Robust Covariate Shift Prediction with Exact Loss Functions

Anqi Liu, Brian D. Ziebart
Department of Computer Science
University of Illinois at Chicago
Chicago, IL, 60607
{aliu33,bziebart}@uic.edu

Abstract

Covariate shift problems widely exist in machine learning since the *independent and identically distributed* (IID) assumption is often violated by the testing input distributions differing from the training input distribution. Recently developed methods that construct a predictor that is inherently robust to the difficulties of learning under covariate shift are restricted to minimizing expected worst-case testing logloss. However, it is important for many application that the loss functions optimized in learning align with the desired application performance metric. In this paper, we address this limitation by proposing a general formulation that robustly minimizing various loss functions, including non-convex ones, under the testing distribution. This generalization makes robust covariate shift prediction applicable to more task scenarios. We demonstrate the benefits on covariate shift classification tasks.

1 Introduction

The *independent and identically distributed* (IID) assumption employed widely across machine learning methods requires the testing data distribution to be the same as the training data distribution. This is quite restrictive in the sense that shift can occur between the training distribution and testing distribution in many settings, which makes models built on the IID assumption inappropriate. Specifically, the predictor minimizing (regularized) loss defined on the training samples provides no performance guarantees when applied on the testing distribution [1, 2]. Though nothing can be learned when the shift between training and testing data is arbitrary, certain assumptions about how training and testing distributions differ allow reasonable adaptive learning methods to be derived [3]. One of the common assumptions is that the bias just comes from the input variables. In this setting, which is known as **covariate shift**, only the distribution of inputs, $P_{\text{train}}(\mathbf{x})$ and $P_{\text{test}}(\mathbf{x})$, differ, while the conditional label distribution, $P(y|\mathbf{x})$, is the same under both the training and the testing distributions. This assumption is much weaker than the IID assumption and covers a broad range of real application scenarios.

The most prevalent methods for addressing covariate shift attempt to debias the training data by reweighting it using a density ratio, $P_{\text{test}}(\mathbf{x})/P_{\text{train}}(\mathbf{x})$. This approach tends to work well when the training and the testing distributions are fairly similar and large amounts of training samples are available. It also enjoys consistency guarantees when provided with an infinite amount of training data. However, when these conditions are violated, i.e., there is only a limited amount of training data and/or significant differences between the training and testing distributions, some of the density ratios for training examples can be extremely large. This leads to high-variance estimates that extrapolate heavily from scant amounts of training data and a lack of generalization guarantees for the resulting predictor [4, 5].

Recently developed robust covariate shift methods take a worst-case approach, constructing a predictor that (approximately) matches training data statistics, but is otherwise the most uncertain on the testing distribution [6, 7]. These methods were built by minimizing the worst case expected target logloss and obtain a parametric form of the predicted output labels' probability distributions. Unfortunately, log loss may not be of interest for many applications and robust accuracy maximization is instead desired, for example. In this paper, we generalize robust covariate shift classification framework to robustly minimize other loss functions, like the 0-1 loss, under covariate shift. We show that even though we cannot obtain parametric forms of the predictor generally, we are able to solve the problem by stochastic (sub-)gradient descent in most cases, with error in the test worst-case loss bounded. We demonstrate the effectiveness of our method using both synthetic examples and real datasets.

2 Background

Under covariate shift, the training distribution and testing distribution share the same conditional label distribution, $P(y|\mathbf{x})$, but have differing distributions over inputs: $P_{\text{train}}(\mathbf{x}, y) = P_{\text{train}}(\mathbf{x})P(y|\mathbf{x})$; $P_{\text{test}}(\mathbf{x}, y) = P_{\text{test}}(\mathbf{x})P(y|\mathbf{x})$. The most prevalent approach for addressing covariate shift attempts to remove the bias between the training and testing distributions [8, 9, 10]. Under this perspective, minimizing the importance-weighted loss of (n) training examples,

$$\lim_{n\to\infty} \min_{\hat{f}} \mathbb{E}_{(\mathbf{X},Y)\sim\tilde{P}^{(n)}_{\text{train}}} \left[\frac{P_{\text{test}}(\mathbf{X})}{P_{\text{train}}(\mathbf{X})} \text{loss}(\hat{f}(\mathbf{X}), Y) \right] = \min_{\hat{f}} \mathbb{E}_{(\mathbf{X},Y)\sim P_{\text{test}}} \left[\text{loss}(\hat{f}(\mathbf{X}), Y) \right], \quad (1)$$

where $\hat{f}$ is estimated predictor and $\tilde{p}$ is the empirical distribution of data, asymptotically minimizes the testing distribution loss, so long as $P_{\text{test}}(\mathbf{x}) > 0 \implies P_{\text{train}}(\mathbf{x}) > 0$.

