Maximum Margin Semi-supervised Learning with Irrelevant Data

包含无关数据的最大间隔半监督学习

Irwin King with Haiqin Yang

king@cse.cuhk.edu.hk
http://www.cse.cuhk.edu.hk/~king

Department of Computer Science & Engineering
The Chinese University of Hong Kong

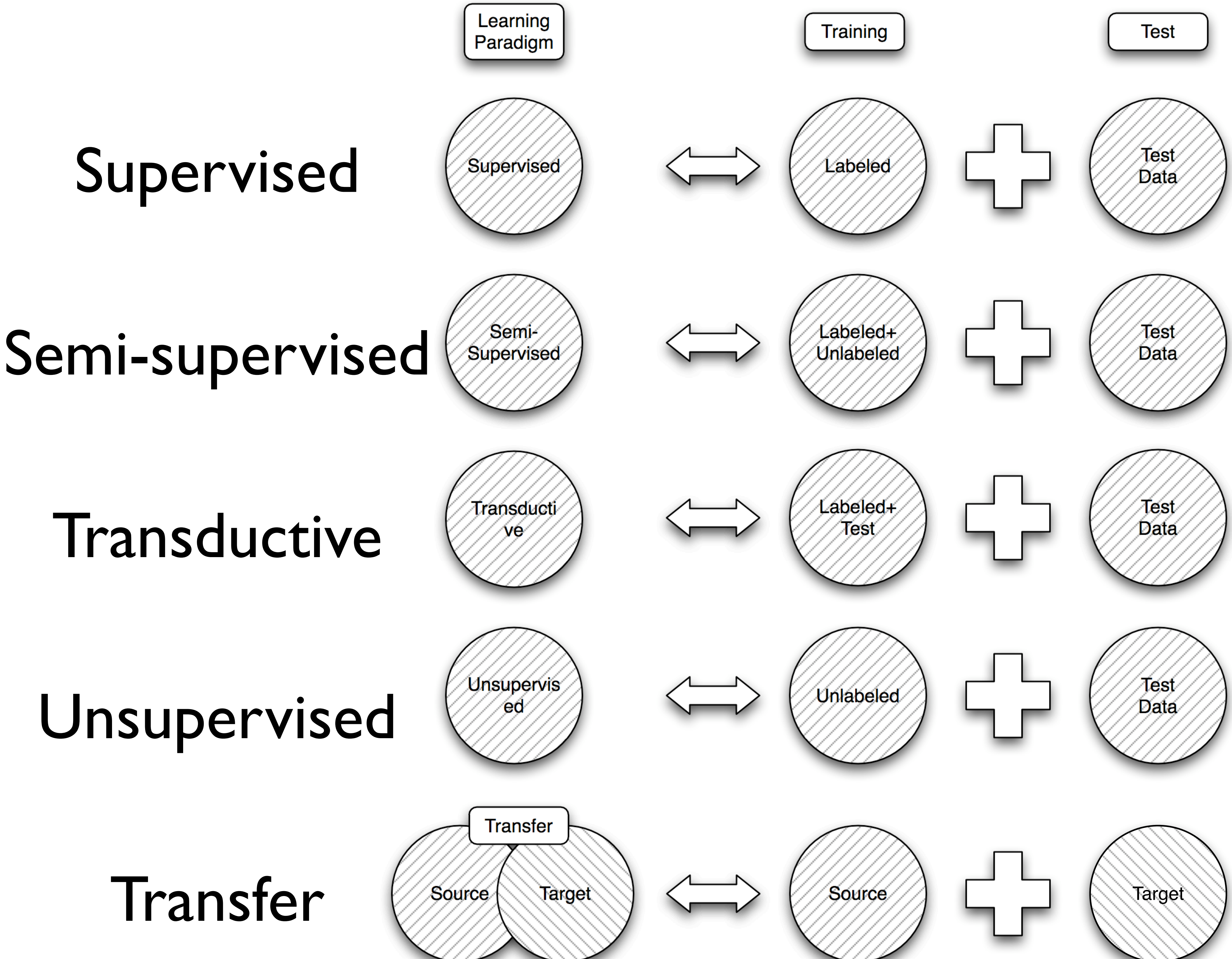
Example Datasets

USPS

MNIST

Other Applications

- Website categorization
  - Classify “sports news” vs. “financial news” with crawling
- Medical diagnosis
  - Classify one disease against another with the presence of other disease types
- Binary classification in multi-class classification tasks

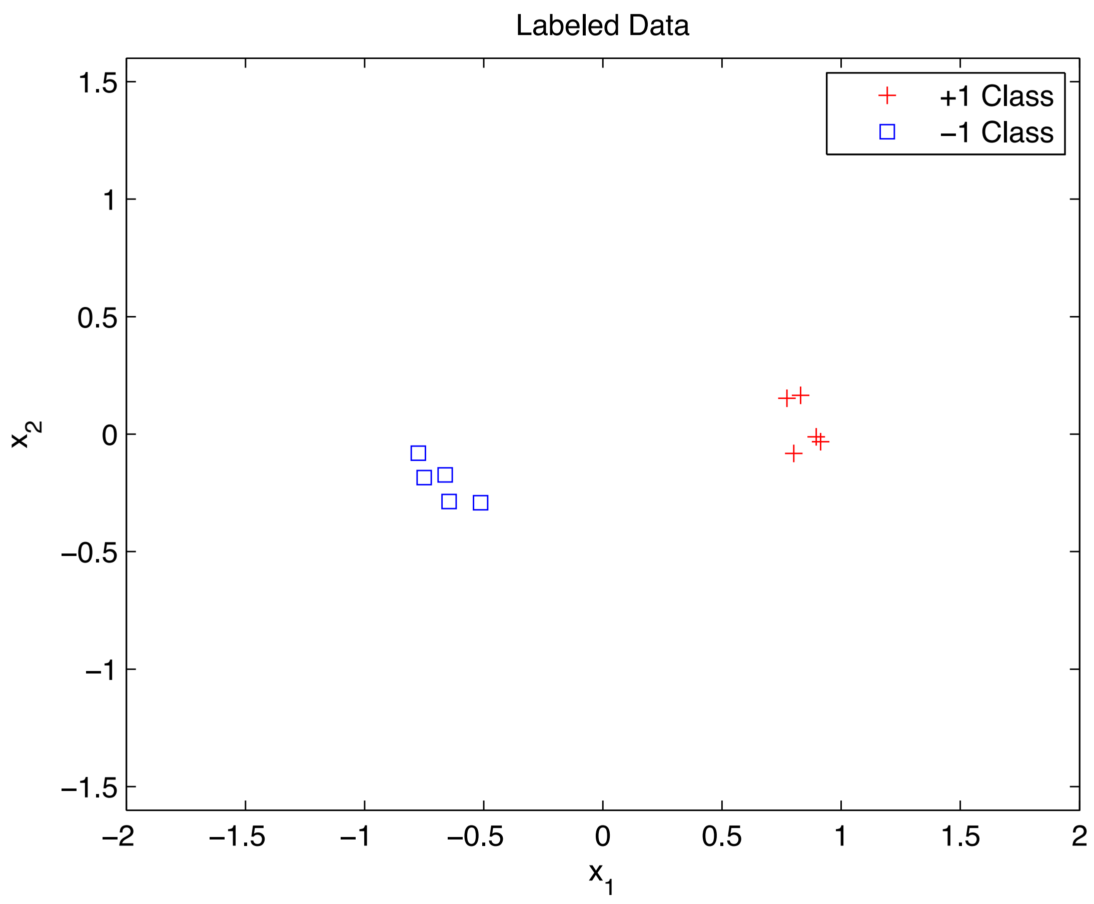

Learning Paradigm
Training
Test
Supervised
Supervised
Labeled
Test Data
Semi-supervised
Semi-Supervised
Labeled+ Unlabeled
Test Data
Transductive
Transductive
Labeled+ Test
Test Data
Unsupervised
Unsupervised
Unlabeled
Test Data
Transfer
Transfer
Source
Target
Source
Target

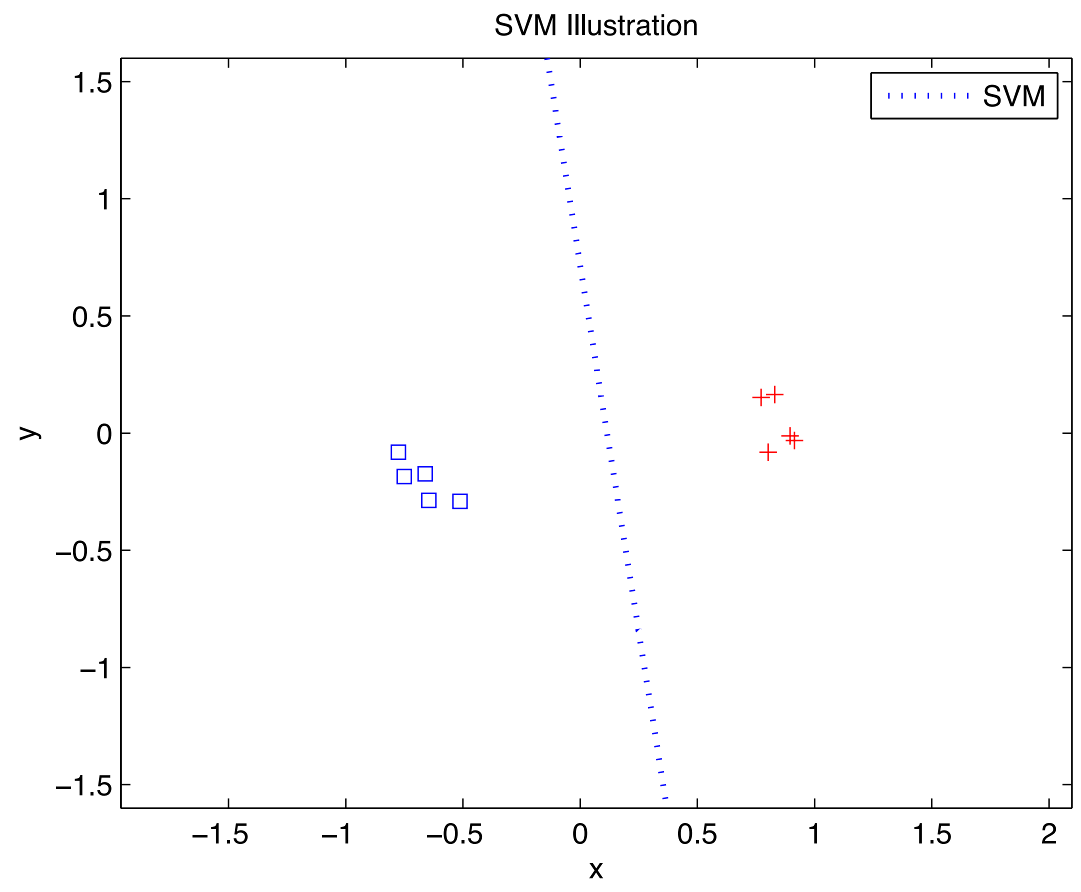
Why Semi-supervised Learning?

