Bayes Classifier cannot be learned from Noisy Responses With Unknown Noise Rate

Joint Work with Subha Maity

Soham Bakshi

PhD Student,
University of Michigan

August 6, 2023

Background & Introduction

- Let $\mathcal{X}$ and $\mathcal{Y}$ be the feature and label spaces respectively
- Clean samples (X_i, Y_i) are drawn independently from $P_{X,Y} \in \Delta(\mathcal{X} \times \mathcal{Y})$ ($\Delta(\mathcal{A})$ is the space of probability measures on $\mathcal{A}$), but the learner only observes the (X_i, Y_i')'s, where Y_i' is corrupted version of Y_i from a conditional distribution: $Y' \mid Y \sim P_{Y'|Y}$
- Learner seeks to estimate the Bayes classifier of Y (the true label)

  $$f_P^\star(x) \triangleq \operatorname{arg\,max}_{y \in \mathcal{Y}} P(Y = y \mid X = x)$$

  from the noisy training data $\{(X_i, Y_i')\}_{i=1}^n$
- Classification with noisy labels problem arises in many areas of science and engineering, including medical image analysis (Karimi et al., 2020) and crowdsourcing (Jiang et al., 2021)
- If the label noise rates/distribution $P_{Y'|Y}$ is known or learnable from external data, we can recover $f_P^\star$ (Bylander, 1994; Cesa-Bianchi et al., 1999; Natarajan et al., 2013)
- Recently, Liu and Guo (2020) propose a method based on *peer prediction*, which provably recovers $f_P^\star$ when there are only two classes and they are balanced (but the label noise distribution is unknown)

Background & Introduction

In our work, we consider the statistical aspects of classification with noisy labels. Our main contributions are the following

- We show that the balanced binary classification problem is the only instance in which $f_P^\star$ can be learned without knowledge of the label noise distribution, while in more general problems, that knowledge is essential
- We develop a new method based on weighted empirical risk minimization (ERM) that learns $f_P^\star$ in the balanced binary classification problem with noisy labels

Identifiability of Bayes Classifier

Setup & Assumptions

- A typical data-point (X, Y, Y') (a triplet of feature, clean label and noisy label) comes from a true distribution $P \equiv P_{X,Y,Y'}$ (unknown)
- Since the learner only observes (X_i, Y_i') pairs we assume that the $P_{X,Y'}$ marginal is known
- Assume that the noise is *instance independent*, i.e, for any $x \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$

$$P(Y' = y' \mid Y = y, X = x) = P(Y' = y' \mid Y = y) \triangleq \epsilon_P(y', y) \tag{1}$$

- Assume that $P_Y = p$, for some $p \in \Delta(\mathcal{Y})$
- Assume the determinant of $E_Q = [[\epsilon_Q(y', y)]]_{y', y \in \mathcal{Y}}$ is positive, which is a rather weak regularity assumption. For binary, the it boils down to $\epsilon_Q(0, 1) + \epsilon_Q(1, 0) < 1$, which is standard in the literature (Liu & Guo, 2020; Natarajan et al., 2013)

Identifiability of Bayes Classifier

Main Result

We define the class $\mathcal{Q}(K, p)$ of all probabilities $Q \equiv Q_{X,Y,Y'} \in \Delta(\mathcal{X} \times \mathcal{Y} \times \mathcal{Y})$ that satisfy our assumptions.

We want to investigate whether the Bayes classifier $f^{\star}_{Q}(x) \triangleq \arg\min_{y \in \mathcal{Y}} Q(Y = y \mid X = x)$ is same for all the $Q \in \mathcal{Q}(K, p)$.

Theorem 1 (Identifiability of the Bayes Classifier)

The Bayes classifier $f^{\star}_{Q}$ is unique for all $Q \in \mathcal{Q}(K, p)$, i.e , $\{f^{\star}_{Q} : Q \in \mathcal{Q}(K, p)\}$ is a singleton set if and only if $K = 2$ and $p = (\frac{1}{2}, \frac{1}{2})$.

