ICML2017
International Conference on Machine Learning 2017
Aug 6 , 2017 - Aug 12, 2017, Sydney, Australia

Reviews For Paper

Paper ID 19

Title Learning to Pivot with Adversarial Networks

Masked Reviewer ID: Assigned_Reviewer_1

Review:

Question	
Summary of the paper (Summarize the main claims/contributions of the paper.)	The paper's idea is to use adversarial training to learn with a pivotal quantity, i.e. learning where one wants to be independent with respect to a certain quantity. Applications are demonstrated on synthetic data and a particle physics problem.
Clarity (Assess the clarity of the presentation and reproducibility of the results.)	Above Average
Clarity - Justification	The paper is generally well-written, and presents ideas logically.
Significance (Does the paper contribute a major breakthrough or an incremental advance?)	Below Average
Significance - Justification	The technical contribution, while intuitive, seems a bit modest when viewed from the perspective of Eqn 8: effectively, one applies a regulariser to the original objective which encourages maximal entropy of the pivotal quantity Z given the model predictions. Similar ideas have been considered in different contexts, e.g. in fairness. The difference in this paper is that the regulariser is expressed as the result of an optimisation problem, following GANs -- certainly a nice approach, but to me a bit unsuprising.
Correctness (Is the paper technically correct?)	Paper is technically correct
Correctness - Justification	I did not find any errors in the paper.
Overall Rating	Weak reject
Detailed comments. (Explain the basis for your ratings while providing constructive feedback.)	The paper's idea is to use adversarial training to learn with a pivotal quantity, i.e. learning where one wants to be independent with respect to a certain quantity. This is intuitive and sensible. Further, the applications where this is demonstrated, particularly in particle physics, are certainly interesting. I feel however that the paper has a couple of limitations: - the technical contribution, while intuitive, seems a bit modest when viewed from the perspective of Eqn 8: effectively, one applies a regulariser to the original objective which encourages maximal entropy of the pivotal quantity Z given the model predictions. Similar ideas have been considered in different contexts, e.g. in fairness under Kamishima et al. Fairness-aware Classifier with Prejudice Remover Regularizer. ECML 2012. In this way, the contribution of the paper appears to be the observation that such a regulariser may be re-expressed as the result of an optimisation problem, following the similar observation in GANs regarding a divergence between distributions. This is indeed a nice observation, but a little unsurprising in my estimation. - the experiments don't appear to compare the proposed approach to any baselines. This means while the results demonstrate that the method indeed works, there isn't a sense of whether some other approach that might also work well (and possibly with significantly reduced computational cost). I'm

	not familiar with the literature on learning with pivotal quantities, so can't comment on appropriate baselines for the general case (if these don't exist, then this must be made explicit). However, a specific case where there is certainly scope for comparison would be the fairness application, which as noted can be viewed as learning with a kind of pivotal quantity; there are several extant methods here (some of which are cited in the Related Work), and these rely on rather different means of achieving the independence desideratum; e.g. the work of Zafar et al., "Learning Fair Classifiers", or the Kamishima et al. method cited above. Other comments: - more discussion of the precise connection to GANs may be prudent; in particular, one could view the proposed approach as a particular min-max problem that leads to an optimal solution with the pivotal independence property. Indeed, the GAN idea of representing a probabilistic model implicitly is absent from this work. To me, the connection is clearest in viewing the conditional entropy as the result of an optimisation (c.f. Jensen-Shannon in the GAN paper). - consider adding appropriate citations in the Introduction, e.g. for pivotal quantities, GANs, et cetera - \mathcal{S} in f : \mathcal{X} -> \mathcal{S} appears undefined - use \rightarrow instead of \mapsto when referring to function specifications. - Eqn 1 implicitly assumes the scores are discrete? - "against an ADVERSARIAL classifier" - Eqn 13 should have conditioning on Z?
Reviewer confidence	Reviewer is knowledgeable

Masked Reviewer ID: Assigned_Reviewer_2

Review:

