Classification from Pairwise Similarity and Unlabeled Data

Han Bao[1,2], Gang Niu[2], Masashi Sugiyama[2,1]

[1]The University of Tokyo, Japan / [2]RIKEN, Japan

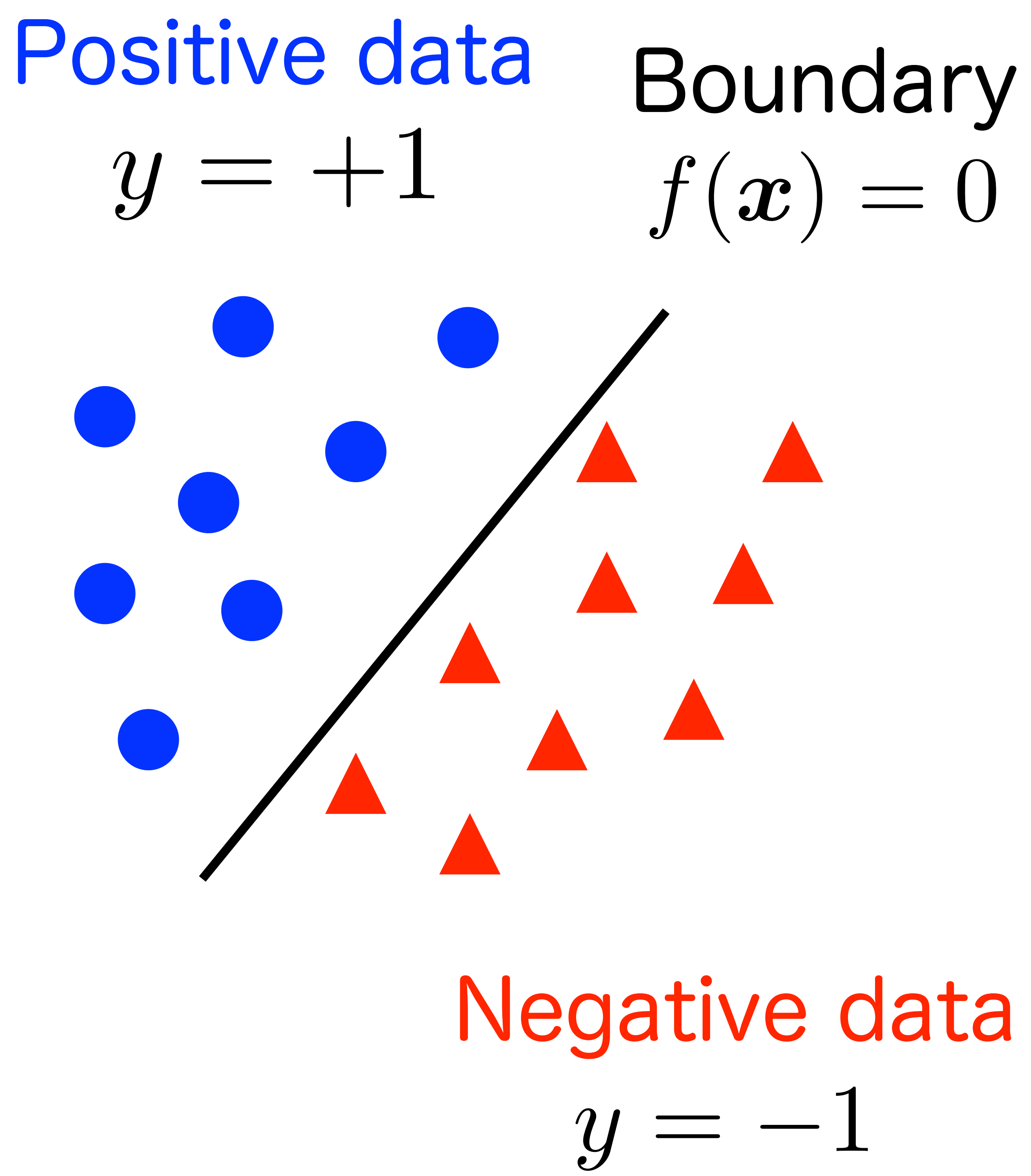
July 13th, 2018

東京大学
THE UNIVERSITY OF TOKYO

RIKEN

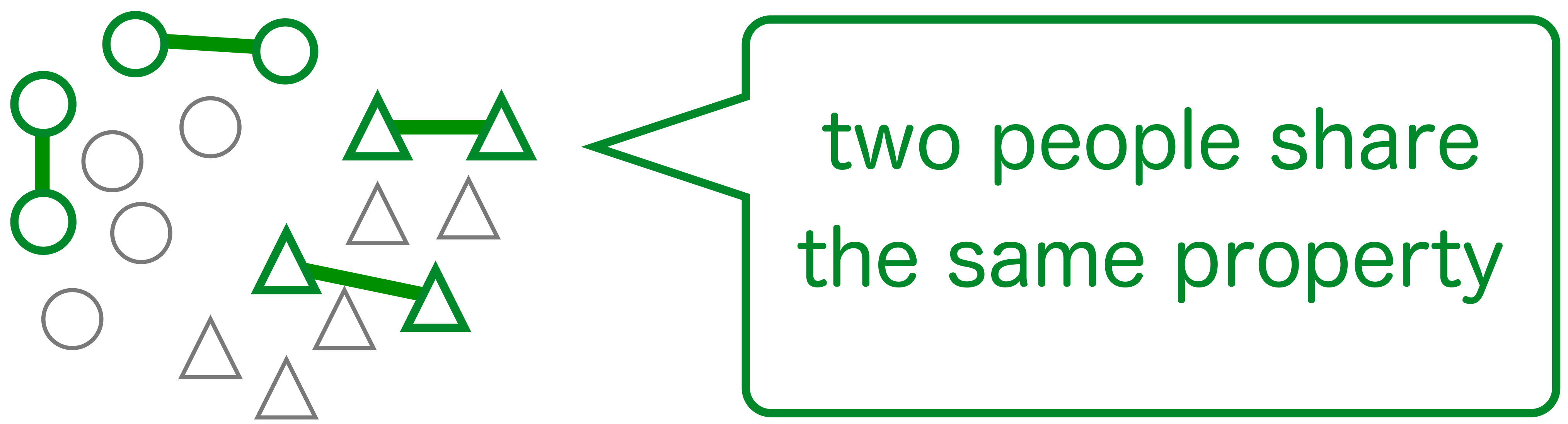
Gentle Start: Binary Classification

- Training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} p(\boldsymbol{x}, y)$
  where data $\boldsymbol{x}_i \in \mathbb{R}^d$ is labeled as $y_i \in \{+1, -1\}$
- Goal: find a classifier $f : \mathbb{R}^d \to \mathbb{R}$
- Method: minimize classification error
  - empirical risk minimization (ERM)



Motivation: Pairwise Information in Classification

- Classification of sensitive matters
  - e.g., politics, religion, opinion on racial issue