- Labeling data are rare, costly, and time consuming to obtain
- Many unlabeled data are easy to collect and may provide useful information

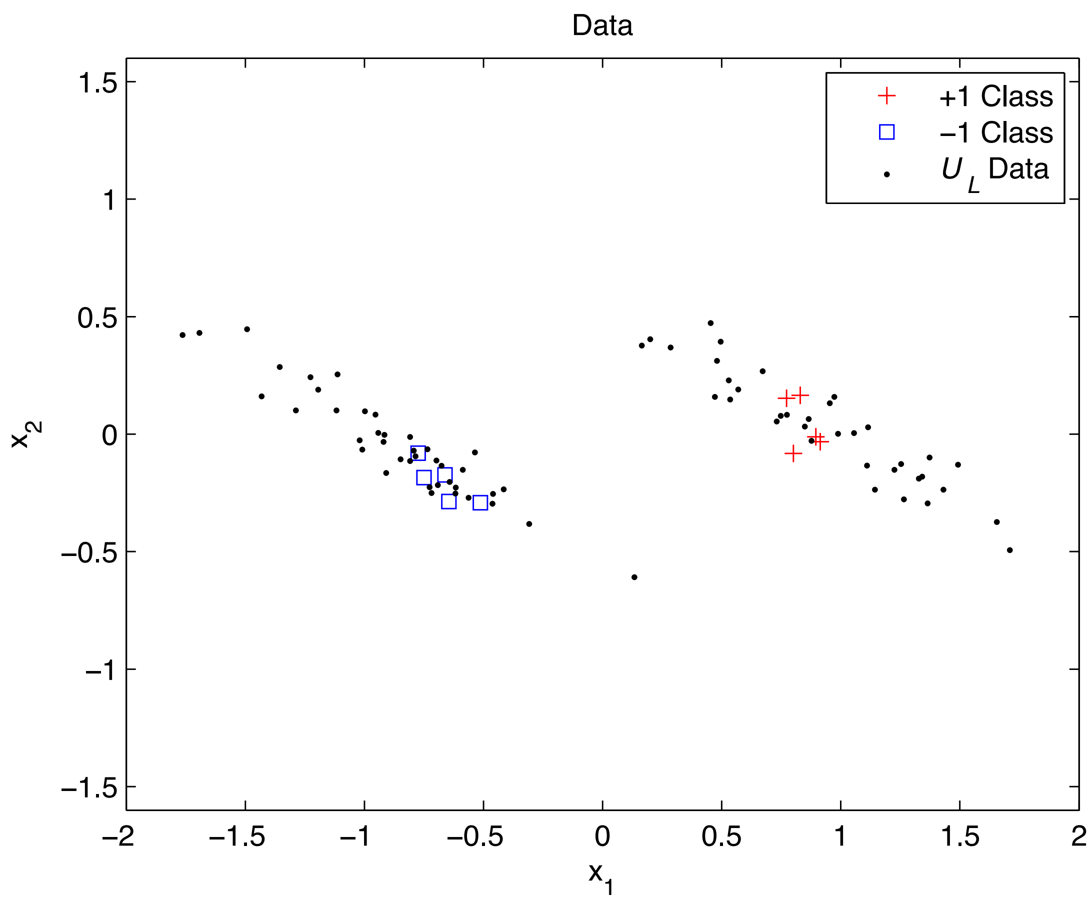
Consider to learn from both labeled and unlabeled data simultaneously!

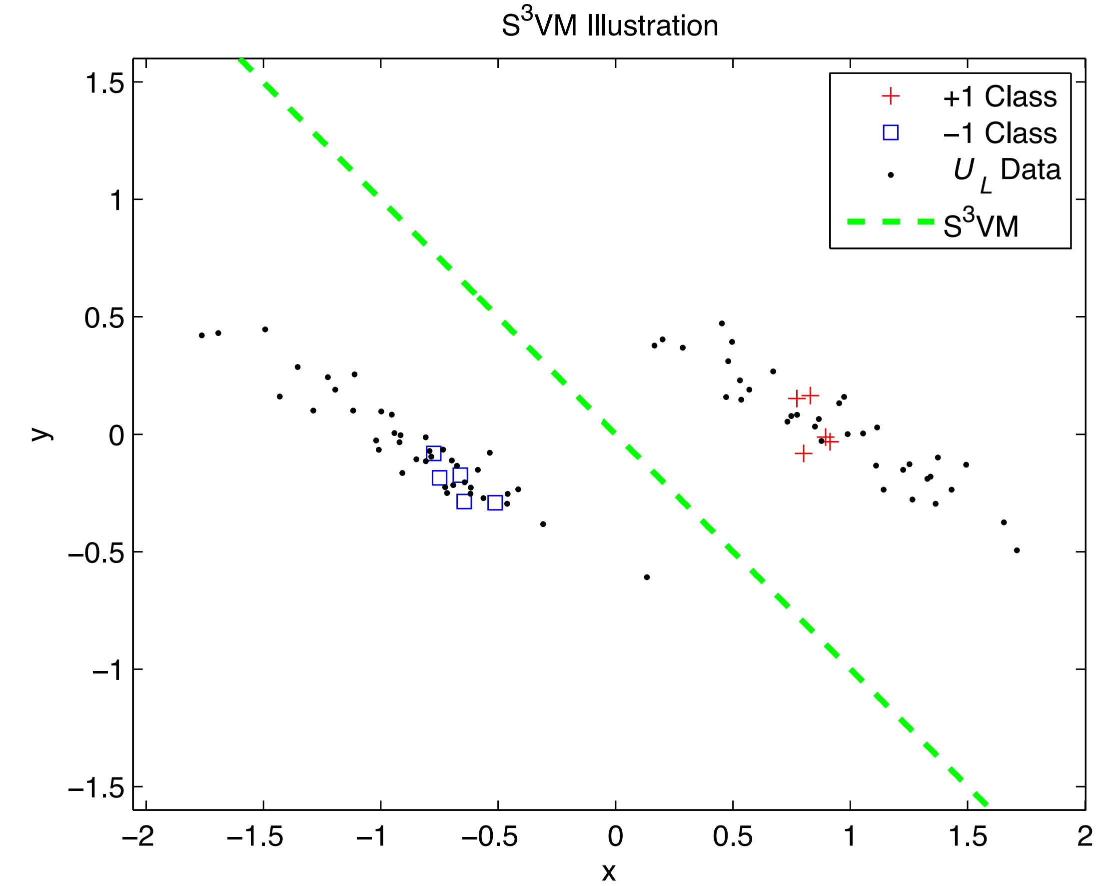
Support Vector Machines (SVMs)



