Positive-Unlabeled Learning

송헌 (songheony@gmail.com)
M.S. candidate at Kyushu University
Computational Learning Theory Team, RIKEN AIP

CONTENTS

PU Learning의 공식화

- Convex formulation for learning from positive and unlabeled data [Du Plessis et al., ICML, 2015]

위 공식의 문제를 해결

- Positive-Unlabeled Learning with Non-Negative Risk Estimator [Kiryo Ryuichi, et al. NIPS, 2017]

PNU Learning의 제안

- Semi-supervised classification based on classification from positive and unlabeled data [Sakai Tomoya, et al. ICML, 2017]

- Theoretical Comparisons of Positive-Unlabeled Learning against Positive-Negative Learning [Niu Gang, et al. NIPS, 2016]

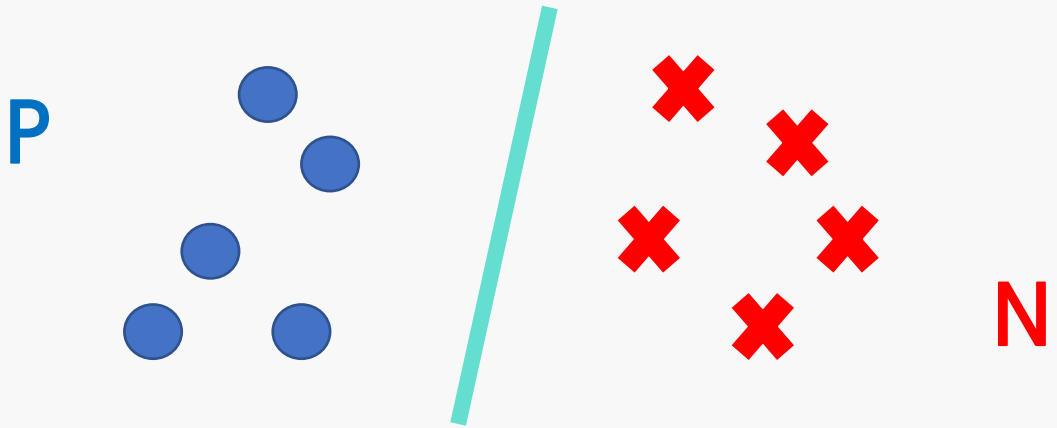
위 연구들의 이론적 분석

CONTENTS

01 Binary Supervised Classification (PN Learning)

- 레이블 $y \in \{+1, -1\}$ 이 붙어있는 샘플들 $\{(x_i, y_i)\}_{i=1}^{n} \sim p(x, y)$ 이 주어진다.

- 리스크 $R(f) = \mathbb{E}_{p(x,y)}\left[\ell\left(yf(x)\right)\right]$ 을 최소화하는 것을 목표로 한다.

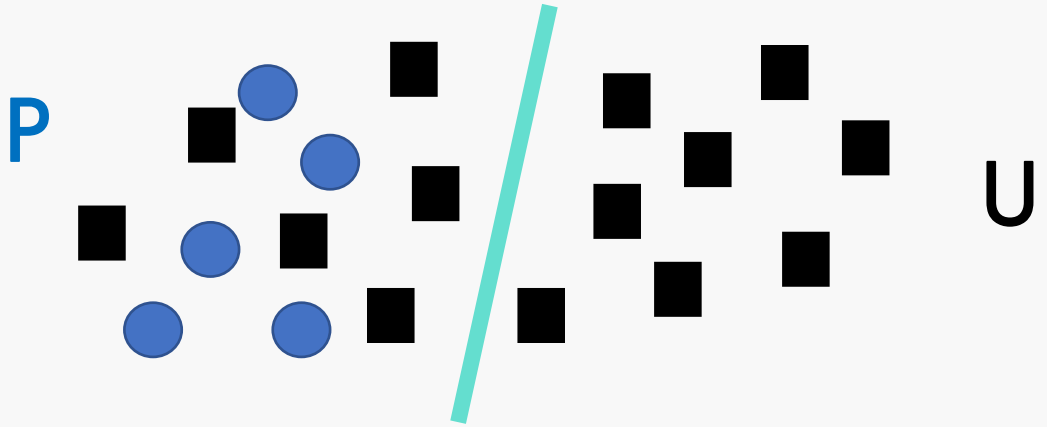


- 이때 Empirical 리스크는 최적의 경우 다음과 같이 나타내어진다.

$$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i f(x_i)\right) = R(f) + O\left(\frac{1}{\sqrt{n}}\right)$$

01 Positive-Unlabeled (PU) Learning

- Positive 샘플들 $\{x_i^P\}_{i=1}^{n_P} \sim p(x|y=+1)$ 과
  Unlabeled 샘플들 $\{x_i^U\}_{i=1}^{n_U} \sim p(x)$ 이 주어진다.



- 데이터 클릭여부, 친구 여부 등 다양한 경우가 있을 수 있음.

Non-convex

- 기존의 방법들은 이론적 분석이 되어있지 않거나, 실제 리스크가 아닌 편향된 목적 함수를 최적화 하는 경우가 많음.

02 Convex formulation for learning from positive and unlabeled data

- 최소화 하려는 리스크는 다음과 같이 표현이 가능함

$$R(f) = \mathbb{E}_{p(x,y)}\left[\ell\left(yf(x)\right)\right]$$
$$= \pi\mathbb{E}_{p(x|y=+1)}\left[\ell\left(f(x)\right)\right] + (1-\pi)\mathbb{E}_{p(x|y=-1)}\left[\ell\left(-f(x)\right)\right]$$

단, $\pi = p(y = +1)$은 Class-prior을 나타내며, 이를 추정하는 많은 방법론들이 제안되어 있음.

