Positive-Unlabeled Classification under Class Prior Shift and Asymmetric Error

Nontawat Charoenphakdee[1,2] and Masashi Sugiyama[2,1]

The University of Tokyo[1]

RIKEN AIP[2]

東京大学
THE UNIVERSITY OF TOKYO

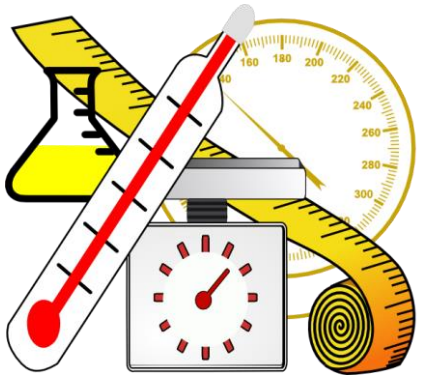
RIKEN

Supervised binary classification (PN classification)

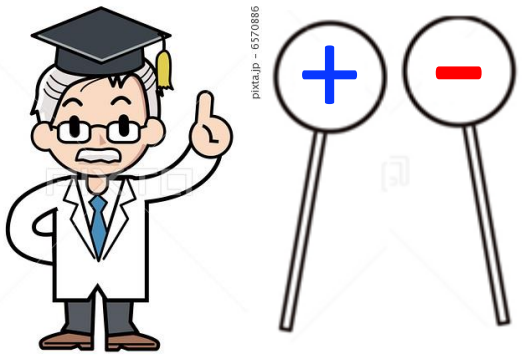
Positive and **Negative** data are given.

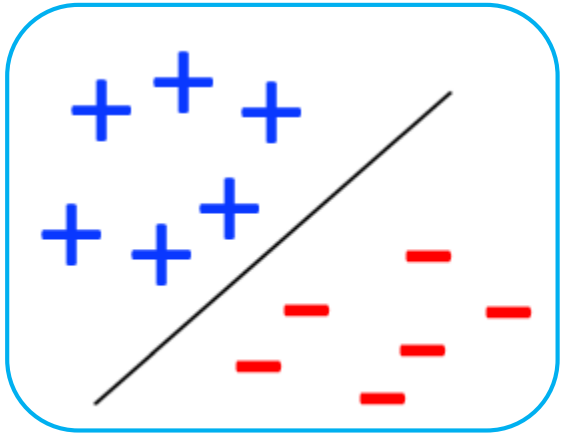
Data collection

Features (input)



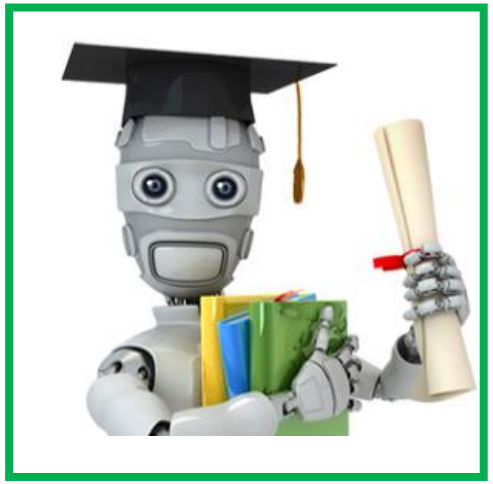
Labels (output)





Machine learning

Binary Classifier



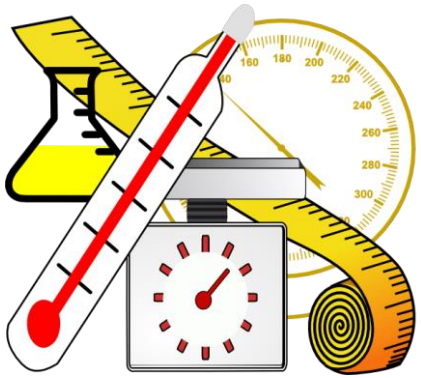
https://t.pimg.jp/006/570/886/1/6570886.jpg
https://www.kullabs.com/uploads/meauring-clip-art-at-clker-com-vector-clip-art-online-royalty-free-H2SJHF-clipart.png
https://d3njjcbhbojbot.cloudfront.net/api/utilities/v1/imageproxy/https://coursera.s3.amazonaws.com/topics/ml/large-icon.png\

Positive-unlabeled classification (PU classification)

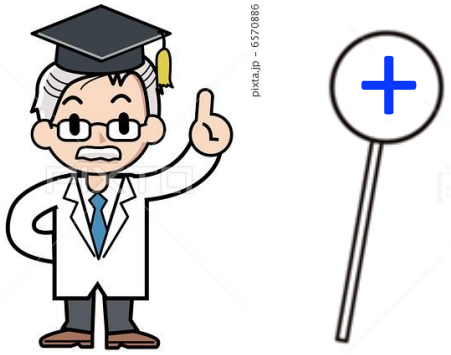
Positive and **Unlabeled** data are given.