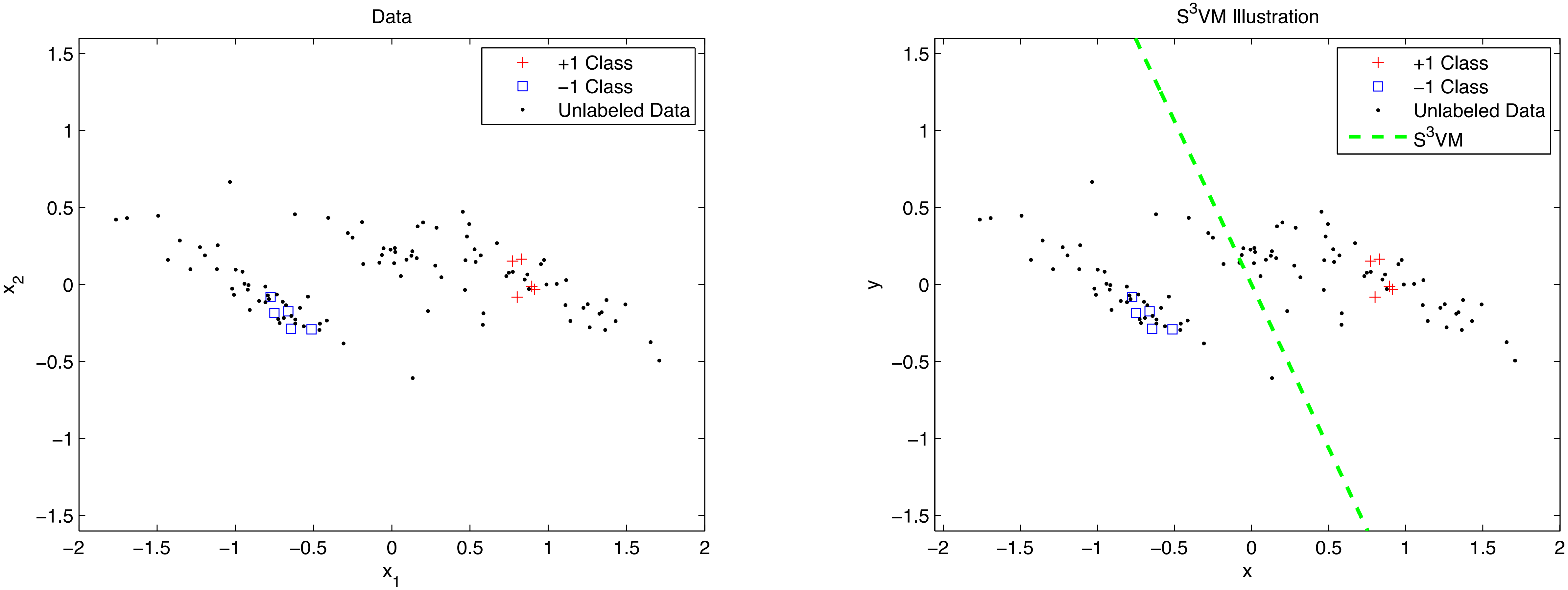


S^3VMs w/ Clean Unlabeled Data

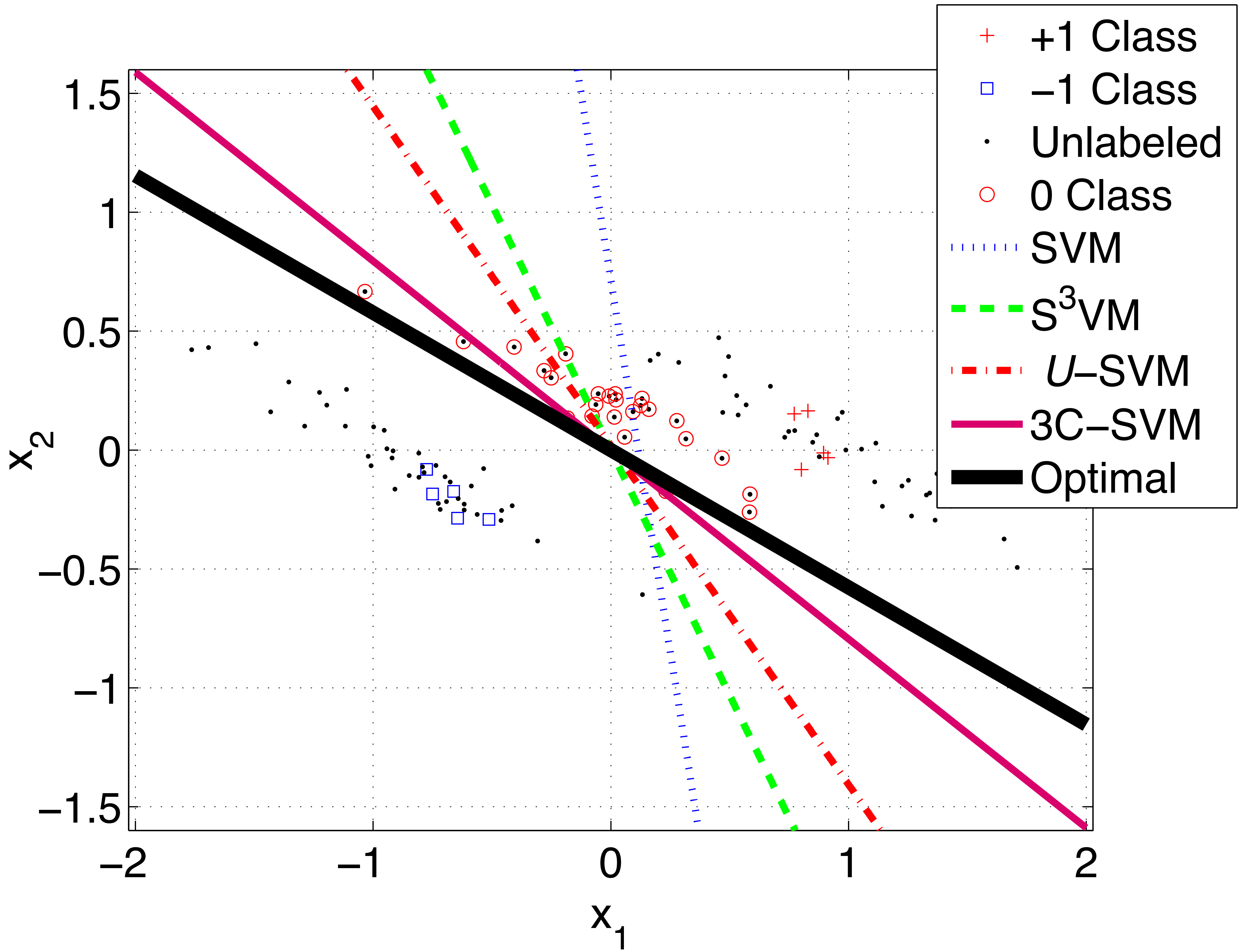




S^3VM w/ Unclean Unlabeled Data



Classifiers



SSL Assumptions

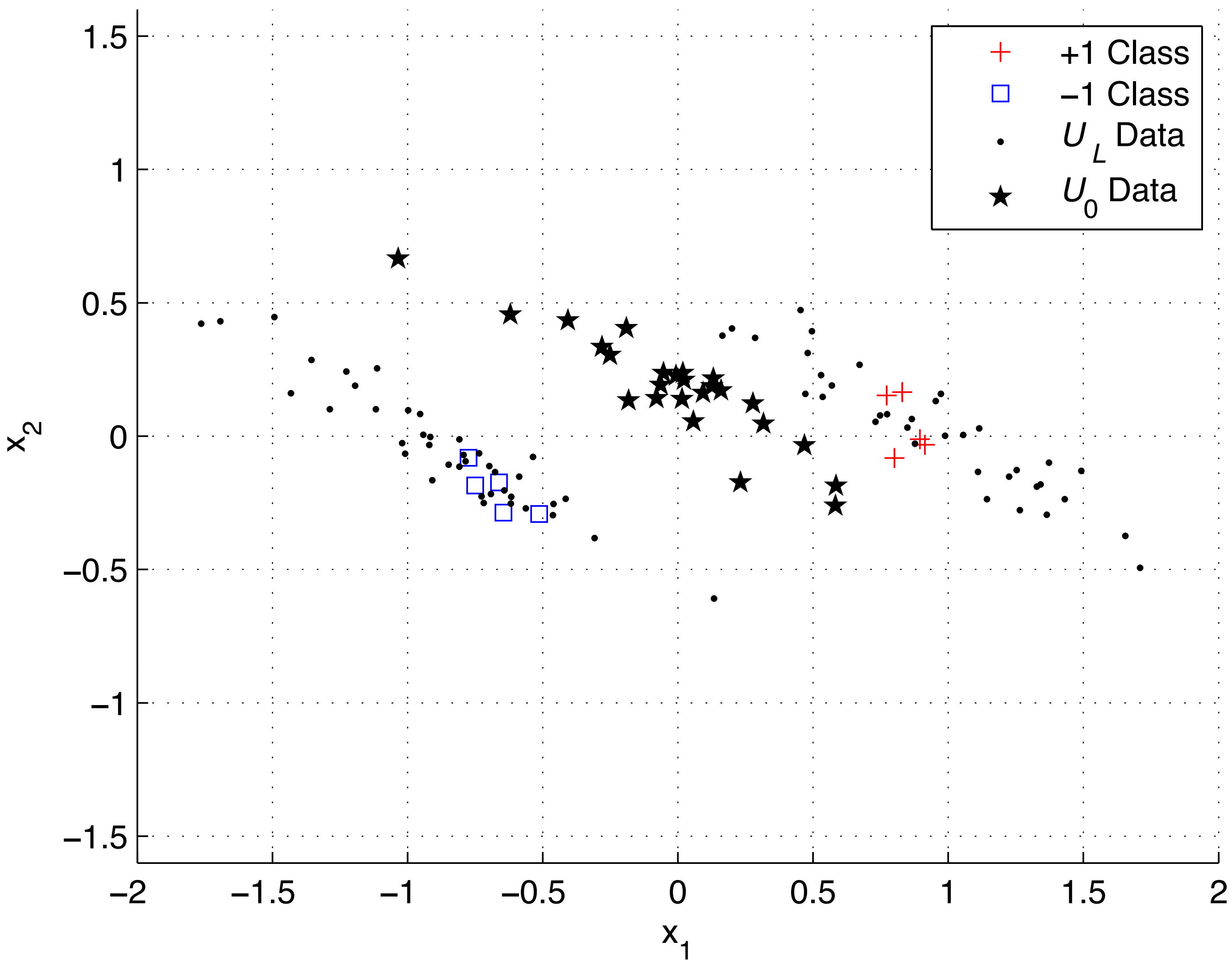
- Unlabeled data are from the same distribution as the labeled data

- What if they are not?

- Unlabeled data may be a mixture of relevant and **irrelevant** data!

- How can we utilize this irrelevant data information to improve performance in SSL?

