Recent Advances in Machine ... m Weak

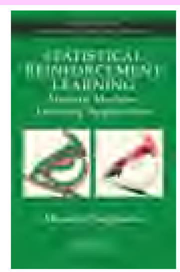
ashi Sugiya

er ... ntell

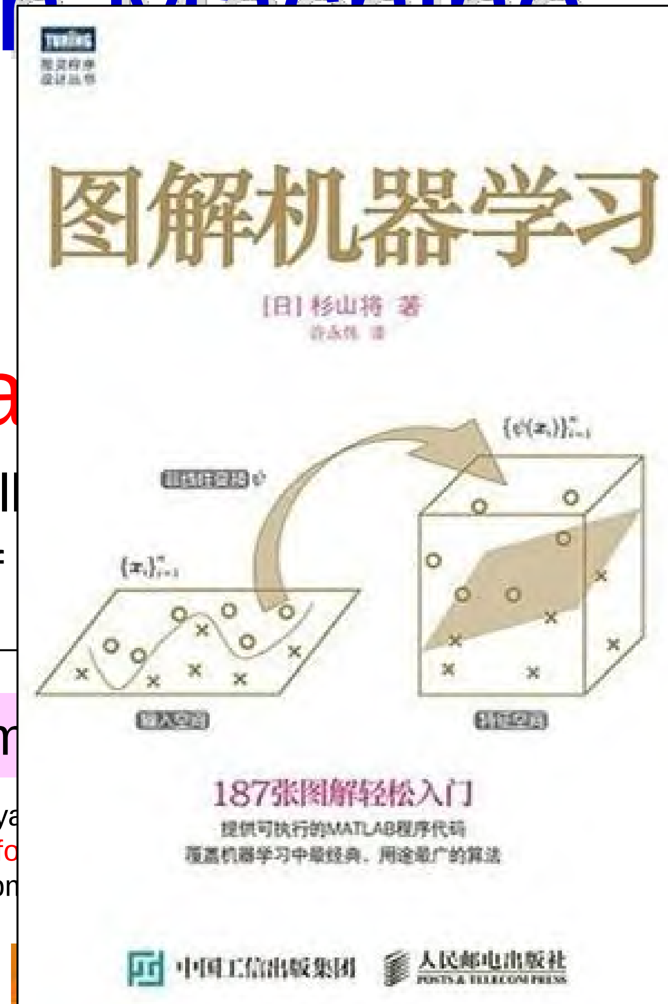
or ... of

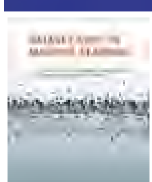
Reinforcem

ervised
ning

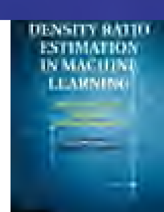


Sugiya
Reinfo
Chapm

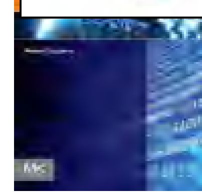




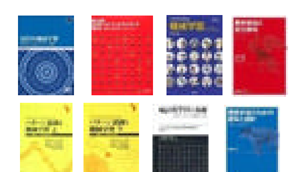
Quiñonero-Sugiyama, Schwaighofer & Lawrence, Dataset Shift in Machine Learning, MIT Press, 2009.



Sugiyama, Suzuki & Kanamori, Density Ratio Estimation in Machine Learning, Cambridge University Press, 2012



Morgan Kaufmann, 2015



In Japanese, (Chinese & Korean)

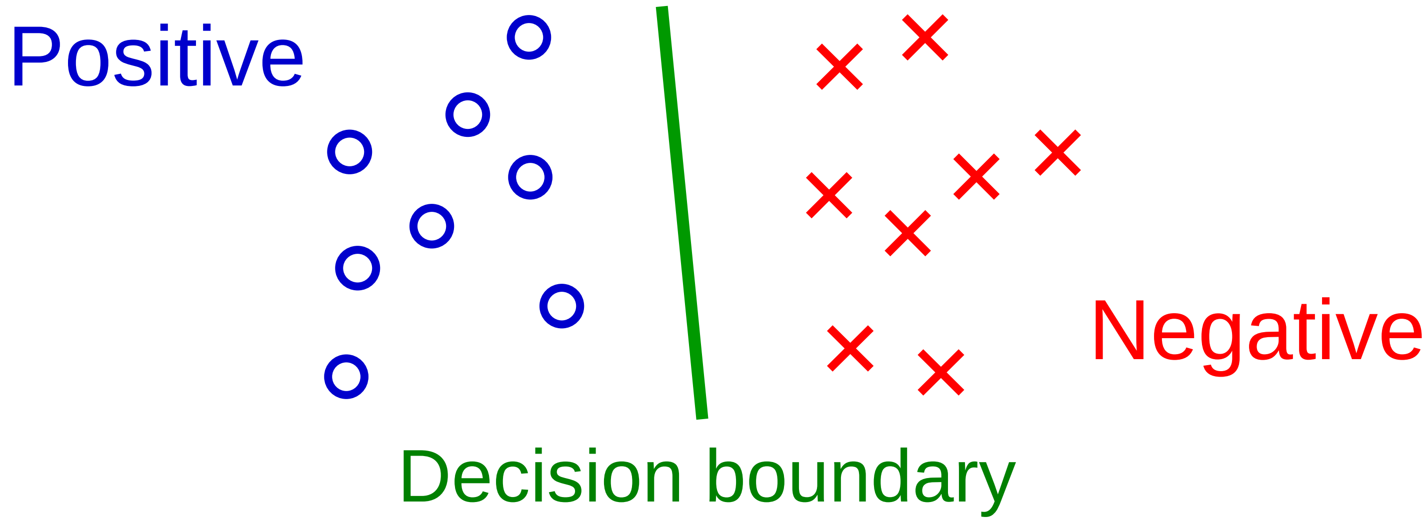
What Is My Talk about?

- Machine learning from big data is successful.
  - Great work on large-scale parallel implementation.
- However, there are various applications where massive labeled data is not available.
  - Medicine, manufacturing, disaster, infrastructure…
- In this talk, I will introduce our recent advances in classification from limited information.

Supervised Classification

- Binary classification from labeled samples:

$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ $\quad \boldsymbol{x} \in \mathbb{R}^d$ $\quad y \in \{+1, -1\}$

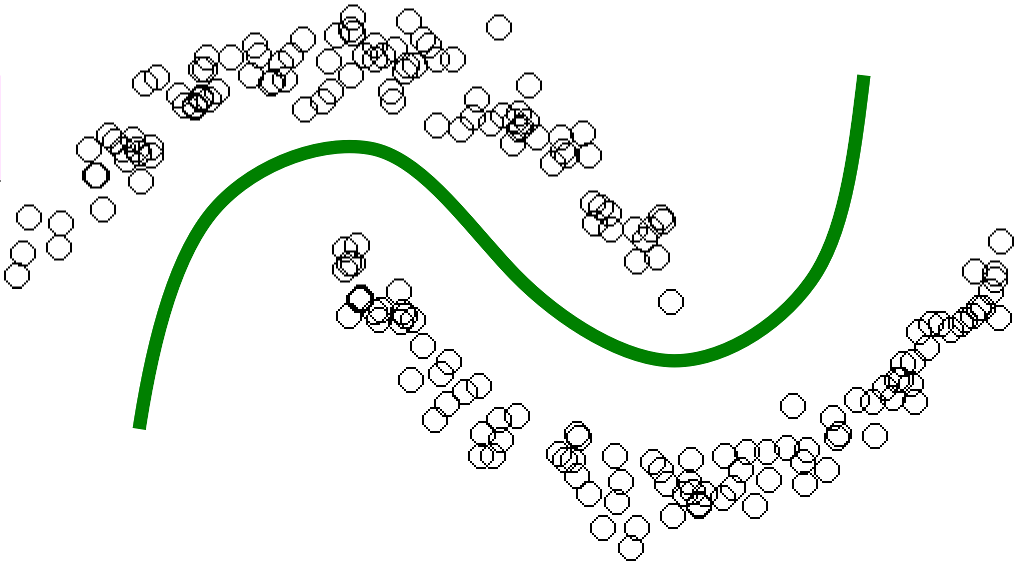


- A large number of labeled samples yield better classification performance.
  - Optimal convergence rate: $\mathcal{O}\left(n^{-1/2}\right)$

Unsupervised Classification

- Since collecting labeled samples is costly, let’s learn a classifier from unlabeled data.

$$\{\boldsymbol{x}'_i\}_{i=1}^{n'}$$

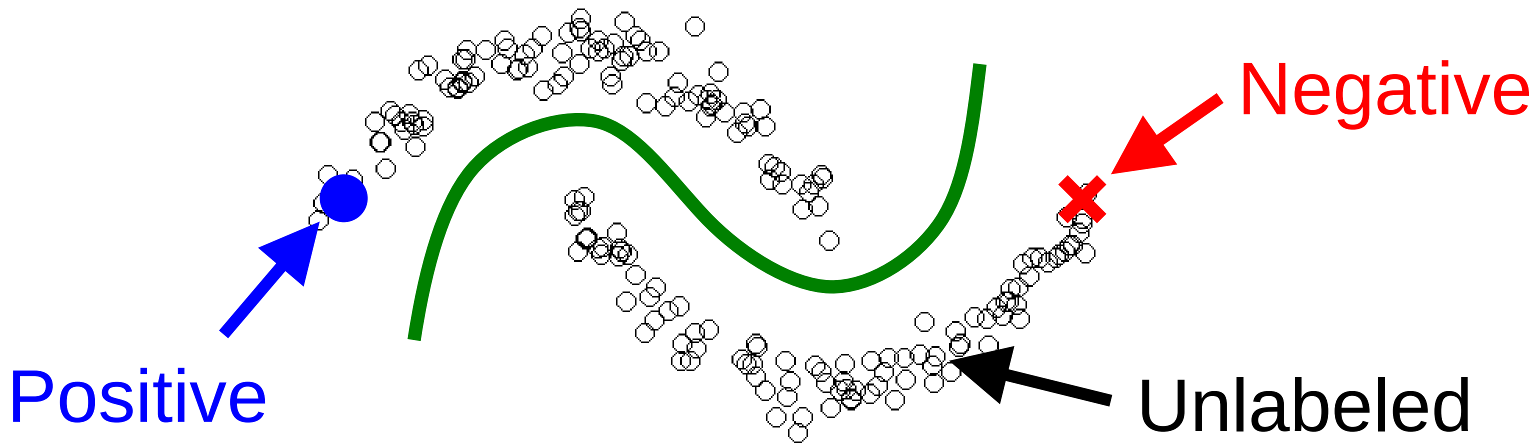


- This is equivalent to clustering.
- To justify this, need the assumption that each cluster corresponds to each class.
  - This is rarely satisfied in practice.