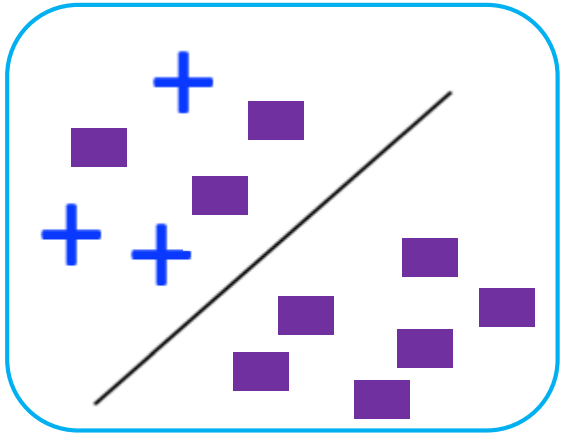
Data collection

Features (input)



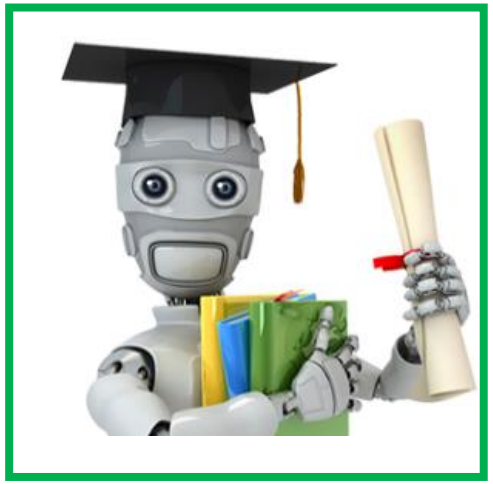
Labels (output)





Machine learning

Binary Classifier



https://t.pimg.jp/006/570/886/1/6570886.jpg
https://www.kullabs.com/uploads/meauring-clip-art-at-clker-com-vector-clip-art-online-royalty-free-H2SJHF-clipart.png
https://d3njjcbhbojbot.cloudfront.net/api/utilities/v1/imageproxy/https://coursera.s3.amazonaws.com/topics/ml/large-icon.png\

Why PU classification?

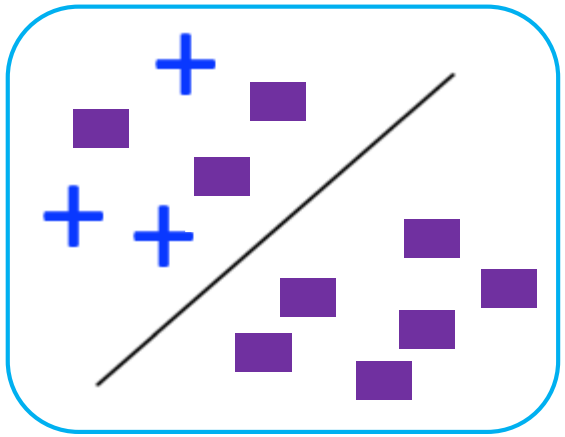
Unlabeled data are **cheaper** to obtain.

Sometimes, **negative** data are **hard to describe**.

In some real-world applications, **collecting negative data is impossible**.

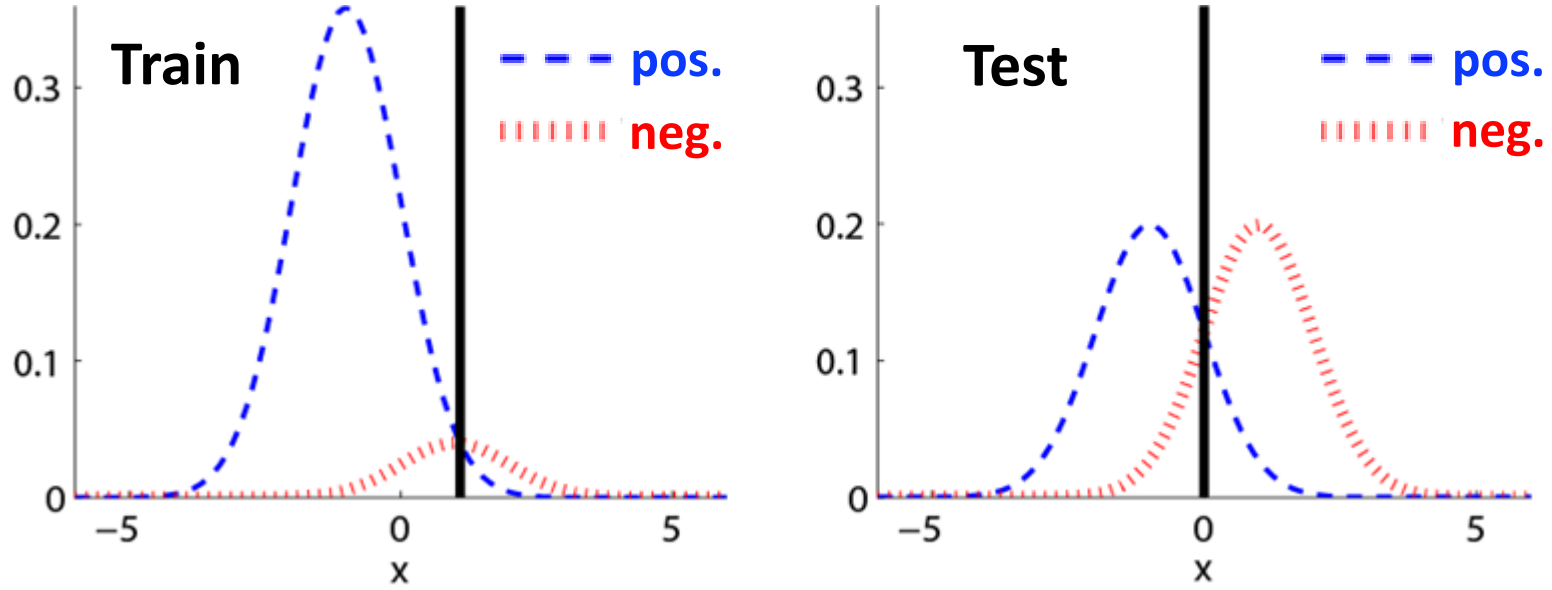
Applications:

- Bioinformatics (Yang+, 2012, Singh-Blom+ 2013, Ren+, 2015)
- Text classification (Li+, 2003)
- Time series classification (Nguyen+, 2011)
- Medical diagnosis (Zuluaga+, 2011)
- Remote-sensing classification (Li+, 2011)



Class prior shift

The ratio of **positive**-**negative** in the **training** and **test** data are different.