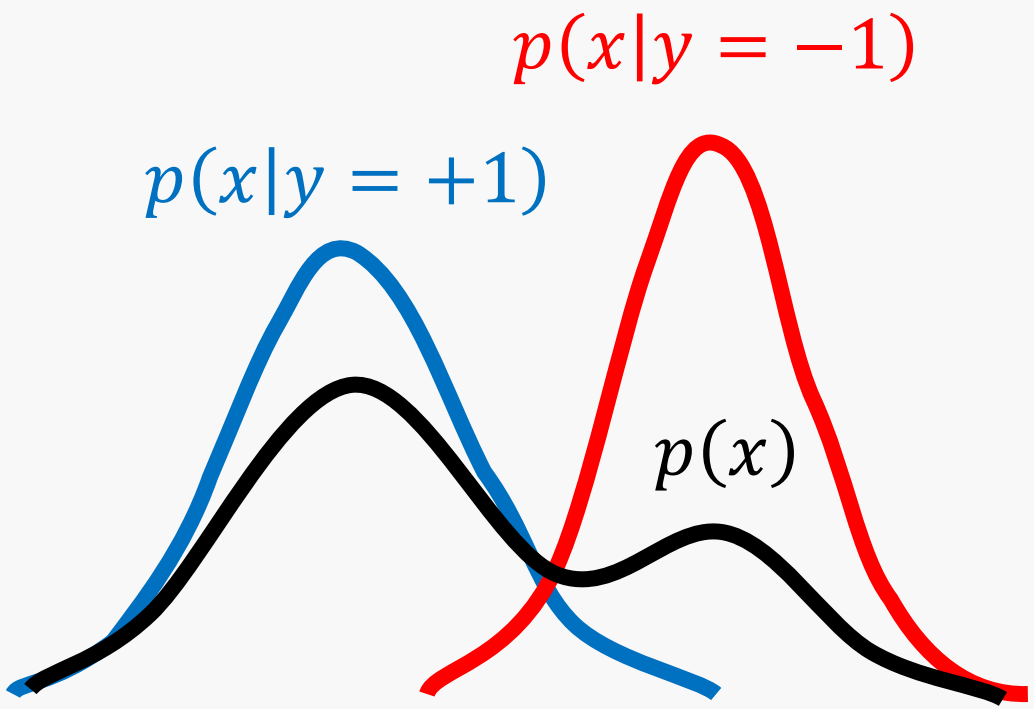
- 이때, Negative 샘플들은 주어지지 않았기 때문에, 해당 리스크를 직접적으로 최소화하는 것은 불가능하다.

- 이를 위해서 Unlabeled 샘플들을 이용해 리스크를 추정한다.

02 Convex formulation for learning from positive and unlabeled data

- Unlabeled 샘플의 분포 $p(x)$는 Positive 샘플의 분포 $p(x|y=+1)$ 와 Negative 샘플의 분포 $p(x|y=-1)$ 가 혼합된 형태이다.

$$p(x) = \pi p(x|y=+1) + (1-\pi)p(x|y=-1)$$



02 Convex formulation for learning from positive and unlabeled data

- 이를 이용하면 다음과 같은 식의 변형이 가능하다

$$p(x) = \pi p(x|y=+1) + (1-\pi)p(x|y=-1)$$

$$\begin{aligned} R(f) &= \mathbb{E}_{p(x)}\left[\ell\left(yf(x)\right)\right] \\ &= \pi\mathbb{E}_{p(x|y=+1)}\left[\ell\left(f(x)\right)\right] + (1-\pi)\mathbb{E}_{p(x|y=-1)}\left[\ell\left(-f(x)\right)\right] \\ &= \pi\mathbb{E}_{p(x|y=+1)}\left[\ell\left(f(x)\right)\right] + \mathbb{E}_{p(x)}\left[\ell\left(-f(x)\right)\right] - \pi\mathbb{E}_{p(x|y=-1)}\left[\ell\left(-f(x)\right)\right] \end{aligned}$$

- 즉, 기존의 리스크에서 편향되지 않은 목적함수를 만들 수 있다.

- 또한 convergence rate가 다음과 같이 나타내어진다.

$$\hat{R}(f) = R(f) + O\left(\frac{2\pi}{\sqrt{n_P}} + \frac{1}{\sqrt{n_U}}\right)$$

02 Convex formulation for learning from positive and unlabeled data

- PN Learning과 convergence rate를 비교해보면 다음과 같다

PN Learning

$$O\left(\frac{\pi}{\sqrt{n_P}}+\frac{1-\pi}{\sqrt{n_N}}\right)$$

$$O\left(\frac{2\pi}{\sqrt{n_P}}+\frac{1}{\sqrt{n_U}}\right)$$

PU Learning

- 즉 다음과 같은 조건을 만족하면 오히려 더 좋은 bound를 얻게 된다

$$\frac{\pi}{\sqrt{n_P}}+\frac{1}{\sqrt{n_U}}<\frac{1-\pi}{\sqrt{n_N}}$$

03 Positive-Unlabeled Learning with Non-Negative Risk Estimator

- 단 이러한 방법에는 문제점이 있는데, 손실함수가 $\ell(x) \geq 0, \forall x$를 만족한다면 다음의 조건 또한 만족해야 한다