  - hard to obtain explicit label
  - instead asking “Which person do you share the same belief as?”
  - cf. randomized response technique [Warner 1965]



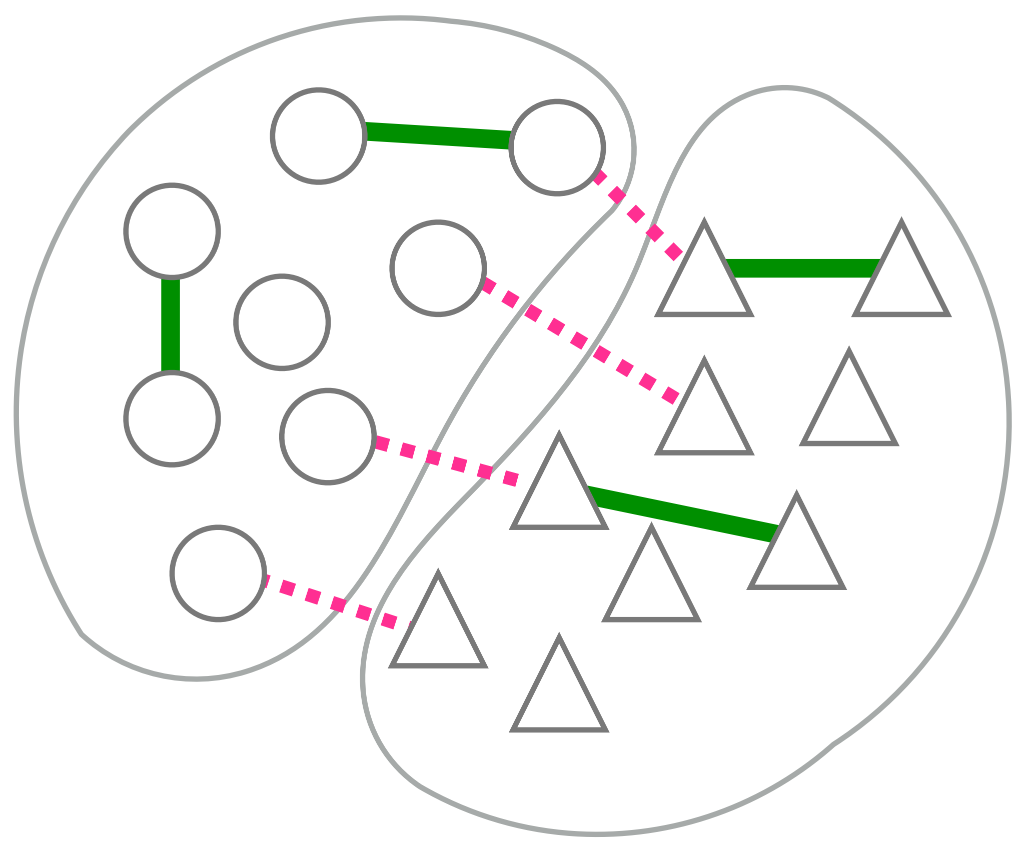
http://leanintokyo.org/wp-content/uploads/2017/12/MeToo.jpg

Warner, S. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 60(309):63–69, 1965.

Related: Semi-supervised Clustering

[Wagstaff+ ICML2001; many other papers]

- Clustering from
  - unlabeled $\mathcal{U} = \{\boldsymbol{x}_i\}$
  - similar $\mathcal{S} = \{(\boldsymbol{x}_i, \boldsymbol{x}'_i)\}$
  - dissimilar $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{x}'_i)\}$



- Offspring of unsupervised clustering
  - Problem: Cluster assumption
    (manifold assumption, low-density separation)
  - does not hold for many datasets

