Learning to Pivot with Adversarial Networks

Gilles Louppe, Michael Kagan and Kyle Cranmer

✉ g.louppe@ulg.ac.be, makagan@slac.stanford.edu, kyle.cranmer@nyu.edu

https://github.com/glouppe/paper-learning-to-pivot/

Abstract

Aim to build a classifier that is robust to changes in the data distribution over a continuous family of data generating processes. Such domain adaptation problems arise in the scientific context when systematic uncertainties define a set of plausible data generating processes, in cases of fairness with respect to continuous attributes, or generally when one desires a classifier to be invariant to changes in a given feature. Robust inference is achieved through building a pivot, a quantity whose distribution does not depend on the unknown values of the nuisance parameter that parametrize the family of data generating processes. We introduce and derive theoretical results for a training procedure based on adversarial networks for enforcing the pivotal property on a predictive model.

Problem Statement

Data generating process $p(X, Y, Z)$
$x \in \mathcal{X}$ are the data
$y \in \mathcal{Y}$ are the target labels
$z \in \mathcal{Z}$ are the nuisance parameters
continuous or categorical

Goal : Learn regression function $f : \mathcal{X} \to \mathcal{S}$ with $s \in \mathcal{S} = \mathbb{R}^{|\mathcal{Y}|}$ with parameters θ_f that minimize loss $\mathcal{L}_f(\theta_f)$ (e.g., the cross-entropy) such that

$$p(f(X;\theta_f) = s|z) = p(f(X;\theta_f) = s|z')$$

Method

$r := p_{\theta_r}(z|f(X;\theta_f) = s)$ is the adversary model with parameters θ_r and loss $\mathcal{L}_r(\theta_f, \theta_r)$. As with generative adversarial networks, f and r are trained simultaneously, which is carried out by considering the value function

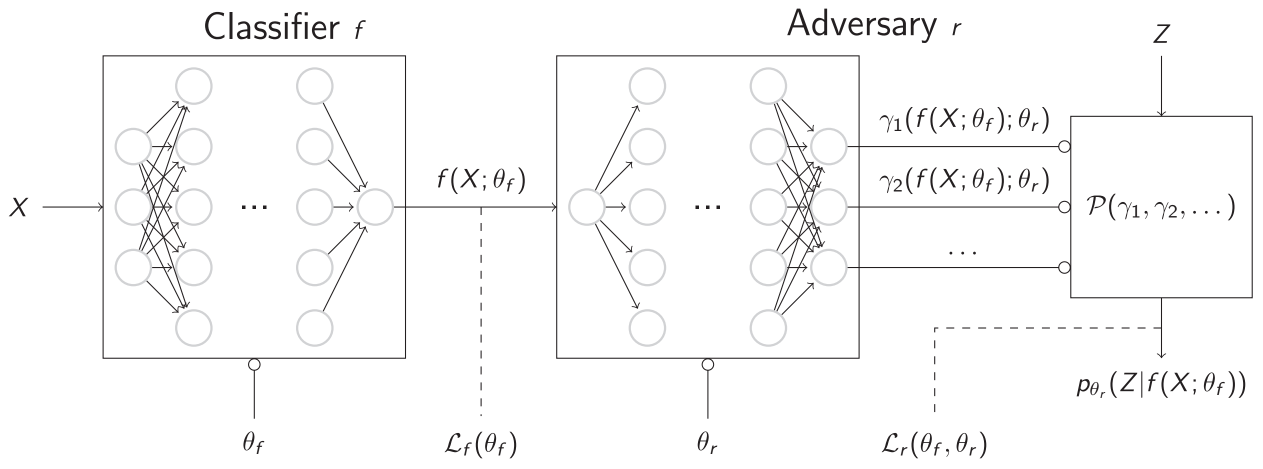
$$E(\theta_f, \theta_r) = \mathcal{L}_f(\theta_f) - \mathcal{L}_r(\theta_f, \theta_r) \tag{1}$$

that we optimize by finding the minimax solution

$$\hat{\theta}_f, \hat{\theta}_r = \arg \min_{\theta_f} \max_{\theta_r} E(\theta_f, \theta_r) \tag{2}$$