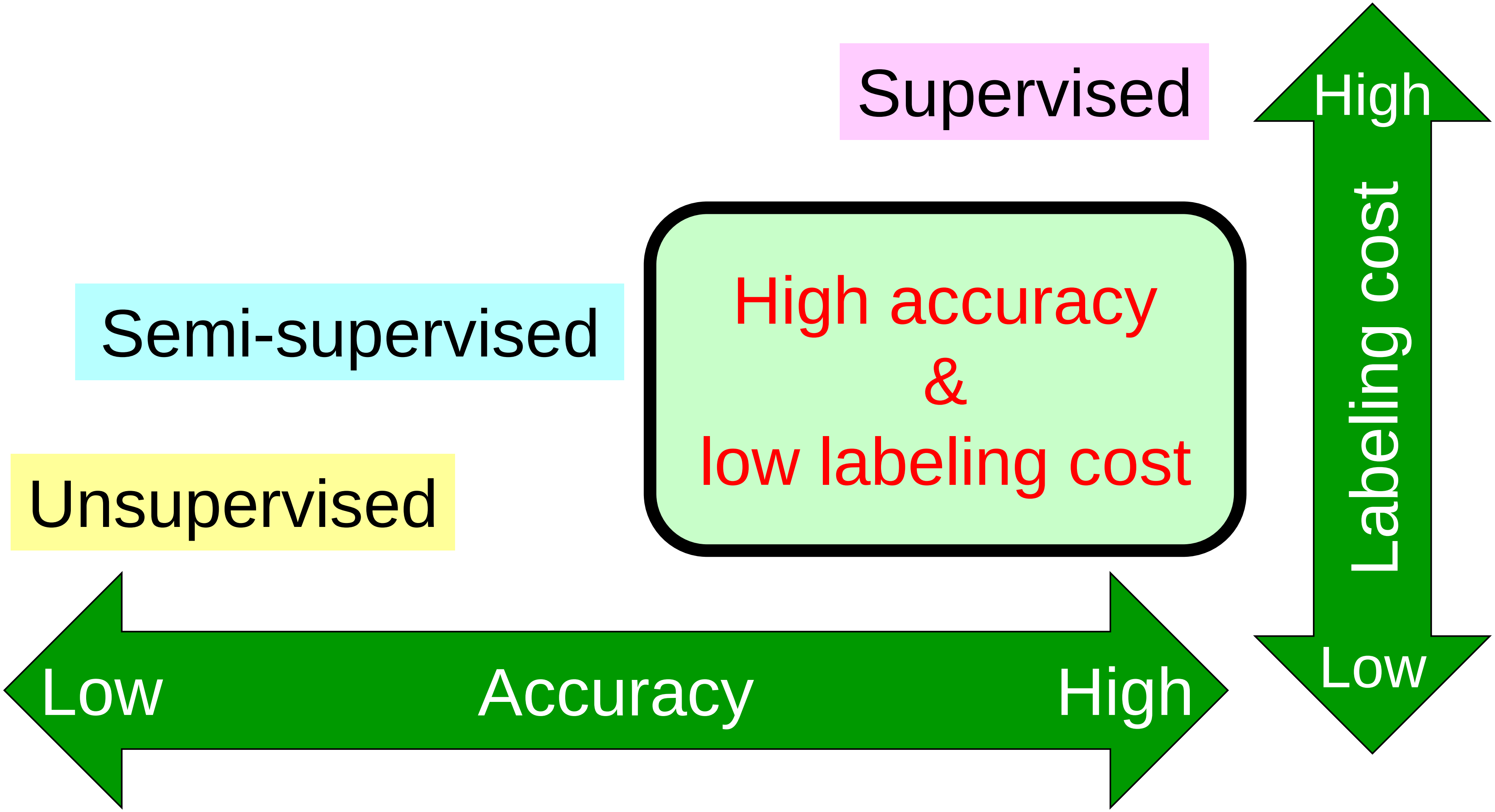
Semi-Supervised Classification

Zhou, Bousquet, Lal, Weston & Schölkopf (NIPS2003) and many

- Use a large number of unlabeled samples and a small number of labeled samples:
- Find a decision boundary along cluster structure induced by unlabeled samples:
  - Sometimes very useful!
  - But same weakness as unsupervised classification.

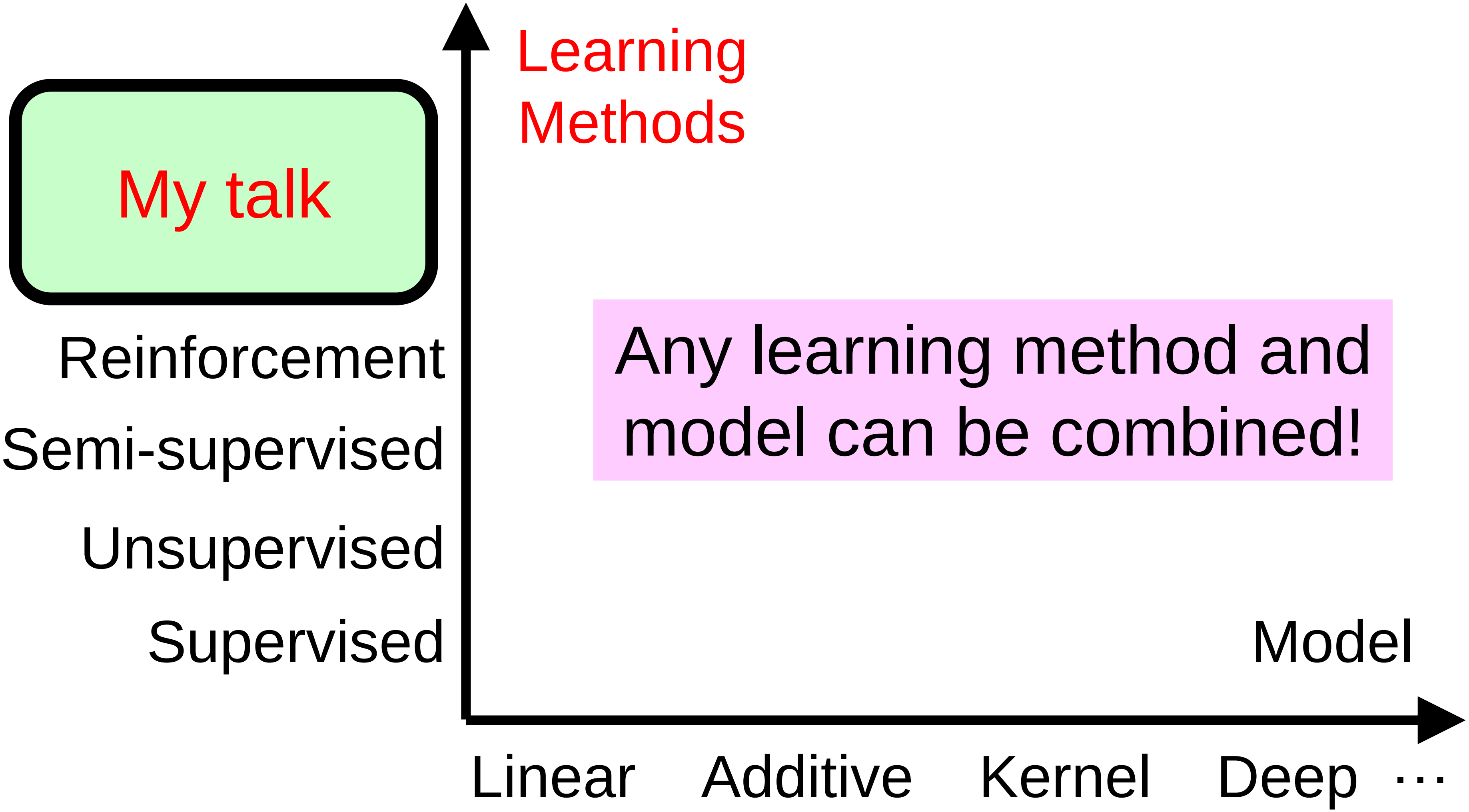


Classification of Classification



- Achieving high classification accuracy with low labeling costs is always a big challenge!