Despite this asymptotic guarantee, predictive performance can be poor when training from finite amounts of samples in both theory and practice. Conceptually, the density ratios of a small number of training examples can become disproportionately large, making the resulting predictor overly sensitive to a small number of training data points—or even one single datapoint. This leads to predictive results with high variance. Indeed, finite generalization bounds for importance-weighted methods require finite second moments: $\mathbb{E}_{P_{\text{train}}(x)}[(P_{\text{test}}(\mathbf{X})/P_{\text{train}}(\mathbf{X}))^2] < \infty$ [4], which is often not satisfied in practice.

Robust covariate shift classification [6] is motivated from minimax robust estimation [11, 12]. The expected testing loss minimization is formulated as a two player game, where the estimator player seeks to minimize the loss function, while an adversarial player tries to maximize it under constraints based on training samples: $\min_{\hat{P}} \max_{\check{P}\in\tilde{\Xi}_{\text{train}}} \mathbb{E}_{\mathbf{X}\sim P_{\text{test}}, \check{Y}|\mathbf{X}\sim\check{P}} \left[-\log \hat{P}(\check{Y}|\mathbf{X}) \right]$. The adversary must choose a distribution $\check{P}$ that is similar to certain measured properties (features), e.g., $\mathbb{E}_{\mathbf{X}\sim\tilde{P}, \check{Y}|\mathbf{X}\sim\check{P}} \left[\phi(\mathbf{X}, \check{Y}) \right] = \mathbb{E}_{(\mathbf{X},Y)\sim\tilde{P}} \left[\phi(\mathbf{X}, Y) \right]$, of the training data. These are denoted by the convex set $\tilde{\Xi}_{\text{train}}$, with ϕ as the feature function. The advantage of this formulation resides in its robustness to the worst possible case of covariate shift and avoidance of huge losses caused by overly optimistic extrapolations. Moreover, the expected testing loss under this formulation is upper bounded by the testing entropy [13]. Recent advances in adversarial loss minimization in the IID setting extend beyond the logloss to non-smooth loss functions, such as the 0-1 loss [14, 15] and ordinal regression loss [16]. In this paper, we establish the general form of the method that minimizes different loss functions under covariate shift.

3 General Loss Formulation

Definition 1. *The* **generalized robust covariate shift classifier** *results from the adversarial loss optimization game with differing training and testing input distributions:*

$$\min_{\hat{P}} \max_{\check{P}} \mathbb{E}_{\mathbf{X}\sim P_{test}} \left[Loss(\check{P}_{\mathbf{X}}, \hat{P}_{\mathbf{X}}) \right] \text{ such that:} \quad (2)$$

$$\mathbb{E}_{\mathbf{X}\sim\tilde{P}_{train}, \check{Y}|\mathbf{X}\sim\check{P}} \left[\phi(\mathbf{X}, \check{Y}) \right] = \mathbb{E}_{(\mathbf{X},Y)\sim\tilde{P}_{train}} \left[\phi(\mathbf{X}, Y) \right],$$

with a loss function that we want to minimize and a feature representation ϕ.

Strong Lagrangian duality holds when $\text{Loss}(\cdot,\cdot)$ is a concave-convex function of $\check{P}$ and $\hat{P}$. This enables us to re-write the game in terms of Lagrangian multipliers θ:

$$\min_{\theta}\min_{\hat{P}}\max_{\check{P}}\mathbb{E}_{\mathbf{X}\sim P_{\text{test}}}\left[\text{Loss}(\check{P}_{\mathbf{X}},\hat{P}_{\mathbf{X}})+\frac{P_{\text{train}}(\mathbf{X})}{P_{\text{test}}(\mathbf{X})}\theta\cdot\phi(\mathbf{X},\check{Y})\right]-\theta\cdot\tilde{\phi}+\epsilon||\theta||_2, \tag{3}$$

where $\tilde{\phi}\triangleq\mathbb{E}_{(\mathbf{X},Y)\sim\tilde{P}_{\text{train}}}[\phi(\mathbf{X},Y)]$ is the feature function evaluated on empirical training data, and we allow ϵ slack for matching the primal constraints, leading to regularization in the dual. The optimization of this objective function is then composed of two steps: first, solve the inner minimax game with respect to $\hat{P}$ and $\check{P}$; second, optimize for θ in the outer minimization to satisfy imposed constraints. We focus our attention on $0-1$ loss, but many other loss functions can also be incorporated.

