Positive-Unlabeled Learning with Non-Negative Risk Estimator

Ryuichi Kiryo\textsuperscript{1,2}  Gang Niu\textsuperscript{1,2}  Martins Christoffel du Plessis  Masashi Sugiyama\textsuperscript{2,1}

\textsuperscript{1} 東京大学  \textsuperscript{2} RIKEN
Binary classification

- Classify input data $X \in \mathbb{R}^d$ to class $Y \in \{+1 \text{ (positive)}, -1 \text{ (negative)}\}$

- Supervised learning:
  Classifier $g: \mathbb{R}^d \rightarrow \mathbb{R}$ is learned from positive data and negative data

![Diagram of binary classification with positive and negative data points]
Supervised binary classification (PN learning)

• **Goal**: minimize expected risk

\[
R(g) = \mathbb{E}_{(X,Y) \sim p(x,y)}[l(g(X), Y)] = \pi_p \mathbb{E}_p[l(g(X), +1)] + \pi_n \mathbb{E}_n[l(g(X), -1)]
\]

risk for positive class  risk for negative class

• Minimize empirical risk: approximation by data in hand

\[
\hat{R}_{pn}(g) = \frac{\pi_p}{n_p} \sum_{x \in \mathcal{X}_p} l(g(x), +1) + \frac{\pi_n}{n_n} \sum_{x \in \mathcal{X}_n} l(g(x), -1)
\]

input data \( X \in \mathbb{R}^d \)

class label \( Y \in \{ \pm 1 \} \)

loss function \( l: \mathbb{R} \times \{ \pm 1 \} \rightarrow \mathbb{R} \)

\[
\begin{align*}
\mathbb{E}_p[\cdot] &:= \mathbb{E}_{X \sim p(x|Y=+1)}[\cdot] \\
\mathbb{E}_n[\cdot] &:= \mathbb{E}_{X \sim p(x|Y=-1)}[\cdot] \\
\pi_p &:= p(Y = +1) \\
\pi_n &:= p(Y = -1) \\
\mathcal{X}_p &= \{x^p_i\}_{i=1}^{n_p} \overset{i.i.d.}{\sim} p(x|Y = +1) \\
\mathcal{X}_n &= \{x^n_i\}_{i=1}^{n_n} \overset{i.i.d.}{\sim} p(x|Y = -1)
\end{align*}
\]
Classification when negative data is unavailable

- Example: click advertisement
  - Clicked: positive
  - Non-clicked: unlabeled (not interesting or unseen)

Learn a **PN** binary classifier from positive and unlabeled data (**PU** learning)
Unbiased PU learning [Natarajan+, NIPS 2013, du Plessis+, ICML 2015]

**Goal:** minimize the same expected risk as PN learning

\[
R(g) = \pi_p \mathbb{E}_p[l(g(X), +1)] + \pi_n \mathbb{E}_n[l(g(X), -1)]
\]

**Idea:** unlabeled data = positive data + negative data

\[
\mathbb{E}_u[l(g(X), -1)] = \pi_p \mathbb{E}_p[l(g(X), -1)] + \pi_n \mathbb{E}_n[l(g(X), -1)]
\]

**Risk can be expressed by only positive and unlabeled data**

\[
R_{pu}(g) = \pi_p \mathbb{E}_p[l(g(X), +1)] + \mathbb{E}_u[l(g(X), -1)] - \pi_p \mathbb{E}_p[l(g(X), -1)]
\]

\[
\mathbb{E}_u[\cdot] = \mathbb{E}_{X \sim p(x)}[\cdot], \quad X_u = \{x_i^u\}_{i=1}^{n_u} \overset{i.i.d.}{\sim} p(x) := \pi_p p(x|Y = +1) + \pi_n p(x|Y = -1)
\]
Theoretical properties of unbiased PU learning
[du Plessis+, NIPS 2014, ICML 2015; Niu+, NIPS 2016]

• Risk estimator is **unbiased**
  \[ \mathbb{E}[\hat{R}_{pu}(g)] = R_{pu}(g) = R(g) \]

• For linear-in-parameter models, estimation error vanishes in the **optimal** parametric rate
  \[ O_p \left( \frac{\pi_p}{\sqrt{n_p}} + \frac{1}{\sqrt{n_u}} \right) \]
  cf. PN learning
  \[ O_p \left( \frac{\pi_p}{\sqrt{n_p}} + \frac{\pi_n}{\sqrt{n_n}} \right) \]

• PU learning can be **better** than PN learning if
  \[ \frac{\pi_p}{\sqrt{n_p}} + \frac{1}{\sqrt{n_u}} < \frac{\pi_n}{\sqrt{n_n}} \]

\[ \pi_p := p(Y = +1), \pi_n := p(Y = -1) \]
How about flexible models, like *deep nets*?

