Attribute Noise Robust Binary Classification (Student Abstract)

Aditya Petety,[1] **Sandhya Tripathi,**[2] **N. Hemachandra**[2]
[1]National Institute of Science Education and Research Bhubaneshwar
[1]aditya.petety@niser.ac.in
[2]Indian Institute of Technology Bombay
[2]{sandhya.tripathi, nh}@iitb.ac.in

Abstract

We consider the problem of learning linear classifiers when both features and labels are binary. In addition, the features are noisy, i.e., they could be flipped with an unknown probability. In Sy-De attribute noise model, where all features could be noisy together with same probability, we show that 0-1 loss (l_{0-1}) need not be robust but a popular surrogate, squared loss (l_{sq}) is. In Asy-In attribute noise model, we prove that l_{0-1} is robust for any distribution over 2 dimensional feature space. However, due to computational intractability of l_{0-1}, we resort to l_{sq} and observe that it need not be Asy-In noise robust. Our empirical results support Sy-De robustness of squared loss for low to moderate noise rates.

Introduction

Quality of data is being compromised as its quantity is getting larger. In classification setup, bad quality data could be due to noise in the labels or noise in the features. Label noise research has gained a lot of attention in last decade (Sastry and Manwani 2016). In contrast, feature or attribute noise is still unexplored. As opposed to continuous valued attributes, noise in categorical features, particularly binary, can drastically change the relative location of a data point and significantly impact the classifier's performance.

(Quinlan 1986) studied the effect of noise when the algorithms are decision trees. (Zhu and Wu 2004; Khoshgoftaar and Van Hulse 2009) study attribute noise from the perspective of detecting noisy data points and correcting them.

Our major contributions lie in identifying loss functions that are robust (or not) to attribute (binary valued) noise in Empirical Risk Minimization (ERM) framework. This has an advantage that there is no need of either knowing the true value or cross-validating over or estimating the noise rates.

Problem Description

Let D be the joint distribution over $\mathbf{X} \times Y$, where $\mathbf{X} \in \mathcal{X} \subseteq \{-1, 1\}^n$ and $Y \in \mathcal{Y} = \{-1, 1\}$. Let the decision function $f : \mathbf{X} \mapsto \mathbb{R}$ be an element of the class of all measurable functions $\mathcal{H}$. We restrict our set of hypothesis to be in linear hypothesis class $\mathcal{H}_{lin} := \{(\boldsymbol{\beta}, c), \boldsymbol{\beta} \in \mathbb{R}^n, c \in \mathbb{R}\}$. Let $\tilde{D}$ denote the distribution on $\tilde{\mathbf{X}} \times Y$ obtained by inducing noise to D with $\tilde{\mathbf{X}} \in \mathcal{X} \subseteq \{-1, 1\}^n$. The corrupted sample is $\tilde{S} := \{(\tilde{\mathbf{x}}_1, y_1), \ldots, (\tilde{\mathbf{x}}_m, y_m)\} \sim \tilde{D}^m$. The probability that the value of i^{th} attribute is flipped is given by $p_i = (\tilde{\mathbf{X}} = -x | \mathbf{X} = x, Y = y), i \in [n]$ where $[n] = \{1, \ldots, n\}$. We assume that the class/label is not affected by attribute noise.

Based on the flipping probability and the dependence between events of flipping for different attributes, we *identify* two attribute noise models. If all the attribute values are flipped together with same probability p, then it is referred to as the **symmetric dependent attribute noise model (Sy-De)**. If each attribute j flips with probability p_j independently of any other attribute $k \in [n] \backslash \{j\}$, then it is referred to as the **asymmetric independent attribute noise model (Asy-In)**. Even though Sy-De attribute noise model is simple, it cannot be obtained by taking $p_i = p_j, \ \forall i, j \in [n]$ in Asy-In attribute noise model. Real world example of Sy-De (or Asy-In) noisy attributes: Consider a room with many sensors connected in series (or with individual battery) measuring temperature, humidity, etc., as binary value, i.e., high or low. A power failure (or battery failures) will lead to all (or individual) sensors/attributes providing noisy observations with same (or different) probability.