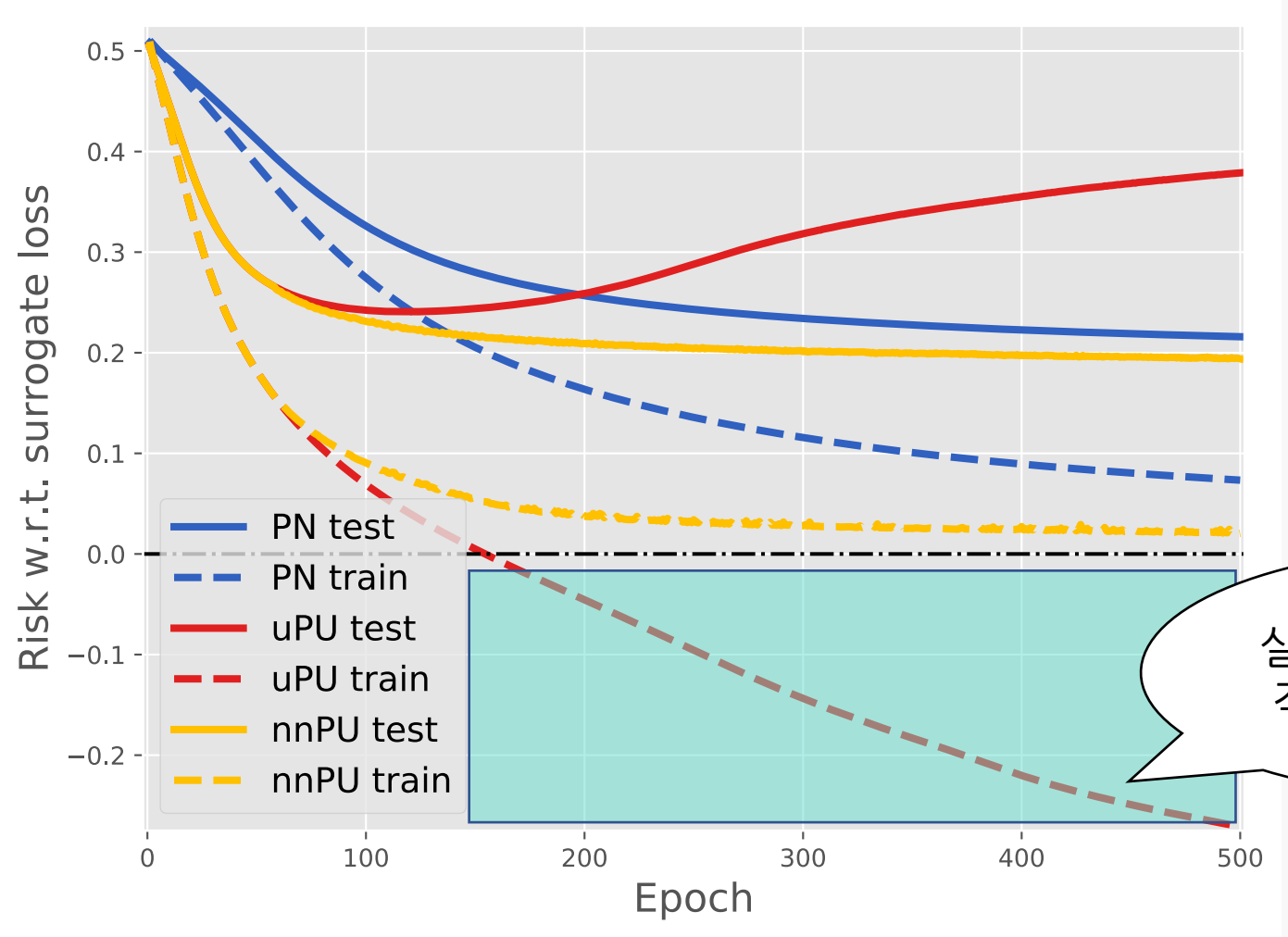
$$\mathrm{R}^{-}(\mathrm{f}) = (1-\pi)\mathbb{E}_{p(x|y=-1)}\left[\ell\left(-f(x)\right)\right] \geq 0$$

- 하지만 실제로 학습을 할 경우에는 Empirical 리스크를 다루기 때문에 이러한 조건을 만족하지 못하는 경우가 발생한다

$$\hat{R}^{-}(\mathrm{f}) = \frac{1}{n_U}\sum_{i=1}^{n_U}\ell\left(-f\left(x_i^U\right)\right) - \frac{\pi}{n_P}\sum_{i=1}^{n_P}\ell\left(-f\left(x_i^P\right)\right) \ngeq 0$$

03 Positive-Unlabeled Learning with Non-Negative Risk Estimator

- 이러한 문제로 인해, 오버피팅이 일어나고 제대로 예측을 못하게 된다.



03 Positive-Unlabeled Learning with Non-Negative Risk Estimator

- 논문에서는 다음과 같은 굉장히 간단한 리스크로 목적함수를 변형시킴

$$\tilde{R}(f) = \frac{\pi}{n_P}\sum_{i=1}^{n_P}\ell\left(f\left(x_i^P\right)\right) + \max\left\{0, \frac{1}{n_U}\sum_{i=1}^{n_U}\ell\left(-f\left(x_i^U\right)\right) - \frac{\pi}{n_P}\sum_{i=1}^{n_P}\ell\left(-f\left(x_i^P\right)\right)\right\}$$

- 변형된 목적함수는 약간의 편향을 갖게 되지만, 샘플수에 따라 편향이 지수적으로 작아지게 된다.

- 또한 여전히 다음과 같은 convergence rate를 가지게 된다

$$\tilde{R}(f) = R(f) + O\left(\frac{\pi}{\sqrt{n_P}} + \frac{1}{\sqrt{n_U}}\right)$$

03 Positive-Unlabeled Learning with Non-Negative Risk Estimator

- 논문에서는 해당 목적함수를 최소화 하기 위하여 다음과 같이, Negative sample에 대한 리스크가 양수가 되도록 조정함.

- If $\frac{1}{n_U}\sum_{i=1}^{n_U}\ell\left(-f\left(x_i^U\right)\right)-\frac{\pi}{n_P}\sum_{i=1}^{n_P}\ell\left(-f\left(x_i^P\right)\right)\geq-\beta$