Setup of Tri-Class SVM (3C-SVM)



$$\mathcal{L} = \{(\mathbf{x}_i,\, y_i)\}_{i=1}^{L}$$

$$\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d, y_i \in \{-1, 0, 1\}$$

$$\mathcal{U} = \mathcal{U}_{\mathcal{R}} \cup \mathcal{U}_0 = \{\mathbf{x}_i\}_{i=1}^{U}$$

Objective: seek

$$f_\vartheta(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b, \vartheta = (\mathbf{w}, b)$$

to separate the binary class data correctly with the help of (mixed) unlabeled data

Function Definition

- **Objective function:**

$$\min_{\vartheta} \quad \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{\mathbf{x}_i \in \mathcal{L}} r_i \ell_L(f_\vartheta(\mathbf{x}_i), y_i) + \sum_{\mathbf{x}_i \in \mathcal{U}} r_i \ell_U(f_\vartheta(\mathbf{x}_i)),$$

Margin | **Empirical Risk Labeled Data** | **Empirical Risk Unlabeled Data**

- **Facts:** if $f_\vartheta(\mathbf{x}_i) \gg 0$, more confident on $+1$-class

  Facts: if $f_\vartheta(\mathbf{x}_i) \ll 0$, more confident on -1-class

- **Principle:** rely more on labeled and relevant data (symmetrical hinge loss)

  Principle: ignore irrelevant data (ϵ-insensitive loss)

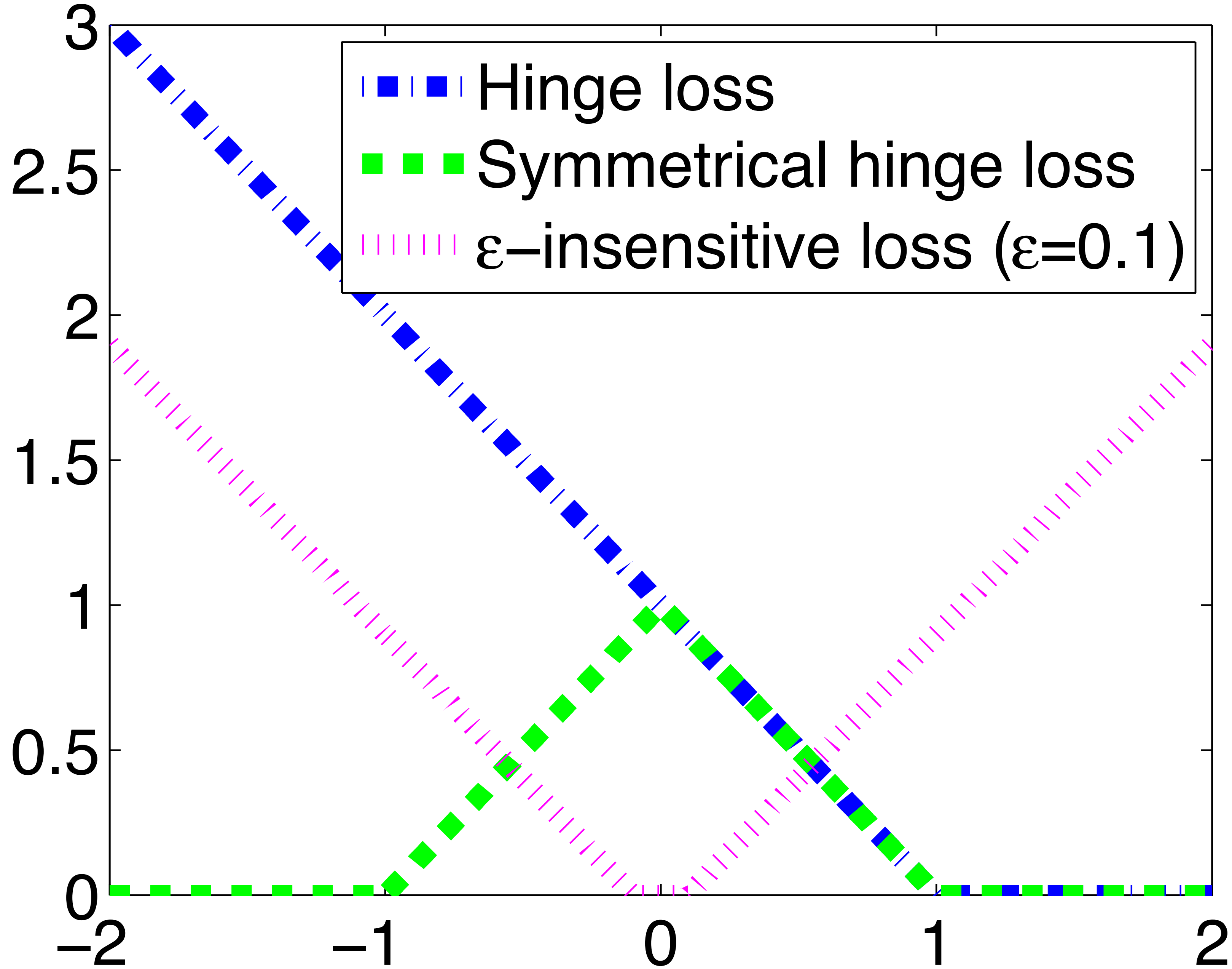
Function Definition

$$\min_{\vartheta} \quad \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{\mathbf{x}_i \in \mathcal{L}} r_i \ell_L(f_\vartheta(\mathbf{x}_i), y_i) + \sum_{\mathbf{x}_i \in \mathcal{U}} r_i \ell_U(f_\vartheta(\mathbf{x}_i)),$$

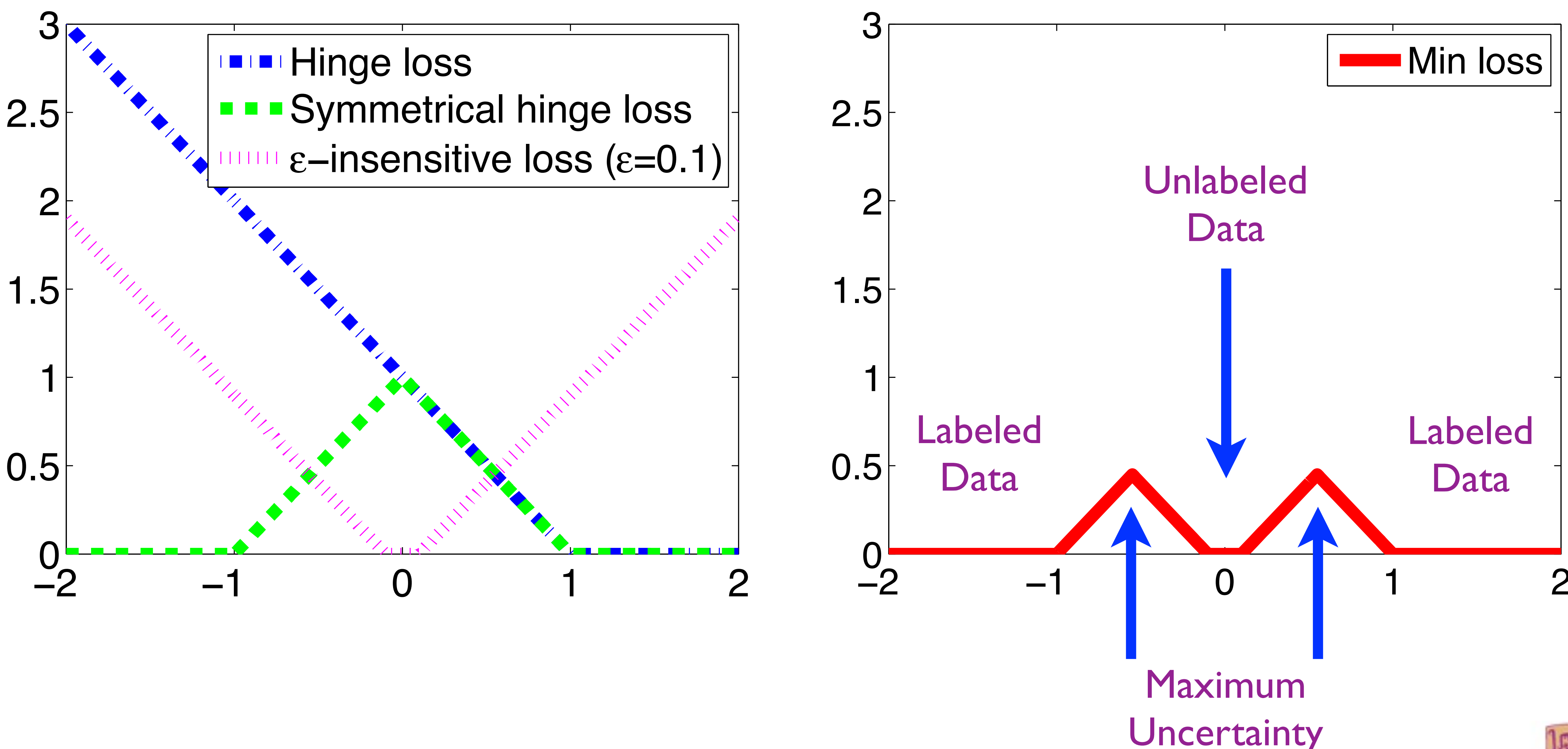
$$\min_{\vartheta} \quad \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{\mathbf{x}_i \in \mathcal{L}_{\pm 1}} r_i H_1(y_i f_\vartheta(\mathbf{x}_i)) + \sum_{\mathbf{x}_i \in \mathcal{L}_0} r_i I_\varepsilon(f_\vartheta(\mathbf{x}_i))$$
$$+ \sum_{\mathbf{x}_i \in \mathcal{U}} r_i \min\{H_1(|f_\vartheta(\mathbf{x}_i)|), I_\varepsilon(|f_\vartheta(\mathbf{x}_i)|)\}.$$

$$H_1(z) = \max\{0,\, 1 - z\}, \quad I_\varepsilon(z) = \max\{0,\, |z| - \varepsilon\}.$$

Loss Functions



Our Loss Function



Model Relationships

$$\min_{\vartheta} \quad \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{\mathbf{x}_i \in \mathcal{L}_{\pm 1}} r_i H_1(y_i f_\vartheta(\mathbf{x}_i)) + \sum_{\mathbf{x}_i \in \mathcal{L}_0} r_i I_\varepsilon(f_\vartheta(\mathbf{x}_i))$$
$$+ \sum_{\mathbf{x}_i \in \mathcal{U}} r_i \min\{H_1(|f_\vartheta(\mathbf{x}_i)|), I_\varepsilon(|f_\vartheta(\mathbf{x}_i)|)\} .$$

3C-SVM

$\mathcal{L}$	-1	0	1
$\mathcal{U}$	-1	0	1

SVM

$\mathcal{L}$	-1		1
$\mathcal{U}$			

S^{3}VM

$\mathcal{L}$	-1		1
$\mathcal{U}$	-1		1

$\mathcal{U}$-SVM

$\mathcal{L}$	-1	0	1
$\mathcal{U}$			

Theorem

- Objective function:

$$\min_{\vartheta} \quad \frac{\lambda}{2}\|\mathbf{w}\|^2 + \sum_{\mathbf{x}_i \in \mathcal{L}_{\pm 1}} r_i H_1(y_i f_\vartheta(\mathbf{x}_i)) + \sum_{\mathbf{x}_i \in \mathcal{L}_0} r_i I_\varepsilon(f_\vartheta(\mathbf{x}_i))$$

$$+ \sum_{\mathbf{x}_i \in \mathcal{U}} r_i \min\{H_1(|f_\vartheta(\mathbf{x}_i)|), I_\varepsilon(|f_\vartheta(\mathbf{x}_i)|)\}\,.$$

- 3C-SVM

Suppose $r_i = \infty$ for unlabeled data and $\epsilon = 0$.