Decision boundary is also shifted → **Lead to low accuracy!**

Examples:

- Collect unlabeled data from **the internet**.
- Collect unlabeled data from **all users/patients/etc.** for **personalized application**.

Class prior shift (cont.)

Existing **PU classification** work assumes class prior of **training** and **test** data are the same (du Plessis+, 2014 2015, Kiryo+, 2017).

Existing class prior shift work is not applicable since they require **positive**-**negative** data (Saerens, 2002, du Plessis+, 2012).

PU classification under class prior shift

Observed

Given: Two sets of data

$$\pi : p(y = 1)$$
$$\mathrm{pos}(\boldsymbol{x}) : p(\boldsymbol{x}|y = 1)$$
$$\mathrm{neg}(\boldsymbol{x}) : p(\boldsymbol{x}|y = -1)$$

Positive $X_{\mathrm{P}} := \{\boldsymbol{x}_i^{\mathrm{P}}\}_{i=1}^{n_{\mathrm{P}}} \overset{\text{i.i.d.}}{\sim} \mathrm{pos}(\boldsymbol{x})$

Unlabeled $X_{\mathrm{U}} := \{\boldsymbol{x}_j^{\mathrm{U}}\}_{j=1}^{n_{\mathrm{U}}} \overset{\text{i.i.d.}}{\sim} \pi_{\mathrm{tr}}\mathrm{pos}(\boldsymbol{x}) + (1 - \pi_{\mathrm{tr}})\,\mathrm{neg}(\boldsymbol{x})$

$\pi_{\mathrm{tr}} \neq \pi_{\mathrm{te}}$: **Class prior shift!**

Unobserved

Test $X_{\mathrm{te}} := \{\boldsymbol{x}_k^{\mathrm{te}}\}_{k=1}^{n_{\mathrm{te}}} \overset{\text{i.i.d.}}{\sim} \pi_{\mathrm{te}}\mathrm{pos}(\boldsymbol{x}) + (1 - \pi_{\mathrm{te}})\,\mathrm{neg}(\boldsymbol{x})$

Q: Does class prior shift heavily degrade the performance?

Classifier may fail miserably under class prior shift…

Accuracy reported in mean and std. error of 10 trials with density ratio method.

Accuracy drops heavily!! **Our method**

Dataset	Accuracy (no shift)	Accuracy (shifted)	Accuracy (shifted)
banana	90.1 (0.6)	82.3 (0.5)	87.9 (0.3)
ijcnn1	72.9 (0.4)	**37.8 (0.7)**	71.7 (0.3)
MNIST	86.0 (0.4)	**69.8 (0.7)**	82.5 (0.6)
susy	79.5 (0.5)	**57.5 (0.9)**	75.9 (0.5)
cod-rna	87.4 (0.6)	78.5 (0.6)	84.7 (0.4)
magic	76.7 (0.5)	**60.6 (1.4)**	**79.0 (0.5)**

No shift: $\pi_{\mathrm{tr}} = \pi_{\mathrm{te}} = 0.3$

Shift! $\pi_{\mathrm{tr}} = 0.7, \pi_{\mathrm{te}} = 0.3$

Problem setting

$$\pi : p(y=1) \qquad \mathbb{E}_{\mathrm{P}}[\cdot] : \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathrm{pos}(\boldsymbol{x})}[\cdot]$$
$$\mathrm{pos}(\boldsymbol{x}) : p(\boldsymbol{x}|y=1)$$
$$\mathrm{neg}(\boldsymbol{x}) : p(\boldsymbol{x}|y=-1) \qquad \mathbb{E}_{\mathrm{N}}[\cdot] : \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathrm{neg}(\boldsymbol{x})}[\cdot]$$

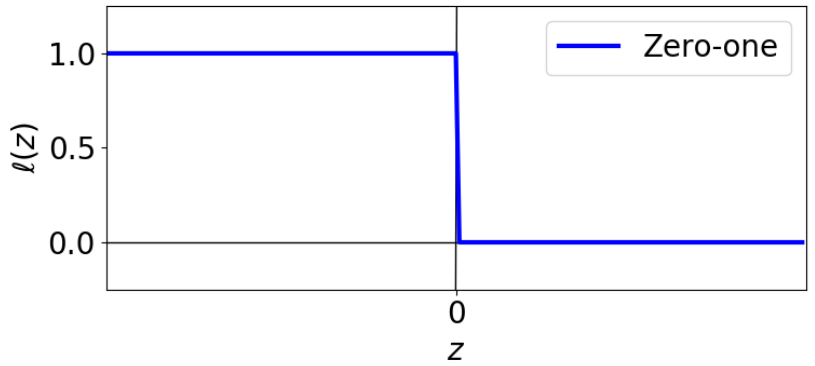
- **Given**: Two sets of data and test class prior π_{te}

Positive $X_{\mathrm{P}} := \{\boldsymbol{x}_i^{\mathrm{P}}\}_{i=1}^{n_{\mathrm{P}}} \overset{\text{i.i.d.}}{\sim} \mathrm{pos}(\boldsymbol{x})$

