Addressing Covariate Shift in Active Learning with Adversarial Prediction

Anqi Liu
Kaiser Asif
Wei Xing
Sima Behpour
Brian D. Ziebart
Lev Reyzin

University of Illinois at Chicago, 851 S. Morgan St., Chicago, IL 60607 USA

Abstract
Active learning approaches used in practice are generally optimistic about their certainty with respect to data shift between labeled and unlabeled data. They assume that unknown datapoint labels follow the inductive biases of the active learner. As a result, the most useful datapoint labels—ones that refute current inductive biases—are rarely solicited. We propose an adversarial approach to active learning that assumes the worst-case about the unknown conditional label distribution under covariate shift. This closely aligns model uncertainty and expected error with generalization error, enabling more useful label solicitation. We investigate the benefits of this approach on classification tasks.

1. Introduction
Active learning has the potential to significantly improve data-efficiency beyond what is possible with randomly provided labels (Angluin, 1988; Balcan et al., 2010). However, data produced from an active learner violates the independent and identically distributed (IID) data property broadly assumed by supervised machine learning techniques (Sugiyama & Kawanabe, 2012). Existing active learning techniques that wrap around supervised classification algorithms either ignore this covariate shift property or employ reweighting methods (Shimodaira, 2000) that still heavily extrapolate from small sets of labeled examples to unlabeled data samples. As a result, they are generally optimistic when dealing with uncertainty. Unfortunately, the combination of this optimistic extrapolation and the ability to influence future training data leads not only to inefficient learning, but also to extreme inaccuracies. This is shown in Figure 1a where the active learner solicits labels where it is uncertain rather than where it is incorrect.

This paper summarizes our recent work on adversarial classification (Asif et al., 2015), covariate shift (Liu & Ziebart, 2014), and active learning (Liu et al., 2015). By explicitly considering batch active learning (Lewis & Gale, 1994) as a special case of learning under covariate shift (Sugiyama & Kawanabe, 2012; Shimodaira, 2000), we develop a pessimistic approach (Topsøe, 1979; Grünwald & Dawid, 2004) to active learning that avoids inefficiencies created by the combination of optimism and non-representative label solicitation. Our approach leverages a recently developed model for learning from biased source sample data by assuming the worst-case about the unknown conditional label distribution (Liu & Ziebart, 2014). Under this approach, common label solicitation strategies guided by model uncertainty or expected loss tend to directly improve the model’s predictive performance. Figure 1b shows the key difference from previous methods: the limited, more...
2. Approach

2.1. Robust bias-aware prediction

We employ a recently-developed approach for robust prediction in settings with dataset shift (Liu & Ziebart, 2014). It provides robust bias-aware (RBA) predictions in the active learning setting for data distribution according to the full data distribution $P_D(x)$, given labeled data samples (denoted with empirical measure $P_L(x)$) treated as being drawn from a labeled data distribution $P_L(x)$.

**Formulation.** We consider a minimax game between estimator choosing $\hat{P}(y|x)$ and constrained adversary choosing evaluation distribution $\hat{P}(y|x)$:

$$
\min_{\hat{P}(y|x)} \max_{P_L(x)} \mathbb{E}_{P_D(x)P(y|x)} \left[ \text{loss}(\hat{P}(Y|X), \hat{Y}) \right],
$$

(1)

The set $\Xi$ constrains the adversary to (approximately) match a set of its statistics, $\mathbb{E}_{P_L(x)\hat{P}(y|x)}[f(X,Y)]$ with sample statistics $\mathbb{E}_{\hat{P}_L(x)\hat{P}(y|x)}[f(X,Y)]$ from the labeled data distribution.

**Logarithmic loss.** Under the logarithmic loss, $-\log \hat{P}(Y = \hat{y}|x)$, estimator and adversary predictions are equivalent ($\hat{P}(y|x) = P_{\theta}(y|x)$) and the dual problem selects model parameters, $\theta$, by maximizing a regularized full data distribution likelihood:

$$
\theta^* = \arg\max_{\theta} \mathbb{E}_{P_D(x,y)} \left[ \log \hat{P}_\theta(Y|X) \right] - \lambda \|\theta\|_1,
$$

(2)

with the conditional label distribution estimate’s form as:

$$
\hat{P}_\theta(y|x) = e^{\frac{P_L(x)}{P_D(x)}} \theta \cdot \sum_{y \in Y} e^{P_L(x)} \cdot \theta \cdot f(x,y').
$$

(3)

**Zero-one loss.** The expected zero-one loss, $\sum_{y} \hat{P}(y|x)I(y \neq \hat{y})$, can be adversarially minimized under this formulation (Asif et al., 2015). The resulting distribution does not have a parametric form and instead is obtained numerically as the solution to a zero-sum game:

$$
\min_{\theta} \mathbb{E}_{P_D(x,y)\hat{P}(y|x)} \left[ \max_{P_L(x)} \min_{\hat{P}(y|x)} \sum_{y \neq \hat{y}} \hat{P}(\hat{y}|X)\hat{P}(\hat{y}|X)I(y \neq \hat{y}) \right. \left. + \frac{P_L(X)}{P_D(X)} \cdot \left( f(X,Y) - f(X,\hat{Y}) \right) \right],
$$

(4)

where parameters $\theta$ are estimated according to the outer convex optimization problem.