We consider ERM framework for classification. A natural choice for loss function is 0-1 loss, i.e., $l_{0-1}(f(\mathbf{x}, y)) = \mathbf{1}_{[f(\mathbf{x})y<0]}$. Bayes classifier $f^* = \arg\min_{f \in \mathcal{H}} R_D(f)$ and Bayes risk is $R_D(f^*) = \min_{f \in \mathcal{H}} R_D(f)$ where $R_D(f) = E_D[\mathbf{1}_{[f(\mathbf{x})y<0]}]$. Corresponding quantities for noisy distribution $\tilde{D}$ are $R_{\tilde{D}}(\tilde{f}^*) = \min_{\tilde{f} \in \mathcal{H}} E_{\tilde{D}}[\mathbf{1}_{[\tilde{f}(\tilde{\mathbf{x}})y<0]}]$ and $\tilde{f}^* = \arg\min_{\tilde{f} \in \mathcal{H}} R_{\tilde{D}}(\tilde{f})$.

Non-convex nature of 0-1 loss makes it difficult to optimize and hence convex upper bounds (surrogate losses) are used in practice. In this work, we consider the squared loss $l_{sq}(f(\mathbf{x}, y)) = (y - f(\mathbf{x}))^2$, a differentiable and convex surrogate loss function. Our restriction of hypothesis to linear class $\mathcal{H}_{lin}$ can be interpreted as a form of regularization. Expected squared clean and corrupted risks are $R_{D,l_{sq}}(f) = E_D[(y - f(\mathbf{x}))^2]$ and $R_{\tilde{D},l_{sq}}(\tilde{f}) = E_{\tilde{D}}[(y - \tilde{f}(\tilde{\mathbf{x}}))^2]$. Hypothesis in $\mathcal{H}_{lin}$ minimizing these clean and corrupted risks

are denoted by $f^*_{sq,lin}$ and $\tilde{f}^*_{sq,lin}$. Next, we define attribute noise robustness of risk minimization scheme A_l.

Definition 1. *Let $f^*_{A_l}$ and $\tilde{f}^*_{A_l}$ be obtained from clean and corrupted distribution D and $\tilde{D}$ using any arbitrary scheme A_l. Then, scheme A_l is said to be attribute noise robust if*

$$R_D(f^*_{A_l}) = R_D(\tilde{f}^*_{A_l}).$$

Also, l is said to be an attribute noise robust loss function.

Attribute Noise Robust Loss Functions

We, first, consider Sy-De attribute noise model and present a counter example (Example 1) to show that 0-1 loss need not be robust to Sy-De attribute noise. To circumvent this problem, we provide a positive result by showing that squared loss is Sy-De attribute noise robust with origin passing linear classifiers (Theorem 1). Details of examples and proofs are available in Supplementary Material (SM)[1].

Example 1. *Consider a population of two data points (in 1-D) (x, y) as $(-1, 1)$ and $(1, -1)$ with probability 0.25 and 0.75 with a classifier $f_{lin}(x) = bx + c$. Then, the l_{0-1} optimal clean classifier is $f^*_{lin} = (b^*, c^*) = (-1, -0.1)$ with $R_D(f^*_{lin}) = 0$. Also, the l_{0-1} optimal Sy-De attribute noise ($p = 0.4$) corrupted classifier is $\tilde{f}^*_{lin} = (\tilde{b}^*, \tilde{c}^*) = (1, -2)$ with $R_D(\tilde{f}^*_{lin}) = 0.25$. Since, $R_D(\tilde{f}^*_{lin}) \neq R_D(f^*_{lin})$, 0-1 loss function need not be Sy-De attribute noise robust.*

Theorem 1. *Consider a clean distribution D on $\mathbf{X} \times Y$ and Sy-De attribute noise corrupted distribution $\tilde{D}$ on $\tilde{\mathbf{X}} \times Y$ with noise rate $p < 0.5$. Then, squared loss l_{sq} with origin passing linear classifiers is Sy-De attribute noise robust,i.e.,*

$$R_D(\tilde{f}^*_{lin,l_{sq}}) = R_D(f^*_{lin,l_{sq}}) \tag{1}$$

where $f^*_{lin,l_{sq}} = (\beta^*_1, \ldots, \beta^*_n)$ and $\tilde{f}^*_{lin,l_{sq}} = (\tilde{\beta}^*_1, \ldots, \tilde{\beta}_n*)$ correspond to optimal linear classifiers learnt using squared loss on clean (D) and corrupted ($\tilde{D}$) distribution.

Remark 1. *Sy-De robustness of squared loss is an interesting result because given an attribute noise corrupted dataset, obtaining a linear classifier entails solving only a linear system of equations. (Demonstrated on UCI datasets.)*

Now, we consider Asy-In attribute noise model and show that 0-1 loss is robust to this noise with non-origin passing classifiers when $n = 2$ (Theorem 2). As l_{0-1} based ERM is computationally intractable, we consider l_{sq} and present a counter example to show that l_{sq} need not be Asy-In noise robust (Example 2).