Classification Losses: Letting $\text{Loss}(\check{P}_{\mathbf{X}},\hat{P}_{\mathbf{X}})=\hat{P}^T C\check{P}$—a bilinear and therefore concave-convex function of $\check{P}$ and $\hat{P}$—allows many classification losses to be represented in the cost matrix C. We can reformulate the inner minimax game as $\min_{\hat{P}}\max_{\check{P}}\mathbb{E}_{\mathbf{X}}[\hat{P}_{\mathbf{X}}^T C'\check{P}_{\mathbf{X}}]$, where $C'=C+\frac{P_{\text{train}}(\mathbf{X})}{P_{\text{test}}(\mathbf{X})}\theta\cdot\phi(\mathbf{X},\check{Y})$. The inner minimax game, which is a two player zero sum game, can be solved by linear programming. Another way to find the equilibrium of the inner minimax game for the special case of 0-1 loss is by seeking an analytical form of the game value as in [14], which brings more computational efficiency. For the outer minimization, we take the subgradient with respect to θ, which we approximate using training samples,

$$\mathbb{E}_{\mathbf{X}\sim P_{\text{train}},\check{Y}|\mathbf{X}\sim\check{P}}\left[\phi(\mathbf{X},\check{Y})\right]-\tilde{\phi}+2\epsilon\theta, \tag{4}$$

and perform subgradient descent.

We show two illustrative examples in Figure 1, where training distribution (solid line) and testing distribution(dashed line) is overlapping in different ways. The prediction color map shows a similar uncertain prediction with logloss-based classifier where there is not enough training data support, like the top right corner in the first figure. Moreover, the 0-1 loss provides more certain prediction in the overlapped region while logloss-based classifier's prediction changes gradually in certainty from the most supported region to the least.

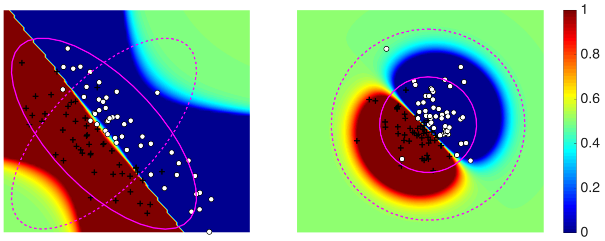

Figure 1: Prediction colormap with robust classifier using 0-1 loss. The colormap shows the $P(Y=$'+'$|\mathbf{x})$. Training data with 5% noise is also shown.

Bounding Expected Worst Case Test Loss: One significant difference between the robust covariate shift methods in this paper and empirical risk minimization based methods is that we directly minimize the worst case expected loss under the test distribution. The reason why this works is because the (sub-)gradient in our formulation only depends on the training distribution, so we are able to use training data to approximate it. On the contrary, the ERM based methods directly approximate the expected loss function using limited data as in Eq.(1). Despite this difference, we can easily control the error in our (sub-)gradients and therefore bound the error in the optimized worst case expected target loss.

We first define the notation $\text{WCLoss}(\theta)$ as the (regularized) worst case loss under the testing distribution, which is equivalent with the Lagrangian form of the optimization game of robust covariate shift classifier in (3). Note that $\text{WCLoss}(\theta)$ differs in meaning from the $\text{Loss}(\hat{P},\check{P})$ we used to optimize in the original framework in Definition (2). For example, in logloss case, the worst case testing loss is obtained from the solved parametric form of $\hat{P}$ and $\check{P}$, which is the worst case predictor $P_\theta(\hat{Y}|\mathbf{X})$, to the Loss$(\hat{P}, \check{P})$: $\mathbb{E}_{P_{\text{test}(X,\hat{Y})}}[-\log P_\theta(\hat{Y}|\mathbf{X})]$.

Theorem 1. *Given m training samples with n dimensional features, if we assume that: the Lagrangian form of the robust covariate shift classifier (3) is strongly convex (with strong convexity constant M) in terms of θ, all density estimation is accurate, and the inner minimax game in (3) is solved exactly, then the expected loss on the testing distribution of the robust covariate shift classifier is bounded, with probability $1-\delta$:*

$$\mathbb{E}_{P_{test}(\mathbf{X})}[WCLoss(\hat{\theta})] \leq \mathbb{E}_{P_{test}(\mathbf{X})}[WCLoss(\theta^*)] + \frac{n \log \frac{2n}{\delta}}{4Mm}.$$

This bound indicates the distance between the expected target loss induced by our learned model from m training data and the optimal target loss decreases with rate $\mathcal{O}(\frac{1}{m})$. Note that the strong convexity condition is easy to satisfy even with non-smooth loss functions with L_2 regularization.