Under each loss function, the density ratio, $\frac{P_L(x)}{P_D(x)}$, moderates the predictions to be less certain wherever the labeled data underrepresents the full data distribution and more certain wherever the labeled data overrepresents it. The logloss learner is much more averse to making strong predictions without supporting data due to the unboundedness of the logloss.

**Parameter estimation.** Training the RBA predictor appears difficult because the dual objective function maximizes the log likelihood (2) or minimizes the game value (4) of the full data, which is partially unlabeled. However, each objective function’s (sub)gradient is based on labeled data distribution statistics,

$$
\mathbb{E}_{P_L(x)\hat{P}(y|x)}[f(X,Y)] - \mathbb{E}_{P_L(x)\hat{P}(y|x)}[f(X,Y)] - \lambda \nabla_{\theta} \|\theta\|_1,
$$

which can be safely approximated using labeled data distribution samples from $\hat{P}_L(x)$.

2.2. Active learning using robust predictions

Conditional label distribution estimates guide label solicitation within an active learner as shown in Algorithm 1.

**Algorithm 1** Label solicitation for pool-based active learner with covariate shift correction

**Input:** unlabeled pool dataset $\mathcal{U}$, labeled dataset $\mathcal{L}$

**Output:** example $x_i \in \mathcal{U}$ to solicit label

- Estimate labeled data density $P_L(x)$
- Estimate full data density $P_D(x)$ ($D = \mathcal{U} \cup \mathcal{L}$)
- Estimate $\hat{P}(y|x)$ from dataset $\mathcal{L}$, $P_L(x)$, and $P_D(x)$.
- Compute value $v_i = \text{metric}(\hat{P}, x_i, D, \mathcal{U})$ for each $x_i \in \mathcal{U}$
- Return $x_{\text{argmax}_i v_i}$ (example label to solicit)

The metrics used to evaluate different unlabeled datapoints for many label solicitation strategies, including uncertainty sampling, are heuristic/greedy methods for minimizing model uncertainty/expected loss. Theorem 1 connects this...
uncertainty/expected loss of the RBA predictor to generalization loss.

**Theorem 1.** Assuming that the actual label distribution \( P(y|x) \) is within the set \( \Xi \), the adversarial loss of the RBA predictor upper bounds its generalization loss:

\[
\mathbb{E}_{P_D(x)P(y|x)} \left[ \text{loss}(P(\hat{Y}|X), \hat{Y}) \right] \geq \mathbb{E}_{P_D(x)P(y|x)} \left[ \text{loss}(P(\hat{Y}|X), \hat{Y}) \right].
\]

In contrast, under the importance reweighting approach, for any hypothesis \( h \) the generalization loss \( \hat{R}(h) \) can be bounded as a function of the empirical importance-weighted loss \( \hat{R}_w(h) \) as follows:

\[
\hat{R}(h) \leq \hat{R}_w(h) + O \left( \sqrt{\mathbb{E}_{P_{E(x)}(X)} w(x)^2} \right)^{3/8} \frac{p \log \left( \frac{m}{p} \right) + \log \left( \frac{1}{\delta} \right)}{m},
\]

where \( w(x) = P_L(x)/P_D(x) \), \( \mathbb{E}_{P_{E(x)}(X)} \left[ w(X)^2 \right] \) is finite, \( p \) upper bounds the pseudo-dimension, a notion of dimension for the hypothesis space, and \( 1 - \delta \) is the bound confidence (Cortes et al., 2010).

Popular label solicitation strategies (Settles, 2012) tend to choose labels with the goal of greedily or approximately minimizing the (importance-weighted) empirical loss \( \hat{R}_w(h) \). Often do they do so without appropriately bounding the density ratio, \( \mathbb{E}_{P_{E(x)}(X)} \left[ \left( P_D(X)/P_L(X) \right)^2 \right] \), which is needed for the empirical loss to generalize from the subset of labeled datapoints to the rest of the dataset. For reasonably large confidence values \( 1 - \delta \), this bound can be looser than agnostic predictions (uniform over labels with logloss of \( \log_2 |\mathcal{Y}| \)) in such cases.

Unfortunately, minimizing the prediction loss for a non-representative labeled data distribution provides no guarantees for the prediction loss on the broader data distribution (Sugiyama & Kawanabe, 2012). Thus, active learners that minimize the uncertainty of the logistic regression model should instead solicit labels from representative datapoints to provide any theoretical performance guarantees.

