The Risks of Invariant Risk Minimization

Elan Rosenfeld  Pradeep Ravikumar  Andrej Risteski
Machine Learning Department, Carnegie Mellon University

Invariance Causal Prediction

Q: How can we train a predictor to ignore non-causal features whose correlation with the target may not hold at test time?

Assume training data can be partitioned into distinct environments. Each environment represents an intervention, inducing a different joint distribution.

Across environments, the causal mechanism $P(Y \mid \text{Parents}(Y))$ remains fixed. Recovering precisely the parents of $Y$ ensures that predictions are invariant and therefore minimax across all possible interventions.

Deep Invariant Feature Learning

Q: How can we accomplish this when the features are latent?

Learn a feature embedder $\Phi$ such that the resulting feature distribution $\Phi(X)$ induces invariance. Some recent suggestions:

\[ \text{IRM: Invariant } \mathbb{E}[Y \mid \Phi(X)] \]

Require optimal regression vector $\beta$ on top of features to be invariant.

\[ \text{REx: Invariant } \mathbb{V}[Y \mid \Phi(X)] \]

Require equal risk across environments.

A Formal Model of Latent Invariant Features

\[ y = \begin{cases} +1, \text{ w.p. } \eta, \\ -1, \text{ otherwise} \end{cases} \]

\[ z = \begin{cases} z_c \sim \mathcal{N}(\mu_c(y), \sigma^2_c I), \\ z_r \sim \mathcal{N}(\mu_r(y), \sigma^2_r I) \end{cases} \]

\[ x = f(z) \]

We present the first analysis under nonlinear observations of the latent variables $p(y, z_c)$ is invariant, capturing the intuition behind invariant feature learning.

Under this model, our goal is to learn a feature embedder $\Phi^*$ which recovers just the invariant features: $\Phi^*(x) = z_c$. We would then also learn the regression vector $\beta^* = \arg\min_{\beta} R(\beta \ast \Phi^*)$.

We call the predictor $\beta^* \ast \Phi^*$ the optimal invariant predictor (OIP). We assume we observe $E$ environments, with infinite samples. Let $d'_{\beta}$ be the dimensionality of $z_c$. Typically expect $d'_{\beta} \gg E$.

Linear Observations

For conditionally Gaussian features, optimal classifier is $\beta^* = \Sigma^{-1}(\mu_1 - \mu_0)$.

So long as this vector is the same for all environments, the solution is feasible under the IRM objective.

If $E \leq d_c$: We construct a feasible linear $\Phi$ which recovers $z_c$ plus an additional set of features which depend on the non-invariant latents $z_r$.

Provably has lower training risk than the OIP.

If $E > d_c$: We prove that any feasible linear $\Phi$ can only depend on $z_c$.

Among such $\Phi$, the OIP has lowest training risk.

Corollary: the optimal invariant predictor is the global minimum of the IRM objective if and only if $E > d_c$.  

Nonlinear Observations

We study IRMv1, a regularized form of IRM used in practice which penalizes the environmental gradient norm of the classifier.

We construct a predictor $\beta \ast \Phi$ with the following properties:

- Penalty term is exponentially small in dimension $d'$.
- Exactly equivalent to the OIP on all but an exponentially small fraction of the training distribution.
  - polynomial # of samples \( \Rightarrow \) indistinguishable!
- On any environment slightly different from the training environments, it is exactly equivalent to the ERM-optimal solution on all but an exponentially small fraction of the test data.

Implication: the solution behaves just like ERM at test time! Furthermore, results apply to all recently proposed variants.