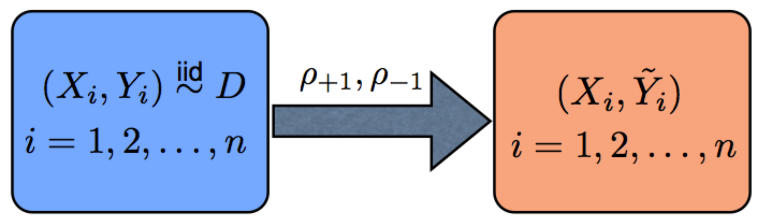
Learning with Noisy Labels

Nagarajan Natarajan [1], Inderjit S. Dhillon [1], Pradeep Ravikumar [1] and Ambuj Tewari [2]

[1]Department of Computer Science, The University of Texas at Austin [2]Dept. of Statistics & Dept. of EECS, University of Michigan, Ann Arbor

Introduction

- Goal: Guarantees for risk minimization, when the training labels (binary classification) are not reliable.
- The label noise is:
  1. ***random***: labels are flipped iid.
  2. ***class-conditional***: flip probability (noise rate) depends on the label (given by ρ_{+1} and ρ_{-1}).



- The class-conditional random noise model is given by:

$$P(\tilde{Y} = -1 | Y = +1) = \rho_{+1}, P(\tilde{Y} = +1 | Y = -1) = \rho_{-1}, \text{ and } \rho_{+1} + \rho_{-1} < 1$$

- ***No*** assumptions on the true distribution. ρ_{+1}, ρ_{-1} assumed to be known.
- The empirical risk minimization we consider is:

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \tilde{\ell}(f(X_i), \tilde{Y}_i). \qquad (1)$$

where $\tilde{Y}_i$ denote "noisy" labels and $\tilde{\ell}$ denotes a modified loss function.
- We develop two methods for obtaining $\tilde{\ell}$ given loss ℓ; our results generalize some existing results for the uniform noise model ($\rho_{+1} = \rho_{-1}$).

Empirical Risk Minimization (ERM)

- Let $f : \mathcal{X} \to \mathbb{R}$ be a real-valued decision function.
- The *risk* of f w.r.t. the 0-1 loss is given by $R_D(f) = \mathbb{E}_{(X,Y)\sim D}\left[1_{\{\text{sign}(f(X)) \neq Y\}}\right]$.
- Bayes optimal

$$f^*(x) = \operatorname*{argmin}_f R_D(f) = \text{sign}(\eta(x) - 1/2)$$

where $\eta(x) = P(Y = 1|x)$ and Bayes risk $R^* = R_D(f^*)$.
- Empirical $\tilde{\ell}$-risk on the observed sample: $\hat{R}_{\tilde{\ell}}(f) := \frac{1}{n}\sum_{i=1}^{n} \tilde{\ell}(f(X_i), \tilde{Y}_i)$.
- ℓ-risk under the "clean" distribution D: $R_{\ell,D}(f) := \mathbb{E}_{(X,Y)\sim D}[\ell(f(X), Y)]$.
- We want to bound $R_D(\hat{f}) - R^*$, i.e. the excess misclassification rate (under the clean distribution) of the ***empirical minimizer obtained using noisy data***.

Method I: Unbiased Estimators

- Let $\ell(t, y)$ be any bounded loss function. Then, if we define,

$$\tilde{\ell}(t, y) := \frac{(1 - \rho_{-y})\,\ell(t, y) - \rho_y\,\ell(t, -y)}{1 - \rho_{+1} - \rho_{-1}}$$