Unlabeled $X_{\mathrm{U}} := \{\boldsymbol{x}_j^{\mathrm{U}}\}_{j=1}^{n_{\mathrm{U}}} \overset{\text{i.i.d.}}{\sim} \pi_{\mathrm{tr}}\mathrm{pos}(\boldsymbol{x}) + (1-\pi_{\mathrm{tr}})\,\mathrm{neg}(\boldsymbol{x})$

- **Goal**: Find a prediction function g that minimizes

$$R^{\ell_{0\text{-}1}}_{\mathrm{Shift}}(g) = \pi_{\mathrm{te}}\,\mathbb{E}_{\mathrm{P}}\left[\ell_{0\text{-}1}(g(\boldsymbol{x}))\right] + (1-\pi_{\mathrm{te}})\,\mathbb{E}_{\mathrm{N}}\left[\ell_{0\text{-}1}(-g(\boldsymbol{x}))\right]$$



Proposed methods

We proposed two approaches for **PU classification** under **class prior shift:**

- **Risk minimization approach**:

  Learn a classifier based on **empirical risk minimization** principle (Vapnik, 1998).

- **Density ratio approach**:

  1. Estimate a **density ratio** of **positive** and **unlabeled** densities.
  2. Use an appropriate threshold to classify.

Later, we will show that our methods are also applicable for **PU classification** with **asymmetric error**.

Risk minimization approach

Consider the following classification risk:

$$R^{\ell_{0\text{-}1}}_{\mathrm{Shift}}(g) = \pi_{\mathrm{te}}\,\mathbb{E}_{\mathrm{P}}\left[\ell_{0\text{-}1}(g(\boldsymbol{x}))\right] + (1-\pi_{\mathrm{te}})\,\mathbb{E}_{\mathrm{N}}\left[\ell_{0\text{-}1}(-g(\boldsymbol{x}))\right]$$

With $\mathbb{E}_{\mathrm{u}}[\cdot] = \pi_{\mathrm{tr}}\,\mathbb{E}_{\mathrm{P}}[\cdot] + (1-\pi_{\mathrm{tr}})\mathbb{E}_{\mathrm{N}}[\cdot]$, we can rewrite $R^{\ell_{0\text{-}1}}_{\mathrm{Shift}}(g)$ as

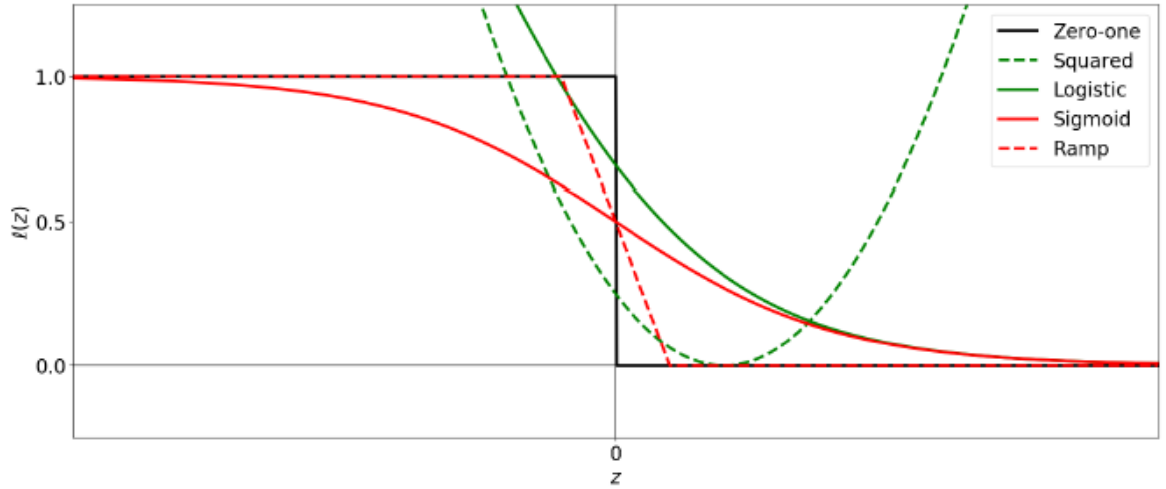
$$R^{\ell_{0\text{-}1}}_{\mathrm{Shift}}(g) = \mathbb{E}_{\mathrm{P}}\left[\pi_{\mathrm{te}}\ell_{0\text{-}1}(g(\boldsymbol{x})) - \frac{\pi_{\mathrm{tr}}(1-\pi_{\mathrm{te}})}{1-\pi_{\mathrm{tr}}}\ell_{0\text{-}1}(-g(\boldsymbol{x}))\right] + \frac{1-\pi_{\mathrm{te}}}{1-\pi_{\mathrm{tr}}}\mathbb{E}_{\mathrm{u}}\left[\ell_{0\text{-}1}(-g(\boldsymbol{x}))\right]$$

Equivalent to existing methods (du Plessis+, 2015) if $\pi_{\mathrm{tr}} = \pi_{\mathrm{te}}$.

