Binary Classification from Positive-Confidence Data

Takashi Ishida[1,2] **Gang Niu**[2] **Masashi Sugiyama**[2,1]

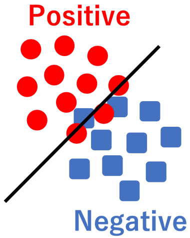
[1] The University of Tokyo
[2] RIKEN

NeurIPS 2018, Canada, December 6th, 2018

Introduction

- Ordinary classification:

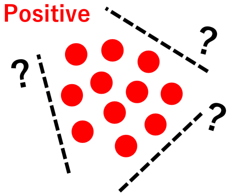
  Learn a binary classifier with *both* **positive** and **negative** training data.



- Research question:

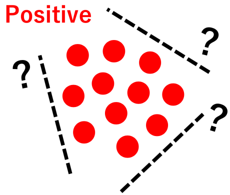
  Can we learn a binary classifier from *only* **positive** data?

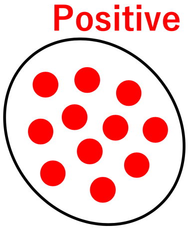
  Without any **negative** data, or even **unlabeled** data?



How About One-Class Methods?

- With *only* **positive** data: We do not know the direction of the negative distribution.
- One-class methods: **Describe** the positive class by clustering-related methods.
- Does not have the ability to tune hyper-parameters for maximizing the generalization ability.
- Aim is *not* on **discriminating** **positive** and **negative** classes!





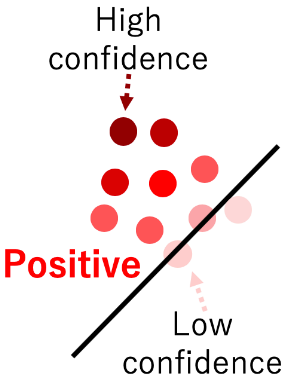
Main Idea

Equip positive data with confidence:

- Example: 95% DOG (5% WOLF)

Main message of the paper:

- If you can equip positive data with confidence (positive-confidence), you can learn a binary classifier with **optimal convergence rate**!
- Positive-confidence includes the information of the negative distribution → allows us to discriminate between positive/negative classes.
- *Positive-confidence (Pconf) classification*.



Notations/Settings for Binary Classification

- Input is $\boldsymbol{x} \in \mathbb{R}^d$ and its class label $y \in \{\pm 1\}$ follows unknown distribution with density $p(\boldsymbol{x}, y)$.
- Goal: Train a binary classifier $g(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}$ so that the classification risk $R(g)$ is minimized:

$$R(g) = \mathbb{E}_{p(\boldsymbol{x},y)}[\ell(yg(\boldsymbol{x}))]$$

  - $\mathbb{E}_{p(\boldsymbol{x},y)}$ denotes expectation over $p(\boldsymbol{x}, y)$.
  - $\ell(z)$ is a loss function that typically takes a large value for small z.

$\rightarrow$ **Empirical risk minimization** (ERM) approach

Notations/Settings for Pconf Classification

- Goal is still the same: minimize $R(g) = \mathbb{E}_{p(\boldsymbol{x},y)}[\ell(yg(\boldsymbol{x}))]$.
- Only have positive samples equipped with *confidence*:

$$\mathcal{X} := \{\boldsymbol{x}_i, r_i\}_{i=1}^n$$

  - $\boldsymbol{x}_i$ is positive data drawn independently from $p(\boldsymbol{x}|y=+1)$.
  - r_i is the positive-confidence given by $r_i = p(y=+1|\boldsymbol{x}_i)$.

Serious issue: We cannot directly employ the standard ERM approach!

Theorem

Classification risk can be expressed as

$$R(g) = p(y = +1) \cdot \mathbb{E}_{p(\boldsymbol{x}|y=+1)}\left[\ell(g(\boldsymbol{x})) + \frac{1 - r(\boldsymbol{x})}{r(\boldsymbol{x})}\ell(-g(\boldsymbol{x}))\right],$$

if we have $p(y = +1|\boldsymbol{x}) \neq 0$ *for all* $\boldsymbol{x}$ *sampled from* $p(\boldsymbol{x})$.

- $p(y = +1)$ can be regarded as a constant when $R(g)$ is minimized w.r.t. g and can be safely ignored.

Comparing Proposed and Naive Formulation

Proposed method:

$$\min_g \sum_{i=1}^{n} \left[\ell(g(\boldsymbol{x}_i)) + \frac{1 - r_i}{r_i} \ell(-g(\boldsymbol{x}_i)) \right]$$

Weighted method (naive):

$$\min_g \sum_{i=1}^{n} \left[r_i \ell(g(\boldsymbol{x}_i)) + (1 - r_i) \ell(-g(\boldsymbol{x}_i)) \right]$$

- Weighted version seems more natural and straightforward, but we show in the paper that it is not an unbiased estimator of the risk.

Conclusions

- Proposed a novel problem setting and algorithm for binary classification from positive-confidence data.
- Showed that an unbiased estimator of the classification risk can be obtained in a model- and optimization-independent way.

Come see our poster @ Room 210 & 230 AB #97 for more!

- **Theoretical work** on estimation error bounds
- **Experiments** on synthetic and benchmark datasets
- **Potential applications** for Pconf classification