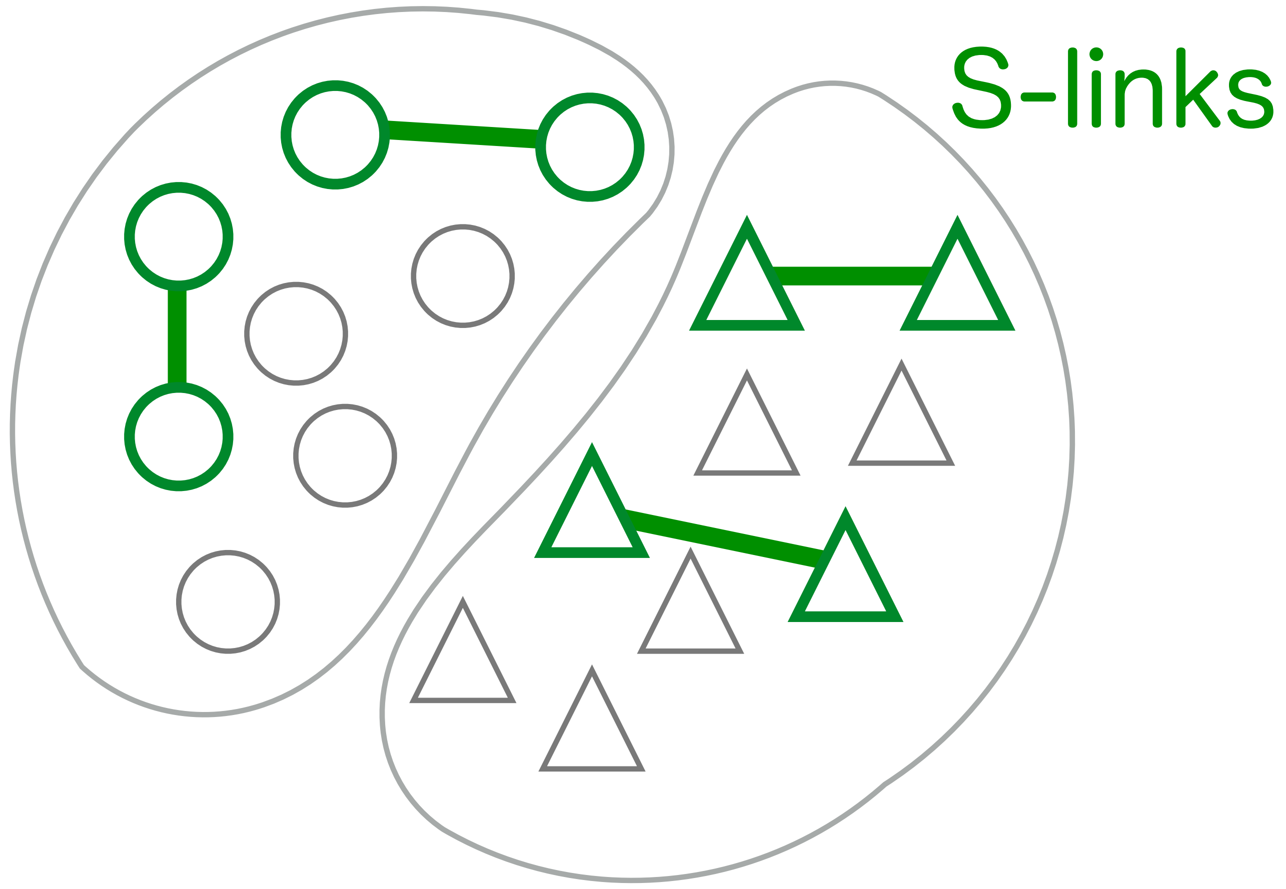
Wagstaff, K., Cardie, C., Rogers, S., and Schrödl, S. Constrained k-means clustering with background knowledge. In ICML, pp. 577–584, 2001.

Our work: SU Classification

- Binary classification from
  - S(imilar) data $\mathcal{S} = \{(\underline{\boldsymbol{x}_i, \boldsymbol{x}'_i})\}$
  - U(nlabeled) data $\mathcal{U} = \{\boldsymbol{x}_i\}$