  목적함수의 기울기 $\nabla\tilde{R}(f)$를 계산

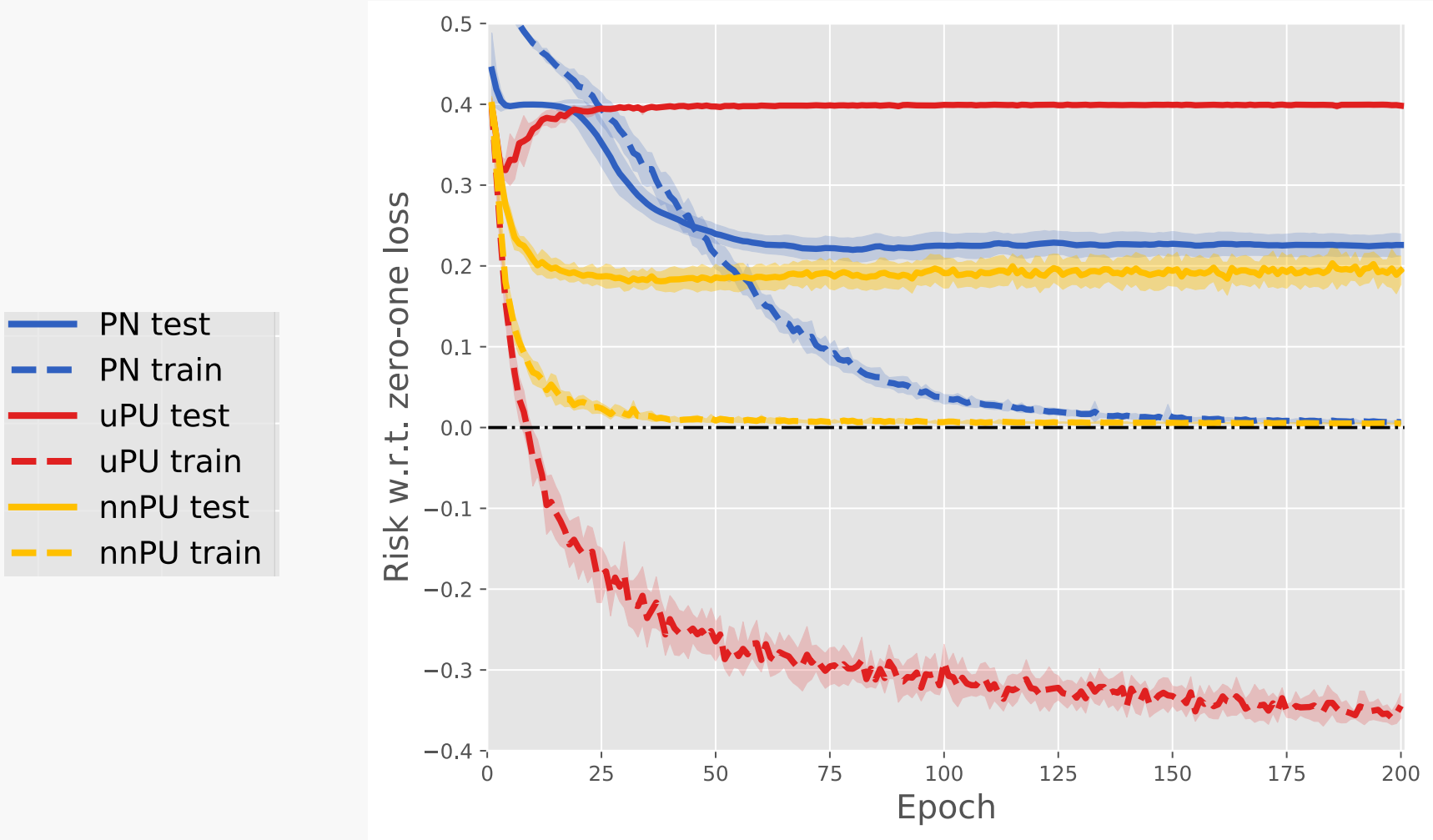
- Else

  기울기 $-\nabla\left(\frac{1}{n_U}\sum_{i=1}^{n_U}\ell\left(-f\left(x_i^U\right)\right)-\frac{\pi}{n_P}\sum_{i=1}^{n_P}\ell\left(-f\left(x_i^P\right)\right)\right)$를 계산

- 계산된 기울기를 바탕으로 Gradient Descent 알고리즘을 실행

03 Positive-Unlabeled Learning with Non-Negative Risk Estimator

- 실제로 충분히 많은 Unlabeled 샘플을 이용하여 실험한 결과 PN Learning 보다 좋은 실험 결과를 얻었다.

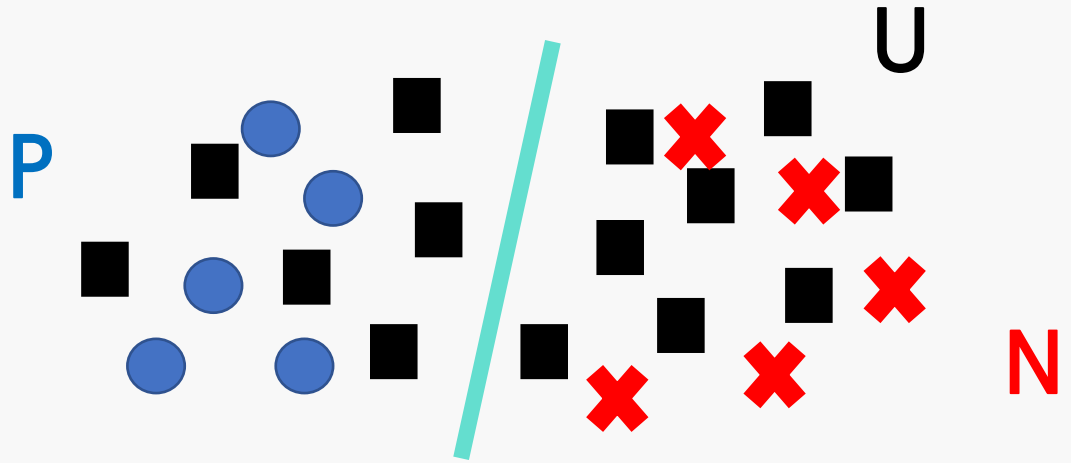


04 Semi-supervised classification based on classification from positive and unlabeled data

- Positive 샘플들 $\{x_i^P\}_{i=1}^{n_P} \sim p(x|y=+1)$,

  Negative 샘플들 $\{x_i^N\}_{i=1}^{n_N} \sim p(x|y=-1)$,

  Unlabeled 샘플들 $\{x_i^U\}_{i=1}^{n_U} \sim p(x)$ 이 주어졌을 때에는?