Proof Idea/Sketch

Say, $\alpha_k(x) = Q_{(X,Y)}(x,k)$ and $a_k(x) = Q_{(X,Y')}(x,k)$. Then

$$\begin{bmatrix} a_1(x) \\ a_2(x) \\ \vdots \\ a_K(x) \end{bmatrix} = E \begin{bmatrix} \alpha_1(x) \\ \alpha_2(x) \\ \vdots \\ \alpha_K(x) \end{bmatrix}, \text{ where } E = \begin{bmatrix} \epsilon(1,1) & \epsilon(1,2) & \cdots & \epsilon(1,K) \\ \epsilon(2,1) & \epsilon(2,2) & \cdots & \epsilon(2,K) \\ \vdots & \vdots & \ldots & \vdots \\ \epsilon(K,1) & \epsilon(K,2) & \cdots & \epsilon(K,K) \end{bmatrix}$$

$$\implies [\alpha_1(x), \alpha_2(x), \ldots, \alpha_K(x)]^\top = E^{-1}[a_1(x), a_2(x), \ldots, a_K(x)]^\top.$$

Now, the vector $[a_1(x), a_2(x), \ldots, a_K(x)]^\top$ is known to us through the distribution of $P_{X,Y'}$. Additionally, we know p' (distribution of Y') and for all the distributions $Q \in \mathcal{Q}(K,p)$ the distribution of Y is p.

We construct two E_1 and E_2 that are (1) stochastic (non-negative entries with column sum one) (2) positive determinant (3) satisfies $p' = E_1 p$ and $p' = E_2 p$ and (4) produces Bayes decision boundaries.

The Identifiable Case

Weighted ERM

$p = (1/2, 1/2)$ is the special case when the *Bayes classifier is unique*, when the noise rates are not required for learning $f_P^\star$. We provide an alternative to the peer loss framework (Liu & Guo, 2020), that $f_P^\star$ can be learned from an weighted ERM:

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} p_n \left(1 - Y_i'\right) \ell \left(f\left(X_i\right), Y_i'\right),$$

where $\mathcal{F}$ is a set of probabilistic classifiers such that for some $\eta^* \in \mathcal{F}$ we have $f_P^*(x) = \mathbb{1}[\eta^*(x) \geq \frac{1}{2}]$, ℓ is the loss and $p_n(y) \triangleq 1/n \sum_{i=1}^{n} \mathbb{1}\left\{Y_i' = y\right\}$.

Lemma 1 (Weighted ERM)

Let $Y, Y' \in \{0, 1\}$, $P(Y = 1) = 1/2$, $\mathcal{F}$ be the set of all binary classifiers on $\mathcal{X}$, $\ell(f(x), y) = \mathbb{1}\{f(x) \neq y\}$ for $f \in \mathcal{F}$ and $\epsilon_P(0, 1) + \epsilon_P(1, 0) < 1$. If $p'(1) = P\left(Y' = 1\right)$ and $p'(0) = P\left(Y' = 0\right)$ then regardless the values of $\epsilon_P(0, 1)$ and $\epsilon_P(1, 0)$ the Bayes classifier is recovered, i.e.

$$f_P^\star(x) = \arg\min_{f \in \mathcal{F}} \mathbb{E}_P \left[p' \left(1 - Y'\right) \ell\left(f(X), Y'\right)\right].$$

The Non-Identifiable Cases

- For imbalanced binary classification (or) for classification tasks with more than two classes *the Bayes classifier is not identifiable* when the noise rates $\left(P_{Y'|Y}\right)$ are unknown
- For establishing the proof of Theorem 1 we construct two different $P_{Y'|Y}$'s that are compatible with the marginals $P_{X,Y'}$ and P_Y but have different Bayes decision boundaries
- This is the problematic case, where it is statistically impossible to learn the Bayes classifier owing to lack of identifiability, and an additional knowledge on $P_{Y'|Y}$ is essential for developing meaningful procedures