Unlabeled data $\mathbf{x}_j$ satisfies
(a) $|\mathbf{w}^T\phi(\mathbf{x}_j) + b| \geq 1 \Rightarrow$ data lie on or out of the margin gap,
or
(b) $\mathbf{w}^T\phi(\mathbf{x}_j) + b = 0 \Rightarrow \mathbf{w}^T(\phi(\mathbf{x}_j) - \phi(\mathbf{x}_0)) = 0,\ \mathbf{x}_j, \mathbf{x}_0 \in \mathcal{U}_0$

Removing Min-Terms

$$\min_{\vartheta,\mathbf{d}} \quad \frac{\lambda}{2}\|\mathbf{w}\|^2+\sum_{\mathbf{x}_i\in\mathcal{L}_{\pm 1}} r_i H_1(y_i f_\vartheta(\mathbf{x}_i))+\sum_{\mathbf{x}_i\in\mathcal{L}_0} r_i I_\varepsilon(f_\vartheta(\mathbf{x}_i))$$

$$+\sum_{\mathbf{x}_{k+L}\in\mathcal{U}} r_{k+L}\left(\underbrace{H_1(|f_\vartheta(\mathbf{x}_i)|+D(1-d_k))}_{Q_1}+\underbrace{I_\varepsilon(|f_\vartheta(\mathbf{x}_i)|-Dd_k)}_{Q_2}\right),$$

- $d_k = 0 \Rightarrow Q_1 = 0$,
- $d_k = 1 \Rightarrow Q_2 = 0$,
- $H_1(|z|+a)$: non-convex, approximated by ramploss, $H_{1-a}(z)-H_\kappa(z)+H_{1-a}(-z)-H_\kappa(-z)$,
- $I_\varepsilon(|z|-b) = H_{-\varepsilon-b}(-z) + H_{-\varepsilon-b}(z)$,
- $H_1(|z|+a)$ and $I_\varepsilon(|z|-b)$ are symmetrical loss.

Concave-Convex Procedure

- Objective function: $Q^{\kappa}(\vartheta, \mathbf{d}) = Q_{vex}(\vartheta, \mathbf{d}) + Q^{\kappa}_{cav}(\vartheta)$

Taylor Expansion

- Each step

$$\boldsymbol{\vartheta}^{t+1} = \arg\min_{\boldsymbol{\vartheta}} \left(Q_{vex}(\boldsymbol{\vartheta}, \mathbf{d}^t) + \frac{\partial Q^{\kappa}_{cav}(\boldsymbol{\vartheta}^t)}{\partial \boldsymbol{\vartheta}} \cdot \boldsymbol{\vartheta} \right),$$

Lagragian Multiplier

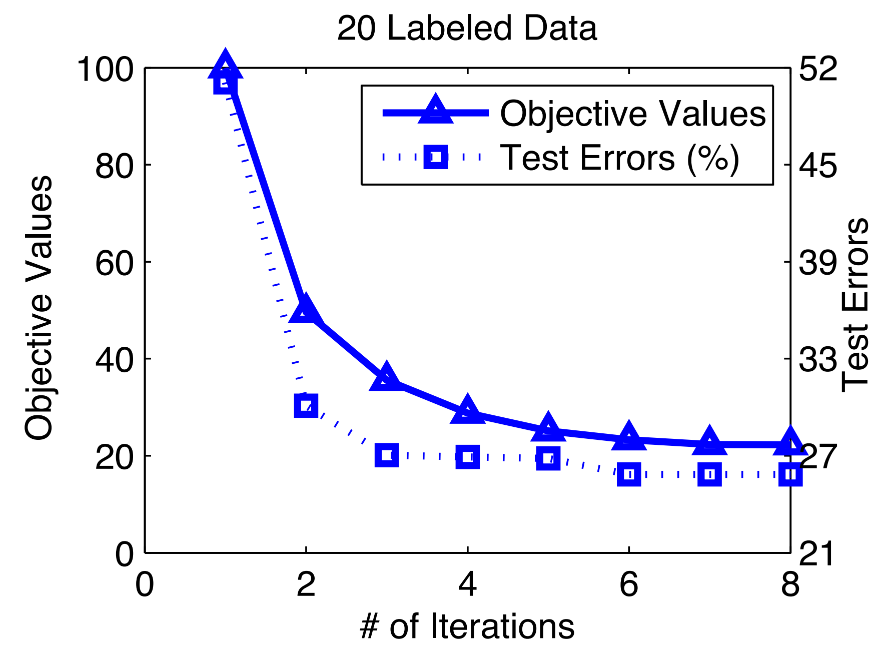
$$\overset{\text{Dual}}{\underset{\text{QP}}{\Longleftrightarrow}} \begin{cases} \max\limits_{\boldsymbol{\alpha}, \boldsymbol{\alpha}^*} & -\frac{\lambda}{2} \|\mathbf{w}(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*)\|^2 + \boldsymbol{\varrho}(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*) \\ \text{s.t.} & \mathbf{A}_e[\boldsymbol{\alpha}; \boldsymbol{\alpha}^*] = \boldsymbol{\mu}^T \mathbf{Y}_{\bullet U}, \\ & \mathbf{A}[\boldsymbol{\alpha}; \boldsymbol{\alpha}^*] \le \mathbf{0}, \\ & \mathbf{0} \le \boldsymbol{\alpha}, \boldsymbol{\alpha}^* \le \mathbf{r}. \end{cases}$$

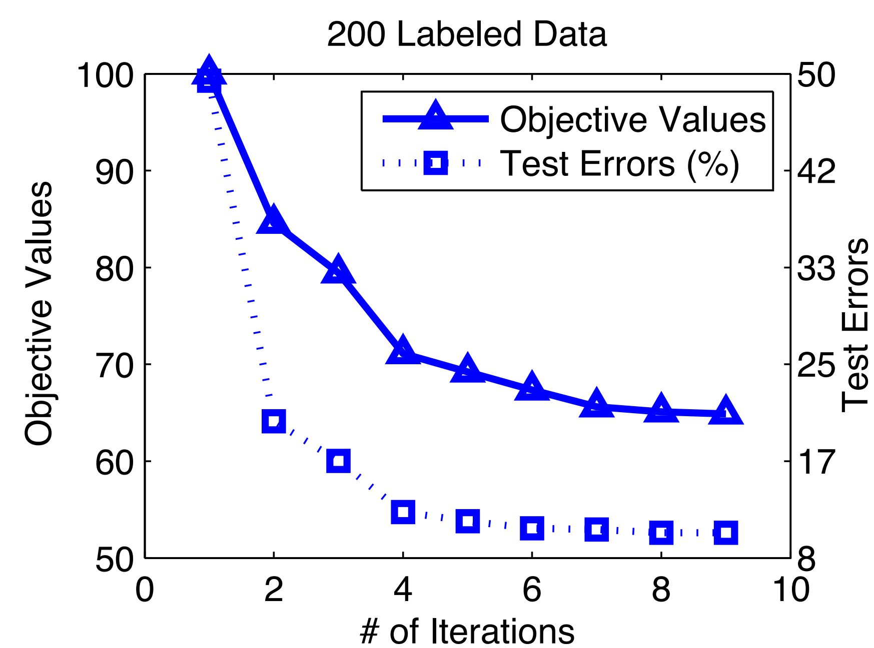
$$d_k = \begin{cases} 1 & \text{if} \quad \xi_k \le \xi_k^* \\ 0 & \text{otherwise} \end{cases}, \quad \begin{array}{l} \xi_k = H_1(|f_{\boldsymbol{\vartheta}}(\mathbf{x}_{k+L})|), \\ \xi_k^* = I_{\varepsilon}(|f_{\boldsymbol{\vartheta}}(\mathbf{x}_{k+L})|), \quad k = 1, \ldots, U. \end{array}$$

