 1448-1458.
  - https://proceedings.mlr.press/v119/chang20c/chang20c.pdf

Thank you!
Questions?