Relation to Deep Learning



Organization

1. Classification of classification
2. Classification from UU data
3. Classification from PU data
4. Classification from PNU data
5. Classification from complementary labels
6. Introduction RIKEN Center for AIP

UU Classification: Setup

du Plessis, Niu & Sugiyama (TAAI2013)

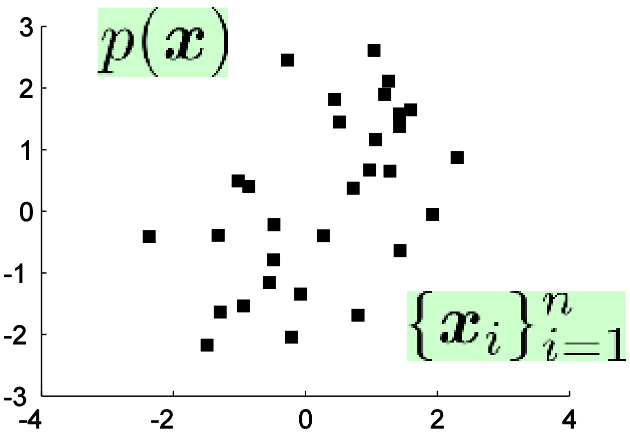
- Given: Two sets of unlabeled data

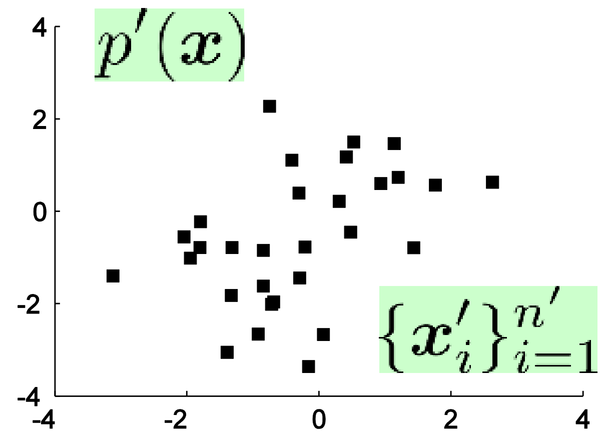
$$\{\boldsymbol{x}_i\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}) \qquad \{\boldsymbol{x}'_i\}_{i=1}^{n'} \overset{\text{i.i.d.}}{\sim} p'(\boldsymbol{x})$$

- Assumption: Only class-priors are different

$$p(y) \neq p'(y) \qquad p(\boldsymbol{x}|y) = p'(\boldsymbol{x}|y)$$

- Goal: Obtain a classifier





Optimal UU Classifier

du Plessis, Niu & Sugiyama (TAAI2013)

- Sign of the difference of class-posteriors:

$$g(\boldsymbol{x}) = \mathrm{sign}[p(y=+1|\boldsymbol{x}) - p(y=-1|\boldsymbol{x})]$$

- Under equal test class-prior $q(y=+1)=1/2$,

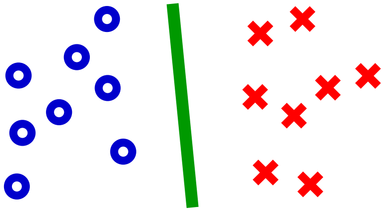
$$g(\boldsymbol{x}) = C\mathrm{sign}[p(\boldsymbol{x}) - p'(\boldsymbol{x})]$$

$$C = \mathrm{sign}[p(y=+1) - p'(y=+1)]$$

- Sign of C is unknown, but just knowing

$$\mathrm{sign}[p(\boldsymbol{x}) - p'(\boldsymbol{x})]$$

still allows optimal separation!



UU Classifier Training

$$\mathrm{sign}[p(\boldsymbol{x}) - p'(\boldsymbol{x})]$$

Difference of kernel density estimators:

- Estimate $p(\boldsymbol{x}), p'(\boldsymbol{x})$ from $\{\boldsymbol{x}_i\}_{i=1}^{n}, \{\boldsymbol{x}'_i\}_{i=1}^{n'}$, separately.
- Simple but systematic under-estimation of $p(\boldsymbol{x}) - p'(\boldsymbol{x})$.

Anderson, Hall & Titterington (J. Multivariate Analysis 1994)

Direct estimation of density-difference:

Kim & Scott (IEEE-TPAMI2010)
Sugiyama, Suzuki, Kanamori, du Plessis, Liu & Takeuchi (NIPS2012, NeCo2013)

- Fit model $f(\boldsymbol{x})$ to $p(\boldsymbol{x}) - p'(\boldsymbol{x})$ directly without estimating $p(\boldsymbol{x}), p'(\boldsymbol{x})$.
- Linear least-squares formulation yields global analytic solution!

$$\min_f \int \Big(f(\boldsymbol{x}) - \big\{ p(\boldsymbol{x}) - p'(\boldsymbol{x}) \big\} \Big)^2 \mathrm{d}\boldsymbol{x}$$

Direct estimation of sign of density-difference:

du Plessis, Niu & Sugiyama (TAAI2013)

- Most direct approach (following Vapnik's principle!).
- Non-convex optimization is involved (use, e.g., CCCP).

Experiments

Misclassification error rate: average (std)

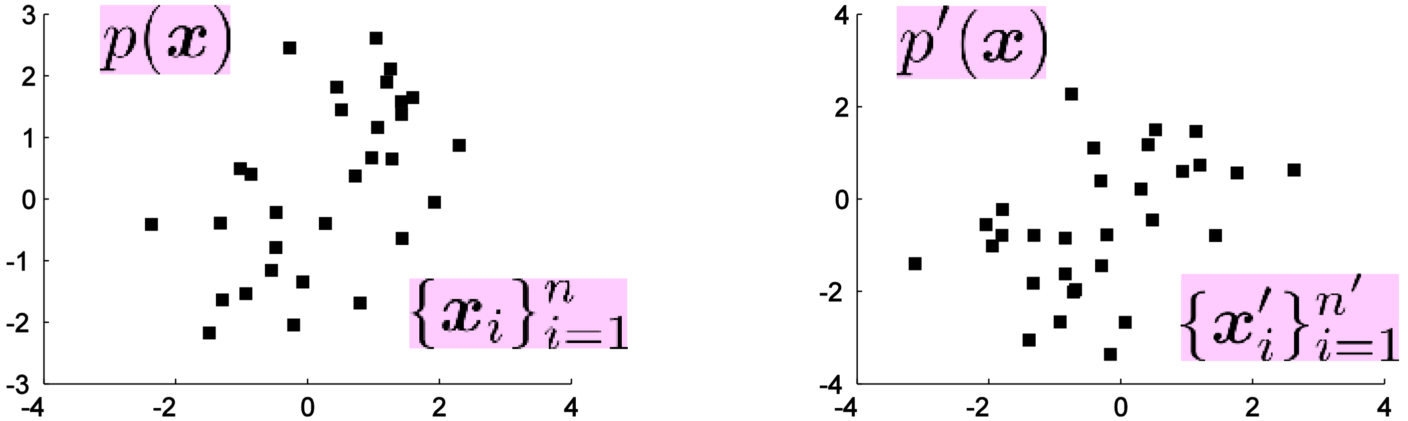
5% t-test

	UU classification			Clustering		
	$\text{sign}[p(\boldsymbol{x}) - p'(\boldsymbol{x})]$	$p(\boldsymbol{x}) - p'(\boldsymbol{x})$	$p(\boldsymbol{x}), p'(\boldsymbol{x})$	k-means	Spectral Ng et al. (NIPS2001)	Infomax Sugiyama et al. (ICML2011)
Dataset	**DSDD**	**LSDD**	**KDE**	**KM**	**SC**	**SMIC**
australian	**.244(.116)**	.259(.088)	.355(.104)	.265(.080)	.376(.065)	.308 (.107)
banana	**.338(.094)**	**.339(.100)**	.365(.067)	.433(.049)	.427(.069)	.424 (.070)
diabetes	**.340(.075)**	.361(.124)	.345(.034)	.373(.063)	.380(.048)	.371 (.114)
german	.375(.042)	.380(.093)	**.354(.057)**	.437(.024)	.445(.057)	.438 (.041)
heart	.270(.133)	**.247(.084)**	.354(.052)	.264(.059)	.315(.081)	.327 (.089)
image	**.331(.078)**	.350(.067)	.350(.039)	.384(.031)	.354(.049)	.382 (.050)
ionosphere	**.291(.099)**	.356(.066)	.345(.048)	.330(.070)	.322(.058)	.314 (.107)
saheart	.378(.093)	**.353(.057)**	.363(.066)	.419(.082)	.395(.022)	.385 (.040)
thyroid	**.227(.098)**	.251(.087)	.302(.022)	.326(.061)	.329(.047)	.307 (.076)
twonorm	.164(.188)	.153(.121)	.352(.096)	**.036(.053)**	.042(.122)	.049 (.120)

$n = n' = 40$ $\quad p(y=+1) = 0.35$ $\quad p'(y=+1) = 0.65$

- UU classification with direct estimation of (sign of) density difference works well !

UU Classification: Summary



- Given two unlabeled datasets with different class-priors, we estimate the sign of difference of class-posteriors: $\mathrm{sign}[p(\boldsymbol{x}) - p'(\boldsymbol{x})]$
- Same convergence rate as fully supervised case can be achieved! $\mathcal{O}(n^{-1/2})]$
- Unlike classification from label proportions, we do not have to know class priors.