Algorithm

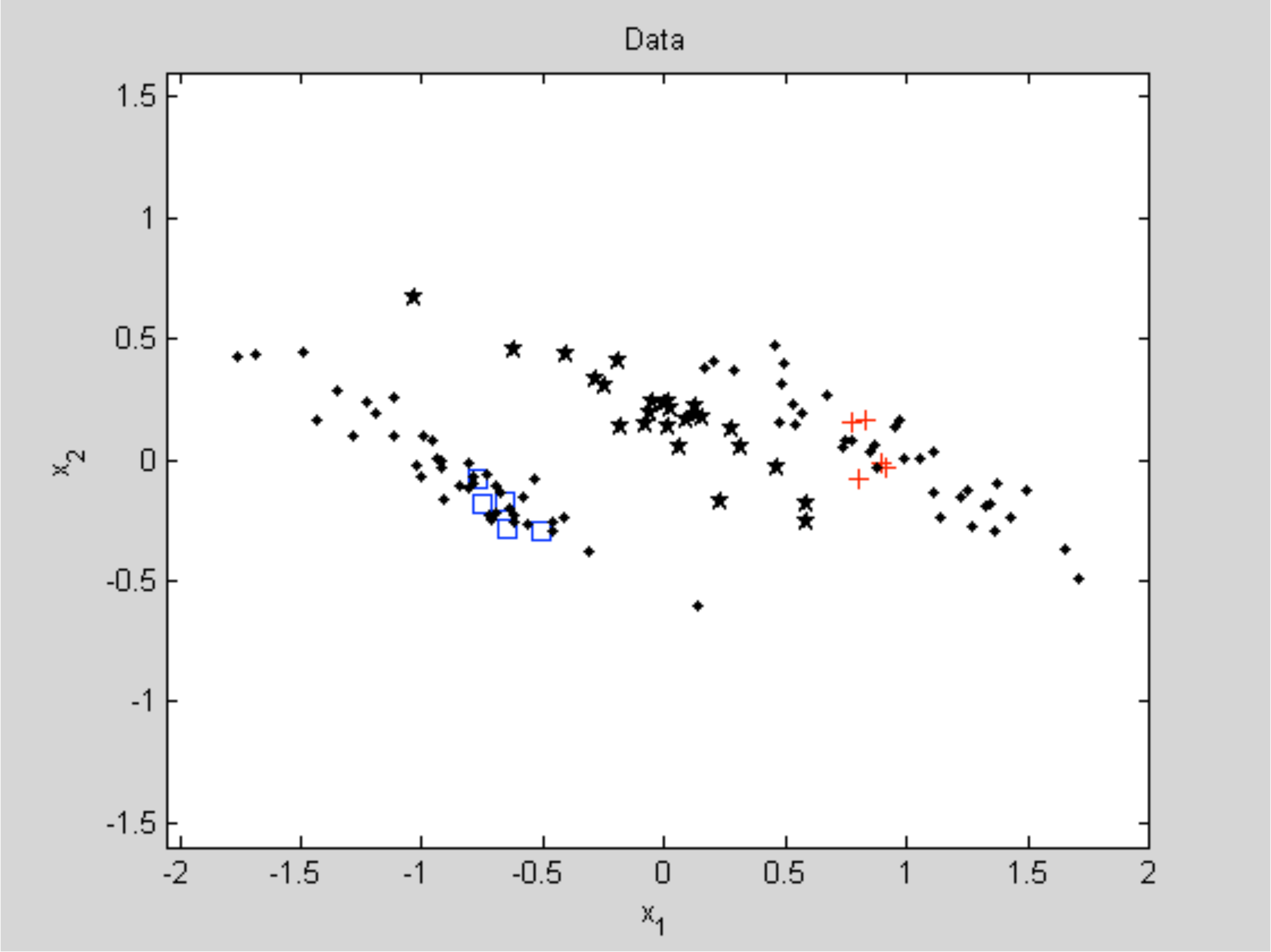
Algorithm 1 CCCP for 3C-SVMs

Initialization:
 $t = 0$;
 Calculate $\vartheta^0 = (\mathbf{w}^0, b^0)$ from a $\mathcal{U}$-SVM solution on the labeled/unlabeled data;
Compute
$$\mu_i^0 = \begin{cases} r_i & \text{if } \ y_i f_{\vartheta^0}(\mathbf{x}_i) < \kappa \text{ and } i \geq L+1 \\ 0 & \text{otherwise} \end{cases};$$
repeat
 $t \leftarrow t + 1$;
 Solve the optimization in (6) to obtain ϑ^t;
 Update $\mathbf{d}^t$ from (4);
 Update $\boldsymbol{\mu}^t$ from (5);
 if $Q^\kappa(\vartheta^t, \mathbf{d}^t) > Q^\kappa(\vartheta^{t-1}, \mathbf{d}^{t-1})$ **then**
  Let $\mathbf{d}^t = \mathbf{d}^{t-1}$;
  Solve the optimization in (6) to obtain ϑ^t by fixing $\mathbf{d}^{t-1}$;
  Update $\boldsymbol{\mu}^t$ from (5);
 end if
until $|\boldsymbol{\mu}^{t+1} - \boldsymbol{\mu}^t| \leq \epsilon$.

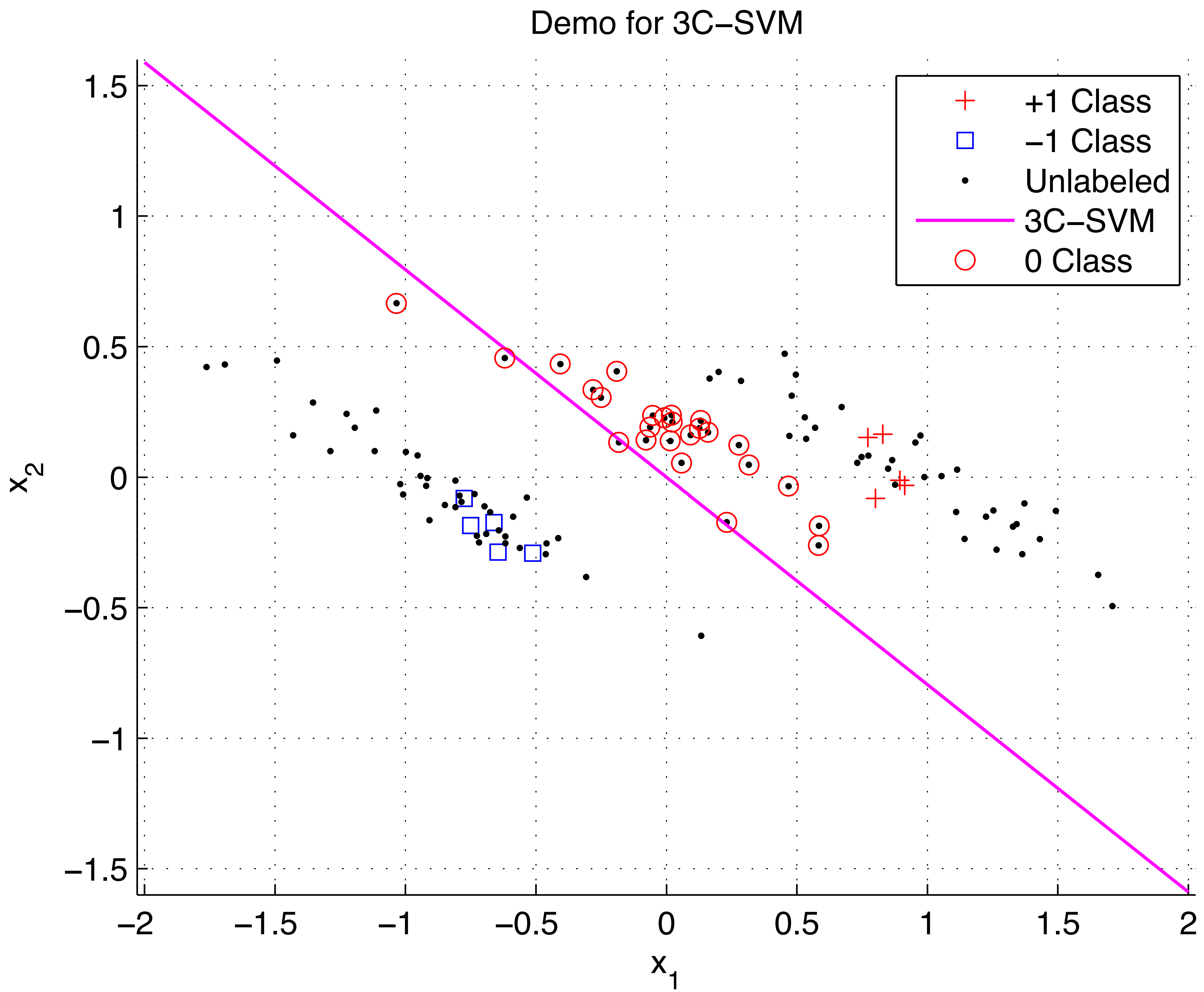




3C-SVM Demo



3C-SVM Result



Experimental Set-up

Comparing Algorithms

- SVMs
- S^3VMs
- *U*-SVMs
- 3C-SVMs

Platform Used

- Matlab 7.3
- MOSEK 5.0

Data Generation

- Follow scheme from Sinz et al., 2008.
- ± 1-class: $c_i^{\pm} = \pm 0.3$, $i = 1, \ldots, 50$, $\sigma_{1,2}^2 = 0.08$, $\sigma_{3,\ldots,50}^2 = 10$.
- Two Gaussians with the Bayes risk being approximately 5%.
- First $\mathcal{U}_0$: zero mean, $\sigma_{1,2}^2 = 0.1$, $\sigma_{3,\ldots,50}^2 = 10$.
- Second $\mathcal{U}_0$: variance values are the same as ± 1-class data, mean is $t \cdot \mathbf{c}^+$, $t = 0.5$.