- 기존에는 Manifold smoothing 등 정규화를 위해 Unlabeled 샘플들을 이용하였음

04 Semi-supervised classification based on classification from positive and unlabeled data

- PN, PU, NU를 조합하여 Unlabeled 샘플들을 직접적으로 이용

- Unlabeled 샘플의 수에 따른 convergence rate의 비교:

  1) n_U가 클 때: PU (NU) > PN > NU (PU)

  2) n_U가 작을 때: PU (NU) > NU (PU) > PN

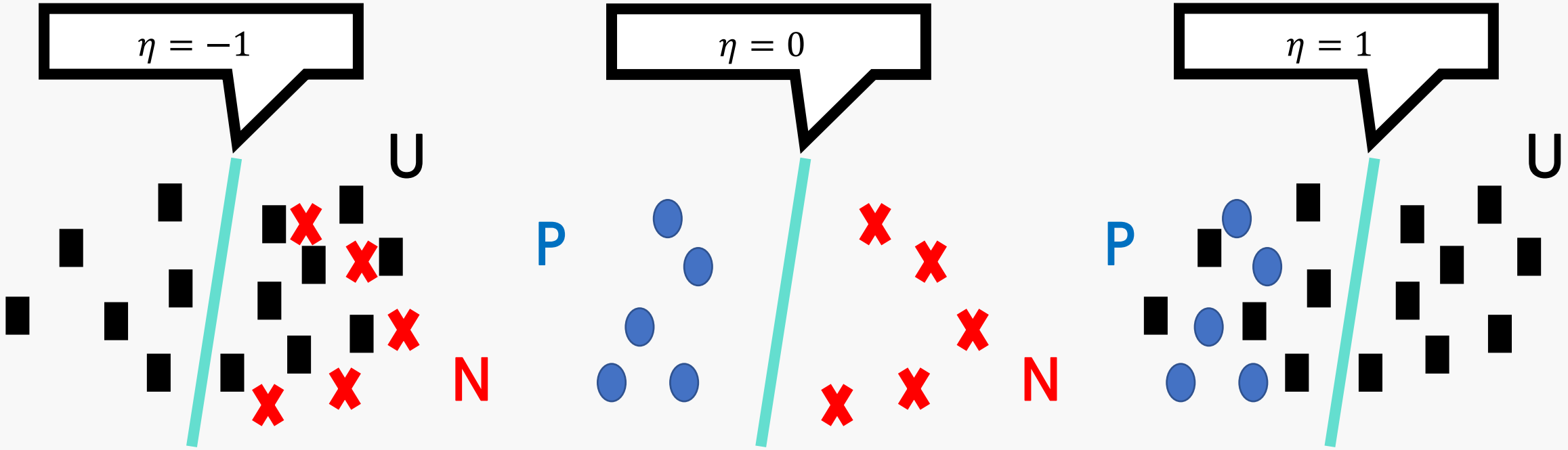
- 즉, PN + PU 혹은 PN + NU가 어떠한 경우에도 좋은 결과를 기대할 수 있음

- 위의 조합을 이용하여 문제를 해결

04 Semi-supervised classification based on classification from positive and unlabeled data

- 따라서 목적함수를 다음과 같이 설계

$$R^{\eta}_{PNU}(f) = \begin{cases} (1-\eta)R_{PN}(f) + \eta R_{PU}(f), & 0 \le \eta \le 1 \\ (1+\eta)R_{PN}(f) - \eta R_{NU}(f), & 0 > \eta \ge -1 \end{cases}$$



- 다음과 같은 convergence rate가 얻어진다

$$\hat{R}_{PNPU}(f) = 2R_{PNPU}(f) + O\left(\frac{(1+\eta)\pi}{\sqrt{n_P}} + \frac{(1-\eta)(1-\pi)}{\sqrt{n_N}} + \frac{\eta}{\sqrt{n_U}}\right)$$

  즉, Unlabeled 샘플들이 학습에 충분히 도움을 줌

- 또한, 특정 η에 대해서는 다음과 같은 관계를 가짐

$$\mathrm{Var}\left[\hat{R}_{PNPU}(f)\right] < \mathrm{Var}\left[\hat{R}_{PN}(f)\right]$$