**3. Experiments**

**3.1. Classification tasks**

We evaluate the performance of different active learning approaches using four datasets from the UCI repository (Bache & Lichman, 2013). We reduce multi-class datasets to binary classification tasks by using the Setosa, Type “1”, Class “0”, and Cytoplasm classes as positive examples and the remaining classes as negative in the respective datasets. In each of our experiments, we divide the dataset into a training set (80% of data) and a testing set (the remaining 20%). We refer the reader to the full version of this work (Liu et al., 2015) for complete details on the features, density estimators, and regularization choices employed.

**3.2. Learning methods**

We apply four models for estimating the conditional label distribution: (1) **Logistic regression** (abbreviated as standard in this section) (Schein & Ungar, 2007); (2) **Sample reweighted logistic regression** (reweighted) (Shimodaira, 2000; Zadrozny, 2004) minimizes the importance weighted estimate of the target logarithmic loss, \( \min_{\theta} \mathbb{E}_{P(x)} P(y|x) \left[ \frac{P_D(x)}{P_L(x)} \log \frac{P_D(y|x)}{P_L(y|x)} \right] + \lambda ||\theta||_1 \); (3) **Robust bias-aware log loss prediction (robust 0-1)** using Eqs. (2) and (3); and (4) **Robust bias-aware zero-one loss prediction (robust 0-1)** using Eq. (4).

We employ two label solicitation strategies for each model: **Uncertainty sampling** (abbreviated as active) selects the example with the largest value-conditioned entropy from the unlabeled dataset. The first datapoint label solicited is selected uniformly at random (the same first datapoint as passive learners); and **Random sampling** (abbreviated as passive) selects each datapoint uniformly at random from the unlabeled dataset. In addition, we apply a density-ratio-based strategy with our robust approach: **Density-ratio sampling** (abbreviated as active density) selects the example with the highest \( P_D(x)/P_L(x) \) under the estimated distribution.

We conduct 30 experiments with each learner on randomized training/testing splits of each dataset and report the mean and the 95% confidence interval of the predictive performance after every data point solicited in the first 20 steps, corresponding to 0.05 significance level in student t-test. We focus on the first 20 examples because real applications require good predictive performance with limited labeled data.

**3.3. Empirical comparisons**

We show that the optimistic active learning methods (active standard and active reweighted) often perform worse than IID logistic regression (passive standard) in Figures 3a and 3b, and in the first 10 steps of Figure 3c. This frequent poor performance results from the active learners getting “stuck” soliciting labels suggested by its optimistic biases to be useful rather than labels that would correct its incorrect beliefs. Further prediction improvements often require first exhausting from the pool of examples that conform to the learner’s incorrect beliefs. Only when the inductive biases of labeled data match those of the unlabeled data, as in active methods for the E. coli dataset, will the optimistic active learner not provide high logloss in the initial steps of active learning.

In contrast, active robust and active density robust using log loss perform comparably to or better than any other methods for all amounts of available data, as shown in Figure 3. Small error bars reflect high stability compared to other methods. In contrast, IID learning methods are quite unsta-
Addressing Covariate Shift in Active Learning with Adversarial Prediction

Figure 3. Logloss of optimistic active learning versus passive (IID) learning (top) and shift-pessimistic active learning versus passive (IID) (bottom) for the first 20 datapoints of learning averaged over 30 randomized withheld evaluation dataset splits with 95% confidence intervals.

Figure 4. Classification error rate of all learning methods for the first 20 datapoints of learning averaged over 30 randomized withheld evaluation dataset splits. The legend is shared for all datasets. Active standard and active reweighted overlap in (a).

ble especially at the beginning due to the bias of a small, randomly chosen sample. Active density robust cannot significantly compete with passive robust because it only considers densities when soliciting labels. Passive robust outperforms passive standard and reweighted, which shows that robust bias-aware prediction effectively controls the extent to which the prediction should generalize. However, since the inductive biases from labeled data tend to generalize accurately using the passive standard and reweighted methods on the E.coli dataset, they exceed the passive robust method given 20 labeled examples.

Though all the algorithms evaluated thus far do not minimize classification error directly, the log loss upper bounds the non-convex classification error (0-1 loss). Thus, one might expect that efficiently reducing log loss in the active learning setting will lead to low classification error. We investigate this in Figure 4, comparing the classification error rate of all seven methods on each dataset. The active robust approach provides the highest prediction accuracy for almost all amounts of available labeled data. In contrast, the high log loss predictions of active standard and active reweighted in Iris and Seed translate to poor classification error rates. We include preliminary experiments with robust zero-one loss minimization that outperforms for some datasets and significantly underperforms for others. Additional investigation is needed to understand and improve upon these differences in performance.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. #1227495.
References