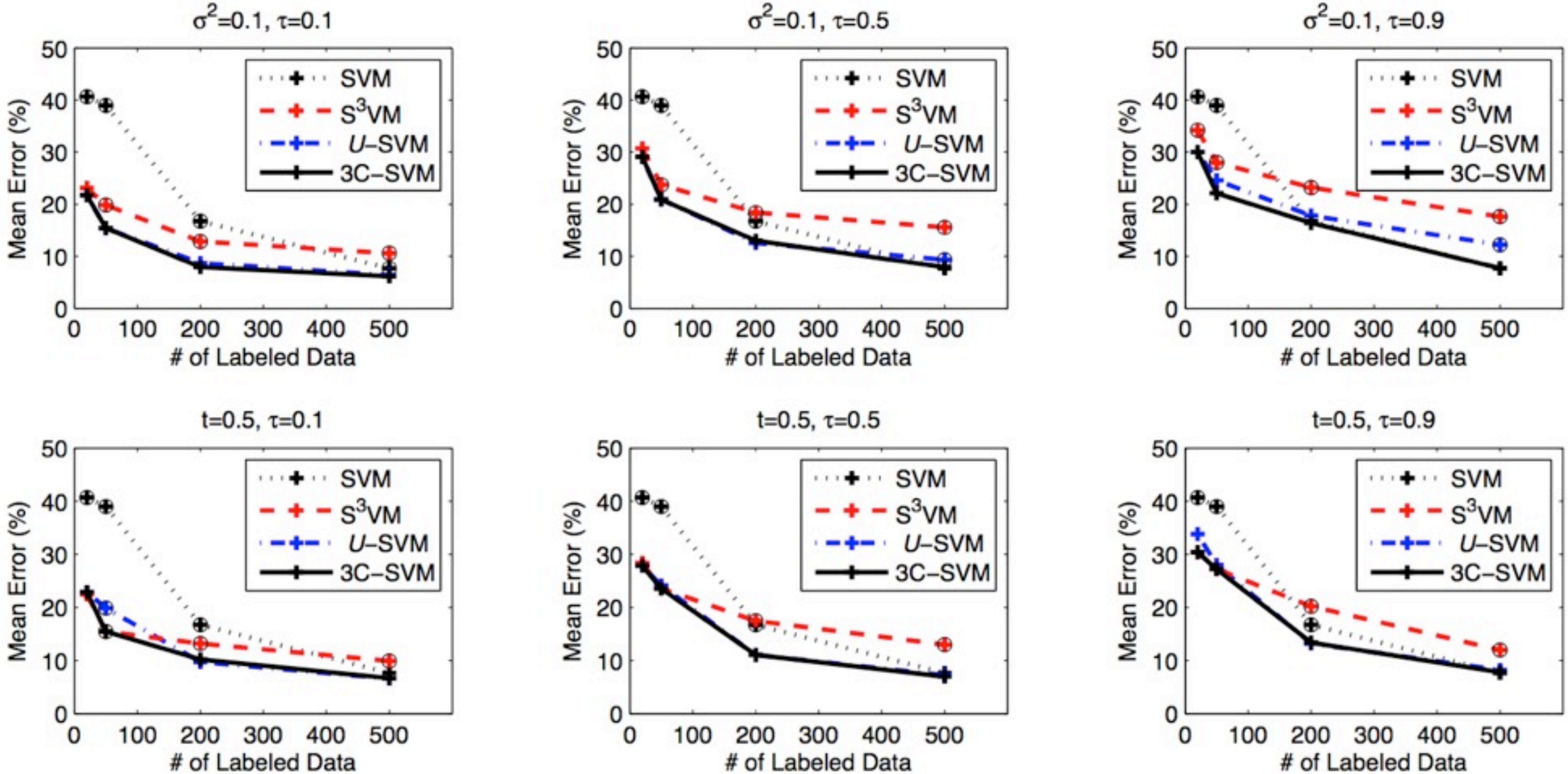
Test Procedure

- $L = 20, 50, 200, 500$
- $U = 500 = (\tau U, (1-\tau)U)$, $\tau = 0.1, 0.5, 0.9$
- Labeled + Unlabeled/500 Test, ten-run average
- Hyperparameters
  - Linear kernel
  - Regularized parameters, forward tuning

	$C_{\mathcal{L}}$	$C_{\mathcal{U}}$	ε	κ
SVM	$\checkmark$	$\times$	$\times$	$\times$
$\mathcal{U}$-SVM	–	$\checkmark$	$\checkmark$	$\times$

  - Further tune on S^3VM
  - 3C-SVM uses the same parameters of other models

Accuracy



σ²=0.1, τ=0.1
σ²=0.1, τ=0.5
σ²=0.1, τ=0.9
t=0.5, τ=0.1
t=0.5, τ=0.5
t=0.5, τ=0.9
SVM
S³VM
U–SVM
3C–SVM
Mean Error (%)
# of Labeled Data

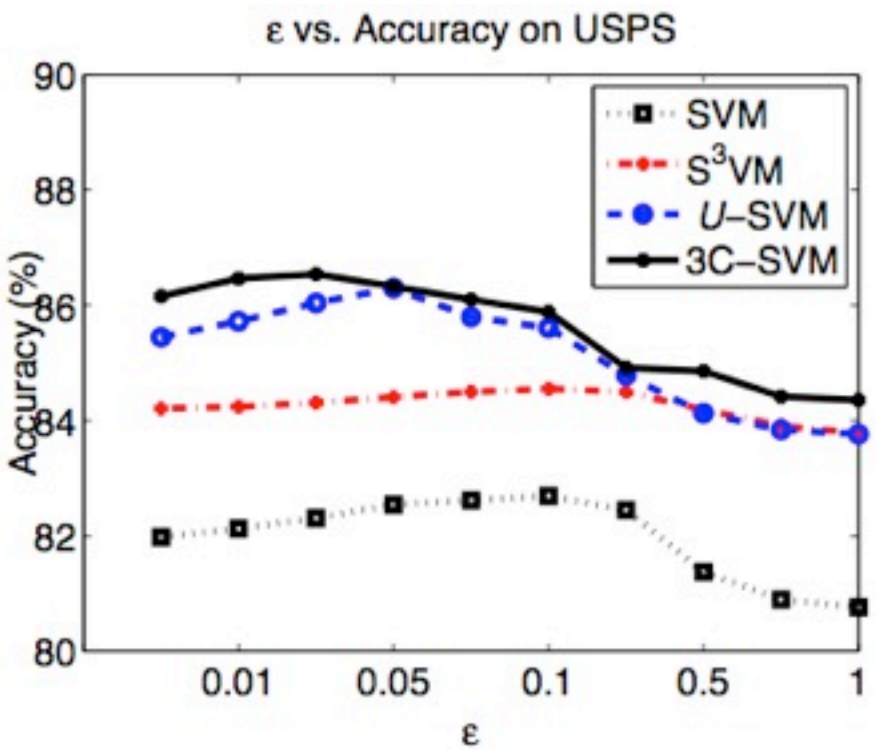
Real-World Data Description

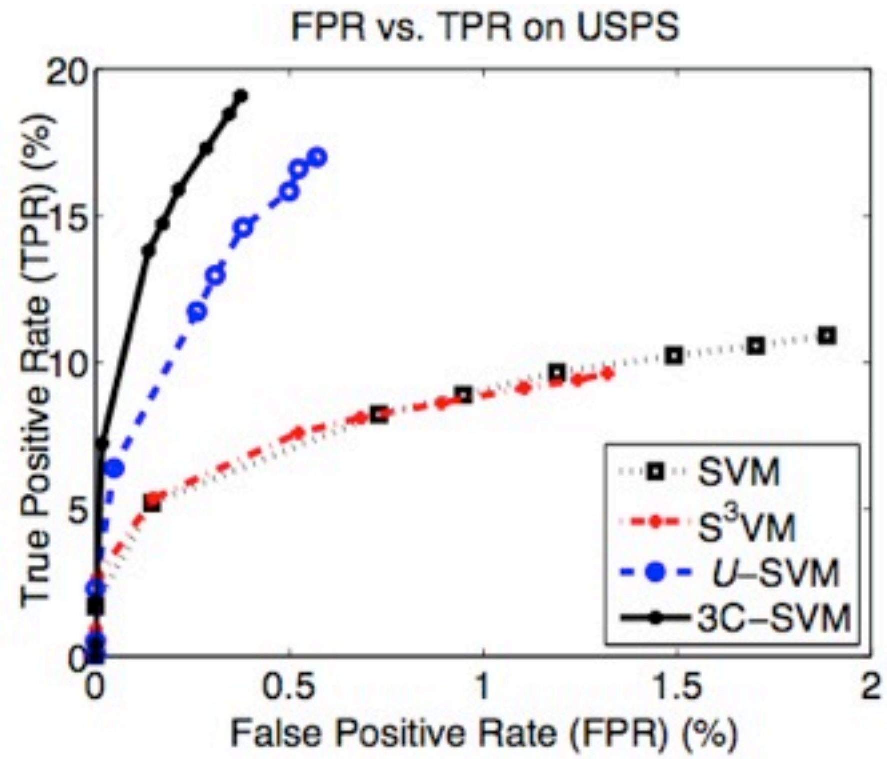
- Datasets:
  - Small size: USPS
  - Large size: MNIST
- Setup
  - ± 1-class: Digits "5" and "8"
  - $\mathcal{U}_0$: Other digits
  - $L = 20$
  - $U = 500 = (\tau U, (1-\tau)U)$, $\tau = 0.1, 0.5, 0.9$
  - RBF kernel: $K(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2)$, $\gamma = \frac{1}{0.3d}$
  - Other hyperparameters are set similar to those in the synthetic datasets