we have, for any t, y, $\quad \mathbb{E}_{\tilde{y}}\left[\tilde{\ell}(t, \tilde{y})\right] = \ell(t, y)$.
- We can learn a predictor with strong risk bounds (Main Result 1) using the constructed $\tilde{\ell}$ in ERM (1).
- The idea is well-known in online learning, where examples arrive sequentially (Bounds akin to Main Result 1 can be obtained using a stochastic descent procedure).
- Suppose $\ell(t, y)$ is ***convex and twice differentiable*** almost everywhere in t (for every y) and also satisfies the symmetry property $\ell''(t, y) = \ell''(t, -y), \forall t \in \mathbb{R}$, then $\tilde{\ell}(t, y)$ ***is also convex in*** t.
- Examples: the squared ℓ_{sq}, the logistic $\ell_{\log}$ and the Huber ℓ_{Hub} losses.
- Hinge loss does not satisfy the above property; but, with the class of bounded-norm hyperplanes $\{\mathbf{w} \in \mathcal{W}\}$, we can approximately minimize the non-convex $F(\mathbf{w}) = \hat{R}_{\tilde{\ell}}(\mathbf{w})$ by minimizing the biconjugate F^{**}.

Method II: Weighted losses

- For $\alpha > 0$, define ℓ_α as an α-weighted margin loss function of the form:

$$\ell_\alpha(t, y) = (1 - \alpha)1_{\{y=1\}}\ell(t) + \alpha 1_{\{y=-1\}}\ell(-t)$$

- Using ℓ_α in (1), we have the following ERM problem with noisy labels:

$$\hat{f}_\alpha = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell_\alpha(f(X_i), \tilde{Y}_i).$$

- For the hinge loss, letting $\mathcal{F}$ to be bounded-norm hyperplanes, we will get the ***biased SVM*** formulation used in Positive-Unlabeled learning:

$$\min_{\mathbf{w}, b} \ \frac{1}{2}\|\mathbf{w}\|^2 + C_+ \sum_{i:\tilde{y}_i = 1} \xi_i + C_- \sum_{i:\tilde{y}_i = -1} \xi_i$$

$$\text{s.t.} : \xi_i \geq 0, \text{ and } \tilde{y}_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \text{for } i = 1, 2, \ldots, n$$

- **Key observation**: The Bayes optimal under the noisy distribution, $\tilde{f}^* = \operatorname{argmin}_f E_{(X,\tilde{Y})\sim D_\rho}\left[1_{\{\text{sign}(f(X)) \neq \tilde{Y}\}}\right]$ is given by

$$\tilde{f}^*(x) = \text{sign}(\tilde{\eta}(x) - 1/2) = \text{sign}\left(\eta(x) - \frac{1/2 - \rho_{-1}}{1 - \rho_{+1} - \rho_{-1}}\right).$$

i.e. ***f^* and $\tilde{f}^*$ differ only in the threshold***.

Main Result: Method I

- $\hat{f}$: obtained using (1) with the unbiased estimate $\tilde{\ell}$ of loss ℓ.
- L_ρ: Lipschitz constant of $\tilde{\ell}$.

MAIN RESULT 1. With probability at least $1 - \delta$,

$$R_{\ell,D}(\hat{f}) \leq \min_{f \in \mathcal{F}} R_{\ell,D}(f) + 4L_\rho \mathfrak{R}(\mathcal{F}) + 2\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Furthermore, if ℓ is *classification-calibrated*, there exists a nondecreasing function ζ_ℓ with $\zeta_\ell(0) = 0$ such that,

$$R_D(\hat{f}) - R^* \leq \zeta_\ell\left(\min_{f \in \mathcal{F}} R_{\ell,D}(f) - \min_f R_{\ell,D}(f) + 4L_\rho \mathfrak{R}(\mathcal{F}) + 2\sqrt{\frac{\log(1/\delta)}{2n}}\right).$$

where the RHS consists of approximation and estimation errors.