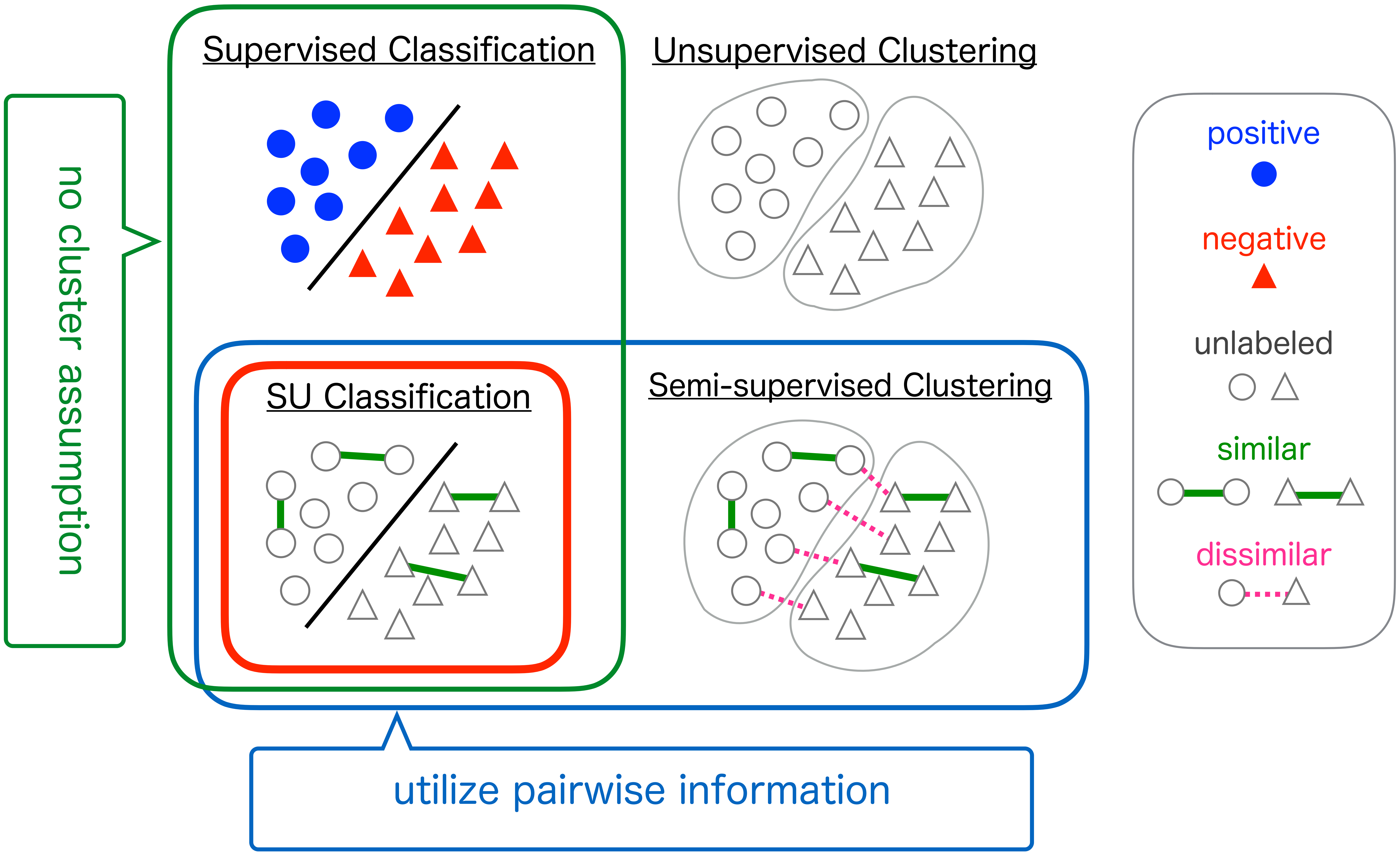
belong to the same class

- Formulation based on classification error is available



- Situation: prediction of sensitive matters
  - e.g., politics, religion
  - "Which person do you share the same belief as?"

Summary



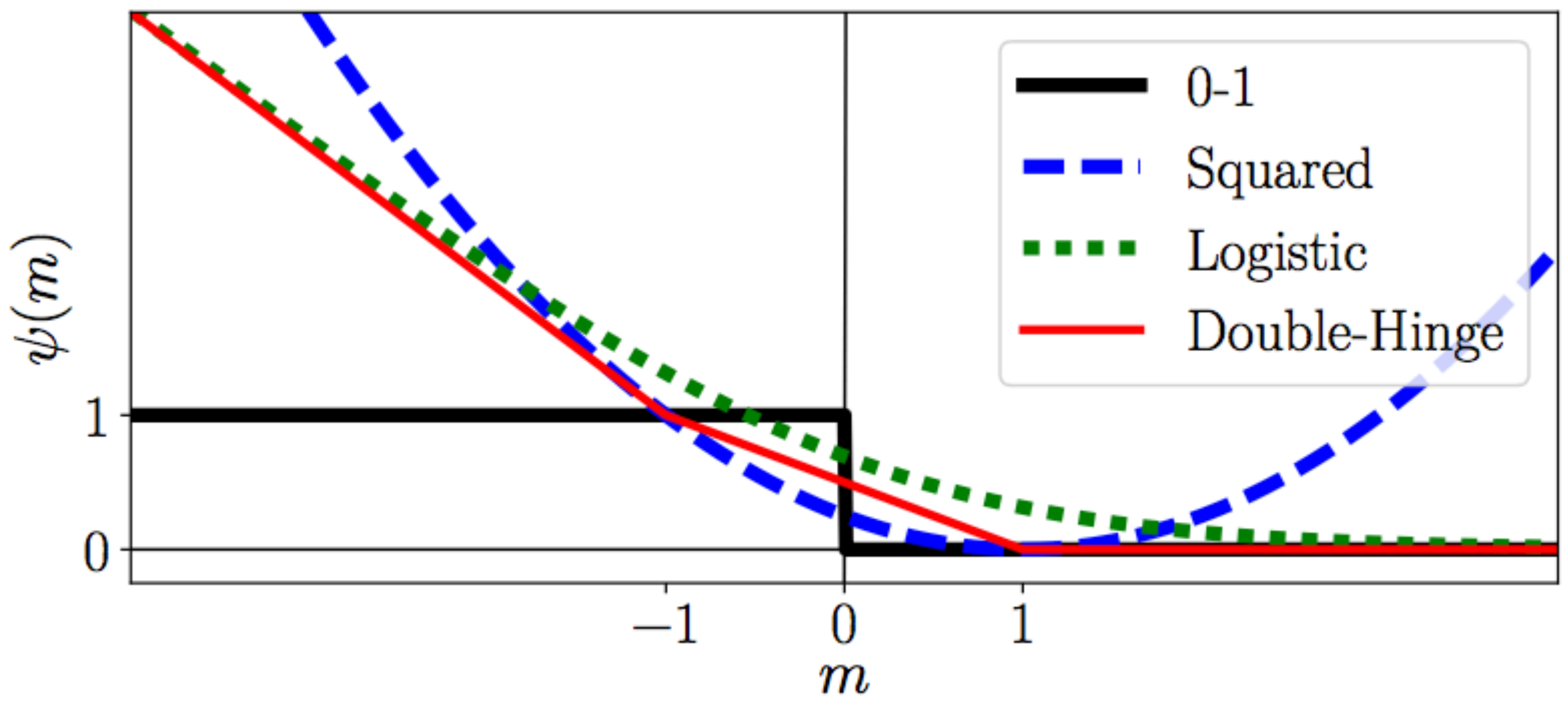
Supervised Classification
Unsupervised Clustering
no cluster assumption
SU Classification
Semi-supervised Clustering
positive
negative
unlabeled
similar
dissimilar
utilize pairwise information