Quadrianto, Smola, Caetano & Le (JMLR2009)

Organization

1. Classification of classification
2. Classification from UU data
3. Classification from PU data
4. Classification from PNU data
5. Classification from complementary labels
6. Introduction RIKEN Center for AIP

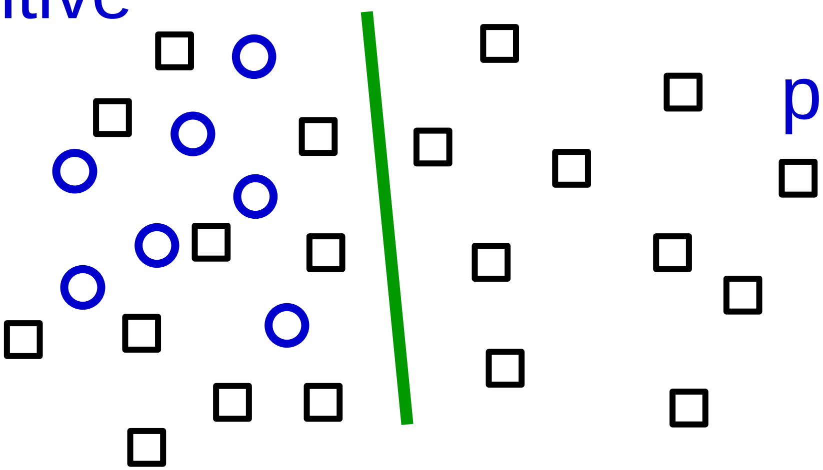
PU Classification: Setup

- **Given:** Positive and unlabeled samples

$$\{(\boldsymbol{x}_i, y_i = +1)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}|y = +1)$$

$$\{\boldsymbol{x}'_i\}_{i=1}^{n'} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x})$$

- **Goal:** Obtain an (ordinary) PN classifier



Examples:

- Click vs. non-click
- Friend vs. non-friend

Classification Risk

- Risk of classifier f :

$$R(f) = \mathbb{E}_{p(\boldsymbol{x},y)}\Big[\ell\Big(yf(\boldsymbol{x})\Big)\Big]$$

$\mathbb{E}$: Expectation

ℓ : Loss

$$= \pi\mathbb{E}_{p(\boldsymbol{x}|y=+1)}\Big[\ell\Big(f(\boldsymbol{x})\Big)\Big] + (1-\pi)\mathbb{E}_{p(\boldsymbol{x}|y=-1)}\Big[\ell\Big(-f(\boldsymbol{x})\Big)\Big]$$

Risk for P data — Risk for N data

$\pi = p(y=+1)$: Class-prior probability (assumed known; can be estimated)

Scott & Blanchard (AISTATS2009)
Blanchard, Lee & Scott (JMLR2010)
du Plessis, Niu & Sugiyama (IEICE2014, MLJ2017)

- Since we do not have N data in the PU setting, the risk cannot be directly estimated.

PU Unbiased Risk Estimation

Natarajan, Dhillon, Ravikumar & Tewari (NIPS2013)
du Plessis, Niu & Sugiyama (ICML2015)

$$R(f) = \underbrace{\pi \mathbb{E}_{p(\boldsymbol{x}|y=+1)}\Big[\ell\Big(f(\boldsymbol{x})\Big)\Big]}_{\text{Risk for P data}} + \underbrace{(1-\pi)\mathbb{E}_{p(\boldsymbol{x}|y=-1)}\Big[\ell\Big(-f(\boldsymbol{x})\Big)\Big]}_{\text{Risk for N data}}$$

- U-density is a mixture of P- and N-densities:

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

- Eliminating the N-density yields

$$R(f) = \pi \mathbb{E}_{p(\boldsymbol{x}|y=+1)}\Big[\ell\Big(f(\boldsymbol{x})\Big)\Big] + \mathbb{E}_{p(\boldsymbol{x})-\pi p(\boldsymbol{x}|y=+1)}\Big[\ell\Big(-f(\boldsymbol{x})\Big)\Big]$$

  - Unbiased risk estimation is possible only from PU data!

Theoretical Analysis

Niu, du Plessis, Sakai, Ma & Sugiyama (NIPS2016)

- Estimation error bounds:

$$R(\widehat{f}_{\mathrm{PU}}) - R(f^*) \leq C(\delta)\left(2\pi/\sqrt{n_{\mathrm{P}}} + 1/\sqrt{n_{\mathrm{U}}}\right)$$

$$R(\widehat{f}_{\mathrm{PN}}) - R(f^*) \leq C(\delta)\left(\pi/\sqrt{n_{\mathrm{P}}} + (1-\pi)/\sqrt{n_{\mathrm{N}}}\right)$$

with probability $1-\delta$

$$R(f) = \mathbb{E}_{p(\boldsymbol{x},y)}\left[\ell\Big(yf(\boldsymbol{x})\Big)\right]$$

$$f^* = \mathrm{argmin}_f\, R(f)$$

$n_{\mathrm{P}}, n_{\mathrm{N}}, n_{\mathrm{U}}$: # of positive, negative and unlabeled samples

  - PU (and PN) achieve optimal convergence rate.

- Comparison: PU bound is smaller than PN if

$$\pi/\sqrt{n_{\mathrm{P}}} + 1/\sqrt{n_{\mathrm{U}}} < (1-\pi)/\sqrt{n_{\mathrm{N}}}$$

  - PU can be better than PN provided a large number of PU data!

Further Correction

Kiryo, Niu, du Plessis & Sugiyama (arXiv2017)

- PN formulation:

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

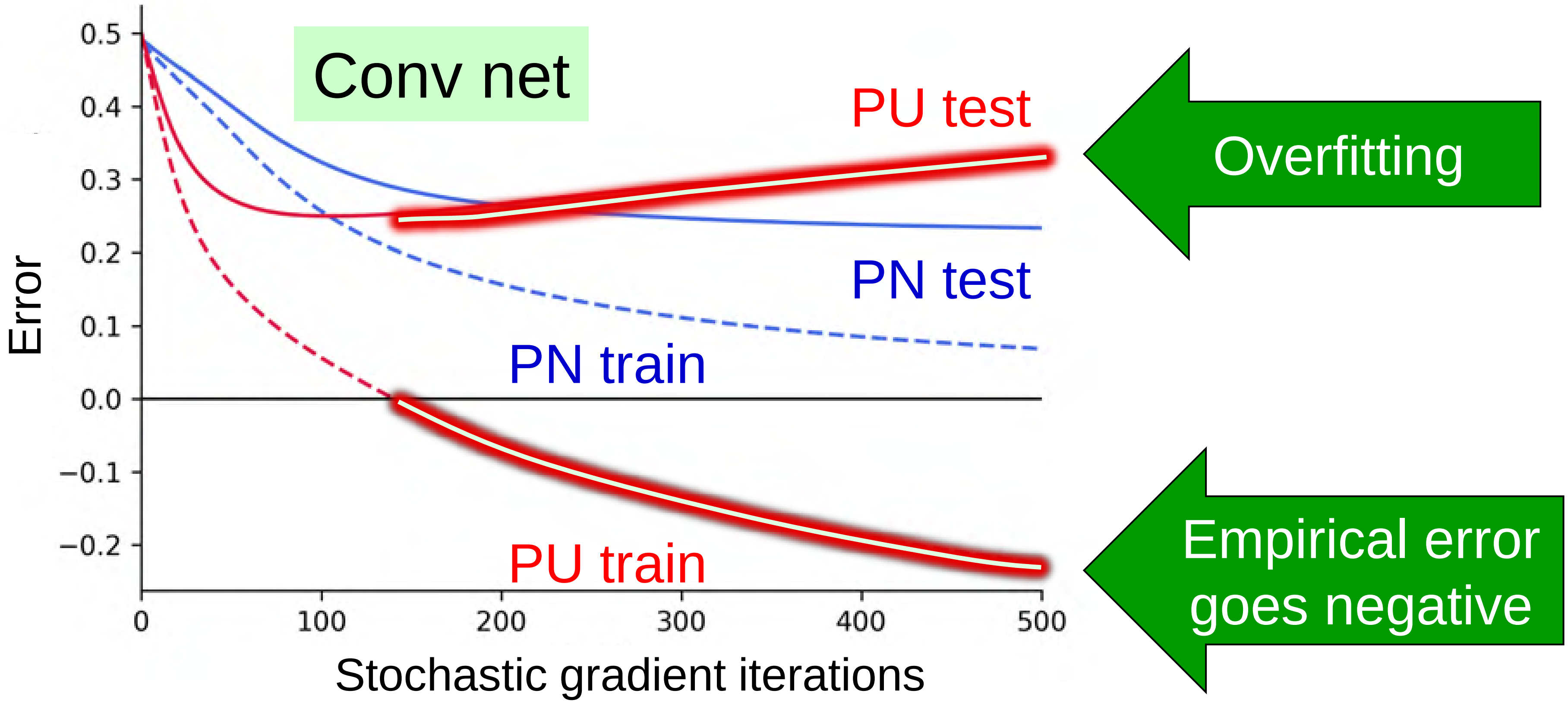
Risk for P data | Risk for N data

- PU formulation: $p(\boldsymbol{x}) = \pi p(\boldsymbol{x}|y=+1) + (1-\pi)p(\boldsymbol{x}|y=-1)$

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

- Risk for N data is non-negative by definition, but its approximation from PU samples can be negative due to “difference of approximations”.
  - In particular, for flexible models such as deep nets.

Non-Negative PU Classification



- We constrain the sample approximation term to be non-negative through back-prop training:

$$\widehat{R}(f) = \pi \widehat{\mathbb{E}}_{p(\boldsymbol{x}|y=+1)}\Big[\ell\Big(f(\boldsymbol{x})\Big)\Big] + \max\Big\{0,\ \widehat{\mathbb{E}}_{p(\boldsymbol{x})-\pi p(\boldsymbol{x}|y=+1)}\Big[\ell\Big(-f(\boldsymbol{x})\Big)\Big]\Big\}$$

  - Now the risk estimator is biased. Is it really good?