$\mathcal{L}_f$ and $\mathcal{L}_r$ are respectively set to the expected value of the negative log-likelihood of $Y|X$ under f and of $Z|f(X;\theta_f)$ under r :

$$\mathcal{L}_f(\theta_f) = \mathbb{E}_{x\sim X}\mathbb{E}_{y\sim Y|x}[-\log p_{\theta_f}(y|x)] \tag{3}$$
$$\mathcal{L}_r(\theta_f, \theta_r) = \mathbb{E}_{s\sim f(X;\theta_f)}\mathbb{E}_{z\sim Z|s}[-\log p_{\theta_r}(z|s)] \tag{4}$$


Classifier f
Adversary r
X
$\cdots$
$f(X;\theta_f)$
$\cdots$
$\gamma_1(f(X;\theta_f);\theta_r)$
$\gamma_2(f(X;\theta_f);\theta_r)$
$\dots$
Z
$\mathcal{P}(\gamma_1, \gamma_2, \dots)$
$p_{\theta_r}(Z|f(X;\theta_f))$
θ_f
$\mathcal{L}_f(\theta_f)$
θ_r
$\mathcal{L}_r(\theta_f, \theta_r)$

Theoretical results

Proposition 1. *If there exists a minimax solution $(\hat{\theta}_f, \hat{\theta}_r)$ for Eqn. (2) such that $E(\hat{\theta}_f, \hat{\theta}_r) = H(Y|X) - H(Z)$, then $f(\cdot;\hat{\theta}_f)$ is both an optimal classifier and a pivotal quantity.*

Proof. (*sketch*)

(i) For fixed θ_f, adversary r is optimal at $\hat{\hat{\theta}}_r = \arg\max_{\theta_r} E(\theta_f, \theta_r) = \arg\min_{\theta_r} \mathcal{L}_r(\theta_f, \theta_r)$, in which case $p_{\hat{\hat{\theta}}_r}(z|f(X;\theta_f) = s) = p(z|f(X;\theta_f) = s)\ \forall z, s$, and $\mathcal{L}_r \to \mathbb{E}_{s\sim f(X;\theta_f)}[H(Z|f(X;\theta_f) = s)]$;

(ii) Value function E can be restated as a function of θ_f only : $E'(\theta_f) = \mathcal{L}_f(\theta_f) - H(Z|f(X;\theta_f))$;

(iii) Have lower bound $H(Y|X) - H(Z) \leq \mathcal{L}_f(\theta_f) - H(Z|f(X;\theta_f))$ where the equality holds at $\hat{\theta}_f = \arg\min_{\theta_f} E'(\theta_f)$ when :
- $\hat{\theta}_f$ minimizes the negative log-likelihood of $Y|X$ under f, which happens when $\hat{\theta}_f$ are the parameters of an optimal classifier. Then, $\mathcal{L}_f$ reduces to minimum value $H(Y|X)$.
- $\hat{\theta}_f$ maximizes the conditional entropy $H(Z|f(X;\theta_f))$, since $H(Z|f(X;\theta)) \leq H(Z)$.

$\Rightarrow$ By assumption, the lower bound is active, thus we have $H(Z|f(X;\theta_f)) = H(Z)$ because of the second condition, which happens exactly when Z and $f(X;\theta_f)$ are independent variables. In other words, the optimal classifier $f(\cdot;\hat{\theta}_f)$ is also a pivotal quantity. □

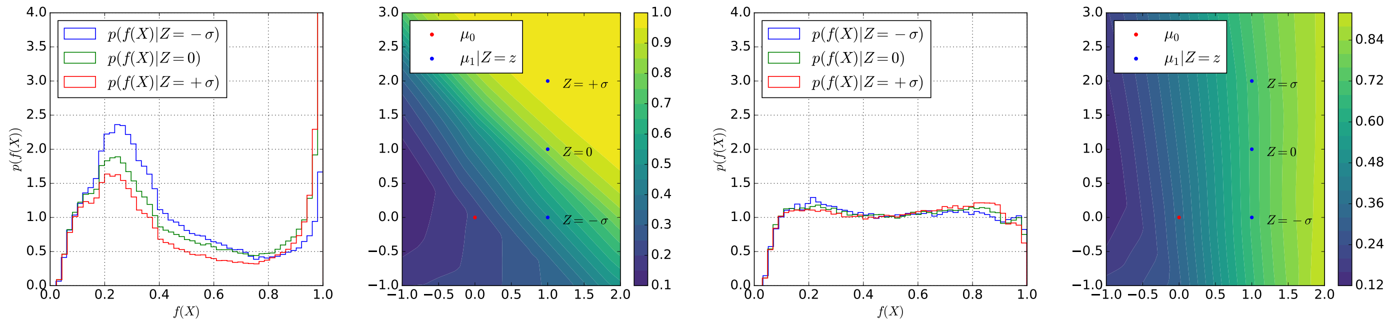
Use in Practice

The assumption of existence of an optimal and pivotal classifier may not hold because the nuisance parameter directly shapes the decision boundary. In this case, the lower bound is strict : f can either be an optimal classifier or a pivotal quantity, but not both simultaneously. In this situation, it is natural to rewrite the value function E as