  즉, 더 안정적으로 학습이 가능하며, cross-validation에 유용함

04 Semi-supervised classification based on classification from positive and unlabeled data

- 실험결과, 제안방법이 기존의 방법들과 PU + NU보다 좋은 성능을 보였음

Data set	n_L	PNU	PUNU	ER	LapSVM	SMIR	WellSVM	S4VM
Banana	10	**30.1 (1.0)**	**32.1 (1.1)**	35.8 (1.0)	36.9 (1.0)	37.7 (1.1)	41.8 (0.6)	45.3 (1.0)
$d = 2$	50	**19.0 (0.6)**	26.4 (1.2)	20.6 (0.7)	21.3 (0.7)	21.1 (1.0)	42.6 (0.5)	38.7 (0.9)
Phoneme	10	32.5 (0.8)	33.5 (1.0)	33.4 (1.2)	36.5 (1.5)	36.4 (1.2)	**28.4 (0.6)**	33.7 (1.4)
$d = 5$	50	28.1 (0.5)	32.8 (0.9)	27.8 (0.6)	27.0 (0.8)	28.6 (1.0)	26.8 (0.4)	**25.1 (0.2)**
Magic	10	**31.7 (0.8)**	34.1 (0.9)	34.2 (1.1)	37.9 (1.3)	36.0 (1.2)	**30.1 (0.8)**	33.3 (0.9)
$d = 10$	50	**29.9 (0.8)**	33.4 (0.9)	30.9 (0.5)	31.0 (0.9)	**30.8 (0.9)**	**28.8 (0.8)**	**29.2 (0.4)**
Image	10	**29.8 (0.9)**	**31.7 (0.8)**	33.7 (1.1)	36.6 (1.2)	36.7 (1.2)	34.7 (1.1)	35.9 (1.0)
$d = 18$	50	**20.7 (0.8)**	26.6 (1.1)	**20.8 (0.8)**	**20.3 (1.0)**	**20.9 (0.9)**	27.2 (1.0)	23.2 (0.7)
Susy	10	**44.6 (0.6)**	**45.0 (0.6)**	47.7 (0.4)	48.2 (0.4)	**45.1 (0.7)**	48.0 (0.3)	46.8 (0.3)
$d = 18$	50	**38.9 (0.6)**	41.5 (0.6)	**37.9 (0.7)**	43.1 (0.6)	43.9 (0.8)	43.8 (0.7)	42.1 (0.4)
German	10	**40.8 (0.9)**	**42.4 (0.7)**	43.6 (0.9)	45.9 (0.7)	46.2 (0.8)	**42.4 (0.8)**	**42.0 (0.7)**
$d = 20$	50	**36.2 (0.8)**	39.0 (0.8)	38.9 (0.6)	40.6 (0.6)	38.4 (1.1)	38.5 (1.0)	**34.9 (0.5)**
Waveform	10	**17.4 (0.6)**	**18.0 (0.9)**	18.5 (0.6)	24.9 (1.4)	**18.0 (1.0)**	**16.7 (0.6)**	20.8 (0.8)
$d = 21$	50	16.3 (0.6)	23.7 (1.2)	**14.2 (0.4)**	18.1 (0.8)	**15.4 (0.6)**	15.5 (0.5)	15.3 (0.3)
ijcnn1	10	43.6 (0.6)	**40.3 (1.0)**	49.7 (0.1)	49.2 (0.3)	44.0 (1.0)	45.9 (0.7)	49.3 (0.8)
$d = 22$	50	**34.5 (0.8)**	37.1 (0.9)	**35.5 (0.8)**	**33.4 (1.1)**	49.4 (0.3)	46.2 (0.8)	48.6 (0.4)
g50c	10	11.4 (0.6)	12.5 (0.6)	23.3 (2.3)	39.8 (1.6)	21.9 (1.3)	**6.6 (0.4)**	27.0 (1.4)
$d = 50$	50	12.5 (1.1)	10.1 (0.6)	8.7 (0.4)	22.5 (1.5)	10.6 (0.6)	**7.4 (0.4)**	12.1 (0.5)
covtype	10	**46.2 (0.4)**	**46.0 (0.4)**	**46.0 (0.5)**	47.1 (0.5)	47.9 (0.5)	**46.9 (0.6)**	**46.4 (0.4)**
$d = 54$	50	**41.3 (0.5)**	42.3 (0.5)	**41.0 (0.4)**	**41.5 (0.5)**	46.2 (0.8)	43.6 (0.6)	**40.8 (0.4)**
Spambase	10	27.2 (0.9)	28.1 (1.1)	31.8 (1.4)	39.7 (1.4)	30.9 (1.3)	**23.8 (0.8)**	36.1 (1.5)
$d = 57$	50	23.4 (1.0)	26.6 (1.0)	22.1 (0.7)	28.5 (1.3)	20.9 (0.5)	**19.1 (0.4)**	24.5 (0.9)
Splice	10	**38.3 (0.8)**	**39.3 (0.8)**	43.9 (0.8)	47.9 (0.5)	41.6 (0.7)	42.0 (1.0)	42.4 (0.6)
$d = 60$	50	**30.6 (0.8)**	34.7 (0.9)	**30.9 (0.8)**	38.8 (1.0)	**30.6 (0.9)**	40.9 (0.8)	35.9 (0.7)
phishing	10	**24.2 (1.2)**	**25.8 (1.0)**	**27.3 (1.6)**	37.2 (1.6)	27.6 (1.6)	27.5 (1.4)	31.7 (1.3)
$d = 68$	50	**15.8 (0.6)**	18.3 (0.8)	**15.4 (0.5)**	21.1 (1.3)	**14.7 (0.8)**	17.2 (0.7)	16.7 (0.8)
a9a	10	**31.4 (0.9)**	**31.3 (1.0)**	34.3 (1.2)	41.0 (1.1)	37.3 (1.3)	**33.1 (1.2)**	34.3 (1.2)
$d = 83$	50	27.9 (0.6)	29.9 (0.8)	28.6 (0.7)	33.3 (1.0)	**26.9 (0.7)**	28.9 (0.8)	**26.2 (0.4)**
Coil2	10	**38.7 (0.8)**	**40.1 (0.8)**	42.8 (0.7)	43.9 (0.8)	43.2 (0.8)	**39.1 (0.9)**	44.0 (0.8)
$d = 241$	50	**23.2 (0.6)**	30.5 (0.9)	**23.6 (0.9)**	**22.8 (0.9)**	25.1 (0.9)	**22.6 (0.8)**	25.4 (0.8)
w8a	10	**35.9 (0.9)**	**33.6 (1.0)**	41.6 (1.0)	46.6 (0.8)	39.4 (0.9)	42.1 (0.8)	43.0 (0.8)
$d = 300$	50	**28.1 (0.7)**	**27.6 (0.6)**	**27.0 (0.9)**	38.7 (0.8)	**28.0 (0.9)**	33.7 (0.8)	35.2 (1.0)