Empirical Risk Minimization

- Goal: Minimize classification risk (= error rate)

Classification risk $R_{\mathrm{PN}}(f) = \underbrace{\mathbb{E}[\ell(yf(\boldsymbol{x}))]}_{\text{expectation over } p(\boldsymbol{x}, y)}$

Empirical risk $\widehat{R}_{\mathrm{PN}}(f) = \frac{1}{n}\sum_{i=1}^{n}\ell(y_i f(\boldsymbol{x}_i))$



example of loss functions

- Margin loss ℓ: give small/large penalty for positive/negative value of $y\,f(\boldsymbol{x})$
  - $\mathrm{sign}(y) = \mathrm{sign}(f(\boldsymbol{x}))$ ⇔ correct
  - $\mathrm{sign}(y) \neq \mathrm{sign}(f(\boldsymbol{x}))$ ⇔ incorrect

Assumption: Data Generating Process

- Pairwise Similarity (S): belong to the same class

$$\{(\boldsymbol{x}_{\mathrm{S},i}, \boldsymbol{x}'_{\mathrm{S},i})\}_{i=1}^{n_{\mathrm{S}}} \sim \; p_{\mathrm{S}}(\boldsymbol{x}, \boldsymbol{x}') = p(\boldsymbol{x}, \boldsymbol{x}' \mid \boxed{y = y' = +1 \vee y = y' = -1})$$

$$= \frac{\pi_+^2 p_+(\boldsymbol{x}) p_+(\boldsymbol{x}') + \pi_-^2 p_-(\boldsymbol{x}) p_-(\boldsymbol{x}')}{\pi_+^2 + \pi_-^2}$$

- Unlabeled Data (U)

$$\{\boldsymbol{x}_{\mathrm{U},i}\}_{i=1}^{n_{\mathrm{U}}} \sim p(\boldsymbol{x})$$

conditional density	class prior
$p_+(\boldsymbol{x}) \triangleq p(\boldsymbol{x} \mid y = +1)$	$\pi_+ \triangleq p(y = +1)$
$p_-(\boldsymbol{x}) \triangleq p(\boldsymbol{x} \mid y = -1)$	$\pi_- \triangleq p(y = -1)$

Classification risk can be estimated from SU data

Theorem:

$$\widehat{R}_{\mathrm{SU},\ell}(f) = \underbrace{\frac{\pi_{\mathrm{S}}}{n_{\mathrm{S}}} \sum_{i=1}^{n_{\mathrm{S}}} \frac{\mathcal{L}_{\mathrm{S},\ell}(f(\boldsymbol{x}_{\mathrm{S},i})) + \mathcal{L}_{\mathrm{S},\ell}(f(\boldsymbol{x}'_{\mathrm{S},i}))}{2}}_{\text{risk for S data}} + \underbrace{\frac{1}{n_{\mathrm{U}}} \sum_{i=1}^{n_{\mathrm{U}}} \mathcal{L}_{\mathrm{U},\ell}(f(\boldsymbol{x}_{\mathrm{U},i}))}_{\text{risk for U data}}$$

no explicit labels needed

is unbiased to $R_{\mathrm{PN},\ell}(f) = \mathbb{E}\left[\ell(yf(\boldsymbol{x}))\right]$

(original classification risk)

- minimize $\widehat{R}_{\mathrm{SU},\ell}(f)$ ⇒ minimize classification risk
- $\widehat{R}_{\mathrm{SU},\ell}(f)$ can be computed only from SU data

N.B.: π_{S} can be estimated

$\pi_+ \triangleq p(y=+1)$

$\pi_- \triangleq p(y=-1)$

$\pi_{\mathrm{S}} \triangleq \pi_+^2 + \pi_-^2$

ℓ : surrogate loss

$$\mathcal{L}_{\mathrm{S},\ell}(z) \triangleq \frac{\ell(z) - \ell(-z)}{2\pi_+ - 1} \qquad \mathcal{L}_{\mathrm{U},\ell}(z) \triangleq \frac{-\pi_-\ell(z) + \pi_+\ell(-z)}{2\pi_+ - 1}$$