Theoretical Analysis

$$\widehat{R}(f) = \pi \widehat{\mathbb{E}}_{p(\boldsymbol{x}|y=+1)}\Big[\ell\Big(f(\boldsymbol{x})\Big)\Big] + \max\Big\{0,\ \widehat{\mathbb{E}}_{p(\boldsymbol{x})-\pi p(\boldsymbol{x}|y=+1)}\Big[\ell\Big(-f(\boldsymbol{x})\Big)\Big]\Big\}$$

- $\widehat{R}(f)$ is still consistent and its bias decreases exponentially: $\mathcal{O}(\exp(-1/n_{\mathrm{P}} + 1/n_{\mathrm{U}}))$

  $n_{\mathrm{P}}, n_{\mathrm{U}}$: # of positive and unlabeled samples

  - In practice, we can ignore the bias of $\widehat{R}(f)$!
- Mean-squared error of $\widehat{R}(f)$ is not more than the original one.
  - In practice, $\widehat{R}(f)$ is more reliable!
- Risk of $\mathrm{argmin}_f \widehat{R}(f)$ for linear models converges with optimal parametric order: $\mathcal{O}(1/\sqrt{n_{\mathrm{P}}} + 1/\sqrt{n_{\mathrm{U}}})$
  - Learned function is optimal.

Experiments

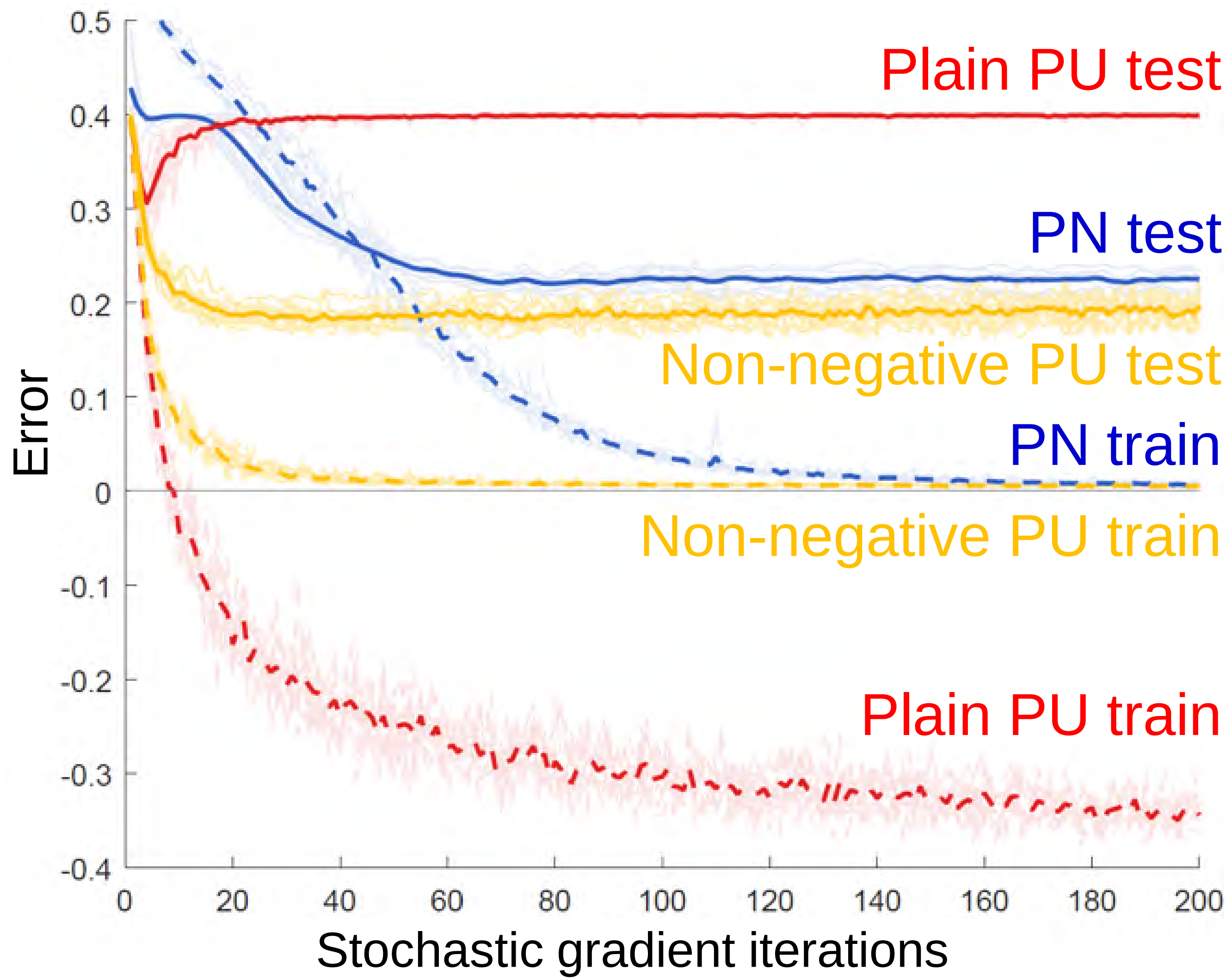
- With a large number of unlabeled data, non-negative PU can even outperform PN!

- Binary CIFAR-10: Positive (airplane, automobile, ship, truck) Negative (bird, cat, deer, dog, frog, horse)
- 13-layer CNN with ReLU

$n_{\mathrm{P}} = 1000$

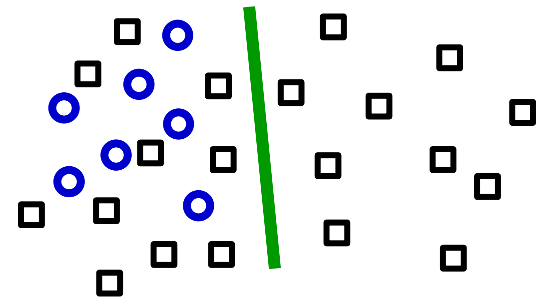
$n_{\mathrm{U}} = 50000$

$\pi = 0.4$



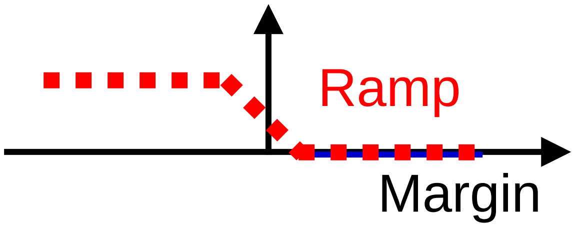
PU Classification: Summary

- Just separating P and U is biased.



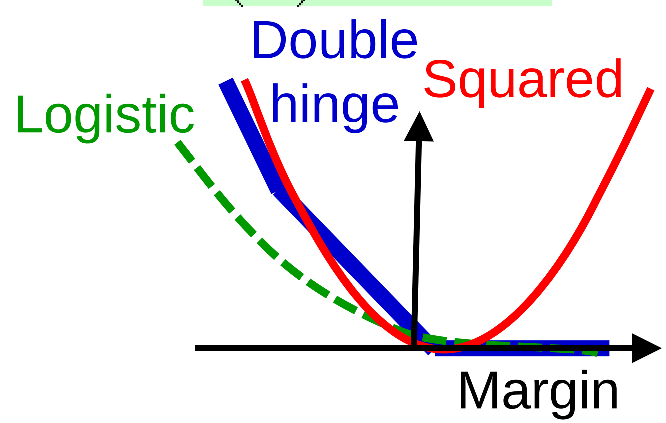
- To be unbiased, use composite loss $\widetilde{\ell}(m) = \ell(m) - \ell(-m)$ for P data.
  Natarajan, Dhillon, Ravikumar & Tewari (NIPS2013)
  - Optimal convergence rate achieved.
    Niu, du Plessis, Sakai, Ma & Sugiyama (NIPS2016)

$\ell(m) + \ell(-m) = \mathrm{Const.}$



- If $\ell(m) + \ell(-m) = \mathrm{Const.}$, the same loss for P and U data.
  du Plessis, Niu & Sugiyama (NIPS2014)

$\widetilde{\ell}(m) = am + b$



- If $\widetilde{\ell}(m) = am + b$, optimization becomes convex.
  du Plessis, Niu & Sugiyama (ICML2015)
- For deep nets, roundup the empirical false negative error.
  Kiryo, Niu, du Plessis & Sugiyama (arXiv2017)