Synthetic Experiment

SetUp

We empirically investigate the classification approach in on a synthetic dataset for binary classification, whose description follows:

- $Y \sim \text{Bernoulli}(p)$
- $X \mid (Y = y) \sim \mathcal{N}_2 \left(\frac{2y-1}{2} \mathbf{1}_2, I_2 \right)$
- $P(Y' = 0 \mid Y = 1) = \epsilon_1, P(Y' = 1 \mid Y = 0) = \epsilon_0$
- We consider two situations: (1) the balanced case ($p = 1/2$), when the Bayes classifier is identified, and (2) an imbalanced case with $p = 0.35$, when the Bayes classifier is not identified
- For both the cases we set $\epsilon_1 = \{0.4 - (1 - \epsilon_0)(1 - p)\} / p$
- Note that, for such a choice $P(Y' = 0) = P(Y' = 0 \mid Y = 1) p + P(Y' = 0 \mid Y = 0)(1 - p) = \epsilon_1 p + (1 - \epsilon_0)(1 - p) = 0.4$
- We compare our classification approach with two baselines: (1) the oracle (trained with the true Y) and (2) the baseline (trained with noisy Y' without any adaptation). We use logistic regression model for all the classifiers.

Synthetic Experiment

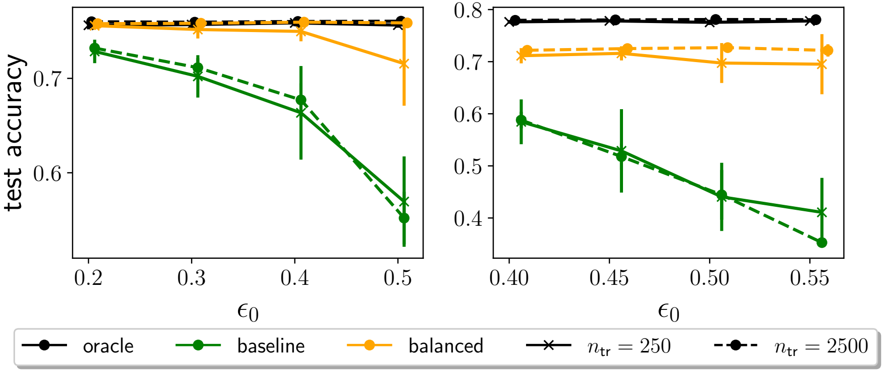

Figure. The consistency (resp. inconsistency) of class balancing approach in for balanced (resp. imbalanced) binary classification, as observed in left (resp. right) plot.

In the balanced case, our ERM approach has identical performance to the oracle case for large sample sizes ($n_{\mathrm{tr}} = 2500$). On contrary, for imbalanced case there is a gap between the class balancing approach and the oracle, even for large sample sizes.

References I

Bylander, T. (1994). Learning linear threshold functions in the presence of classification noise. *Proceedings of the seventh annual conference on Computational learning theory*, 340–347.

Cesa-Bianchi, N., Dichterman, E., Fischer, P., Shamir, E., & Simon, H. U. (1999). Sample-efficient strategies for learning in the presence of noise. *Journal of the ACM (JACM)*, *46*(5), 684–719.

Jiang, L., Zhang, H., Tao, F., & Li, C. (2021). Learning from crowds with multiple noisy label distribution propagation. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(11), 6558–6568.

Karimi, D., Dou, H., Warfield, S. K., & Gholipour, A. (2020). Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical image analysis*, *65*, 101759.

Liu, Y., & Guo, H. (2020). Peer loss functions: Learning from noisy labels without knowing noise rates. *International conference on machine learning*, 6226–6236.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., & Tewari, A. (2013). Learning with noisy labels. *Advances in neural information processing systems*, *26*.