Question	
Summary of the paper (Summarize the main claims/contributions of the paper.)	This paper attempts to learn predictions invariant to a set of known "nuisance parameters," which are assumed to be provided during training but unavailable at test time; the goal is to make test-time predictions robust with respect to the value of these parameters. To accomplish this, the authors take a GAN-type approach in which an adversary attempts to reconstruct the value of the nuisance parameters from the predictions; the predictor attempts to simultaneously produce good predictions as well as preventing the adversary from inferring the values of the nuisance parameters based on those predictions.
Clarity (Assess the clarity of the presentation and reproducibility of the results.)	Above Average
Clarity - Justification	The paper's development is reasonable and fairly straightforward. A few minor suggestions for improvement: Figure 1 should clarify what the \gamma_i are, since they are not defined until the next page. Adding the domains of various parameters here might help as well. On line 138, S should be explicitly defined, and "of parameters \theta_f" should be rephrased to make it clearer that \theta_f are properties of the regression function f, not the outputs of the regression model somehow; just changing to "with parameters \theta_f" would probably help.
Significance (Does the paper contribute a major breakthrough or an incremental advance?)	Below Average
Significance - Justification	This framing of nuisance parameters with respect to scientific inference is very interesting; however, it seems from the point of the models very similar to much of the work on fair learning algorithms. You discuss this relationship briefly, but it would be useful to know: what is the fundamental difference between your work and especially that of Edwards and Storkey (2015)? Is it only the support for continuous variables, or is there something more fundamental? Generally, the model seems reasonable and I like the reframing of nuisance parameters, but its

	extension over related work should be clearer, the theoretical result is very limited, and the experimental results not that informative.
Correctness (Is the paper technically correct?)	Paper is technically correct
Correctness - Justification	The proof of Proposition 1 seems reasonable, though I did not verify it in great detail.
Overall Rating	Weak reject
Detailed comments. (Explain the basis for your ratings while providing constructive feedback.)	One interesting question which you did not seem to address: how does the prior distribution on Z affect the final predictions? One of the advantages of toy models is that you can often compare your model's behavior to that of the theoretically optimal model, or a simple similar model, to gain insight about what the model picks up on and what it doesn't. In the toy experiment of your Section 5.1: - You could plot the true posterior density calculated with the training data's prior on Z. Does that look essentially like the non-adversarially trained classifier's output? - It seems like the outputs of your model, both the adversarial and non-adversarial ones, are not too different from the output of a logistic regression model. It seems likely that you could relatively easily compute the relevant conditional entropies for a logistic regression model -- or perhaps a probit regression might be easier -- and relate the behavior of that model to yours in this case. This could help understand the space of tradeoffs between optimal prediction and pivotal-ness in this problem, and how well your adversary tracks the true conditional entropies in this case. - You could maybe even just discretize the problem and compute the conditional entropies for any arbitrary conditional density, to see if any possible output could be both pivotal and optimal. This optimization problem might be too difficult, though. For the physics experiment: you show that better final performance (in terms of the approximate median significance) is possible for certain values of lambda, but that performance is worse both for too-small and too-large values. Is it reasonable to choose a value of lambda based on performance on a validation set, as is typically done for choosing regularization values? How reliable is that estimate? An experiment doing so would be helpful.
Reviewer confidence	Reviewer is knowledgeable

Masked Reviewer ID: Assigned_Reviewer_3

Review:

Question	
Summary of the paper (Summarize the main claims/contributions of the paper.)	This paper proposes a method to learn a regression function f(x) from X to Y, where the decision boundary is robust to a nuisance random variable Z. In order to do so, an adversarial training procedure is used to force the distribution of f(X) to be independent of Z.
Clarity (Assess the clarity of the presentation and reproducibility of the results.)	Excellent (Easy to follow)
Clarity - Justification	The paper was clear and easy to follow.
Significance (Does the paper contribute a major breakthrough or an incremental advance?)	Above Average
Significance - Justification	See the detailed comments.
Correctness (Is the	Paper is technically correct

paper technically correct?)	
Correctness - Justification	See the detailed comments.
Overall Rating	Strong accept
Detailed comments. (Explain the basis for your ratings while providing constructive feedback.)	The problem of learning representation that is independent of nuisance variable is important with many practical applications. This problem has been studied in the context of domain adaptation and learning fair representation. While I am not expert in the field of fair representation, I don't think the proposed adversarial training approach is a huge leap forward in this field. For example, the idea of trying to achieve independence by a distribution matching objective is used in [1] where an MMD objective is used on the posterior distribution of a VAE to achieve independence. However it was interesting to see that replacing the MMD objective with the adversarial training as done in this paper would work as well. Another complaint that I have is that the experimental results are not compared against any other method in the literature. On the positive side, this paper has some nice theoretical results with an experiment of a real world application showing the effectiveness of this approach. Also the toy example was very nice and gives insight to how the method works. In short, I don't think this idea is a huge leap forward in the field, but this is a nice paper and I recommend acceptance. [1] The Variational Fair AutoEncoder
Reviewer confidence	Reviewer is knowledgeable