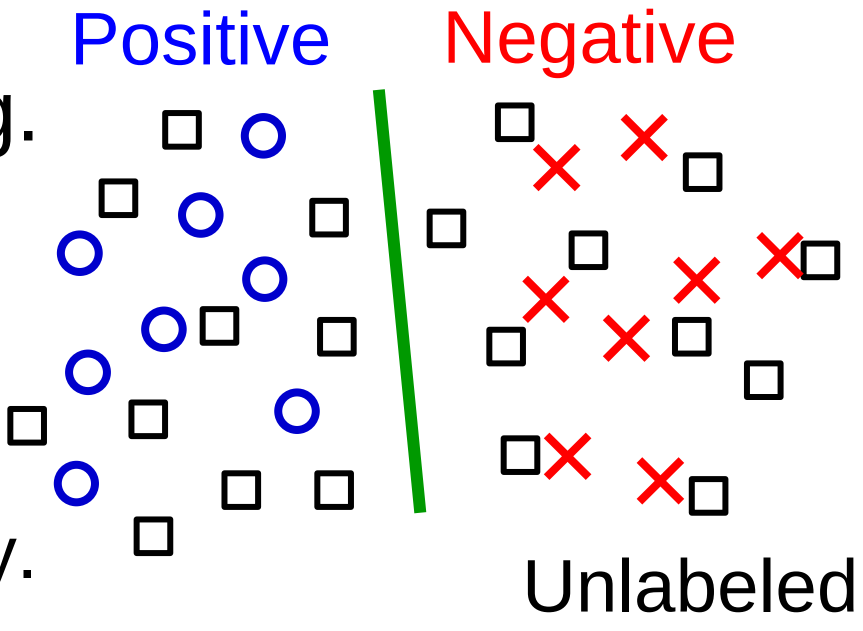
Organization

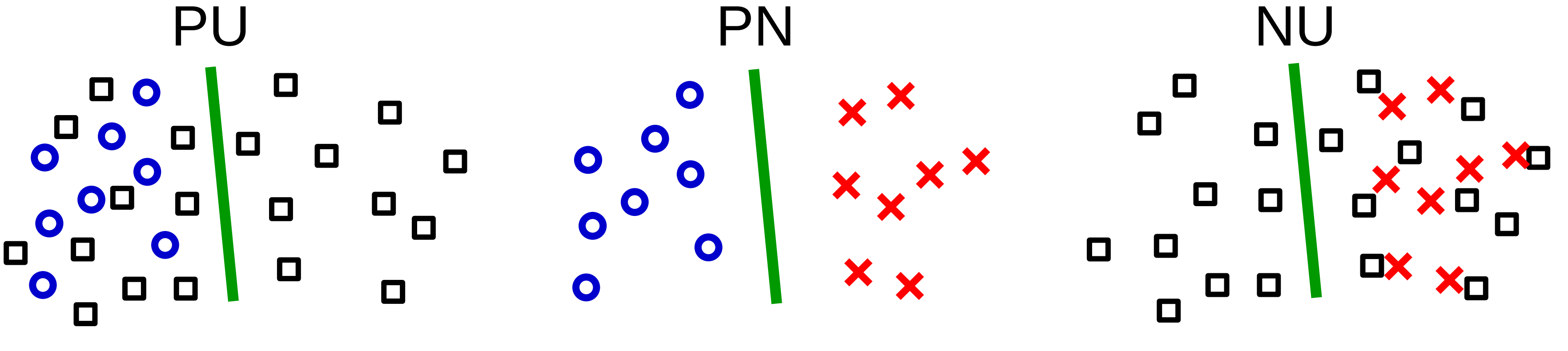
1. Classification of classification
2. Classification from UU data
3. Classification from PU data
4. Classification from PNU data
5. Classification from complementary labels
6. Introduction RIKEN Center for AIP

PNU Classification

Sakai, du Plessis, Niu & Sugiyama (ICML2017)

- PNU classification is semi-supervised learning.
- Let's decompose this into PU, PN, and NU classification:
  - Each can be solved easily.
  - Combine two of them!

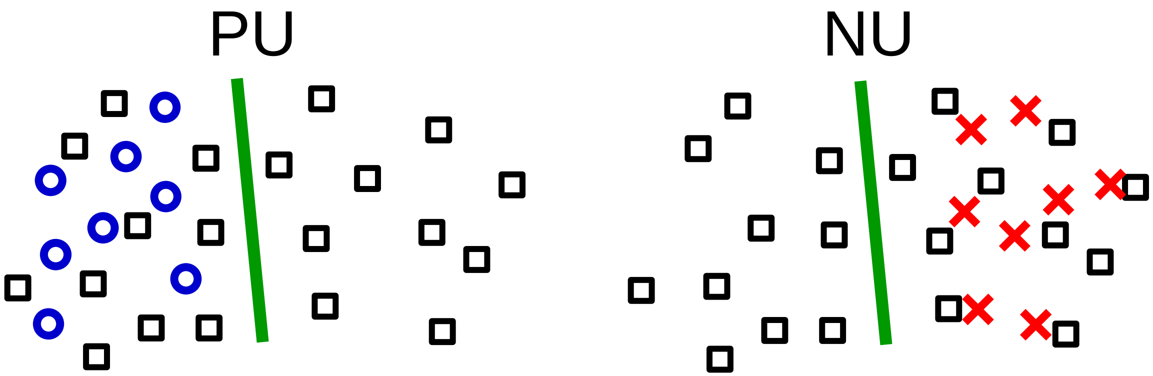




PU+NU Classification

- Natural choice: Combine PU & NU (symmetric).

$$R_{\mathrm{PU+NU}}(f) = (1-\gamma)R_{\mathrm{PU}}(f) + \gamma R_{\mathrm{NU}}(f) \qquad 0 \leq \gamma \leq 1$$



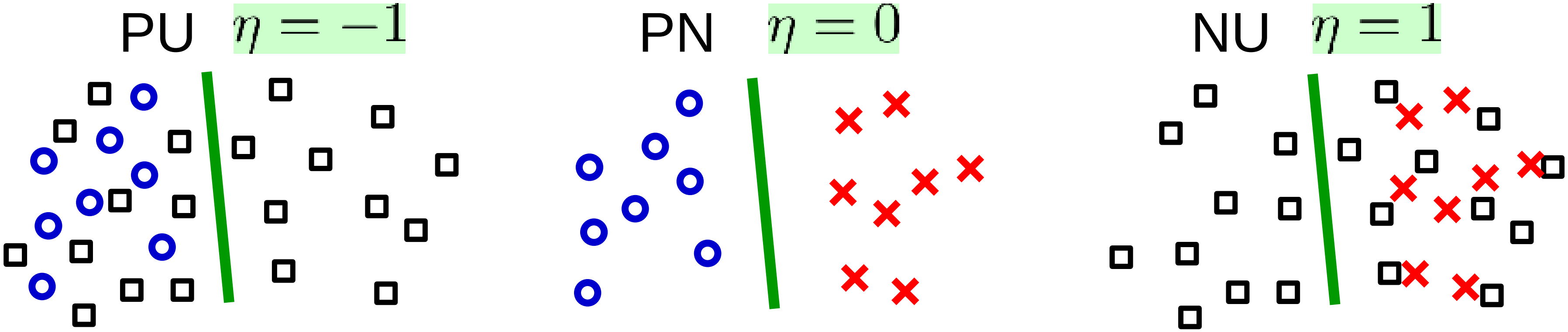
- Theoretical risk analysis:

Niu, du Plessis, Sakai, Ma & Sugiyama (NIPS2016)

  - When PU<NU, PU<PN<NU or PN<PU<NU.
  - When NU<PU, NU<PN<PU or PN<NU<PU.
- PU+NU is not the best possible combination.
- PU+PN & NU+PN are the best combinations.

PN+PU & PN+NU Classification

Proposed method: Combine best methods:



$$R^{\eta}_{\mathrm{PNU}}(f) = \begin{cases} R^{\eta}_{\mathrm{PN+PU}}(f) & (\eta \geq 0) \\ R^{-\eta}_{\mathrm{PN+NU}}(f) & (\eta < 0) \end{cases}$$

$-1 \leq \eta \leq 1$

- PN+PU classification:

$$R^{\gamma}_{\mathrm{PN+PU}}(f) = (1-\gamma)R_{\mathrm{PN}}(f) + \gamma R_{\mathrm{PU}}(f)$$

$0 \leq \gamma \leq 1$

- PN+NU classification:

$$R^{\gamma}_{\mathrm{PN+NU}}(f) = (1-\gamma)R_{\mathrm{PN}}(f) + \gamma R_{\mathrm{NU}}(f)$$

$0 \leq \gamma \leq 1$

Theoretical Analysis

- Generalization error bounds:

$$R_{0/1}(f) \leq 2\widehat{R}^{\gamma}_{\mathrm{PN+PU}}(f) + \mathcal{O}(1/\sqrt{n_{\mathrm{P}}} + 1/\sqrt{n_{\mathrm{N}}} + 1/\sqrt{n_{\mathrm{U}}})$$

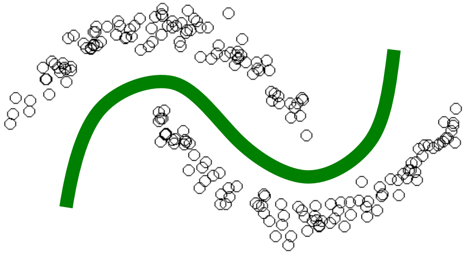
$$R_{0/1}(f) \leq 2\widehat{R}^{\gamma}_{\mathrm{PN+NU}}(f) + \mathcal{O}(1/\sqrt{n_{\mathrm{P}}} + 1/\sqrt{n_{\mathrm{N}}} + 1/\sqrt{n_{\mathrm{U}}})$$

$$R_{0/1}(f) = \mathbb{E}_{p(\boldsymbol{x},y)}\left[\ell_{0/1}\Big(yf(\boldsymbol{x})\Big)\right]$$

$\widehat{R}$: Empirical version of R

$n_{\mathrm{P}}, n_{\mathrm{N}}, n_{\mathrm{U}}$: # of positive, negative and unlabeled samples

  - Unlabeled data always helps without cluster assumptions!



- We use unlabeled data for loss evaluation, not for regularization (as manifold smoothing).
  - Label information is extracted from unlabeled data!

Experiments

Misclassification error rate: average (std)

5% t-test

Dataset	n_u	π	$\widehat{\pi}$	Proposed	EntReg (Grandvalet & Bengio, NIPS2004)	LapSVM (Belkin et al., JMLR2006)	SMIR (Niu et al., ICML2013)	WellSVM (Li et al., JMLR2013)
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

- Proposed PN+PU & PN+NU works well!