Unbiased PU learning works well in linear-in-parameter models experimentally.
[du Plessis+, NIPS 2014, ICML 2015; Niu+, NIPS 2016]

\[
\phi(x) \xrightarrow{w} y
\]

Unbiased PU learning with flexible model

- Classify even digits and odd digits of MNIST by 3-layer MLP
- $n_p = 100$, $n_n = 50$, $n_u = 59900$
Overfitting and negative risk of unbiased PU

If classifier can perfectly separate P and U, training error w.r.t. 0-1 loss is:

\[
\pi_p \cdot 0 + \frac{1}{n_u} \sum_{x \in X_u} 0 \cdot (1 - l(g(x), -1)) - \frac{\pi_p}{n_p} \sum_{x \in X_p} l(g(x), -1) < 0
\]

- Risk for positive class
- Risk for negative class
PU learning with non-negative risk estimator (non-negative PU learning)

• Idea:

Round-up risk for negative class to zero

• We propose the new risk estimator which is always non-negative

\[
\tilde{R}_{pu}(g) = \frac{\pi_p}{n_p} \sum_{x \in X_p} l(g(x), +1) + \max \left\{ 0, \frac{1}{n_u} \sum_{x \in X_u} l(g(x), -1) - \frac{\pi_p}{n_p} \sum_{x \in X_p} l(g(x), -1) \right\}
\]

classifier \( g: \mathbb{R}^d \to \mathbb{R} \)
loss function \( l: \mathbb{R} \times \{\pm 1\} \to \mathbb{R} \)

\[
\begin{align*}
\pi_p &:= p(Y = +1) \\
\pi_n &:= p(Y = -1) \\
\mathbb{E}_n[\cdot] &:= \mathbb{E}_{X \sim p(x|Y = -1)}[\cdot]
\end{align*}
\]

positive dataset \( X_p = \{x^p_i\}_{i=1}^{n_p} \) i.i.d. \( p(x|Y = +1) \)
unlabeled dataset \( X_u = \{x^u_i\}_{i=1}^{n_u} \) i.i.d. \( p(x) \)
Theoretical analysis of non-negative PU learning

- Risk estimator is **consistent** and its bias decreases **exponentially**
  - Bias is negligible in practice

\[ O_p \left( \exp\left( -1/\left( \pi_p^2/n_p + 1/n_u \right) \right) \right) \]

- Risk estimator may reduce mean squared error
  - Non-negative risk estimator is **more stable**

\[ \mathbb{E}_{x_p,x_u} \left[ \left( \hat{R}_{pu}(g) - R(g) \right)^2 \right] \leq \mathbb{E}_{x_p,x_u} \left[ \left( \hat{R}_{pu}(g) - R(g) \right)^2 \right] \]

  non-negative PU  unbiased PU

- For linear-in-parameter models, estimation error vanishes in the **optimal** parametric rate

\[ O_p \left( \frac{\pi_p}{\sqrt{n_p}} + \frac{1}{\sqrt{n_u}} \right) \]
Large-scale algorithm

• Want to use mini-batch SGD

Our objective function

\[
\frac{\pi_p}{n_p} \sum_{i=1}^{N} \sum_{x \in \mathcal{X}_p^i} l(g(x), +1) + \max \left\{ 0, \sum_{i} \left( \frac{1}{n_u} \sum_{x \in \mathcal{X}_u^i} l(g(x), -1) - \frac{\pi_p}{n_p} \sum_{x \in \mathcal{X}_p} l(g(x), -1) \right) \right\}
\]

• Sum of risks in mini-batches

\[
\sum_{i=1}^{N} \left[ \frac{\pi_p}{n_p} \sum_{x \in \mathcal{X}_p^i} l(g(x), +1) + \max \left\{ 0, \frac{1}{n_u} \sum_{x \in \mathcal{X}_u^i} l(g(x), -1) - \frac{\pi_p}{n_p} \sum_{x \in \mathcal{X}_p} l(g(x), -1) \right\} \right]
\]

\((\mathcal{X}_p^i, \mathcal{X}_u^i): i\text{-th mini-batch } (i = 1, \ldots, N)\)
CIFAR10 experiment

Achieved smaller test error than unbiased PU and even PN!

Training error → 0

• $P=$artifacts
• $N=$animals
• 13-layer CNN
  [Springenberg+, ICLR2015]
• $n_p = 1000$
• $n_n = 562$
• $n_u = 50000$
Conclusions

• We proposed a non-negative risk estimator for PU learning which improves on the state-of-the-art unbiased risk estimators.

• The new risk estimator is more robust against overfitting, and training very flexible model given limited P data becomes possible.

• A large-scale PU learning algorithm was also developed.

• Extensive theoretical analyses were presented.

• Intensive experiments were carried out as well.