No access to distribution: we minimize **empirical error** (Vapnik, 1998):

$$\widehat{R}^{\ell_{0\text{-}1}}_{\mathrm{PU\text{-}shift}}(g) = \frac{1}{n_{\mathrm{P}}}\sum_{i=1}^{n_{\mathrm{P}}}\left[\pi_{\mathrm{te}}\ell_{0\text{-}1}(g(\boldsymbol{x}^{\mathrm{p}}_i)) - \frac{\pi_{\mathrm{tr}}(1-\pi_{\mathrm{te}})}{1-\pi_{\mathrm{tr}}}\ell_{0\text{-}1}(-g(\boldsymbol{x}^{\mathrm{p}}_i))\right] + \frac{1}{n_{\mathrm{U}}}\frac{1-\pi_{\mathrm{te}}}{1-\pi_{\mathrm{tr}}}\sum_{j=1}^{n_{\mathrm{U}}}\ell_{0\text{-}1}(-g(\boldsymbol{x}^{\mathrm{u}}_j))$$

Surrogate losses for binary classification



Directly minimize 0-1 loss is difficult.

- NP-Hard, discontinuous, not differentiable (Ben-david+, 2003, Feldman+, 2012)

In practice, minimize a **surrogate loss** (regularization can also be added):

$$\widehat{R}^{\ell}_{\text{PU-shift}}(g) = \frac{1}{n_{\text{P}}}\sum_{i=1}^{n_{\text{P}}}\left[\pi_{\text{te}}\ell(g(\boldsymbol{x}_i^{\text{P}})) - \frac{\pi_{\text{tr}}(1-\pi_{\text{te}})}{1-\pi_{\text{tr}}}\ell(-g(\boldsymbol{x}_i^{\text{P}}))\right] + \frac{1}{n_{\text{U}}}\frac{1-\pi_{\text{te}}}{1-\pi_{\text{tr}}}\sum_{j=1}^{n_{\text{U}}}\ell(-g(\boldsymbol{x}_j^{\text{u}}))$$

Density ratio estimation

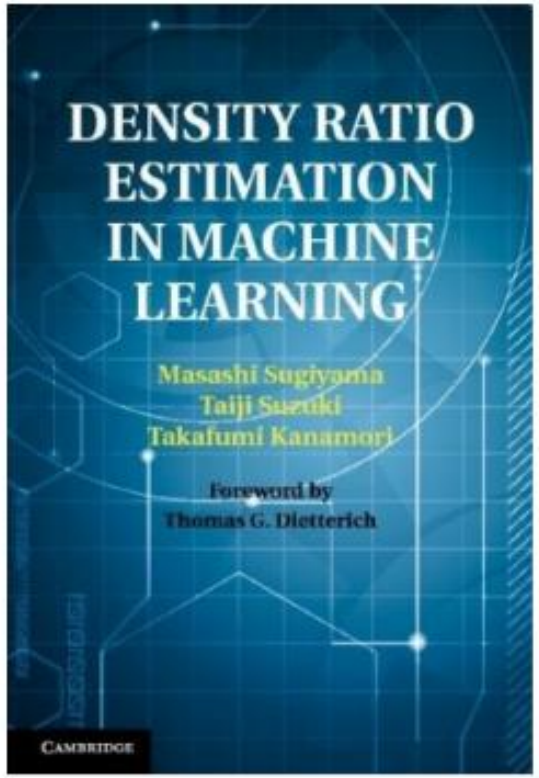
Goal: Estimate the density ratio:

$$r(\boldsymbol{x}) = \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})}$$

from two sets of data

$$X_{\mathrm{nu}} := \{\boldsymbol{x}_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{nu}}(\boldsymbol{x})$$

$$X_{\mathrm{de}} := \{\boldsymbol{x}_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}} \overset{\text{i.i.d.}}{\sim} p_{\mathrm{de}}(\boldsymbol{x})$$



Please check this book to learn more about density ratio estimation (Sugiyama+, 2012)

Applications: outlier detection (Hido+, 2011), change-point detection (Liu+, 2013), robot control (Hachiya+, 2009) event detection in images/movies/text (Yamanaka, 2011, Matsugu, 2011, Liu, 2012), etc.

Naïve approach: estimate $\hat{p}_{\mathrm{nu}}(\boldsymbol{x})$, $\hat{p}_{\mathrm{de}}(\boldsymbol{x})$ separately then perform division $\frac{\hat{p}_{\mathrm{nu}}(\boldsymbol{x})}{\hat{p}_{\mathrm{de}}(\boldsymbol{x})}$.

Does not work well (estimation error is amplified from division operation).

Unconstrained least-squares important fitting (uLSIF)

Goal: Estimate the density ratio: $r(\boldsymbol{x}) = \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})}$ (Kanamori+, 2012)

How: estimate $\widehat{r}$ by minimizing squared loss objective:

$$\mathrm{SQ}(\widehat{r}) = \int \left(\widehat{r}(\boldsymbol{x}) - r(\boldsymbol{x})\right)^2 p_{\mathrm{de}}(\boldsymbol{x}) d\boldsymbol{x}$$

Squared loss decomposition:

$$\mathrm{SQ}(\widehat{r}) = \int \left(\widehat{r}(\boldsymbol{x})\right)^2 p_{\mathrm{de}}(\boldsymbol{x}) d\boldsymbol{x} - 2 \int \widehat{r}(\boldsymbol{x}) p_{\mathrm{nu}}(\boldsymbol{x}) d\boldsymbol{x} + \mathrm{Constant}$$

Empirical minimization (constant can be safely ignored):

$$\widehat{\mathrm{SQ}}(\widehat{r}) = \frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} \left(\widehat{r}(\boldsymbol{x}_j^{\mathrm{de}})\right)^2 - \frac{2}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} \widehat{r}(\boldsymbol{x}_i^{\mathrm{nu}})$$

Unconstrained least-squares important fitting (cont.)