Misclassification error rate

04 Semi-supervised classification based on classification from positive and unlabeled data

- 또한, 실제로 class-prior를 추론하여 실험하였을 때도 좋은 결과를 얻었음

Misclassification error rate

Data set	n_U	π	$\hat{\pi}$	PNU	ER	LapSVM	SMIR	WellSVM
Arts	1000	0.50	0.49 (0.01)	**27.4 (1.3)**	**26.6 (0.5)**	**26.1 (0.7)**	40.1 (3.9)	**27.5 (0.5)**
	5000	0.50	0.50 (0.01)	**24.8 (0.6)**	26.1 (0.5)	26.1 (0.4)	30.1 (1.6)	N/A
	10000	0.50	0.52 (0.01)	**25.6 (0.7)**	**25.4 (0.5)**	**25.5 (0.6)**	N/A	N/A
Deserts	1000	0.73	0.67 (0.01)	**13.0 (0.5)**	15.3 (0.6)	16.7 (0.8)	17.2 (0.8)	18.2 (0.7)
	5000	0.73	0.67 (0.01)	**13.4 (0.4)**	**13.3 (0.5)**	16.6 (0.6)	24.4 (0.6)	N/A
	10000	0.73	0.68 (0.01)	**13.3 (0.5)**	**13.7 (0.6)**	16.8 (0.8)	N/A	N/A
Fields	1000	0.65	0.57 (0.01)	**22.4 (1.0)**	26.2 (1.0)	26.6 (1.3)	28.2 (1.1)	26.6 (0.8)
	5000	0.65	0.57 (0.01)	**20.6 (0.5)**	22.6 (0.6)	24.7 (0.8)	29.6 (1.2)	N/A
	10000	0.65	0.57 (0.01)	**21.6 (0.6)**	**22.5 (0.6)**	25.0 (0.9)	N/A	N/A
Stadiums	1000	0.50	0.50 (0.01)	**11.4 (0.4)**	**11.5 (0.5)**	12.5 (0.5)	**17.4 (3.6)**	**11.7 (0.4)**
	5000	0.50	0.50 (0.01)	**11.0 (0.5)**	**10.9 (0.3)**	**11.1 (0.3)**	13.4 (0.7)	N/A
	10000	0.50	0.51 (0.00)	**10.7 (0.3)**	**10.9 (0.3)**	**11.2 (0.2)**	N/A	N/A
Platforms	1000	0.27	0.33 (0.01)	**21.8 (0.5)**	23.9 (0.6)	24.1 (0.5)	30.1 (2.3)	26.2 (0.8)
	5000	0.27	0.34 (0.01)	**23.3 (0.8)**	**24.4 (0.7)**	**24.9 (0.7)**	26.6 (0.3)	N/A
	10000	0.27	0.34 (0.01)	**21.4 (0.5)**	24.3 (0.6)	24.8 (0.5)	N/A	N/A
Temples	1000	0.55	0.51 (0.01)	**43.9 (0.7)**	**43.9 (0.6)**	**43.4 (0.6)**	50.7 (1.6)	**44.3 (0.5)**
	5000	0.55	0.54 (0.01)	**43.4 (0.9)**	**43.0 (0.6)**	**43.1 (1.0)**	**43.6 (0.7)**	N/A
	10000	0.55	0.50 (0.01)	**45.2 (0.8)**	**44.4 (0.8)**	**44.2 (0.7)**	N/A	N/A

05 Class-prior estimation for learning from positive and unlabeled data

- Class-prior π를 추론하는 간단한 방법 중 하나는 다음과 같은 분포를 찾는 것[*]

$$q(x;\theta) = \theta p(x|y=+1) + (1-\theta)p(x|y=-1)$$

- 그리고 $q(x;\theta)$와 $p(x)$를 가깝게 하는 θ를 찾아서 π를 추정

$$\theta = \underset{0\leq\theta\leq1}{\operatorname{argmin}} \int f\left(\frac{q(x;\theta)}{p(x)}\right) p(x)dx$$

  단, f는 convex 함수이며, $f(1) = 0$을 만족

- 보통 f에는 KL-Divergence, Pearson Divergence 등이 사용되어짐

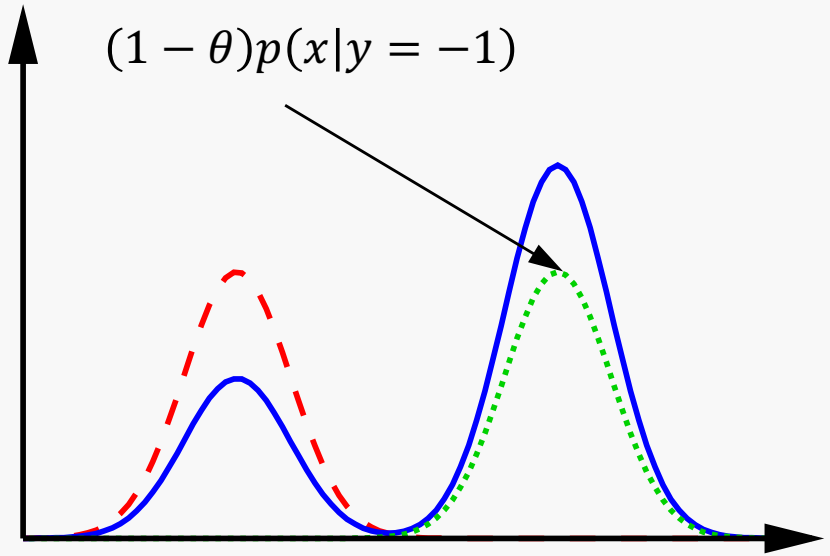
[*] Analysis of Learning from Positive and Unlabeled Data [Du Plessis et al., NIPS, 2014]