Organization

1. Classification of classification
2. Classification from UU data
3. Classification from PU data
4. Classification from PNU data
5. Classification from complementary labels
6. Introduction RIKEN Center for AIP

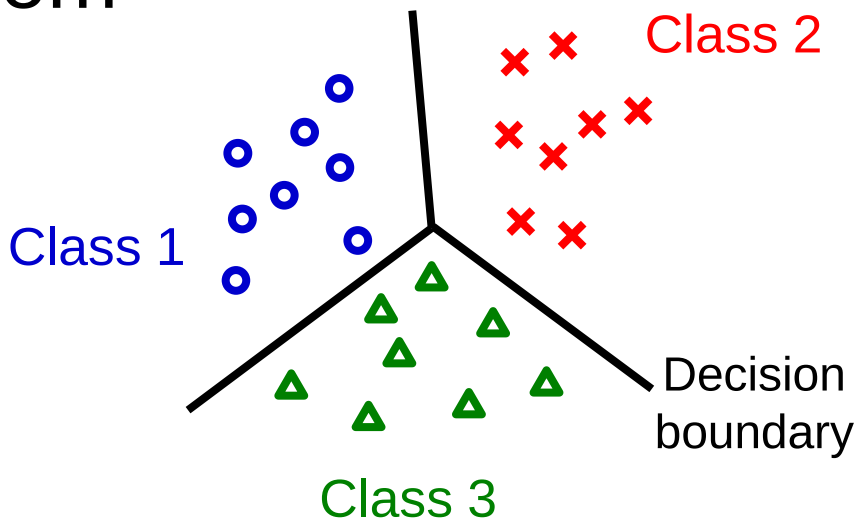
Classification from Complementary Labels

Ishida, Niu & Sugiyama (arXiv2017)

- **Complementary label**: $\bar{y} \in \{1, 2, \ldots, c\}$
  - Pattern $\boldsymbol{x}$ does not belong to class $\bar{y}$.
  - Choosing a complementary class is less laborious than choosing an ordinary class label for large c.
- **Goal**: Learn a classifier from complementary labels.

$$\{(\boldsymbol{x}_i, \bar{y}_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} \bar{p}(\boldsymbol{x}, \bar{y})$$

$$\bar{p}(\boldsymbol{x}, \bar{y}) = \frac{1}{c-1} \sum_{y \neq \bar{y}} p(\boldsymbol{x}, y)$$



Possible Approaches

$$\{(\boldsymbol{x}_i, \bar{y}_i)\}_{i=1}^{n} \stackrel{\text{i.i.d.}}{\sim} \bar{p}(\boldsymbol{x}, \bar{y})$$

- Approach 1: Classification from partial labels

  Cour, Sapp & Taskar (JMLR2011)

  - Multiple candidate classes are provided for each $\boldsymbol{x}_i$.
  - Complementary labels are the extreme case of partial labels given to all $c-1$ classes other than $\bar{y}_i$.

- Approach 2: Multi-label classification
  - Each $\boldsymbol{x}_i$ can belong to multiple classes.
  - Negative label for $\bar{y}_i$ and positives for the rest.

- We want a more direct approach!

Unbiased Risk Estimation with Complimentary Labels

Ishida, Niu & Sugiyama (arXiv2017)

- c-class classifier: $f(\boldsymbol{x}) = \underset{y \in \{1,\ldots,c\}}{\operatorname{argmax}} \; g_y(\boldsymbol{x})$

  $g_y(\boldsymbol{x})$: 1-vs-rest classifier for y

- Classification risk:

$$R(f) = \mathbb{E}_{p(\boldsymbol{x},y)}\left[\sum_{y \neq y'} \ell\Big(g_y(\boldsymbol{x}) - g_{y'}(\boldsymbol{x})\Big)\right]$$

  $\mathbb{E}$: Expectation

- For pairwise symmetric loss, risk is

$$R(f) = \frac{1}{c-1}\mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}\left[\sum_{y \neq \bar{y}} \ell\Big(g_y(\boldsymbol{x}) - g_{\bar{y}}(\boldsymbol{x})\Big)\right] - \text{Const.}$$

  $\ell(m) = 1/\Big(1 + e^m\Big)$

  - Unbiased risk estimation is possible from $\{(\boldsymbol{x}_i, \bar{y}_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \bar{p}(\boldsymbol{x}, \bar{y})$!

Theoretical Analysis

- Estimation error:

$$R\left(f^*\right) - R\left(\widehat{f}\right) = \mathcal{O}_p\left(n^{-1/2}\right)$$

$$R(f) = \mathbb{E}_{p(\boldsymbol{x},y)}\left[\mathcal{L}\left(f(\boldsymbol{x}), y\right)\right]$$

$$f^* = \operatorname*{argmin}_f \mathbb{E}_{p(\boldsymbol{x},y)}\mathcal{L}\left(f(\boldsymbol{x}), y\right) \qquad \widehat{f} = \operatorname*{argmin}_f \sum_{i=1}^{n} \bar{\mathcal{L}}\left(f(\boldsymbol{x}_i), \bar{y}_i\right)$$

- Optimal parametric convergence rate!

Experiments

Correct classification rate: average (std)

5% t-test

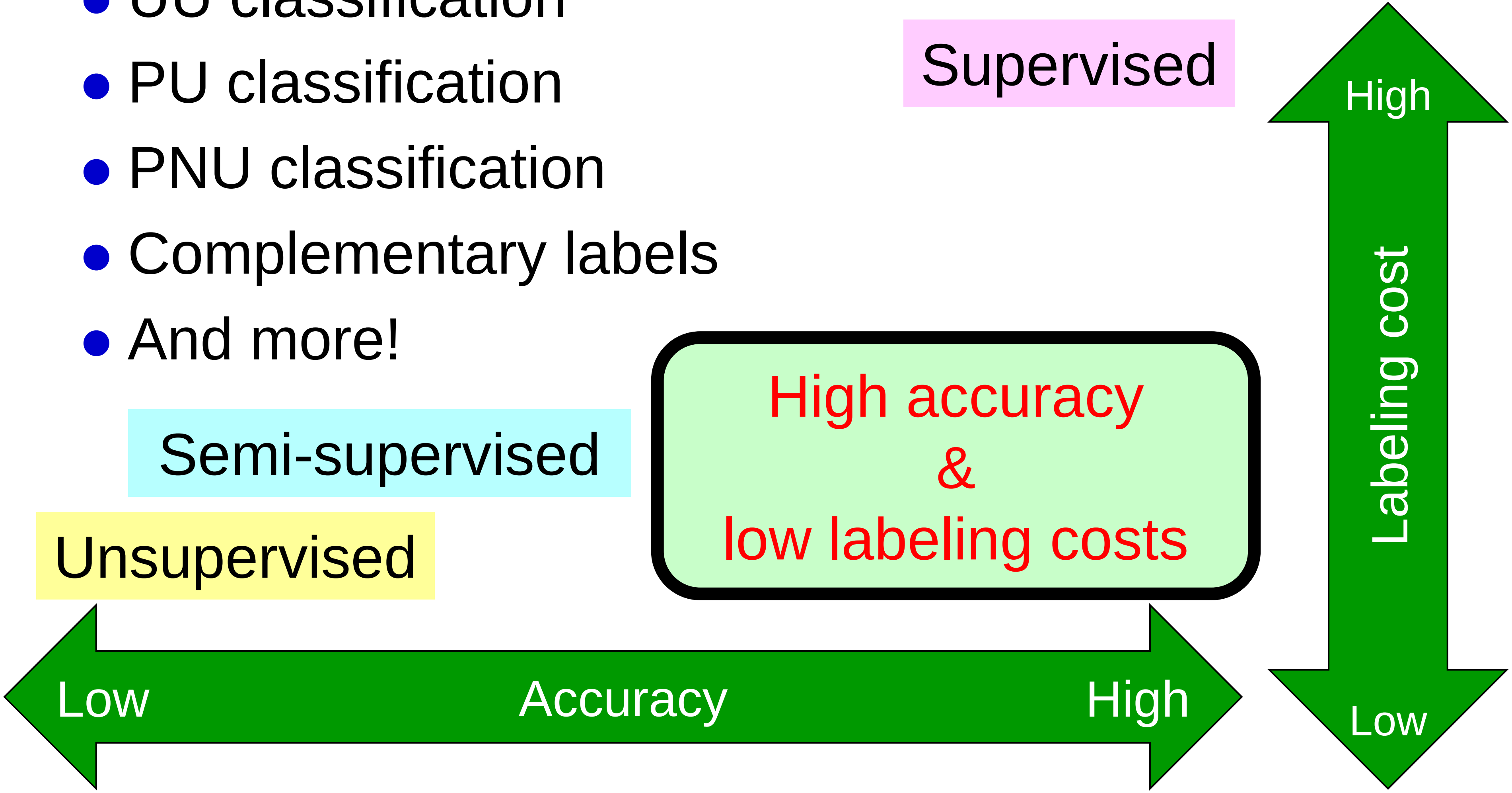
Dataset	Class	# train	# test	Proposed	Partial-label	Multi-label	Ordinary label
WAVEFORM1	$1 \sim 3$	1230	406	**85.7(0.9)**	84.1(1.5)	84.7(1.6)	85.8(0.9)
WAVEFORM2	$1 \sim 3$	1221	400	**84.4(1.3)**	**83.1(2.7)**	81.8(2.3)	86.7(1.8)
SATIMAGE	$1 \sim 7$	415	211	**67.2(7.0)**	54.8(6.8)	51.6(6.0)	67.9(4.2)
SHUTTLE	1, 4, 5	2458	809	94.9(9.7)	**97.5(0.7)**	90.4(11.8)	97.5(0.8)
SEGMENTATION	$1 \sim 7$	29	299	**36.1(6.8)**	31.7(5.8)	26.6(5.4)	58.6(4.5)
PENDIGITS	$1 \sim 5$	719	336	**79.4(9.5)**	73.2(6.4)	**75.9(7.7)**	78.8(2.9)
	$6 \sim 10$	719	335	**77.7(3.8)**	65.5(6.4)	72.0(8.6)	74.7(4.6)
	even #	719	335	**74.0(7.3)**	58.5(9.9)	65.7(6.3)	74.8(5.5)
	odd #	719	336	**88.5(5.9)**	74.6(4.4)	79.1(6.1)	84.0(8.8)
MNIST	$1 \sim 5$	5842	980	**88.4(4.2)**	71.5(7.4)	56.6(12.4)	77.9(0.4)
	$6 \sim 10$	5421	892	**83.4(2.6)**	67.4(8.1)	50.5(13.7)	77.0(4.5)
	even #	5421	892	**85.3(2.2)**	70.4(6.7)	61.7(11.1)	76.7(1.4)
	odd #	5842	958	**85.0(3.7)**	67.3(8.6)	57.3(13.0)	76.5(0.7)
DRIVE	$1 \sim 5$	3931	1280	**87.6(5.9)**	72.7(7.0)	64.2(12.6)	79.3(5.1)
	$6 \sim 10$	3958	1318	**84.9(5.7)**	73.1(5.8)	69.7(9.3)	81.6(2.9)
	even #	3932	1295	**82.4(5.6)**	72.9(6.6)	63.2(12.8)	83.5(5.3)
	odd #	3931	1310	**76.9(8.0)**	60.0(6.9)	51.6(9.3)	65.4(3.3)
LETTER	$1 \sim 5$	565	171	**79.6(5.5)**	67.6(6.0)	71.0(9.3)	82.2(4.3)
	$6 \sim 10$	550	178	**73.2(6.3)**	63.9(6.1)	61.2(10.6)	75.9(5.6)
	$11 \sim 15$	556	177	**73.3(5.9)**	66.6(3.4)	59.0(10.1)	75.4(5.0)
	$16 \sim 20$	550	184	**71.5(5.9)**	64.9(5.2)	63.5(7.0)	73.9(5.3)
	$21 \sim 25$	585	167	**76.2(6.0)**	68.3(8.1)	63.1(11.2)	77.1(5.1)
VOWEL	$1 \sim 5$	48	42	**35.6(9.0)**	**37.0(9.3)**	31.5(6.7)	54.9(6.7)
	$6 \sim 10$	48	42	**32.6(7.5)**	**34.1(7.7)**	**30.0(9.8)**	53.0(4.4)
	even #	48	42	**36.6(9.0)**	**39.9(10.5)**	**33.3(7.8)**	62.1(5.6)
	odd #	48	42	**28.2(9.0)**	**28.8(7.2)**	23.2(4.8)	54.0(5.5)
USPS	$1 \sim 5$	652	166	**70.1(5.2)**	62.8(7.2)	45.8(5.9)	76.2(2.3)
	$6 \sim 10$	542	147	**64.3(4.7)**	**61.4(5.9)**	41.7(5.3)	76.9(5.1)
	even #	556	147	**70.6(5.4)**	63.7(7.2)	48.4(5.3)	75.7(2.7)
	odd #	542	166	**63.1(4.3)**	57.8(6.8)	37.8(5.7)	73.6(3.4)