4 Experiments

We conduct experiments on real datasets and investigate the performance of robust 0-1 loss classifier in the framework. We chose four datasets from the UCI repository [17, 18] for the experiments. In order to create covariate shift, we synthetically generate 30 separate experiments in each dataset by drawing 100 training samples and 100 testing data samples from it, following an existing sampling procedure [9]. For each method, the regularization weights are chosen by 5-fold cross validation or importance weighted cross validation(IWCV). We use a discriminative classifier—logistic regression—as the density (ratio) estimator. We evaluate three methods:

- **Robust bias aware 0-1 classifier (Robust 0-1)** utilizes the general robust covariate shift classification framework (2) with Loss$(\check{P}_\mathbf{X}, \hat{P}_\mathbf{X}) = \hat{P}^T C \check{P}$ with C as the 0-1 loss matrix.
- **Adversarial 0-1 classifier (Adv 0-1)** minimizes expected 0-1 loss on the training distribution and has an optimization objective of: $\min_\theta \min_{\hat{P}} \max_{\check{P}} \mathbb{E}_{\mathbf{X} \sim P_{\text{train}}}[\hat{P}^T C \check{P} + \theta \cdot \phi(X, \check{Y})]$ $-\theta \cdot \tilde{\phi} + \epsilon||\theta||_2$, where $\tilde{\phi} = \mathbb{E}_{(\mathbf{X},Y) \sim \tilde{P}_{\text{train}}}[\phi(\mathbf{X}, Y)]$ here.
- **Importance Weighted SVM (IW-SVM)** reweights the training data with $\frac{P_{\text{test}}(X)}{P_{\text{train}}(X)}$ and minimizes reweighted multiclass hinge loss [19] on training data.

We show the comparison of accuracy in Table 1 and highlight methods that are either the best under paired t-test or not statistically distinguishable with significance level 0.1 in bold. We can see that Robust 0-1 performs better than other methods except in Seed, where it is statistically no worse than others. And Robust 0-1 can improve from Adv 0-1 at most times. That means minimizing worst case test loss using the adversarial game formulation (2) under covariate shift is better than minimizing training loss and ignoring the bias using the same formulation.

Table 1: Average Accuracy Comparison for UCI datasets

Datasets	**Robust 0-1**	**Adv 0-1**	**IW-SVM**
Seed	**0.834**	**0.820**	**0.820**
Vertebral	**0.823**	0.805	0.748
Vehicle	**0.547**	0.535	0.497
Spam	**0.757**	0.711	0.724

5 Conclusion

Covariate shift classification is an important but difficult task for machine learning in non-stationary environments when testing sample labels are not available during training. The original robust bias-aware classifier only optimizes logloss. We developed a general robust covariate shift classification framework that is flexible enough to minimize various loss functions. We used UCI under bias to demonstrate the model performance. We also investigate the theoretical property of the framework and are able to bound the worse case testing loss of our model properly.

References

[1] Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the International Conference on Machine Learning*, pages 903–910. ACM, 2004.

[2] Wei Fan, Ian Davidson, Bianca Zadrozny, and Philip S. Yu. An improved categorization of classifier's sensitivity on sample selection bias. In *Proc. of the IEEE International Conference on Data Mining*, pages 605–608, 2005.

[3] John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. In *Advances in neural information processing systems*, pages 129–136, 2008.

[4] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems*, pages 442–450, 2010.

[5] Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *Proceedings of the International Conference on Algorithmic Learning Theory*, pages 38–53, 2008.

[6] Anqi Liu and Brian D. Ziebart. Robust classification under sample selection bias. In *Advances in Neural Information Processing Systems*, pages 37–45, 2014.

[7] Xiangli Chen, Mathew Monfort, Anqi Liu, and Brian D Ziebart. Robust covariate shift regression. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 1270–1279, 2016.

[8] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[9] Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Schölkopf. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, pages 601–608, 2006.

[10] Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul V. Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems*, pages 1433–1440, 2008.

[11] Flemming Topsøe. Information theoretical optimization techniques. *Kybernetika*, 15(1):8–27, 1979.

[12] Peter D. Grünwald and A. Phillip Dawid. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics*, 32:1367–1433, 2004.

[13] Anqi Liu, Lev Reyzin, and Brian D Ziebart. Shift-pessimistic active learning using robust bias-aware prediction. In *AAAI*, pages 2764–2770, 2015.

[14] Rizal Fathony, Anqi Liu, Kaiser Asif, and Brian Ziebart. Adversarial multiclass classification: A risk minimization perspective. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 559–567. Curran Associates, Inc., 2016.

[15] Farzan Farnia and David Tse. A minimax approach to supervised learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4240–4248. Curran Associates, Inc., 2016.

[16] Rizal Fathony, Mohammad Ali Bashiri, and Brian Ziebart. Adversarial surrogate losses for ordinal regression. In *Advances in Neural Information Processing Systems 30*. 2017.

[17] D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz. UCI repository of machine learning databases, 1998.

[18] J P Siebert. Vehicle recognition using rule based methods. Technical report, Mar 1987.

[19] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, 2(Dec):265–292, 2001.