05 Class-prior estimation for learning from positive and unlabeled data

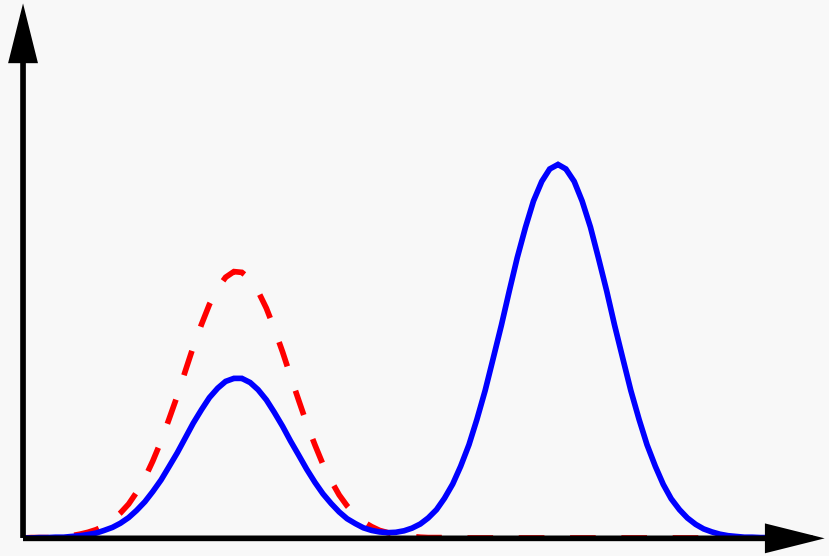
- 하지만, $p(x|y=-1)$의 샘플들은 주어지지 않기 때문에 다음을 탐색

$$q(x;\theta) = \theta p(x|y=+1)$$

$$\theta = \underset{0\leq\theta\leq 1}{\operatorname{argmin}} \int f\left(\frac{q(x;\theta)}{p(x)}\right) p(x)dx$$



$$q(x;\theta) = \theta p(x|y=+1) + (1-\theta)p(x|y=-1)$$

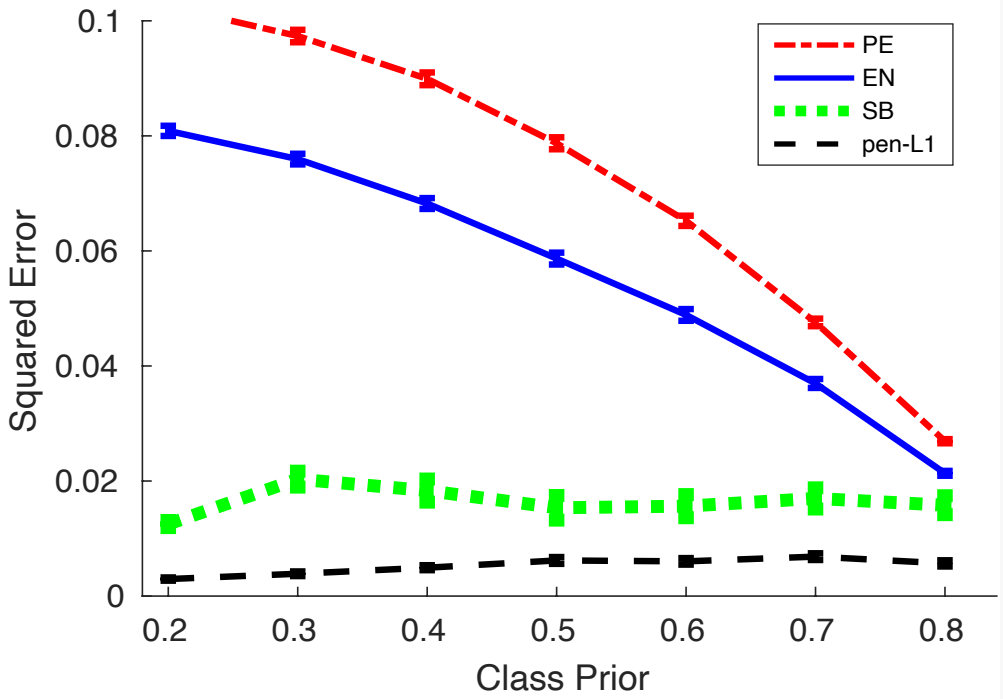


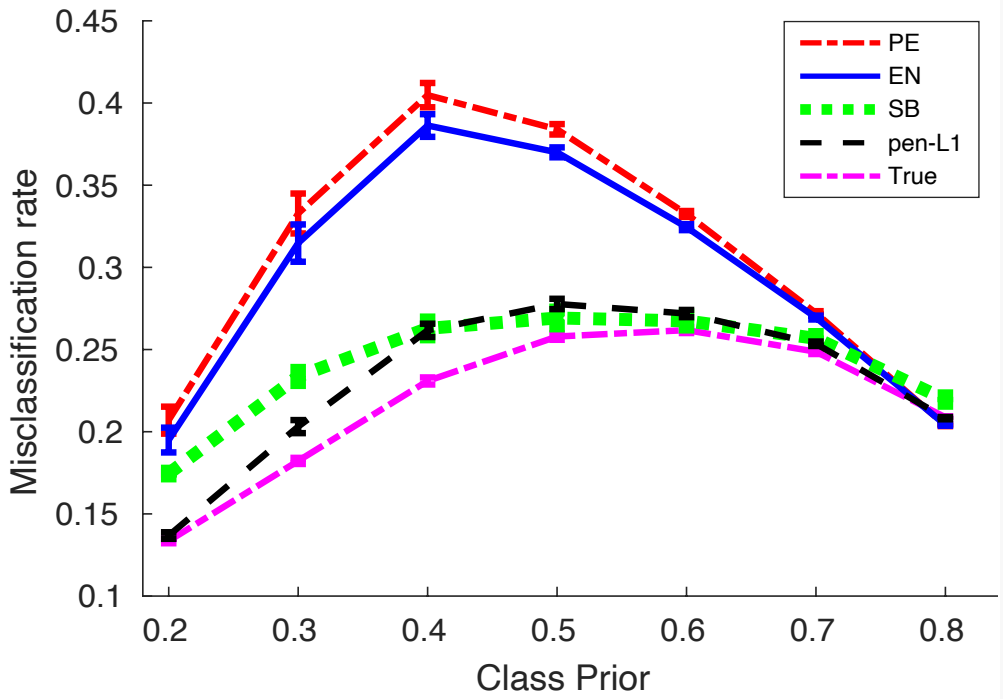
$$q(x;\theta) = \theta p(x|y=+1)$$

05 Class-prior estimation for learning from positive and unlabeled data

- 이로인해 π를 overestimate 하는 경향을 보임

- 논문에서는 L1-distance를 이용하는 것으로 이러한 문제를 해결





Q & A

Q&A