Model: linear-in parameter model (Kanamori+, 2012)

$$\hat{r}(\boldsymbol{x}) = \sum_b \theta_b \phi_b(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x})$$

$\phi_b(\boldsymbol{x})$: basis function (e.g., Gaussian kernel)

Objective:

$$\min_{\boldsymbol{\theta}} \left[\frac{1}{2} \boldsymbol{\theta}^\top \widehat{\boldsymbol{H}} \boldsymbol{\theta} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\theta} + \frac{\lambda}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta} \right]$$

$$\widehat{\boldsymbol{H}} = \frac{1}{n_{\mathrm{de}}} \sum_{j=1}^{n_{\mathrm{de}}} \phi(\boldsymbol{x}_j^{\mathrm{de}}) \phi(\boldsymbol{x}_j^{\mathrm{de}})^\top$$

$$\widehat{\boldsymbol{h}} = \frac{1}{n_{\mathrm{nu}}} \sum_{i=1}^{n_{\mathrm{nu}}} \phi(\boldsymbol{x}_i^{\mathrm{nu}})$$

λ : regularization parameter

$\boldsymbol{I}$: identity matrix

Global solution can be computed **analytically:** $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I})^{-1} \widehat{\boldsymbol{h}}$

Parameter tuning (regularization, basis) can be done by **cross-validation.**

Density ratio approach

$$\mathrm{pos}(\boldsymbol{x}) : p(\boldsymbol{x}|y=1)$$
$$\mathrm{neg}(\boldsymbol{x}) : p(\boldsymbol{x}|y=-1)$$
$$\mathrm{unl}(\boldsymbol{x}) = \pi_{\mathrm{tr}}\mathrm{pos}(\boldsymbol{x}) + (1-\pi_{\mathrm{tr}})\mathrm{neg}(\boldsymbol{x})$$

Consider Bayes-optimal classifier of binary classification (no prior shift)

$$f^*_{\mathrm{Bayes}}(\boldsymbol{x}) = \mathrm{sign}\left[p(y=+1|\boldsymbol{x}) - \frac{1}{2}\right]$$

We can rewrite it as

$$f^*_{\mathrm{Bayes}}(\boldsymbol{x}) = \mathrm{sign}\left[\pi_{\mathrm{tr}}\frac{\mathrm{pos}(\boldsymbol{x})}{\mathrm{unl}(\boldsymbol{x})} - \frac{1}{2}\right]$$

Density ratio!

Another formulation is

$$f^*_{\mathrm{Bayes}}(\boldsymbol{x}) = \mathrm{sign}\left[\pi_{\mathrm{tr}} - \frac{1}{2}\frac{\mathrm{unl}(\boldsymbol{x})}{\mathrm{pos}(\boldsymbol{x})}\right]$$

Q1: How to modify when class prior shift occurs?

Q2: Which formulation is preferable?

Q1: Density ratio approach (shift)

Consider Bayes-optimal classifier of binary classification

$$f^*_{\mathrm{Bayes}}(\boldsymbol{x}) = \mathrm{sign}\left[p(y=+1|\boldsymbol{x}) - \frac{1}{2}\right]$$

$$\mathrm{pos}(\boldsymbol{x}) : p(\boldsymbol{x}|y=1)$$
$$\mathrm{neg}(\boldsymbol{x}) : p(\boldsymbol{x}|y=-1)$$
$$\mathrm{unl}(\boldsymbol{x}) = \pi_{\mathrm{tr}}\mathrm{pos}(\boldsymbol{x}) + (1-\pi_{\mathrm{tr}})\mathrm{neg}(\boldsymbol{x})$$

We can rewrite it as

$$f^*_{\mathrm{Bayes}}(\boldsymbol{x}) = \mathrm{sign}\left[\pi_{\mathrm{tr}}\frac{\mathrm{pos}(\boldsymbol{x})}{\mathrm{unl}(\boldsymbol{x})} - \frac{\pi_{\mathrm{tr}}(1-\pi_{\mathrm{te}})}{\pi_{\mathrm{te}}+\pi_{\mathrm{tr}}-2\pi_{\mathrm{tr}}\pi_{\mathrm{te}}}\right]$$

Density ratio!

Another formulation is

$$f^*_{\mathrm{Bayes}}(\boldsymbol{x}) = \mathrm{sign}\left[\frac{\pi_{\mathrm{te}}+\pi_{\mathrm{tr}}-2\pi_{\mathrm{tr}}\pi_{\mathrm{te}}}{(1-\pi_{\mathrm{te}})} - \frac{\mathrm{unl}(\boldsymbol{x})}{\mathrm{pos}(\boldsymbol{x})}\right]$$

Simply modifying the threshold can solve this problem!

Q2: Difficulty of density ratio estimation

In general, density ratio is **unbounded**. ☹

$$r(\boldsymbol{x}) = \frac{p_{\mathrm{nu}}(\boldsymbol{x})}{p_{\mathrm{de}}(\boldsymbol{x})}$$

$r(\boldsymbol{x})$ is unbounded when $p_{\mathrm{de}}(\boldsymbol{x}) = 0$.

This raises issues of robustness and stability.