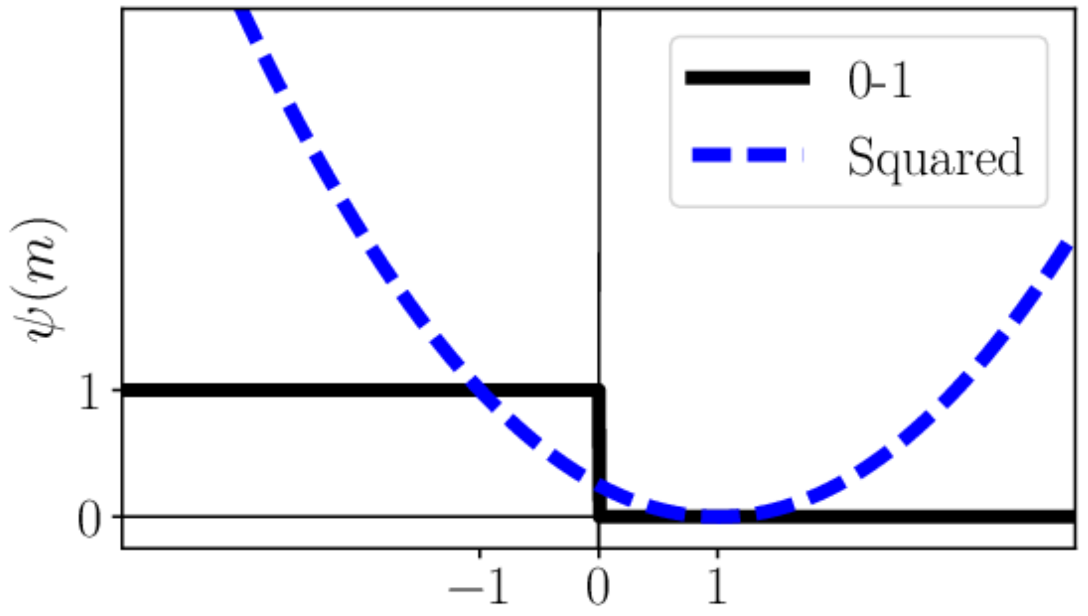
Good loss function gives convex objective

- Theorem: If the loss function ℓ satisfies

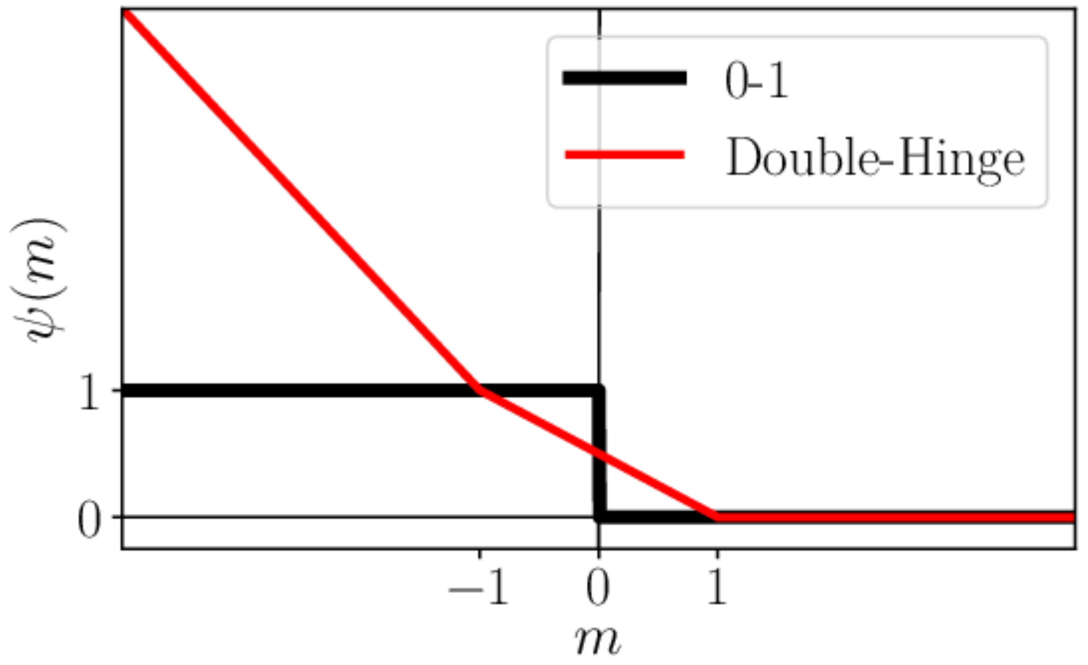
$$\ell(z) - \ell(-z) = -z\,,$$

then the objective (w/ model $f(\boldsymbol{x}) = \boldsymbol{w}^\top \phi(\boldsymbol{x})$ and l_2-reguralization) is convex.

- Example: squared loss, double-hinge loss, logistic loss



squared loss
⇒ analytical solution (linear system)