Main Results: Method II

- Consider the risk of f w.r.t. the α-weighted 0-1 loss under noisy distribution D_ρ:

$$R_{\alpha,D_\rho}(f) = E_{(X,\tilde{Y})\sim D_\rho}\left[(1 - \alpha)1_{\{\tilde{Y}=1\}}1_{\{f(X)\leq 0\}} + \alpha 1_{\{\tilde{Y}=-1\}}1_{\{f(X)>0\}}\right].$$

MAIN RESULT 2. For the choices,

$$\alpha^* = \frac{1 - \rho_{+1} + \rho_{-1}}{2}, \text{ and } A_\rho = \frac{1 - \rho_{+1} - \rho_{-1}}{2},$$

there exists a constant B_X that is independent of f such that, for all functions f,

$$R_{\alpha^*,D_\rho}(f) = A_\rho R_D(f) + B_X.$$

Therefore:

$$\operatorname*{argmin}_f R_{\alpha^*,D_\rho}(f) = \operatorname*{argmin}_f R_D(f) = \text{sign}(\eta(x) - 1/2).$$

- Let ℓ be *classification-calibrated*, with Lipschitz constant L.
- $\hat{f}_{\alpha^*}$: obtained using (1) with weighted loss ℓ_{α^*} of loss ℓ.

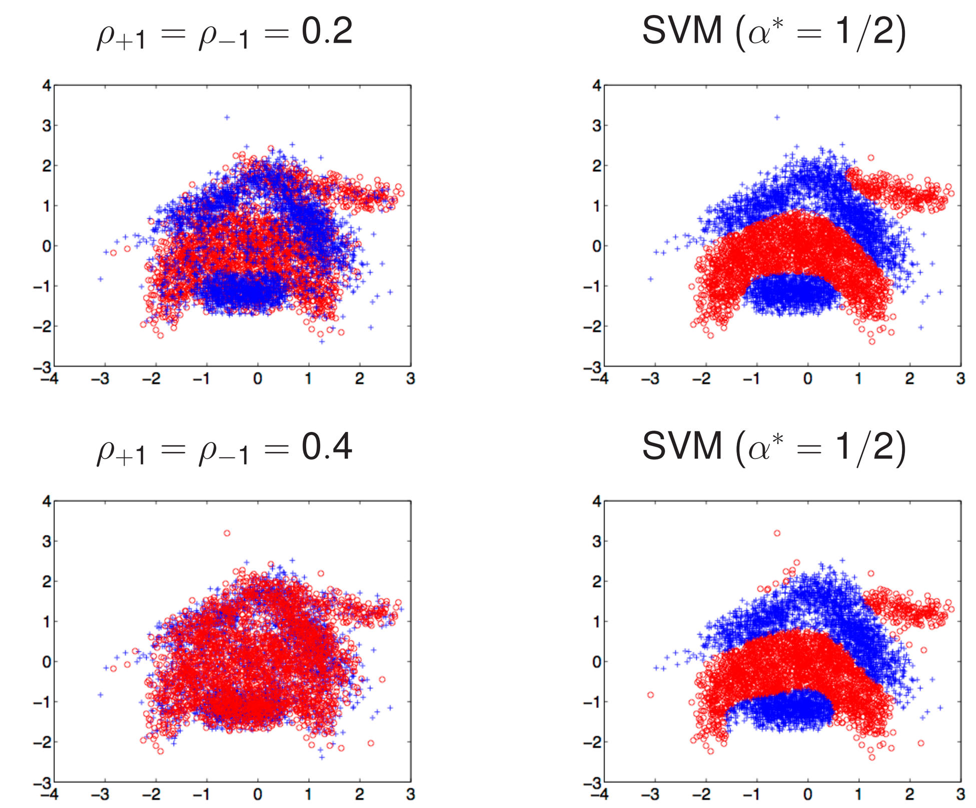
MAIN RESULT 3. For the choices α^* and A_ρ above, there exists a nondecreasing function $\zeta_{\ell_{\alpha^*}}$ with $\zeta_{\ell_{\alpha^*}}(0) = 0$, such that, with probability at least $1 - \delta$:

$$R_D(\hat{f}_{\alpha^*}) - R^* \leq A_\rho^{-1}\zeta_{\ell_{\alpha^*}}\left(\min_{f \in \mathcal{F}} R_{\alpha^*,D_\rho}(f) - \min_f R_{\alpha^*,D_\rho}(f) + 4L\mathfrak{R}(\mathcal{F}) + 2\sqrt{\frac{\log(1/\delta)}{2n}}\right).$$

- $\rho_{+1} = \rho_{-1} \implies \alpha^* = 1/2 \implies$ the original loss suffices.