We show that the density ratio $\frac{\mathrm{pos}(\boldsymbol{x})}{\mathrm{unl}(\boldsymbol{x})}$ is bounded in PU classification. ☺

Q2: Density ratio in PU

$$\mathrm{pos}(\boldsymbol{x}) : p(\boldsymbol{x}|y=1)$$
$$\mathrm{neg}(\boldsymbol{x}) : p(\boldsymbol{x}|y=-1)$$
$$\mathrm{unl}(\boldsymbol{x}) = \pi_{\mathrm{tr}}\mathrm{pos}(\boldsymbol{x}) + (1-\pi_{\mathrm{tr}})\mathrm{neg}(\boldsymbol{x})$$

In **PU classification**, density ratio $\frac{\mathrm{pos}(\boldsymbol{x})}{\mathrm{unl}(\boldsymbol{x})}$ is bounded.

$$0 \leq \frac{\mathrm{pos}(\boldsymbol{x})}{\mathrm{unl}(\boldsymbol{x})} \leq \frac{1}{\pi_{\mathrm{tr}}}$$ Lower and upper bounded ☺

$$\pi_{\mathrm{tr}} \leq \frac{\mathrm{unl}(\boldsymbol{x})}{\mathrm{pos}(\boldsymbol{x})}$$ Unbounded from above ☹

Insight: estimate $\frac{\mathrm{pos}(\boldsymbol{x})}{\mathrm{unl}(\boldsymbol{x})}$ is preferable.

Our experimental results agree with this observation.

Experiments: class prior shift train 0.7 -> test 0.3

Datasets: ***banana, ijcnn1, MNIST, susy, cod-rna, magic***

Methods:

- Density ratio $\frac{\mathrm{pos}(\boldsymbol{x})}{\mathrm{unl}(\boldsymbol{x})}$ **($\frac{\boldsymbol{p}}{\boldsymbol{u}}$uLSIF)**
- Density ratio $\frac{\mathrm{unl}(\boldsymbol{x})}{\mathrm{pos}(\boldsymbol{x})}$ **($\frac{\boldsymbol{u}}{\boldsymbol{p}}$ uLSIF)**
- Linear-in input model (Lin): Double hinge loss **(DH-Lin)**, squared loss **(Sq-Lin)**
- Kernel model (Ker): Double hinge loss **(DH-Ker)** , squared loss **(Sq-Ker)**

Parameter selection: (regularization, kernel width) 5-fold cross-validation.

We also investigated when wrong test class prior is given.

Results reported in mean and std. error of accuracy of 10 trials.
Outperforming methods are bolded based on one-sided t-test with significance level 5%.
Dataset information and more experiments and can be found in the paper.

Results: class prior shift $\pi_{\mathrm{tr}} = 0.7, \pi_{\mathrm{te}} = 0.3$

Dataset	π^{g}	$\frac{u}{p}$uLSIF	$\frac{p}{u}$uLSIF	DH-Lin	DH-Ker	Sq-Lin	Sq-Ker
banana		83.0(1.0)	**86.4 (0.5)**	70.2(0.5)	78.3(1.0)	70.0(0.0)	83.4(0.4)
ijcnn1		70.8(0.6)	**74.2 (0.7)**	70.0(0.1)	69.8(0.2)	71.5(0.3)	69.2(0.5)
MNIST	π'	79.3(0.5)	**81.7 (0.5)**	74.0(1.1)	**82.4 (1.0)**	52.3(1.4)	**83.4 (0.9)**
susy	0.3	74.3(0.5)	**76.0 (0.3)**	72.7(0.6)	70.0(0.0)	**75.5 (1.4)**	74.7(0.7)
cod-rna		82.1(1.0)	82.8(0.8)	**87.3 (0.7)**	77.3(0.8)	**85.2 (1.1)**	80.2(1.0)
magic		71.5(0.7)	**75.8 (0.6)**	72.7(1.1)	70.8(0.4)	**75.0 (1.0)**	72.9(0.7)
banana		84.7(1.1)	**88.7 (0.7)**	54.9(1.4)	81.7(1.6)	53.6(1.2)	83.8(1.3)
ijcnn1		**64.9 (1.4)**	**66.6 (1.0)**	60.4(1.4)	51.6(3.0)	62.2(1.2)	48.2(2.8)
MNIST	0.5	81.9(0.4)	**84.1 (0.6)**	72.5(1.0)	82.5(0.7)	52.9(1.1)	81.9(0.9)
susy		**75.9 (1.1)**	**77.0 (0.6)**	67.5(1.4)	75.5(0.6)	71.6(1.0)	72.8(1.1)
cod-rna		**85.3 (0.7)**	**85.4 (0.5)**	**86.2 (0.7)**	80.1(1.1)	**86.5 (0.9)**	81.2(1.2)
magic		67.6(0.8)	**73.6 (0.9)**	**72.6 (0.7)**	62.4(1.9)	**71.8 (0.7)**	68.9(0.8)
banana		**80.6 (1.3)**	**82.1 (1.1)**	31.8(0.9)	48.9(1.5)	30.0(0.0)	69.9(1.1)
ijcnn1		35.2(1.4)	**42.4 (0.9)**	30.0(0.0)	30.0(0.0)	32.4(0.5)	30.9(0.4)
MNIST	π	**79.9 (0.7)**	72.6(0.6)	71.1(1.1)	64.8(1.1)	64.0(0.6)	74.2(1.0)
susy	0.7	35.6(3.1)	**44.2 (2.9)**	30.0(0.0)	30.0(0.0)	**42.0 (1.5)**	36.8(1.3)
cod-rna		**77.7 (2.2)**	**77.8 (2.1)**	**79.6 (0.7)**	67.8(0.8)	78.2(0.5)	68.3(1.0)
magic		51.6(0.3)	**60.3 (1.5)**	**56.2 (2.7)**	32.8(0.7)	**58.7 (1.4)**	50.1(1.6)