double-hinge loss
⇒ quadratic program

computationally efficient to obtain solution

Estimation Error Bound

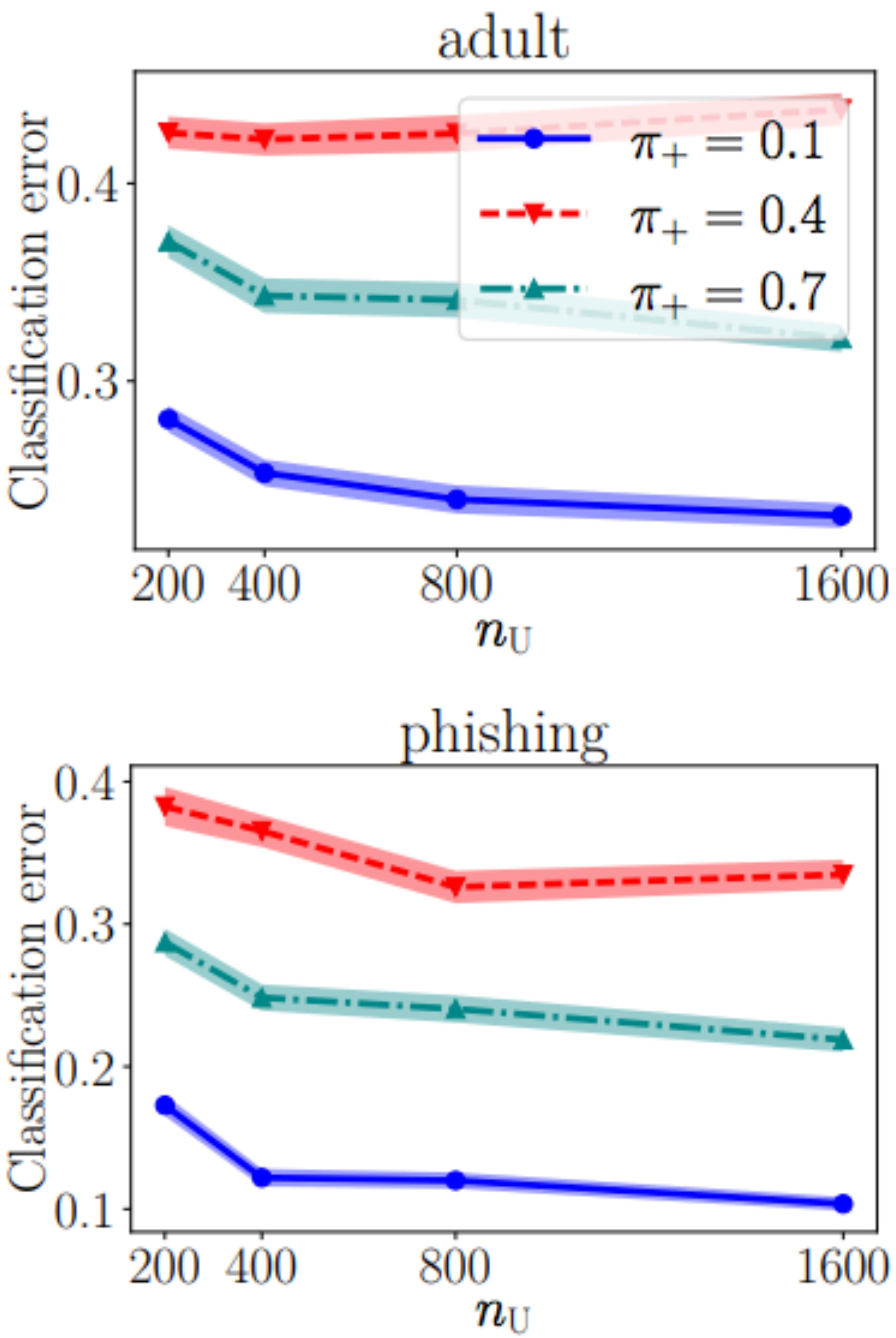
$$\underbrace{R_{\mathrm{PN}}(\hat{f}) - R_{\mathrm{PN}}(f^*)}_{\text{estimation error of risk of empirical minimizer}} = \mathcal{O}_p\left(\frac{1}{\sqrt{2\underbrace{n_{\mathrm{S}}}_{\text{\# S data}}}} + \frac{1}{\sqrt{\underbrace{n_{\mathrm{U}}}_{\text{\# U data}}}}\right)$$

- consistency: estimation error $\to 0$ as $n \to \infty$
- optimal parametric convergence rate [Mendelson 2008]

$$R_{\mathrm{PN}}(f) = \mathbb{E}[\ell(yf(\boldsymbol{x}))]$$

$$f^* = \operatorname*{argmin}_f R_{\mathrm{PN}}(f) \quad \text{true minimizer}$$

$$\hat{f} = \operatorname*{argmin}_f \widehat{R}_{\mathrm{SU}}(f) \quad \text{empirical minimizer}$$



Mendelson, S. Lower bounds for the empirical minimization algorithm. IEEE Transactions on Information Theory, 54 (8):3797–3803, 2008.

Experiments (Benchmark Datasets)