Experimental Results

- Results on UCI 'banana' dataset (uniform label noise setting, regular SVM is robust):



- Existing methods for dealing with label noise (no theoretical guarantees):
  - PAM: Perceptron Algorithm with Margin (Khardon et al., JMLR, 2007)
  - NHERD (Crammer et al., NIPS 2010)
  - Random Projection (Stempfel et al., ALT 2007)
- We report the average accuracy of classification (%) over three-fold tests.
- Proposed methods (Unbiased estimator $\tilde{\ell}_{\log}$ and Biased SVM) are competitive with the state-of-the-art:

Dataset (d, n_+, n_-)	Noise Rates	$\tilde{\ell}_{\log}$	Biased SVM	PAM	NHERD	RP
Breast cancer (9, 77, 186)	$\rho_{+1} = \rho_{-1} = 0.2$	70.12	67.85	69.34	64.90	69.38
	$\rho_{+1} = 0.3, \rho_{-1} = 0.1$	70.07	67.81	67.79	65.68	66.28
	$\rho_{+1} = \rho_{-1} = 0.4$	67.79	67.79	67.05	56.50	54.19
Diabetes (8, 268, 500)	$\rho_{+1} = \rho_{-1} = 0.2$	76.04	66.41	69.53	73.18	75.00
	$\rho_{+1} = 0.3, \rho_{-1} = 0.1$	75.52	66.41	65.89	74.74	67.71
	$\rho_{+1} = \rho_{-1} = 0.4$	65.89	65.89	65.36	71.09	62.76
Thyroid (5, 65, 150)	$\rho_{+1} = \rho_{-1} = 0.2$	87.80	94.31	96.22	78.49	84.02
	$\rho_{+1} = 0.3, \rho_{-1} = 0.1$	80.34	92.46	86.85	87.78	83.12
	$\rho_{+1} = \rho_{-1} = 0.4$	83.10	66.32	70.98	85.95	57.96
German (20, 300, 700)	$\rho_{+1} = \rho_{-1} = 0.2$	71.80	68.40	63.80	67.80	62.80
	$\rho_{+1} = 0.3, \rho_{-1} = 0.1$	71.40	68.40	67.80	67.80	67.40
	$\rho_{+1} = \rho_{-1} = 0.4$	67.19	68.40	67.80	54.80	59.79
Heart (13, 120, 150)	$\rho_{+1} = \rho_{-1} = 0.2$	82.96	61.48	69.63	82.96	72.84
	$\rho_{+1} = 0.3, \rho_{-1} = 0.1$	84.44	57.04	62.22	81.48	79.26
	$\rho_{+1} = \rho_{-1} = 0.4$	57.04	54.81	53.33	52.59	68.15
Image (18, 1188, 898)	$\rho_{+1} = \rho_{-1} = 0.2$	82.45	91.95	92.90	77.76	65.29
	$\rho_{+1} = 0.3, \rho_{-1} = 0.1$	82.55	89.26	89.55	79.39	70.66
	$\rho_{+1} = \rho_{-1} = 0.4$	63.47	63.47	73.15	69.61	64.72

Open Problems

- Constant-Partition Classification Noise (CPCN) model: Richer model, with noise rates specific to discrete regions in $\mathcal{X} \times \{\pm\}$. It is open if theoretical guarantees for ERM are possible under CPCN model, even with assumptions on noise rates or distribution.
- Computing the biconjugate of sum of non-convex losses, with bounded-norm constraint, is easy only in limiting cases (norm constraint presents a problem); It is open if the biconjugate of $F(\mathbf{w}) = \hat{R}_{\tilde{\ell}}(\mathbf{w})$ admits a closed form or can be efficiently minimized.

Mail: {naga86,inderjit,pradeepr}@cs.utexas.edu, tewaria@umich.edu