Theorem 2. *Consider a clean distribution D with probabilities $\{d_1, d_2, d_3, d_4\}$ on $\mathbf{X} \times Y$ with $n = 2$ (population of 2^n data points) and Asy-In attribute noise corrupted distribution $\tilde{D}$ on $\tilde{\mathbf{X}} \times Y$ with noise rates $p_1 < 0.5$ and $p_2 < 0.5$. Then, 0-1 loss with non-origin passing linear classifiers is Asy-In attribute noise robust, i.e.,*

$$R_D(\tilde{f}^*_{lin,l_{0-1}}) = R_D(f^*_{lin,l_{0-1}}) \tag{2}$$

where $f^*_{lin,l_{0-1}} = (\beta^*_1, 1, c^*)$ and $\tilde{f}^*_{lin,l_{0-1}} = (\tilde{\beta}^*_1, 1, \tilde{c}^*)$ correspond to optimal linear classifiers learnt using 0-1 loss on clean (D) and corrupted ($\tilde{D}$) distribution respectively.

[1]Available at https://arxiv.org/abs/1911.07875

Example 2. *Consider a population of 3 data points (in 2-D) (x_1, x_2, y) as $(-1, 1, 1)$, $(-1, -1, -1)$, and $(1, -1, 1)$ with probabilities as $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ with a classifier $f_{lin}(\mathbf{x}) = b_1 x_1 + b_2 x_2$. Then, l_{sq} optimal clean classifier is $f^*_{lin,l_{sq}} = (b^*_1, b^*_2) = (0.5, 0.5)$ with $R_D(f^*_{lin,l_{sq}}) = 0.5$. Also, l_{sq} optimal Asy-In attribute noise ($p_1 = 0.1$, $p_2 = 0.2$) corrupted classifier is $\tilde{f}^*_{lin,l_{sq}} = (\tilde{b}^*_1, \tilde{b}^*_2) = (0.4, 0.3)$ with $R_D(\tilde{f}^*_{lin,l_{sq}}) = 0.25$. Since, $R_D(f^*_{lin,l_{sq}}) \neq R_D(\tilde{f}^*_{lin,l_{sq}})$, squared loss need not be Asy-In attribute noise robust.*

Experiments

Figure 1 demonstrates Sy-De attribute noise robustness of squared loss on 3 UCI datasets (Dheeru and Karra Taniskidou 2017); details in SM. As SPECT dataset is imbalanced, in addition to accuracy, we also report arithmetic mean (AM). To account for randomness in noise, results are averaged over 15 trials of train-test partitioning (80-20). The low accuracy in comparison to clean classifier can be attributed to the finite samples available for learning the classifiers.

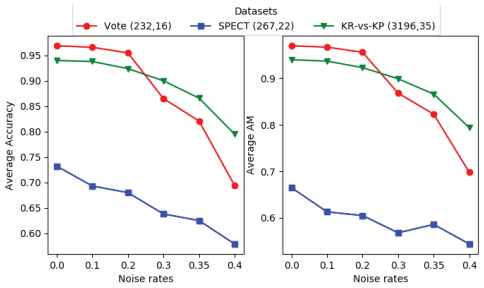

Figure 1: Test data performance of l_{sq} with Sy-De attribute noise.

Looking Forward

Our work is an initial attempt in binary valued attribute noise; an extension to general discrete valued attributes would be interesting. Asy-In attribute noise model raises some non-trivial questions w.r.t. choice of loss functions like robustness of 0-1 for $n > 2$, explanation for the surprising non-robustness of squared loss as compared to robustness of a difficult to deal 0-1 loss, search for other surrogate loss functions that are robust. Finally, we believe that attribute dimension n could have a role to play in noise robustness.

References

Dheeru, D., and Karra Taniskidou, E. 2017. UCI machine learning repository.

Khoshgoftaar, T. M., and Van Hulse, J. 2009. Empirical case studies in attribute noise detection. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 39(4):379–388.

Quinlan, J. R. 1986. The effect of noise on concept learning. *Machine learning: An artificial intelligence approach* 2:149–166.

Sastry, P. S., and Manwani, N. 2016. *Robust Learning of Classifiers in the Presence of Label Noise*. chapter 6, 167–197.

Zhu, X., and Wu, X. 2004. Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review* 22(3):177–210.