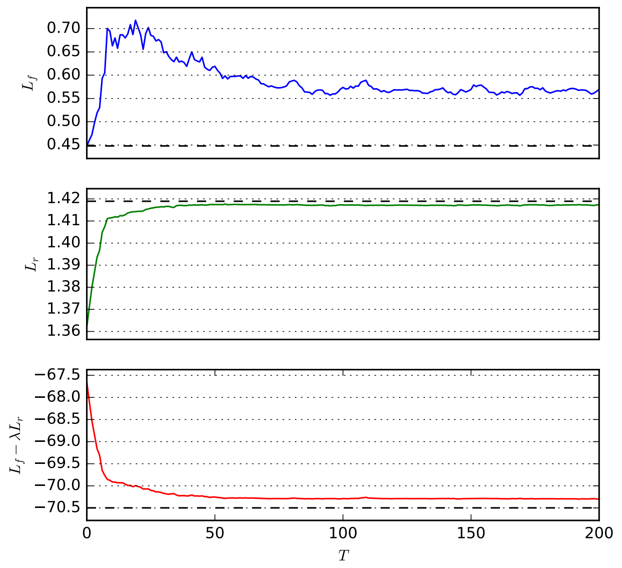
$$E_\lambda(\theta_f, \theta_r) = \mathcal{L}_f(\theta_f) - \lambda\mathcal{L}_r(\theta_f, \theta_r) \tag{5}$$

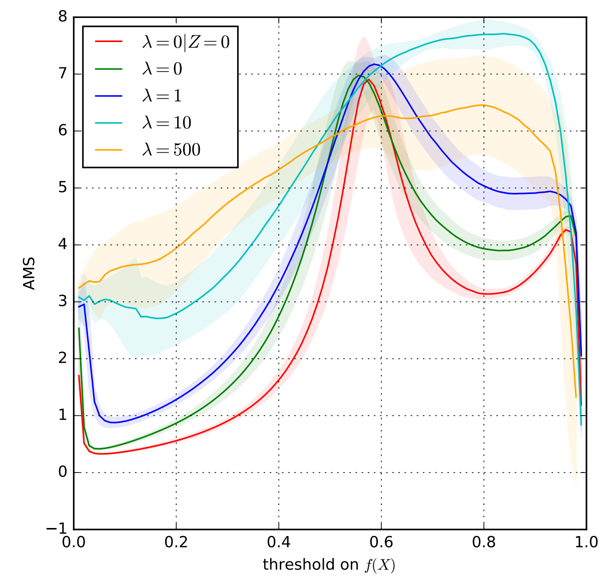
where $\lambda \geq 0$ is a hyper-parameter controlling the trade-off between the performance of f and its independence with respect to the nuisance parameter. Setting λ to a large value preferably enforces f to be pivotal while setting λ close to 0 constrains f to be optimal.

Illustration



Physics Example





Conclusions

✓ Proposed a flexible learning procedure for building a predictive model that is independent of continuous or categorical nuisance parameters by jointly training two neural networks in an adversarial fashion.

✓ Motivated the proposed algorithm by showing that the minimax value of its value function corresponds to a predictive model that is both optimal and pivotal (if that models exists) or for which one can tune the trade-off between power and robustness.

✓ Empirically point confirmed the effectiveness of our method on a toy example and a particle physics example.

□ Proposed solution can be used in any situation where the training data may not be representative of the real data the predictive model will be applied to in practice.

□ In the scientific context, the presence of systematic uncertainty can be incorporated by considering a family of data generation processes.

□ The approach also extends to cases where independence of the predictive model with respect to observed random variables is desired, as in fairness for classification.

KC and GL are both supported through NSF ACI-1450310, additionally KC is supported through PHY-1505463 and PHY-1205376. MK is supported by the US Department of Energy (DOE) under grant DE-AC02-76SF00515 and by the SLAC Panofsky Fellowship.