Correct test prior is given

Wrong test prior is given

Traditional PU

Preferable method in our experiments

(density ratio $\frac{p}{u}$uLSIF)

PU classification with asymmetric error

$\pi_{\mathrm{tr}} : p(y=1)$
$\mathrm{pos}(\boldsymbol{x}) : p(\boldsymbol{x}|y=1)$
$\mathrm{neg}(\boldsymbol{x}) : p(\boldsymbol{x}|y=-1)$
$\alpha \in (0,1)$: false negative error

$\mathbb{E}_{\mathrm{P}}[\cdot] : \mathbb{E}_{\boldsymbol{x}\sim\mathrm{pos}(\boldsymbol{x})}[\cdot]$
$\mathbb{E}_{\mathrm{N}}[\cdot] : \mathbb{E}_{\boldsymbol{x}\sim\mathrm{neg}(\boldsymbol{x})}[\cdot]$

- **Given**: Given two sets of sample:

Positive $X_{\mathrm{P}} := \{\boldsymbol{x}_i^{\mathrm{P}}\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} \mathrm{pos}(\boldsymbol{x})$

Unlabeled $X_{\mathrm{U}} := \{\boldsymbol{x}_i^{\mathrm{U}}\}_{i=1}^{n'} \overset{\text{i.i.d.}}{\sim} \pi_{\mathrm{tr}}\mathrm{pos}(\boldsymbol{x}) + (1-\pi_{\mathrm{tr}})\,\mathrm{neg}(\boldsymbol{x})$

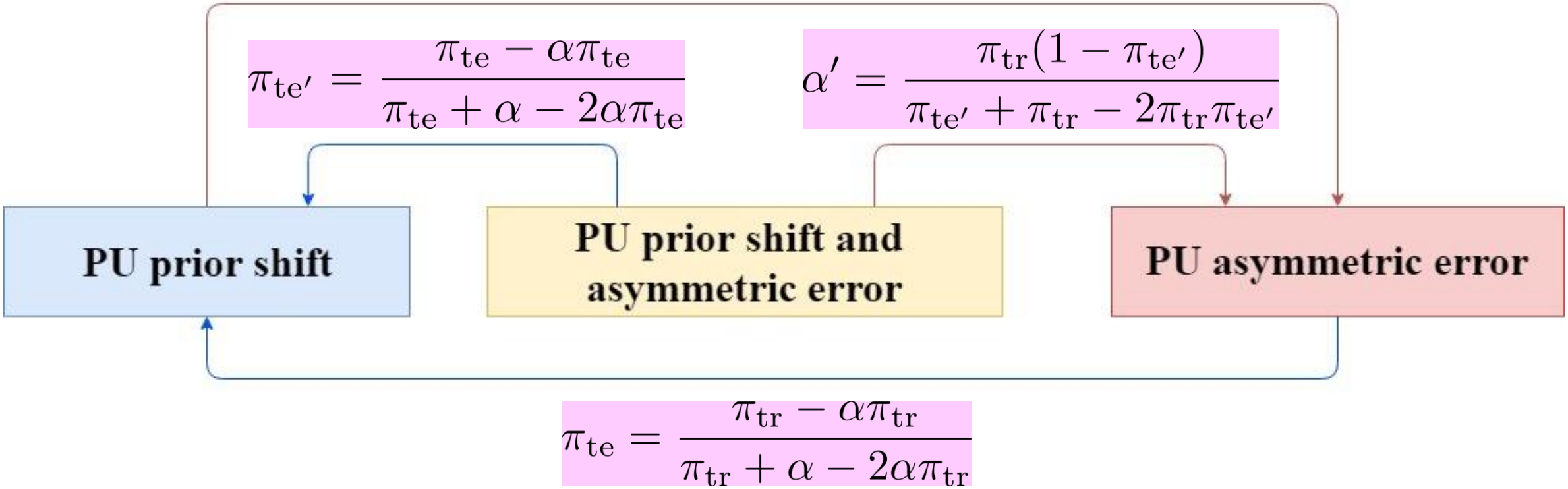
- **Goal**: Find a prediction function g that minimizes

$$R^{\ell}_{\mathrm{Asym}}(g) = (1-\alpha)\pi_{\mathrm{tr}}\,\mathbb{E}_{\mathrm{P}}\left[\ell(g(\boldsymbol{x}_{\mathrm{P}}))\right] + \alpha(1-\pi_{\mathrm{tr}})\,\mathbb{E}_{\mathrm{N}}\left[\ell(-g(\boldsymbol{x}_{\mathrm{N}}))\right]$$

Reduce to symmetric error when $\alpha = 0.5$

The equivalence of prior shift and asymmetric error

$$\alpha = \frac{\pi_{\mathrm{tr}}(1-\pi_{\mathrm{te}})}{\pi_{\mathrm{te}}+\pi_{\mathrm{tr}}-2\pi_{\mathrm{tr}}\pi_{\mathrm{te}}}$$



We can relate these problems based on the analysis of Bayes-optimal classifier.

Conclusion

Class prior shift may **heavily degrade** the performance of positive-unlabeled classification **(PU classification)**.

- Proposed **two approaches** for handling this problem **effectively**:
  - Risk minimization approach
  - Density ratio approach
- Showed the **equivalence** of **class prior shift** and **asymmetric error** problems in **PU classification**.
  - Our methods are **applicable** for **both problems**.
  - Also **applicable** when considering **both problems simultaneously**.
- Poster: #31: May 2nd from 7:00-9:00PM