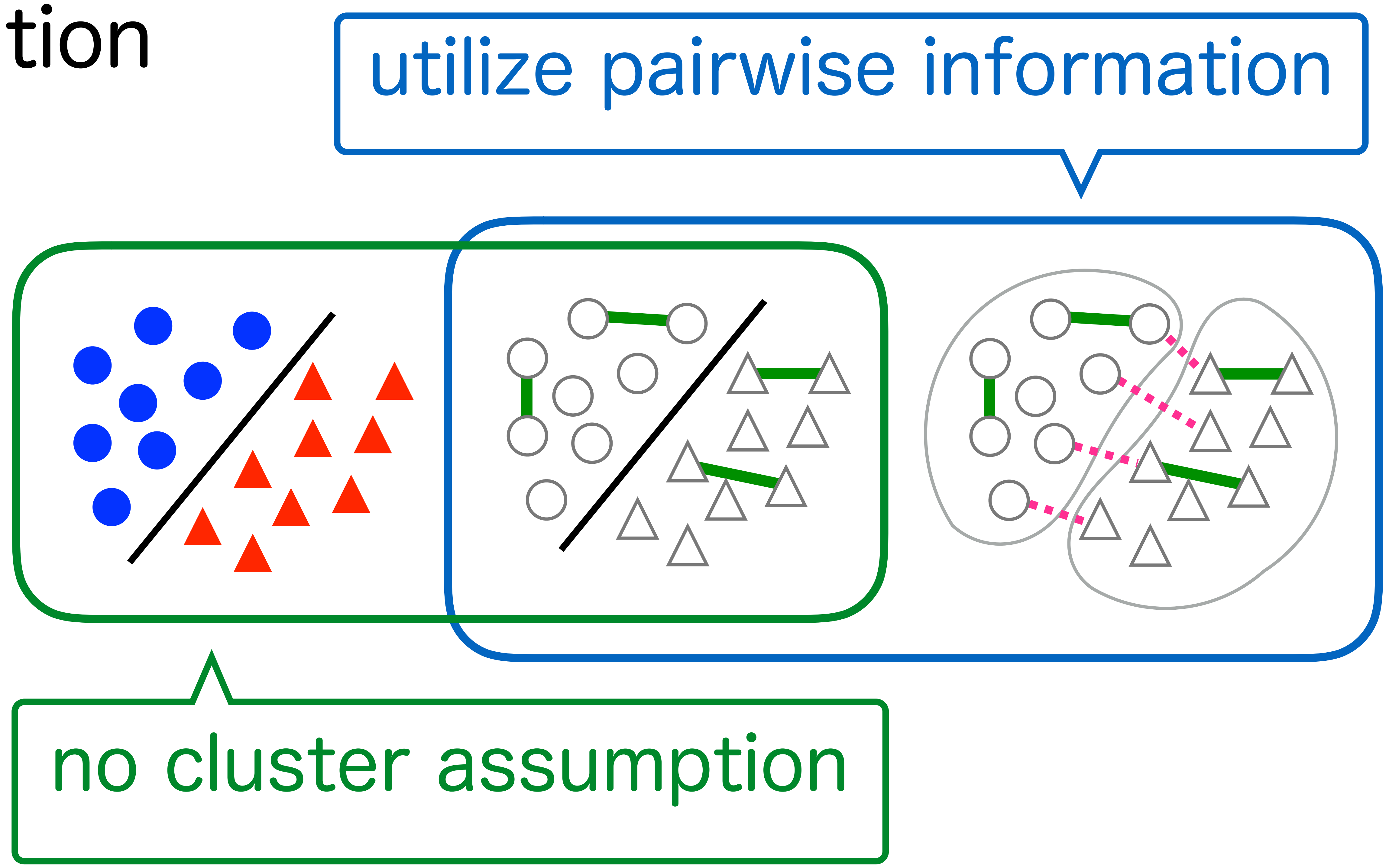
classification accuracies are shown

Dataset	Dim	SU(proposed)		Baselines					
		Squared	Double-hinge	KM	ITML	SERAPH	3SMIC	DIMC	IMSAT(linear)
adult	123	64.5 (1.2)	**84.5 (0.8)**	58.1 (1.1)	57.9 (1.1)	66.5 (1.7)	58.5 (1.3)	63.7 (1.2)	69.8 (0.9)
banana	2	**67.5 (1.2)**	**68.2 (1.2)**	54.3 (0.7)	54.8 (0.7)	55.0 (1.1)	61.9 (1.2)	64.3 (1.0)	**69.8 (0.9)**
cod-rna	8	**82.8 (1.3)**	71.0 (0.9)	63.1 (1.1)	62.8 (1.0)	62.5 (1.4)	56.6 (1.2)	63.8 (1.1)	69.1 (0.9)
higgs	28	55.1 (1.1)	**69.3 (0.9)**	**66.4 (1.6)**	**66.6 (1.3)**	63.4 (1.1)	57.0 (0.9)	65.0 (1.1)	**69.7 (1.4)**
ijcnn1	22	65.5 (1.3)	**73.6 (0.9)**	54.6 (0.9)	55.8 (0.7)	59.8 (1.2)	58.9 (1.3)	66.2 (2.2)	68.5 (1.1)
magic	10	66.0 (2.0)	**69.0 (1.3)**	53.9 (0.6)	54.5 (0.7)	55.0 (0.9)	59.1 (1.4)	63.1 (1.1)	**70.0 (1.1)**
phishing	68	75.0 (1.4)	**91.3 (0.6)**	64.4 (1.0)	61.9 (1.1)	62.4 (1.1)	60.1 (1.3)	64.8 (1.4)	69.4 (0.8)
phoneme	5	**67.8 (1.5)**	**70.8 (1.0)**	65.2 (0.9)	66.7 (1.4)	**69.1 (1.4)**	61.3 (1.1)	64.5 (1.2)	**69.2 (1.1)**
spambase	57	69.7 (1.4)	**85.5 (0.5)**	60.1 (1.8)	54.4 (1.1)	65.4 (1.8)	61.5 (1.3)	63.6 (1.3)	70.5 (1.1)
susy	18	59.8 (1.3)	**74.8 (1.2)**	55.6 (0.7)	55.4 (0.9)	58.0 (1.0)	57.1 (1.2)	65.2 (1.0)	70.4 (1.2)
w8a	300	62.1 (1.5)	**86.5 (0.6)**	71.0 (0.8)	69.5 (1.5)	0.0 (0.0)	60.5 (1.5)	65.0 (2.0)	70.2 (1.2)
waveform	21	77.8 (1.3)	**87.0 (0.5)**	56.1 (0.8)	54.8 (0.7)	56.5 (0.9)	56.5 (0.9)	65.0 (0.9)	69.7 (1.1)

- linear-in-input model / $n_U = n_S = 500$ / l_2-reguralization
- baseline: unsupervised / semi-supervised clustering

Summary

- Motivation



- SU classification does not need explicit labels
- Properties: convexity, estimation error bound

Poster: #67

arXiv URL