Fast Rates for a k-NN Classifier Robust to Unknown Asymmetric Label Noise

Henry W J Reeve and Ata Kabán

University of Birmingham, United Kingdom

International Conference on Machine Learning 2019

Pacific Ballroom #187

Learning with asymmetric label noise

Suppose we have a distribution $\mathbb{P}$ over $(X, Y) \in \mathcal{X} \times \{0, 1\}$

Our goal is to obtain a classifier $\phi : \mathcal{X} \to \{0, 1\}$ which minimizes

$$\mathcal{R}(\phi) = \mathbb{P}\left[\phi(X) \neq Y\right]$$

We would like uncorrupted data:

$$\mathcal{D} = \{(X_1, Y_1), \cdots, (X_1, Y_n)\} \overset{\text{i.i.d.}}{\sim} \mathbb{P}$$

Instead, we have corrupted data:

$$\mathcal{D}_{\text{corr}} = \{(X_1, \tilde{Y}_1), \cdots, (X_1, \tilde{Y}_n)\} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\text{corr}}$$

Learning with asymmetric label noise

There exist label noise probabilities $p_0, p_1 \in (0,1)$ with $p_0 + p_1 < 1$

$$(X, \tilde{Y}) \sim \mathbb{P}_{\text{corr}}$$

$$\Updownarrow$$

1. $$(X, Y) \sim \mathbb{P}$$

2. $$\tilde{Y} = \begin{cases} Y & \text{with probability } 1 - p_Y. \\ 1 - Y & \text{with probability } p_Y \end{cases}$$

Samples $(X_i, \tilde{Y}_i)$ consist of a feature vector X_i and a noisy label $\tilde{Y}_i$.

Applications

Asymmetric class-conditional label noise occurs in numerous applications:

- Nuclear particle classification - distinguishing neutrons from gamma rays

  (Blanchard et al., 2016)

- Protein classification and other problems with Positive and Unlabelled data

  (Elkan & Noto, 2009)

The Robust k-NN classifier of Gao et al. (2018)

Let $\hat{\eta}_{\mathrm{corr}}$ be the k-nearest neighbors regression estimator based on $\mathcal{D}_{\mathrm{corr}}$

1) Estimate the label noise probabilities p_0, p_1

$$\hat{p}_0 := \min_{i \in [n]} \{\hat{\eta}_{\mathrm{corr}}(X_i)\} \qquad \hat{p}_1 := 1 - \max_{i \in [n]} \{\hat{\eta}_{\mathrm{corr}}(X_i)\}$$

2) Binary k-nearest neighbor prediction with a label noise dependent threshold:

$$\hat{\phi}_{n,k}(x) := 1\left\{\hat{\eta}_{\mathrm{corr}}(x) \geq \frac{1}{2} \cdot (1 + \hat{p}_0 - \hat{p}_1)\right\}$$

The Robust k-NN classifier of Gao et al. (2018)

The Robust k-NN classifier was introduced by Gao et al. (2018) who:

1) Conducted a comprehensive empirical study which demonstrates that the method typically outperforms a range of competitors.

2) Proved finite sample bounds. However,

   a) Fast rates ($o(n^{-1/2})$) have not been established.

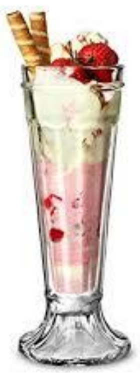
   b) The bounds assume prior knowledge of the label noise p_0, p_1.

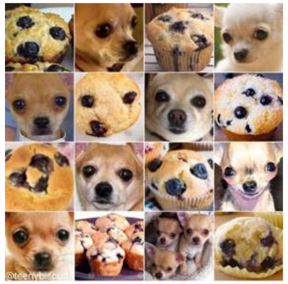
In our work the label noise probabilities are **unknown!**

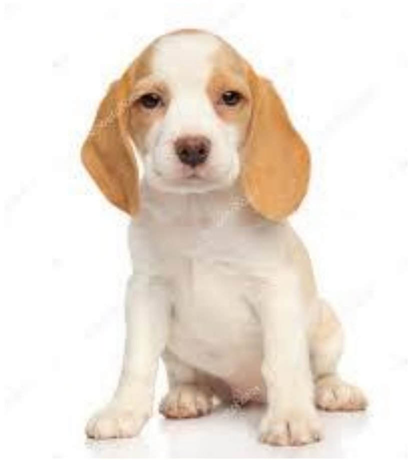
Range assumption

We adopt the range assumption of Menon et al. (2015):

$$\eta(x) := \mathbb{P}[Y = 1 | X = x]$$







$$\inf_{x \in \mathcal{X}} \{\eta(x)\} = 0$$

$$\sup_{x \in \mathcal{X}} \{\eta(x)\} = 1$$

Non-parametric assumptions

We also adopt the following non-parametric assumptions:

A) Measure-smoothness assumption :

$$|\eta(x_0) - \eta(x_1)| \leq \omega \cdot \mu \left(B_{\|x_0 - x_1\|}(x_0) \right)^{\lambda}$$

B) Tysbakov's margin assumption :

$$\mu\left(\left\{x \in \mathcal{X} : \left|\eta(x) - \frac{1}{2}\right| < \Delta\right\}\right) \leq C_\alpha \cdot \Delta^\alpha$$

Fast rates for the Robust k-NN classifier

Main result (Reeve & Kabán, 2019)

Suppose that $\mathbb{P}$ satisfies (1) the range assumption,

(2) the measure-smoothness assumption,

(3) Tsybakov's margin assumption.

With probability at least $1-\delta$ over the corrupted sample $\mathcal{D}_{\mathrm{corr}}$, the Robust k-Nearest Neighbor classifier satisfies

$$\mathcal{R}\left(\hat{\phi}_{n,k(n)}\right)-\mathcal{R}\left(\phi^{*}\right) \leq C\cdot\left(\frac{\log(n/\delta)}{n}\right)^{\frac{\lambda(1+\alpha)}{2\lambda+1}}+\delta.$$

Matches the minimax optimal rate for the noise free setting (up to log factors)!

Conclusions

Pacific Ballroom #187

- We established fast rates for the Robust k-NN classifier of Gao et al. (2016)
- A high probability bound is established for **unknown** asymmetric label noise
- The finite sample rates match the minimax optimal rates for the label-noise free setting up to logarithmic factors (e.g. Audibert & Tsybakov, 2006)
- As a biproduct of our analysis we provide a high probability bound for determining the maximum of a noisy function with minimal assumptions.

Thank you for listening!