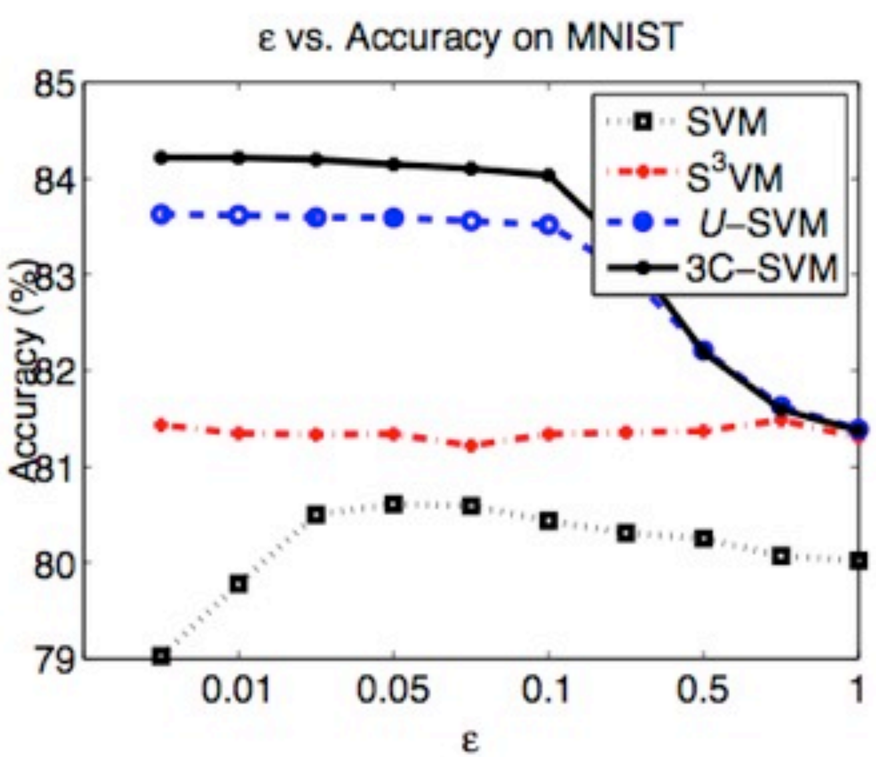
Accuracy Results

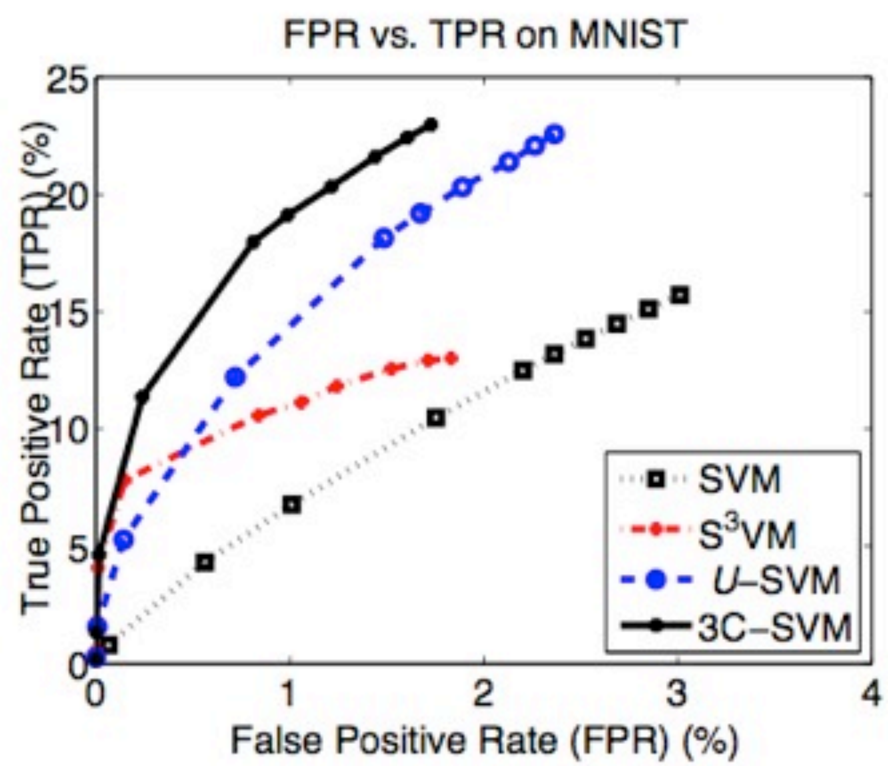
Dataset	Algorithm	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$
USPS	SVM	72.4 ± 15.9 (**0.7**)	72.4 ± 15.9 (**9.5**)	72.4 ± 15.9 (53.1)
	S^{3}VM	63.6 ± 8.9 (**0.0**)	68.2 ± 8.0 (**2.2**)	73.2 ± 7.0 (**9.5**)
	$\mathcal{U}$-SVM	83.1 ± 2.5 (**0.0**)	73.4 ± 4.4 (**0.0**)	64.2 ± 3.6 (**0.0**)
	3C-SVM	**87.2**± 2.3	**80.6**± 4.8	**75.4**± 7.3
MNIST	SVM	70.9 ± 11.4 (**0.3**)	70.9 ± 11.4 (**0.8**)	70.9 ± 11.4 (13.6)
	S^{3}VM	70.9 ± 10.5 (**0.7**)	72.4 ± 10.1 (**1.0**)	75.7 ± 9.1 (**9.8**)
	$\mathcal{U}$-SVM	84.2 ± 2.2 (**0.2**)	80.0 ± 4.6 (**0.9**)	75.0 ± 3.9 (**1.0**)
	3C-SVM	**85.3**± 1.6	**82.8**± 2.9	**77.6**± 3.9

Accuracy on Detecting 0-class









Conclusions

- A novel maxi-margin classifier, 3C-SVM, can distinguish data into -1, +1, and 0, three categories.
- The model incorporates standard SVMs, S^3VMs, and U-SVMs as specific cases.
- Introduce a min-loss function which combines loss for relevant and irrelevant data.
- Present transformations so it can be solved by CCCP, a high efficiency algorithm.
- Effectiveness and efficiency are demonstrated through synthetic and real-world data.

Future Work

- Model speed-up
- Multi-class extension
- Theoretical analysis, generalization bound, etc.

References

1. Chapelle, O., Schölkopf, B., and Zien, A. (Eds.). *Semi-supervised learning.* Cambridge, MA: MIT Press. 2006

2. Collobert, R., Sinz, F., Weston, J., and Bottou, L. Large scale transductive svms. *Journal of Machine Learning Research*, *7*, 1687–1712. 2006.

3. Weston, J., Collobert, R., Sinz, F. H., Bottou, L., and Vapnik, V. Inference with the universum. In ICML'06, 1009–1016. 2006.

4. Vapnik, V., and Kotz, S. *Estimation of dependences based on empirical data: Empirical inference science (information science and statistics).* Secaucus, NJ, USA: Springer-Verlag New York, Inc. 2nd edition. 2006.

5. Sinz, F. H., Chapelle, O., Agarwal, A., and Schölkopf, B. An analysis of inference with the universum. In *Proceedings of the 21th Neural Information Processing Systems Conference*, 1369–1376. Cambridge, MA, USA: MIT Press. 2008.

On-Going Research

Machine Learning

- Heavy-Tailed Symmetric Stochastic Neighbor Embedding (NIPS'09)
- Adaptive Regularization for Transductive Support Vector Machine (NIPS'09)
- Direct Zero-norm Optimization for Feature Selection (ICDM'08)
- Semi-supervised Learning from General Unlabeled Data (ICDM'08)
- Learning with Consistency between Inductive Functions and Kernels (NIPS'08)
- An Extended Level Method for Efficient Multiple Kernel Learning (NIPS'08)
- Semi-supervised Text Categorization by Active Search (CIKM'08)
- Transductive Support Vector Machine (NIPS'07)
- Global and local learning (ICML'04, JMLR'04)

On-Going Research

Web Intelligence/Information Retrieval

- A Generalized Co-HITS Algorithm and Its Application to Bipartite Graphs (KDD'09)
- Entropy-biased Models for Query Representation on the Click Graph (SIRIR'09)
- Effective Latent Space Graph-based Re-ranking Model with Global Consistency (WSDM'09)
- Formal Models for Expert Finding on DBLP Bibliography Data (ICDM'08)
- Learning Latent Semantic Relations from Query Logs for Query Suggestion (CIKM'08)
- RATE: a Review of Reviewers in a Manuscript Review Process (WI'08)
- MatchSim: link-based web page similarity measurements (WI'07)
- Diffusion rank: Ranking web pages based on heat diffusion equations (SIGIR'07)
- Web text classification (WWW'07)

On-Going Research

Recommender Systems/Collaborative Filtering

- Learning to Recommend with Social Trust Ensemble (SIRIR'09) Semi-Nonnegative Matrix Factorization with Global Statistical Consistency in Collaborative Filtering (CIKM'09)
- Recommender system: accurate recommendation based on sparse matrix (SIGIR'07)
- SoRec: Social Recommendation Using Probabilistic Matrix Factorization (CIKM'08)

Human Computation

- An Analytical Study of Puzzle Selection Strategies for the ESP Game (WI'08)
- An Analytical Approach to Optimizing The Utility of ESP Games (WI'08)

Acknowledgments

- Prof. Michael R. Lyu
- Dr. Hongbo Deng (Ph.D.)
- Baichuan Li (Ph.D.)
- Zhenjiang Lin (Ph.D.)
- Hao Ma (Ph.D.)
- Mingzhen Mo (M.Phil.)
- Dingyan Wang (M.Phil.)
- Wei Wang (M.Phil.)
- Haiqin Yang (Ph.D.)
- Connie Yuen (Ph.D.)
- Xin Xin (Ph.D.)
- Chao Zhou (Ph.D.)
- Yi Zhu (Ph.D.)

VeriGuide

- Similarity text detection system
- Developed at CUHK
- Promote and uphold academic honesty, integrity, and quality
- Support English, Traditional and Simplified Chinese
- Handle .doc, .txt, .pdf, .html, etc. file formats
- Generate detailed originality report including readability

VeriGuide Free Trial

http://www.cse.cuhk.edu.hk/~king

Q & A

http://www.cse.cuhk.edu.hk/~king