Use only 1/(c-1) times less samples since 1 ordinary label corresponds to (c-1) complementary labels

Proposed method works well!

Summary

- We need continuous effort to achieve high classification accuracy with low labeling!
  - UU classification
  - PU classification
  - PNU classification
  - Complementary labels
  - And more!



Organization

1. Classification of classification
2. Classification from UU data
3. Classification from PU data
4. Classification from PNU data
5. Classification from complementary labels
6. Introduction RIKEN Center for AIP

RIKEN Center for AIP

- RIKEN founded Center for Advanced Intelligence Project (AIP) in 2016.
- Our missions:

1. Development of next-generation AI technology (understand deep learning and go beyond)
2. Acceleration of scientific research (iPS cells, manufacturing, materials…)
3. Contribution to solving socially critical problems (healthcare for super-aged society, disaster resilience, infrastructure management…)
4. Study of ethical, legal and social issues of AI.
5. Human resource development (academia & industry)

Organization of AIP Center

2017 June 1st

Over 200 researchers!

Various application domains
(companies, universities, research institutes, etc.)

Goal-Oriented Technology Research Group:
Abstract complex real-world problems into solvable forms
(22 teams)

Generic Technology Research Group:
Develop fundamental theory and algorithms for abstracted problems
(18 teams)

Artificial Intelligence in Society Research Group:
Analyze the influence of AI spreading in society
(8 teams)

NEC/ Fujitsu/ Toshiba Collaboration Centers

International Partners

China
- Peking University
- Nanjing University
- Shanghai University
- Hong Kong University of Science and Technology

Korea
- KAIST
- Postech
- Artificial Intelligence Research Institute

Singapore
- National University of Singapore

US
- Toyota Technological Institute at Chicago
- University of Pennsylvania

Germany
- Berlin Big Data Center
- Technische Universitaet Darmstadt

UK
- Edinburgh Center for Robotics

Finland
- Aalto University

More coming soon!

Computational Resources

With Dr. Bill Dally (NVIDIA SVP) (Feb. 27, 2017)
https://blogs.nvidia.co.jp/2017/03/06/fujitsu-ai-supercomputer/

FUJITSU NVIDIA

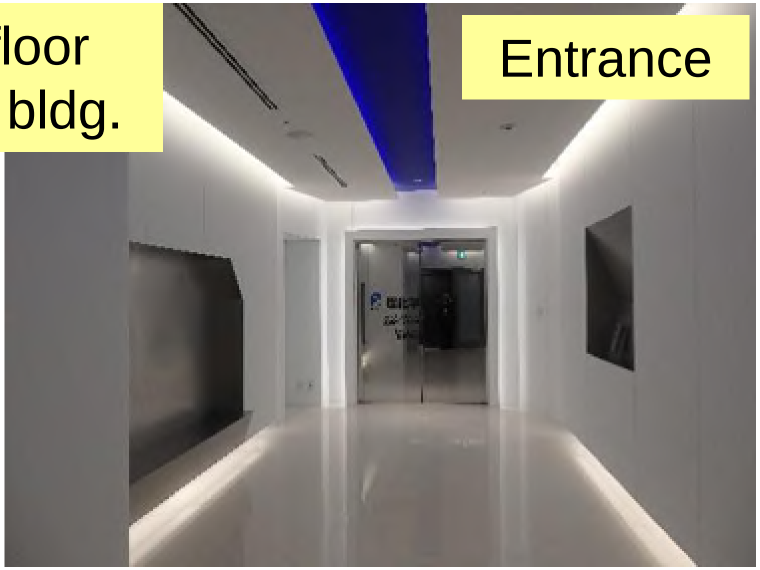
- 24 x NVIDIA DGX-1 (half-precision 4PFLOPS)
  - The largest customer installation of DGX-1 systems in March 2017.
- Ranked 4th in the Green500 List (June 2017)
  - 10.602GFLOPS/W

Our Office in the Heart of Tokyo!

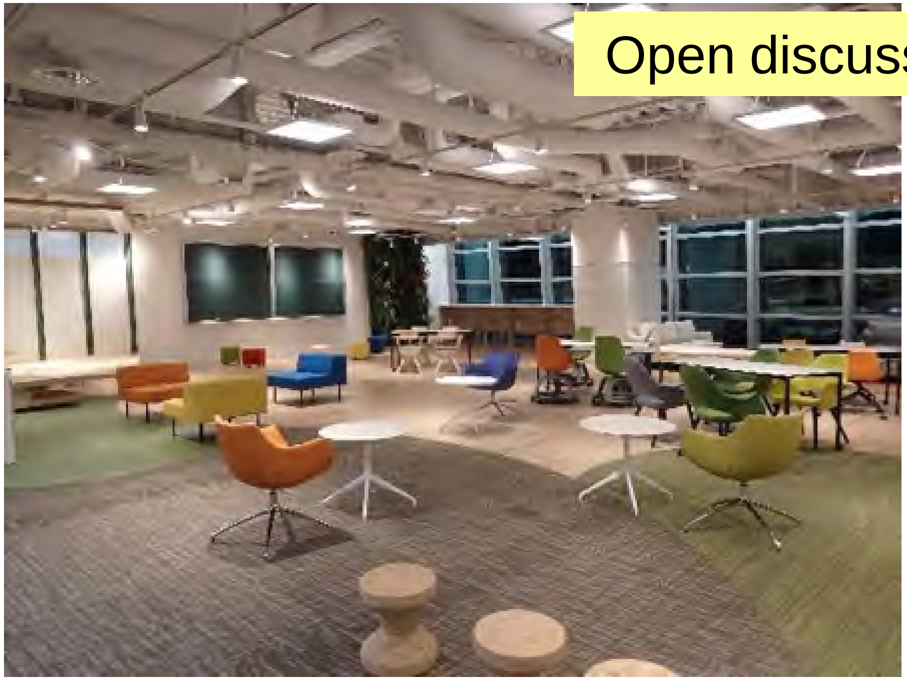
- Directly connected to Nihonbashi Station.
- Walking distance from Tokyo Station.

Visit us!

15th floor of this bldg.



Entrance



Open discussion space